IEICE TRANSACTIONS

on Fundamentals of Electronics, Communications and Computer Sciences

DOI:10.1587/transfun.2024EAL2015

Publicized:2024/04/08

A PUBLICATION OF THE ENGINEERING SCIENCES SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

LETTER

Spatial extrapolation of early room impulse responses with noise-robust physics-informed neural network

Izumi TSUNOKUNI[†a)], ***Member***, **Gen SATO**[†], **Yusuke IKEDA**[†b)], ***Nonmembers***, ***and*** **Yasuhiro OIKAWA**[††], ***Member***

SUMMARY This paper reports a spatial extrapolation of the sound field with a physics-informed neural network. We investigate the spatial extrapolation of the room impulse responses with physics-informed SIREN architecture. Furthermore, we proposed a noise-robust extrapolation method by introducing a tolerance term to the loss function.

key words: *Deep neural network, Wave equation, SIREN, PI–SIREN*

1. Introduction

Measurement of the room impulse response (RIR), which represents the sound propagation characteristics in a room, is essential in various applications, such as sound field reproduction, visualization, and acoustic design. The RIR depends on the position of the microphone because each RIR is measured using a single loudspeaker and microphone assuming the linear time-invariant system. To obtain the spatial difference in sound propagation, RIRs must be measured at multiple points in a wider region, particularly in the early part of the RIR, which is highly dependent on the measurement position.

Recently, several reconstruction methods for sound fields and RIRs at multiple points based on physical models have been proposed. In early RIRs, reconstruction methods based on physical models and compressed sensing have been extensively studied [1]–[4]. Mignot *et al.* proposed a method to estimate the early RIRs using the sparsity of the early parts of the RIRs in the time domain [5]. In [6], a sound field was interpolated and extrapolated using the superposition of a sparse set of plane waves. In addition, extrapolation methods that are more difficult than interpolation methods have been investigated [6], [7]. In [8], [9], a method for extrapolating the early RIR around locally-located microphones by superposing sparse point sources was proposed.

With the development of machine learning, deep learning-based reconstruction methods for sound fields have been proposed [10], [11]. However, in machine learning, the computed results are not guaranteed to satisfy the physical properties. In 2019 [12], a physics-informed neural network (PINN) was proposed to introduce a governing equation for the loss function. PINN can obtain an output that satisfies physical laws. Recently, the PINN was introduced to the problem of sound field reconstruction [13]–[15]. However, in general machine learning, it is difficult to introduce higher-order derivatives, which are also included in the wave equation, into the cost function because of the commonly used activation function, such as rectified linear unit (ReLU).

The sinusoidal representation network (SIREN) architecture [16] allows higher-order differentiation by introducing periodic functions into the activation function. In [14], [17], the early part of RIRs was reconstructed by introducing the wave equation into the cost function of the SIREN architecture, which is called "Physics-informed SIREN (PI–SIREN)" in [14]. PI–SIREN applies SIREN to the inverse problem of sound propagation, which has a wide range of applications but has not been fully investigated, particularly in the extrapolation problem.

Furthermore, general data-driven methods avoid overfitting by adding noise to the input signals in the training dataset or by data augmentation [18]. However, PI–SIREN cannot use data augmentation because it is not a data-driven method. Using only a single set of microphone signals, PI–SIREN solves the inverse problem of estimating signals at specified positions based on the wave equation.

In this study, we investigated the use of PI–SIREN for spatial extrapolation of early RIRs in a two-dimensional sound field. In addition, we propose a noise-robust method using PI–SIREN by dynamically changing the loss function considering the error tolerance of the microphone signals.

We emphasize two main differences between the proposed method and PI–SIREN [14], [17]: first, in this study, we introduced the error tolerance of microphone signals in the loss function; second, we contributed to the investigation of the spatial extrapolation of early RIRs in the two-dimensional sound field by comparing the estimation accuracies with and without the use of physical laws.

The remainder of this paper is organized as follows. Section 2 introduces the problem statement, SIREN, PI–SIREN, and proposed method. Section 4 presents the simulation results. Finally, we conclude the paper in Section 5.

2. Method

2.1 Problem statement

As shown in Fig. 1, we consider the reconstruction of a two-dimensional sound field Ω from an internal sound field $\Omega_{\text{in}} (\in \Omega)$. In the region Ω, we assume that the sound field

[†]The authors are with the Tokyo Denki Univ., Tokyo, Japan.
[††]The author is with the Waseda Univ., Tokyo, Japan.
a) E-mail: 21udc02@ms.dendai.ac.jp
b) E-mail: yusuke.ikeda@mail.dendai.ac.jp

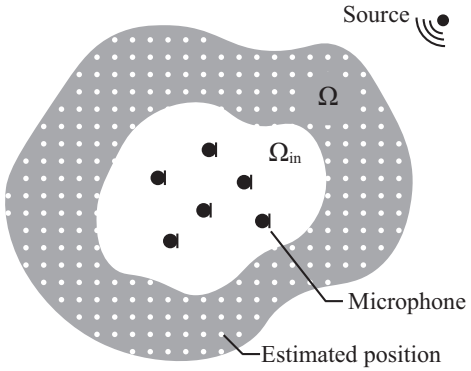

Fig. 1 Extrapolation problem of sound field.

is homogeneous and no sound source exists. Thus, for the sound pressure $p(\mathbf{x})|\mathbf{x} \in \Omega$, the following wave equation holds:

$$\frac{1}{c^2}\frac{\partial^2 p(\mathbf{x})}{\partial t^2} - \nabla^2 p(\mathbf{x}) = 0 \quad (\mathbf{x} \in \Omega), \tag{1}$$

where c is the speed of sound, t is time, and ∇ is the gradient.

The sound field Ω is discretized at M points and indexed as $\mathcal{M}(= 1, 2, \ldots, M)$. The $\tilde{M}$ measurement points are part of the discretized positions ($\tilde{\mathcal{M}} \in \mathcal{M}$) and are positioned in the region Ω_{in}. In this study, we reconstructed the signals $p(\mathbf{x}_m)|m \in \mathcal{M}$ at all discrete points from the signals $p(\mathbf{x}_m)|m \in \tilde{\mathcal{M}}$ at the measurement points.

2.2 SIREN and PI–SIREN

Using multilayer perceptron (MLP), the function $f(\cdot)$ of the neural network for RIR reconstruction is as follows:

$$f(\mathbf{w}, \mathbf{x}) = (\phi_N \circ \phi_{N-1}, \ldots, \phi_1)(\mathbf{x}), \tag{2}$$

where $\mathbf{x}$ is the input to the network and $\mathbf{w}$ is the set of learnable parameters. ϕ_n is the function of n-th layer of MLP. To represent the wave equations shown in Eq.(1), obtaining the second derivatives of the function $f(\cdot)$ is necessary. To determine this using a synthetic derivative, the derivative of the activation function must be determined. However, general neural network models use activation functions that cannot be differentiated into higher orders.

SIREN [16] is an effective network architecture representing higher-order derivative signals. To maintain the information contained in higher-order derivatives, SIREN uses a periodic function as the activation function of the MLP. Consequently, the SIREN derivative becomes its phase-shifted output. SIREN uses the sinusoidal activation function and n-th layer function ϕ_n as follows:

$$\phi_n(\mathbf{x}_n) = \sin(\omega \mathbf{x}_n^{\mathsf{T}} \mathbf{w}_n + \mathbf{b}_n), \tag{3}$$

where $\mathbf{x}_n$, $\mathbf{w}_n$, and $\mathbf{b}_n$ are the input signal, weight, and bias of the n-th layer, respectively. ω is a hyperparameter [16].

In PI–SIREN [14], SIREN was applied to solve the inverse problem of sound field reconstruction by introducing the wave equation Eq.(1) to the loss function L_{ps} as follows:

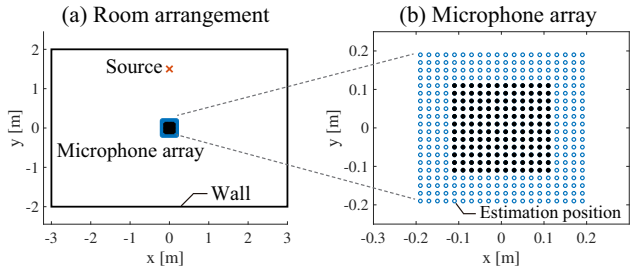

Fig. 2 (a) Arrangement of measurement positions and room geometry. (b) Detailed arrangement of measurement and estimation positions. The solid black circle is the position of microphone, and the blue circle is the estimation position.

$$L_{\text{ps}} = L_{\text{err}} + \lambda L_{\text{wave}}, \tag{4}$$

where,

$$L_{\text{err}} = \frac{1}{\tilde{M}} \sum_{m=1}^{\tilde{M}} \|\hat{\mathbf{h}}_m - \mathbf{h}_m\|_2^2, \tag{5}$$

$$L_{\text{wave}} = \frac{1}{M} \sum_{m=1}^{M} \|\frac{1}{c^2}\frac{\partial^2 \hat{\mathbf{h}}_m}{\partial t^2} - \nabla^2 \hat{\mathbf{h}}_m\|_2^2. \tag{6}$$

$\|\cdot\|_2$ is the ℓ_2-norm, and $\hat{\mathbf{h}}_m$ and $\mathbf{h}_m (\in \mathbb{R}^{T\times 1})$ are the estimated and measured RIRs of the m-th microphone, respectively. T is the number of samples. λ is a weight parameter that controls the balance between L_{err} and L_{wave}. The loss function L_{wave} allows the solution of the neural network to follow physics laws. Note that SIREN and PI–SIREN are not data-driven methods, and the network inputs are fixed positions for the measurement and estimation. In addition, $\mathbf{h}_m$ in Eq.(5) is only a single set of measurement signals.

3. Proposed method

In this study, we considered early RIR measurements in a noisy environment. As shown in Eq. (5), the loss function L_{err} determines a solution that minimizes the differences between the microphone signals and the output. When the microphone signals are contaminated by noise, the estimation accuracy degrades because of the overfitting of the microphone signals with noise. Thus, the proposed method introduces error tolerance into the loss function, as follows:

$$L = \begin{cases} L_{\text{err}} + \lambda L_{\text{wave}}, & \text{if } L_{\text{err}} > \epsilon \\ \epsilon + \lambda L_{\text{wave}}, & \text{if otherwise} \end{cases} \tag{7}$$

The error tolerance ϵ is determined by the noise energies of all the microphones. When the loss L_{err} becomes smaller than the error tolerance ϵ, the loss function becomes only the loss L_{wave} in the wave equation.

In [19], a similar method was used to dynamically change the cost function with respect to the cost based on physical laws to make it easier to solve the extrapolation problem. However, the proposed method changes the cost function with respect to microphone errors to introduce microphone-error tolerance.

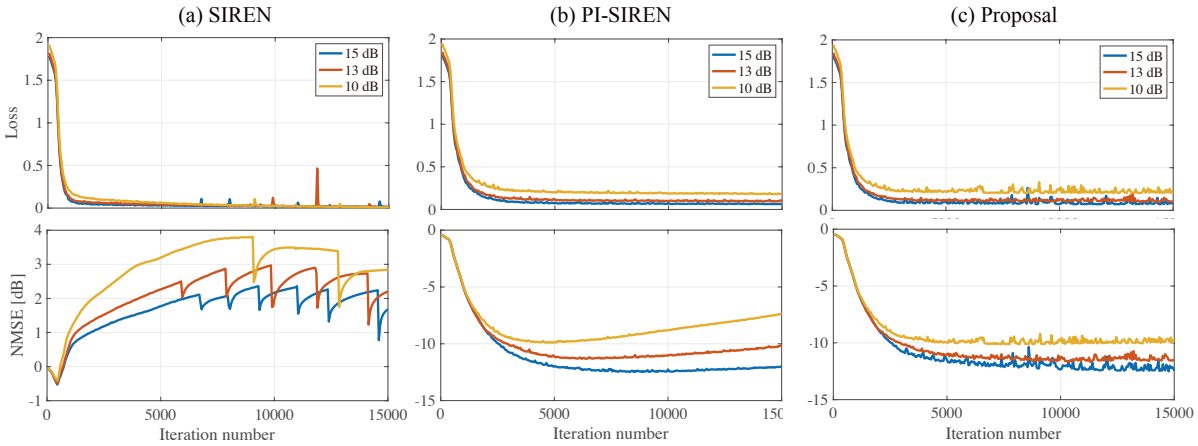

Fig. 3 Comparison of loss and NMSE with SIREN, PI–SIREN, and proposed method. The upper and bottom figures show the loss functions and NMSEs for microphones with SNRs of 10, 13, and 15 dB, respectively.

Table 1 Comparison of NMSE at 5k, 10k, and 15k iterations. $\mathrm{NMSE}_{\mathrm{all}}$ indicates the overall estimation error. $\mathrm{NMSE}_{\mathrm{est}}$ and $\mathrm{NMSE}_{\mathrm{mic}}$ indicate NMSEs for the estimation and microphone positions, respectively.

		SNR 10 dB		SNR 13 dB		SNR 15 dB	
Iteration No.		PI–SIREN	Proposal	PI–SIREN	Proposal	PI–SIREN	Proposal
5000	$\mathrm{NMSE}_{\mathrm{all}}$	-9.8	-10.0	-11.1	-11.0	-12.0	-11.5
	$\mathrm{NMSE}_{\mathrm{est}}$	-8.1	-8.3	-9.4	-9.3	-10.2	-9.9
	$\mathrm{NMSE}_{\mathrm{mic}}$	-19.9	-18.6	-20.6	-19.6	-22.2	-17.8
10000	$\mathrm{NMSE}_{\mathrm{all}}$	-8.8	-9.8	-11.0	-11.2	-12.3	-12.3
	$\mathrm{NMSE}_{\mathrm{est}}$	-7.0	-8.2	-9.2	-9.7	-10.6	-10.5
	$\mathrm{NMSE}_{\mathrm{mic}}$	-21.1	-17.2	-23.6	-17.8	-21.2	-24.2
15000	$\mathrm{NMSE}_{\mathrm{all}}$	-7.4	-10.0	-10.1	-11.5	-12.0	-12.5
	$\mathrm{NMSE}_{\mathrm{est}}$	-5.5	-8.1	-8.4	-9.7	-10.2	-10.6
	$\mathrm{NMSE}_{\mathrm{mic}}$	-19.6	-21.3	-19.6	-22.6	-23.4	-24.2

4. Simulation experiment

4.1 Simulation condition

Simulation experiments were conducted to evaluate the noise robustness of the proposed method compared to SIREN and PI–SIREN. As demonstrated in [14], SIREN uses L_{err} (Eq.5) for the loss function. The RIRs were analytically calculated with the sfs-toolbox [20] and Pyroomacoustics [21] using the image source method [22]. Gaussian white noise was added to the microphone signals with 10, 13, and 15 dB signal-to-noise ratio (SNR). The room size was 6 m × 4 m and the line source was positioned at $(0, 1.5)$. Figure 2 shows the arrangement of the experiment; 256 points around the microphones were extrapolated from 144 microphone signals at 0.02 m intervals, which correspond to the half-wavelength at the sampling frequency 8 kHz. We used the first 200 samples of RIRs, including the second-order reflections.

The network architecture comprises 3 hidden layers with 256 neurons, the last layer of which is linear. The hyperparameter of SIREN was $\omega = 12$ for the first and the hidden layers. The parameter λ in Eq. (6) was set to $\lambda = 1.0 \times 10^{-6}$, and the network was trained for 15000 iterations. The optimizer was Adam. The learning rate was 1.0×10^{-4}.

The estimation accuracy was evaluated using the normalized mean square error (NMSE), defined as

$$\mathrm{NMSE} = 10 \log_{10} \frac{1}{M} \sum_{m=1}^{M} \frac{\|\hat{\mathbf{h}}_m - \mathbf{h}_m\|_2^2}{\|\mathbf{h}_m\|_2^2}, \tag{8}$$

where $\hat{\mathbf{h}}_m$ and $\mathbf{h}_m$ denote m-th estimated signal and ground truth, respectively.

4.2 Result

Figure 3 compares the loss functions and NMSEs at SNR 10, 13, and 15 dB for the SIREN, PI–SIREN, and the proposed methods, respectively.

In SIREN, from Figs. 3(a), the loss function converged and decayed as the number of iterations increased. However, NMSE did not improve with increasing iterations and was above 0 dB for nearly all the iterations. In PI–SIREN, as shown in Figs. 3(b), at SNR of 15 dB, both the loss function and NMSE decreased and converged. At SNRs of 13 and 10 dB, NMSE decreased once, however, the estimation accuracy degraded as learning proceeded. The NMSE was degraded owing to the overfitting of the microphone signal with noise.

Early stopping of learning is a method used to prevent overfitting [23]. In these methods, the data used is a single

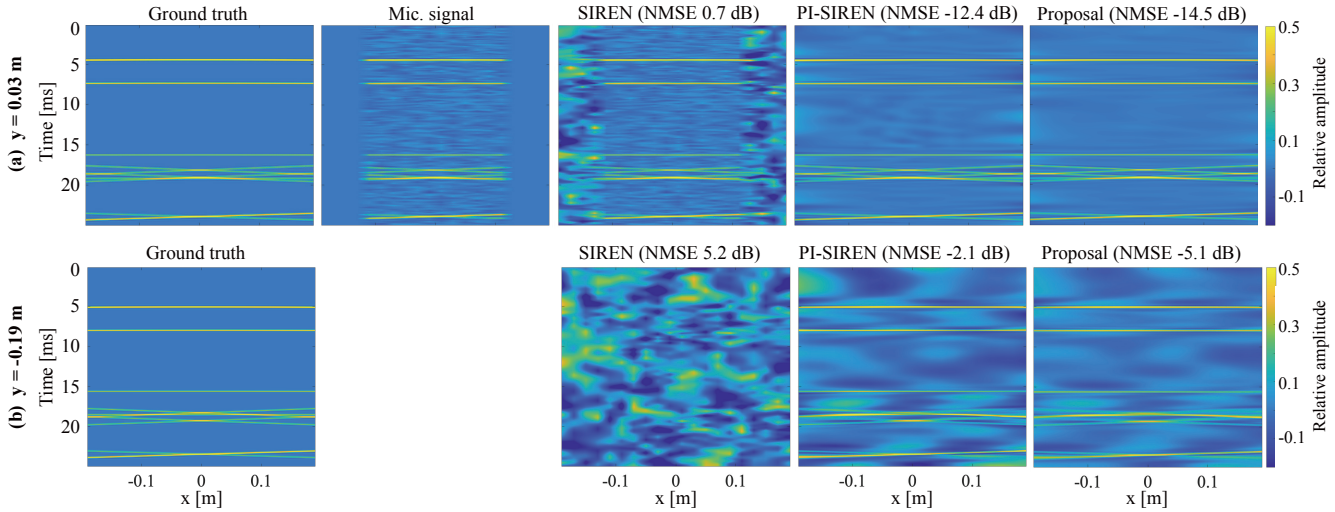

Fig. 4 Estimated signals at 15K iterations with SNR 10 dB. The upper figures show signals at $y = 0.03$ m (including the microphone positions). The bottom figures show signals at $y = -0.19$ m (not including the microphone positions).

data set. In addition, it is not possible to obtain all the data required for the NMSE calculation in advance. Thus, the timing of early stopping must be determined by only the loss function. From Fig. 3(b), the convergence of the loss function in PI–SIREN exhibited the same trend regardless of the magnitude of the noise, but the timings of the convergences of the NMSEs were different. Moreover, if the stop of learning is delayed, the overfitting to the microphone noise will begin. Therefore, it is difficult to determine when to stop learning based on the loss function only.

As shown in Figs. 3(c), NMSE remained constant as the number of training iterations increased at 13 dB and 10 dB SNR for the proposed method. Thus, the estimation of the proposed method was stable, regardless of the timing of learning stops after convergence. After the loss converged, it changed repeatedly and rapidly compared to PI–SIREN. This is because the proposed method changes the loss when the microphone error is less than the error tolerance.

As shown in Table 1, we compared NMSEs for 5K, 10K, and 15K iterations. In Table 1, $\mathrm{NMSE}_{\mathrm{all}}$ indicates the overall estimation error. $\mathrm{NMSE}_{\mathrm{est}}$ and $\mathrm{NMSE}_{\mathrm{mic}}$ indicate NMSEs for the estimation and microphone positions, respectively. At SNR 15 dB, both PI–SIREN and the proposed method exhibited similar accuracy, which did not depend on the number of training. However, at 15K iterations of SNR 13 dB, the NMSE and $\mathrm{NMSE}_{\mathrm{est}}$ were approximately 1 dB larger than those of the 5K iterations, whereas the $\mathrm{NMSE}_{\mathrm{all}}$ and $\mathrm{NMSE}_{\mathrm{est}}$ were lower in the proposed method. In particular, the degradation of estimation accuracy in PI–SIREN was most noticeable at SNR 10 dB. The NMSE of PI–SIREN degraded by approximately 1–2 dB for each additional 5K of training, whereas the NMSE of the proposed method remained constant even as the training iterations increased.

Therefore, in noisy environments, it was appropriate for the PI–SIREN to complete the learning as soon as the loss function decreased. In the proposed method, because the estimation accuracy does not depend significantly on the timing of training, the training can be completed at any time after the loss function is sufficiently lowered.

Finally, the 15K-th estimated signals at SNR 10 dB are compared at $y = 0.03$ m and $y = -0.19$ m. Figure 4(a) shows the estimated signals for $y = 0.03$ m. SIREN can reconstruct signals only at the microphone positions; however, it also includes noises. By contrast, the proposed method and PI–SIREN achieved approximate extrapolation and denoising of the microphone signals. The NMSE was approximately 2.1 dB lower for the proposed method compared to PI–SIREN. Figure 4(b) shows the estimated signals at $y = -0.19$ m. SIREN could not estimate all the signals because the microphone positions were not included. PI–SIREN and the proposed method could estimate the direct and reflected sounds, although the estimation accuracy was degraded compared with when $y = 0.03$ m. Therefore, the proposed method can stabilize the estimation accuracy against noise components as the number of training iterations increases.

5. Conclusion

In this study, we investigated the spatial extrapolation accuracies of two-dimensional RIRs using PI–SIREN and proposed a noise-robust method by introducing error tolerance into the loss function of PI–SIREN. In the simulation experiments, the sound field was extrapolated using noise-containing microphone signals. PI–SIREN achieved spatial extrapolation of early RIRs in a 2D sound field. Furthermore, the proposed method achieved constant estimation accuracy even when noises were added to the microphone signals. In future studies, we will extend the proposed method to three-dimensional sound fields. In addition, we will compare the conventional analytical methods and the proposed deep learning method.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 22KJ2786.

References

[1] R.G. Baraniuk, "Compressive sensing [lecture notes]," IEEE Signal Processing Magazine, vol.24, no.4, pp.118–121, 2007.

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

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

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

[5] R. Mignot, L. Daudet, and F. Ollivier, "Room reverberation reconstruction: Interpolation of the early part using compressed sensing," IEEE Transactions on Audio, Speech, and Language Processing, vol.21, no.11, pp.2301–2312, Nov. 2013.

[6] 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.

[7] E. Fernandez-Grande, D. Caviedes-Nozal, M. Hahmann, X. Karakonstantis, and S.A. Verburg, "Reconstruction of room impulse responses over extended domains for navigable sound field reproduction," 2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA), pp.1–8, 2021.

[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 method," Applied Acoustics, vol.179, p.108027, 2021.

[9] I. Tsunokuni, H. Matsuhashi, and Y. Ikeda, "Spatial extrapolation of early room impulse responses with source radiation model based on equivalent source method," Audio Engineering Society Convention 152, May 2022.

[10] F. Lluís, P. Martínez-Nuevo, M. Bo Møller, and S. 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.

[11] 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, 02 2023.

[12] 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.

[13] K. Shigemi, S. Koyama, T. Nakamura, and H. Saruwatari, "Physics-informed convolutional neural network with bicubic spline interpolation for sound field estimation," 2022 International Workshop on Acoustic Signal Enhancement (IWAENC), pp.1–5, 2022.

[14] 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," Proc. Forum Acusticum, 2023.

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

[16] V. Sitzmann, J.N. Martel, A.W. Bergman, D.B. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," Proc. NeurIPS, 2020.

[17] K. Xenofon and F.G. Efren, "Room impulse response reconstuction using physics-constrained neural networks," Proc. Forum Acusticum, 2023.

[18] L. Holmstrom and P. Koistinen, "Using additive noise in back-propagation training," IEEE Transactions on Neural Networks, vol.3, no.1, pp.24–38, 1992.

[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," 35th AAAI Conference on Artificial Intelligence, AAAI 2021, p.8146 – 8154, 2021.

[20] H. Wierstorf and S. Spors, "Sound Field Synthesis Toolbox," 132nd Convention of the Audio Engineering Society, 2012.

[21] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.351–355, 2018.

[22] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," The Journal of the Acoustical Society of America, vol.65, no.4, pp.943–950, 1979.

[23] S. Theodoridis, Machine Learning: A Bayesian and Optimization Perspective 2nd Edition, Elsevier Ltd., 2020.