Olivieri *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2024) 2024:42
https://doi.org/10.1186/s13636-024-00366-2

EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY

Open Access

Physics-informed neural network for volumetric sound field reconstruction of speech signals

Marco Olivieri[1*], Xenofon Karakonstantis[2*], Mirco Pezzoli[1], Fabio Antonacci[1], Augusto Sarti[1] and Efren Fernandez-Grande[2]

Abstract

Recent developments in acoustic signal processing have seen the integration of deep learning methodologies, alongside the continued prominence of classical wave expansion-based approaches, particularly in sound field reconstruction. Physics-informed neural networks (PINNs) have emerged as a novel framework, bridging the gap between data-driven and model-based techniques for addressing physical phenomena governed by partial differential equations. This paper introduces a PINN-based approach for the recovery of arbitrary volumetric acoustic fields. The network incorporates the wave equation to impose a regularization on signal reconstruction in the time domain. This methodology enables the network to learn the physical law of sound propagation and allows for the complete characterization of the sound field based on a limited set of observations. The proposed method's efficacy is validated through experiments involving speech signals in a real-world environment, considering varying numbers of available measurements. Moreover, a comparative analysis is undertaken against state-of-the-art frequency domain and time domain reconstruction methods from existing literature, highlighting the increased accuracy across the various measurement configurations.

Keywords Sound field reconstruction, Physics-informed neural network, Time domain, Speech signals

1 Introduction

Sound field reconstruction (SFR) is one of the main challenges in the literature of acoustic signal processing. It consists of the estimation of the acoustic field over an extended area using a limited set of multichannel acquisitions, typically obtained with distributed microphones or arrays. SFR is of paramount importance in augmented and virtual reality applications, where the user experience relies on immersive audio environments [1]. A complete acquisition of an acoustic environment, i.e., satisfying the Nyquist-Shannon sampling theorem in the audible range up to 20 kHz, imposes an extremely dense sampling of the space with a distance lower than 1 cm for two consecutive microphones. This condition limits the practical implementation of such applications raising the interest in SFR for reducing the number of required measurements. Beyond its critical role in augmented and virtual reality applications, SFR finds application across diverse problems including sound field control [2], source separation [3], and localization [4].

Several solutions to SFR have been introduced in the literature, predominantly tackling the reconstruction in terms of an inverse problem. The inference relies on a few selected measurement points, from which the field at other locations can be estimated, via interpolation or

*Correspondence:
Marco Olivieri
marco1.olivieri@polimi.it
Xenofon Karakonstantis
xenoka@dtu.dk
[1] Dipartimento di Elettronica, Informatica e Bioingegneria, Politecnico di Milano, Milano, Italy
[2] Department of Electrical & Photonics Engineering, Technical University of Denmark, Kongens Lyngby, Denmark

extrapolation. In the literature, we can identify three-main classes of SFR methodologies: non-parametric or expansion-based [5–8], parametric [9–11], and deep learning (DL) [12–14].

In order to estimate the coefficients of the expansion-based representation, most techniques rely on compressed sensing principles [15] exploiting a priori assumptions on the sparsity of the data in time, space, or frequency. In [5], the equivalent source method [16] (ESM) is adopted in order to model the direct sound of sound sources through a sparse set of Green's functions, while reverberation is described by means of a dictionary of plane waves and a low-rank residual. The solution in [5] is found using sparsity-based optimization. A similar strategy, defined in the spherical harmonics domain, is proposed in [8]. The authors of [17], instead, exploit the ESM and the sparsity of room impulse response (RIR) signals in the time domain to represent the far-field sound field in a source-free volume, leading to the time equivalent source method (TESM). In contrast, [18] further extends this approach to dynamically regularize the solution based on the reverberation characteristics present in a room. More recently, the utilization of dictionary learning featuring more spatially constrained representations of the acoustic environment has demonstrated efficacy in achieving generalization across diverse sound fields [19].

Another popular strategy relies on the so-called kernel ridge regression (KRR) [20] which provides a least-square-based solution to the reconstruction. One advantage of KRR lies in the lightweight implementation provided by linear filtering. In [6], a sound field interpolation strategy based on KRR has been introduced and has since been extended to multiple arrays [21]. This method exploits the expansion into infinite sum plane wave functions in order to constrain the interpolation to satisfy the Helmholtz equation, inherently imposing a physical prior to the solution. Variations to the kernels function have been proposed, including prior information on the source direction [22], constraints on the reciprocity principle [23], or mixed models for the reverberation [24]. In [25], the authors adopt a probabilistic approach to sound field reconstruction based on Gaussian process regression which similarly to KRR exploits a kernel function to model the spatial correlation of the sound field.

Unlike expansion-based solutions described above, parametric approaches [9, 10, 26–28] are not aimed at numerical reconstruction of the sound field. In fact, parametric techniques rely on simplified signal models that describe the sound field by the means of a few parameters, e.g., the source location, its direct signal, and the diffuse field with the goal to convey a perceptually convincing reconstruction. Typically, the parameters are estimated from the microphone signals using beamforming [29] or other linear filtering [30, 31], and the sound field reconstruction is provided at the user through the adopted signal model.

Following the success of deep neural network (DNN) in several problems of acoustics [12, 13, 30, 32–37], deep learning solutions became a popular approach for SFR. A first work in [12] adopted a convolutional neural network (CNN) for the reconstruction of room transfer functions. The main limitation in applying classical DNN in SFR is represented by the large amount of data required for the training of the model. This typically translates in a reconstruction limited to low frequencies or lacking of generalization to unseen room conditions [12]. In order to overcome such limitations, two main strategies can be identified in the literature. On the one side, generative models [38–40] have been introduced to improve the ability of the network in reconstructing meaningful sound field. On the other side, domain knowledge has been employed to help the neural network to follow the mathematical laws of the underlying physical problem, e.g., for acoustic propagation phenomena.

In [38], a generative adversarial network (GAN) has been trained to solve SFR problem. Different GAN designs have been tested in [38], and the results revealed improved reconstruction and bandwidth extension with respect to the customary plane wave decomposition both on simulated and real data sets. More recently, GANs has been combined with the physical prior given by plane wave expansion in order to exploit acoustic models in the generation. In particular, in [39], the authors use a generator network to compute the coefficients of the plane wave expansion whose reconstruction is evaluated by the discriminator network for the adversarial training. Unlike the aforementioned techniques that work in the frequency domain, a time domain model has been introduced in [14], where a deep-prior paradigm [41] for the generation of RIR is proposed. The deep prior technique employs the structure of CNN as a regularizer of the SFR solution, hence does not require a training data set, but the network is rather trained using a pre-element training. Although very accurate also when a small set of pressure data is available, the applicability of such techniques strongly depends on the training set with limitations in the generalization to different rooms, e.g., non-rectangular rooms with complex boundary conditions [12, 13].

More recently, the use of physics-informed neural network (PINN) has been investigated to bridge the gap between model-based solutions, which are constrained by the underlying modeling assumptions, and data-driven methods, whose solutions largely depend on training data. PINNs aim to regularize the estimates of a neural network to follow a given partial differential

equation (PDE) governing the system under analysis, thus providing physics-informed solutions. In the context of SFR, authors in [42] proposed a CNN whose loss function penalizes deviation from the Helmholtz equation [43]. The selected PDE is computed through numerical methods to avoid the spatial discretization of the pressure distribution. More computationally efficient approaches take advantage of the automatic differentiation framework [44] underlying the training procedure of neural networks [45].

Different PINN approaches for the reconstruction of sound fields have been presented in [46–48]. Inspired by the sinusoidal representation networks (SIREN) [49], different studies [46–48] proposed a PINN-based approach for the reconstruction of sound fields to efficiently recover the RIRs in time domain starting from a small number of available observations. Moreover, in [47, 48], PINNs are employed to comprehensively characterize all acoustic quantities in the sound field. This includes the inference of the pressure field, particle velocity field, and sound intensity flows, by exploiting the automatic differentiation principle of neural network. As do the majority of DL-based methods, they focus on the reconstruction of height-invariant scenarios, typically by limiting to the case with sources and the emitters placed in the same plane.

In this manuscript, we aim at extending the methodology introduced in [46, 47] to reconstruct arbitrarily sound field in 3D target regions starting from different sets of available recordings. We exploit the potential of DL estimates with the regularization provided by the physical knowledge of the acoustic wave propagation. Different from considering only the RIRs, we consider speech signal in real environments, with a relevant step towards practical application scenarios. Specifically, the devised DNN is driven by the available measurements and constrained to satisfy the physical PDE of the wave equation. Numerous strategies for computing the unknown pressure function exist, ranging from implicit solutions to the homogeneous wave equation [6, 17, 36, 50, 51] to those focusing on data fidelity [9, 12, 14] and even combinations of both [39]. However, explicitly incorporating the wave equation as a constraint follows recent developments in the literature. Moreover, we directly reconstruct the signals in time domain, thus fully characterizing the sound pressure at the desired locations. The reconstruction of the proposed method has been evaluated using real data, i.e., MeshRIR dataset [52], measured in acoustically treated rooms. Starting from different random subset of available observations, results show improved reconstruction performances with respect to frequency domain [6] and time domain [17] state-of-the-art approaches. Notice that the proposed PINN approach is optimized on available data, with the PDE enforcing regularization, thus eliminating the need for training on large data sets to learn the correct solution. Moreover, PINN provides a continuous sound field representation, allowing arbitrary coordinate inputs, unlike standard networks that use discretized grids. The promising results of the devised method prove how PINN represents an appealing solution to generalize the reconstruction of sound field towards practical scenarios, thanks to the advantages of the acoustic physical priors and the potential of DL strategies to infer representations from real data. Moreover, the PDE loss in PINNs introduces soft constraints, allowing deviations from the exact solution and the desired one of the real scenario. This approach makes such solution more flexible than traditional physics-constrained methods, which do not admit deviations from the assumed model, potentially offering better performance in real-world applications. Furthermore, we show the computation of the reconstructed intensity field in the desired volumetric region.

The rest of the paper is organized as follows. In Section 2, we define the data model 2.1 and the problem formulation 2.2. Section 3 presents the proposed method based on PINN and details information about the network architecture and training procedure are described in Section 3.1. Results are reported in Section 4 along with the the description of setup and evaluation metrics in Sections 4.1 and 4.2, respectively. Moreover, the validation is presented in Section 4.3 with the state-of-the-art comparison and the computation of the intensity field in Sections 4.4 and 4.5, respectively. Finally, Section 5 draws some conclusion and future developments.

2 Data model and problem formulation

2.1 Data model

Let us consider an acoustic source located at an arbitrary position $\boldsymbol{r}_\mathrm{s} = [x_\mathrm{s}, y_\mathrm{s}, z_\mathrm{s}]^T$ and a set $\mathcal{M}$ of M measurements acquiring the generated sound field at positions $\boldsymbol{r}_m = [x_m, y_m, z_m]^T$ with $m = 1, \ldots, M$. Under the assumption of a linear time-invariant (LTI) acoustic system and in the absence of noise, the acoustic pressure acquired by the m^{th} sensor can be expressed as

$$p(\boldsymbol{r}_m, t) = h_{m,\mathrm{s}}(t) * s(t), \tag{1}$$

where symbol * denotes the linear convolution operator, $p(\boldsymbol{r}_m, t)$ is the time domain sound pressure measured at location $\boldsymbol{r}_m$, $s(t)$ is the signal emitted by the source, and $h_{m,\mathrm{s}}(t)$ is the room impulse response (RIR) function that describes the transfer path of sound from the source in $\boldsymbol{r}_\mathrm{s}$ to the receiver at $\boldsymbol{r}_m$. Notice that due to the LTI assumption and in ideal conditions with unbounded domain, the RIR in (1) is solution of the inhomogeneous wave equation

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) p(\boldsymbol{r}_m, t) = s(\boldsymbol{r}_\mathrm{s}, t), \tag{2}$$

where c is the speed of sound, ∇^2 is the Laplacian operator, and the source term s is considered to be dependent on position $\boldsymbol{r}_\mathrm{s}$ in this instance.

Obtaining a RIR experimentally typically entails the use of a receiver-emitter combination, the first of which records the variation in sound pressure over time produced by the latter through the principle of transduction (i.e., microphone and loudspeaker), as well as a post-processing stage, which requires the deconvolution of the acquired signal with respect to the emitted source signal [53, 54]. In practical applications, the sound pressure is acquired at M discrete sensor positions and is commonly organized in a $N \times M$ matrix defined as

$$\mathbf{P} = \tilde{p}(\boldsymbol{r}_m, t_n) = [\mathbf{p}_1, \ldots, \mathbf{p}_m, \ldots, \mathbf{p}_M], \tag{3}$$

where $\mathbf{p}_m \in \mathbb{R}^{N\times 1}$ is the vector containing the N-length sampled pressure (1) at position $\boldsymbol{r}_m$ and time $t_n \subset t$.

2.2 Problem formulation

This section addresses the challenge of determining a function that accurately represents the pressure field $p(\boldsymbol{r}, t)$ in a source-free region based on a restricted set of observations $\tilde{p}(\boldsymbol{r}_m, t_n)$. With $\mathcal{M}$ denoting the set of sensors or measurements and M is the cardinality of $\mathcal{M}$ (i.e., $M = |\mathcal{M}|$), the objective is delineated by

$$\begin{aligned} \hat{\boldsymbol{\beta}}_\mathrm{opt} = \arg\min_{\boldsymbol{\beta}} & \sum_{m\in\mathcal{M}} \sum_{n} \left(|p(\boldsymbol{\beta}, \boldsymbol{r}_m, t_n) - \tilde{p}(\boldsymbol{r}_m, t_n)|^2 \right) \\ \text{s.t.} & \\ & \left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) p(\boldsymbol{\beta}, \boldsymbol{r}, t) = 0, \end{aligned} \tag{4}$$

where $\boldsymbol{\beta}$ are the model parameters. The problem is constrained by the homogeneous wave equation since the bounded region does not include any sources [43]. Thus, the aim is to minimize the discrepancy between the estimated pressure function and the observed data while adhering to the wave equation constraint either implicitly (analytic basis function expansions) or explicitly. To explore the sensitivity of a model to sensor array decimation, we limit the set of measurements to a subset of the original set, denoted as $\tilde{\mathcal{M}}$, where $\tilde{\mathcal{M}} \subseteq \mathcal{M}$ and $|\tilde{\mathcal{M}}| = \tilde{M} < M$.

3 Proposed method

This study formulates the constrained optimization problem in (4) using a PINN, thus adopting a neural network $\mathcal{N}(\cdot)$ that takes as input the signal domain, i.e., the scalar time and position values where the pressure is to be evaluated, and provides as output an estimate of the pressure field. This implies that the function of pressure is given by

$$p(\boldsymbol{r}, t) = \mathcal{N}(\boldsymbol{\Theta}, \boldsymbol{r}, t), \tag{5}$$

where $\boldsymbol{\Theta}$ are the neural network parameters. Therefore, we can rewrite the optimization in (4) in order to find the optimal weights $\boldsymbol{\Theta}_\mathrm{opt}$ that parameterize the network as

$$\begin{aligned} \boldsymbol{\Theta}_\mathrm{opt} = \arg\min_{\boldsymbol{\Theta}} & \sum_{m\in\mathcal{M}} \sum_{n} |\mathcal{N}(\boldsymbol{\Theta}, \boldsymbol{r}_m, t_n) - \tilde{p}(\boldsymbol{r}_m, t_n)|^2 \\ \text{s.t.} & \\ & \left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) \mathcal{N}(\boldsymbol{\Theta}, \boldsymbol{r}, t) = 0. \end{aligned} \tag{6}$$

Notice that although the data-fidelity term in (6) is computed only with the available observations in $\mathcal{M}$, the physical regularization given by the wave equation is applied for all positions $\boldsymbol{r}$ in the domain.

Given that the optimal weights $\boldsymbol{\Theta}_\mathrm{opt}$ have been obtained, one can obtain the particle velocity via Euler's equation of motion (a result of conservation of momentum) [48]

$$\begin{aligned} \boldsymbol{u}(r, t) &= -\frac{1}{\rho}\int_{t_0}^{t} \nabla p(\boldsymbol{r}, t)\partial t \\ &= -\frac{1}{\rho}\int_{t_0}^{t} \nabla \mathcal{N}(\boldsymbol{\Theta}_\mathrm{opt}, \boldsymbol{r}, t)\partial t, \end{aligned} \tag{7}$$

between time t_0 and t, where the gradient $\nabla\mathcal{N}(\boldsymbol{\Theta}_\mathrm{opt}, \boldsymbol{r}, t)$ is obtained via automatic differentiation, while ρ represents the density of the fluid medium.

Together with the pressure and the particle velocity, the instantaneous intensity field is obtained from the product

$$\boldsymbol{i}(\boldsymbol{r}, t) = \boldsymbol{u}(\boldsymbol{r}, t) \cdot p(\boldsymbol{r}, t), \tag{8}$$

which allows for a complete characterization of the reconstructed sound field anywhere in the domain of interest.

3.1 Neural network architecture description

In the following, we describe in detail the proposed model $\mathcal{N}(\cdot)$, along with the definition of the adopted loss function and training procedure.

3.1.1 SIREN-inspired neural network

The sinusoidal representation networks (SIREN) proposed in [49] aim at leveraging periodic activation functions for implicit neural representations. Based on the PINN framework, these models are ideally suited for representing complex natural signals and their derivatives, e.g., representation of images, wavefields, video,

and sound [49]. Recent works proved the ability of SIREN approach also in the context of SFR [46, 47]. Therefore, inspired by such solutions, we propose PINN-SIREN architecture to recover audio signals in a 3D volume.

The devised architecture parameterized by the learnable weights $\mathbf{\Theta}$ and the input x has the structure of a multilayer perceptron (MLP) [55] with I layers, whose structure can be expressed as

$$\mathcal{N}(\mathbf{\Theta}, \mathsf{x}) = (\Phi_I \circ \Phi_{I-1} \circ \cdots \circ \Phi_1)(\mathsf{x}), \tag{9}$$

where symbol $\circ$ denotes the function composition operation. As for the SIREN network [49], the i^{th} layer is characterized by a sinusoidal activation function, namely

$$\Phi_i(\mathsf{x}_i) = \sin\left(\omega_0 \mathsf{x}_i^T \mathbf{\Theta}_i + \mathsf{b}_i\right), \tag{10}$$

where x_i, $\mathbf{\Theta}_i$, and b_i are the input vector, the weights and the biases of the i^{th} layer, respectively, and ω_0 represents an initialization parameter [49].

3.1.2 Input/output data

The network input is the $N \times M \times 4$ tensor representing the combination of all the possible points of the domain, i.e., time samples and 3D spatial coordinate, of the desired microphones, namely

$$[t_n, x_m, y_m, z_m] \quad n \in \{1, \ldots, N\}, m \in \mathcal{M}. \tag{11}$$

On the other hand, the output of the network is the scalar pressure values of $\hat{p}(t_n, x_m, y_m, z_m)$ which can then be reorganized into the N-length time domain pressure signals as given by Eq. (3). Moreover, in order to constrain the network to satisfy the wave Eq. (6), we compute the time and space derivatives of the output with respect to the input coordinates through the automatic differentiation principle [44].

It is worth noticing that authors in [49] normalized the input of SIREN architecture, both time and space coordinates, in the range $[-1, 1]$ to converge towards a correct numerical solution. Moreover, they showed how the initialization of the parameter ω_0 in (10) for different values spans multiple periods of the sinusoidal activations over $[-1, 1]$ and how it affects the frequency content of the reconstructed signals.

In this work, differently from the experiments presented in [49] that consider a single time domain signal, the devised network structure is modified in order to deal with 4D input domain (11). Moreover, although we consider the space components, i.e., x, y, and z, of the input in the range $[-1, 1]$ as in [49], we increase the range of the time component in $[-100, 100]$. Imposing such a different time range with respect to the space coordinates is equivalent to adding a constant multiplication term to the weights of the input layer of the network, as demonstrated in the supplementary material of [49]. This different normalization enables the network to separately process the frequency contents in time and space domains. Indeed, the problem under consideration requires to focus on high frequencies for the estimation of the speech signal and on low spatial frequencies to correctly interpolate the sound field in the desired region.

3.1.3 MLP structure

The proposed architecture (9) has been implemented in PyTorch and it is composed of $I = 5$ layers of 512 neurons, as depicted in Fig. 1. The input layer and the three hidden layers are characterized by sinusoidal activation

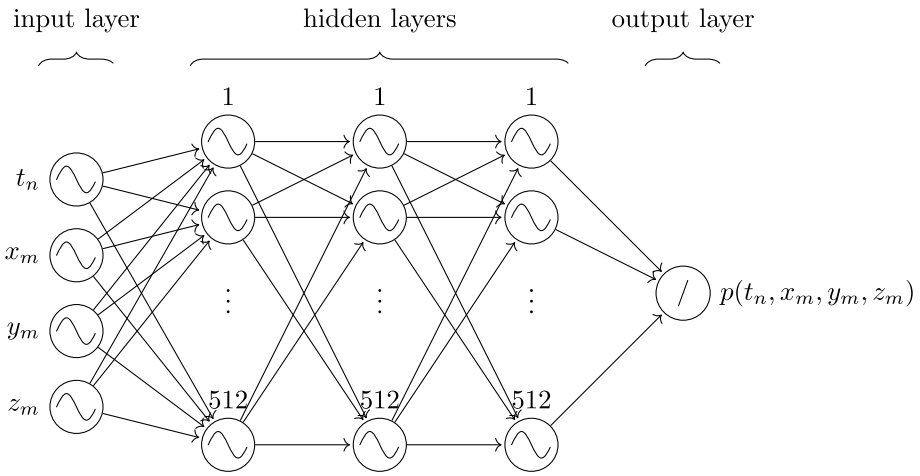

Fig. 1 Schematic of the proposed MLP architecture. The time and spatial domain points are depicted in input, while the estimate of the acoustic pressure is represented in output. The 3 hidden layers with 512 neurons are characterized by sinusoidal activation functions, while the last layer is linear

functions (10), while the output layer has a linear activation function, thus achieving a total number of learnable parameters around 790 000. Moreover, the initialization frequency ω_0 in (10) is experimentally set to 0.5 for the first layer, while $\omega_0 = 30$ for the hidden layers. Finally, due to the scaling of the input collocation points, we also scale the speed of sound to reflect the scaled travel time during training.

3.1.4 Training procedure

To incorporate the physical constraints outlined in the wave Eq. (6) to the network, we employ the following loss function

$$\mathcal{L} = \frac{1}{\mathcal{M}} \sum_{m\in\mathcal{M}} \sum_{n} \left\| \hat{p}(\boldsymbol{r}_m, t_n) - \tilde{p}(\boldsymbol{r}_m, t_n) \right\|^2 + \lambda \frac{1}{Q} \sum_{\mathrm{q}=1}^{Q} \sum_{n} \left\| \nabla^2 \hat{p}(\boldsymbol{r}_\mathrm{q}, t_n) - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \hat{p}(\boldsymbol{r}_\mathrm{q}, t_n) \right\|^2 . \tag{12}$$

Here, $\hat{p}$ and $\tilde{p}$ denote the network estimate and the measured sound pressure, respectively. The summation over Q positions is performed to ensure the fulfillment of the wave equation in a batch-like setting. Essentially, Eq. (12) represents the Lagrangian form of the constrained objective detailed in Eqs. (4) and (6). The parameter λ in (12) serves to balance the contributions of the two terms in the loss function and has been maintained at a constant value of $1 \cdot 10^{-5}$ throughout this study selecting the best between the values in the range $[10^{-2}, 10^{-8}]$ through a Grid Search approach.

We train the devised model for 3000 iterations on a single NVIDIA Titan RTX GPU with 24 GB of memory, and we adopt Adam optimizer [56] with learning rate equal to $5 \cdot 10^{-5}$ to compute the gradient descent algorithm.

4 Results

In the following, we present the results of the sound field reconstruction method considering different subsets of available observations. Firstly, we provide the setup adopted and examples of acoustic field reconstruction in the desired 3D region. Then, we compare the performance with respect to the kernel interpolation method [6] and the TESM [17] approaches.

4.1 Setup

We consider the MeshRIR dataset [52] to evaluate the devised method with measured data. This dataset collects RIRs inside a 3D region measured from a single source position placed outside the cuboidal space, as depicted in Fig. 2a.

The RIRs at each position have been measured with an omnidirectional microphone (Primo, EM272J) by recording the signal emitted with a loudspeaker (DIATONE,

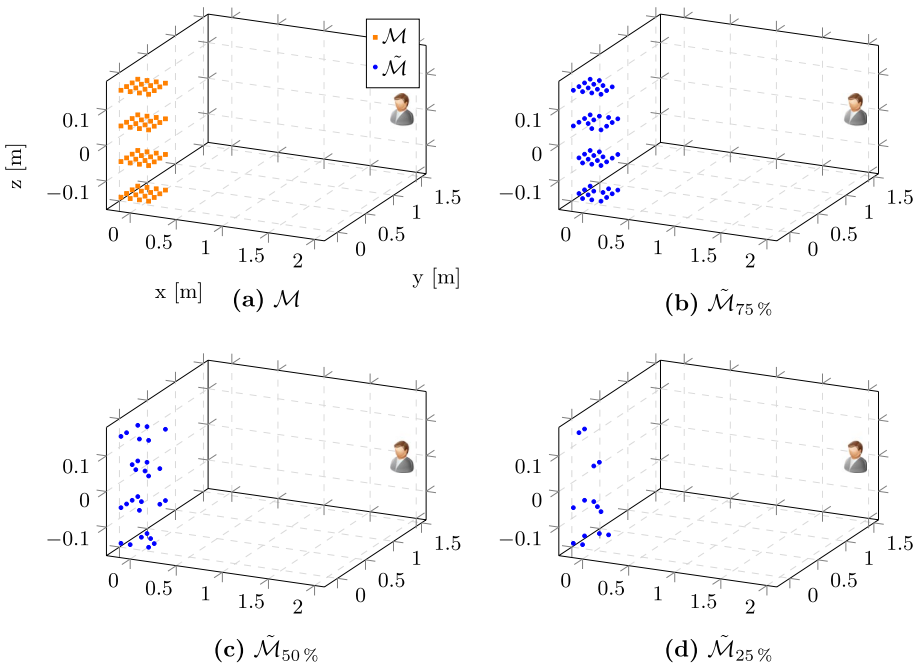

Fig. 2 Setups adopted for the devised 3D sound field reconstruction method. The square-shape orange marks are microphones in $\mathcal{M}$, while the circle-shaped blue marks represents the available observations $\tilde{\mathcal{M}}$. In **a**, the 64 positions of the acoustic field are depicted, while **b**, **c**, and **d** show the different setups with 75 %, 50 %, and 25 % of available microphones, respectively. The head icon represents the source position

DS-7), as described in [52]. The measurements have been conducted inside a room with dimensions $7 \times 6.4 \times 2.7$ m, reverberation time $T_{60} = 0.38$ s, and temperature around 26.3 °C; thus, the estimated sound speed is $c = 346.8$ m/s.

We consider different subsets of the MeshRIR data to retrieve the acoustic field in $M = 64$ positions arranged in an equally spaced grid of $4 \times 4 \times 4$ points, with inter-sensor distance $d = 0.1$ m in the x, y, and z dimension. The desired speech signals are obtained according to (1) with the convolution operator between the measured RIRs and a clean speech recording, thus obtaining the time domain signals sampled at 16 000 Hz. Moreover, due to the capacity limitations of the GPU memory, we consider signals with $N = 800$ samples, and we recover the acoustic fields for different randomly selected windows of the speech signals. Therefore, we collect the pressures in the matrix $\mathbf{P} \in \mathbb{R}^{800\times 64}$ (3), which represents also the ground truth of the reconstruction method.

With reference to Fig. 2, we consider three different scenarios in which 3/4, 1/2, and 1/4 of the total sensors $\mathcal{M}$ are available. In the following, we will denote with $\tilde{\mathcal{M}}_{75\,\%}$, $\tilde{\mathcal{M}}_{50\,\%}$, and $\tilde{\mathcal{M}}_{25\,\%}$ the set of available observations corresponding to 48, 32, and 16 randomly selected microphones, respectively.

Figure 2a shows the set $\mathcal{M}$ of microphones corresponding to the desired $|\mathcal{M}| = 64$ sensors. The different setups with $\tilde{\mathcal{M}}_{75\,\%}$, $\tilde{\mathcal{M}}_{50\,\%}$, and $\tilde{\mathcal{M}}_{25\,\%}$ of the available observations are depicted in Fig. 2b, c, and d, respectively. Aiming to recover the time domain pressure signals in the 64 positions of the 3D region from different available observations, we denote the pressure estimations with different subscripts to identify the dimension of observation subset in input, namely $\hat{\mathbf{P}}_{75\,\%}$, $\hat{\mathbf{P}}_{50\,\%}$, and $\hat{\mathbf{P}}_{25\,\%}$. The PDE term in (12) is instead evaluated over the entire set of available points $Q = M$ ensuring that the wave equation is satisfied over the entire data set.

4.2 Evaluation metrics

The performance of the reconstructions is assessed in terms of the normalized mean square error (NMSE), expressed in decibel, with perfect reconstruction when NMSE is at $-\infty$. In particular, we compute the error between the estimated data and the reference acoustic field as

$$\mathrm{NMSE}(\hat{\mathbf{P}}, \mathbf{P}) = 10 \log_{10} \frac{1}{M} \sum_{m=1}^{M} \frac{\left\|\hat{\mathbf{p}}_m - \mathbf{p}_m\right\|^2}{\|\mathbf{p}_m\|^2}. \quad (13)$$

Moreover, to evaluate the reconstruction accuracy in the positions that do not belong to the available observations, i.e., $m \notin \tilde{\mathcal{M}}$, we define with

$$\mathrm{NMSE}_{\mathrm{VAL}} = 10 \log_{10} \frac{1}{M - \tilde{M}} \sum_{m \notin \tilde{\mathcal{M}}} \frac{\left\|\hat{\mathbf{p}}_m - \mathbf{p}_m\right\|^2}{\|\mathbf{p}_m\|^2} \quad (14)$$

the NMSE of the missing measurement positions, and with

$$\mathrm{NMSE}_{\mathrm{SIG}} = 10 \log_{10} \frac{1}{\tilde{M}} \sum_{\tilde{m} \in \tilde{\mathcal{M}}} \frac{\left\|\hat{\mathbf{p}}_{\tilde{m}} - \mathbf{p}_{\tilde{m}}\right\|^2}{\|\mathbf{p}_{\tilde{m}}\|^2} \quad (15)$$

we consider the error between the available observation and the fitted data.

4.3 Validation

In order to assess the effectiveness of the proposed physics-informed methodology, we evaluate the reconstruction performance in ten different randomly selected time windows of the speech signals. Therefore, we evaluate $\hat{\mathbf{P}}$ multiple times considering all the time windows and the three different scenarios of microphone setups.

In Fig. 3, we show different examples of acoustic field reconstruction for different time snapshots. The first column depicts the ground truth $\mathbf{P}$ of the sound field, while the second, third, and last columns show the reconstructions $\hat{\mathbf{P}}_{75\,\%}$, $\hat{\mathbf{P}}_{50\,\%}$, and $\hat{\mathbf{P}}_{25\,\%}$, respectively, computed from the devised network (9). Notice that the 3D sound field is depicted in the xy-plane at different z elevations and time snapshots. Moreover, the pressure fields of each row are normalized in the range $[-1, 1]$.

Inspecting the results, we observe accurate reconstructions of the sound fields for all the different microphone setups considered, as confirmed in Fig. 3. As expected, $\hat{\mathbf{P}}_{75\,\%}$ achieves the best performance with an average NMSE $= -20.71$ dB, while the average NMSE (13) decreases to -18.81 dB and -13.88 dB when considering $\tilde{\mathcal{M}}_{50\,\%}$ and $\tilde{\mathcal{M}}_{25\,\%}$ of microphones, respectively. As a matter of fact, the different numbers of observations that sample the acoustic field highly impact the overall reconstructions in the desired 3D region. However, even starting from 16 microphones, the reconstruction is close to the ground truth, as can be observed in Fig. 3d.

4.4 SOTA comparison

In order to validate the devised methodology also with respect to state-of-the-art approaches, we compare the resulting estimations with two model-based methods and a data-driven method for sound field reconstruction. We compute the acoustic reconstructions in the desired 3D region with the kernel method [6] and the TESM [17]. To compare with a data-driven approach, we follow [39] and train a GAN on random wave fields, with a multi-layer perceptron generator network and a convolutional discriminator network. This GAN is

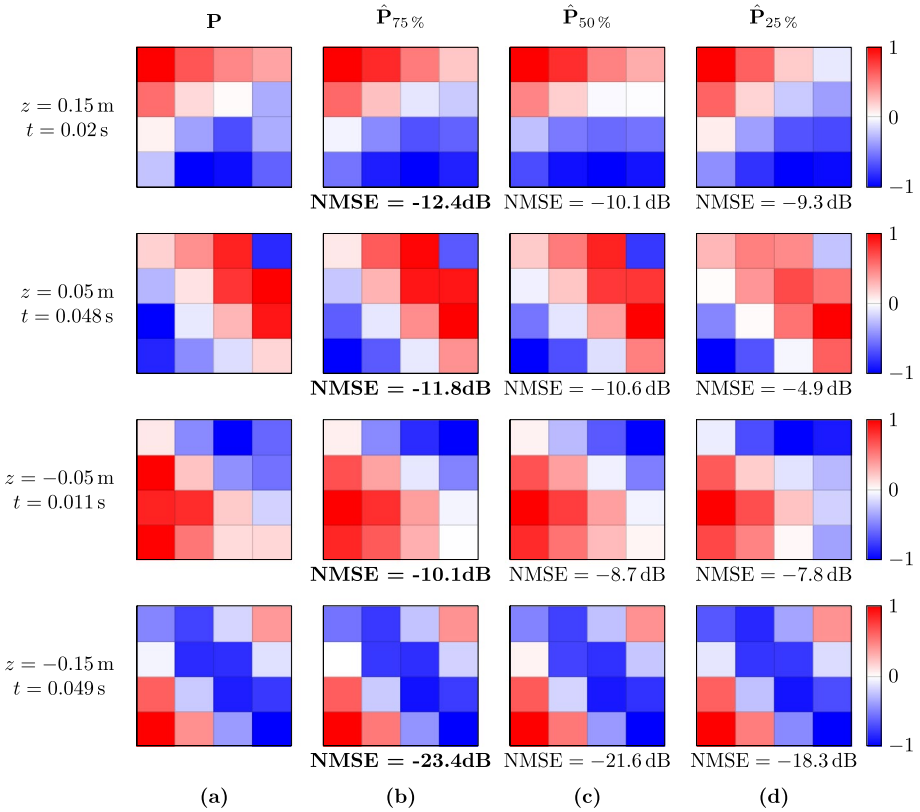

Fig. 3 Reconstruction examples at different time snapshots computed from different subsets of observations. The normalized magnitude of the acoustic field is depicted in the xy-plane at the different elevations of the desired grid points. Column **a** show the ground truth of the sound field. In column **b**, **c**, and **d**, the reconstructions computed from 75 %, 50 %, and 25 % of microphones are depicted along with the NMSE value with respect to the recovered acoustic field and the ground truth in the plane. The sound field in each row is normalized in the range $[-1, 1]$

then "inverted" to obtain the wave-field coefficients that explain the data, serving as a data-driven prior for sound field reconstruction.

The three-dimensional reconstructed sound fields using kernel ridge regression [6] are obtained using $\lambda = 10^{-3}$ as regularization parameter, a sampling rate of 16 000 Hz, and a kernel filter defined in 1025 points. On the other hand, differently from the original TESM introduced in [17], we modified the number and position of the equivalent sources. Following [18], the TESM is designed solved using variational Bayes to approximate the posterior coefficients of the equivalent sources, where we apply a zero-mean Laplace distribution over the coefficients with a scale of $\sigma_{\text{Lap}} = 0.1$, thereby promoting sparsity in the solution. Finally, the parameters of the variational distribution are optimized using the Adam optimizer with a learning rate of $\eta = 0.01$. Specifically, we arrange 400 equivalent sources in a sphere with radius 0.72 m. Although defined in frequency domain and time domain, respectively, both kernel interpolation and TESM approaches are constrained to satisfy the physical law of wave propagation (4). In general, they are able to provide good reconstructions of the acoustic fields for each of the considered subsets of observations. The GAN-based approach is trained in a similar manner to [39] in terms of adversarial objective and training hyperparameters, except in this case we implement the generator network as a multi-layer perceptron (MLP) consisting of 6 linear layers, each followed by leaky-ReLU activations. The input layer consists of 100 neurons, while the output layer corresponds to coefficients of discrete plane waves, distributed uniformly in all directions. In particular, the output layer consists of 1800 neurons, corresponding to the number of waves. Each hidden layer is composed of 512 neurons.

In Fig. 4, we show an example of reconstruction of time domain signal in a position that does not belong to the available observations, and we compared it with the ground truth pressure p. The estimate of the proposed method p^{PINN} is depicted above, while the kernel and TESM reconstruction, denoted as p^{Kernel} and p^{TESM}, respectively, are below. In general, the three reconstruction methods fit the ground truth signal. Notice that differently from the p^{PINN}, some errors in the initial and

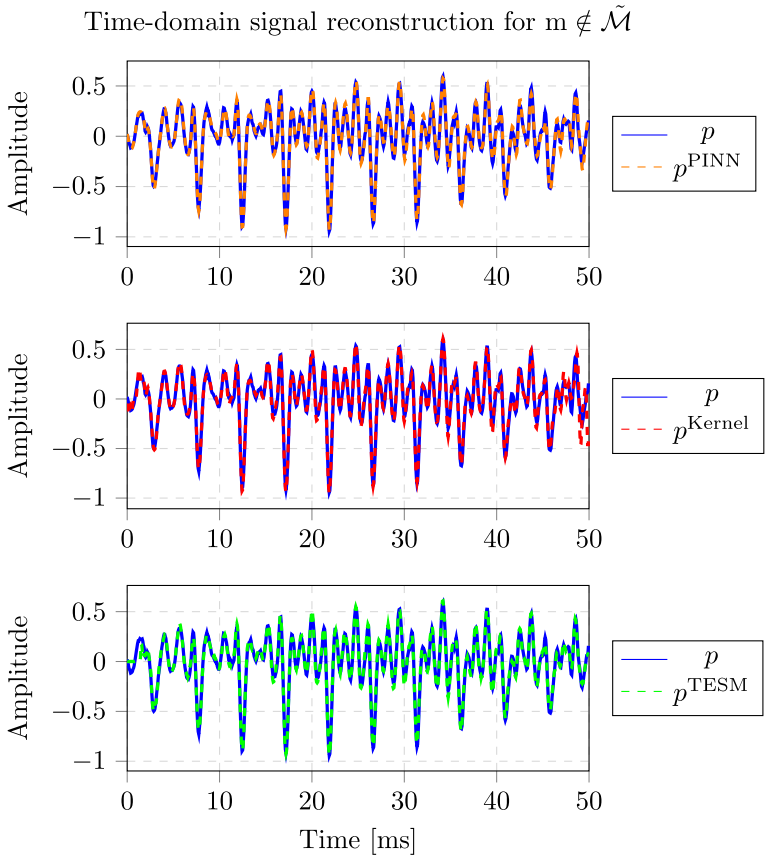

Fig. 4 Example of reconstruction comparison of an unknown signal in time domain. From above to the bottom, the estimate of the proposed method p^{PINN} and the reconstructions obtained with the kernel method p^{Kernel} and TESM p^{TESM}. Each diagram is depicted with the ground truth p

final parts of the signal are present for p^{Kernel} and p^{TESM}. This is due to the filter in the frequency domain for the kernel method [6] and to the causal convolution operator for TESM [17]. Indeed, by computing the MSE between the different approach and the ground truth pressure of the example reported in Fig. 4, we achieve −60.07 dB for the proposed approach and −49.94 dB and −55.68 dB of MSE for the kernel method and TESM, respectively.

To provide a quantitative evaluation of the proposed technique, we compute the reconstruction accuracy for all the considered time windows of the signal. Figure 5 shows the mean and standard deviation error of the NMSE for the proposed PINN method and the two model-based approaches for the three different observation setups $\tilde{\mathcal{M}}$. In particular, the NMSE_{VAL} (14) computed for $m \notin \tilde{\mathcal{M}}$ and NMSE_{SIG} (15) computed for $m \in \tilde{\mathcal{M}}$ are depicted in the Fig. 5a and b, respectively.

From Fig. 5a, we can notice that the proposed approach and the kernel method outperform TESM in average for all the configurations. As a matter of fact, the mean NMSE_{VAL} for TESM are −10.03 dB, −11.6 dB, and −12.05 dB for $\tilde{\mathcal{M}}_{25\,\%}$, $\tilde{\mathcal{M}}_{50\,\%}$, and $\tilde{\mathcal{M}}_{75\,\%}$ experiments, respectively. Although the devised method and the kernel method achieve similar reconstruction results around $\text{NMSE}_{\text{VAL}} = -12$ dB for the case $\tilde{\mathcal{M}}_{25\,\%}$, hence with $\tilde{M} = 16$ microphone observations, the proposed PINN approach reaches better reconstruction performance with a NMSE_{VAL} difference of 2.16 dB and 6.71 dB with respect to $\tilde{\mathcal{M}}_{50\,\%}$ and $\tilde{\mathcal{M}}_{75\,\%}$, respectively. Moreover, notice that the standard deviation of the proposed PINN and kernel methods are similar in the three conditions with values around 3.5 dB, 3.8 dB, and 3.88 dB when moving from $\tilde{\mathcal{M}}_{75\,\%}$ to $\tilde{\mathcal{M}}_{25\,\%}$.

Investigating the results, we believe the similar NMSE_{VAL} performances for $\tilde{\mathcal{M}}_{25\,\%}$ are due to the relatively small dimension of the 3D region. With an aperture of 0.3 m in all the three dimensions, the kernel method is able to provide good interpolations of the available signals [6]. Moreover, inspecting the NMSE_{SIG} in Fig. 5b, we can notice that the kernel method retrieves the best fit of the available observations with NMSE_{SIG} equal to −26.72 dB, −24.46 dB, and −23 dB for the 1/4, 1/2, and 3/4 of measurements. This is due to the design model of the interpolation problem, while the PINN method

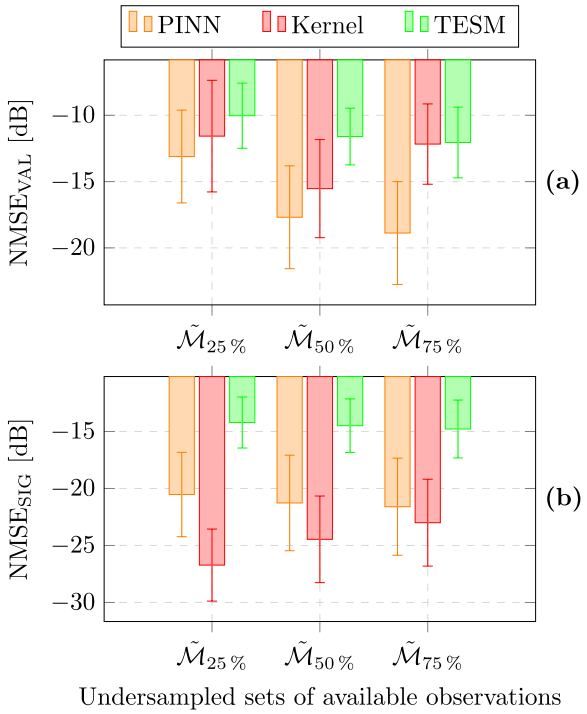

Fig. 5 Mean (column heights) and standard deviation (whiskers) comparison of the NMSE between the proposed PINN approach, the kernel, and TESM methods for different sets of available microphones. **a** The metric considering the unknown pressure signals. **b** The metrics relative to the available observations

behaves with a similar trend as the reconstruction for the unknown signals with a NMSE$_{\text{SIG}}$ around -21 dB. As a matter of fact, to provide the desired wave equation regularization, the physical term in the loss function of the proposed network (12) spreads the error on the whole desired domain without focusing only on the available observations.

Although the devised PINN method is designed in time domain, we report in Fig. 6 a frequency domain example of reconstructed signal for an unknown observation, i.e., m $\notin \tilde{\mathcal{M}}$. The magnitude of the frequency domain pressure for the three reconstructions $|P|^{\text{PINN}}$, $|P|^{\text{Kernel}}$, and $|P|^{\text{TESM}}$ are depicted in dB along with the ground truth $|P|$. We can notice that the proposed method reaches the best accuracy with respect to the two model-based methods. Although the trough around $f = 80$ Hz is not detected, at low frequencies the PINN method is able to match the desired reconstruction somewhat better than the two baselines, providing an overall improvement of the MSE between the estimate and the ground truth spectra of 5.3 dB and 3.2 dB with respect to the kernel method and TESM, respectively. Moreover, the devised solution can capture the frequency contents up to $f = 2500$ Hz, while both kernel and TESM reconstructions decrease the accuracy after 1000 Hz and 2000 Hz, respectively. Remaining in the frequency domain, Fig. 7 compares the performance of the proposed method with a GAN-based reconstruction, trained on random wave fields. The GAN method, which necessitates an iterative approach for pressure reconstruction, can be cumbersome when dealing with wideband signals. This limitation underscores the attractiveness of the PINN's time domain formulation. The results demonstrate that while both methods perform comparably at low frequencies, with the GAN-based approach even outperforming on the PINN in some cases (i.e., for NMSE$_{\text{VAL}}$ $\tilde{\mathcal{M}}_{50\,\%}$ at 200 Hz). However, the GAN exhibits performance degradation at higher frequencies. This is likely due to the low-frequency bias inherent in neural networks with ReLU activation functions [57] requiring more specialized architectures to overcome this limitation.

4.5 Intensity field computation

In addition to describing the pressure field, the PINN provides a thorough characterization of the sound field through the reconstruction of intensity flows, governed by Eq. (8) and illustrated in Fig. 8 Here, the intensity is projected on the three facets of a cuboid volume and the trajectories of air particles within the room are delineated by streamlines, with directional arrows indicating their path. Additionally, the contours convey the magnitude of the reconstructed intensity in watts/m^2.

At time instant $t = 1.612$ s (Fig. 8a), the snapshot reveals heightened intensity emanating from the speaker's direction $\vec{r} = (0.15, 0.15, 0.)$ m, along with lateral room reflections (approximately $\vec{r} = (-0.15, 0.1, 0.08)$ m and $\vec{r} = (-0.15, 0, -0.05)$ m). Predominantly, energy is directed from the speaker to the origin and from the rear room corner $\vec{r} = (-0.15, -0.15, -0.15)$ m.

The subsequent snapshot at time $t = 1.774$ s (Fig. 8b) illustrates the phenomenon of destructive interference between oppositely propagating wave fronts. Given that the intensity vector is normal to the propagation direction, energy convergence towards zero along the y axis is evident. This observation is further reinforced by the periodic nature of the signal (spoken voice), indicating wavefronts where pressure and velocity are in phase.

Finally, at time $t = 1.94$ s (Fig. 8c), substantial lateral energy is observed from all directions, corresponding to reflections propagating in all directions, contributing to a diffuse-like energy distribution.

5 Conclusions

In this manuscript, we proposed an approach for the volumetric reconstruction of speech signals with a PINN framework. We proved the devised architecture to learn an implicit representation of the acoustic field

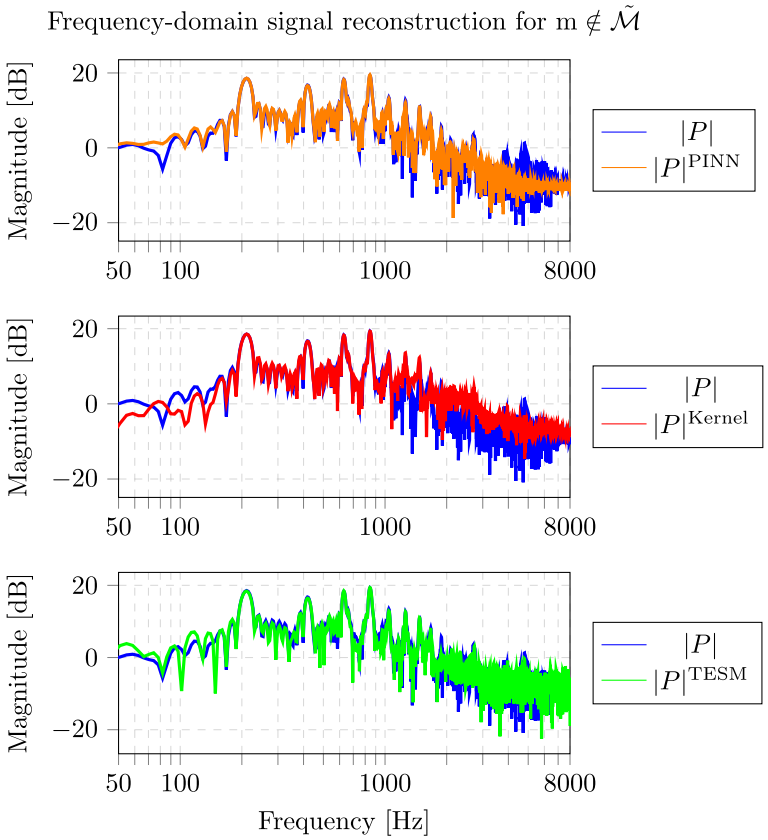

Fig. 6 Example of reconstruction comparison of an unknown signal in frequency domain. From above to the bottom, the magnitude estimate of the proposed method $|P|^{\text{PINN}}$ and the reconstructions obtained with the kernel method $|P|^{\text{Kernel}}$ and TESM $|P|^{\text{TESM}}$. Each diagram is depicted with the ground truth $|P|$ in dB scale

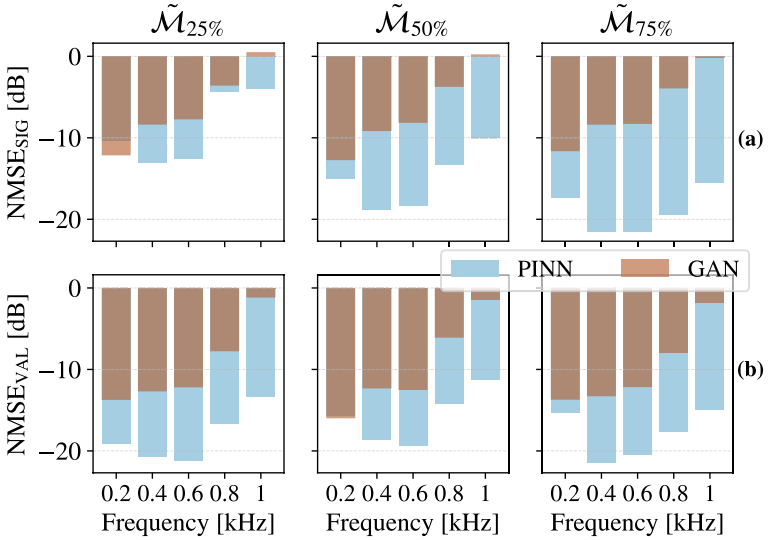

Fig. 7 NMSE of narrowband sound fields between the proposed PINN approach and GAN approach [39] with **a** the metric considering the unknown pressure signals and **b** the metrics relative to the available observations

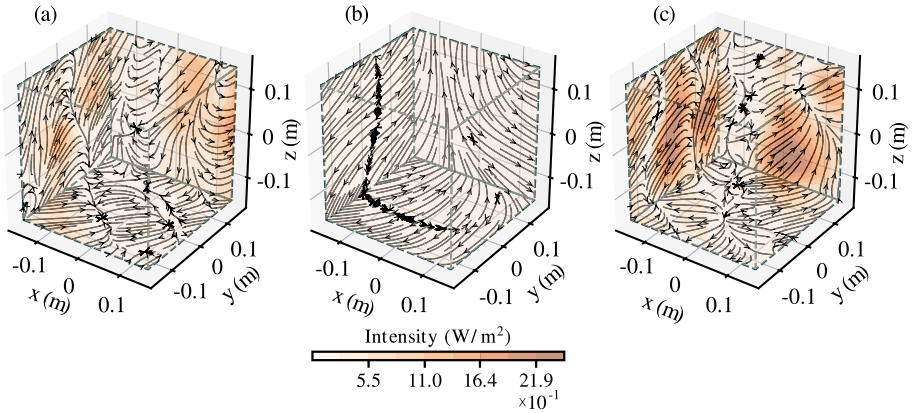

Fig. 8 Snapshots of the instantaneous intensity vector plotted on the three rear facets of a cuboid sized 0.15^3 m^3 at time **a** $t = 1.612$ s, **b** $t = 1.774$ s, and **c** $t = 1.94$ s

in the target 3D region from a few and sparse microphone observations and regularizing the estimates to follow the physical prior knowledge of the acoustic propagation, i.e., the wave equation.

Estimating the time domain signals of the pressure field, the devised method is able to recover arbitrary acoustic fields, with no constraints on the geometry and acoustic conditions of the environments. We validated the proposed approach considering a real dataset and considering different subsets of available observations. The performance has been evaluated in terms of normalized mean square error between the sound pressure ground truth and the reconstructions. Moreover, we compared the results with respect to two state-of-the-art methods for the reconstruction of sound field in time domain and frequency domain.

We proved that our solution outperformed both the baselines preserving the frequency contents of the acoustic field. Furthermore, to show the potentiality of the proposed PINN method, we presented an example of intensity field retrieved from the network estimates, by exploiting the automatic differentiation principle.

The results prove how the combination of data-driven approach with the regularization provided by the physical propagation prior of acoustic fields can increase the accuracy of state-of-the-art models, both in time domain and in frequency domain. Nevertheless, we believe the potentiality of such PINN approach in the context of sound field reconstruction can be further exploited for the processing of large 3D region, such as in rooms or concert hall. Moreover, from this preliminary study, we plan to extend the methodology considering arbitrary sound fields generated from different acoustic sources, e.g., music or noise signals, in different acoustic environments.

Acknowledgements
This work was made possible through the support of various entities. Firstly, the European Union contributed under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU partnership regarding "Telecommunications of the Future" (PE00000001 - program "RESTART"). Additionally, the "REPERTORIUM project" provided support with grant agreement number 101095065 under Horizon Europe. Cluster II. Culture, Creativity and Inclusive Society, call HORIZON-CL2-2022-HERITAGE-01-02. Lastly, the VILLUM Foundation supported this work under grant number 19179 for the project titled "Large-scale acoustic holography."

Authors' contributions
Conceptualization, M.O. and M.P.; methodology, M.O., X.K, and M.P.; software, M.O. and X.K.; validation, M.O. and X.K.; formal analysis, M.O., X.K., and M.P.; investigation, M.O. and X.K.; data curation, M.O. and X.K.; writing—original draft preparation, M.O., X.K. and M.P.; writing—review and editing, M.O., X.K., M.P., F.A., A.S. and E.F.; visualization, M.O. and X.K.; supervision, F.A., A.S. and E.F.; project administration, F.A., A.S. and E.F.; funding acquisition, F.A., A.S. and E.F. All authors have read and agreed to the published version of the manuscript.

Availability of data and materials
This manuscript has no associated data.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 30 January 2024 Accepted: 4 August 2024
Published online: 09 September 2024

References

1. M. Vorländer, D. Schröder, S. Pelzer, F. Wefers, Virtual reality for architectural acoustics. J. Build. Perform. Simul. **8**(1), 15–25 (2015)
2. M. Tohyama, T. Koike, J.F. Bartram. Fundamentals of acoustic signal processing (2000)
3. M. Pezzoli, J.J. Carabias-Orti, M. Cobos, F. Antonacci, A. Sarti, Ray-space-based multichannel nonnegative matrix factorization for audio source separation. IEEE Signal Process. Lett. **28**, 369–373 (2021)
4. M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, B. Lee et al., A survey of sound source localization methods in wireless acoustic sensor networks. Wirel. Commun. Mob. Comput. **2017** (2017)
5. S. Koyama, L. Daudet, Sparse representation of a spatial sound field in a reverberant environment. IEEE J. Sel. Top. Signal Process. **13**(1), 172–184 (2019)

6. N. Ueno, S. Koyama, H. Saruwatari, in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. Kernel ridge regression with constraint of Helmholtz equation for sound field interpolation (IEEE, 2018), pp. 436–440
7. N. Ueno, S. Koyama, H. Saruwatari, Sound field recording using distributed microphones based on harmonic analysis of infinite order. IEEE Signal Process. Lett. **25**(1), 135–139 (2017)
8. M. Pezzoli, M. Cobos, F. Antonacci, A. Sarti, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Sparsity-based sound field separation in the spherical harmonics domain (IEEE, 2022), pp. 1051–1055
9. M. Pezzoli, F. Borra, F. Antonacci, A. Sarti, S. Tubaro, in *2018 26th European Signal Processing Conference (EUSIPCO)*. Reconstruction of the virtual microphone signal based on the distributed ray space transform (IEEE, 2018), pp. 1537–1541
10. V. Pulkki, S. Delikaris-Manias, A. Politis, *Parametric time-frequency domain spatial audio* (Wiley, 2018)
11. G. Del Galdo, O. Thiergart, T. Weller, E. Habets, in *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays*. Generating virtual microphone signals using geometrical information gathered by distributed arrays (IEEE, 2011), pp. 185–190
12. F. Lluis, P. Martinez-Nuevo, M. Bo Møller, S. Ewan Shepstone, Sound field reconstruction in rooms: Inpainting meets super-resolution. J. Acoust. Soc. Am. **148**(2), 649–659 (2020)
13. M.S. Kristoffersen, M.B. Møller, P. Martínez-Nuevo, J. Østergaard, Deep sound field reconstruction in real rooms: Introducing the isobel sound field dataset. (2021). arXiv preprint arXiv:2102.06455
14. M. Pezzoli, D. Perini, A. Bernardini, F. Borra, F. Antonacci, A. Sarti, Deep prior approach for room impulse response reconstruction. Sensors **22**(7), 2710 (2022)
15. D.L. Donoho, Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
16. S. Lee, The use of equivalent source method in computational acoustics. J. Comput. Acoust. **25**(01), 1630001 (2017)
17. N. Antonello, E. De Sena, M. Moonen, P.A. Naylor, T. Van Waterschoot, Room impulse response interpolation using a sparse spatio-temporal representation of the sound field. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(10), 1929–1941 (2017)
18. D. Caviedes-Nozal, E. Fernandez-Grande, Spatio-temporal Bayesian regression for room impulse response reconstruction with spherical waves. IEEE/ACM Trans. Audio Speech Lang. Process. **31**, 3263–3277 (2023)
19. M. Hahmann, S.A. Verburg, E. Fernandez-Grande, Spatial reconstruction of sound fields using local and data-driven functions. J. Acoust. Soc. Am. **150**(6), 4417–4428 (2021)
20. V. Vovk, *Kernel Ridge Regression* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013), pp.105–116
21. A. Figueroa-Durán, E. Fernandez-Grande, in *10th Convention of the European Acoustics Association*. Room impulse response reconstruction from distributed microphone arrays using kernel ridge regression (European Acoustics Association, 2023)
22. J.G. Ribeiro, N. Ueno, S. Koyama, H. Saruwatari, Region-to-region kernel interpolation of acoustic transfer functions constrained by physical properties. IEEE/ACM Trans. Audio Speech Lang. Process. **30**, 2944–2954 (2022)
23. J.G. Ribeiro, N. Ueno, S. Koyama, H. Saruwatari, in *2020 IEEE 11th Sensor Array and Multichannel Signal Processing Workshop (SAM)*. Kernel interpolation of acoustic transfer function between regions considering reciprocity (IEEE, 2020), pp. 1–5
24. J.G. Ribeiro, S. Koyama, H. Saruwatari, in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Kernel interpolation of acoustic transfer functions with adaptive kernel for directed and residual reverberations (IEEE, 2023), pp. 1–5
25. D. Caviedes-Nozal, N.A. Riis, F.M. Heuchel, J. Brunskog, P. Gerstoft, E. Fernandez-Grande, Gaussian processes for sound field reconstruction. J. Acoust. Soc. Am. **149**(2), 1107–1119 (2021)
26. L. McCormack, A. Politis, R. Gonzalez, T. Lokki, V. Pulkki, Parametric ambisonic encoding of arbitrary microphone arrays. IEEE/ACM Trans. Audio Speech Lang. Process. **30**, 2062–2075 (2022). https://doi.org/10.1109/TASLP.2022.3182857
27. M. Pezzoli, F. Borra, F. Antonacci, S. Tubaro, A. Sarti, A parametric approach to virtual miking for sources of arbitrary directivity. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 2333–2348 (2020)
28. O. Thiergart, G. Del Galdo, M. Taseska, E.A. Habets, Geometry-based spatial sound acquisition using distributed microphone arrays. IEEE Trans. Audio Speech Lang. Process. **21**(12), 2583–2594 (2013)
29. S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, A consolidated perspective on multimicrophone speech enhancement and source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(4), 692–730 (2017)
30. R. Mignot, G. Chardon, L. Daudet, Low frequency interpolation of room impulse responses using compressed sensing. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(1), 205–216 (2013)
31. W. Jin, W.B. Kleijn, Theory and design of multizone soundfield reproduction using sparse methods. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(12), 2343–2355 (2015)
32. M. Olivieri, M. Pezzoli, F. Antonacci, A. Sarti, A physics-informed neural network approach for nearfield acoustic holography. Sensors **21**(23), 7834 (2021)
33. M. Olivieri, R. Malvermi, M. Pezzoli, M. Zanoni, S. Gonzalez, F. Antonacci, A. Sarti, Audio information retrieval and musical acoustics. IEEE Instrum. Meas. Mag. **24**(7), 10–20 (2021)
34. M. Olivieri, L. Comanducci, M. Pezzoli, D. Balsarri, L. Menescardi, M. Buccoli, S. Pecorino, A. Grosso, F. Antonacci, A. Sarti, in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Real-time multichannel speech separation and enhancement using a beamspace-domain-based lightweight cnn (IEEE, 2023), pp. 1–5
35. X. Karakonstantis, E. Fernandez Grande, in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*. Sound field reconstruction in rooms with deep generative models, vol. 263 (Institute of Noise Control Engineering, 2021), pp. 1527–1538
36. E. Zea, Compressed sensing of impulse responses in rooms of unknown properties and contents. J. Sound Vib. **459**, 114871 (2019)
37. M.J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M.A. Roch, S. Gannot, C.A. Deledalle, Machine learning in acoustics: Theory and applications. J. Acoust. Soc. Am. **146**(5), 3590–3628 (2019)
38. E. Fernandez-Grande, X. Karakonstantis, D. Caviedes-Nozal, P. Gerstoft, Generative models for sound field reconstruction. J. Acoust. Soc. Am. **153**(2), 1179–1190 (2023)
39. X. Karakonstantis, E. Fernandez-Grande, Generative adversarial networks with physical sound field priors. J. Acoust. Soc. Am. **154**(2), 1226–1238 (2023)
40. F. Miotello, L. Comanducci, M. Pezzoli, A. Bernardini, F. Antonacci, A. Sarti, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Reconstruction of sound field through diffusion models (IEEE, 2024), pp. 1476–1480
41. D. Ulyanov, A. Vedaldi, V. Lempitsky, in *Proceedings of the IEEE conference on computer vision and pattern recognition*. Deep image prior (2018), pp. 9446–9454
42. K. Shigemi, S. Koyama, T. Nakamura, H. Saruwatari, in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. Physics-informed convolutional neural network with bicubic spline interpolation for sound field estimation (IEEE, 2022), pp. 1–5
43. E.G. Williams, *Fourier acoustics: Sound radiation and nearfield acoustical holography* (Academic Press, London, 1999)
44. A.G. Baydin, B.A. Pearlmutter, A.A. Radul, J.M. Siskind, Automatic differentiation in machine learning: A survey. J. Mach. Learn. Res. **18**, 1–43 (2018)
45. M. Raissi, P. Perdikaris, 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. **378**, 686–707 (2019)
46. M. Pezzoli, F. Antonacci, A. Sarti, Implicit neural representation with physics-informed neural networks for the reconstruction of the early part of room impulse responses. Forum Acusticum 2023. 2177–2184 (2023)
47. X. Karakonstantis, E. Fernandez-Grande, Room impulse response reconstuction using physics-constrained neural networks. Forum Acusticum 2023. 3181–3188 (2023)
48. X. Karakonstantis, D. Caviedes-Nozal, A. Richard, E. Fernandez-Grande, Room impulse response reconstruction with physics-informed deep learning. J. Acoust. Soc. Am. **155** (2): 1048–1059 (2024)

49. V. Sitzmann, J. Martel, A. Bergman, D. Lindell, G. Wetzstein, Implicit neural representations with periodic activation functions. Adv. Neural Inf. Process. Syst. **33**, 7462–7473 (2020)
50. S. Damiano, F. Borra, A. Bernardini, F. Antonacci, A. Sarti, in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Soundfield reconstruction in reverberant rooms based on compressive sensing and image-source models of early reflections (IEEE, 2021), pp. 366–370
51. E. Fernandez-Grande, D. Caviedes-Nozal, M. Hahmann, X. Karakonstantis, S.A. Verburg, in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. Reconstruction of room impulse responses over extended domains for navigable sound field reproduction (IEEE, 2021), pp. 1–8
52. S. Koyama, T. Nishida, K. Kimura, T. Abe, N. Ueno, J. Brunnström, in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. MeshRIR: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods (IEEE, 2021), pp. 1–5
53. G. Stan, J. Embrechts, D. Archambeau, Comparison of different impulse response measurement techniques. J. Audio Eng. Soc. **50**(4), 249–262 (2002)
54. A. Farina, in *Audio engineering society convention 122*. Advancements in impulse response measurements by sine sweeps (Audio Engineering Society, 2007)
55. K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989)
56. D.P. Kingma, J. Ba, in *ICLR (Poster)*. Adam: A method for stochastic optimization (2015)
57. N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, A. Courville, in *International conference on machine learning*. On the spectral bias of neural networks (PMLR, 2019), pp. 5301–5310

Publisher's Note