Noise-Robust Estimation of Early-part Room Impulse Responses based on Physics-Informed Neural Network with Dynamic Pulling Method

Ken Kurata*, Gen Sato*, Izumi Tsunokuni* and Yusuke Ikeda*
* Tokyo Denki University, JAPAN
E-mail: {24fmi09, 24udc03, 21udc02}@ms.dendai.ac.jp, yusuke.ikeda@mail.dendai.ac.jp

***Abstract*—Recently, physics-informed neural networks (PINN) has been applied to the problem of estimating the early room impulse responses (RIR) from a small number of microphones. PINN uses two types of loss functions corresponding to physical laws and data errors. Thus, PINN has the challenge that when gradients associated with two different loss functions conflict with each other, they are not properly learned. In this study, we propose a method for estimating early-part RIRs by introducing the dynamic pulling method (DPM) into the SIREN architecture with residual connections, using noisy microphone signals. From the simulation experiments on a two-dimensional sound field, the proposed method was able to estimate RIRs with higher accuracy in noisy environments than conventional methods.**

I. Introduction

The room impulse response (RIR) represents the sound propagation in a room from a loudspeaker to a microphone under the assumption of linear time invariance to the sound field. In particular, the early part of the RIR is important because it represents the early reflected sound components, which have more energy than the late reflections and have a significant effect on the sound image and timbre impression of sound source [1].

In applications such as sound field control and sound field visualization, the variation in the early part of the RIR between different microphone positions is significant. Consequently, it is necessary to conduct measurements at multiple points rather than at a single point. However, when RIR measurements are required over large regions or at high spatial densities, obtaining RIR measurements for a large number of points can be challenging.

In recent years, many methods have been proposed to estimate the RIRs at more points using signals obtained from a limited number of microphones. In particular, nonparametric methods for numerically estimating the sound field have been studied. For example, compressed sensing (CS) methods [2], [3] have been proposed using plane waves [4], spherical harmonic functions [5], [6], and the modal expansion [7], which are solutions to the wave equation. The equivalent source method [8], [9] represents the sound field by the superposition of point sources selected and weighted under the assumption of sparsity in time signal, space and other factors.

On the other hand, RIR estimation methods using deep learning have also attracted much attention in recent years. Convolutional neural networks (CNN) [10] and generative adversarial Network (GAN) [11] can model complicated relationships between inputs and outputs and have been applied to RIR estimation. Lluís *et al*. proposed a U-net architecture trained on the size of the sound field obtained using the greens function in a rectangular enclosure [12]. Similarly, E. Fernandez-Grande *et al*. employed a GAN-based approach to estimate RIRs from limited measurements and achieved improved estimation accuracy compared to conventional plane wave regression methods. [13].

Furthermore, a deep learning model called physics-informed neural networks (PINN) [14], which introduces physical laws into the loss function, has been proposed. It can be trained so that the output is adapted to the partial differential equations governing the system of interest. In particular, SIREN [15], a network of multi layer perceptrons with sinusoidal activation functions, has proven to be an effective architecture for learning neural implicit representations of various signals and solving wave equations. Additionally, PINN and SIREN were applied to the inverse problem, and a method for estimating RIRs using them was proposed [16], [17]. In [16], a comparison of the PINN and CS methods [18] showed that the PINN method had improved estimation accuracy.

PINN generally uses two types of loss functions: losses for data errors and losses for physical laws. This leads to the known problem of conflicting gradient vectors from the two loss functions, resulting in insufficient learning. To solve this problem, several methods have been proposed. Dynamic pulling method (DPM) [19] adjusts the composition of the two gradients when the gradients of the two loss functions conflict. In [20], the self-adaptive loss balanced method have been proposed to determine the composition of two gradients at each learning step. In [17], the self-adaptive loss balanced method was adapted to the RIR estimation problem with SIREN. In particular, when solving inverse problems, such as RIR estimation, overfitting to the data can degrade estimation accuracy if the data contain significant noise. In addition, if the learning process fully utilizes physical laws, it can also be expected to be effective in removing noise from the data in accordance with these laws. In [21], an RIR estimation method have been proposed that is robust to noise in the microphone signals by dynamically switching the loss functions to tolerate

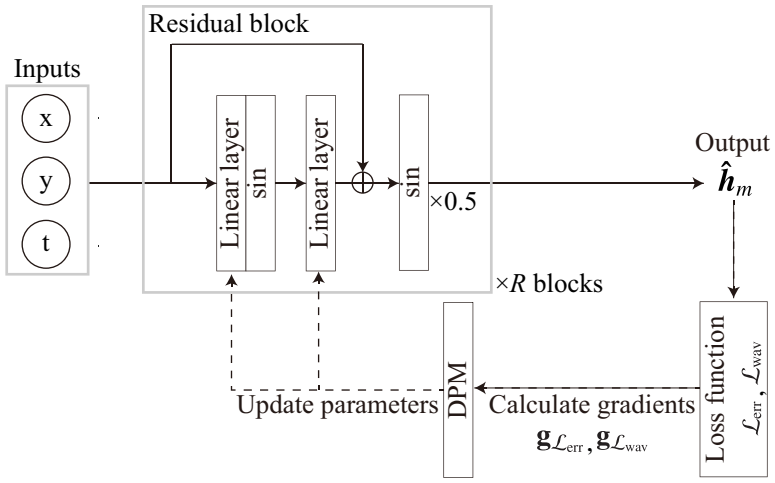

Fig. 1. Network architecture. Inputs are coordinates (x, y) and time t. The network is composed of R layers of residual blocks. The loss function is calculated by the output RIRs $\hat{\boldsymbol{h}}_m (m \in \mathcal{M})$ and measured RIRs $\boldsymbol{h}_m (m \in \mathcal{M})$, and gradients $\mathbf{g}_{\mathcal{L}_{\text{err}}}$ and $\mathbf{g}_{\mathcal{L}_{\text{wav}}}$ are computed to use DPM. For updating the network parameters, the gradients are modified by DPM. $\mathcal{M}$ is the location of all estimated points.

data errors. However, RIR estimation in the presence of large noise in the microphone signals requires further study to improve estimation accuracy.

In this study, to improve the RIR estimation accuracy in a noisy environment, we propose the estimation method of early RIRs based on PINN by introducing DPM and residual connections [22]. The proposed method uses DPM to prevent overfitting to the data and take advantage of losses due to the wave equation to improve estimation accuracy. In addition, this study uses residual connections to improve estimation accuracy by utilizing deeper network layers not considered in conventional RIR estimation methods.

The remainder of this paper is organized as follows. Section II describes the network architecture and DPM in the proposed method. Section III presents the simulation experiments in a two-dimensional sound field to compare the estimation accuracies between the conventional methods and the proposed methods. Finally, we conclude the paper in Sect. IV.

II. Proposed methods

A. Physics-informed neural network using residual connection

A deep learning method using physical laws in loss functions is known as a physics-informed neural network (PINN). In the case of solving acoustical problem, the wave equation with second derivatives is often used as the physical laws. However, general activation functions such as ReLU not allow second-order derivatives in the automatic differentiation. In this study, a SIREN network [15] based on a multi-layer perceptron (MLP) that allows for higher-order differentiation by using a sine function as activation function. In addition, a residual connection is employed to enhance the stability of learning in deeper networks. [22].

The proposed network is illustrated in the Fig. 1. The network is represented as a composite function of each network layer ϕ as follows

$$f(\mathbf{w}, \mathbf{x}) = (H_R \circ H_{R-1} \circ \cdots H_r \circ \cdots H_1)(\mathbf{x}) \tag{1}$$

$$H_r = \frac{1}{2} \sin(\omega_0(\phi_{2r-1}(\mathbf{x}_{2r-1})\mathbf{w}_{2r} + \mathbf{b}_{2r}) + \mathbf{x}_{2r-1}) \tag{2}$$

$$\phi_{2r-1}(\mathbf{x}_{2r-1}) = \sin(\omega_0(\mathbf{x}_{2r-1}^{\top}\mathbf{w}_{2r-1} + \mathbf{b}_{2r-1})) \tag{3}$$

H_r is the residual block in the r-th layer $(r = 1, \ldots, R)$. The $\mathbf{w}$ and $\mathbf{b}$ are the weights and biases of the network, respectively. The training proceeds with the input $\mathbf{x}$. ω_0 is the SIREN weights.

In this study, the early-part RIRs is estimated using the coordinate (x, y) and time information t as input $\mathbf{x} = (x, y, t)$. Note that SIREN multiplies the residual computed output by 1/2 to ensure that the output remains within the range of $[-1, 1]$ [15].

The loss function can be represented using the wave equation as follows

$$\mathcal{L} = \mathcal{L}_{\text{err}} + \lambda \mathcal{L}_{\text{wav}} \tag{4}$$

$$\mathcal{L}_{\text{err}} = \frac{1}{\tilde{M}} \sum_{m \in \tilde{\mathcal{M}}} \|\hat{\boldsymbol{h}}_m - \boldsymbol{h}_m\|_2^2 \tag{5}$$

$$\mathcal{L}_{\text{wav}} = \frac{1}{M} \sum_{m=1}^{M} \|\frac{1}{c}\frac{\partial^2 \hat{\boldsymbol{h}}_m}{\partial t^2} - \nabla^2 \hat{\boldsymbol{h}}_m\|_2^2 \tag{6}$$

where λ is a balance parameter between the data loss function L_{err} and the loss function of wave equation $\mathcal{L}_{\text{wav}}$, c is the speed of sound, t is time, and ∇ is the gradient. $\hat{\boldsymbol{h}}_m$, $\boldsymbol{h}_m$ denote estimated RIRs and measured RIRs, respectively. $\tilde{\mathcal{M}}$ is the set of microphone indices, and M is the total number of measurement positions including microphone positions and evaluation points.

B. Dynamic pulling method

When using a loss function that combines multiple types of losses, the gradients from each loss may conflict, which could disturb the learning process. Thus, in this study, we use

the following algorithm based on the dynamic pulling method (DPM) [19], which dynamically adjusts the combination of gradients from the loss functions.

$$\mathbf{g}^{(k)} = \begin{cases} \mathbf{g}_{\mathcal{L}}^{(k)}, \text{ if } \mathbf{g}_{\mathcal{L}_{\text{err}}}^{(k)} \cdot \mathbf{g}_{\mathcal{L}_{\text{wav}}}^{(k)} \geq 0 \\ v + \mathbf{g}_{\mathcal{L}}^{(k)}, \text{ otherwise} \end{cases} \tag{7}$$

$$v = \frac{-\mathbf{g}_{\mathcal{L}}^{(k)} \cdot \mathbf{g}_{\mathcal{L}_{\text{wav}}}^{(k)} + \delta^{(k)}}{\|\mathbf{g}_{\mathcal{L}_{\text{wav}}}^{(k)}\|_2^2} \mathbf{g}_{\mathcal{L}_{\text{wav}}}^{(k)} \tag{8}$$

$$\delta^{(k+1)} = \begin{cases} w\delta^{(k)}, & \text{if } \mathcal{L}_{\text{wav}}^{(k)} > \epsilon \\ \frac{\delta^{(k)}}{w}, & \text{if } \mathcal{L}_{\text{wav}}^{(k)} \leq \epsilon. \end{cases} \tag{9}$$

$\mathbf{g}^{(k)}$ is the gradient at the k-th step, and $\mathbf{g}_{\mathcal{L}}^{(k)}$, $\mathbf{g}_{\mathcal{L}_{\text{err}}}^{(k)}$, and $\mathbf{g}_{\mathcal{L}_{\text{wav}}}^{(k)}$ are the gradients of the losses $\mathcal{L}$, $\mathcal{L}_{\text{err}}$, and $\mathcal{L}_{\text{wav}}$, respectively. w, ϵ, and δ are hyperparameters. And v is the gradient change vector, which is updated for each learning. That is, as shown in Fig. 1, after obtaining the gradient from each loss function, this algorithm is used to adjust the gradient and update the network.

III. Simulation Experiment

A. Simulation conditions

Simulation experiments were conducted to estimate the early part of RIRs from a small number of microphone signals in a two-dimensional sound field. The estimation accuracy of four methods was compared: two conventional methods, CS [8] and PINN [16], and two proposed methods, PINN with DPM (DPM) and PINN with DPM and residual connections (DPM–Res).

Fig. 2 shows the arrangement of measurement points (12×12) and estimation points (20×20) . The Gaussian noises were added to the microphone signals with $\text{SNR} = 10$ dB. The sound source was a line source at $(0, 1.5)$ m. The length of microphone signal was 0.02 s. The sampling frequency was 8 kHz. The early part of RIRs were simulated by SFS toolbox [23] including up to second-order reflections.

The deep learning Optimizer was adam, and the learning late was set to 10^{-5}. For DPM-Res, the number of block was R=2. For DPM, the number of hidden layers was four. For simple PINN, the number of hidden layers was three. For initial parameters, the balance parameter was $\lambda = 10^{-6}$. ω_0 = 7 for DPM and DPM-Res and ω_0 = 12 for PINN.

The estimation accuracy was evaluated by the normalized mean square error (NMSE) as follows.

$$\text{NMSE} = 10 \log_{10} \frac{1}{M} \sum_{m=1}^{M} \frac{\|\hat{\boldsymbol{h}}_m - \boldsymbol{h}_m\|_2^2}{\|\hat{\boldsymbol{h}}_m\|_2^2}. \tag{10}$$

B. Results

We compared NMSEs of the estimated sound fields by the two conventional methods and two proposed methods. Fig. 3 shows the estimated sound field at $\text{y} = 0.03$ m.

From Fig. 3(c), the amplitude of estimated sound field by CS was significantly small. The microphone signal was noisy with an SNR of 10 dB, making it difficult to estimate the sound field using CS. From Fig. 3(d)-(f), it can be observed that both the amplitude and phase estimated by deep learning methods (PINN, DPM, DPM-Res) are very close to the ground truth. In addition, Fig. 3(d), (e), and (f) show that the estimation errors were reduced in the order of PINN, DPM, and DPM-Res. This indicates that DPM and DPM-Res are more robust to microphone noise than the conventional PINN.

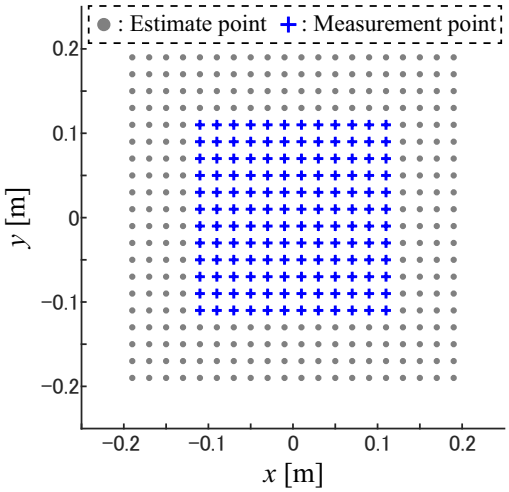

Fig. 2. Arrangement of measurement points (crosses) and estimation points (circles) (each point is 0.02 m apart, placed in the center of a 6 m × 4 m rectangular room. 2-dimensional sound field with a reflectance of 1 on the four walls)

Next, the NMSEs were compared between the conventional and proposed methods. Fig. 4 shows the distributions of NMSE for CS, PINN, DPM, and DPM–Res. Fig. 4(a) and (b) show the results of the conventional methods. PINN improved the estimation accuracy by approximately 4.0 dB NMSE compared to CS. Fig. 4(b) and (c) show that the proposed DPM method achieved an improvement of approximately 3.0 dB in NMSE estimation accuracy compared to PINN. Furthermore, DPM-Res improves the estimation accuracy by approximately 1.2 dB in NMSE compared to DPM and by approximately 4.2 dB in NMSE compared to PINN.

Table I shows comparison of the mean NMSEs at the estimation points only, excluding the microphone positions i.e., points outside the microphones. PINN improved by approximately 2.5 dB in mean NMSE compared to CS. DPM and DPM-Res achieved an improvement in estimation accuracy of approximately 5.6 dB and 6.8 dB in the average NMSE, respectively, compared to CS.

Furthermore, Fig. 5 shows NMSE at each training step using deep learning methods. From the figure, it is evident that PINN exhibits overfitting to noise. In contrast, DPM effectively mitigates overfitting. Moreover, DPM-Res demonstrates a faster convergence rate compared to DPM.

These results show that the proposed methods (DPM and DPM-Res) significantly improve the estimation accuracy over conventional methods for noise removal for RIRs at microphone points and for RIR estimation at points outside

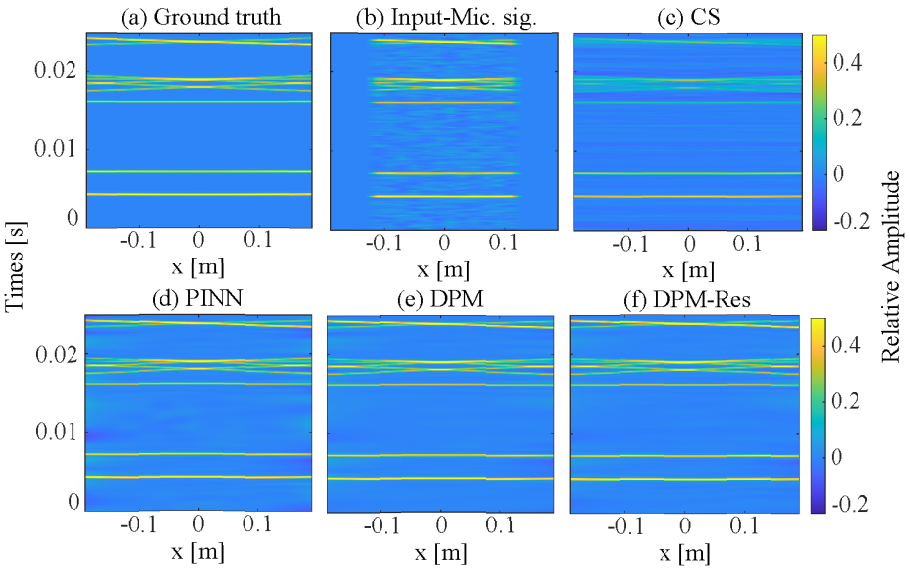

Fig. 3. Estimated RIR signals at y = 0.03 m (a) Desired sound field (b) Microphone signals for estimation (c)–(f) Estimated RIRs by CS, PINN, DPM and DPM-Res, respectively.

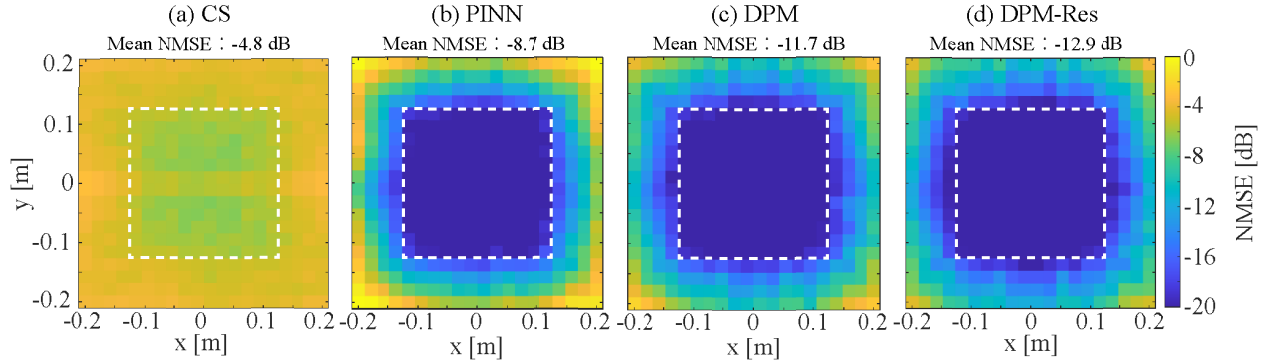

Fig. 4. Time-averaged NMSE maps with CS, PINN, DPM and DPM-Res. The white dot-lined borders indicates the region of microphone positions. (a)–(b) Conventional methods (CS, PINN) (c)–(d) Proposed methods (DPM, DPM-Res)

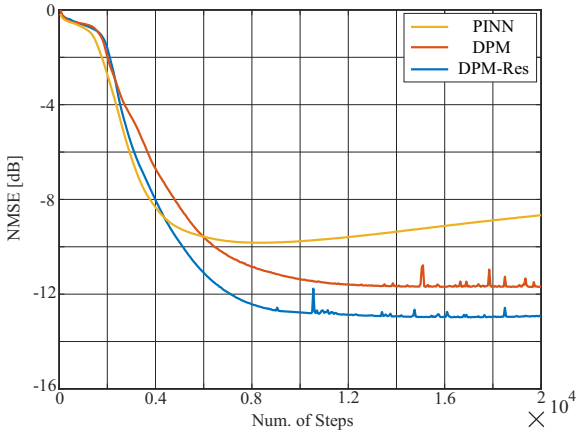

Fig. 5. Comparison of NMSEs in each learning steps with PINN, DPM, and DPM-Res.

the microphones. In particular, DPM-Res achieves the most accurate estimation even in noisy environments.

TABLE I
MEAN NMSE AT THE ESTIMATION POINTS ONLY, EXCLUDING THE MICROPHONE POSITIONS

Method	CS	PINN	DPM	DPM–Res
Mean NMSE [dB]	-4.3	-6.8	-9.9	-11.1

IV. CONCLUSIONS

In this study, we proposed the estimation of early-part RIR based on PINN using DPM and residual connections. From the simulation experiments, the proposed method acheived an improvement in estimation accuracy over conventional PINN and CS. In future work, we will study the configuration of the microphone array and further improve the PINN network to reduce the number of microphones and improve the estimation accuracy of the early-part RIR estimation.

ACKNOWLEDGMENT

This work was partially supported by Research Institute for Science and Technology of Tokyo Denki University Grant Number Q24D-03 / Japan.

References

[1] T. Gotoh, Y. Kimura, A. Kurahashi, and A. Yamada, "A consideration of distance perception in binaural hearing," *The Journal of the Acoustical Society of Japan*, vol. 33, no. 12, pp. 667–671, 1977.

[2] D. L. Donohoh, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[4] S. A. Verburg and E. Fernandez-Grande, "Reconstruction of the sound field in a room using compressive sensing," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3770–3779, 2018.

[5] M. Pezzoli, M. Cobos, F. Antonacci, and A. Sarti, "Sparsity-based sound field separation in the spherical harmonics domain," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1051–1055.

[6] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala, "Sound field separation in a mixed acoustic environment using a sparse array of higher order spherical microphones," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 151–155.

[7] O. Das, P. Calamia, and S. V. Amengual Gari, "Room impulse response interpolation from a sparse set of measurements using a modal architecture," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 960–964.

[8] 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 methods," *Applied Acoustics*, vol. 179, p. 108 027, 2021.

[9] N. Antonello, E. D. Sena, M. Moonen, P. A. Naylor, and T. V. Waterschoot, "Room impulse response interpolation using a sparse spatio-temporal representation of the sound field," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1929–1941, 2017.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84–90, 2012.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[12] F. Lluís, P. Martínez-Nuevo, M. B. Møller, and S. E. 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.

[13] E. Fernandez-Grande, X. Karakonstantis, D. Caviedes-Nozal, and P. Gerstoft, "Generative models for sound field reconstruction," *The Journal of the Acoustical Society of America*, vol. 153, no. 2, pp. 1179–1190, 2023.

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

[15] V. Sitzmann, J. N. P. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Curran Associates Inc., 2020.

[16] M. Pezzoli and a. A. S. F. Antonacci, "Implicit neural representation with physics-informed neural networks for the reconstruction of the early part of room impulse responses," in *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, 2023, pp. 2177–2184.

[17] X. Karakonstantis and E. Fernandez-Grand, "Room impulse response reconstuction using physics-informed neural networks," in *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, vol. 378, 2023, pp. 3181–3188.

[18] E. Zea, "Compressed sensing of impulse responses in rooms of unknown properties and contents," *Journal of Sound and Vibration*, vol. 459, p. 114 871, 2019.

[19] J. Kim, K. Lee, D. Lee, S. Y. Jhin, and N. Park, "Dpm: A novel training method for physics-informed neural networks in extrapolation," in *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, Association for the Advancement of Artificial Intelligence, 2021, pp. 8146–8154.

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

[21] I. Tsunokuni, G. Sato, Y. Ikeda, and Y. Oikawa, "Spatial extrapolation of early room impulse responses with noise-robust physics-informed neural network," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E107-A, no. 9, 2024.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] H. Wierstorf and S. Spors, "Sound field synthesis toolbox," *132nd Convention of the Audio Engineering Society*, 2012.