MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Physics-Informed Direction-Aware Neural Acoustic Fields

Masuyama, Yoshiki; Germain, François G; Wichern, Gordon; Ick, Christopher; Le Roux, Jonathan

TR2025-142 October 14, 2025

Abstract

This paper presents a physics-informed neural network (PINN) for modeling first-order Ambisonic (FOA) room impulse responses (RIRs). PINNs have demonstrated promising performance in sound field interpolation by combining the powerful modeling capability of neural networks and the physical principles of sound propagation. In room acoustics, PINNs have typically been trained to represent the sound pressure measured by omnidirectional microphones where the wave equation or its frequency-domain counterpart, i.e., the Helmholtz equation, is leveraged. Meanwhile, FOA RIRs additionally provide spatial characteristics and are useful for immersive audio generation with a wide range of applications. In this paper, we extend the PINN framework to model FOA RIRs. We derive two physics-informed priors for FOA RIRs based on the correspondence between the particle velocity and the (X, Y, Z)-channels of FOA. These priors associate the predicted W - channel and other channels through their partial derivatives and impose the physically feasible relationship on the four channels. Our experiments confirm the effectiveness of the proposed method compared with a neural network without the physics-informed prior.

IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2025

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Physics-Informed Direction-Aware Neural Acoustic Fields

Yoshiki Masuyama[1], *François G. Germain*[1], *Gordon Wichern*[1], *Christopher Ick*[1,2], *Jonathan Le Roux*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]Music and Audio Research Laboratory, New York University, NY, USA

***Abstract*—This paper presents a physics-informed neural network (PINN) for modeling first-order Ambisonic (FOA) room impulse responses (RIRs). PINNs have demonstrated promising performance in sound field interpolation by combining the powerful modeling capability of neural networks and the physical principles of sound propagation. In room acoustics, PINNs have typically been trained to represent the sound pressure measured by omnidirectional microphones where the wave equation or its frequency-domain counterpart, i.e., the Helmholtz equation, is leveraged. Meanwhile, FOA RIRs additionally provide spatial characteristics and are useful for immersive audio generation with a wide range of applications. In this paper, we extend the PINN framework to model FOA RIRs. We derive two physics-informed priors for FOA RIRs based on the correspondence between the particle velocity and the (X, Y, Z)-channels of FOA. These priors associate the predicted W-channel and other channels through their partial derivatives and impose the physically feasible relationship on the four channels. Our experiments confirm the effectiveness of the proposed method compared with a neural network without the physics-informed prior.**

1. INTRODUCTION

Room impulse responses (RIRs) capture sound propagation in indoor environments and serve as a fundamental component of room acoustics [1]. In addition to the direct sound, an RIR involves reflections and diffraction depending on the materials of the room surface and the positions of the sound source and microphone. RIRs dramatically vary in each room, and their precise modeling is crucial for high-quality immersive audio experiences in mixed reality applications [2], [3]. Furthermore, RIR modeling is useful to enhance various downstream tasks, including dereverberation [4]. RIRs are typically measured by using a loudspeaker and one or more microphones placed within a room. The measurement process requires an enormous amount of time and effort to cover a large area with dense spatial grids. To alleviate this issue, spatial interpolation from a set of sparse measurements is in great demand, and various techniques have been developed [5]–[9], notably using compressed sensing [10]–[12].

Recently, neural networks have been leveraged in RIR interpolation and generation, motivated by their adaptive and powerful modeling capabilities [13]–[18]. In particular, neural fields (NFs) have gained increasing attention due to their flexibility [19]–[22]. An NF is trained to predict the sound pressure from a given microphone position and optionally other contexts, e.g., the source position, on the sparse measurements. Once it is trained, we can predict RIRs at arbitrary positions in a grid-less manner. However, NFs are typically trained from scratch for each room, which leads to overfitting and results in poor generalization capability to unmeasured positions. More fundamentally, NFs do not respect the physical principles of sound propagation when trained solely to minimize reconstruction error.

Physics-informed neural networks (PINNs) [23], [24] have mitigated this issue by enforcing the network output to follow the physics of sound propagation [25]–[30]. By exploiting automatic differentiation, PINNs efficiently calculate the derivatives of the network outputs with respect to the inputs, e.g., the microphone position. During training, the NF is then regularized to minimize the discrepancy of the network outputs and their derivatives from the governing partial differential equation (PDE). For the sound field reconstruction, the time-domain wave equation or its frequency-domain counterpart, i.e., the Helmholtz equation, have been widely used.

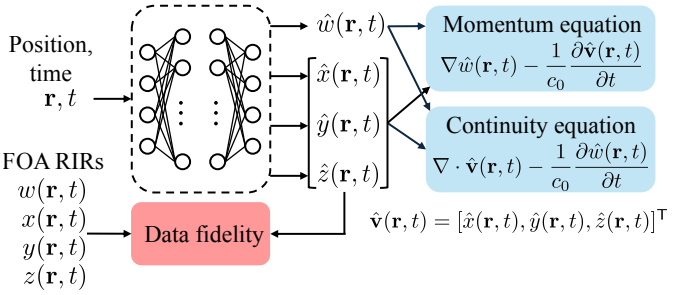

Fig. 1: Illustration of the proposed PINN for FOA RIR modeling, PI-DANF. The blue boxes show the physics-informed priors tailored for Ambisonics.

Along with the modeling of omnidirectional RIRs, which is now a mature field, the modeling of spatial RIRs has gained increasing attention to render binaural audio according to listener orientation [31]–[33]. In particular, first-order Ambisonics (FOA) [34]–[36] encodes the directional characteristics of sound field based on spherical harmonic decomposition [37]. The zeroth-order component known as the W-channel corresponds to the sound pressure measured by an omnidirectional microphone. Meanwhile, the first-order components, i.e., the (X, Y, Z)-channels, encode the particle velocity along the Cartesian directions. The generation and interpolation of FOA RIRs have improved with the use of neural networks [38], [39]. In particular, the direction-aware neural acoustic field (DANF) learns to predict the FOA RIR, i.e., the (W, X, Y, Z)-channels, from the microphone position and the context, facilitating grid-less reconstruction of spatial RIRs [39]. Consequently, DANF can be used to efficiently decode binaural audio depending on listener orientation at any position. DANF optimization has relied solely on data fidelity to sparsely measured FOA RIRs, and it should be beneficial to consider physical principles.

In this paper, we propose PI-DANF, a physics-informed extension of DANF to generate acoustically valid FOA RIRs. We leverage the analytical relationship between the (X, Y, Z)-channels and the particle velocity, and derive physics-informed priors from the linearized momentum equation and the continuity equation as illustrated in Fig. 1. Since the wave equation is derived by combining these PDEs, PI-DANF can be interpreted as a natural extension of the PINN used for omnidirectional RIR [28]. Our experiments on simulated data validate the effectiveness of PI-DANF compared to the vanilla DANF and to DANF with the previous wave-equation-based prior on the W-channel.

2. PRELIMINARIES

2.1. Sound Field Reconstruction Using Neural Networks

Let us denote the sound pressure $p(\mathbf{r}, t) \in \mathbb{R}$ as a function of position $\mathbf{r} \in \mathbb{R}^3$ in Cartesian coordinates and time $t \in \mathbb{R}$. In particular, we consider the sound field in a region $\Omega \subset \mathbb{R}^3$ that does not contain sources. The sound pressure is assumed to follow two PDEs under

regular conditions [40]. The first equation is the linearized momentum equation given by

$$\nabla p(\mathbf{r},t) + \rho_0 \frac{\partial \mathbf{u}(\mathbf{r},t)}{\partial t} = \mathbf{0}, \tag{1}$$

where $\mathbf{u}(\mathbf{r},t) \in \mathbb{R}^3$ is the particle velocity, ∇ denotes the gradient with respect to $\mathbf{r}$, and ρ_0 is the density of the air. The other equation is derived from the continuity equation as follows:

$$\rho_0 \nabla \cdot \mathbf{u}(\mathbf{r},t) + \frac{1}{c_0^2}\frac{\partial p(\mathbf{r},t)}{\partial t} = 0, \tag{2}$$

where $\nabla\cdot$ denotes the divergence, and c_0 is the sound speed. On the basis of (1)–(2), the wave equation is derived as follows:

$$\Delta p(\mathbf{r},t) - \frac{1}{c_0^2}\frac{\partial^2 p(\mathbf{r},t)}{\partial t^2} = 0, \tag{3}$$

where Δ denotes the Laplacian.

We aim to reconstruct the continuous sound field $p(\mathbf{r},t)$ from its sparse measurements at $\mathbf{r}_d \in \Omega$, where $d = 0,\dots,D-1$ is the index of the microphone position. Sound field reconstruction can be realized by optimizing an NF to map the given position and time to the corresponding sound pressure [20], [21]:

$$p(\mathbf{r},t) \approx \hat{p}(\mathbf{r},t) = \mathtt{NF}_\theta(\mathbf{r},t), \tag{4}$$

where θ denotes the parameters of the NF. The NF is trained to minimize the following reconstruction error of the measurements:

$$\mathcal{L}_{\text{data}} = \frac{1}{DL}\sum_{d=0}^{D-1}\sum_{l=0}^{L-1} |\hat{p}(\mathbf{r}_d,t_l) - p(\mathbf{r}_d,t_l)|, \tag{5}$$

where $l = 0,\dots,L-1$ is the sample index of the measured RIRs. Once the NF is trained, it can predict sound pressure at any position and time in a grid-less manner.

The training of NFs solely with the loss function in (5) does not consider the properties of sound propagation, resulting in poor generalization capability to unmeasured positions. The PINN framework mitigates this issue by incorporating a prior term derived from the physical principles of the sound propagation. Specifically, the prior term derived from the wave equation in (3) has been widely used for time-domain modeling [27], [28]:

$$\mathcal{L}_{\text{wave}} = \mathbb{E}_{\mathbf{r}\in\Omega}\mathbb{E}_{t\in[0,T]}\left|\Delta\hat{p}(\mathbf{r},t) - \frac{1}{c_0^2}\frac{\partial^2\hat{p}(\mathbf{r},t)}{\partial t^2}\right|, \tag{6}$$

where $\mathcal{L}_{\text{wave}}$ measures the discrepancy from the wave equation. This prior term can be efficiently computed at any position $\mathbf{r} \in \Omega$ and time $t \in [0,T]$ by using automatic differentiation, where $T \in \mathbb{R}_+$ is for limiting the time range. PINNs have demonstrated promising performance under various conditions compared with compressed sensing and other deep learning methods [27].

2.2. First-Order Ambisonics (FOA)

FOA, often referred to as simply Ambisonics, is an audio format that captures the spatial characteristics of a sound field by using the spherical harmonics up to the first order [34]–[36]. This representation is beneficial for a wide range of applications, including mixed reality, and has been used for 3D spatial audio coding [41]. FOA encodes the sound field into four channels, where the W-channel matches the sound pressure measured by a virtual omnidirectional microphone up to a multiplicative constant. Meanwhile, the (X,Y,Z)-channels are derived from the spherical harmonics of the first order and can be interpreted as measurements from virtual dipole microphones along the Cartesian directions. Under the SN3D normalization [42], the sound pressure $p(\mathbf{r},t)$ is equivalent to $w(\mathbf{r},t)$, and the particle velocity $\mathbf{u}(\mathbf{r},t)$ is given by

$$\mathbf{u}(\mathbf{r},t) = -\frac{1}{\rho_0 c_0}\mathbf{v}, \tag{7}$$

$$\mathbf{v}(\mathbf{r},t) = [x(\mathbf{r},t), y(\mathbf{r},t), z(\mathbf{r},t)]^\mathsf{T}, \tag{8}$$

where $w(\mathbf{r},t)$, $x(\mathbf{r},t)$, $y(\mathbf{r},t)$, and $z(\mathbf{r},t)$ are the channels of the FOA RIR measured at position $\mathbf{r}$ and time t [31].

Recently, deep generative models have been developed for audio in the FOA format [43]–[45]. Related to our work, a paper proposes a generative adversarial network for FOA RIRs [38]. These previous studies mainly focus on generating plausible audio, while we aim to interpolate FOA RIRs in a physically feasible manner.

3. PROPOSED METHOD

3.1. Direction-Aware Neural Field (DANF)

While the (X,Y,Z)-channels of FOA RIRs can be computed using a PINN for sound pressure in (4) through the linearized momentum equation in (1), this requires the numerical integration of the spatial gradient over time, which makes binaural decoding too complex [28]. As an alternative, DANF has been proposed to extend an NF from predicting omnidirectional RIRs to representing FOA RIRs [39]:

$$\hat{w}(\mathbf{r},t), \hat{\mathbf{v}}(\mathbf{r},t) = \mathtt{DANF}_\theta(\mathbf{r},t). \tag{9}$$

The network is optimized by minimizing the reconstruction error of sparse measurements[1]:

$$\mathcal{L}_{\text{data}} = \frac{1}{DL}\sum_{d=0}^{D-1}\sum_{l=0}^{L-1}(|\hat{w}(\mathbf{r}_d,t_l) - w(\mathbf{r}_d,t_l)| + \|\hat{\mathbf{v}}(\mathbf{r}_d,t_l) - \mathbf{v}(\mathbf{r}_d,t_l)\|_1), \tag{10}$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm. DANF directly provides spatial information at the given position, which allows binaural decoding depending on listener orientation at any position.

In the original paper [39], DANF struggled to predict reliable FOA RIRs at unmeasured positions when trained from scratch with a limited number of measurements. Although pre-training DANF on large-scale RIR datasets was proposed to alleviate this issue, it does not guarantee adherence to the physical principles of sound propagation. We expect the PINN framework to address this problem.

3.2. Physics-Informed Prior for Ambisonics

A naive way to adopt a PINN framework for DANF is to apply the wave-equation-based prior in (6) to the W-channel because that channel corresponds to the sound pressure. This training, however, only regularizes the W-channel and leaves the (X,Y,Z)-channels unconstrained. Thus, it may not be sufficient to interpolate FOA RIRs with appropriate spatial characteristics.

We thus propose two physics-informed priors tailored for FOA RIRs. As described in Section 2.2, the (X,Y,Z)-channels are analytically related to the particle velocity in (7). By substituting $\hat{w}(\mathbf{r},t)$ and $-\hat{\mathbf{v}}(\mathbf{r},t)/(c_0\rho_0)$ respectively into $p(\mathbf{r},t)$ and $\mathbf{u}(\mathbf{r},t)$ in the linearized momentum equation (1), the first penalty term is formulated as

$$\mathcal{L}_{\text{momentum}} = \mathbb{E}_{\mathbf{r}\in\Omega}\mathbb{E}_{t\in[0,T]}\left\|\nabla\hat{w}(\mathbf{r},t) - \frac{1}{c_0}\frac{\partial\hat{\mathbf{v}}(\mathbf{r},t)}{\partial t}\right\|_1. \tag{11}$$

[1]The original paper also employs additional losses in the short-time Fourier transform domain, but all the losses concern the fidelity to the measured FOA RIRs [39].

The other prior is derived by substituting the network outputs into the continuity equation (2) and multiplying by c_0:

$$\mathcal{L}_{\text{continuity}} = \mathbb{E}_{\mathbf{r}\in\Omega}\mathbb{E}_{t\in[0,T]}\left|\nabla\cdot\hat{\mathbf{v}}(\mathbf{r},t) - \frac{1}{c_0}\frac{\partial\hat{w}(\mathbf{r},t)}{\partial t}\right|. \quad (12)$$

These proposed priors connect the predictions of all channels, in contrast to the wave-equation-based prior which focuses solely on the W-channel. Furthermore, they penalize the absolute value of four scalars per position $\mathbf{r}$ and time t, while the previous one penalizes one scalar. That is, the proposed priors more precisely reflect the physical principles of FOA RIRs.

3.3. Network Architecture and Training Setup

To realize PI-DANF, we follow the network architecture and training scheme of the PINN for sound pressure [28]. For the network, the modified multi-layer perceptron is used to perform adaptive processing. The network input $\mathbf{h}_0 = [\mathbf{r}^\mathsf{T}, t]^\mathsf{T}$ is passed to two encoders:

$$\mathbf{f} = \texttt{SIREN}_f(\mathbf{h}_0), \quad (13)$$

$$\mathbf{g} = \texttt{SIREN}_g(\mathbf{h}_0), \quad (14)$$

where SIREN is the fully connected layer with sine activation [46]. This is advantageous for PINN because the periodic activation function provides well-behaved derivatives. Then, the forward process of the k-th hidden layer is defined as

$$\tilde{\mathbf{h}}_k = \texttt{SIREN}_k(\mathbf{h}_{k-1}), \quad (15)$$

$$\mathbf{h}_k = (1-\tilde{\mathbf{h}}_k)\odot\mathbf{f} + \tilde{\mathbf{h}}_k\odot\mathbf{g}, \quad (16)$$

where $\odot$ denotes Hadamard product, and $k = 1,\dots,K$. The output of the final hidden layer $\mathbf{h}_K$ is then projected to obtain the FOA RIR estimate $[\hat{w}(\mathbf{r},t), \hat{\mathbf{v}}(\mathbf{r},t)^\mathsf{T}]^\mathsf{T}$.

The network is then optimized to minimize the data fidelity term and the sum of the proposed priors with variable weights $(\epsilon_{\text{data}}, \epsilon_{\text{prior}})$:

$$\mathcal{L}_{\text{all}} = \frac{1}{2\epsilon_{\text{data}}^2}\mathcal{L}_{\text{data}} + \frac{1}{2\epsilon_{\text{prior}}^2}(\mathcal{L}_{\text{momentum}} + \mathcal{L}_{\text{continuity}}) + \log(\epsilon_{\text{data}}\epsilon_{\text{prior}}), \quad (17)$$

where the weights are adaptively updated together with the network parameters θ [47]. The limit of $\epsilon_{\text{prior}} \to \infty$ corresponds to the vanilla DANF without the physics-informed priors. For the priors, we stochastically take the position $\mathbf{r} \in \Omega$ and time $t \in [0, T]$ in each iteration by using Latin hypercube sampling as in [28].

4. EXPERIMENTS

4.1. Dataset and experimental setup

We evaluate the proposed PI-DANF with FOA RIRs simulated by HARP[2] [48]. HARP is built on top of Pyroomacoustics [49] and generates high-order Ambisonics RIRs by combining the image source method with specific directivity patterns. While the original dataset provides $100,000$ RIRs under various room conditions, we simulated FOA RIRs for 10 rooms with a random geometric setup and used their early part up to 100 ms with a sampling rate of 8 kHz. Each room was a shoebox as depicted in Fig. 2, and the target region Ω was a cuboid of dimensions 1.0 m × 1.0 m × 1.0 m. The sound source was randomly placed anywhere in the room except for the red area near the walls. We simulated impulse responses on a grid of 5 cm width in Cartesian coordinates, resulting in $21 \times 21 \times 21 = 9,261$ measurements. We randomly chose $\{250, 500\}$ measurements to train the NFs and used 50 other measurements for validation. We evaluated the network on the remaining $8,711$ measurements.

[2]https://github.com/whojavumusic/HARP

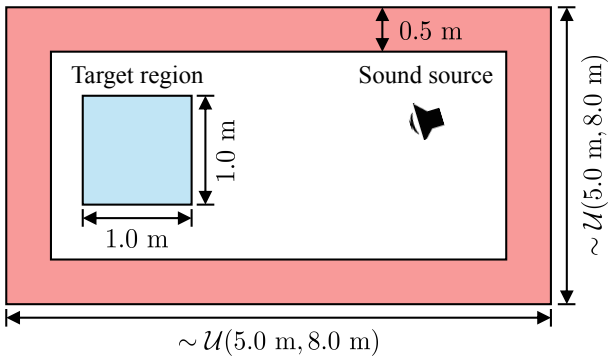

Fig. 2: Room geometry for simulation. The room height was sampled from $\mathcal{U}(2.5\text{ m}, 4.0\text{ m})$.

Our network architecture and training setups mainly followed an implementation of the PINN for omnidirectional RIRs[3] [28]. The NFs consisted of 3 hidden layers, i.e., $K = 3$, with 512 units. They were optimized with the Adam optimizer and cosine annealing. In each iteration, we used all the measured positions $(\mathbf{r}_d)_{d=0}^{D-1}$ but randomly sampled 250 discrete times t_l to compute the data fidelity term. Meanwhile, for the physics-informed priors, $25,000$ pairs of $\mathbf{r}$ and t were taken by Latin hypercube sampling. We set the initial values of ϵ_{data} and ϵ_{prior} to 1.0 and 0.1, respectively. The NFs were trained for $100,000$ iterations.

The predicted FOA RIRs were evaluated by the normalized mean squared error (NMSE) for the W-channel and (X, Y, Z)-channels:

$$\text{NMSE}(w,\hat{w}) = 10\log_{10}\left(\frac{\sum_{\tilde{d}=0}^{\tilde{D}-1}\|\hat{\mathbf{w}}(\mathbf{r}_{\tilde{d}}) - \mathbf{w}(\mathbf{r}_{\tilde{d}})\|_2^2}{\sum_{\tilde{d}=0}^{\tilde{D}-1}\|\mathbf{w}(\mathbf{r}_{\tilde{d}})\|_2^2}\right), \quad (18)$$

where $\tilde{d} = 0,\dots,\tilde{D}$ indexes the evaluation positions, $\mathbf{w}(\mathbf{r}_{\tilde{d}}) = [w(\mathbf{r}_{\tilde{d}}, t_0),\dots,w(\mathbf{r}_{\tilde{d}}, t_{L-1})]^\mathsf{T}$, and $\|\cdot\|_2$ denotes the ℓ_2 norm. A recent study indicates that the correlation to the reference RIR at the W- and (X, Y, Z)-channels match the listening quality and perceptual localization accuracy [50]. We thus measured the Pearson's correlation coefficients as follows:

$$\text{PCC}(w,\hat{w}) = \frac{1}{\tilde{D}}\sum_{\tilde{d}=0}^{\tilde{D}-1}\frac{[\hat{\mathbf{w}}(\mathbf{r}_{\tilde{d}}) - \underline{\hat{\mathbf{w}}}(\mathbf{r}_{\tilde{d}})]^\mathsf{T}[\mathbf{w}(\mathbf{r}_{\tilde{d}}) - \underline{\mathbf{w}}(\mathbf{r}_{\tilde{d}})]}{\|\hat{\mathbf{w}}(\mathbf{r}_{\tilde{d}}) - \underline{\hat{\mathbf{w}}}(\mathbf{r}_{\tilde{d}})\|_2\|\mathbf{w}(\mathbf{r}_{\tilde{d}}) - \underline{\mathbf{w}}(\mathbf{r}_{\tilde{d}})\|_2}, \quad (19)$$

where $\underline{\hat{\mathbf{w}}}(\mathbf{r}_{\tilde{d}})$ and $\underline{\mathbf{w}}(\mathbf{r}_{\tilde{d}})$ are the time average of $\hat{w}(\mathbf{r}_{\tilde{d}}, t_l)$ and $w(\mathbf{r}_{\tilde{d}}, t_l)$, respectively. The evaluation was performed with the checkpoint that achieved the best NMSE for the W-channel on the validation set.

4.2. Results

Figure 3 compares the sound pressure predicted by the DANF with the wave-equation-based prior on the W-channel (Naive) and the proposed PI-DANF (Prop). These NFs were optimized with 250 measurements. According to the leftmost panels, both NFs successfully captured the direct sound and early reflection. Meanwhile, the naive method resulted in lower power and even missed some reflection after 25 ms. The proposed method, on the other hand, achieved better reconstruction.

Figure 4 and Figure 5 depict the NMSE and Pearson's correlation coefficient for the W- and (X, Y, Z)-channels. Here, we averaged the channel-wise scores for the (X, Y, Z)-channels. The naive method outperformed the vanilla DANF that was trained solely with the data fidelity term. Interestingly, the performance on the (X, Y, Z)-channels was also improved even though the wave-equation-based prior

[3]https://github.com/xefonon/RIRPINN/tree/main

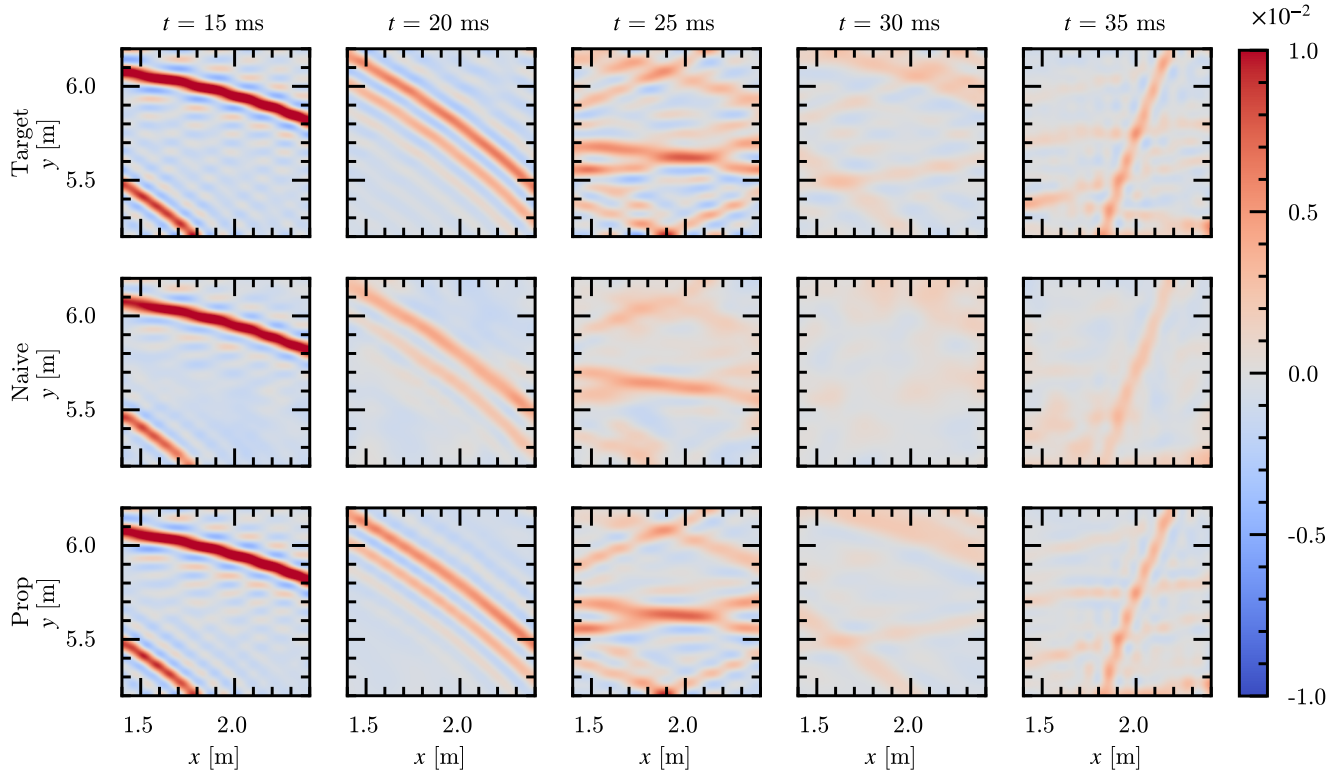

Fig. 3: Visualization of the original and predicted sound pressure, i.e., W-channel, within the target region at 2.30 m height. The sound source was located at (0.85 m, 1.95 m, 1.31 m) inside a room of dimensions of (6.65 m, 7.15 m, 3.71 m).

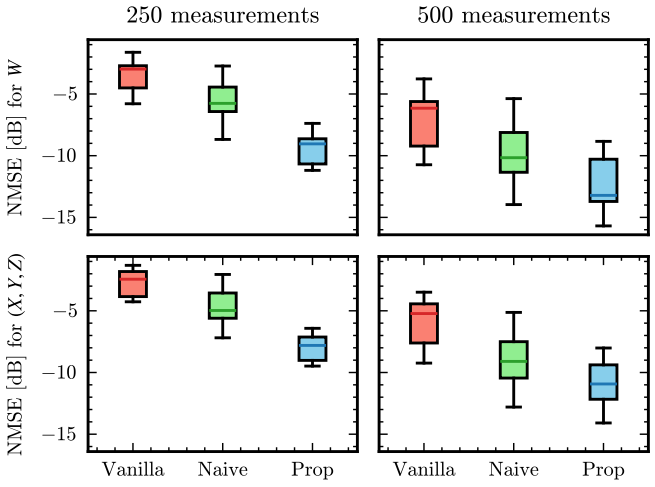

Fig. 4: Boxplots of NMSE over 10 rooms (lower is better).

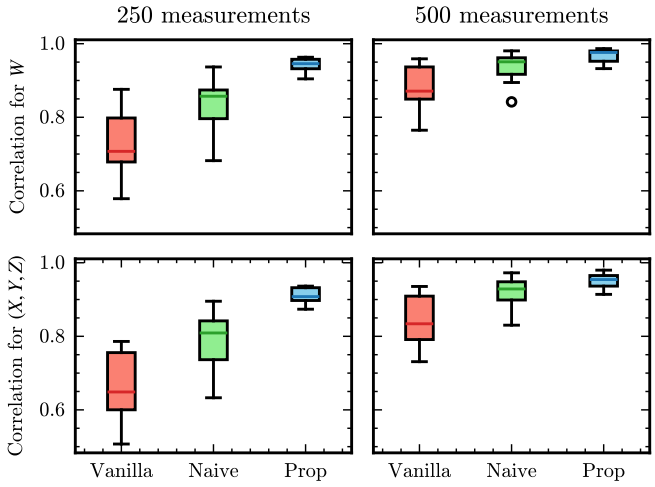

Fig. 5: Boxplots of Pearson's correlation coefficients (higher is better).

regularizes only the W-channel. This could be because the network is shared across different channels except for the final projection layer. The PI-DANF consistently performed the best owing to the proposed priors tailored for FOA RIRs.

5. CONCLUSION

We proposed PI-DANF, a PINN framework for predicting FOA RIRs. Based on the linearized momentum equation and the continuity equation, we derived two physics-informed priors to enforce the network outputs to follow the physical principles of FOA RIRs. These priors connect the W-channel and (X, Y, Z)-channels through their partial derivatives while the wave-equation-based prior has been applied to only the sound pressure, i.e., W-channel. We have confirmed the effectiveness of the PI-DANF through numerical experiments compared with the vanilla DANF and its naive extension. Future work will focus on scaling PI-DANF to model complete FOA RIRs, instead of only the early part, in realistic rooms and aim to generate physically feasible and perceptually natural RIRs.

REFERENCES

[1] H. Kuttruff, *Room Acoustics*. CRC Press, 2016.

[2] T. Lentz, D. Schröder, M. Vorländer, and I. Assenmacher, "Virtual reality system with integrated sound field simulation and reproduction," *EURASIP J. Adv. Signal Process.*, vol. 2007, pp. 1–19, 2007.

[3] S. Serafin, M. Geronazzo, C. Erkut, N. C. Nilsson, and R. Nordahl, "Sonic interactions in virtual reality: State of the art, current challenges, and future directions," *IEEE Comput. Graph. Appl.*, vol. 38, no. 2, pp. 31–43, 2018.

[4] L. Bahrman, M. Fontaine, J. Le Roux, and G. Richard, "Speech dereverberation constrained on room impulse response characteristics," in *Proc. Interspeech*, 2024, pp. 622–626.

[5] N. Ueno, S. Koyama, and H. Saruwatari, "Kernel ridge regression with constraint of helmholtz equation for sound field interpolation," in *Proc. IWAENC*, 2018.

[6] S. Lee, "The use of equivalent source method in computational acoustics," *J. Comput. Acoust.*, vol. 25, no. 1, p. 1630001, 2017.

[7] I. Tsunokuni, K. Kurokawa, H. Matsuhashi, Y. Ikeda, and N. Osaka, "Spatial extrapolation of early room impulse responses in local area using sparse equivalent sources and image source method," *Appl. Acoust.*, vol. 179, p. 108027, 2021.

[8] M. Jälmby, F. Elvander, and T. Van Waterschoot, "Low-rank room impulse response estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 957–969, 2023.

[9] D. Sundström, F. Elvander, and A. Jakobsson, "Optimal transport based impulse response interpolation in the presence of calibration errors," *IEEE Trans. Signal Process.*, vol. 72, pp. 1548–1559, 2024.

[10] R. Mignot, L. Daudet, and F. Ollivier, "Room reverberation reconstruction: Interpolation of the early part using compressed sensing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 11, pp. 2301–2312, 2013.

[11] N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, and T. Van Waterschoot, "Room impulse response interpolation using a sparse spatio-temporal representation of the sound field," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1929–1941, 2017.

[12] S. A. Verburg and E. Fernandez-Grande, "Reconstruction of the sound field in a room using compressive sensing," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3770–3779, 2018.

[13] N. J. Bryan, "Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation," in *Proc. ICASSP*, 2020.

[14] F. Lluis, P. Martinez-Nuevo, M. Bo Møller, and S. Ewan Shepstone, "Sound field reconstruction in rooms: Inpainting meets super-resolution," *J. Acoust. Soc. Am.*, vol. 148, no. 2, pp. 649–659, 2020.

[15] A. Ratnarajah, Z. Tang, and D. Manocha, "IR-GAN: Room impulse response generator for far-field speech recognition," in *Proc. Interspeech*, 2021, pp. 286–290.

[16] M. Pezzoli, D. Perini, A. Bernardini, F. Borra, F. Antonacci, and A. Sarti, "Deep prior approach for room impulse response reconstruction," *Sens.*, vol. 22, no. 7, p. 2710, 2022.

[17] A. Ratnarajah, Z. Tang, R. Aralikatti, and D. Manocha, "MESH2IR: Neural acoustic impulse response generator for complex 3d scenes," in *Proc. ACM Int. Conf. Multimed.*, 2022, pp. 924—933.

[18] Z. Chen, I. D. Gebru, C. Richardt, A. Kumar, W. Laney, A. Owens, and A. Richard, "Real acoustic fields: An audio-visual room acoustics dataset and benchmark," in *Proc. CVPR*, 2024.

[19] A. Luo, Y. Du, M. Tarr, J. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," in *Proc. NeurIPS*, vol. 35, 2022, pp. 3165–3177.

[20] A. Richard, P. Dodds, and V. K. Ithapu, "Deep impulse responses: Estimating and parameterizing filters with deep networks," in *Proc. ICASSP*, 2022, pp. 3209–3213.

[21] K. Su, M. Chen, and E. Shlizerman, "INRAS: Implicit neural representation for audio scenes," in *Proc. NeurIPS*, vol. 35, 2022, pp. 8144–8158.

[22] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu, "AV-NeRF: Learning neural fields for real-world audio-visual scene synthesis," in *Proc. NeurIPS*, 2023.

[23] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *J. Comput. phys.*, vol. 378, pp. 686–707, 2019.

[24] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nat. Rev. Phys.*, vol. 3, no. 6, pp. 422–440, 2021.

[25] M. Olivieri, M. Pezzoli, F. Antonacci, and A. Sarti, "A physics-informed neural network approach for nearfield acoustic holography," *Sens.*, vol. 21, no. 23, p. 7834, 2021.

[26] X. Chen, F. Ma, A. Bastine, P. Samarasinghe, and H. Sun, "Sound field estimation around a rigid sphere with physics-informed neural network," in *Proc. APSIPA*, 2023, pp. 1984–1989.

[27] M. Pezzoli, F. Antonacci, and A. Sarti, "Implicit neural representation with physics-informed neural networks for the reconstruction of the early part of room impulse responses," *arXiv preprint arXiv:2306.11509*, 2023.

[28] X. Karakonstantis, D. Caviedes-Nozal, A. Richard, and E. Fernandez-Grande, "Room impulse response reconstruction with physics-informed deep learning," *J. Acoust. Soc. Am.*, vol. 155, no. 2, pp. 1048–1059, 2024.

[29] G. Sato and Y. Ikeda, "Data-driven physics-informed neural network for sound field estimation in rooms of arbitrary size," in *Proc. APSIPA*, 2024.

[30] S. Koyama, J. G. C. Ribeiro, T. Nakamura, N. Ueno, and M. Pezzoli, "Physics-informed machine learning for sound field estimation: Fundamentals, state of the art, and challenges," *IEEE Signal Process. Mag.*, vol. 41, no. 6, pp. 60–71, 2024.

[31] V. Merimaa, Juha; Pulkki, "Spatial impulse response rendering I: Analysis and synthesis," *J. Audio Eng. Soc.*, vol. 53, pp. 1115–1127, 2005.

[32] P. Coleman, A. Franck, P. J. Jackson, R. J. Hughes, L. Remaggi, and F. Melchior, "Object-based reverberation for spatial audio," *J. Audio Eng. Soc.*, no. 9731, 2017.

[33] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3616–3627, 2018.

[34] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, pp. 2–10, 1973.

[35] J. Daniel, J.-B. Rault, and J.-D. Polack, "Ambisonics encoding of other audio formats for multiple listening conditions," *J. Audio Eng. Soc.*, no. 4795, 1998.

[36] D. Arteaga, "Introduction to ambisonics," Lecture Notes on "Audio 3D", Universitat Pompeu Fabra, 2023.

[37] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography*. Academic Press, 1999.

[38] H. Ren, C. Ritz, J. Zhao, X. Zheng, and D. Jang, "Towards a B-format ambisonic room impulse response generator using conditional generative adversarial network," in *Proc. APSIPA*, 2024.

[39] C. Ick, G. Wichern, Y. Masuyama, F. G. Germain, and J. Le Roux, "Direction-aware neural acoustic fields for few-shot interpolation of ambisonic impulse responses," in *Interspeech*, 2025.

[40] D. T. Blackstock, *Fundamentals of physical acoustics*. John Wiley & Sons, 2000.

[41] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H audio—the new standard for universal spatial/3d audio coding," *J. Audio Eng. Soc.*, vol. 62, pp. 821–830, 2014.

[42] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format," in *Proc. AES Int. Conf.*, 2003.

[43] M. Heydari, M. Souden, B. Conejo, and J. Atkins, "ImmerseDiffusion: A generative spatial audio latent diffusion model," in *Proc. ICASSP*, 2025.

[44] S. S. Kushwaha, J. Ma, M. R. P. Thomas, Y. Tian, and A. Bruni, "Diff-SAGe: End-to-end spatial audio generation using diffusion models," in *Proc. ICASSP*, 2025.

[45] J. Kim, H. Yun, and G. Kim, "ViSAGe: Video-to-spatial audio generation," in *Proc. ICLR*, 2025.

[46] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Proc. NeurIPS*, 2020, pp. 7462–7473.

[47] Z. Xiang, W. Peng, X. Liu, and W. Yao, "Self-adaptive loss balanced physics-informed neural networks," *Neurocomput.*, vol. 496, pp. 11–34, 2022.

[48] S. Saini and J. Peissig, "HARP: A large-scale higher-order ambisonic room impulse response dataset," *arXiv preprint arXiv:2411.14207*, 2024.

[49] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. ICASSP*, 2018, pp. 351–355.

[50] H. Ren, C. Ritz, J. Zhao, and D. Jang, "Towards an objective quality metric for interpolated directional room impulse responses," in *Proc. ICASSP*, 2024, pp. 8205–8209.