Learning-integrated Methods for Spatial Audio Capture and Reproduction

Xingyu Chen

A thesis submitted for the degree of

Master of Philosophy

The Australian National University

Research School of Engineering

Australian National University

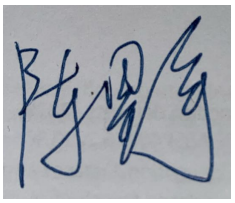
July 2024

To family, friends, and the wonderful universe we are intimately connected with

Disclaimer

I hereby declare that the work undertaken in this thesis has been conducted by me alone, except where indicated in the text. I conducted this work between August 2022 and June 2024, during which period I was an MPhil student at the Australian National University. This thesis, in whole or any part of it, has not been submitted to this or any other university for a degree. Part of this thesis is based on my publications which are submitted to international conferences.

- [1] **Chen, Xingyu**, Fei Ma, Amy Bastine, Prasanga Samarasinghe, and Huiyuan Sun. "Sound Field Estimation around a Rigid Sphere with Physics-informed Neural Network." In 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1984-1989. IEEE, 2023.
- [2] **Chen, Xingyu**, Fei Ma, Yile Zhang, Amy Bastine, and Prasanga N. Samarasinghe. "Head-Related Transfer Function Interpolation with a Spherical CNN." arXiv preprint arXiv:2309.08290 (2023).
- [3] **Chen, Xingyu**, Hanwen Bi, Wei-Ting Lai, and Fei Ma. "Monaural speech enhancement on drone via Adapter based transfer learning." In 2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE.



Xingyu Chen

13 May 2025

Acknowledgements

This work would not have been possible without the support of numerous colleagues, friends, and family members. I am deeply grateful to all those who contributed to my journey. Specifically, I would like to acknowledge and thank the following individuals:

- Foremost, my sincere thanks go to my supervisors for their invaluable guidance and mentorship.
- Fei Ma, for his continued encouragement and support, and for introducing me to the world of research.
- Prasanga Samarasinghe, for her guidance on the structure of my thesis, monitoring my progress, and consistently demonstrating rock-solid reliability.
- Aimee Zhang, for providing valuable suggestions.
- My fellow students in the Audio and Acoustics Signal Processing Group for their collaboration.
- Amy Bastine and Huiyuan Sun, for teaching me algorithms and providing meticulous feedback and suggestions on my papers.
- Wei-ting, Angela, and Frank, for their engagement in many enriching technical discussions.
- My family and friends, for their eternal love, motivation, and unwavering support.

Abstract

This thesis presents a study on advancing spatial audio technologies using deep learning methods, with a focus on three major areas: sound field estimation, Head-Related Transfer Function(HRTF) interpolation, and monaural speech enhancement on drones.

Spatial audio is integral to immersive applications such as entertainment, teleconferencing, and AR/VR applications including gaming and defense applications. However, challenges persist due to technological constraints like the limited spatial resolution of microphone arrays, which impede the full capture and reproduction of sound scenes. As a result, there is a notable drive to enhance spatial audio technologies to improve user experience.

The integration of deep learning in spatial audio is transformative, paralleling its impact in fields like image processing and computer vision. Deep learning's capability to manage complex, non-linear problems has started to revolutionize spatial audio, though its application in this field is still in the early stages of development.

This thesis introduces innovative learning-based methods to address challenges in spatial audio. It employs Physics-Informed Neural Networks for sound field estimation to achieve accurate sound pressure estimation. For HRTF interpolation, it presents a spherical convolutional neural network designed to enhance interpolation accuracy. Furthermore, the thesis explores the application of transfer learning to speech enhancement on drones, focusing on improving clarity and reducing drone noise.

The findings of this thesis not only contribute to the theoretical advancements in the field of spatial audio but also have practical implications for enhancing user experience in many immersive audio applications, including those using VR/AR techniques. It also enables future use of drone-embedded microphone arrays for spatial audio capture. This work contributes insights into the application of deep learning technologies in spatial audio and offers ideas for future studies to build on and improve these methods.

Contents

Chapter 1

Introduction

Spatial audio is an interdisciplinary field that integrates expertise from numerous domains including audio engineering, acoustics, computer science, applied psychoacoustics, and beyond. Spatial audio research aims to recreate or synthesize acoustic environments by adeptly combining sound recording, processing, and reproduction techniques. This involves preserving the fidelity and spatial attributes of the audio content—reflecting the actual locations of sound sources and the characteristics of the acoustic environment—and extends into diverse related disciplines. These encompass signal processing and machine learning for audio transformation, physics, fluid dynamics for modeling sound propagation, and electrical, mechanical, and materials engineering for designing audio hardware. Additionally, the fields of language and psychology play crucial roles in understanding human communication and the perception of sound. Consequently, spatial audio stands as a pivotal concept with broad relevance to both research and commercial industries, significantly impacting human auditory experiences and communication.

1.1 Background

Fig. 1.1 shows a general spatial audio pipeline consisting of three main stages: recording, processing, and reproduction of sound. In the recording phase, spatial audio often employs spherical microphone arrays to capture the sound from all directions, and may also use head-shaped microphones for binaural recordings. These techniques are designed to preserve the nuanced interactions of sound with the human head and ears, crucial for creating a realistic auditory experience.

During the processing stage, the spatial characteristics of the captured sound are either enhanced or manipulated to extract information that is not directly measurable. Techniques such as ambisonics are pivotal here; they encode sound sources within a spherical coordinate system, allowing for an immersive audio playback that remains consistent regardless of listener orientation. This phase also considers vari-

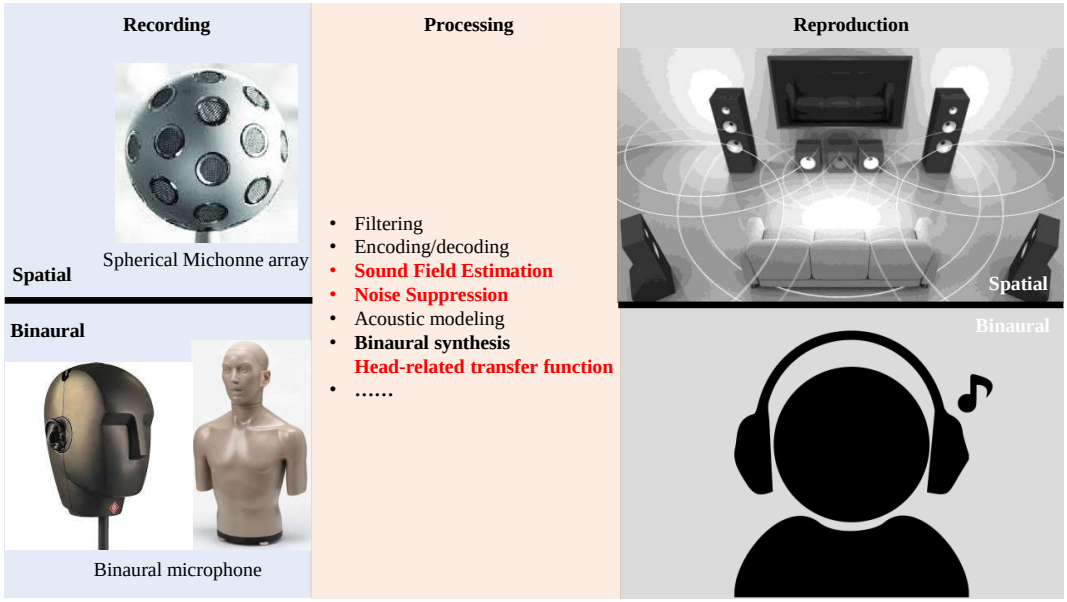


Figure 1.1: A general spatial audio pipeline

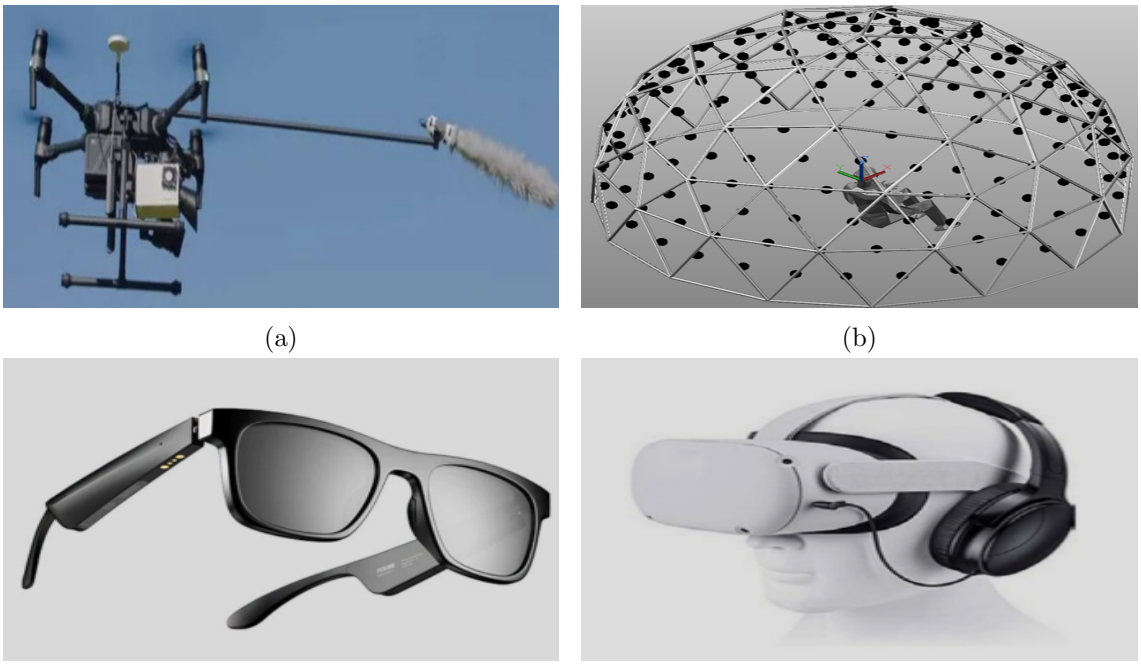
ous acoustic factors such as distance, direction, and environmental reflections, which are essential for accurately positioning sound sources within a virtual space.

Finally, the reproduction stage auralizes the processed sound scene through multiple speakers or headphones. This includes the application of the Head-Related Transfer Function (HRTF) to convert spatial recordings for headphone listening through binaural synthesis, effectively mimicking the natural hearing process. Such advanced reproduction methods transcend traditional stereo panning by reconstructing a full three-dimensional audio environment, significantly enhancing the depth and realism of the listener's auditory experience.

1.2 Motivation

Accurate spatial audio rendering requires precise sound field estimation and Head-Related Transfer Function (HRTF) interpolation. Traditional methods for sound field estimation, which measure the distribution of sound pressure or intensity, often face limitations due to complex array setups restricted by spatial resolution and environmental constraints. Additionally, HRTF interpolation, essential for personalized auditory experiences in virtual reality, confronts challenges from the high variability in individual ear and head anatomies that affect auditory perception.

Looking toward future advancements, the field of spatial audio is increasingly exploring unconventional and dynamic environments for sound capture and reproduction. Figure 1.2(a) illustrates aerial audio recording with a flying drone, presenting



(a) (b)

(c) (d)

Figure 1.2: Futuristic Technologies in Spatial Audio: (a) Aerial Audio Recording Scenario with a Flying Drone (b) High-Density Hemispherical Array of Loudspeakers for Enhanced Sound Field Control (c) Open Ear Style Smart Glasses (d) VR Devices

new possibilities and challenges for capturing audio in difficult-to-access or rapidly changing environments. Figure 1.2(b) features a High-Density Hemispherical Array of Loudspeakers, designed for precise sound field control. Additionally, Open Ear Style Smart Glasses and VR devices (Figures 1.2(c) and (d)) integrate spatial audio into wearable technologies, enhancing user experiences. Among these innovations, drones pose a unique challenge due to their noise characteristics, which can significantly impact the accuracy of sound recording.

This thesis explores three topics related to immersive audio: sound field estimation, Head-Related Transfer Function interpolation, and speech enhancement. Each topic addresses specific challenges, including data limitations and the integration of acoustic and audio domain knowledge with advanced deep learning models. For sound field estimation, this thesis proposes a method using Physics-Informed Neural Networks to achieve accurate sound pressure estimation. For HRTF interpolation, it introduces a spherical convolutional neural network to enhance the interpolation accuracy. Lastly, the thesis discusses the application of transfer learning to speech enhancement on drones, focusing on improving clarity and reducing drone noise. Although this task is monaural, it supports spatial audio workflows by enhancing signal quality for subsequent spatial encoding and reproduction, especially in dynamic recording environments.

1.3 Thesis Outline

The structure of this thesis is organized as follows:

Chapter 2 provides a foundation on spherical transformation, soundwave propagation, sound recording, and deep learning, setting the stage for the methods applied in subsequent chapters.

Chapter 3 presents a method for sound field estimation, emphasizing the application of physics-informed neural networks.

Chapter 4 presents a method for HRTF upsampling, introducing spherical convolution and spherical harmonic transformations to enhance spatial representation and interpolation accuracy.

Chapter 5 presents a method for speech enhancement on drones, exploring transfer learning, especially the adapter-based tuning method, to improve speech clarity and reduce drone noise.

Chapter 6 synthesizes the research findings, discusses the implications of this work, and outlines potential future directions for immersive audio research.

Chapter 2

Background

This chapter introduces the foundational concepts and methodologies essential for analyzing acoustic and audio research. It begins by explaining Cartesian and spherical coordinate systems for describing three-dimensional sound fields. It then explores the spherical harmonics transform for function decomposition and reconstruction on the sphere. The discussion extends to sound wave propagation and provides a general solution to the wave equation. It then covers spatial sound recording, highlighting the roles of the Head-Related Transfer Function and spatial microphone arrays. Finally, the chapter introduces deep learning concepts like supervised learning, neural networks, and transfer learning, which provide innovative tools to address challenges in this field.

2.1 Coordinate Systems

This thesis employs both Cartesian and spherical coordinate systems to describe the spatial attributes of soundfields within three-dimensional spaces, as shown in Fig. 2.1. In a spherical coordinate system, a point in space is defined by the position vector $\boldsymbol{x} = (r, \theta, \phi)$, where $r = \|\boldsymbol{x}\|$ represents the radial distance from the origin, with $\|\cdot\|$ indicating the Euclidean norm. The polar angle θ is measured from the vertical axis within the range $0 \leq \theta \leq \pi$, and the azimuthal angle ϕ is measured in the horizontal plane from the origin, where $0 \leq \phi < 2\pi$. These spherical coordinates are related to Cartesian coordinates (x, y, z) via

$$\begin{aligned} x &= r \sin\theta \cos\phi, \\ y &= r \sin\theta \sin\phi, \\ z &= r \cos\theta. \end{aligned} \tag{2.1}$$

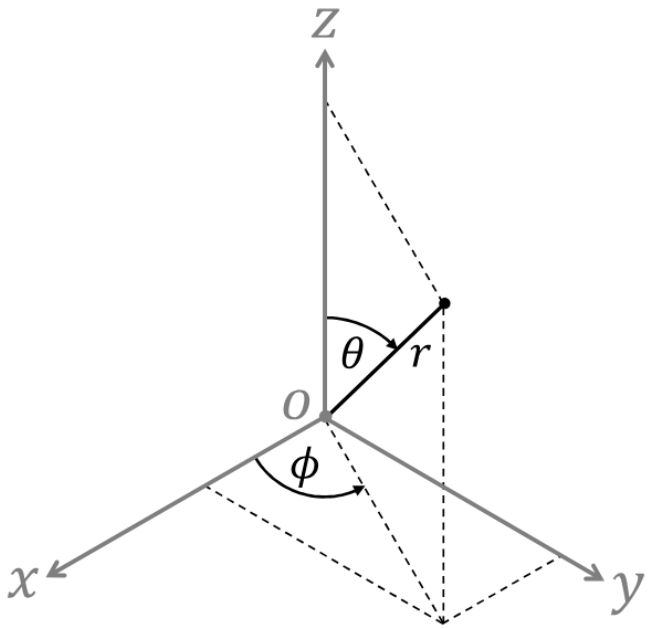


Figure 2.1: Spherical coordinate system defined relative to the Cartesian coordinate system

Vector directions can be expressed using the unit vector notation

$$\hat{\boldsymbol{x}} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}, \tag{2.2}$$

or alternatively, through the polar and azimuthal angles (θ, ϕ). This flexibility allows for the interchangeable use of position and direction in function parameterizations, where $f(\boldsymbol{x}) = f(r, \theta, \phi)$ and $g(\hat{\boldsymbol{x}}) = g(\theta, \phi)$.

2.2 Spherical Harmonics Transform

Spherical harmonics, $Y_n^m : S^2 \to \mathbb{C}$, form an orthonormal basis for $L^2(S^2)$, the space of square integrable functions on the sphere. The degree $m \in \mathbb{Z}$ and the order $n \in \mathbb{N}$ define the spherical harmonics. Any function $f : S^2 \to \mathbb{C}$ in $L^2(S^2)$ can be decomposed into this basis via the spherical harmonics transform (SHT) as shown in Eq. (2.3), and reconstructed via its inverse (ISHT) as shown in Eq. (2.4).

The SHT for spherical data, akin to a generalized Fourier Transform, projects functions defined on the sphere into the spherical harmonics space. The SHT for a function $g(\theta, \phi)$ is given by [4]:

$$\mathsf{a}_n^m = \int_0^{2\pi} \int_0^{\pi} g(\theta, \phi) \overline{Y_n^m(\theta, \phi)} \sin\theta \, d\theta \, d\phi, \tag{2.3}$$

where a_n^m are the spherical harmonic coefficients, $\overline{Y_n^m(\theta,\phi)}$ indicates the complex conjugate of the spherical harmonic function.

The inverse transformation reconstructs the original function from its spherical harmonic coefficients:

$$g(\theta,\phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \mathsf{a}_n^m Y_n^m(\theta,\phi), \tag{2.4}$$

This method enables synthesizing the function f on the sphere utilizing the coefficients derived from the SHT. For instance, one can capture the sound pressure over the entire sphere when employing a spherical microphone array. This application will be discussed in detail in Chapter 4.

2.3 Sound Wave Propagation in Free Field

This section covers the physics of sound waves in an idealized free field. A free field is defined as a special sound environment in a uniform and isotropic medium, typically air, where boundary reflections are absent, making conditions like those in an anechoic chamber or certain open regions ideal for such analysis.

2.3.1 Wave Equation

Sound waves generated by sound sources propagate in a space through a medium, creating a sound field. This field is a region within the medium where sound waves can be described by the spatial and temporal distribution of sound pressure $\hat{p}(\boldsymbol{x},t)$ in the time domain at location $\boldsymbol{x}$. In a source-free region, sound pressure in air satisfies the homogeneous wave equation [5]:

$$\nabla^2\hat{p} - \frac{1}{c^2}\frac{\partial^2\hat{p}}{\partial t^2} = 0, \tag{2.5}$$

where ∇^2 denotes the Laplacian operator and c denotes the speed of sound.

2.3.2 Helmholtz Equation

Assuming a time-harmonic solution, we have:

$$\hat{p}(\boldsymbol{x},t) = p(\boldsymbol{x},\omega)e^{i\omega t}, \tag{2.6}$$

where $p(x, \omega)$ is the space dependent part of the soundfield and $i = \sqrt{-1}$. Substitution of Eq. (2.6) into Eq. (2.5) yields the Helmholtz equation

$$\nabla^2 p(\boldsymbol{x}, k) + k^2 p(\boldsymbol{x}, k) = 0, \tag{2.7}$$

where k is the wavenumber, which is related to the angular frequency ω and wavelength λ through

$$k = \frac{\omega}{c} = \frac{2\pi}{\lambda}. \tag{2.8}$$

2.3.3 Spherical Harmonics Decomposition

Within a source-free region Ω, characterized by a radius R, the sound pressure at any point $\boldsymbol{x}$ relative to the origin Ω_O can be modeled using spherical harmonics. We can solve the Helmholtz equation Eq. (2.7), using spherical coordinates to describe the space-dependent acoustic pressure [6]:

$$p(\boldsymbol{x}, k) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \alpha_n^m(k) j_n(kr) Y_n^m(\theta, \phi), \tag{2.9}$$

where $j_n(kr)$ denotes the n-th order spherical Bessel function of the first kind.

Sound pressure can be determined by solving either the wave equation in the time domain, the Helmholtz equation in the frequency domain, or using spherical harmonics decomposition in the spatial domain, depending on the initial and boundary conditions.

2.4 Boundary Conditions for Sound Pressure

While sound wave propagation in a free field represents an ideal scenario, most practical situations involve reflective boundaries. At a receiver position, sound waves are a composite of direct sound from the source and reflections from boundaries. Boundary conditions are crucial for accurately modeling acoustic wave propagation, particularly at the domain's boundaries. These conditions can vary: Dirichlet conditions specify the sound pressure; Neumann conditions involve the derivative of sound pressure; and Robin conditions incorporate elements of both [5].

Sound Pressure at a Rigid Boundary

A key scenario in acoustic analysis is the interaction of sound waves with rigid bodies, such as spheres, where specific boundary conditions apply. At a rigid boundary, no

air particle displacement through the surface is allowed, leading to a Neumann boundary condition [7]:

$$\frac{\partial p}{\partial v}(\boldsymbol{x}, k) = 0 \quad \text{for} \quad \boldsymbol{x} \in S, \tag{2.10}$$

where S denotes the surface of a rigid sphere, $\frac{\partial p}{\partial v}$ represents the derivative of the sound pressure in the direction normal (perpendicular) to the surface of the sphere. This condition ensures that the normal component of the velocity at the boundary is zero, crucial for studying scattering phenomena and resonance effects.

The presence of the rigid sphere alters the sound field, creating patterns of interference and diffraction that can be precisely analyzed using the Helmholtz equation in conjunction with the specified boundary condition.

2.5 Spatial Sound Recording

Sound recording involves capturing sound waves using microphones, which convert these mechanical vibrations into electrical signals. However, a single microphone alone cannot capture the complex spatial nature of real-world acoustic environments. This section introduces spatial sound recording techniques including spherical microphone arrays and noise reduction for recording.

2.5.1 Recording with Spherical Microphone Arrays

Spherical microphone arrays provide the flexibility to account for head rotation and personalized head effects. These arrays utilize the higher order ambisonics format, which encodes the sound field using a weighted sum of spatial basis functions, often referred to as higher order spherical harmonics. This technique allows for comprehensive capture, processing, and reproduction of sound across various acoustic environments.

In Eq. (2.9), the summation of order up to an infinite order is not practical. Hence, a truncation of this summation at a maximum order N is common, due to the low pass nature of the spherical Bessel functions [8]. Thus, Eq. (2.9) can be approximated as

$$p(\boldsymbol{x}, k) \approx \sum_{n=0}^{N} \sum_{m=-n}^{n} \alpha_n^m(k) j_n(kr) Y_n^m(\theta, \phi), \tag{2.11}$$

where $N = \lceil kr \rceil$ represents the truncation order [8], with $\lceil \cdot \rceil$ denoting the floor function, which rounds down to the nearest integer. This truncation is essential for practical computations as it limits the infinite series to a finite number, reducing complexity and computational demands while maintaining sufficient accuracy

for modeling the sound field within the defined frequency and spatial resolution limits.

For a rigid spherical array of radius r_0, $\alpha_{nm}(k)$ can be extracted by integrating the recording over the spherical surface, which gives

$$\alpha_n^m(k) = \frac{1}{b_n(kr_0)} \int_0^{2\pi} \int_0^{\pi} p(r_0, \theta, \phi, k) \overline{Y_n^m}(\phi, \theta) \sin\phi \, d\phi \, d\theta. \tag{2.12}$$

where

$$b_n(kr) = j_n(kr) - \frac{j_n'(kr_0)}{h_n'(kr_0)} h_n(kr), \tag{2.13}$$

where $h_n(\cdot)$ is the spherical Hankel function, $(\cdot)'$ refers to the first order derivative. In practice, this integration is achieved using an equivalent discrete summation of spatial samples over the sphere, where at least $Q = (N+1)^2$ samples are necessary. We rewrite it as

$$\alpha_n^m(k) \approx \frac{1}{b_n(kr_0)} \sum_{q=1}^{Q} p(r_0, \theta_q, \phi_q, k) \overline{Y_n^m}(\theta_q, \phi_q) \mathcal{X}(q). \tag{2.14}$$

where θ_q, ϕ_q is the elevation and azimuth angle of the q-th microphone, respectively, and $\mathcal{X}(q)$ is the weighting coefficients specific to the sampling scheme of the microphone array. For regularly distributed sampling, $\mathcal{X}(q) = 1$ for all the microphone samples.

2.5.2 Noise Reduction

One of the challenges in sound recording is the presence of unwanted ambient noise which can obscure or distort the sound signal intended to be captured. The microphone recording can be represented as

$$y(t) = s(t) * h(t) + v(t), \tag{2.15}$$

where $s(t)$ denotes the clean sound with t being the time index, $h(t)$ is the impulse response between the human and the microphone, $v(t)$ denotes the noise, and $*$ denotes the convolution operation. Noise reduction methods are therefore employed to enhance the clarity of the recordings. These techniques involve various signal processing methods that identify and mitigate noise components from the captured audio. By improving the signal-to-noise ratio, these processes ensure that the true essence of the sound, including its spatial characteristics as captured by the microphones, is preserved.

2.6 Binaural Reproduction

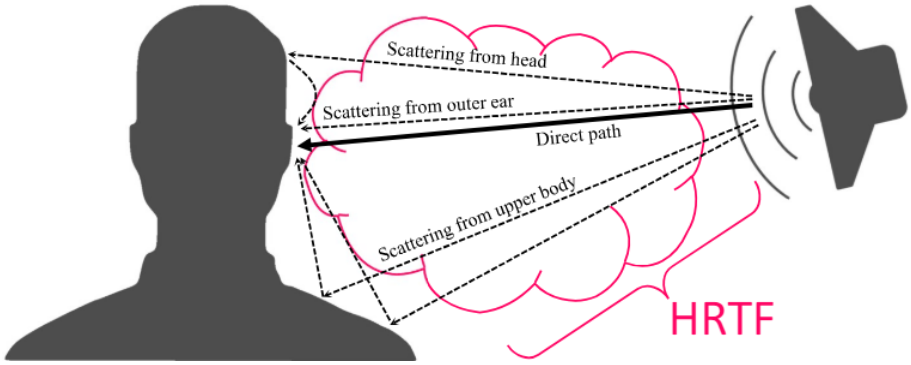


Figure 2.2: Illustration of the HRTF, which describes how sound is transformed from the sound source before reaching the listener's eardrum by reflecting off their upper body, head, and outer ear

Human auditory perception operates through a complex mechanism involving two ears. Ears interact with various anatomical structures, including the head, pinnae, and torso, which significantly scatter and diffract incoming sound waves. These interactions are crucial as they modulate the sound waves reaching each ear, imparting essential spatial information that facilitates directional hearing. This complex interaction is described by the Head-Related Transfer Function (HRTF), which quantifies the overall filtering effects these anatomical features impose on incoming sounds [9].

HRTFs are individually specified for each ear through the following

$$H_L = H_L(r_s, \theta_s, \phi_s, f, a) = \frac{P_L(r_s, \theta_s, \phi_s, f, a)}{P_{\text{free}}(r_s, f)}, \tag{2.16}$$

$$H_R = H_R(r_s, \theta_s, \phi_s, f, a) = \frac{P_R(r_s, \theta_s, \phi_s, f, a)}{P_{\text{free}}(r_s, f)}, \tag{2.17}$$

where (r_s, θ_s, ϕ_s) denotes the sound source position relative to the listener; f is frequency; P_L and P_R are sound pressures at the left and right ears, respectively; and P_{free} is the sound pressure in a free field at the head's center when the head is absent. Each HRTF is inherently unique to an individual, reflecting the specific anatomical structures and dimensions of each person, represented by the parameter a. At the far-field distance with $r_s > 1.0$ m, an HRTF is nearly distance independent and termed the far-field HRTF. In this thesis, we consider far-field HRTF.

Recording with head-shaped microphones involves placing two microphones at the ear positions of a human-like dummy to capture binaural signals, which inherently include the dummy's HRTF. This technique allows for direct playback through head-

phones, providing listeners with an auditory experience as if they were present at the original recording scene. However, this method inherently assumes a generic auditory perspective, aligning the listener's experience with that of the dummy's head orientation during the recording. Additionally, it fails to account for individual variations in HRTF, which can significantly affect the realism and personalization of the audio experience.

To overcome these limitations, a more flexible approach involves using spherical microphone arrays for spatial recording. This method not only captures sound from all directions but also allows for post-processing adjustments to integrate personalized HRTF data. Such a process enables the tailoring of the auditory experience to individual listeners, enhancing the overall realism and immersion. Consequently, while direct binaural recording with dummy heads offers immediate and straightforward playback, the integration of spherical microphone recordings with personalized HRTF provides a superior and more adaptable spatial audio solution.

2.7 Deep Learning

This section introduces fundamental deep learning concepts. Deep learning is a subset of machine learning where artificial neural networks learn from large amounts of data. Deep learning can be thought of as a way to automatically learn to recognize patterns and make decisions based on data. The capability of deep learning models to perform feature extraction progressively increases their complexity and abstraction, making them highly effective for large-scale and complex tasks.

2.7.1 Supervised Learning

This thesis focuses on supervised learning, which involves learning a function from labeled training data by defining a mapping $\mathtt{y} = \mathtt{F}(\mathtt{x}; \Phi)$. The model learns to approximate this function through the parameters Φ [10].

The process involves several key steps. Initially, a dataset comprising input features and their corresponding target labels is compiled. Subsequently, the choice of model is important and depends on the nature of the task and the data available. The model is then trained using back-propagation to optimize the parameters and minimize the discrepancies between predicted and actual values. After training, the model is evaluated on a validation set to assess its performance on new, unseen data, ensuring that it can reliably generalize the patterns it has learned.

A loss function, such as mean squared error for regression tasks or cross-entropy

loss for classification, is employed to assess the accuracy of the model's predictions against the actual labels. The optimization process, typically executed using algorithms like gradient descent, is crucial for training the model. The optimization process is

$$L(\Phi) = \frac{1}{\mathtt{N}} \sum_{i=1}^{\mathtt{N}} \ell(\mathtt{F}(\mathtt{x}^{(i)}; \Phi), \mathtt{y}^{(i)}), \tag{2.18}$$

where $L(\Phi)$ represents the loss function, $\mathtt{N}$ is the number of samples in the dataset, ℓ is the loss per sample, $\mathtt{F}$ is the model parameterized by Φ, $\mathtt{x}^{(i)}$ are the input features, and $\mathtt{y}^{(i)}$ are the true labels. It aims to minimize the loss function to enhance the model's accuracy.

Backpropagation calculates the gradient of the loss function for each model parameter. These gradients are then used to update the parameters using [11]:

$$\Phi \leftarrow \Phi - \eta \nabla_{\Phi} L(\Phi), \tag{2.19}$$

where η is the learning rate and $\nabla_{\Phi} L(\Phi)$ represents the gradient of the loss function for the parameters Φ.

2.7.2 Fully Connected Neural Networks

Fully Connected Neural Networks, also known as Multilayer Perceptrons, are a fundamental type of neural network in deep learning where each neuron in one layer is connected to every neuron in the subsequent layer, as illustrated in Figure 2.3. The structure of a fully connected layer can be represented as:

$$\mathbf{a}^{(l+1)} = \delta(\mathbf{W}^{(l)}\mathbf{a}^{(l)} + \mathbf{b}^{(l)}), \tag{2.20}$$

where $\mathbf{a}^{(l)}$ is the vector of activations from the previous layer, $\mathbf{W}^{(l)}$ represents the weight matrix for layer l, $\mathbf{b}^{(l)}$ is the bias vector, and δ denotes the activation function applied element-wise. Figure 2.3 illustrates a typical fully connected layer.

Activation functions introduce non-linearity, allowing the network to learn complex patterns without altering the scale of the input. One of the most popular activation functions used in these networks is the Rectified Linear Unit (ReLU), defined as [12]:

$$\mathrm{ReLU}(\mathtt{x}) = \max(0, \mathtt{x}). \tag{2.21}$$

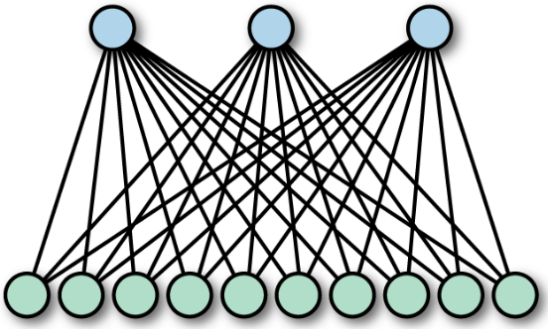


Figure 2.3: fully connected layer

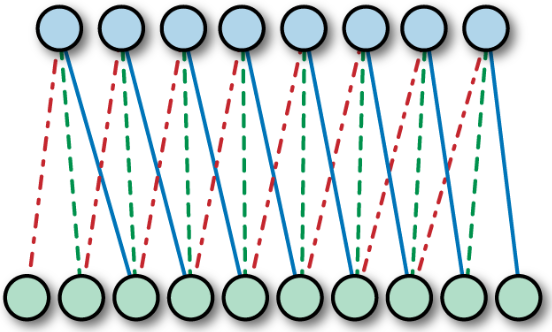


Figure 2.4: convolutional layer

2.7.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a class of deep neural networks designed to process data that has a grid-like topology, such as images or time series data. Time series data is treated as a 1-dimensional grid sampled at regular intervals, and image data as a 2-dimensional grid of pixels.

The architecture of CNNs fundamentally differs from that of fully connected networks by the pattern of connections between layers. Unlike fully connected networks where each neuron is connected to every neuron in the previous layer, neurons in a CNN are connected only to a local region of the previous layer, utilizing the same weights across the entire layer. This setup, known as convolution, involves applying a small window, or filter, across the input data, as illustrated in Figure 2.4. This operation extracts local feature representations from the input.

The effectiveness of CNNs is most apparent when processing two-dimensional spatial data, such as images [13]. The convolutional operation can be represented as:

$$I * \mathtt{k}(\mathtt{x}) = \sum_{\mathtt{y} \in \mathbb{Z}^2} I(\mathtt{y})\mathtt{k}(\mathtt{x} - \mathtt{y}), \tag{2.22}$$

where I represents the input image on the 2D plane $\mathbb{Z}^2$, and $\mathtt{k}$ denotes the convolutional kernel. The convolution exhibits shift equivariance, expressed as:

$$[L_t I] * \mathtt{k} = L_t[I * \mathtt{k}], \tag{2.23}$$

where L_t represents a translation operation, and $L_t I$ is the shifted image. This property implies that the convolutional kernel $\mathtt{k}$ uses the same weights across different regions of the image I, effectively reducing the number of parameters and enhancing the model's generalization capacity.

Parameter sharing and the convolution operation replace traditional matrix mul-

tiplication in at least one of the layers within a CNN. This not only reduces the number of parameters significantly compared to fully connected networks but also makes CNNs equivariant to translations. In practical terms, if the input data shifts, the output feature map shifts accordingly. These characteristics make CNNs exceptionally suitable for tasks requiring the recognition and processing of localized spatial features.

2.7.4 Transfer Learning

Transfer learning is a powerful strategy in deep learning where knowledge acquired from one task is leveraged on a different, but related, task [14]. This strategy is particularly useful for enhancing performance on tasks with limited data availability by utilizing models pre-trained on large datasets.

The main methods in transfer learning are fine-tuning and feature extraction. Fine-tuning involves making minor adjustments to a pre-trained model to tailor it to a new dataset, while feature extraction uses the learned features of a pre-trained model as the foundation for training a new, simpler model. These strategies not only save computational resources but also reduce the risk of overfitting on smaller datasets.

Overall, transfer learning boosts learning efficiency and improves model performance by applying existing knowledge to new challenges, which is crucial when data is limited or computational power is restricted.

Chapter 3

Sound field estimation

Accurate estimation of the sound field around a rigid sphere necessitates adequate sampling on the sphere, which may not always be possible. To overcome this challenge, this thesis proposes a method for sound field estimation based on a physics-informed neural network. This method integrates physical knowledge into the architecture and training process of the network. In contrast to other learning-based methods, the proposed method incorporates additional constraints derived from the Helmholtz equation and the zero radial velocity condition on the rigid sphere. Consequently, it can generate physically feasible estimations without requiring a large dataset. In contrast to the spherical harmonic-based method, the proposed method has better fitting abilities and circumvents the ill condition caused by truncation. Simulation results demonstrate the effectiveness of the proposed method in achieving accurate sound field estimations from limited measurements, outperforming the spherical harmonic method and plane-wave decomposition method.

3.1 Introduction

Sound field analysis is a key topic of research with a wide range of applications, including spatial audio processing [15, 16], augmented reality/virtual reality [17], and active noise control [18, 19]. To determine the distribution of acoustic pressure in a region, researchers often use microphone arrays. However, microphone arrays cause reflections and scattering, which potentially impacts the accuracy of the sound field estimation. Therefore, understanding the scattering field is essential to ensure the accuracy of the sound field estimation.

Rigid spherical microphone arrays are extensively employed in sound field analysis due to their theoretically derived scattering fields, enabling the estimation of sound fields in three-dimensional space surrounding the array [20–22]. Conventional methods, such as the spherical harmonic (SH) method [23], represent the sound field by a set of orthonormal modes. However, these methods have inherent limitations

in practical applications. For instance, the limited number of microphones leads to a discrete representation through integration, resulting in truncation errors [8]. Plane-wave decomposition method assumes that the sound field only consists of plane waves [24]. It neglects the spatial variation within each wavefront, which may not fully capture the intricacies of the scattering fields.

Learning-based methods have been employed in scattering problems, and have demonstrated effectiveness in fitting nonlinear mappings [25, 26]. Among these methods, Convolutional Neural Networks (CNNs) have gained popularity due to their ability to promote parameter sharing and reduce the number of trainable parameters [27]. However, in the context of sound field estimation, convolutional and max-pooling layers introduce pixel spaces that can lead to discontinuities in sound pressure continuity. This discontinuity potentially undermines the physical interpretation of the data [28–30]. Additionally, the heavy reliance on training data often necessitates a large simulated dataset.

Recently, researchers introduced Physics-Informed Neural Networks (PINNs) as a method to incorporate physical knowledge into the network architecture and training [30–37]. This integration allows PINNs to achieve physically feasible estimations and requires less training data. Fully connected Neural Networks (FNNs) are commonly employed as part of PINN due to their ability to approximate any continuous function, as guaranteed by the Universal Approximation Theorem [38].

In this thesis, we propose a method for sound field estimation around a rigid sphere using PINNs. We employ a small FNN network structure and incorporate two constraints: the Helmholtz equation and the fact that sound pressure cannot generate radial motion at the sphere boundary. We use the FNN to model the mapping between Euclidean space positions and sound pressure and use these constraints as a loss function to train the network. Notably, our method relies solely on measured data from microphones, eliminating the need for setting up a grid and using simulated datasets. Simulation results demonstrate the effectiveness of our method, outperforming the spherical harmonic method and the plane-wave decomposition method.

3.2 Problem formulation for sound field reconstruction

We introduce two coordinate systems, as shown in Figure 3.1, where $\boldsymbol{x} = (x, y, z)$ and $\boldsymbol{r} = (r, \theta, \phi)$ represent Cartesian and spherical coordinates, respectively. Our

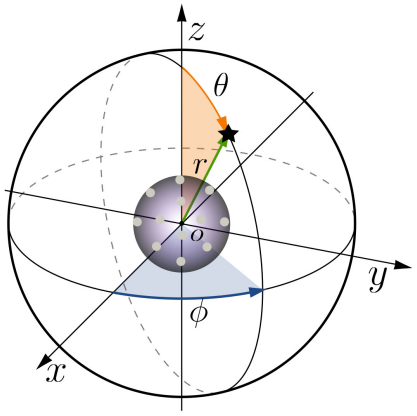


Figure 3.1: System setup: A rigid sphere with microphones showing Cartesian and spherical coordinates. • represent the microphones, and ⋆ represents a reference point.

study focuses on a sound field $\mathcal{V} \subset \mathbb{R}^3$, which includes a rigid sphere located at the origin O with a radius of r_0. We aim to estimate the sound field $P(\boldsymbol{r}, k)$ at position $\boldsymbol{r}$ and wavenumber k around the sphere $r \geq r_0$, by using measurements from Q microphones mounted on the rigid sphere.

For a homogeneous sound field (absence of sound sources) around a rigid sphere, theoretical solutions for the sound field have been established [20]. There are three issues with the implementation of the theoretical method. First, the infinite series in Eq. (2.11) needs to be truncated at limited order N [8]. Second, the pressure on the sphere is sampled at a limited (Q) number of positions. Hence, the integral in Eq. (2.12) needs to be discretely approximated as Eq. (2.14). Third, when the distance r increases ($r > r_0$), the reconstruction of the sound field becomes an ill-posed problem, leading to large variations due to small noise [22]. Consequently, small variations resulting from truncation errors or noise will be magnified.

We propose a neural network approach integrated with physics laws to overcome these challenges and obtain more accurate estimations. This method aims to establish a regression relationship between coordinates and sound pressure while satisfying the governing Helmholtz equation and the boundary condition.

3.3 Physics-informed neural network (PINN) method

In this section, we present the PINN method, which consists of three steps: (1) constructing an FNN using microphone positions and measured sound pressure as training pairs, (2) incorporating physical knowledge as a constraint in the network,

and (3) adjusting the weight of the loss function to facilitate convergence. The workflow of the PINN method is presented in Figure 3.2.

FNN is a composition of multiple layers, and the nodes between adjacent layers are fully connected. A single layer of the network is denoted as

$$F(\boldsymbol{x}) = \sigma\left(\boldsymbol{x}^{\intercal}\mathbf{w} + b\right), \tag{3.1}$$

where $\boldsymbol{x} \in \mathbb{R}^3$ is the inputs, which in this case are the Cartesian positions in the sound field $\mathcal{V}$, $\mathbf{w}$ represents the weights, b is the bias, and σ is the activation function. The FNN is a composition of multiple layers

$$\mathtt{F}(\boldsymbol{x}, k; \Phi) = \left(F_n \circ \ldots \circ F_2 \circ F_1\right)(\boldsymbol{x}), \tag{3.2}$$

where Φ represents the set of all trainable parameters $\mathbf{w}$ and b, and we use the notation ; to distinguish between the input variable $\boldsymbol{x}_i$ and the model parameters Φ. Hereafter, k, the wavenumber, is omitted in $\mathtt{F}(\boldsymbol{x}, k; \Phi)$ for notional simplicity. Φ is learned by minimizing the mean squared error (MSE) loss $\mathcal{L}_{\text{data}}$ with respect to the training data

$$\Phi^* = \arg\min_{\Phi} \mathcal{L}_{\text{data}}(\boldsymbol{x}; \Phi), \tag{3.3}$$

where $\mathcal{L}_{\text{data}}$ is defined as

$$\mathcal{L}_{\text{data}}(\boldsymbol{x}_i; \Phi) = \frac{1}{Q}\sum_{i=1}^{Q}\left(P(\boldsymbol{x}_i, k) - \mathtt{F}\left(\boldsymbol{x}_i; \Phi\right)\right)^2. \tag{3.4}$$

It fits the regression relationship between the spatial coordinates of the microphones $\boldsymbol{x}$ and the corresponding sound pressure values $P(\boldsymbol{x}, k)$. The network results $\mathtt{F}(\boldsymbol{x}; \Phi)$ represent the estimated sound pressure $\hat{P}(\boldsymbol{x}, k)$.

The aforementioned approach represents a purely data-driven method. It needs a large dataset, which is not always possible. We incorporate domain knowledge to reduce the data requirements and generate physically meaningful estimations.

For sound field estimation around a rigid sphere, we can leverage two domain knowledge to improve the estimation accuracy. Firstly, sound fields are governed by the Helmholtz equation Eq. (2.7). Therefore, we require the PINN estimation to satisfy

$$\left(\nabla^2 + k^2\right)\mathtt{F}(\boldsymbol{x}; \Phi) = 0. \tag{3.5}$$

Secondly, radial velocity is zero on the rigid sphere:

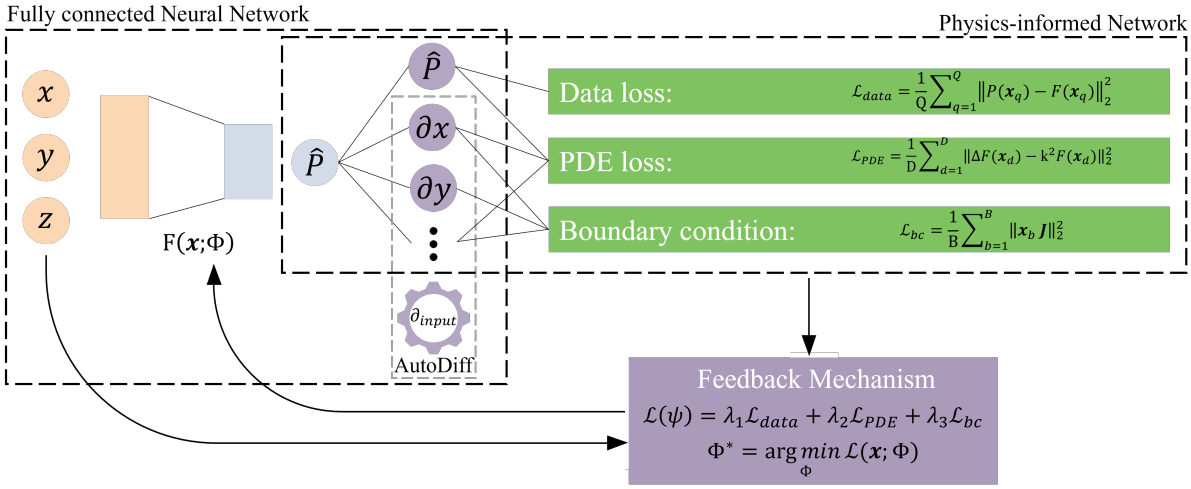


Figure 3.2: The workflow includes three components: an FNN for mapping inputs to outputs with trainable parameters Φ, a Physics-Informed Network that incorporates governing Helmholtz equations and boundary conditions, and a feedback mechanism to minimize the loss function and update Φ.

$$\left.\frac{\partial P(\boldsymbol{r},k)}{\partial r}\right|_{r=r_0} = 0. \tag{3.6}$$

It represents the derivative of the sound pressure for the radius of the sphere and can be written in terms of derivatives for the Cartesian coordinates as

$$\frac{\partial P(\mathbf{r},k)}{\partial r} = \frac{\partial P(\mathbf{r},k)}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial P(\mathbf{r},k)}{\partial y}\frac{\partial y}{\partial r} + \frac{\partial P(\mathbf{r},k)}{\partial z}\frac{\partial z}{\partial r}. \tag{3.7}$$

Conduct the substitutions $\frac{\partial x}{\partial r} = \frac{x}{r}$, $\frac{\partial y}{\partial r} = \frac{y}{r}$, $\frac{\partial z}{\partial r} = \frac{z}{r}$, (3.6) can be expressed in a matrix form

$$\left.\frac{\partial P(\boldsymbol{r},k)}{\partial r}\right|_{r=r_0} = \begin{bmatrix} x & y & z \end{bmatrix} \begin{bmatrix} \frac{\partial P(\mathbf{r},k)}{\partial x} \\ \frac{\partial P(\mathbf{r},k)}{\partial y} \\ \frac{\partial P(\mathbf{r},k)}{\partial z} \end{bmatrix} = 0. \tag{3.8}$$

We denote $\boldsymbol{x_b}$ as a point on the rigid sphere with radius r_0 and $\begin{bmatrix} \frac{\partial \mathtt{F}(\boldsymbol{x};\Phi)}{\partial x} & \frac{\partial \mathtt{F}(\boldsymbol{x};\Phi)}{\partial y} & \frac{\partial \mathtt{F}(\boldsymbol{x};\Phi)}{\partial z} \end{bmatrix}^{\mathsf{T}}$ as $\boldsymbol{J}$. Therefore, we require the PINN estimation to satisfy (3.8), i.e.,

$$\boldsymbol{x_b J} = 0. \tag{3.9}$$

At present, we have constructed an FNN and acquired physical knowledge, which allows us to develop a PINN. Leveraging automatic differentiation within the neural network, we can efficiently compute the first-order and second-order derivatives of $\mathtt{F}(\boldsymbol{x};\Phi)$ as required in Eq. (3.5) and Eq. (3.9). These derivatives are essential for updating the network parameters Φ using gradient descent.

We formulate the loss function to include three terms, each contributing to the overall training process

$$\begin{aligned}\mathcal{L}(\Phi) = \lambda_1 \underbrace{\frac{1}{Q}\sum_{q=1}^{Q} \|P(\boldsymbol{x}_q) - \mathtt{F}(\boldsymbol{x}_q;\Phi)\|_2^2}_{\mathcal{L}_{\text{data}}} \\ + \lambda_2 \underbrace{\frac{1}{D}\sum_{d=1}^{D} \left\|\nabla^2 \mathtt{F}(\boldsymbol{x}_d;\Phi) + k^2 \mathtt{F}(\boldsymbol{x}_d;\Phi)\right\|_2^2}_{\mathcal{L}_{\text{PDE}}} \\ + \lambda_3 \underbrace{\frac{1}{B}\sum_{b=1}^{B} \|\boldsymbol{x}_b \boldsymbol{J}\|_2^2}_{\mathcal{L}_{\text{bc}}},\end{aligned} \tag{3.10}$$

where $\|\cdot\|_2$ is the 2-norm. $\mathcal{L}_{\text{data}}$ represents the discrepancy between the PINN estimations $\mathtt{F}(\boldsymbol{x})$ and the actual measurements $P(\boldsymbol{x}_q)$ obtained from microphones. $\mathcal{L}_{\text{PDE}}$ ensures that the Helmholtz equation is satisfied at D discrete positions $\boldsymbol{x}_d$ within the domain $\mathcal{V}$, and $\mathcal{L}_{\text{bc}}$ imposes the boundary condition of zero radial velocity at B discrete positions $\boldsymbol{x}_b$ on the rigid sphere.

To control the relative importance of these terms and ensure similar influences on the training process, we introduce weight factors λ_1, λ_2, and λ_3. The weights in the loss function significantly influence the convergence and accuracy of the PINN. By balancing loss weights, a compromise can be achieved between accurately fitting the microphone measurements and adhering to the physical constraints. Direct normalization is not feasible since all three losses aim to approach zero. $\mathcal{L}_{\text{data}}$ is influenced by data noise. To address the issue, we balance $\mathcal{L}_{\text{PDE}}$ and $\mathcal{L}_{\text{bc}}$ to the same order of magnitude during the initial stages of training. Specifically, We set λ_1 to 1, λ_2 to k^2, and λ_3 to r_0 , which are empirically chosen to match each loss term's scales and ensure balanced convergence during training.

3.4 Experimental validation on simulated sound field

We conduct simulations in a 3D free field to evaluate the proposed method regarding the sound field estimation surrounding a rigid sphere and compare the performance of three methods: the spherical harmonic method (SH) [23], the plane wave decomposition method (PL) [39], and the proposed method (PINN).

We uniformly mount 32 omnidirectional microphones on a rigid sphere with a radius

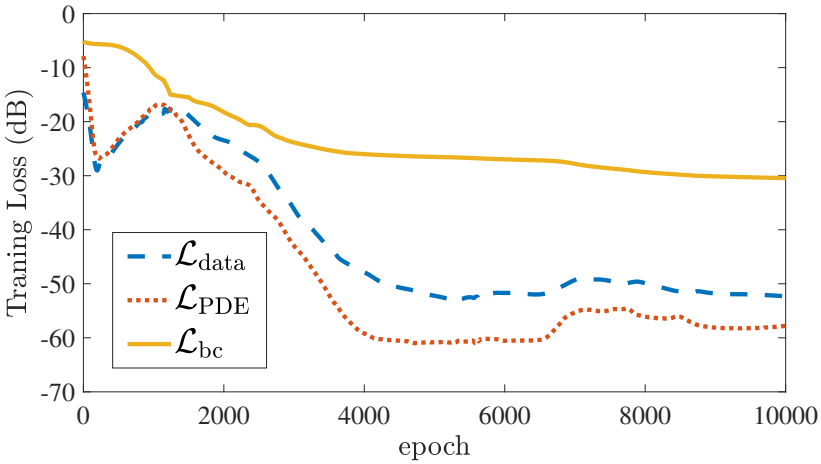


Figure 3.3: PINN training loss: $\mathcal{L}_{\text{data}}$, $\mathcal{L}_{\text{PDE}}$, and $\mathcal{L}_{\text{bc}}$.

of $r_0 = 0.042$ m as shown in Fig. 3.1. We simulate the transfer functions between the sound source and the microphones on the rigid sphere according to [40]. Specifically, we place two point sources at coordinates $(2.5, 0.8, 0.0)$ m and $(-2.0, -0.6, 1.2)$ m, both with an amplitude of 1. The speed of sound is set to $c = 343$ m/s. To account for realistic conditions, we add white Gaussian noise to the observation signals, resulting in a signal-to-noise ratio (SNR) of 30 dB. The pressure values are then normalized within the range of $[-1, 1]$.

We construct a 3-layer FNN with 4 nodes with the hyperbolic tangent (tanh) activation function for the PINN method. To train the PINN, we employ the Adam optimizer [41] with a learning rate of 10^{-5} for a total of 10,000 epochs. We set $\mathcal{L}_{\text{data}}$ based on the 32 microphone measurements, $\mathcal{L}_{\text{PDE}}$ based on 1000 randomly distributed points surrounding the sphere, and $\mathcal{L}_{\text{bc}}$ based on 500 uniformly distributed sampling points on the rigid sphere. As we train the network in Cartesian coordinates, we calculate the spherical coordinates $\boldsymbol{r}$ and then convert them to Cartesian coordinates $\boldsymbol{x}$.

In Fig. 3.3, we illustrate the training process. It demonstrates the simultaneous and nearly synchronized reduction of the training error $\mathcal{L}_{\text{PDE}}$ and $\mathcal{L}_{\text{bc}}$ in the initial stages of training for the first 2000 epochs. Subsequently, a gradual decline in $\mathcal{L}_{\text{data}}$ is observed. $\mathcal{L}_{\text{bc}}$ remains consistently higher than the other two losses throughout training, likely due to the higher spatial variability and numerical magnitude of the velocity gradients compared to the pressure field. Despite this, the overall downward trend of $\mathcal{L}_{\text{bc}}$ indicates that the network is still able to incorporate boundary information effectively. This observation suggests that the weight configurations of $\lambda_1 = 1$, $\lambda_2 = k^2$, and $\lambda_3 = r$ achieve a good balance.

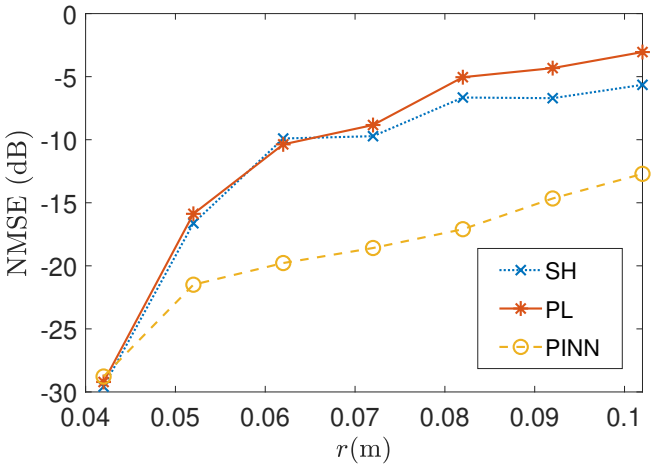


Figure 3.4: The sound field estimation error NMSE at 1000 Hz as a function of radius.

We define the pressure error between the true and estimated sound fields as

$$\text{error} := \|P(\boldsymbol{x}_g, \omega) - \hat{P}(\boldsymbol{x}_g, \omega)\|, \tag{3.11}$$

where $P(\boldsymbol{x}_g, \omega)$ and $\hat{P}(\boldsymbol{x}_g, \omega)$ represent the true and estimated pressures, respectively.

To evaluate the overall performance, we defined the normalized mean squared error (NMSE) as

$$\text{NMSE} := 10 \log_{10} \frac{\sum_{g=1}^{10000} \|P(\boldsymbol{x}_g, \omega) - \hat{P}(\boldsymbol{x}_g, \omega)\|_2^2}{\sum_{g=1}^{10000} \|P(\boldsymbol{x}_g, \omega)\|_2^2}, \tag{3.12}$$

The estimation was performed at 10,000 evaluation positions uniformly sampled within a sphere with a specific radius and frequency.

Fig. 3.4 illustrates the relationship between the radius r of the estimation sphere and the NMSE at a frequency of 1000 Hz. When $r = r_0$ corresponds to estimating the sound pressure on the rigid sphere, all three methods yield comparable results. However, as the radius increases, the SH and PL methods show a significant surge in error. In contrast, the PINN method exhibits a more gradual increase in error as the radius increases. It has a lower NMSE of $7 - 10$ dB for $r > 0.05$ m than the SH and PL methods. This indicates that the PINN method performs better compared to the conventional methods.

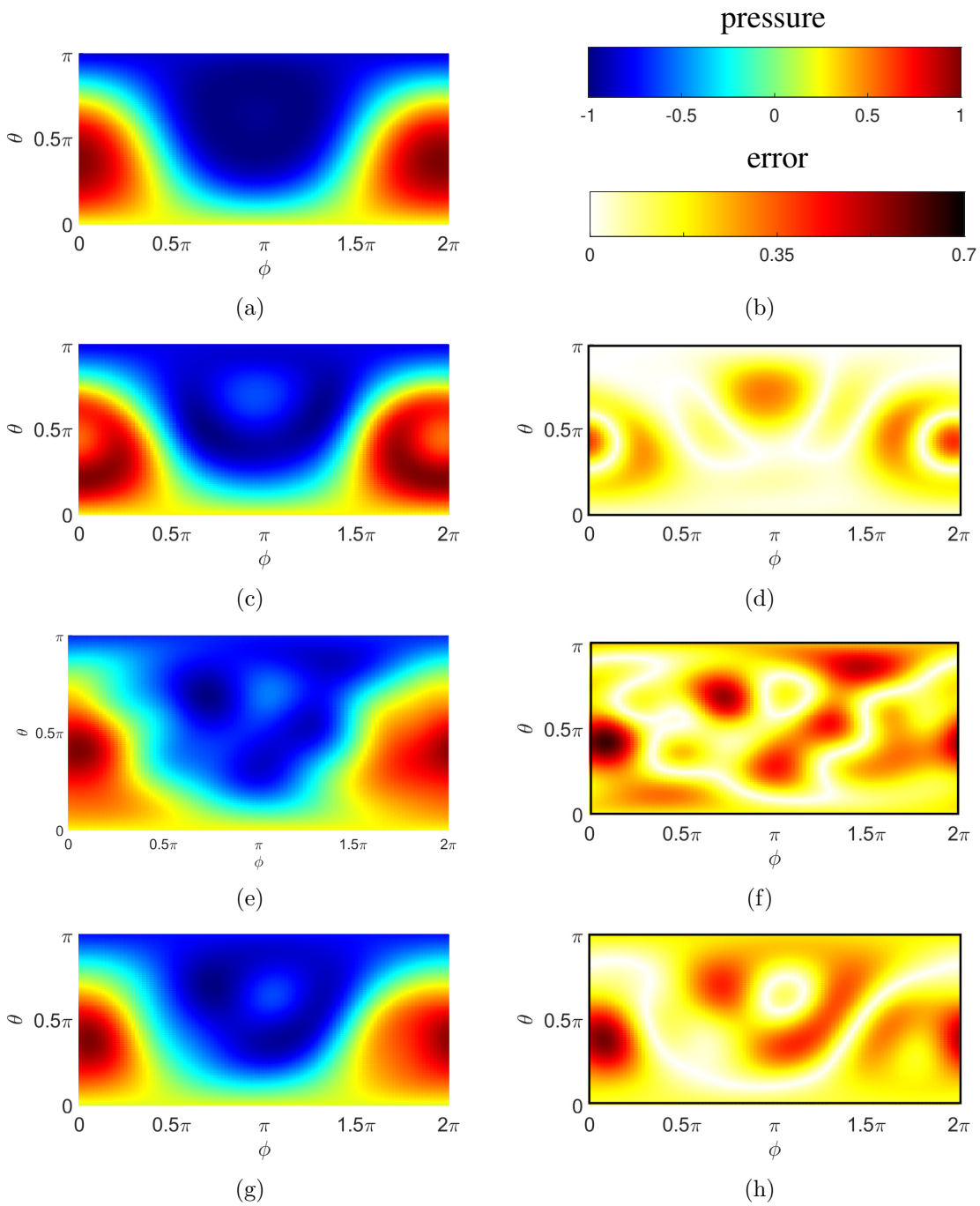


Figure 3.5: The sound field at a frequency of 1000 Hz and $r = 0.072$ m.(a) Ground truth, (b) Color bar, (c) PINN estimation, (d) PINN error, (e) PL estimation, (f) PL error, (g) Spherical Harmonics estimation, (h) Spherical Harmonics error.

Fig. 3.5 presents the ground truth, estimated sound pressure using these three methods, and error at a frequency of 1000 Hz and a radius of 0.072 m. It is observed that the SH method exhibits higher errors around $\theta = 0.5\pi$, the PL method performs the worst among the three due to its assumption of far-field conditions, which ignore the scattering information. The PINN method achieves the best results overall, although it still experiences relatively higher errors at $\theta = 0.5\pi$.

3.5 Conclusions

In this study, we propose a PINN method to estimate the sound field around a rigid sphere based on the measurement of a small number of microphones mounted on the rigid sphere. Our method employs an FNN to model the mapping between space positions and sound pressure. We incorporate the Helmholtz equation and the fact that sound pressure cannot generate radial motion at the sphere boundary as part of the loss function. This integration of physical knowledge into the network architecture and training process enhances the accuracy of estimation. A major advantage of our approach is that it relies solely on measured data, eliminating the need for grid setup and simulated datasets. Simulation results demonstrate that our method outperforms the spherical harmonic method and the plane-wave decomposition method.

It is hard to derive theoretical solutions for the scattering sound of complex geometries, but PINN can use physical and geometric constraints to fit the sound pressure distribution. It allows our method to estimate the sound field of arbitrary geometry potentially.

Chapter 4

Head-related transfer functions upsampling

Head-related transfer functions (HRTFs) are transfer functions between point sound sources and the listener's ear canal and are crucial for spatial sound field reproduction in virtual reality applications. HRTFs are usually measured in discrete and finite directions, necessitating upsampling for unmeasured directions. While deep learning approaches have recently advanced HRTF upsampling, they often overlook the spherical nature of the data. This thesis proposes a novel deep learning-based interpolation method that incorporates spherical convolution to extract spatial features directly on the sphere using spherical harmonics (SHs), thereby circumventing the limitations of conventional plane-based CNNs. Specifically, the convolution process is implemented as part of a neural network architecture by decomposing and reconstructing HRTFs through SH representations. The SHs, an orthogonal function set defined on a sphere, allow the spherical convolution layers to effectively capture the spatial features of HRTFs that are sampled on a sphere. Simulation results demonstrate the effectiveness of the proposed method in achieving more accurate interpolation from sparse measurements.

4.1 Introduction

The growing popularity of virtual reality and augmented reality applications has sparked an interest in spatial audio rendering methods [42]. One of the most popular methods is binaural reproduction, which relies on the head-related transfer function (HRTF) [43]. HRTFs represent how sound waves are modified as they propagate from a sound source to the listener's ears. This acoustic path is affected by several factors, including sound diffraction, reflection, and absorption by the listener's head, shoulders, and torso, as well as the resonances and shape of the listener's outer ears (pinnae). makes it highly individual [44] This acoustic path is affected by several

factors, including sound diffraction, reflection, and absorption by the listener's head, shoulders, torso, and outer ears (pinnae), making it highly individualized [44].

Individual HRTFs can be obtained through measurement, which is the most common and accurate method [45]. HRTFs are measured in discrete and finite directions, sampled in directions around a spatial spherical surface or a horizontal circle. HRTFs in unmeasured directions can be estimated from the measured data by using various interpolation methods.

The main interpolation methods can be divided into methods based on nearest measurements and based on basis function decomposition. Bilinear interpolation [46], Spherical triangular interpolation [47], and Barycentric interpolation [48] use several neighboring points around the point to be interpolated to calculate a weighted sum or average. Spherical harmonic decomposition [49, 50] and spatial principal component decomposition [51, 52] decompose HRTFs into a weighted sum of basis functions, and HRTFs on different directions are separately represented by variations in basis functions and weights. These methods only rely on the data of a single subject, it becomes much more inaccurate when interpolating sparser measurements.

More recently, deep learning methods have been introduced for HRTF interpolation. These methods extract common features among subjects from datasets, significantly reducing the required density of HRTF measurements needed to meet a given interpolation accuracy requirement. (Methods of using anthropometric features are outside the scope of this thesis.) [53] uses an autoencoder conditioned on source positions and obtains the source position-independent representation by using an aggregation module between the encoder and decoder, aggregating latent variables of a given source position. [54, 55] use neural field, where the HRTF is implicitly represented as a function of the sound source direction.

In learning-based methods, the Convolutional layer also plays a key role. Researchers found that HRTF interpolation is similar to the image super-resolution task, both upsampling from sparse sampling to dense. [56, 57] use the U-NET [58] to map sparse HRTFs to the dense HRTFs. [59] uses a convolutional super-resolution generative adversarial network. [60, 61] use a CNN and Transformer architecture to derive high-resolution HRTFs from sparse data, by mapping latent sequences that capture spatial and spectral correlations. However, CNN was primarily designed to process signals in a 2D plane. There exists the need to address the plane-sphere projection while interpolating sparsely measured HRTF with a CNN. Thus, to process HRTF which is defined on a sphere, CNN-based methods require complicated plane-sphere projection which does not necessarily capture the spatial features of HRTF [62].

Spherical CNNs have been proposed to support a spherical form of data with spherical convolution. Spherical CNNs define their convolutional filters via either known spatial tessellation [63–65] or spherical harmonic decomposition [66–69]. Successful applications have been demonstrated in tasks such as 3D shape analysis [70], medical imaging [71–73], Direction of Arrival [74]. To the best of our knowledge, our work was the first to apply spherical CNNs to the processing of HRTFs. Thuillier et al. [75] have subsequently employed a Neural Process model based on spherical CNNs.

We propose a spherical CNN method for HRTF interpolation. The spherical CNN method exploits the spherical harmonics, an orthogonal function set defined on a sphere, to resolve the plane-sphere projection problem. We transform the magnitude of sparsely measured HRTF into SH coefficients, which are convoluted with kernel functions expressed in the SH domain. The kernel function captures the spatial features of HRTF. Then we convert the convolution result into the magnitude of dense interpolated HRTF.

4.2 Problem formulation for HRTF interpolation

HRTF depends on the anatomy of a subject and the position of a sound source [43]. Let $H_i^{\text{left(right)}}(\Omega, l)$ denote the HRTF for the left (right) ear of subject $i \in \{1, \ldots, I\}$, source position Ω, and frequency bin $l \in \{1, \ldots, L\}$, where I is the total number of subjects and L is the number of frequency bins.

Source position $\Omega = (r, \theta, \phi)$ is defined by spherical coordinates, the radius r, the elevation $\theta = [-\frac{\pi}{2}, \frac{\pi}{2}]$ and the azimuth $\phi = [0, 2\pi)$. Hereafter, the superscript for the left(right) ear and the subscript for the subject number are omitted for notation simplicity.

In this paper, we focus on the far-field HRTF magnitude spectra $H_{\text{M}}(\Omega, l)$, i.e., the sound source is at a distance $r > 1.2$ m from the center of subject head [43],

$$H_{\text{M}}(\Omega, l) = 20 \log_{10} \left(|H(\Omega, l)| \right), \tag{4.1}$$

which is a logarithmic scale similar to human auditory perception [76]. The far-field HRTF magnitude spectra $H_{\text{M}}(\Omega, l)$ exhibits minimal dependence on r [43], and thus we redefine $\Omega = (\theta, \phi)$. In this regard, we treat $H_{\text{M}}(\Omega, l)$ as a function of spherical signal $H : \mathbb{S}^2 \to \mathbb{R}^L$, where $\mathbb{S}^2$ is a two-dimensional manifold parameterized by (θ, ϕ).

The problem of interest is to interpolate spatially dense HRTFs $\left\{H_{\mathrm{M}}(\Omega, l), \Omega \in \Omega^{\mathrm{dense}}\right\}$ from spatially sparse HRTFs $\{H_{\mathrm{M}}(\Omega, l), \Omega \in \Omega^{\mathrm{sparse}}\}$, where $\Omega^{\mathrm{sparse}} \subset \Omega^{\mathrm{dense}}$. The phase of the head-related impulse responses (HRIRs) can be derived by applying minimum phase reconstruction to the estimated magnitudes of the HRTF, ensuring it closely resembles the phase of the original HRTF [77].

4.3 Spherical CNN-based model

4.3.1 Spherical convolution

Given a sparse set of HRTF measurements, provide a reconstructed HRTF set in high-spatial resolution. HRTF interpolation has previously been considered as an image super-resolution task, applying CNNs [57, 59]. Conventional CNNs, originally designed for planar images, exhibit shift equivariance, which reduces model parameters and enhances generalization. Specifically, each convolutional kernel $\mathcal{K}$ shares weight parameters across different image regions. However, HRTFs, being defined on a spherical domain $\mathbb{S}^2$, do not maintain shift equivariance, making direct application of conventional CNNs impractical and necessitating complex projections of HRTF data [57, 59].

Considering the limitations of existing methods, we introduce a spherical convolution approach for HRTF interpolation, which effectively captures and processes spherical features. Spherical convolution generalizes CNN to function on data distributed over a sphere by using spherical convolution.

$$(g * \mathcal{K})(\Omega) = \int_{\mathbf{R} \in \mathbf{SO}(3)} g(\mathbf{R}\eta)\mathcal{K}\left(\mathbf{R}^{-1}\Omega\right) \mathrm{d}\mathbf{R}, \tag{4.2}$$

where g and $\mathcal{K}$ are spherical functions, $\mathcal{K}$ is a learnable convolutional kernel, $\mathbf{R}$ is the rotational operator in $\mathbf{SO}(3)$, the group of all 3D rotation matrices that preserve orientation and distance, and η is the north pole [67]. This operation, Eq. (4.2), exhibits rotational equivariance,

$$[\mathbf{R}g] * \mathcal{K} = \mathbf{R}[g * \mathcal{K}], \tag{4.3}$$

allowing spherical CNNs to share kernel weights across different spherical regions. When considering H_{M}, our goal is to use $H_{\mathrm{M}} * \mathcal{K}$ to capture spatial features effectively.

The spherical convolution, (4.2), does not possess a closed-form solution and is

computationally realized through spectral transformation [67], employing a linear combination of spherical harmonics and corresponding coefficients α_{nm} of g and β_{n0} of $\mathcal{K}$ [78],

$$(g * \mathcal{K})(\Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 2\pi \sqrt{\frac{4\pi}{2n+1}} \alpha_{nm} \beta_{n0} Y_{nm}(\Omega), \tag{4.4}$$

where $Y_{nm}(\Omega)$ is real spherical harmonics defined as

$$\begin{aligned} Y_{nm}(\Omega) = & \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) \\ & \times \begin{cases} \sin(|m|\phi) & m < 0, \\ \cos(m\phi) & m \geq 0, \end{cases} \end{aligned} \tag{4.5}$$

where $(\cdot)$! represents the factorial function, $P_n^m(\cdot)$ is the associated Legendre function, $n \geq 0$ and $-n \leq m \leq n$ denote the order and degree, respectively. The real spherical harmonics have the same orthonormality properties as the complex spherical harmonics.

4.3.2 Spherical harmonic decomposition

When considering H_{M}, our goal is use $H_{\mathrm{M}} * \mathcal{K}$ to effectively capture spatial features. The H_{M} coefficients a_{nm} can be obtained through spherical harmonic decomposition. We decompose HRTF magnitude spectra $H_{\mathrm{M}}(\Omega, l)$ onto the SHs expansion a finite order of truncation,

$$H_{\mathrm{M}}(\Omega, l) \approx \sum_{n=0}^{N_H} \sum_{m=-n}^{n} a_{nm}(l) Y_{nm}(\Omega), \tag{4.6}$$

where N_H denotes the highest decomposition order. Given P HRTF measurements, Eq.(4.6) can be expressed in a matrix form as inverse Spherical Harmonic Transform (ISHT)

$$\mathbf{H} = \mathbf{Y}\mathbf{a}, \tag{4.7}$$

where

$$\mathbf{H} = \begin{bmatrix} H_{\mathrm{M}}(\Omega_1, 1) & \cdots & H_{\mathrm{M}}(\Omega_1, L) \\ \vdots & \ddots & \vdots \\ H_{\mathrm{M}}(\Omega_P, 1) & \cdots & H_{\mathrm{M}}(\Omega_P, L) \end{bmatrix}$$

is a $P \times L$ matrix.

$$\mathbf{Y} = \begin{bmatrix} Y_{00}(\Omega_1) & \cdots & Y_{N_H N_H}(\Omega_1) \\ \vdots & \ddots & \vdots \\ Y_{00}(\Omega_P) & \cdots & Y_{N_H N_H}(\Omega_P) \end{bmatrix}$$

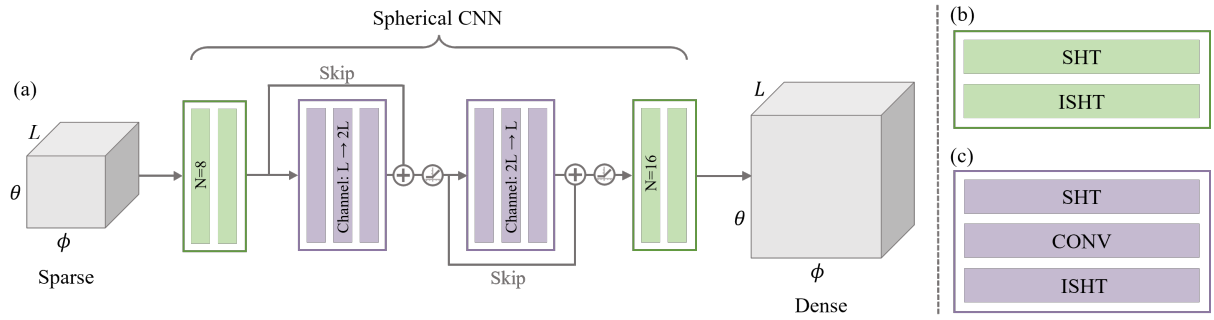


Figure 4.1: The network structure of the model. (a): It starts with a Mapping Block for sparse to dense signal transformation, followed by two Convolutional Blocks for spatial feature learning, and ends with a Mapping Block for dense domain mapping. (b): The Mapping block: processes the spherical signal through SHT to convert it into SH coefficients, followed by ISHT maps back to the spherical signal region. (c): The Convolutional block: performs SHT on the spherical signal, followed by coefficients multiplication and ISHT.

is the $Y_{nm}(\cdot)$ of order n and degree m, at P directions. It is a $P \times (N_H+1)^2$ matrix containing the real spherical harmonic (SH) basis functions.

$$\mathbf{a} = \begin{bmatrix} a_{00}(1) & \cdots & a_{00}(L) \\ \vdots & \ddots & \vdots \\ a_{N_H N_H}(1) & \cdots & a_{N_H N_H}(L) \end{bmatrix}$$

is a $(N_H+1)^2 \times L$ matrix containing the SH coefficients. To accurately derive these coefficients, we apply the Spherical Harmonic Transform (SHT) using a least-squares method,

$$\mathbf{a} = (\mathbf{Y}^\intercal \mathbf{Y})^{-1} \mathbf{Y}^\intercal \mathbf{H}. \tag{4.8}$$

This approach is particularly effective when the number of HRTF measurements significantly exceeds the number of coefficients required for the chosen truncation order, $P \gg (N_H+1)^2$, mitigating inaccuracies in sparse data interpolation.

4.3.3 Spherical Convolutional Block

We employ spherical CNNs to learn the coefficients β_{n0} of kernel functions $\mathcal{K}$. These kernels capture spherical features in HRTFs across subjects, thereby facilitating the transformation of low-order coefficients into high-order coefficients to achieve accurate interpolation of dense HRTFs.

In the first convolutional block, after applying the Spherical Harmonic Transform (SHT) as Eq.(4.8) with a low-order $N = 8$, we obtain a feature map $U \in \mathbb{R}^{(N+1)^2 \times L}$, where N and L denote the order and frequency dimensions, respectively. The frequency dimension serves a role similar to that of the channel dimension. To compensate for the limited spatial resolution of low-order SH, we apply a kernel that

expands the number of channels to $2L$, yielding $U' \in \mathbb{R}^{(N+1)^2 \times 2L}$. Subsequently, through the Inverse Spherical Harmonic Transform (ISHT) as Eq.(4.7), we derive $\mathbf{H}' \in \mathbb{R}^{P \times 2L}$.

In the second convolutional block, after applying the SHT with a high-order $N' = 16$, we achieve a feature map $U \in \mathbb{R}^{(N'+1)^2 \times L}$. Applying a kernel with L channels results in the output $U' \in \mathbb{R}^{(N'+1)^2 \times L}$. Then, following the ISHT with $\mathbf{Y}'$, a $P' \times (N'+1)^2$ matrix, we obtain $\mathbf{H}' \in \mathbb{R}^{P' \times L}$. At this stage, the channel number is reduced back to L to match the original scale and to avoid overparameterization.

For example, for a 120×93 matrix $\mathbf{H}$, the first convolutional block transforms it into a 120×186 matrix $\mathbf{H}'$. After processing through the second convolutional block, it results in a 480×93 matrix $\mathbf{H}'$. Fig. 4.1 illustrates the detailed network structure.

4.4 Evaluation on HRTF dataset

In this section, we conducted experiments on HRTF interpolation to assess the effectiveness of our proposed method, and compare it with the SH method [50], the SH+DNN method [79], HRTF field method [54], and the CNN+GAN method [59].

4.4.1 Data pre-processing

We chose to conduct experiments on the HUTUBS dataset due to its comprehensive subject coverage [80]. This dataset includes HRIR measurements from 94 unique subjects, after excluding any duplicates. We divided these subjects into three groups for our experiments: 77 for training, 10 for validation, and 7 for testing. Fig. 4.2 shows the direction of 120 known HRTFs and 360 unknown HRTFs.

We used ground truth 16-order SH coefficients to generate HRTFs from 480 directions and across $L = 93$ frequency bins, spanning from 172 Hz to 16 kHz.

4.4.2 Training

The Log Spectral Distortion LSD (Eq. (4.9)) serves as the loss function for the training process.

$$\mathrm{LSD} = \sqrt{\frac{1}{PL} \sum_{p=1}^{P} \sum_{l=1}^{L} \left| H_{\mathrm{M}}(\Omega_p, l) - \hat{H}_{\mathrm{M}}(\Omega_p, l) \right|^2}. \tag{4.9}$$

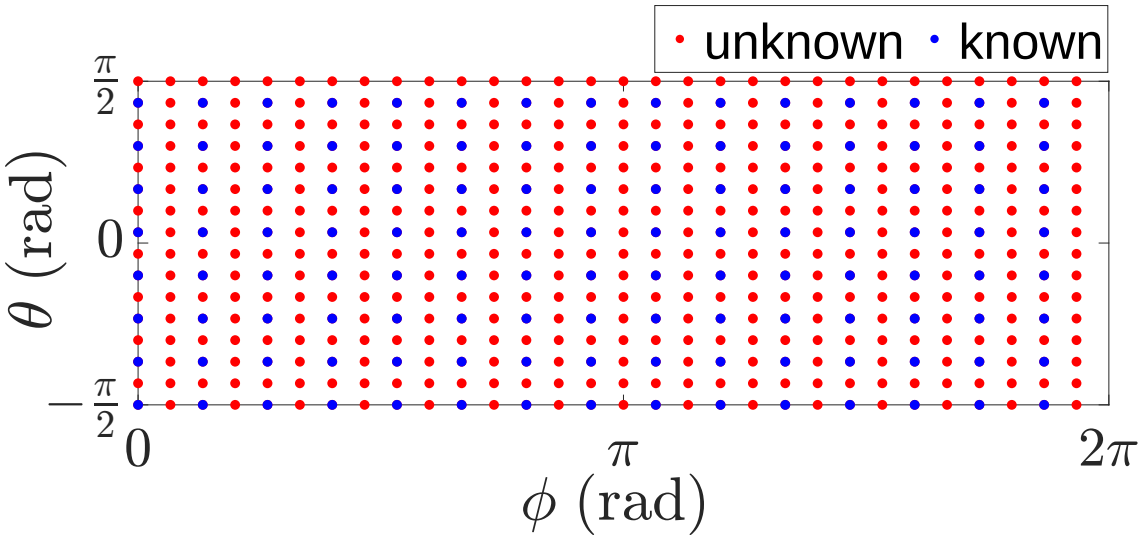


Figure 4.2: Directions of 360 unknown and 120 known HRTFs.

Table 4.1: Comparison of model parameters and average LSD for test Subject IDs 88–94, averaged over 360 unknown spatial directions.

HRTF model	Parameters	Unknown
SH N = 8 [50]	-	7.24
SH+DNN [79]	803241	5.11
HRTF field [54]	4458589	4.19
CNN+GAN [59]	100827332	4.36
Proposed	**484530**	**2.17**

The training procedure used the Adam optimizer [81] with a batch size of 14. Our model was trained on a V100 GPU for about 700 epochs using a learning rate of 0.001, with early stopping based on validation data.

4.4.3 Data post-processing

To obtain the head-related impulse responses (HRIRs) of the estimated HRTF magnitudes, we can use a method of assigning a phase to the HRTF magnitude. For example, a method based on minimum-phase reconstruction provides a phase similar to that of the original HRTFs in sound image localization [77]. The minimum phase reconstruction method is applied to HRTF magnitudes to generate HRIRs.

4.4.4 Result and discussion

We conducted a comparative analysis by adopting interpolation methods, including SH, SH+DNN, HRTF field, and CNN+GAN. Table 4.1 provides insights into the number of parameters and average LSD of learning-based methods for unknown HRTFs across Subjects ID 88-94. Notably, the proposed method exhibits the lowest

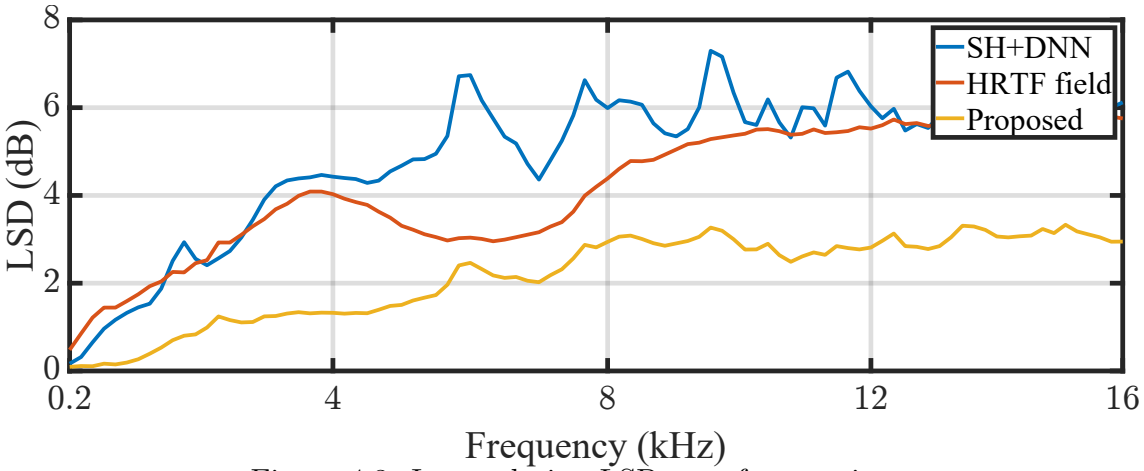


Figure 4.3: Interpolation LSD over frequencies.

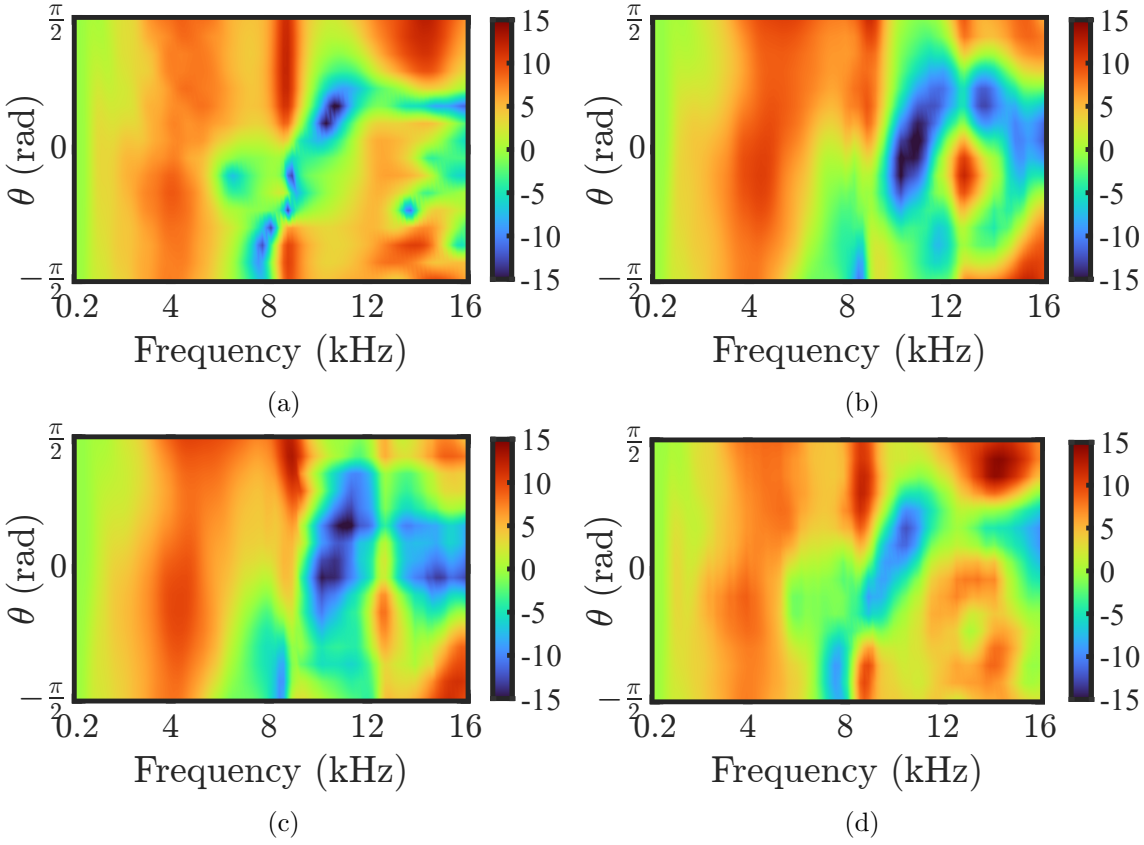


Figure 4.4: The left ear HRTF at $\phi = \pi$ of Subject ID 89. (a) Ground truth, (b) SH+DNN, (c) HRTF field, (d) Proposed method.

parameter count due to its streamlined architecture consisting of only two convolution blocks. In contrast, both SH+DNN and HRTF field methods rely on fully connected networks. SH+DNN requires sub-networks training for each frequency bin while CNN+GAN employs a generator with 11 conventional CNN layers. This observation highlights the efficiency of our approach. Furthermore, our method achieves the lowest LSD. The SH method with 8th-order coefficients performed worse than

SH+DNN with 4th-order coefficients, due to its inability to accurately compute 8th-order coefficients with 120 known HRTFs. The 8th order was chosen as it provides a higher spatial resolution and is commonly used in SH-based HRTF representations. However, higher-order SH requires more dense spatial sampling to achieve accurate reconstruction, which may not be satisfied with the limited number of input HRTFs. It indicates that the proposed method learns from convolution blocks, rather than solely relying on direct inferences through the mapping blocks.

We compare the HRTF interpolation performance of the proposed model with the SH+DNN and HRTF field model which has fewer parameters. Interpolating the high-frequency part of HRTF (above 8 kHz) is challenging. Fig. 4.3 demonstrates that our method performs well in high frequencies. Fig. 4.4 showcases HRTF interpolation at $\phi = \pi$. It provides a visual assessment of how well the methods fit the *notch*, which represents a notable decrease at high-frequency, crucial for sound source direction detection [43]. The ground truth exhibits a peak in the 8 to 12 kHz frequency range. The other two methods show a different spectral shape compared to the ground truth. In contrast, our proposed method appropriately captures the peaks and notches of the ground truth.

4.5 Conclusions

This paper proposed a spherical CNN method to interpolate HRTFs based on sparse measurements. By leveraging the SH transformation, the proposed method directly exploits spherical information, eliminating the need for plane-sphere projection. Simulation results demonstrated that our method outperforms both conventional SH methods and learning-based methods. This method not only enhances HRTF interpolation but also offers a compelling alternative to the widely employed SH method in soundfield reproduction tasks.

Chapter 5

Monaural speech enhancement on drones

Monaural Speech enhancement on drones is challenging because the ego-noise from the rotating motors and propellers leads to extremely low signal-to-noise ratios at onboard microphones. Although recent masking-based deep neural network methods excel in monaural speech enhancement, they struggle in the challenging drone noise scenario. Furthermore, existing drone noise datasets are limited, causing models to overfit. Considering the harmonic nature of drone noise, this thesis proposes a frequency-domain bottleneck adapter to enable transfer learning. Specifically, the adapter's parameters are trained on drone noise while retaining the parameters of the pre-trained Frequency Recurrent Convolutional Recurrent Network (FRCRN) fixed. Evaluation results demonstrate that the proposed method can effectively enhance speech quality. Moreover, it is a more efficient alternative to fine-tuning models for various drone types, which typically requires substantial computational resources.

5.1 Introduction

Speech enhancement using a drone-mounted microphone enables services in diverse areas, such as search and rescue missions, video capture, and filmmaking [82]. However, the microphone's proximity to the drone's noise sources, such as motors and propellers, results in an extremely low signal-to-noise ratio (SNR) of recordings, typically ranging from -25 to -5 dB [83]. This severely limits the drone audition applications.

While multi-channel microphone array methods are commonly used to improve audio quality, they often require specialized hardware that is either too large or heavy for most drones. For instance, drones used in the DREGON dataset [84] are equipped with an 8-channel microphone array and have a total weight of up to 1.68 kg, and

Wang *et al.* [85] employ a drone with a 0.2 m diameter circular microphone array. Moreover, the performance of these methods degrades significantly in dynamic scenarios with moving microphones or sound sources [86]. Thus, developing efficient single-channel solutions is preferred for extending the applicability of drone audition.

Monaural speech enhancement, or single-channel noise suppression, has been extensively studied for decades. Traditional approaches include spectral subtraction [87] and filtering [88], while recent advances use deep neural networks (DNNs). Particularly, masking-based DNN methods showcased promising results in recent Deep Noise Suppression Challenges (DNS) [89]. These methods often use Convolutional Recurrent Networks (CRNs) to extract features from temporal-spectral patterns, and then predict a ratio mask for noisy speech on the time-frequency spectrum [90–92]. However, these models are typically optimized for scenarios with relatively high input SNRs (e.g., > -5 dB). Applying these pre-trained models directly to drone noise often results in sub-optimal enhancement performance.

Research specifically addresses drone noise suppression is still in its early stages [93], and public drone noise datasets are relatively small [84, 86, 94, 95], containing only a few types of drones and lacking the diversity needed for training. Mukhutdinov *et al.* [83] comprehensively evaluated the performance of DNNs in monaural speech enhancement on drones. They rely on limited drone noise datasets, leading to the overfitting to specific drone types.

To address challenges in drone audition, we propose a frequency-domain bottleneck adapter for transfer learning, specifically designed to capture the harmonic nature of drone noise. This adapter-based tuning method selectively trains adapter parameters while retaining the pre-trained model's parameters [96, 97], which prevents overfitting with the small-scale training data. We apply this method by fine-tuning a pre-trained FRCRN on drone noise datasets. The proposed method efficiently adapts to the distinct acoustic characteristics of various drone types, thereby enhancing speech quality and intelligibility in drone recordings.

While this work focuses on monaural enhancement due to the hardware constraints of drones, we note that lightweight and scalable multi-channel designs may be considered in future work to further improve spatial awareness, provided they meet drone payload limitations.

5.2 Problem formulation for drone noise suppression

Consider a single microphone mounted on a flying drone to capture human speeches as shown in Fig. 5.1a. The microphone recording can be represented as

$$y(t) = s(t) * h(t) + v(t), \tag{5.1}$$

where $s(t)$ denotes the clean speech with t being the time index, $h(t)$ is the impulse response between the human and the microphone, $v(t)$ denotes the noise, and $*$ denotes the convolution operation. $v(t)$ is mainly contributed by the drone propeller rotations and vibrations of the structure and is dominated by multiple narrow-band noises [98, 99], as demonstrated in Fig. 5.1b. As the drone-mounted microphone is close to the noise sources (i.e. motors and propellers), the SNR of $y(t)$ is very low, resulting in the extreme difficulty in uncovering $s(t)$ from $y(t)$. This is demonstrated in Fig. 5.1c.

To leverage information from both the time and frequency domains, we apply the Short-Time Fourier Transform (STFT) to Eq.(5.1), yielding the Time-Frequency (T-F) domain representation:

$$Y(K, l) = S(K, l)H(K, l) + V(K, l), \tag{5.2}$$

where $Y(K, l), S(K, l)$, and $V(K, l)$ are the complex spectra of $y(t), s(t)$, and $v(t)$, respectively, with K denotes the frequency bin index and l denotes the time frame index. Fig. 5.1d illustrates a typical $Y(K, l)$, highlighting the time-frequency sparsity of harmonic noise and energy concentrated at isolated frequency bins.

The noisy speech can be enhanced by complex masking [100]. In the ideal case, the enhanced speech is obtained by

$$\hat{S}(K, l) = c\,\mathrm{IRM}(K, l) \odot Y(K, l), \tag{5.3}$$

where $c\,\mathrm{IRM}(K, l)$ is the complex Ideal Ratio Mask (cIRM), and $\odot$ is element-wise complex multiplication. The cIRM is formulated by

$$c\,\mathrm{IRM}(K, l) = M_{\mathrm{r}}(K, l) + iM_{\mathrm{i}}(K, l), \tag{5.4}$$

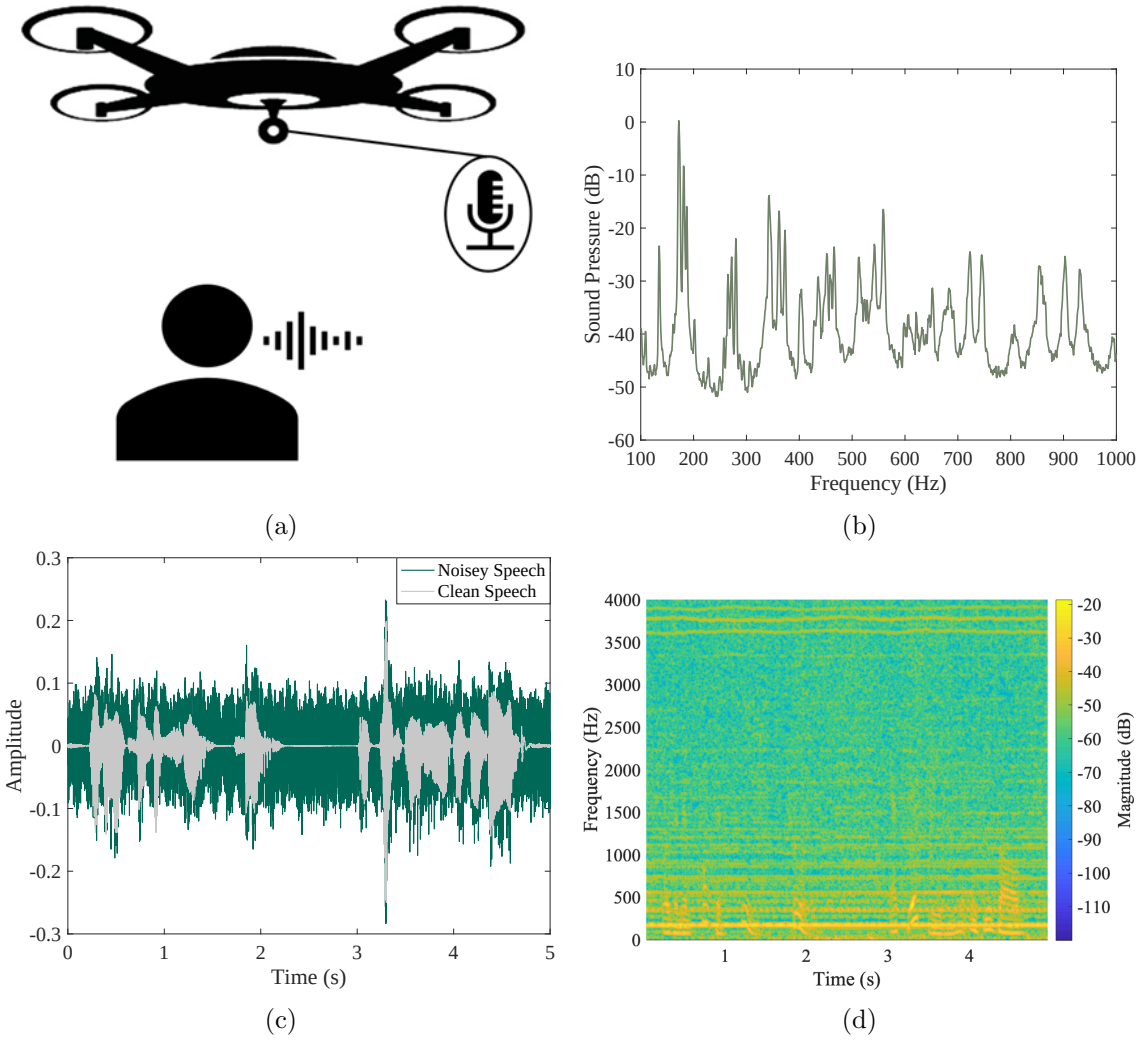


Figure 5.1: Problem setup: (a) illustration of the monaural speech recording scenario over a flying drone (b) drone ego noise in the frequency-domain (c) noisy speech in the time domain (d) noisy speech in the time-frequency-domain.

where $M_{\mathrm{r}}(K,l)$ and $M_{\mathrm{i}}(K,l)$ are, respectively, given by

$$\begin{aligned} M_{\mathrm{r}}(K,l) &= \frac{Y_{\mathrm{r}}(K,l)S_{\mathrm{r}}(K,l) + Y_{\mathrm{i}}(K,l)S_{\mathrm{i}}(K,l)}{Y_{\mathrm{r}}^2(K,l) + Y_{\mathrm{i}}^2(K,l)}, \\ M_{\mathrm{i}}(K,l) &= \frac{Y_{\mathrm{r}}(K,l)S_{\mathrm{i}}(K,l) - Y_{\mathrm{i}}(K,l)S_{\mathrm{r}}(K,l)}{Y_{\mathrm{r}}^2(K,l) + Y_{\mathrm{i}}^2(K,l)}, \end{aligned} \tag{5.5}$$

where the subscripts r and i indicate the real and imaginary components of the spectra, respectively. Since $S(K,l)$ is unknown, directly deriving $c\,\mathrm{IRM}(K,l)$ is not feasible. Then, the research problem reduces to estimate a complex mask $\hat{M}(K,l)$ that approximates $c\,\mathrm{IRM}(K,l)$ using the single channel recording and the pre-knowledge about drone noise.

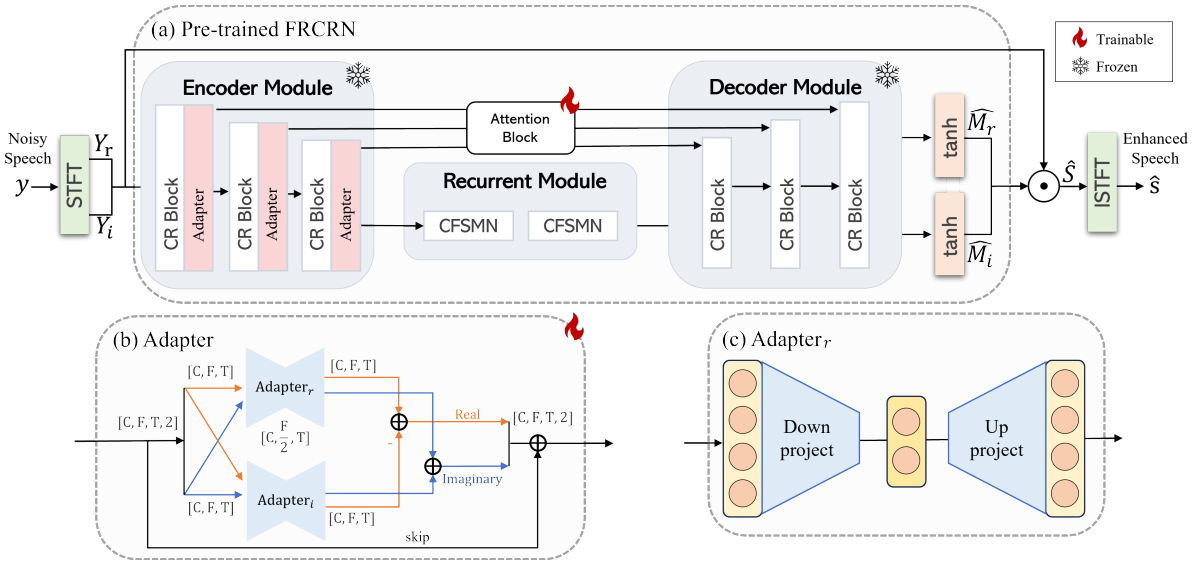


Figure 5.2: The overview of the Adapter pipeline. (a) Pre-trained FRCRN with adapter embedded; (b) Adapter; (c) Adapter$_r$.

5.3 Design of frequency-domain adapter

5.3.1 Transfer learning with FRCRN

To enhance monaural speech for drones, we develop a transfer learning-based method to estimate the $\hat{M}(K,l)$ using the FRCRN. FRCRN achieves state-of-the-art (SOTA) results in the recent DNS challenge [89], which includes a wide variety of noise types. Fig. 5.2a shows the FRCRN architecture with convolutional recurrent (CR) blocks, each comprised of a convolutional layer and a Feedforward Sequential Memory Network (FSMN) layer. This design enables the model to capture local temporal-spectral structures and long-range frequency dependencies, making it well-suited for exploiting the long-range frequency correlations found in drone noise [98].

5.3.2 Adapter tuning on FRCRN

Transfer learning includes fine-tuning and adapter tuning, both of which involve copying the parameters from a pre-trained FRCRN [1] and tuning them on the target data. The fine-tuning is trained on a subset of pre-trained model parameters, while the adapter-based tuning method only is trained on the adapter parameters but keeps the pre-trained model's parameters fixed, making adapter tuning more parameter-efficient.

We propose a frequency-domain bottleneck adapter to learn drone noise character-

[1] https://github.com/modelscope/modelscope

istics. Figure 5.2a shows the adapter embedded in the encoder module, positioned after each CR block. When tuning the drone dataset, the pre-trained model's parameters are frozen, and the adapter and the attention block parameters are fine-tuned to learn. The attention block acts as the skip pathway to facilitate information flow, hence it is kept unfrozen.

Figure 5.2b illustrates the structure of the adapter. The adapter processes features in the frequency domain and maintains consistent input and output dimensions. It includes a skip connection which helps in relaying the original information directly to subsequent layers, thus preserving its integrity and enhancing the model's expressive capacity. The adapter parameters are initially set to zero, allowing the adapter to function as an approximate identity function initially. This design ensures that the adapter can be seamlessly integrated into the model, gradually learning and adapting to new tasks or data without disrupting the performance of the pre-trained model.

The detailed operations of the adapter are as follows. The adapter consists of two cells for the real and imaginary parts of the input. We adopt a transformation that mimics complex multiplication in Cartesian form to preserve the algebraic structure of complex numbers. Specifically, given input components $\mathbf{A}_{\mathrm{r}}^{\mathrm{in}}$ and $\mathbf{A}_{\mathrm{i}}^{\mathrm{in}}$, the outputs are computed as:

$$\begin{aligned} \mathbf{A}_{\mathrm{r}}^{\mathrm{out}} &= \mathrm{Adapter}_{\mathrm{r}}\left(\mathbf{A}_{\mathrm{r}}^{\mathrm{in}}\right) - \mathrm{Adapter}_{\mathrm{i}}\left(\mathbf{A}_{\mathrm{i}}^{\mathrm{in}}\right), \\ \mathbf{A}_{\mathrm{i}}^{\mathrm{out}} &= \mathrm{Adapter}_{\mathrm{r}}\left(\mathbf{A}_{\mathrm{i}}^{\mathrm{in}}\right) + \mathrm{Adapter}_{\mathrm{i}}\left(\mathbf{A}_{\mathrm{r}}^{\mathrm{in}}\right). \end{aligned} \tag{5.6}$$

Fig. 5.2c illustrates the real cell operation. For a real part $U_r \in \mathbb{R}^{C\times F\times T}$ of a feature map $U \in \mathbb{R}^{C\times F\times T\times 2}$, C, F, T denote channel, frequency, and time frame dimensions, respectively, , applying the real cell of the adapter, $\mathrm{Adapter}_r$, yields the output $U_r' \in \mathbb{R}^{C\times F\times T}$:

$$\begin{aligned} \mathbf{h} &= \mathrm{ReLU}(\mathbf{W}_1 U_r + \mathbf{b}_1), \\ U_r' &= \mathbf{W}_2\mathbf{h} + \mathbf{b}_2. \end{aligned} \tag{5.7}$$

Here, ReLU represents the activation function [12]. $\mathbf{W}_1$ performs a frequency-domain downward projection, reducing F to $F/2$, and $\mathbf{W}_2$ is a frequency-domain upward projection, expanding $F/2$ back to F. The terms $\mathbf{b}_1$ and $\mathbf{b}_2$ are biases. Complex cell shares the same structure as the real cell.

5.3.3 Loss function

The original loss function for FRCRN combines scale-invariant SNR (SI-SNR) and the mean squared error (MSE) losses of cIRM estimates. In our method, we use only

SI-SNR as the loss function to avoid the need to balance two losses. The SI-SNR loss $\mathcal{L}(s, \hat{s})$ is defined as [101]:

$$\mathcal{L}(s, \hat{s}) = 10 \log_{10} \left(\frac{\| s_{\text{target}} \|_2^2}{\| e_{\text{noise}} \|_2^2} \right) \tag{5.8}$$

with $s_{\text{target}} = (\langle \hat{s}, s \rangle \cdot s)/\|s\|_2^2$ and $e_{\text{noise}} = \hat{s} - s_{\text{target}}$. Here, $\langle \cdot, \cdot \rangle$ denotes dot product and $\| \cdot \|_2$ is the L2 norm.

5.4 Evaluation on drone noise data

5.4.1 Dataset

We use clean speech and drone ego-noise to generate clean-noisy pairs for training, validation, and testing. Clean speech data is sourced from DNS-2022 [89] and LibriSpeech [102]. Table 5.1 details the drone ego-noise datasets, which include AS [86], AVQ [95], DREGON [84] and samples using DJI Phantom 2. The IDs for AS and AVQ are based on data entries from a public repository[2], while the IDs for DREGON and DJI describe flight states. The noise types are categorized based on flight conditions, specifically constant denotes noise levels are relatively stable, and dynamic denotes noise levels varying. For multichannel recordings, only single-channel data is used. Some audio samples have been trimmed to remove takeoff sequences, and all audio samples are resampled at 16 kHz.

The training set is generated using clean speech from DNS-2022 and drone ego-noise from AVQ (excluding n116), DREGON, and DJI. The validation set is generated using clean speech from DNS-2022 and noise from AVQ's n116. The test set is created with clean speech from LibriSpeech and drone ego-noise from AS. Clean speech segments and noise segments are randomly cropped to the same length and then mixed at an SNR varying from -25 to -5 dB. In total, we generate 5 hours of training data, 1 hour of validation data, and 1 hour of testing data. The average SNR is -15 dB. Considering that the duration of existing drone ego-noise datasets is short, we did not generate longer-duration data.

5.4.2 Evaluation

We compare the proposed method (**Adapter tuning**) with 3 methods: (i) a pre-trained FRCRN without tuning (**w/o tuning**); (ii) an untrained FRCRN, trained on the training data (**w/o pre-trained**); (iii) a fine-tuning method (**Fine-tuning**)

[2]https://zenodo.org/records/4553667

Table 5.1: Drone ego-noise dataset

Dataset	**ID**	**Noise type**	**Length [s]**
AS [86]	n121	constant	130
	n122	dynamic	140
AVQ [95]	n116	constant	120
	n117	constant	120
	n118	constant	40
	n119	constant	210
	n120	dynamic	214
DREGON [84]	Free Flight	dynamic	72
	Hovering	constant	25
	Up&Down	dynamic	28
	Rectangle	dynamic	25
	Spinning	constant	23
DJI	Free Flight	dynamic	60
	Hovering	constant	60

Table 5.2: Evaluation results

	Trainable	Evaluation Metrics		
	Params(m)	PESQ	ESTOI	SI-SNR
Noisy speech	-	1.06	0.09	-14.65
w/o tuning	-	1.04	0.32	-8.49
w/o pre-trained	14	1.25	0.36	2.56
Fine-tuning	1.2	1.12	0.32	2.64
Adapter tuning	**0.3**	**1.31**	**0.47**	**2.87**

where only the encoder module and attention block are trainable. To conduct a comprehensive evaluation, multiple evaluation metrics are used, including the Perceptual Evaluation of Speech Quality (PESQ) [103], Extended Short-Time Objective Intelligibility Measure (ESTOI) [104], and SI-SNR [101]. Additionally, we compare the efficiency of different methods in terms of the number of trainable parameters.

The results are presented in Table 5.2. Overall, **Adapter tuning** outperforms the competing methods for all metrics. The **w/o tuning** shows limited effectiveness, as it is not designed for drone scenarios. **w/o pre-trained** demonstrates improved performance, especially in PESQ and SI-SNR. This indicates the significant difference between drone noise and the noise in the normal training dataset. Although the FSMN layer in the encoder module processes frequency-domain information as the proposed adapter, **Fine-tuning** does not yield optimal results. This may be because **Fine-tuning** leads to the forgetting of previously learned information, whereas **Adapter tuning** preserves all pre-trained parameters. The trainable parameters indicate the proposed method's efficiency, surpassing **Fine-tuning** with

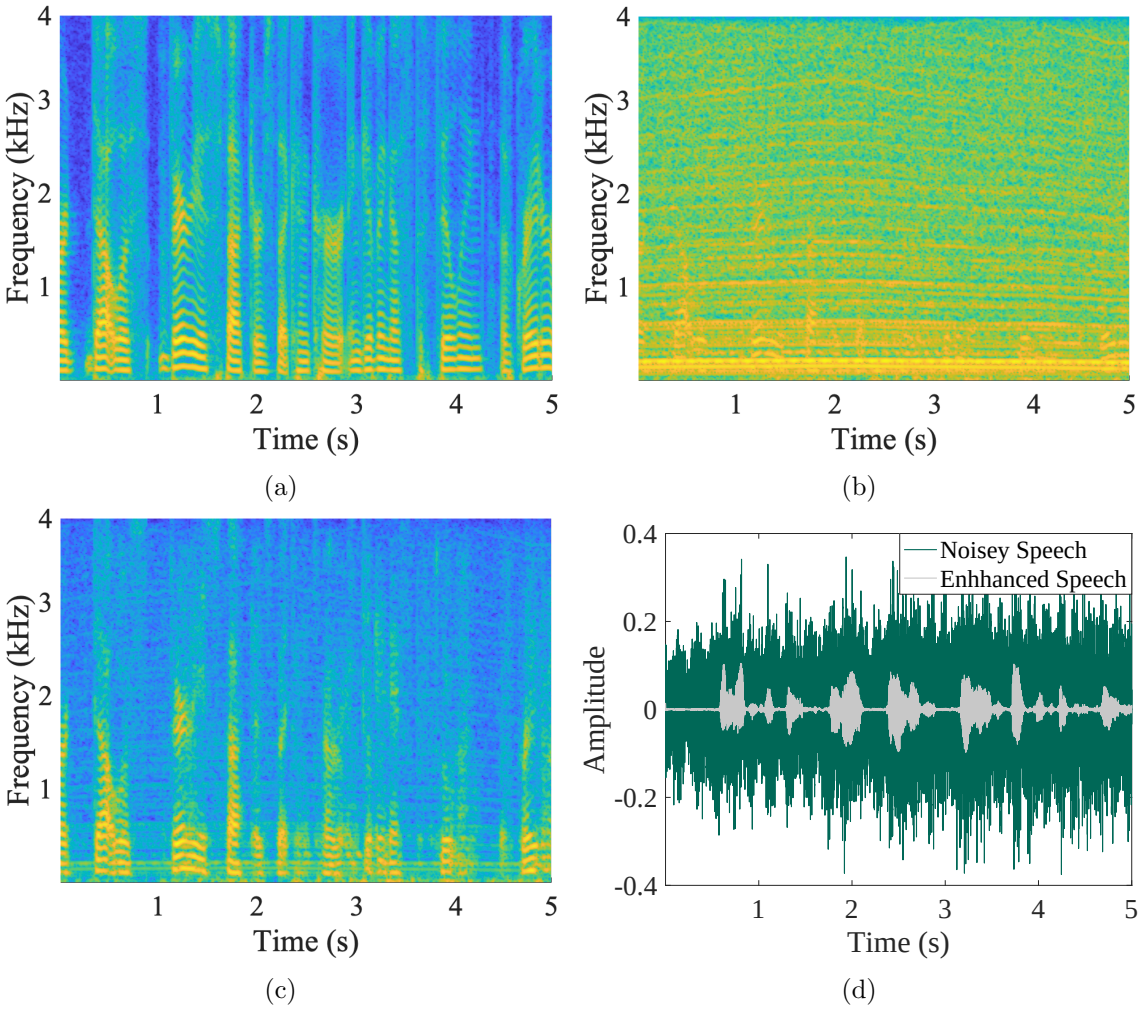


Figure 5.3: Results comparison: (a) clean speech (b) noisy speech (c) adapter tuning enhanced speech (d) noisy speech and adapter tuning enhanced speech in the time domain.

much less trainable parameters.

Figures 5.3a, 5.3b, and 5.3c illustrate the time-frequency spectra for segments of clean speech, noisy speech, and adapter-enhanced speech, respectively. Figure 5.3d illustrates the time-domain plot of speech before and after enhancement. As shown in Fig. 5.3b, in the noisy speech, the speech is obscured in the drone noise, and sound power is mainly concentrated in the harmonic frequencies. A comparison between Figures 5.3c and 5.3a reveals that the enhanced speech retains a sound power distribution similar to that of the clean speech, demonstrating the effectiveness of the proposed method. Fig. 5.3d shows that the enhanced speech is significantly clearer, with an 18 dB improvement in Signal-to-Noise Ratio (SNR).

5.5 Conclusions

Due to the cost and weight constraints, monaural speech enhancement for drones is preferred over multichannel speech enhancement. However, the absence of spatial information and the low SNR makes monaural speech enhancement challenging. By exploiting the harmonic nature of the drone ego noise, this thesis developed a frequency-domain bottleneck adapter for transfer learning. The method takes advantage of transfer learning to re-use knowledge learned from large data, thereby achieving effective training on small data. The proposed method enhances speech quality and intelligibility in drone recordings, paving the way for broader drone audition applications.

Chapter 6

Conclusions and Future Work

In this chapter, we state the general conclusions drawn from this thesis. We also outline some future research directions arising from this work.

6.1 Conclusions

This thesis has explored the integration of deep learning methods into acoustic and audio research, addressing several challenging aspects of spatial audio technologies. Starting with a foundation in spherical transformations, sound wave propagation, and deep learning, established in Chapter 2, each subsequent chapter has built upon this groundwork and developed innovative solutions for sound field estimation, HRTF interpolation, and speech enhancement.

Chapter 3 introduced a Physics-Informed Neural Network (PINN) method for estimating the sound field around a rigid sphere. This method leverages a Feedforward Neural Network (FNN) to approximate the function between spatial positions and sound pressure, incorporating the Helmholtz equation and boundary conditions directly into the loss function. By addressing the limitations inherent in traditional sound field estimation methods, which often struggle with spatial resolution constraints, the PINN approach showcases a robust capability.

Chapter 4 detailed the development of a spherical convolutional neural network to interpolate Head-Related Transfer Functions (HRTFs) from sparse measurements. By integrating spherical harmonic transformations directly, our method avoids the complexities of plane-sphere projections, thereby enhancing the interpolation accuracy. This method effectively uses sparse measurements to achieve accurate, dense HRTF interpolations, demonstrating the robustness of adapting to the high variability in individual ear and head anatomies.

In Chapter 5, we tackled the problem of monaural speech enhancement on drones, a field complicated by low signal-to-noise ratios. This study effectively leverages large datasets to facilitate training on limited data by designing a frequency domain

bottleneck adapter that incorporates transfer learning. The developed method not only enhances speech clarity and intelligibility but also specifically addresses the challenges posed by the noise characteristics of drones, which can significantly impact sound recording accuracy.

6.2 Future Work

Several future research problems arise from the work presented in this thesis. We outline potential directions for future research below.

6.2.1 Sound field estimation

Our research has focused on sound field estimation around rigid spheres in free-field conditions. Future work could extend these analyses to more complex acoustic environments. Notably, scenarios involving reverberant conditions, such as those found within enclosed spaces like rooms, or dealing with irregularly shaped objects, present significant challenges and opportunities for further research.

6.2.2 HRTF interpolation

We addressed the magnitude of HRTFs, while the phase component was not considered. Future research could employ spin-weighted spherical convolutions to directly manipulate HRTFs as complex values, enhancing the precision and applicability of our models. Additionally, while we have used sparse HRTF data, incorporating anthropometric features based on simple photos of users' heads could significantly improve personalization and commercial viability. Integrating these features with sparse HRTFs represents a promising direction for making HRTF interpolation more user-specific and practical.

6.2.3 Speech enhancement on drones

Our method of enhancing speech on drones has leveraged frequency domain techniques with transfer learning. Currently, this approach is implemented in a single-channel format as outlined in Chapter 5, serving as a foundational stage for our research. We plan to extend this work to multi-channel systems, aiming to more effectively capture expansive sound fields across multiple drones, thereby enhancing spatial audio recording over larger areas. There is potential to explore fine-tuning time-domain-based learning models, which may offer new insights and improvements. Expanding the use of various transfer learning methods could further enhance

the effectiveness and robustness of speech enhancement technologies, particularly in challenging drone environments.

Bibliography

[1] Xingyu Chen, Fei Ma, Amy Bastine, Prasanga Samarasinghe, and Huiyuan Sun, "Sound field estimation around a rigid sphere with physics-informed neural network," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 1984–1989.

[2] Xingyu Chen, Fei Ma, Yile Zhang, Amy Bastine, and Prasanga N Samarasinghe, "Head-related transfer function interpolation with a spherical cnn," *arXiv preprint arXiv:2309.08290*, 2023.

[3] Xingyu Chen, Hanwen Bi, Wei-Ting Lai, and Fei Ma, "Monaural speech enhancement on drone via adapter based transfer learning," *arXiv preprint arXiv:2405.10022*, 2024.

[4] Rodney A Kennedy and Parastoo Sadeghi, *Hilbert space methods in signal processing*, Cambridge University Press, 2013.

[5] Lawrence E Kinsler, Austin R Frey, Alan B Coppens, and James V Sanders, *Fundamentals of acoustics*, John wiley & sons, 2000.

[6] Earl G Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*, Academic press, 1999.

[7] Daniel P Jarrett, Emanuël AP Habets, and Patrick A Naylor, *Theory and applications of spherical microphone array processing*, vol. 9, Springer, 2017.

[8] Darren B Ward and Thushara D Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on speech and audio processing*, vol. 9, no. 6, pp. 697–707, 2001.

[9] DYN Zotkin, Jane Hwang, R Duraiswaini, and Larry S Davis, "Hrtf personalization using anthropometric measurements," in *2003 IEEE workshop on applications of signal processing to audio and acoustics (IEEE Cat. No. 03TH8684)*. Ieee, 2003, pp. 157–160.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning*, MIT press, 2016.

[11] Paul J Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[12] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[13] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al., "Recent advances in convolutional neural networks," *Pattern recognition*, vol. 77, pp. 354–377, 2018.

[14] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, pp. 1–40, 2016.

[15] Jeroen Breebaart and Christof Faller, *Spatial audio processing: MPEG surround and other applications*, John Wiley & Sons, 2007.

[16] Francis Rumsey, *Spatial audio*, Taylor & Francis, 2012.

[17] Mark A Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of the Audio Engineering Society*, vol. 53, no. 11, pp. 1004–1025, 2005.

[18] Jihui Zhang, Thushara D Abhayapala, Wen Zhang, Prasanga N Samarasinghe, and Shouda Jiang, "Active noise control over space: A wave domain approach," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 26, no. 4, pp. 774–786, 2018.

[19] Fei Ma, Wen Zhang, and Thushara Dheemantha Abhayapala, "Active control of outgoing broadband noise fields in rooms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 529–539, 2020.

[20] Earl G Williams and J Adin Mann III, "Fourier acoustics: sound radiation and nearfield acoustical holography," 2000.

[21] Boaz Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Transactions on speech and audio processing*, vol. 13, no. 1, pp. 135–143, 2004.

[22] Earl G Williams and Kazuhiro Takashima, "Vector intensity reconstructions in a volume surrounding a rigid spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 773–783, 2010.

[23] Thushara D Abhayapala and Darren B Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *2002*

IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2002, vol. 2, pp. II–1949.

[24] Boaz Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2149–2157, 2004.

[25] Zhun Wei and Xudong Chen, "Deep-learning schemes for full-wave nonlinear inverse scattering problems," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 4, pp. 1849–1860, 2018.

[26] Xudong Chen, Zhun Wei, Maokun Li, and Paolo Rocca, "A review of deep learning approaches for inverse scattering problems (invited review)," *Progress In Electromagnetics Research*, vol. 167, pp. 67–81, 2020.

[27] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[28] Francesc Lluis, Pablo Martinez-Nuevo, Martin Bo Moller, and Sven Ewan Shepstone, "Sound field reconstruction in rooms: Inpainting meets super-resolution," *The Journal of the Acoustical Society of America*, vol. 148, no. 2, pp. 649–659, 2020.

[29] Miklas Strøm Kristoffersen, Martin Bo Møller, Pablo Martínez-Nuevo, and Jan Østergaard, "Deep sound field reconstruction in real rooms: introducing the isobel sound field dataset," *arXiv preprint arXiv:2102.06455*, 2021.

[30] Kazuhide Shigemi, Shoichi Koyama, Tomohiko Nakamura, and Hiroshi Saruwatari, "Physics-informed convolutional neural network with bicubic spline interpolation for sound field estimation," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2022, pp. 1–5.

[31] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis, "Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations," *arXiv preprint arXiv:1711.10561*, 2017.

[32] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang, "Physics-informed machine learning," *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, 2021.

[33] Ruixian Liu and Peter Gerstoft, "Sd-pinn: Physics informed neural networks for spatially dependent pdes," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[34] Marco Olivieri, Mirco Pezzoli, Fabio Antonacci, and Augusto Sarti, "A physics-informed neural network approach for nearfield acoustic holography," *Sensors*, vol. 21, no. 23, pp. 7834, 2021.

[35] Nikolas Borrel-Jensen, Allan P Engsig-Karup, and Cheol-Ho Jeong, "Physics-informed neural networks for one-dimensional sound field predictions with parameterized sources and impedance boundaries," *JASA Express Letters*, vol. 1, no. 12, 2021.

[36] Fei Ma, Thushara D Abhayapala, Prasanga N Samarasinghe, and Xingyu Chen, "Physics informed neural network for head-related transfer function upsampling," *arXiv preprint arXiv:2307.14650*, 2023.

[37] Fei Ma, Thushara D Abhayapala, and Prasanga N Samarasinghe, "Circumvent spherical bessel function nulls for open sphere microphone arrays with physics informed neural network," *arXiv preprint arXiv:2308.00242*, 2023.

[38] Franco Scarselli and Ah Chung Tsoi, "Universal approximation using feed-forward neural networks: A survey of some existing methods, and some new results," *Neural networks*, vol. 11, no. 1, pp. 15–37, 1998.

[39] Munhum Park and Boaz Rafaely, "Sound-field analysis by plane-wave decomposition using spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 3094–3103, 2005.

[40] Boaz Rafaely, *Fundamentals of spherical array processing*, vol. 8, Springer, 2015.

[41] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[42] Rishabh Gupta, Jianjun He, Rishabh Ranjan, Woon-Seng Gan, Florian Klein, Christian Schneiderwind, Annika Neidhardt, Karlheinz Brandenburg, and Vesa Välimäki, "Augmented/mixed reality audio for hearables: Sensing, control, and rendering," *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 63–89, 2022.

[43] Song Li and Jürgen Peissig, "Measurement of head-related transfer functions: A review," *Applied Sciences*, vol. 10, no. 14, pp. 5014, 2020.

[44] Robert Pelzer, Manoj Dinakaran, Fabian Brinkmann, Steffen Lepa, Peter Grosche, and Stefan Weinzierl, "Head-related transfer function recommendation based on perceptual similarities and anthropometric features," *The*

Journal of the Acoustical Society of America, vol. 148, no. 6, pp. 3809–3817, 2020.

[45] Corentin Guezenoc and Renaud Seguier, "Hrtf individualization: A survey," in *Audio Engineering Society Convention 145*, 2018.

[46] Durand R Begault and Leonard J Trejo, "3-d sound for virtual reality and multimedia," Tech. Rep., 2000.

[47] Fábio P Freeland, Luiz WP Biscainho, and Paulo SR Diniz, "Interpositional transfer function for 3d-sound generation," *Journal of the Audio Engineering Society*, vol. 52, no. 9, pp. 915–930, 2004.

[48] Klaus Hartung, Jonas Braasch, and Susanne J Sterbing, "Comparison of different methods for the interpolation of head-related transfer functions," in *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*. Audio Engineering Society, 1999.

[49] Dmitry N Zotkin, Ramani Duraiswami, and Nail A Gumerov, "Regularized hrtf fitting using spherical harmonics," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2009, pp. 257–260.

[50] Jens Ahrens, Mark RP Thomas, and Ivan Tashev, "Hrtf magnitude modeling using a non-regularized least-squares fit of spherical harmonics coefficients on incomplete data," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–5.

[51] Bo-Sun Xie, "Recovery of individual head-related transfer functions from a small set of measurements," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 282–294, 2012.

[52] Mengfan Zhang, Zhongshu Ge, Tiejun Liu, Xihong Wu, and Tianshu Qu, "Modeling of individual hrtfs based on spatial principal component analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 785–797, 2020.

[53] Yuki Ito, Tomohiko Nakamura, Shoichi Koyama, and Hiroshi Saruwatari, "Head-related transfer function interpolation from spatially sparse measurements using autoencoder with source position conditioning," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2022, pp. 1–5.

[54] You Zhang, Yuxiang Wang, and Zhiyao Duan, "Hrtf field: Unifying mea-

sured hrtf magnitude representation with neural fields," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[55] Yoshiki Masuyama, Gordon Wichern, François G Germain, Zexu Pan, Sameer Khurana, Chiori Hori, and Jonathan Le Roux, "Niirf: Neural iir filter field for hrtf upsampling and personalization," *arXiv preprint arXiv:2402.17907*, 2024.

[56] Devansh Zurale, Shahrokh Yadegari, and Shlomo Dubnov, "Deep hrtf encoding & interpolation: Exploring spatial correlations using convolutional neural networks," in *19th Sound and Music Computing Conference, SMC 2022*. Sound and Music Computing Network, 2022, pp. 350–357.

[57] Ziran Jiang, Jinqiu Sang, Chengshi Zheng, Andong Li, and Xiaodong Li, "Modeling individual head-related transfer functions from sparse measurements using a convolutional neural network," *The Journal of the Acoustical Society of America*, vol. 153, no. 1, pp. 248–259, 2023.

[58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[59] Aidan OT Hogg, Mads Jenkins, He Liu, Isaac Squires, Samuel J Cooper, and Lorenzo Picinali, "Hrtf upsampling with a generative adversarial network using a gnomonic equiangular projection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[60] Devansh Zurale and Shlomo Dubnov, "Spatial upsampling of sparse head related transfer functions-a vq-vae & transformer based approach," in *Audio Engineering Society Conference: AES 2023 International Conference on Spatial and Immersive Audio*. Audio Engineering Society, 2023.

[61] Devansh Zurale and Shlomo Dubnov, "Learning sub-dimensional hrtf representations towards individualization applications-traditional and deep learning approaches," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2023, pp. 1–5.

[62] Aidan OT Hogg, Mads Jenkins, He Liu, Isaac Squires, Samuel J Cooper, and Lorenzo Picinali, "Hrtf upsampling with a generative adversarial network us-

ing a gnomonic equiangular projection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[63] Chiyu Jiang, Jingwei Huang, Karthik Kashinath, Philip Marcus, Matthias Niessner, et al., "Spherical cnns on unstructured grids," *arXiv preprint arXiv:1901.02039*, 2019.

[64] Nathanaël Perraudin, Michaël Defferrard, Tomasz Kacprzak, and Raphael Sgier, "Deepsphere: Efficient spherical convolutional neural network with healpix sampling for cosmological applications," *Astronomy and Computing*, vol. 27, pp. 130–146, 2019.

[65] Pim De Haan, Maurice Weiler, Taco Cohen, and Max Welling, "Gauge equivariant mesh cnns: Anisotropic convolutions on geometric graphs," *arXiv preprint arXiv:2003.05425*, 2020.

[66] Taco Cohen and Max Welling, "Group equivariant convolutional networks," in *International conference on machine learning*. PMLR, 2016, pp. 2990–2999.

[67] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis, "Learning so (3) equivariant representations with spherical cnns," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 52–68.

[68] Risi Kondor and Shubhendu Trivedi, "On the generalization of equivariance and convolution in neural networks to the action of compact groups," in *International conference on machine learning*. PMLR, 2018, pp. 2747–2755.

[69] Carlos Esteves, Ameesh Makadia, and Kostas Daniilidis, "Spin-weighted spherical cnns," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8614–8625, 2020.

[70] Carlos Esteves, Yinshuang Xu, Christine Allen-Blanchette, and Kostas Daniilidis, "Equivariant multi-view networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1568–1577.

[71] Erik J Bekkers, Maxime W Lafarge, Mitko Veta, Koen AJ Eppenhof, Josien PW Pluim, and Remco Duits, "Roto-translation covariant convolutional networks for medical image analysis," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer, 2018, pp. 440–448.

[72] Chao Zhang, Stephan Liwicki, William Smith, and Roberto Cipolla,

"Orientation-aware semantic segmentation on icosahedron spheres," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3533–3541.

[73] Seungbo Ha and Ilwoo Lyu, "Spharm-net: Spherical harmonics-based convolution for cortical parcellation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2739–2751, 2022.

[74] Tianle Zhong, Israel Mendoza Velázquez, Yi Ren, Héctor Manuel Pérez Meana, and Yoichi Haneda, "Spherical convolutional recurrent neural network for real-time sound source tracking," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 5063–5067.

[75] Etienne Thuillier, Craig Jin, and Vesa Välimäki, "Hrtf interpolation using a spherical neural process meta-learner," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[76] Julius O. Smith, "Techniques for digital filter design and system identification with application to the violin," M.S. thesis, Stanford University, 06/1983 1983.

[77] Doris J Kistler and Frederic L Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1637–1647, 1992.

[78] James R Driscoll and Dennis M Healy, "Computing fourier transforms and convolutions on the 2-sphere," *Advances in applied mathematics*, vol. 15, no. 2, pp. 202–250, 1994.

[79] Jingwei Xi, Wen Zhang, and Thushara D Abhayapala, "Magnitude modelling of individualized hrtfs using dnn based spherical harmonic analysis," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 266–270.

[80] F. Brinkmann, M. Dinakaran, R. Pelzer, J. J. Wohlgemuth, F. Seipel, D. Voss, P. Grosche, and S. Weinzierl, "The hutubs head-related transfer function (hrtf) database," 2019.

[81] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[82] Meysam Basiri, Felix Schill, Pedro U Lima, and Dario Floreano, "Robust acoustic source localization of emergency signals from micro air vehicles," in

2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2012, pp. 4737–4742.

[83] D. Mukhutdinov, A. Alex, A. Cavallaro, and L. Wang, "Deep learning models for single-channel speech enhancement on drones," *IEEE Access*, vol. 11, pp. 22993–23007, 2023.

[84] Martin Strauss, Pol Mordel, Victor Miguet, and Antoine Deleforge, "Dregon: Dataset and methods for uav-embedded sound source localization," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.

[85] Lin Wang and Andrea Cavallaro, "Ear in the sky: Ego-noise reduction for auditory micro aerial vehicles," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2016, pp. 152–158.

[86] Lin Wang and Andrea Cavallaro, "Acoustic sensing from a multi-rotor drone," *IEEE Sensors Journal*, vol. 18, no. 11, pp. 4570–4582, 2018.

[87] Israel Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on speech and audio processing*, vol. 11, no. 5, pp. 466–475, 2003.

[88] S. S. Haykin, *Adaptive filter theory*, Pearson Education India, 2002.

[89] H. Dubey, V. Gopal, R. Cutler, S. Matusevych, S. Braun, E. S. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "Icassp 2022 deep noise suppression challenge," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2022.

[90] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Interspeech*, 2018, vol. 2018, pp. 3229–3233.

[91] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv preprint arXiv:2008.00264*, 2020.

[92] Shengkui Zhao, Bin Ma, Karn N Watcharasupat, and Woon-Seng Gan, "Frcrn: Boosting feature representation using frequency recurrence for monaural speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9281–9285.

[93] Zhihao Wang, Jian Chen, and Steven CH Hoi, "Deep learning for image super-resolution: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.

[94] O. Ruiz-Espitia, J. Martinez-Carranza, and Rasco., "Aira-uas: an evaluation corpus for audio processing in unmanned aerial system," in *2018 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE, 2018, pp. 836–845.

[95] Lin Wang, Ricardo Sanchez-Matilla, and Andrea Cavallaro, "Audio-visual sensing from a quadcopter: dataset and baselines for source localization and sound enhancement," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5320–5325.

[96] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[97] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, De L.Q., M. Gesmundo, A.and Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning (ICML)*. PMLR, 2019, pp. 2790–2799.

[98] Giorgia Sinibaldi and Luca Marino, "Experimental analysis on the noise of propellers for small uav," *Applied Acoustics*, vol. 74, no. 1, pp. 79–88, 2013.

[99] Hanwen Bi, Fei Ma, Thushara D Abhayapala, and Prasanga N Samarasinghe, "Spherical array based drone noise measurements and modelling for drone noise reduction via propeller phase control," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 286–290.

[100] Donald S Williamson, Yuxuan Wang, and DeLiang Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[101] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "Sdr–half-baked or well done?," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[102] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE*

international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015, pp. 5206–5210.

[103] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. IEEE, 2001, vol. 2, pp. 749–752.

[104] Jesper Jensen and Cees H Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.