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

Velocity Potential Neural Field for Efficient Ambisonics Impulse Response Modeling

Masuyama, Yoshiki; Germain, François G; Wichern, Gordon; Hori, Chiori; Le Roux, Jonathan

TR2026-033 March 19, 2026

Abstract

First-order Ambisonics (FOA) is a standard spatial audio format based on spherical harmonic decomposition. Its zeroth- and first- order components capture the sound pressure and particle velocity, respectively. Recently, physics-informed neural networks have been applied to the spatial interpolation of FOA signals, regularizing the network outputs based on soft penalty terms derived from physical principles, e.g., the linearized momentum equation. In this paper, we reformulate the task so that the predicted FOA signal automatically satisfies the linearized momentum equation. Our network approximates a scalar function called velocity potential, rather than the FOA signal itself. Then, the FOA signal can be readily recovered through the partial derivatives of the velocity potential with respect to the network inputs (i.e., time and microphone position) according to physics of sound propagation. By deriving the four channels of FOA from the single-channel velocity potential, the reconstructed signal follows the physical principle at any time and position by construction. Experimental results on room impulse response reconstruction confirm the effectiveness of the proposed framework.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2026

© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

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

VELOCITY POTENTIAL NEURAL FIELD FOR EFFICIENT AMBISONICS IMPULSE RESPONSE MODELING

Yoshiki Masuyama, François G. Germain, Gordon Wichern, Chiori Hori, Jonathan Le Roux

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

ABSTRACT

First-order Ambisonics (FOA) is a standard spatial audio format based on spherical harmonic decomposition. Its zeroth- and first-order components capture the sound pressure and particle velocity, respectively. Recently, physics-informed neural networks have been applied to the spatial interpolation of FOA signals, regularizing the network outputs based on soft penalty terms derived from physical principles, e.g., the linearized momentum equation. In this paper, we reformulate the task so that the predicted FOA signal automatically satisfies the linearized momentum equation. Our network approximates a scalar function called velocity potential, rather than the FOA signal itself. Then, the FOA signal can be readily recovered through the partial derivatives of the velocity potential with respect to the network inputs (i.e., time and microphone position) according to physics of sound propagation. By deriving the four channels of FOA from the single-channel velocity potential, the reconstructed signal follows the physical principle at any time and position by construction. Experimental results on room impulse response reconstruction confirm the effectiveness of the proposed framework.

***Index Terms*—** Ambisonics, room impulse response interpolation, physics-informed neural network, velocity potential

1. INTRODUCTION

Sound field reconstruction aims to predict a sound field (typically a sound pressure field) at any point in a given spatial region based on a finite set of measurements [1, 2]. This task has many applications in spatial audio such as immersive audio generation [3] and sound field control [4]. Although dense measurements could yield high-quality reconstruction, acquiring them requires an enormous amount of time and effort. Consequently, various advanced methods have been developed to achieve high-quality reconstruction even from a limited number of measurements [5–13]. In this paper, we focus on the spatial interpolation of room impulse responses (RIRs), as they fully characterize the relation between any source and receiver positions and unlock many spatial audio applications [14].

Traditional RIR interpolation techniques typically leverage physics of sound propagation such as geometrical acoustics [15, 16] and wave-based acoustics [10–12]. More recently, neural-network-based approaches have shown promising performance thanks to their powerful modeling capability [17–20]. In particular, neural fields have been actively used in RIR interpolation, where the network characterizes the sound field as a function of time, source position, and/or microphone position [21–23]. Once the neural fields are trained, we can predict RIRs at arbitrary positions in a grid-less manner. These two directions have been combined with the concept of physics-informed neural networks (PINNs) [24–28]. A PINN for sound field reconstruction is a neural field whose training is regularized by a physics prior. Specifically, the derivatives of the predicted sound pressure with respect to the inputs (e.g., time and microphone position) are penalized towards following the governing partial differential equations, e.g., the wave equation [26].

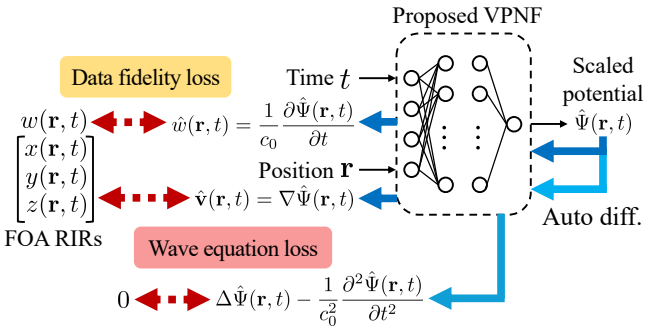

Fig. 1. Overview of the proposed VPNF, where bold solid arrows indicate automatic differentiation. We predict the zeroth- and first-order components of Ambisonics by computing the partial derivatives of the potential with respect to time and position, respectively.

PINNs have been specialized to model directional RIRs [29], in particular the (four-channel) first-order Ambisonics (FOA) format [30]. FOA [31–33] is a widely-used format for spatial audio built on spherical harmonic decomposition [34]. Its zeroth-order component corresponds to the sound pressure, and the three remaining first-order components match the particle velocity up to a constant multiplicative factor. Leveraging these relationships, physics-informed direction-aware neural acoustic field (PI-DANF) [30] proposes to regularize the neural field training based on physical principles, e.g., the linearized momentum equation [35]. PI-DANF achieves better reconstruction than a vanilla neural field, but the predicted FOA RIRs are still allowed to deviate from the principles.

To partially overcome this limitation, we reformulate the task so that the predicted FOA RIRs satisfy the linearized momentum equation by construction. Under typical assumptions, sound pressure and particle velocity can be computed as partial derivatives of a single scalar function, called velocity potential [36, 37]. The derivatives satisfy the linearized momentum equation by design. Hence, we aim to model the velocity potential by using a differentiable function.

In this paper, we propose a neural field for velocity potential reconstruction, named a velocity potential neural field (VPNF). As illustrated in Fig. 1, our neural field outputs a scaled version of the velocity potential. Then, leveraging automatic differentiation, we recover the four-channel FOA RIRs based on the partial derivatives of the predicted velocity potential with respect to time and microphone position. This design ensures that the reconstructed sound field follows the linearized momentum equation at any time and position. In addition, we can incorporate a soft penalty term regularizing the pre-

dicted velocity potential towards following the wave equation. Our experiments confirm the effectiveness of VPNF for FOA RIR interpolation compared with a vanilla neural field and PI-DANF when the number of measurements is limited.

2. PRELIMINARIES

2.1. Problem settings

Let the sound pressure and particle velocity be $p(\mathbf{r},t) \in \mathbb{R}$ and $\mathbf{u}(\mathbf{r},t) \in \mathbb{R}^3$, respectively, where $\mathbf{r} \in \mathbb{R}^3$ denotes the position in Cartesian coordinates, and $t \in \mathbb{R}$ is the time. Throughout this paper, we consider a source-free region $\Omega \subset \mathbb{R}^3$ and focus on air sound propagation assuming air is an inviscid fluid. The sound pressure and particle velocity satisfy the following linearized momentum equation [35]:

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

where ∇ denotes the gradient with respect to $\mathbf{r}$, and ρ_0 is the density of the air. They also follow the continuity equation given by

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

where $\nabla\cdot$ is the divergence, and c_0 is the sound speed in the air. The true sound speed and other physical parameters are assumed to be constant and known.

Assuming that air is inviscid, we can define a velocity potential $\Phi(\mathbf{r},t) \in \mathbb{R}$ in the source-free region Ω such that [36, 37]:

$$\mathbf{u}(\mathbf{r},t) = \nabla \Phi(\mathbf{r},t). \tag{3}$$

By combining this definition with (1), we obtain the following relation between the velocity potential and sound pressure:

$$p(\mathbf{r},t) = -\rho_0 \frac{\partial \Phi(\mathbf{r},t)}{\partial t}. \tag{4}$$

Then, combining (2)–(4), we additionally obtain the wave equation for the velocity potential as follows:

$$\Delta \Phi(\mathbf{r},t) - \frac{1}{c_0^2}\frac{\partial^2 \Phi(\mathbf{r},t)}{\partial t^2} = 0, \tag{5}$$

where Δ denotes the Laplacian.

Ambisonics is an audio format capturing the spatial characteristics of a sound field based on spherical harmonics [31–33]. Under the SN3D normalization [38], its W-channel coincides with the sound pressure measured by an omnidirectional microphone, i.e., $w(\mathbf{r},t) = p(\mathbf{r},t)$. Meanwhile, the (X,Y,Z)-channels correspond to the first-order components of spherical harmonics and are proportional to the particle velocity as follows [29]:

$$\mathbf{v}(\mathbf{r},t) = [x(\mathbf{r},t); y(\mathbf{r},t); z(\mathbf{r},t)], \tag{6}$$

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

where $[\cdot;\cdot]$ denotes vertical concatenation. Hence, the (X,Y,Z)-channels capture the gradient of the velocity potential up to a multiplicative constant.

2.2. PI-DANF for interpolating FOA RIRs

RIR interpolation aims to reconstruct spatially-continuous RIRs from sparse measurements at $\mathbf{r}_d \in \Omega$, where $d = 0, \ldots, D-1$ is the index of the microphone position, with the source location being assumed fixed. While various signal-processing-based methods have been developed [7, 10–12, 15], neural fields have recently gained much attention due to their flexibility and powerful modeling capability [21–23]. A neural field for the sound pressure is formulated as follows [22]:

$$p(\mathbf{r},t) \approx \hat{p}(\mathbf{r},t) = \texttt{NF}_\theta(\mathbf{r},t), \tag{8}$$

where θ are the model parameters. Once trained, it can predict sound pressure at any time and position in a grid-less manner.

Going beyond RIRs measured by omnidirectional microphones, neural fields have been applied to interpolate FOA RIRs [30, 39]. The direction-aware neural field (DANF) predicts the four channels of FOA RIRs instead of only the sound pressure $w(\mathbf{r},t)$ [39]:

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

The neural field is optimized to minimize the reconstruction error at the measured positions:

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

where $l = 0, \ldots, L-1$ is the sample index of the measurements, and $\|\cdot\|_1$ denotes the ℓ_1 norm.

Inspired by PINNs [40, 41], DANF was extended to incorporate penalty terms derived from the physical principles of sound propagation, as PI-DANF [30]. Specifically, the first penalty term is derived from the linearized momentum equation in (1) as follows:

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

where $T \in \mathbb{R}_+$ is for limiting the time range. The second term similarly penalizes the discrepancy with the continuity equation in (2). These penalty terms incentivize the outputs of PI-DANF to follow the principles of sound propagation even at unmeasured points. However, it is not guaranteed that the prediction follows the principles strictly.

3. VELOCITY POTENTIAL NEURAL FIELD (VPNF)

3.1. Formulation of VPNF

Our goal is to more explicitly leverage the physical relationship between the sound pressure and particle velocity to train an improved neural field for FOA RIRs. As shown in Section 2.1, the W-channel corresponds to the sound pressure, and the other channels match the particle velocity. Then, from (3)–(4), we obtain the following relation by considering $\Psi(\mathbf{r},t) = -\rho_0 c_0 \Phi(\mathbf{r},t)$:

$$w(\mathbf{r},t) = \frac{1}{c_0}\frac{\partial\Psi(\mathbf{r},t)}{\partial t}, \tag{12}$$

$$\mathbf{v}(\mathbf{r},t) = \nabla\Psi(\mathbf{r},t). \tag{13}$$

That is, we can predict FOA RIRs $[w(\mathbf{r},t); \mathbf{v}(\mathbf{r},t)]$ at any $\mathbf{r} \in \Omega$ and $t \in [0,T]$ by approximating $\Psi(\mathbf{r},t)$ as a function of $\mathbf{r}$ and t and calculating its derivatives.

We thus propose VPNF that implicitly represents $\Psi(\mathbf{r},t)$ using a neural field, i.e.,

$$\Psi(\mathbf{r},t) \approx \hat{\Psi}(\mathbf{r},t) = \texttt{VPNF}_\theta(\mathbf{r},t). \tag{14}$$

Since it is impractical to measure the ground-truth velocity potential, we train VPNF using the following data-fidelity term, matching its derivatives to the FOA RIRs at the measured positions:

$$\mathcal{L}_{\text{data}} = \frac{1}{DL}\sum_{d=0}^{D-1}\sum_{l=0}^{L-1}\Big(\Big|\frac{1}{c_0}\frac{\partial\hat{\Psi}(\mathbf{r},t)}{\partial t} - w(\mathbf{r}_d,t_l)\Big| + \Big\|\nabla\hat{\Psi}(\mathbf{r}_d,t_l) - \mathbf{v}(\mathbf{r}_d,t_l)\Big\|_1\Big), \quad (15)$$

where the partial derivatives are computed using automatic differentiation in deep learning frameworks. In addition, we can incorporate another penalty term derived from the wave equation in (5):

$$\mathcal{L}_{\text{wave}} = \mathbb{E}_{\mathbf{r}\in\Omega}\mathbb{E}_{t\in[0,T]}\Big|\Delta\hat{\Psi}(\mathbf{r},t) - \frac{1}{c_0^2}\frac{\partial^2\hat{\Psi}(\mathbf{r},t)}{\partial t^2}\Big|. \quad (16)$$

We note that this penalty term can be computed at arbitrarily sampled positions $\mathbf{r}\in\Omega$ and $t\in[0,T]$ in a grid-less manner.

The FOA RIRs predicted by VPNF satisfy the linearized momentum equation in (1) at any time and position thanks to their derivation in (12)–(13). Meanwhile, PI-DANF separately predicts the W- and (X,Y,Z)-channels and exploits that equation only as a penalty term [30]. It is thus not guaranteed that the FOA RIRs predicted using PI-DANF satisfy the linearized momentum equation exactly. We would argue that VPNF more strictly integrates the neural field with the physical principles of sound propagation. Conversely, the penalty term in (16), derived from the continuity equation in (2), means that VPNF, like PI-DANF, still does not guarantee that the predicted FOA RIRs exactly follow the continuity equation in contrast to the linearized momentum equation. It would require a more complex design to ensure the prediction simultaneously follows both equations. We leave such a development to future work.

3.2. Network architecture and training setup

To realize VPNF, we mainly follow the network architecture and training strategy of PI-DANF [30]. The network comprises a modified multi-layer perceptron (MLP) with SIREN activations [42], where the periodic activation provides well-behaved derivatives [26]. Then, VPNF is realized as follows:

$$\texttt{VPNF}_\theta(\mathbf{r},t) = \texttt{ModifiedMLP}_\theta^{(1)}(\mathbf{r},c_0t), \quad (17)$$

where the superscript denotes the output dimension of the modified MLP, and the final layer outputs a scalar without any activation. Here, we convert the time t to distance c_0t to align the scale of the network inputs. Furthermore, we experimentally find a simple modification could improve the performance as follows:

$$\texttt{VPNF}_\theta^{+}(\mathbf{r},t) = [c_0t;\mathbf{r}]^{\mathsf{T}}\,\texttt{ModifiedMLP}_\theta^{(4)}(\mathbf{r},c_0t), \quad (18)$$

where the output dimension of the modified MLP is four. We expect this parametrization makes the gradient of $\Psi(\mathbf{r},t)$ close to $\texttt{ModifiedMLP}_\theta^{(4)}(\mathbf{r},c_0t)$ except for the multiplicative factor c_0, especially around $(\mathbf{r},t)=(\mathbf{0},0)$. It may be beneficial as the modified MLP has worked well for modeling single-channel [26] and FOA RIRs [30].

The network is optimized to minimize the data fidelity term in (15) or its sum with the penalty term in (16). When incorporating the penalty term, we balance the two terms adaptively [26, 43]:

$$\mathcal{L}_{\text{all}} = \frac{1}{2\epsilon_{\text{data}}^2}\mathcal{L}_{\text{data}} + \frac{1}{2\epsilon_{\text{wave}}^2}\mathcal{L}_{\text{wave}} + \log(\epsilon_{\text{data}}\epsilon_{\text{wave}}), \quad (19)$$

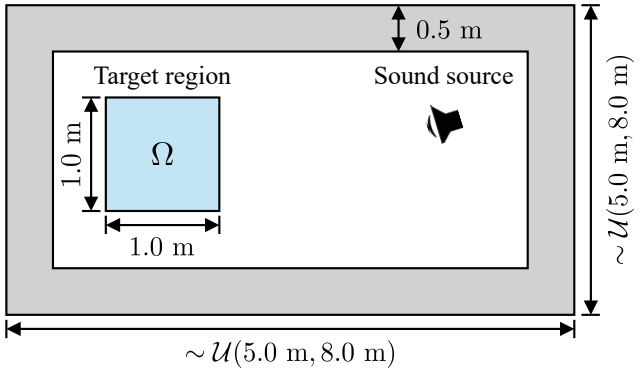

Fig. 2. Overhead view of the room geometry for simulation.

where ϵ_{data} and ϵ_{wave} are the weights for each term, and these parameters are also updated by backpropagation. The position $\mathbf{r}\in\Omega$ and time $t\in[0,T]$ for (16) are randomly sampled at each iteration.

4. EXPERIMENTS

4.1. Dataset and experimental setup

The proposed VPNF is evaluated on FOA RIRs simulated with `HARP`[1] [44], an extension of `Pyroomacoustics` [45] for Ambisonics. The direct sound and early reflections up to 100 ms are simulated at 8 kHz in ten shoebox-shaped rooms with randomly sampled room dimensions and surface materials. An example room configuration is illustrated in Fig. 2, where the room length/width and height were randomly sampled from $[5.0, 8.0)$ m and $[2.5, 4.5)$ m, respectively. The (blue) target region Ω is a cube with 1.0 m size. The cube and a sound source are randomly placed within the room excluding a 0.5 m buffer near the walls (see shaded area in Fig. 2). RIRs were simulated on a grid of 5 cm width in Cartesian coordinates, resulting in 9,261 measurements.

We assess the performance under two different measurement setups. The first one is the reconstruction from a set of FOA RIRs randomly sampled from the 9,261 measurements. We randomly choose a fixed set of $\{30, 50, 70, 100, 150, 200\}$ measurements to train the neural fields and use 50 distinct measurements for validation. The evaluation is performed on the remaining 9,011 positions using the model with the best normalized mean squared error (NMSE) for the W-channel on the validation set. In the second setup, we train the neural field on a fixed set of $\{100, 200\}$ FOA RIRs measured at the surface of the cube. This setup is more challenging than the first one, as we cannot access measurements within the cube. We again use 50 additional data measured on the surface for validation and evaluate the performance of the checkpoint with the best validation score on the remaining 9,011 positions.

Our network and training configuration follow those used in PI-DANF [30], which is based on a prior PINN for omnidirectional RIRs[2] [26]. VPNF and its variants consist of 3 hidden layers with 512 hidden units. They are optimized with the Adam optimizer and cosine annealing. During training, we randomly sample 250 discrete times t_l from all the training-data positions to compute the data fidelity term in (15). When incorporating the wave-equation-based penalty term in (16), we randomly sample 25,000 pairs of $\mathbf{r}$ and t

[1] `https://github.com/whojavumusic/HARP`
[2] `https://github.com/xefonon/RIRPINN/tree/main`

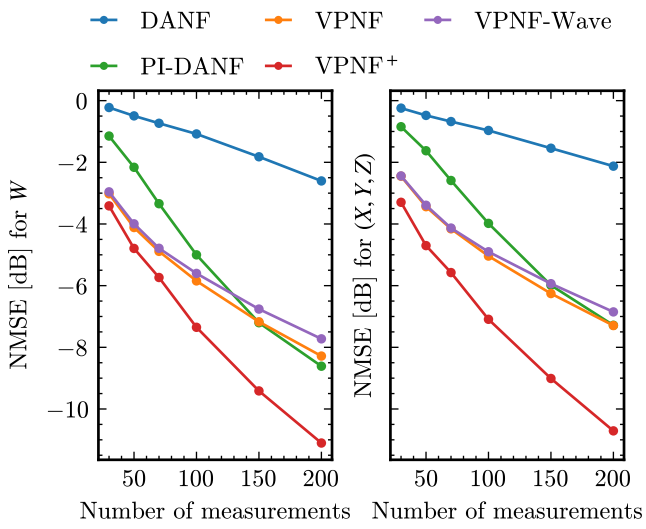

Fig. 3. NMSE [dB] averaged over ten rooms with different numbers of measurements.

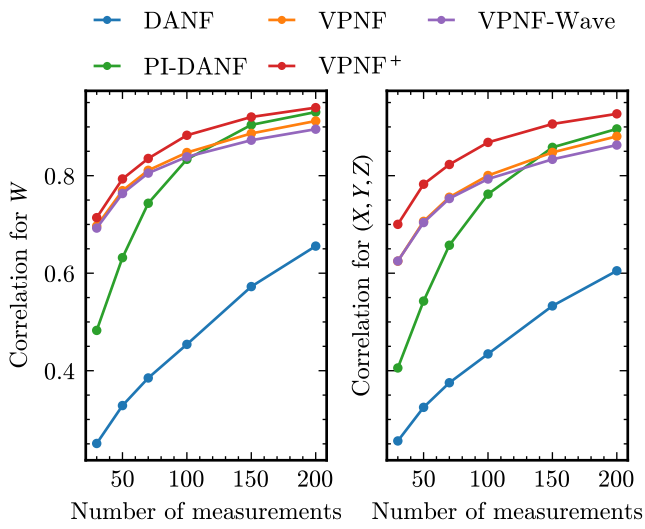

Fig. 4. Average Pearson's correlation coefficients for the W- and (X, Y, Z)-channels.

by Latin hypercube sampling [46]. We set the initial values of ϵ_{data} and ϵ_{wave} to 1.0 and 0.1, respectively. The networks are trained for 100,000 iterations.

The performance is evaluated by NMSE in dB scale for the W-channel and (X, Y, Z)-channels:

$$\text{NMSE}(s, \hat{s}) = 10 \log_{10} \left(\frac{\sum_{\tilde{d}=0}^{\tilde{D}-1} \|\hat{\mathbf{s}}(\mathbf{r}_{\tilde{d}}) - \mathbf{s}(\mathbf{r}_{\tilde{d}})\|_2^2}{\sum_{\tilde{d}=0}^{\tilde{D}-1} \|\mathbf{s}(\mathbf{r}_{\tilde{d}})\|_2^2} \right), \quad (20)$$

where s is any of the four channels of FOA RIRs, $\tilde{d} = 0, \ldots, \tilde{D}-1$ indexes evaluation positions, $\mathbf{s}(\mathbf{r}_{\tilde{d}}) = [s(\mathbf{r}_{\tilde{d}}, t_0), \ldots, s(\mathbf{r}_{\tilde{d}}, t_{L-1})]^\mathsf{T}$, and $\|\cdot\|_2$ denotes the ℓ_2 norm. Another metric is the Pearson's correlation coefficient between the reference and predicted RIRs at each channel, motivated by its correlation with perceptual localization accuracy [47]:

$$\text{PCC}(s, \hat{s}) = \frac{1}{\tilde{D}} \sum_{\tilde{d}=0}^{\tilde{D}-1} \frac{[\hat{\mathbf{s}}(\mathbf{r}_{\tilde{d}}) - \underline{\hat{\mathbf{s}}}(\mathbf{r}_{\tilde{d}})]^\mathsf{T} [\mathbf{s}(\mathbf{r}_{\tilde{d}}) - \underline{\mathbf{s}}(\mathbf{r}_{\tilde{d}})]}{\|\hat{\mathbf{s}}(\mathbf{r}_{\tilde{d}}) - \underline{\hat{\mathbf{s}}}(\mathbf{r}_{\tilde{d}})\|_2 \|\mathbf{s}(\mathbf{r}_{\tilde{d}}) - \underline{\mathbf{s}}(\mathbf{r}_{\tilde{d}})\|_2}, \quad (21)$$

where $\underline{\hat{\mathbf{s}}}(\mathbf{r}_{\tilde{d}})$ and $\underline{\mathbf{s}}(\mathbf{r}_{\tilde{d}})$ are the time average of $\hat{s}(\mathbf{r}_{\tilde{d}}, t_l)$ and $s(\mathbf{r}_{\tilde{d}}, t_l)$, respectively. We took the average of these channel-wise metrics over the (X, Y, Z)-channels.

4.2. Results

Figure 3 shows the NMSE for the W- and (X, Y, Z)-channels with different numbers of measurements under the first condition, where the measurements were randomly sampled from the target region. Here, DANF [39] denotes a neural field for FOA RIRs trained only with the data-fidelity term in (10). PI-DANF additionally exploits the physical principles in (1)–(2) as soft penalties [30]. Consistently with prior trends in [30], PI-DANF reliably outperforms DANF as a result of regularizing the network outputs according to the underlying physics. The proposed VPNF achieves lower NMSE than all baselines when the number of measurements is less than or equal to 100. Its performance is however comparable with PI-DANF when we have more measurements. These results demonstrate the effectiveness of enforcing the linearized momentum equation by modeling velocity potential, especially under the more challenging conditions, i.e., with fewer of measurements. Among variants of VPNF, we find that VPNF$^+$ (see Eq. (18)) substantially improves the performance. Interestingly, using Eq. (16) as a soft penalty, i.e., VPNF-Wave, shows little improvement in our tests. We will analyze this point further in future work. Similar tendencies are found in the Pearson's correlation coefficient results, as seen in Fig. 4.

Table 1. Reconstruction results from $\{100, 200\}$ measurements on the surface of the target region.

	NMSE [dB] (↓)				Correlation (↑)			
	W-ch		(X, Y, Z)-ch		W-ch		(X, Y, Z)-ch	
	100	200	100	200	100	200	100	200
DANF [39]	−0.38	−0.37	−0.40	−0.43	0.31	0.38	0.31	0.36
PI-DANF [30]	−2.03	−3.87	−1.52	−2.99	0.62	0.77	0.53	0.69
VPNF	−3.97	−5.04	−3.26	−4.15	0.75	0.81	0.69	0.75
VPNF-Wave	−3.83	−4.74	−3.29	−4.09	0.74	0.79	0.69	0.74
VPNF$^+$	**−4.50**	**−6.31**	**−4.24**	**−5.92**	**0.78**	**0.86**	**0.76**	**0.83**

Table 1 summarizes the reconstruction performance from measurements on the surface of the cube. Under this more challenging condition, VPNF and its variants show a clear edge over all baselines for all metrics.

5. CONCLUSION

We presented VPNF, a PINN that models the velocity potential field and predicts FOA RIRs by taking the output gradient with respect to the network inputs (i.e., time and microphone position). By design, the predicted FOA RIRs are guaranteed to satisfy the linearized momentum equation at any time and position. We also tested the additional physics-informed penalty term derived from the wave equation for the velocity potential. Our experiments demonstrate the effectiveness of VPNF, especially when the number of measurements is limited. Future work will explore few-shot adaptation of a pretrained VPNF to new rooms to minimize the required measurements.

6. REFERENCES

[1] E. Fernandez-Grande, "Sound field reconstruction using a spherical microphone array," *J. Acoust. Soc. Am.*, vol. 139, no. 3, pp. 1168–1178, 2016.

[2] N. Ueno and S. Koyama, "Sound field estimation: Theories and applications," *Foundations and Trends® in Signal Processing*, vol. 19, no. 1, pp. 1–98, 2025.

[3] M. Vorländer, D. Schröder, S. Pelzer, and F. Wefers, "Virtual reality for architectural acoustics," *J. of Build. Perform. Simul.*, vol. 8, no. 1, pp. 15–25, 2015.

[4] S. Koyama, J. Brunnström, H. Ito, N. Ueno, and H. Saruwatari, "Spatial active noise control based on kernel interpolation of sound field," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3052–3063, 2021.

[5] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. ICASSP*, vol. 2, 2002, pp. II–1781–II–1784.

[6] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *Proc. ICASSP*, vol. 2, 2002, pp. II–1949–II–1952.

[7] 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.

[8] A. Laborie, R. Bruno, and S. Montoya, "A new comprehensive approach of surround sound recording," in *Proc. AES Conv.*, 2003.

[9] P. Samarasinghe, T. Abhayapala, and M. Poletti, "Wavefield analysis over large areas using distributed higher order microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 647–658, 2014.

[10] 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.

[11] N. Ueno, S. Koyama, and H. Saruwatari, "Sound field recording using distributed microphones based on harmonic analysis of infinite order," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 135–139, 2018.

[12] ——, "Kernel ridge regression with constraint of helmholtz equation for sound field interpolation," in *Proc. IWAENC*, 2018.

[13] 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.

[14] S. Koyama, E. De Sena, P. Samarasinghe, M. R. P. Thomas, and F. Antonacci, "Past, present, and future of spatial audio and room acoustics," in *Proc. ICASSP*, 2025.

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

[16] 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.

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

[18] 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.

[19] 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.

[20] S. Della Torre, M. Pezzoli, F. Antonacci, and S. Gannot, "Diffusion-RIR: Room impulse response interpolation using diffusion models," in *Proc. Forum Acusticum*, 2025.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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," in *Proc. Forum Acusticum*, 2023.

[26] 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.

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

[28] 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.

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

[30] Y. Masuyama, F. G. Germain, G. Wichern, C. Ick, and J. Le Roux, "Physics-informed direction-aware neural acoustic fields," *arXiv preprint arXiv:2507.06826*, 2025.

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

[32] J. Daniel, J.-B. Rault, and J.-D. Polack, "Ambisonics encoding of other audio formats for multiple listening conditions," in *Proc. AES Conv.*, 1998.

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

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

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

[36] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of acoustics*. John Wiley & Sons, 2000.

[37] A. D. Pierce, *Acoustics: an introduction to its physical principles and applications*. Springer, 2019.

[38] 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.

[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] 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.

[41] 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.

[42] 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.

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

[44] S. Saini and J. Peissig, "Harp: A large-scale higher-order ambisonic room impulse response dataset," in *Proc. ICASSP Workshop*, 2025.

[45] 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.

[46] M. D. McKay, R. J. Beckman, and W. J. Conover, "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 42, no. 1, pp. 55–61, 2000.

[47] 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.