DOI: 10.61782/fa.2023.1182

forum acusticum 2023

IMPLICIT NEURAL REPRESENTATION WITH PHYSICS-INFORMED NEURAL NETWORKS FOR THE RECONSTRUCTION OF THE EARLY PART OF ROOM IMPULSE RESPONSES

Mirco Pezzoli* **Fabio Antonacci** **Augusto Sarti**
Dipartimento di Elettronica, Informatica e Bioingegneria, Politecnico di Milano, Italy

ABSTRACT

Recently, deep learning and machine learning approaches have been widely employed for various applications in acoustics. Nonetheless, in the area of sound field processing and reconstruction, classic methods based on the solutions of the wave equation are still widespread. Lately, physics-informed neural networks have been proposed as a deep learning paradigm for solving partial differential equations that govern physical phenomena, bridging the gap between purely data-driven and model-based methods. In this study, we exploit physics-informed neural networks to reconstruct the early part of missing room impulse responses in a uniform linear array. This methodology allows us to leverage the underlying law of acoustics, i.e., the wave equation, forcing the neural network to generate physically meaningful solutions given only a limited number of data points. The results from real measurements show that the proposed model achieves accurate reconstruction and performance in line with state-of-the-art deep learning and compressive sensing techniques while maintaining a lightweight architecture.

Keywords: *physics-informed neural network, sound field reconstruction, wave equation.*

1. INTRODUCTION

Sound field reconstruction is fundamental in augmented and virtual reality applications, where users can experience immersive audio environments. To accurately characterize the acoustic properties of a given environment, the acquisition of multichannel signals is necessary. Room impulse responses (RIRs) captured with microphone arrays are particularly useful for this and various tasks such as sound source localization [1, 2], separation [3, 4], and sound field navigation [5–7]. In fact RIRs provide a model of the sound propagation between the acoustic source and the microphone array within an environment.

The reconstruction of RIRs or sound field in general, has been a subject of extensive research, leading to the development of two primary categories of solutions: parametric and non-parametric techniques. Parametric methods [7–12] rely on simplified parametric models of the sound field to convey an effective spatial audio perception to the user. In contrast, non-parametric methods [13–17] aim to numerically estimate the acoustic field. Most of the available techniques in this class are based on compressed sensing principles [18] combined with the solutions of the wave equation [19], i.e., plane wave [20] and spherical wave [17, 21], the modal expansion [15] or the equivalent source method (ESM) [16, 22].

A third category comprising deep learning emerged as an alternative approach for sound field reconstruction and a wide range of problems in the field of acoustics [23–27]. In [28], a convolutional neural network (CNN) has been proposed for the reconstruction of room transfer functions. However, as noted in [28] the model is limited to low frequencies and the generalization is constrained by the available data set. To overcome the frequency and data set limitations, in [29], the authors proposed a *deep prior* approach [30] to RIR reconstruction in time domain. The deep prior paradigm [29] considers the structure of a CNN

**Corresponding author*: mirco.pezzoli@polimi.it.

as a regularization prior to learn a mapping from a random input to the reconstructed RIRs of an Uniform Linear Array (ULA). As a result, no extensive data set is required for training since the optimization is performed over a single ULA.

Recently, in order to exploit the underlying physics of the sound field, a physics-informed neural network (PINN) [31–33] for sound field reconstruction has been introduced in [34]. The main idea of PINN [31, 32] is to force the output of a network to follow the partial differential equations (PDE) governing the system under analysis. In particular, PDE computation is performed exploiting the automatic differentiation framework underlying the training procedure of neural networks. Following the PINN approach, in [34] the authors augmented the loss function used for training a CNN with the computation of the Helmholtz equation [34]. However, differently from standard PINNs [31], the network provides as output an estimate of the derivatives required to compute the Helmholtz equation instead of relying on automatic differentiation. Moreover, the system works at a fixed frequency (300 Hz) and it has been tested only on simulated data.

In this paper, we propose the use of a physics-informed approach for the reconstruction of the early part of RIRs. As a matter of fact, the early part of RIRs provides relevant information on the geometry of environment [35, 36] affecting the timbre and localization of acoustic sources [37]. Therefore, accurate reconstruction of the early part of RIRs [22, 38] is required, while the late reverberation is typically modelled through its statistical characteristics [7, 8, 39]. In order to avoid frequency limitations, we work in the time domain. We adopt a network that takes as input the signal domain i.e., the time and position of the microphone and provides as output an estimate of the RIRs at the given coordinates. In order to improve the performance exploiting prior knowledge on the signal domain, we employed a network structure known as SIREN [40] trained using the PINN paradigm. We refer to the adopted approach to as physics-informed SIREN (PI-SIREN). SIREN demonstrated to be an effective architecture to learn *neural implicit representations* of different signals including audio and for solving the wave equation (direct problem) [40]. However, the adoption of SIREN has not been fully explored yet for solving time-domain inverse problems in the field of multichannel acoustic processing or applying to real acoustic measurements. In this work, we investigate the use of PI-SIREN for the reconstruction of early parts of the RIRs

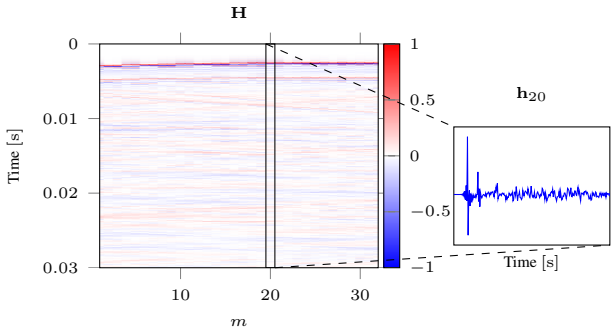


Figure 1. Example of RIRs $\mathbf{H}$ of a $M = 32$ microphones ULA.

acquired by an ULA. Results on simulations revealed that in contrast to classical PINN, PI-SIREN is a suitable architecture for RIR reconstruction. In addition, we compare the reconstructions of PI-SIREN on real data with respect to state-of-the-art solutions based on compressed sensing [41] and deep learning [29] showing improved reconstruction of the early parts of the RIRs in two of the three considered rooms.

2. PROBLEM STATEMENT

2.1 RIR data model

Let us consider an acoustic source located in $\boldsymbol{r}' = [x', y', z']^T$ and a set of M microphones acquiring the generated sound field. Assuming linear acoustics and absence of noise, the sound pressure at the mth sensor can be defined as

$$p(\boldsymbol{r}_m, t) = h(t, \boldsymbol{r}_m, \boldsymbol{r}') * s(t), \quad m = 1, \ldots, M, \quad (1)$$

where $p(\boldsymbol{r}_m, t)$ is the time-domain sound pressure at time instant t and location $\boldsymbol{r}_m$, $s(t)$ is the signal emitted by the source and $*$ denotes the linear convolution operation. The term $h(t, \boldsymbol{r}_m, \boldsymbol{r}')$ in (1) refers to the RIR between the source in $\boldsymbol{r}'$ and the sensor at $\boldsymbol{r}_m$. In general, RIR provides a description of the sound propagation in the environment from a source to a receiver and due to (1), it completely characterizes the spatial properties of the sound field. In ideal conditions with unbounded domain, the RIR is given by the well-known Green's function [19] which is a particular solution of the inhomogenous wave equation [19]

$$\nabla^2 p(\boldsymbol{r}, t) - \frac{1}{c^2} \frac{\partial^2 p(\boldsymbol{r}, t)}{\partial t^2} = \delta(\boldsymbol{r} - \boldsymbol{r}', t), \quad (2)$$

where δ is the Dirac delta function and c is the speed of sound in air.

In this work, we consider the RIRs of an ULA and the location of each mth microphone is given by the distance d between two consecutive sensors as $r_m = [x_a, (m-1)d, z_a]^T$. The values of x_a and z_a are the same for all the sensors in the ULA. It follows that the maximum frequency for aliasing-free sound field acquisition in the ULA is limited by the distance d through

$$F_{\max} = \frac{c}{2d}. \tag{3}$$

In practice, we organize the acquired RIRs in a $N \times M$ matrix defined as

$$\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_M], \tag{4}$$

where $\mathbf{h}_m \in \mathbb{R}^{N\times 1}$ is the vector containing the N-length sampled RIR of the mth microphone. In Fig. 1, an example of RIRs acquired by a ULA is shown.

2.2 RIR reconstruction problem

We assume that a limited subset, indexed as $\tilde{\mathcal{M}}$, of the ULA sensors $\mathcal{M}$ is available, and thus $\tilde{\mathcal{M}} \subseteq \mathcal{M}$ $\left(|\tilde{\mathcal{M}}| = \tilde{M} < M\right)$. The goal of RIR reconstruction is to recover the missing data exploiting the information available from RIRs in the observation points $\{\boldsymbol{r}_{\tilde{m}}\}_{\tilde{m}\in\tilde{\mathcal{M}}}$. Various techniques have been proposed in the literature to address the spatial-sampling requirement for reconstructing RIRs from an undersampled measurement set. In general, this task can be interpreted in the framework of inverse problems, and a solution to the problem can be found through the following minimization

$$\begin{aligned}\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\text{argmin}}\ J(\boldsymbol{\theta}) = \\ E\left(f(\boldsymbol{\theta}, \{\boldsymbol{r}_{\tilde{m}}\}_{\tilde{m}\in\tilde{\mathcal{M}}}) - \mathbf{H}(\{\boldsymbol{r}_{\tilde{m}}\}_{\tilde{m}\in\tilde{\mathcal{M}}})\right),\end{aligned} \tag{5}$$

where $E(\cdot)$ is a data-fidelity term, e.g., the mean squared error, between the estimated data and the observations, and $f(\boldsymbol{\theta}, \boldsymbol{r})$ is a function that generates the estimated RIRs using the parameters $\boldsymbol{\theta}$. The time dependency in (5) has been omitted for notational simplicity. It is worth noting that in (5), the evaluation of f is performed in the observation locations $\{\boldsymbol{r}_{\tilde{m}}\}_{\tilde{m}\in\tilde{\mathcal{M}}}$. However, f must be able to provide a meaningful estimate also in location that are different from the available ones. Therefore, the solution to the ill-posed problem (5) is constrained using regularization strategies. Typical techniques include compressed sensing frameworks based on assumptions about the signal model such as plane and spherical wave expansions [13], ESM [22], or the RIRs structure [41], as well as deep learning approaches [28, 29].

3. PROPOSED METHOD

In this work, we aim at solving the RIR reconstruction problem (5) in order to provide an estimate of the ULA RIRs as

$$\widehat{\mathbf{H}} = f\left(\boldsymbol{\theta}^*, \{\boldsymbol{r}_m\}_{m\in\mathcal{M}}\right), \tag{6}$$

where the function $f(\cdot)$ represents a neural network. In particular, we adopt the structure of a SIREN [40] neural network. SIREN proved to be an effective architecture for learning the so-called *neural implicit representations* of different classes of signals, including audio signals. The proposed model has the structure of a multilayer perceptron (MLP) with sinusoidal activation functions, for which the ith layer can be expressed as

$$\phi_i(\mathbf{x}_i) = \sin\left(\omega_0 \mathbf{x}_i^T \boldsymbol{\theta}_i + \mathbf{b}_i\right), \tag{7}$$

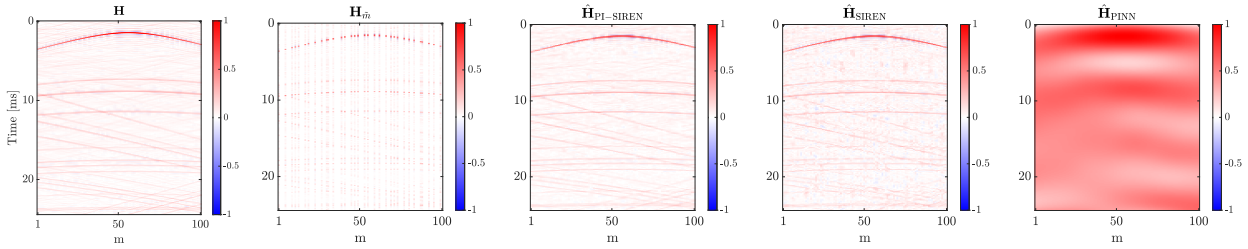
where $\mathbf{x}_i$, $\boldsymbol{\theta}_i$, and $\mathbf{b}_i$ are the input vector, the weights and the biases of the ith layer, respectively, while ω_0 is an initialization hyper-parameter [40]. The adopted SIREN architecture is thus a composition of L layers

$$f(\boldsymbol{\theta}, \mathbf{x}) = (\phi_L \circ \phi_{L-1}, \ldots, \phi_1)(\mathbf{x}), \tag{8}$$

where $\mathbf{x}$ is the input of the network while $\boldsymbol{\theta}$ is the set of learnable parameters. Following the paradigm of neural implicit representations, the SIREN model takes as input the signal domain, namely the sensor position $\boldsymbol{r}_m$ and the time instant t and provides as output an estimate of the RIR $\hat{h}(t, \boldsymbol{r}_m)$. Hence, the role of the network is to provide a parameterized representation of the signals through the parameters of the MLP. Essentially, during the training, the neural network overfits the available signals becoming an alternative implicit representation of the RIRs.

Although we can fit the available RIRs through SIREN, regularization strategies are required in order to provide meaningful results in different points of the domain i.e., to estimate the missing RIRs.

Here, we consider training SIREN using the PINN approach, denoting the solution as PI-SIREN. Using as the target for the training the reconstruction of the observation only, there is no guarantee that the solution follows the physical law of the underlining problem, namely the wave equation [19]. PINN are forced to learn solutions that follows the PDE of the underlying physics in order to obtain improved results. This approach exploits the prior



(a) (b) (c) (d) (e)

Figure 2. (a) Simulated RIRs $\mathbf{H}$. (b) The observation $\mathbf{H}_{\tilde{m}}$ of $\tilde{M} = 33$ microphones employed as input for the networks. The reconstructions obtained using PI-SIREN (c), SIREN (d) and PINN (e).

knowledge on the system in order to regularize the estimation of the neural network. Therefore, we adopted the following loss function for training PI-SIREN which includes a physics-informed term as

$$\mathcal{L} = \frac{1}{\tilde{M}} \sum_{\tilde{m} \in \tilde{\mathcal{M}}} \|\hat{h}_{\tilde{m}} - h_{\tilde{m}}\|_2^2 + \lambda \frac{1}{M} \sum_{m=1}^{M} \left\| \frac{1}{c^2} \frac{\partial^2 \hat{h}_m}{\partial t^2} - \nabla^2 \hat{h}_m \right\|_2^2, \tag{9}$$

where $\|\cdot\|_2$ is the ℓ_2 norm, the first term of the summation represents a distance between the prediction and the available data, while the second term corresponds to the PDE loss given by the wave equation and weighted by parameter λ. While the first part of (9) makes the network fit the observation, the PDE term constraints the output to follow the wave equation. The use of the PDE loss results in a regularized solution since the output conforms with the underlying physical equation. Once trained, PI-SIREN can be used to obtain the RIRs at the missing and available positions of the ULA simply feeding the network with the locations $\boldsymbol{r}_m$, $m = 1, \ldots, M$, and the different time instants t.

4. NUMERICAL EXPERIMENTS

4.1 Setup

We evaluate the performance of PI-SIREN for RIR reconstruction on both simulated and measured data from [41]. We considered an ULA of $M = 100$ microphones with distance $d = 2.02\,\text{cm}$ which gives a maximum frequency (3) $F_{\max} = 8.489\,\text{kHz}$. The simulated RIRs have been computed at sampling rate $8\,\text{kHz}$ using the image source technique [42] for a shoe-box room of dimensions $6\,\text{m} \times 4\,\text{m} \times 3\,\text{m}$ and reverberation time $\text{T}_{60} = 0.5\,\text{s}$. For this work we limit the analysis to the first $20\,\text{ms}$ of the RIRs which corresponds to the early part (direct and early reflections) of the impulse response.

The proposed PI-SIREN architecture is composed of $L = 5$ layers of 256 neurons in which the last layer is linear. The network has a total of 198401 trainable parameters. The initialization frequency ω_0 in (7) is set to 15 for the first layer, while as in [40] $\omega_0 = 30$ for the hidden layers. The network is trained for 2000 iterations using Adam optimizer with learning rate equal to 10^{-4}. The weight parameter in (9) has been experimentally set to $\lambda = 5 \cdot 10^{-15}$. Similarly to [29,41], we evaluate the reconstruction performance in terms of the normalized mean square error (NMSE) between the reconstructed data and the reference RIRs defined as [41]

$$\text{NMSE}\left(\widehat{\mathbf{H}}, \mathbf{H}\right) = 10 \log_{10} \frac{1}{M} \sum_{m=1}^{M} \frac{\|\widehat{\mathbf{h}}_m - \mathbf{h}_m\|^2}{\|\mathbf{h}_m\|^2}, \tag{10}$$

where $\widehat{\mathbf{h}}_m$ is the mth RIR estimate provided by the reconstruction technique. The observations are computed considering $\tilde{M} = \{20, 33\}$ microphones randomly selected as in [41], which corresponds to $1/5$ and $1/3$ of available sensors, respectively.

4.2 Results

In order to assess the effectiveness of the proposed PI-SIREN methodology in terms of architecure and training strategy, we evaluate the reconstruction performance on simulated data. We compare the reconstruction of PI-SIREN with respect to a classical PINN architecture [31] and SIREN trained without the wave equation term in the

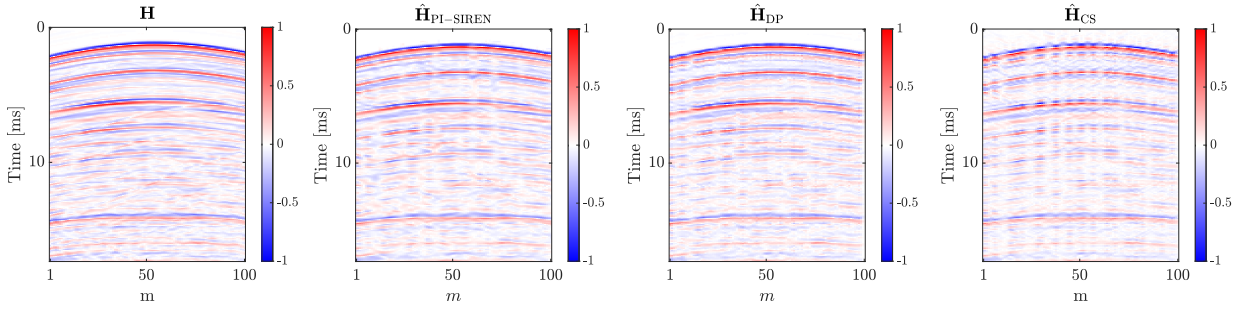


Figure 3. Reconstruction of the RIRs of Munin room using $\tilde{M} = 20$ available sensors. (a) The measured RIRs $\mathbf{H}$. The reconstructions are obtained using the proposed model $\hat{\mathbf{H}}_{\text{PI-SIREN}}$ (b), DP [29] (c) and CS [41] (d).

loss function (9). The PINN shares the same structure of PI-SIREN in terms of layers and parameters, however tanh is adopted as nonlinear function of the neurons. In Fig. 2, the RIRs $\mathbf{H}$ along with the observation $\mathbf{H}_{\tilde{m}}$ with $\tilde{M} = 33$ and the obtained reconstructions are reported.

From Fig. 2(e), we can observe that PINN fails to reconstruct the RIRs, obtaining a $\text{NMSE}_{\text{PINN}} = 14.5\,\text{dB}$. The adoption of the sinusoidal activation function in SIREN (see Fig. 2(d)) determines an improved reconstruction performance with $\text{NMSE}_{\text{SIREN}} = -7.1\,\text{dB}$. Inspecting Fig. 2(d), we can observe that SIREN reconstructs the direct path and the early reflections in $\hat{\mathbf{H}}_{\text{SIREN}}$, filling the missing channels. It follows that SIREN provides an effective implicit representation of the considered signals thanks to the use of the sinusoidal nonlinearity. In [43], the authors show how a two-layers SIREN can be related to a discrete cosine transform (DCT) of the signal. In the context of this work, the consideration in [43] can be loosely interpreted in terms of a real-valued plenacoustic representation [44] of the RIRs. Nonetheless, the reconstruction in Fig. 2(d) contains noisy components and the estimated wave fronts at some of the missing locations are incoherent. In Fig. 2(c), the output of PI-SIREN is depicted. It is possible to note that, differently from the basic SIREN, PI-SIREN is able to estimate the RIRs more accurately, coherently reconstructing the wave fronts at the missing locations. The reconstruction of PI-SIREN achieves a $\text{NMSE}_{\text{PI-SIREN}} = -11.2\,\text{dB}$ which is lower with respect to both SIREN and the PINN. Through the physics-informed loss function in (9), in fact, the output of the network is forced to conform with the physical prior of the wave equation. Therefore, the adoption of the physics-informed loss function in PI-SIREN allows us to obtain an improved performance.

Room	Balder		Freja		Munin	
Mic.	20	33	20	33	20	33
	NMSE [dB]					
CS	-5.87	-11.47	**-5.89**	**-11.01**	-7.52	-15.25
DP	-5.52	-11.44	-4.68	-9.21	-8.98	-16.03
PI-SIREN	**-6.26**	**-11.74**	-5.65	-10.61	**-10.00**	**-16.17**

Table 1. NMSE of the considered techniques at different downsampling conditions for the three rooms.

4.3 Experimental results

We evaluate the performance of PI-SIREN on real RIRs measured in three rooms [41] and we compare the estimated reconstruction with respect to the compressed sensing method (CS) in [41] and the deep prior (DP) methodology of [29]. The employed ULA consists of $M = 100$ sensors with distance $d = 3\,\text{cm}$. The rooms are named "Balder", "Freja" and "Munin" and the estimated reverberation times T_{30} are $0.32\,\text{s}$, $0.46\,\text{s}$ and $0.63\,\text{s}$, respectively.

In Table 4.2, the NMSE obtained for the different rooms are reported. As expected, when a lower number of sensors $\tilde{M} = 20$ is available the reconstruction performance is reduced for all the considered techniques. The performance of the three methods is in line for all the considered scenarios. However, the proposed PI-SIREN is able to achieve lower NMSE in Balder and Munin rooms for both the adopted undersampling conditions. Interestingly, the best reconstruction performance is achieved for

room Munin for every method. This room has the highest T_{30}, but as notices in [41], the density of early reflections is lower with respect to the other rooms, making the reconstruction less challenging. CS obtained the lowest NMSE in room Freja. However, the difference with respect of the proposed model is limited to $0.24\,\mathrm{dB}$ and $0.4\,\mathrm{dB}$ for the $\tilde{M} = 20$ and $\tilde{M} = 33$ scenarios, respectively.

In Fig. 3, the reconstructions of $\mathbf{H}$ for room Munin given $\tilde{M} = 20$ microphones are reported. The reference RIRs are depicted in Fig. 3(a). Inspecting the reconstruction in Fig. 3, we can note that all the three methods managed to reconstruct the main structure of the RIRs. However, the reconstruction provided by CS presents an underestimation of the RIRs at the missing locations which are seen as light vertical stripes in Fig. 3(d). Instead $\mathbf{H}_{\mathrm{PI-SIREN}}$ and $\mathbf{H}_{\mathrm{DP}}$ have a similar performance with a lower reconstruction error ($\mathrm{NMSE}_{\mathrm{PI-SIREN}} = -10\,\mathrm{dB}$) for the proposed model compared to DP ($\mathrm{NMSE}_{\mathrm{DP}} = -8.89\,\mathrm{dB}$).

5. CONCLUSION

In this work we proposed the use of PINNs for the reconstruction of early part of RIRs. The devised architecture consists of a SIREN neural network trained exploiting the physics-informed neural network framework. This allows us to impose the governing wave equation to the solution of the RIR reconstruction. The results show that the SIREN architecture itself provides an implicit representation of the data. Moreover, the adoption of the physics-informed training demonstrated to improve the reconstruction performance. We investigated the application of the proposed model on real data, showing competitive results with state-of-the-art techniques based on compressed sensing and deep learning. The proposed technique is appealing since it synergistically exploits the flexibility of deep learning and the prior knowledge of physics. We foresee the future of this work concerning the network design and the modeling of the whole RIRs that can improve the performance and the applicability with respect to the current results.

6. ACKNOWLEDGMENTS

This work has been funded by "REPERTORIUM project. Grant agreement number 101095065. Horizon Europe. Cluster II. Culture, Creativity and Inclusive society. Call HORIZON-CL2-2022-HERITAGE-01-02."

7. REFERENCES

[1] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, 2017.

[2] M. Cobos, M. Pezzoli, F. Antonacci, and A. Sarti, "Acoustic source localization in the spherical harmonics domain exploiting low-rank approximations," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[3] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[4] M. Pezzoli, J. J. Carabias-Orti, M. Cobos, F. Antonacci, and A. Sarti, "Ray-space-based multichannel nonnegative matrix factorization for audio source separation," *IEEE Signal Processing Letters*, vol. 28, pp. 369–373, 2021.

[5] J. G. Tylka and E. Y. Choueiri, "Fundamentals of a parametric method for virtual navigation within an array of ambisonics microphones," *Journal of the Audio Engineering Society*, vol. 68, no. 3, pp. 120–137, 2020.

[6] L. McCormack, A. Politis, T. McKenzie, C. Hold, and V. Pulkki, "Object-based six-degrees-of-freedom rendering of sound scenes captured with multiple ambisonic receivers," *Journal of the Audio Engineering Society*, vol. 70, no. 5, pp. 355–372, 2022.

[7] M. Pezzoli, F. Borra, F. Antonacci, S. Tubaro, and A. Sarti, "A parametric approach to virtual miking for sources of arbitrary directivity," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2333–2348, 2020.

[8] V. Pulkki, S. Delikaris-Manias, and A. Politis, *Parametric time-frequency domain spatial audio*. Wiley Online Library, 2018.

[9] M. Pezzoli, F. Borra, F. Antonacci, A. Sarti, and S. Tubaro, "Estimation of the sound field at arbitrary positions in distributed microphone networks based on distributed ray space transform," in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, pp. 186–190, IEEE, 2018.

[10] M. Pezzoli, F. Borra, F. Antonacci, A. Sarti, and S. Tubaro, "Reconstruction of the virtual microphone signal based on the distributed ray space transform," in *26th European Signal Processing Conference (EUSIPCO)*, pp. 1537–1541, IEEE, 2018.

[11] O. Thiergart, G. Del Galdo, M. Taseska, and E. A. P. Habets, "Geometry-based spatial sound acquisition using distributed microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2583–2594, 2013.

[12] L. McCormack, A. Politis, R. Gonzalez, T. Lokki, and V. Pulkki, "Parametric ambisonic encoding of arbitrary microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2062–2075, 2022.

[13] S. Koyama and L. Daudet, "Sparse representation of a spatial sound field in a reverberant environment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 172–184, 2019.

[14] J. G. Ribeiro, S. Koyama, and H. Saruwatari, "Kernel interpolation of acoustic transfer functions with adaptive kernel for directed and residual reverberations," *arXiv preprint arXiv:2303.03869*, 2023.

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

[16] 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 Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1929–1941, 2017.

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

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

[19] E. G. Williams, *Fourier Acoustics*. London, UK: Academic Press, 1999.

[20] W. Jin and W. B. Kleijn, "Theory and design of multizone soundfield reproduction using sparse methods," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2343–2355, 2015.

[21] 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)*, pp. 151–155, IEEE, 2017.

[22] 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.

[23] M. Olivieri, M. Pezzoli, F. Antonacci, and A. Sarti, "Near field acoustic holography on arbitrary shapes using convolutional neural network," in *29th European Signal Processing Conference (EUSIPCO)*, pp. 121–125, IEEE, 2021.

[24] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle, "Machine learning in acoustics: Theory and applications," *The Journal of the Acoustical Society of America (JASA)*, vol. 146, no. 5, pp. 3590–3628, 2019.

[25] M. Olivieri, R. Malvermi, M. Pezzoli, M. Zanoni, S. Gonzalez, F. Antonacci, and A. Sarti, "Audio information retrieval and musical acoustics," *IEEE Instrumentation & Measurement Magazine*, vol. 24, no. 7, pp. 10–20, 2021.

[26] C. Campagnoli, M. Pezzoli, F. Antonacci, and A. Sarti, "Vibrational modal shape interpolation through convolutional auto encoder," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 261, (Seoul, Korea), pp. 5619–5626, Institute of Noise Control Engineering, August 2020.

[27] 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.

[28] 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.

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

[30] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9446–9454, 2018.

[31] 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," *Journal of Computational physics*, vol. 378, pp. 686–707, 2019.

[32] S. Cuomo, V. S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli, "Scientific machine learning through physics–informed neural networks: where we are and what's next," *Journal of Scientific Computing*, vol. 92, no. 3, p. 88, 2022.

[33] 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.

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

[35] I. Dokmanić, Y. M. Lu, and M. Vetterli, "Can one hear the shape of a room: The 2-d polygonal case," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 321–324, IEEE, 2011.

[36] F. Antonacci, J. Filos, M. R. Thomas, E. A. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2683–2695, 2012.

[37] 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.

[38] B. Alary and A. Politis, "Frequency-dependent directional feedback delay network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 176–180, IEEE, 2020.

[39] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual evaluation of model-and signal-based predictors of the mixing time in binaural room impulse responses," *Journal of the Audio Engineering Society*, vol. 60, no. 11, pp. 887–898, 2012.

[40] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7462–7473, 2020.

[41] E. Zea, "Compressed sensing of impulse responses in rooms of unknown properties and contents," *Journal of Sound and Vibration*, vol. 459, p. 114871, 2019.

[42] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[43] F. Pistilli, D. Valsesia, G. Fracastoro, and E. Magli, "Signal compression via neural implicit representations," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3733–3737, IEEE, 2022.

[44] T. Ajdler, L. Sbaiz, and M. Vetterli, "The plenacoustic function and its sampling," *IEEE Transactions on Signal Processing*, vol. 54, no. 10, pp. 3790–3804, 2006.