Novel View Synthesis with Physically Consistent Machine Learning for Augmented Listening

Diego Di Carlo, *Member, IEEE,* Shoichi Koyama, *Member, IEEE,* Aditya Arie Nugraha, *Member, IEEE,*
Mathieu Fontaine, *Member, IEEE,* Yoshiaki Bando, *Member, IEEE,* Kazuyoshi Yoshii, *Senior Member, IEEE.*

***Abstract*—**

***Index Terms*—Augmented listening, head-related transfer function (HRTF), Gaussian process, physics-informed neural networks (PINN), spatial audio, array manifold.**

I. Introduction

AUGMENTED listening (AL) is a recent umbrella term for sound field manipulation technologies that enhance human listening abilities by modifying the sounds they hear in real time [1]. Applications of augmented listening often involve creating *personalized sound zones* by manipulating sound fields through addition, removal, or modification of sources tailored to individual preferences [2]. Besides the real-time performances, it focuses on audio analysis (e.g., localization and separation) and synthesis (e.g., rendering) problems at the intersection of acoustics, signal and array processing, machine learning (ML), and human-machine interaction [3].

AL technologies, such as hearing aids and smart headphones, aim to provide users with a clearer and more personalized auditory experience, particularly in challenging environments with multiple sound sources. The emergence of extended reality (XR) technologies, which encompass augmented reality (AR), virtual reality (VR), and mixed reality (MR), has further amplified the demand for sophisticated audio processing methods. XR devices, including smart glasses like Hololens2 and Meta Quest 3, offer immersive user experiences by blending digital and physical soundscapes [4]. However, while image and video processing techniques are well-advanced, achieving a comparable sense of audio immersion remains challenging [5], [6], but also for improving accessibility [7].

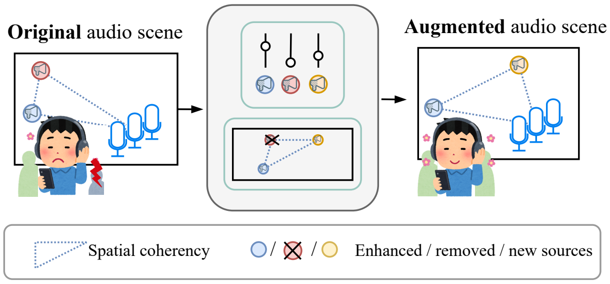
Original audio scene
Augmented audio scene
Spatial coherency
○ / ☒ / ○ Enhanced / removed / new sources

Fig. 1: Augmented listening system "remix" the sound we perceive. Semantic and spatial information of the audio scene can be modified, but the spatial content must remain physical to convey realism.

The SPeech Enhancement for Augmented Reality (SPEAR) Challenge [8], endorsed by the IEEE, highlighted challenges in augmented listening using recordings from a head-worn microphone array in a restaurant-like setting. The goal was to create binaural audio signals based on head orientation and the direction of a target speaker. An end-to-end learning-based method was the top performer in both objective and subjective evaluations, while the isotropic Minimum Variance Distortionless Response (MVDR) beamformer [9, Sec 2.1] ranked second in subjective assessments. The authors of the challenge impute this in the trade-off between interference suppression and speech distortion of most algorithms. However, it underscores the importance and effectiveness of accurate acoustics front-end processing, which boils down to the correct modeling of beamformers' steering vectors.

Acoustics plays a central role since signals of interest are subject to *wave propagation* in space and time. Sound is regarded as a *field*, a continuous function of space and time, whose main interest is its reconstruction at arbitrary points (See [10] for a recent review). This *regression* problem is traditionally solved by finding a linear combination of physics-based basis functions (e.g., spherical harmonics) that satisfy the wave equation by construction. Current advancements use ML to improve convergence and stability in case of ill-posedness or ill-conditionedness due to data scarcity or noise. Notable is the use of Physics-Informed Neural Networks (PINNs) [11] to solve nonlinear regression using the *partial differential equation* (PDE) of the wave equation as regularization [12], [13]. Other successful methods use data-driven methods to retain prior information from huge datasets, typically generated by

Manuscript received xxx yyy, 2024; revised xxx yyy, 2024; accepted xxx yyy, 2024. Date of publication xxx yyy, 2024; date of current version xxx yyy, 2024. This work was supported by JST PRESTO no. JPMJPR20CB, JSPS KAKENHI nos. JP23K16912, JP23K16913, and JP24H00742, ANR Project SAROUMANE (ANR-22-CE23-0011), and Hi! Paris Project MASTER-AI. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. xxx. *(Corresponding author: Diego Di Carlo).*

Diego Di Carlo and Aditya Arie Nugraha are with the Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo 103-0027, Japan (e-mail: diego.dicarlo@riken.jp; adityaarie.nugraha@riken.jp)

Shoichi Koyama is with National Institute of Informatics, Tokyo, Japan (e-mail: XXXXXXXXXXX).

Mathieu Fontaine is with LTCI, Télécom Paris, Institut Polytechnique de Paris, France (e-mail: mathieu.fontaine@telecom-paris.fr).

Yoshiaki Bando is with the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, 135-0064, Japan (e-mail: y.bando@aist.go.jp).

Kazuyoshi Yoshii is with the Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo 103-0027, Japan, and XXXXXXXXX (e-mail: yoshii@i.kyoto-u.ac.jp).

Digital Object Identifier xxx

Source code available https:github.comChutlhuNeuralSteerer physics simulators (virtually supervised learning). State-of-the-art (SOTA) ML works produce physical solutions with increasing levels of accuracy, but they lack two aspects. First, they are good in the "mean-squared error" sense, which may not be optimized for downstream tasks nor reflect the user's preferences [8]. Second, they lack uncertainty quantification, limiting their reliability in downstream tasks.

In signal processing applications, sounds are modeled as signals being modified by filters. The former models *semantic content* (e.g., speech, melodies, etc.), while the latter encodes the propagation effect and *spatial* information (e.g., source and sensor locations, room properties). Audio processing techniques offer practical frameworks to tackle several ill-posed inverse problems for audio analysis and synthesis. Standard techniques, e.g., multi-signal classification (MUSIC) and multichannel Wiener filtering, are still commonly used for fast and reliable source localization and separation, respectively [14]. Lately, ML and statistical optimization advances have revolutionized this field. While end-to-end SOTA ML methods like [15] can effectively separate concurrent speakers in challenging acoustic scenarios, computational complexity, and real-time requirements prevent the deployment of these models on lightweight devices intended for augmented listening. A current trend is to use lightweight hybrid algorithms as *front-end acoustic* processing coupled with more intensive optimization in the back-end [16]. Here, we focus on the acoustic models for front-end processing.

Besides, array processing approaches leverage the spatial diversity of multichannel audio setups and wave propagation properties to enhance signals of interest. Still, SOTA methods rely on a simplified geometrical approximation of the sound propagation, typically assuming anechoic conditions [8]. The challenge is to estimate the full propagation filters, that is, the *Acoustic Impulse Responses (AIR)*, commonly regarded as a discrete *blind channel estimation* problem. Due to the huge variability of possible acoustic conditions, the diversity of the array's geometry configuration, and the precision of the solution so as not to affect downstream tasks, current ML and deep learning methods struggle in practical situations. Novel physics-informed methods developed for sound field reconstruction offer attractive directions [10].

A. Steering vectors

Steering vectors represent the interaction of sound waves with microphone arrays [17]. They are foundational in spatial audio processing, enabling tasks such as speech enhancement [14], sound source localization [18], and sound scene synthesis [19]. While traditionally modeled as algebraic formulations in free-field scenarios, practical applications often require handling complex acoustic environments, accounting for generic acoustic propagation effects as for *acoustic impulse responses* (AIRs), indoor reverberation as for *room impulse responses* (RIR), or the anatomical effects as in *head-related transfer functions* (HRTFs).

In this work, we consider steering vectors as multi-channel mathematical quantities encoding the anechoic sound propagation impinging on a set of microphones from a location to a reference point at a given frequency. This is called *array manifold* in antenna array processing [20]. Specifically, we assume a multi-channel spatiotemporal representation of the sound field, expressed as a collection of AIRs measured on an aperture that constitutes the spatial domain of interest, for instance, the sphere surrounding a listener's head. This representation is most commonly adopted in sound field synthesis and Ambisonics reproduction, where the listener and the valid sound field are delimited by a space enclosed by a loudspeaker array with far-field sources contributing to the overall reconstructed sound pressure [21]. Note that this definition allows us to consider HRTF and multi-channel measurements as steering vectors.

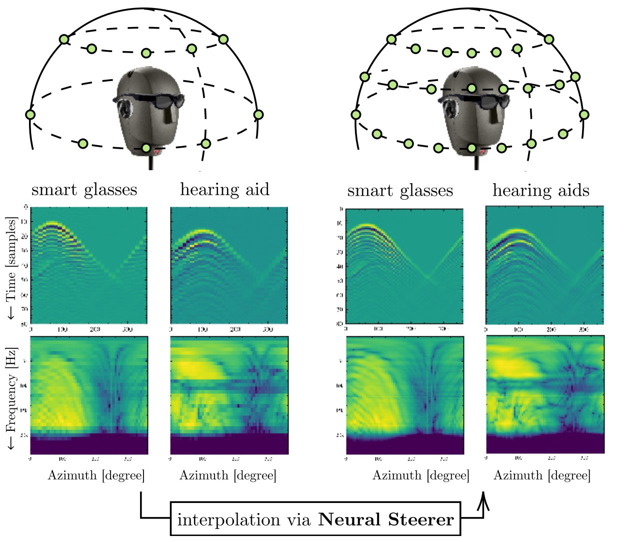

Fig. 2: Visual abstract of continuous steering vector modeling

Steering vectors can be broadly categorized based on the acoustic conditions they encode. In anechoic free-field environments, steering vectors are derived from closed-form expressions knowing the array geometry [17] or relative time-differences of arrival (TDOA) in far-field scenarios [22]. This formulation offer a simple, fast and differentiable computation of the steering vectors, a geometric interpretation, valid in every environment. However it does not model acoustics reflection, which are regarded as noise term in further downstream tasks.

For scenarios involving non-free-field anechoic environments, steering vectors account for additional factors such as scattering objects and microphone directivity (a.k.a., pick-up) patterns. A prominent example is HRTF, which captures the spatial filtering effects caused by human anatomical features, such as pinnae, head and torso. Such vectors are typically estimated through either in-lab measurements or computationally-heavy simulators.

Beyond anechoic conditions, general AIR-based steering vectors represent complex propagation environment that can be computed using model based on numerical acoustic simulators or fast simulators integrating geometric acoustics, stochastic approaches or hybrid methods [23]. While these representations provide precise spatial information, their applicability is often constrained to environments where acoustic properties are well-characterized.

The methods to obtain steering vectors depend on the above condition and available resources. Algebraic approaches and some simulators offer computational efficiency, especially in simple anechoic environments, but require detailed knowledge of the array geometry and environmental parameters. Such methods, while fast and differentiable, may struggle to replicate real-world acoustic profiles, being sensitive to microphone positioning and speed of sound variations [17, Chapter 6.6]. Estimation-based methods treat steering vector acquisition as an inverse problem, utilizing tools like blind source separation [**?**] or relative transfer function (RTF) estimation [14, Section VI.B.3]. Despite progress in this area, robust estimation remains a significant challenge in audio signal processing [24], [25].

Alternatively, steering vectors can be directly measured, which provides the most accurate representation of the actual acoustic environment [19]. However, such measurements are expensive, time-consuming, require expert oversight, and are highly dependent on spatial sampling resolution [26]. Additionally, measured vectors are often stored in lookup tables, making them susceptible to dataset-specific conventions and preprocessing requirements [27].

To reduce the time required and the complexity of measured steering vector setup and to make the method scalable, spatial upsampling has been proposed to generate high-resolution steering vectors. These methods are commonly referred to as upsampling, interpolation, regression, or super-resolution methods, especially in the field of HRTF synthesis and sound field reconstruction. Several methods have been proposed to spatially upsampling measurements improving the spatial resolution of arrays with fewer sensors or offering a cost-effective solution without compromising performance. Considering the steering vectors and this is basically a curve fitting problem.

In this work, we propose a novel method for upsampling measured steering vectors from limited observations, formulated as a Gaussian Process (GP) regression problem. By modeling the relationship between spatial-frequency coordinates and multichannel steering vector measurements, our approach ensures a continuous representation across the channel dimension. This method leverages strong physical priors, striking a balance between soft and hard physics-based constraints to effectively address data scarcity. Notably, our model simultaneously accounts for both the magnitude and phase of steering vectors, ensuring a comprehensive and accurate representation. We further perform an ablation study to assess the contributions of various components within the architecture, providing insights into the design choices that drive its performance.

The evaluation focuses on the challenging task of upsampling steering vectors from sparse measurements, addressing scenarios encountered in augmented listening pipelines. We present a thorough analysis of the proposed method, benchmarking it against state-of-the-art approaches, including Physics-Informed Neural Networks (PINNs), across frequency and spatial dimensions. Our results demonstrate the model's ability to achieve high fidelity in both angular and frequency domains, significantly outperforming baseline methods. Additionally, we extend the evaluation to downstream tasks such as beamforming using real-world data, showcasing the method's utility in end-to-end augmented listening pipelines. These contributions position our approach as a robust solution for steering vector upsampling in both theoretical and applied contexts.

B. Contributions

In some applications, measuring the trade-off between the observed data and the interpolation is preferable. Methods like linear and spline interpolation allow for perfect recall of the "training" data.

Motivation

- Standard approaches do not work with sparse measurements or arbitrary microphone array geometry (the are facing this in the case of steering vectors).
- In our case, the data-fit and the model-based terms do not share the same local minima.
- PINNs only minimize the PDE; the output is not guaranteed to be physically consistent.
- PINNs do not preserve the nature of the observed data
- GP regression is a versatile framework for interpolation that benefits of several properties:
  - lets the user design the data properties through the covariance function
  - solution are "not far" for the RHKS solution. Need citation
  - It is a generative approach; it estimates density; it models uncertainties that can be useful in downstream tasks.
  - vs. PINN: observed points can be "preserved". I need to prove this.
  - vs. PINN: less hyper-params to tune
  - vs. standard methods: expressive, non-orthogonal (can work on any grid), network modularity for later works.
- Downstream tasks
  - Sound field reconstruction
  - RIR interpolation
  - Novel-view audio synthesis

The rest of the paper is organized as follows. Section II presents the state of the art in sound field reconstruction and spatial upsampling. Section III formulate the signal models and the problem of interest. Section V presents the proposed models of continuous steering vectors. Section VI reports some implementation details and presents a series of experiments conducted with the proposed models and their comparison with several state-of-the-art spatial upsampling methods. This includes analysis for sound field reconstruction and the downstream task in speech enhancement. Section VII concludes the paper.

II. Related works

Various techniques have been developed in HRTF upsampling and sound field reconstruction to upsample acoustic measurements around the human head. HRTF upsampling methods typically focus on the magnitude of a minimum-phase representation of the Head-Related Transfer Function (HRTF), while the phase is compensated using analytical steering vectors encoding the interaural level and phase differences, that is, the anechoic propagation. These methods often assume that the contributions from both ears are equal or independent. In contrast, sound field reconstruction methods take a more general approach and rely on precise physical models to model the

		Approaches	Problems related to steering vector interpolation					
			User-centred sound field		Generic sound field			
			HRTF upsampling	Source directivity	SFR	HOA mic	RIR interpolation	Array manifold
		Review papers	[28], [29]		[10], [30]			
Data-driven		Weighted interpolation	[31]–[37]					
		Subspace methods	[38]–[45]					
		Deep learning	[?], [46]–[53]		[30], [54]–[57]			
		. . . with Neural Fields	[27], [47], [58]					
		Manifold Learning	[59]–[61]					
Knowledge-driven		Geometric-based	[?], [62]–[69]		[70]–[73]			
		Parametric (DSP)-based	[?], [74]–[80]	[72]	[81], [82]			
	Physics-based	Physics-constrained	[83]–[91]		[?], [92]–[101]			
		. . . with DNN	[102]		[21], [97], [103]			
		. . . with GP	[104]		[105]–[107]			
		Physics-informed	[12], [108]		[?], [108]–[114]	[115]		

TABLE I: Schematic organization of the literature in acoustic steering vector upsampling.

sound field. They also leverage specific array geometries, such as spherical microphone arrays, or utilize multiple microphones to achieve precise reconstruction at high spatial resolution.

A. HRTF upsampling

Spatial upsampling of head-related transfer functions (HRTFs) measured on a sparse grid is an important issue, particularly relevant when capturing individual datasets. While early studies mostly used nearest-neighbor approaches, ongoing research focuses on interpolation in the spherical harmonics (SH) domain. The interpolation can either be performed on the complex spectrum or separately on magnitude and unwrapped phase. Furthermore, preprocessing methods can be applied to reduce the spatial complexity of the HRTF dataset before interpolation. We compare different methods for the interpolation of HRTFs and show that SH and nearest-neighbor based approaches perform comparably. While generally a separate interpolation of magnitude and unwrapped phase outperforms an interpolation of the complex spectra, this can be compensated by appropriate preprocessing methods.

Authors of [29] recently reviewed the state-of-the-art techniques for HRTF measurement interpolation. In such work, the rich literature is categorized into 3 classes: nearest-neighbor approaches that return a weighted combination of neighboring measurements, functional methods that model a mathematical function of frequency and direction, and methods based on neural networks. Here we propose a different classification based on how much explicit physical knowledge is employed to super-resolve HRTF measurements: data-driven and knowledge-driven approaches, the latter being subdivided into physics-driven and approximated parametric methods. HRTF representation also influences interpolation performances: the results of [31] indicated that a separate interpolation of the magnitude and unwrapped phase of the HRTF performed better than an interpolation of the complex spectrum. This fact, also confirmed in [28], applied for the task of virtual acoustic environments, not for other downstream tasks, such as beamforming, as investigated in this work.

a) Data-driven methods: These methods rely only on information contained in the observations, in prepared training sets or provided from data coming from multi-modalities. Data-driven methods could be local measurements (linear, bilinear, trilinear, barycentric) interpolation or exploit data-driven knowledge from a global set of measurements using methods like PCA, wavelets, or DNN.

Weighted interpolation is the most straightforward approach. This method typically assumes noiseless measurements acquired on a regular grid. These methods have been shown to produce a sufficiently good agreement between measured and interpolated HRTFs when a relatively large number of measurements are still present [33]. In the case where the low-resolution HRTF contains 320 or more source positions, it is preferable to use barycentric interpolation [116]. As explained in [116], the acoustic measurement [117], [118] is still considered the gold standard of these different approaches. The downside to performing acoustic measurements is the expensive custom setup required and the time it takes. This method has been shown to produce good results when the HRTFs contain a relatively large number of IRs [33], for example, with an angular distance of 10–15° between measurements; however, it becomes much less reliable when interpolating sparser measurements (e.g., each 30–40°) [116]. Methods in this category are uses inverse distance [31], bilinear interpolation [32], [41], tetrahedral interpolation [33], barycentric interpolation in spherical interpolation [34] or natural neighbor interpolation [28]. The authors of [37] extends the this methods to non-uniform grid. This methods have been recently evaluated with real data measured on a Kemar mannequin in [29], [35]. According to [29] the best performing methods is the bilinear interpolation to upsample ... not reported??.

Subspace methods aims to reduce the dimensionality of a data set while retaining the primary variation in the data, rather than improve interpolation in case of sparse measurements [28]. In this way, fewer coefficients are interpolated instead of the HRTF itself. Principal Components Analysis (PCA) is a statistical algorithm for deriving spectral shape basis functions and decomposing HRTFs (see [119, Chapter 6.2] for a review). Authors of [38] proved that PCA of the resulting 5300 HRTF magnitude functions revealed that the HRTFs can be modeled as a linear combination of five basic spectral shapes (basis functions), and that this representation accounts for approximately 90% of the variance in the original HRTF magnitude functions. Later works such as [39], [40] that interpolated via bilinear interpolation or multivariate polynomial filtering [42]. Authors of [43] introduce the Spatial PCA which was later used by [44] in conjunction with a deep neural network that processes anthropometric features. The work of [39] tackles the problem in continuous settings proposing a model based on the Karhunen-Loéve transform, while in [45] proposed the interpolation of wavelet coefficients.

Deep Neural Network have been extensively used for HRTF upsampling. This approach is appealing because of good generalization, multi-modal learning and quick inference. A recent review is provided in [29] and [3]. In contrast with local methods, DNN are able to extract both global and local properties (features) that can be used for a fast inference. In the HRTF literature, we can identify the following approaches: DNN architectures, like UNet, inspired from the one used in image super-resolution or inpainting, usually change 2D-CNN layer with TCN. [46], [47], [53], [116]; or architecture that uses other modalities, such as anthropometric features [48], [49], or image of the pinnae [50], [51] with autoencoders models. Finally, recent trends that focus on upsampling sparse measurements utilize Generative Adversarial Networks for spatial upsampling of HRTF [52], [53], [116]. Another approach exploits the natural interpolation effect due to the spectral bias of coordinate-based neural networks [120]. These architectures, called *neural fields*, have been recently proposed in computer vision to model physical *fields*, such as radiance field for novel view synthesis of 3D objects (e.g., in NeRF) and physical quantities (e.g., PINNs) [121]. Such approach has been used also for HRTF interpolation [49], implicit auralization of audio signal [47] and different HRTF datasets with different grid conventions across different subjects [27].

b) Knowledge-driven methods: Methods of this group leverage prior knowledge to super-resolve measurements. The problem is typically formulated as a regression problem whose solutions are constrained by a parametric model or regularized. Methods in this group can be further classified as geometric-based methods that use geometric reasoning to weights the interpellant coefficients [62]–[69], [122], DSP-based methods that approximate the shape of HRTF's spectrum with simple filters whose parameter space can be smoothly interpolated [58], [74]–[80], [123], and finally, Physics-driven methods using the Helmholtz equation or its free-field parameterized solution, e.g., the Green's function, to regularize [12] or constraint the solution [83]–[91], [102], [104].

Geometric-based methods uses simplified spatial model to achieve fast interpolations, assuming the spatial distribution of the directional HRTF on a sphere centered on the user head. The most popular methods is the Vector-Based And Panning (VBAP) [62] which is used in ISO/IEC MPEG-H 3D Audio standard for reproducing immersive audio coding with multiple loudspeakers. This method creates spatialized sound by distributing the audio signal between loudspeakers in a way that gives the impression of a sound source being at an arbitrary position. By considering HRTF as directional audio data, VBAP can be employed to spatialization HRTF using the known measurement as anchor points. VBAP is considered as a local panning technique, because it only drives a small number of loudspeakers (at most three) close to the target direction, as opposed to global panning techniques such as Ambisonics amplitude pannning, which is based on approximated physical modeling, as explaied below. Authors of [63] proposed an extension of VBAP, recasting it as an l1 optimization problem applied to global panning. Alternative global geometric methods relies of smooth interpolation on a sphere [64], [65] using spherical thin-plate splines [124]. This approach was used by [66] to interpolate the PCA coefficient instead of the actual measurements for fast global interpolation. Interestingly, the authors of [67] propose a shallow neural network extending the RBF basis function to represent spherical data, accepting the quering direction as input and returning the value of HRTF for given frequency. The proposed "von Mises Basis Function" is the natural constraint of periodicity and singularity at the poles. This idea could be considered as a precursor of the neural field approach presented below. Gaussian Process have been also applied to continuously model HRTF over direction and frequencies [68]. In particular the authors proposed a stationary covariance function based on a kernel based on the chordal distance to model the sources and a inverse-quadratic kernel function to model the correlation among frequencies, which model of an exponentially decreasing process in the time-domain. To our best knowledge this is the only work that explicitly consider a continuous model (and smoothness) over frequency. This approach was later extended to aggregate heterogeneous HRTF dataset in [125], a task that is later studied with neural fields in [27]. Finally, in the deep learning community, some studies focused on extending CNN layer to accomodate the spherical geometry of HRTF data [69], [122]. Spherical convulutional layers have been proposed to executes rotation-equivariant feature transforms. Interestingly, [69] place the upsampling neural network along side with Neural Gaussian Process to provides uncertainty estimates on the upsampled regions and such estimates could inform the sequential decision problem of acquiring as few correcting HRTF data points as needed to meet a desired level of HRTF individualization accuracy. While the authors claims promising results, this method in inherently discrete, performing upsampling on a given resolution. While this is not issues in real life application, the problem is that such approaches require annotated training data. Interestingly, this work analyses the performances for very few observation, within 5 to 100 ranges.

DSP-driven methods use signal processing techniques and properties to reduce the complexity of the problem. Most of the works model HRTF as cascade of zeros and/or poles filters [74]–[78] to model the peak and notches of HRTFs spectrum. Noticing smoothness in the variation of such parameters with respect to direction, they apply interpolation over the space of the filter's parameters. Later this techniques was extended to more expressive IIR filters with parametric representation [79], [80]. In particular, the authors of [80] propose a neural field for the prediction of such coefficients, using then a cascade of differentiable IIR filters in the spirit of Differentiable DSP neural architecture [126].

Physic-driven methods models regards the HRTF as the sound field around the human head. Physics-driven methods can be broadly subdivided in two categories: physics-constarained and physics-informed methods. The former methods aims at the reconstruction of the sound fields typically relying on the interpolation of the projection of the measured set onto a linear combination of spatial basis functions, a linear combination of some spherical harmonics or plane waves that satisfy of the Helmholtz equation. Therefor the solution is constrained to be in a specific physcal space. These basis functions represent sound propagation in a homogeneous medium where each sound field component is unknown and obtained as part of an optimization problem. This is commonly brought about as a plane wave expansion, considered an implicit, truncated solution to the homogeneous Helmholtz. A sound field can also be then approximated as a finite sum of spherical harmonics. Estimation of the weighting coefficients of the basis function is achieved by direct integration of the measurements [84] or solving regularized least-square optimization problems [85]–[88] with spherical basis function or using pre/post-processing techniques to further reduce the number of measurements needed for good interpolation with LS fitting by spherical harmoncs [89]–[91]. As the SH basis functions form a spatially continuous set of solutions of the wave equation, an interpolation in the SH domain yields a physically correct and spatially continuous HRTF representation as long as $N \geq \kappa r$ [127], with $\kappa = \omega c$, ω the frequency and c the propagation velocity of sound. Meaning that a minimum of Consequently, the interpolation of sparse HRTF sets, i.e., HRTF sets which are measured with a low spatial resolution results in an incomplete description of the spatial and spectral properties and leads to order-limitation artifacts affecting high-frequency components and binaural cues, Effect of truncation order error in [128]. Sampling scheme of the measurement on the sphere also affects the results. While equiangular, Gaussian, Lebedev, or Fliege schemes yelds to similar results [129], it is know that random spare measurement lead to a poor interpolation performances [90], [130]. Also, it is a specific property of SH interpolation that by transforming the sampled functions to the SH domain, errors or inaccuracies of one measured point on the sphere unavoidable affect the entire SH representation [28]. Therefore, regularization or more advance optimization techniques are required, for instance, the coefficients can be predicted by a neural network [**?**] or optimized using Bayesian variational inference [104], or preprocessing, e.g. time-aliment [89], [131] or directional equalization to compensation linear phase component and diffraction [88], [90], [91]. See [91], [131] for recent reviews about methods aiming at reduciting spatial aliasing and order truncation errors of SH-based approaches..

Finally, the physics-informed methods use physics as data augmentation under the paradigm of *virtually supervised* training or "soft" regularization terms as a bias to drive the solution towards a good balance between data and physics using the paradigm of Physics-informed Neural Network [12].

B. Sound Field Reconstruction

The objective of the sound field reconstruction is to continuously reconstruct the pressure field in a given position within a target region so that it is possible to reproduce sound signals at arbitrary positions [97]. The applications span from VR/AR, speech enhancement, acoustic imaging (or holography), and active noise control [10]. Because of its deep connection with physics and the strict hi-fi requirement, most of the work uses prior physical knowledge. An exception is made for a few pure data-driven DNN-based works based on the recent development in deep learning, e.g. [30], [54]–[57] inspired from models of image super-resolution in computer vision.

The work in this field can be broadly classified into two [111]: non-parametric or expansion-based models [70], [92]–[94], [96], [97], [105], parametric models [73], [81], [82], [98]–[101], [132], [133] and a traversal hybrid approaches based on deep learning [13], [21], [30], [54]–[57], [97], [103], [108]–[115] and Gaussian Processes [105]–[107], [134].

A part of the work in [70], that uses natural neighbor interpolation and geometrical alignments [135], non-parametric approaches aim at estimating the acoustic field by expressing it as a linear combination of solutions of the wave equation [136], e.g., plane wave [94]–[99], [105], spherical harmonics [**?**], [92], [93], acoustics modes of indoor spaces [**?**], [101], [132], or equivalent sources generating spherical waves [**?**], [100]. These functions are used as basis functions to compose dictionaries for regularized linear regression [92]–[95] or covariances for kernel ridge regression [96], [97]. The latter approach has recently been extended to the framework Gaussian Process [105], which combines plane-wave-based representation for modeling directional components and statistically-motivated kernel to model the isotropic distribution of the reverberant field. Besides, GP allows for uncertainty quantification, providing more insight into the reconstruction of the field. The works of [98]–[101] instead recast the problem within the framework of compressed sensing, assuming sparsity of the representation of the sound field.

Parametric models: these models rely on approximate sound scene representations or signal models that result in a sound field perceptually similar to the target one [71]. Then, reconstruction of the sound field is obtained by modulating this physical latent variable. Methods in this category include audio analysis methods based on geometrical acoustics-based method [71], [72], [81] that estimates parameters of the audio scene (location, radiance pattern, and signal) or of the room impulse response [73].

Data-driven models: Following the success of deep neural networks (DNN), several works applied these models for SFR. Specifically, some works used UNet-like CNN architectures proposed for image reconstruction and inpainting to reconstruct the low frequencies of the sound field (up to 300 Hz) [54], [55] in a pure data-driven fashion. Recent works show that generative models, e.g., Generative Adversarial Networks [21], [56] and diffusion models [57], improve the performances by being able to recover the high frequencies of the sound field. The work of [30] uses a deep prior model to create a regularized solution of the RIR interpolation problem in a pure data-driven fashion treated as image inpainting. The main limitation of the above methods is their pure data-driven nature, which limits their application to the conditions encoded in the training datasets. While physical simulators can be used to augment training data, these models suffer from the variety of acoustic conditions or setups in practice.

Physics-driven models: To address this issue, physics-driven deep learning methods have been proposed. The ben-

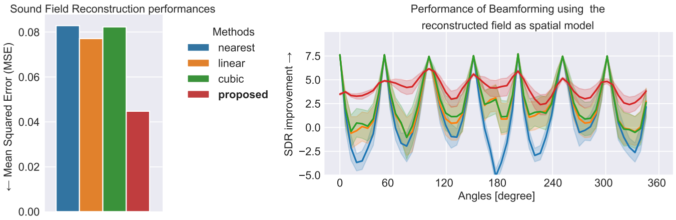

Fig. 3: **Intuition of the limitation of deep learning-based regression and neural fields**: While the proposed method is superior in terms of (low) MSE for the task of sound field reconstruction, traditional methods like nearest neighbor and linear interpolation "retain" the observed data unprocessed. This behavior is difficult to control in deep learning pipelines and can detrimentally affect downstream tasks, such as speech enhancement with beamforming, which requires a spatial model for the sound field.

efit of physics as *inductive bias* is threefold: ensure output structure, overcome data scarcity, and make models less "opaque". Physics-driven ML models can be further categorized into: physics-*informed* [**?**], [108]–[115], [137] and physics-*constrained* models [21], [97], [103].

The main idea is to leverage residuals of the *wave equation* PDE and use them as regularization terms that are minimized during model training in a linear scalarized multi-objective optimization. In most SOTA methods, the parameters are optimized via iterative stochastic optimization for a given weighted combination of tasks, so different combinations require retraining or fine-tuning. A part of the work in [109] that computes the PDE with finite difference scheme, and it is a supervised approach; the other methods extend the unsupervised physics-informed neural network [11], which leverage on the backpropagation to evaluate the PDE residual is a meshless, continuous fashion. These works differ by the setup which consists of several microphones deployed on a confined space (linear [13], planar [109], [112], cubic [111]) or on spherical microphones [108], [110], [115]. Besides, in most of the work, several microphones are deployed, typically more than 32, except [111], [115] that push the analysis to output featuring 16 sparse microphones and 4 microphones, respectively. These strategies produce physical solutions with increasing accuracy as the number of observations, but they lack in several aspects. First, these architectures are hard to train in practice, specifically tuning the hyperparameter of the multi-objective functions; secondly, the physical constraint is only minimized, meaning that no physical guaranties in the structure output is provided; then, the solution is good in the "mean-squared error" sense, which may not be optimized for downstream tasks or distort good observation as exemplified in Figure 3; Finally, this model are intrinsically deterministic and no uncertainty quantification. While the first issue could be addressed with multi-task learning techniques tailored for PINNs [138], our proposed approach aims to solve the remaining ones.

To constraint the solution to be physical, physics-constrained machine learning methods estimated coefficients for basis expansion (e.g., spherical harmonics [21], [106], plane waves [97], [107]) or virtual point sources [103]. These works leverage the power of automatic differentiation of deep learning frameworks to estimate expansion coefficient via neural networks [21], [103], Gaussian process [106], [107], [134] or a combination of the two [97]. Interestingly, the work of [97] also combines kernel methods to compute the directional component of the sound field supported by a shallow-DNN-based ODE solver that returns the physically principled statistical structure of late reverberation in the frequency domain. While this approach outperforms traditional regression methods for interior sound field reconstruction at low frequencies ($<4000\,\mathrm{Hz}$) with several microphones (typically more than 32). This is done to respect the upper bound of the truncation order N of the sound field decomposition via orthonormal bases, $I = 2N + 1$ with $N = \lceil \kappa_{f_{\max}} R \rceil$ [139]. In this work, we are interested in reconstructing the exterior field around the human head, with only 6 microphones, including the high frequencies of typical speech processing ($< 8\,\mathrm{kHz}$).

C. Other related field of research

Some works focus on estimating microphone pickup patterns, referred to as *radiation pattern* in array processing. This problem has attracted some recent interest due to the growing interest in spatial audio recordings (e.g., high-order ambisonic microphones). This application constrained the microphone setup to a specific placement of the capsule. The works problem is usually tackled with the tool discussed above for HRTF upsampling and SFR [70], [72], [115]. The work of [**?**], [140] applies the techniques to the related problem of source directivity estimation with SH interpolation and PINNs, repressively. Interestingly, the works in [70], [115] are the only work that pushes the analysis of spatial upsampling to only 3 and 4 spherical microphones, respectively. Authors of [115] claim the superiority of the PINNs-based approach over the one presented in [70].

The spatial interpolation room impulse response from spare measurements is a special case of SFR. The literature on these two problems has a huge overlap. While the latter has a narrow focus as it seeks higher precision at a low scale, the former tends to aim at the modeling of overall room acoustics on a bigger scale [141] with, typically characterizing acoustic early reflections [100] and late reverberation descriptions [142]. The novel application of augmented listening and immersive navigation of virtual spaces, has driven attention to reducing the computational complexity of RIR interpolation methods, typically relying on machine learning techniques [143]–[147]. As neural fields have become more popular, several works [148]–[151] in recent years proposed neural implicit acoustic fields and achieved state-of-the-art performance. However, these learning-based methods still produce unsatisfying waveform shapes and show weakness in novel impulse response synthesis [141]. To retain physical consistency, geometrical acoustics models, e.g., the image source model [147] or wave equation in [149], are currently promising directions.

Array processing exploits the structure of an array system to enable advanced super-resolution space-time processing. The array manifold is the continuous locus of all the array response vectors (manifold vectors) and maps the geometrical aspect of the array system, typically directional parameters (e.g., angle of

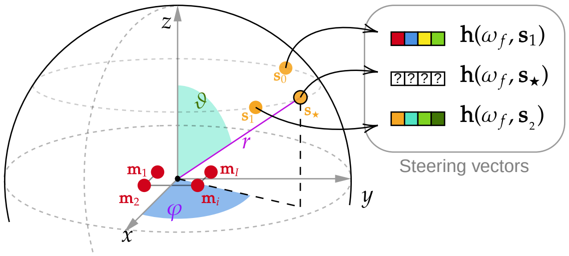

Fig. 4: Measurement grid and reference system used in this work and illustration of the steering vector interpolation problem.

arrival), to the signal environment (steering vectors) [152, Chapter 1.7]. The correct modeling of this quantity enables a high-resolution direction of arrival estimation and signal segregation. Interpolation of array manifold for discrete measurement was addressed in early works [20], but due to methods complexity and suboptimal performances, it evolves into a parameter estimation problem within a calibration problem [153]. As the array's geometry is commonly known a priori, the core modeling of steering vectors can be computed geometrically, leading to *analytical* manifolds [154]. Then, most of the work recasts the problem regarding (self-) array calibration to recover unknown complex gain and phase factors [153] or *mutual coupling matrices* that compensate for the mismatch. We identify that the methods in this field typically rely on an optimization-based scheme that shares similarities with maximum-likelihood-based approaches in beamforming [14]. To the authors' best knowledge, no works address interpolation of array manifold, neither have used the deep learning models to model this continuous object, but we underline the similarity between these two fields and borrow the term for this in our proposed approach.

Physics enters implicitly via training, solutions are data-driven. The network learn the physical operator, but it is not constrained to be the correct one. Works [113]. (DeepONet-based, Fourier Neural Operator, Resolving simple 2D free-field wave equation), [114] (Deep Neural Operator) (Extention to realistic domain, Together with domain decomposition and transfer learning frameworks, Real time and Complex 3D environment, Resolving the low frequency).

III. Problem formulation

A. Free-field sound propagation

In the frequency domain, the homogeneous Helmholtz equation describes the evolution of the complex acoustics pressure field $h \in \mathbb{C}$ as a function of position $\mathbf{q}$ and the angular frequency ω as

$$\nabla^2_{\mathbf{q}} h(\omega, \mathbf{q}) + \frac{\omega^2}{c^2} h(\omega, \mathbf{q}) = 0, \tag{1}$$

where $\nabla^2_{\mathbf{q}}$ is the 3-dimensional Laplacian operator and c is the speed of sound with respect to the space. This equation is linear with respect to h, implying that the pressure field is the sum of the pressure fields resulting from multiple sound sources.

Assuming free space propagation, a single frequency point source at position $\mathbf{s} \in \mathbb{R}^3$ emits a pressure wave in the form of

$$h(\omega, \mathbf{q} \mid \mathbf{s}) = \frac{1}{\sqrt{4\pi r}} e^{-\jmath \omega r / c} \tag{2}$$

where $r = \|\mathbf{q} - \mathbf{s}\|_2$ is the distance between the source and the measurement location and $\jmath = \sqrt{-1}$. This equation is the solution of Eq. (1) in ideal free-field propagation, which is also known as *Green's function*, or free-field *acoustic impulse response* (AIR) in the signal processing vocabulary.

Therefore, the acoustic sound field $x(\omega, \mathbf{q})$ measured at $\mathbf{q}$ produced by a sound source emitting a signal $s(\omega, \mathbf{s})$ at location $\mathbf{s}$, can be computed as

$$x(\omega, \mathbf{q}) = h(\omega, \mathbf{q} \mid \mathbf{s}) s(\omega, \mathbf{s}). \tag{3}$$

B. Spherical representation

The sound pressure field can represented as a linear combination of spatial basis functions [136]. Regarding free-field sound propagation on the sphere surface, spherical harmonics are solutions of the PDE in eq. (1) and are commonly used to describe a pressure field. A function $H(\Omega, \omega)$ that is square integrable on the surface $\Omega = (\varphi, \vartheta) \in \mathbb{S}^2$ of a 3D-sphere that is centered around the coordinate origin can be represented by the coefficients $H_{nm}(\omega)$ of a series of spherical harmonics $Y_{nm}(\Omega)$ [155]. Applying the Helmholtz reciprocity principle, we describe the sound pressure of a sound source in the ear of the subject on the surface of a sphere.

$$x(\omega, \mathbf{q}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} c_{lm}(\omega, \mathbf{q}_0) \bar{Y}_l^m(\omega, \mathbf{q} - \mathbf{q}_0) \tag{4}$$

where $\kappa = \omega / c$ is the wave number. Here c_{lm} are the expansion coefficient of order l and degree m, $\mathbf{q}_0$ is the expansion center, and $\bar{Y}_m^l(\cdot)$ are the corresponding modified spherical harmonics.

Depending on the nature of the sound field of interest, the modified spherical harmonics can be further developed as

$$\bar{Y}_l^m(\omega, \mathbf{q}) = \begin{cases} Y_m^l \left(\frac{\mathbf{q}}{\|\mathbf{q}\|_2} \right), & \textit{in general case} \\ h_l^1(\kappa \|\mathbf{q}\|_2) Y_l^m \left(\frac{\mathbf{q}}{\|\mathbf{q}\|_2} \right), & \textit{exterior field} \\ j_l^1(\kappa \|\mathbf{q}\|_2) Y_l^m \left(\frac{\mathbf{q}}{\|\mathbf{q}\|_2} \right), & \textit{interior field} \end{cases} \tag{5}$$

where $\bar{Y}_m^l$ are spherical harmonics of order l and degree m accepting as argument the azimuthal and the polar coordinate of the unit vector.

The surface spherical harmonics $Y_l^m(\cdot)$ are a complete and orthonormal set that can be defined as

$$Y_n^m(\varphi, \vartheta) = (-1)^m \sqrt{\frac{(2l+1)}{4\pi} \frac{(n-|m|)!}{(n-|m|)!}} P_n^{|m|}(\cos \varphi) e^{\jmath m \vartheta} \tag{6}$$

where $P_l^m(\cdot)$ denotes mth-order the associated Legendre function of n-th degree, ϕ denotes the azimuth and θ the colatitude.

IV. Sound field regression

Let $y_n := y(\mathbf{z}_n)$ denote a noise measurement of the sound field h generated from a single source, such that

$$y_n = h(\mathbf{z}_n) + \varepsilon_n \tag{7}$$

where ε_n models noise and $\mathbf{z}_n = [\omega_n, \mathbf{q}_n] \subset \mathbb{R} \times \mathbb{S}^2$ is the vector of measurement collocation point, frequency and space. Let $\mathbf{y} = [y(\mathbf{z}_1), \dots, y(\mathbf{z}_N)]^\mathsf{T} \in \mathbb{C}^{FI}$ and $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_N]^\mathsf{T} \in \mathbb{R}^{N \times 4}$ be the vector of measurements the sound field measured and coordinates at F frequencies and I positions, respectively. Given $\mathbf{y}$ and $\mathbf{Z}$, we here consider the problem of estimating the underling continuous function h, or, similarly, the reconstruction of the sound field at another frequency and position $\mathbf{z}_*$. We will refer to this problem as regression, that is interpolation in presence of noise [156].

The estimation of the continuous function $f : \mathbb{R}^P \to \mathbb{K}$ ($\mathbb{K}$ is $\mathbb{R}$ or $\mathbb{C}$) from discrete observation $\mathbf{y} \in \mathbb{K}^N$ at the sampling points $\{\mathbf{z}_i\}_{n=1}^N$ is achieve by representing f with some model with parameters $\boldsymbol{\theta}$ and solving the following optimization problem

$$\operatorname*{arg\,min}_{\boldsymbol{\theta}} \mathcal{L}\left(\mathbf{y}, \mathbf{f}(\{\mathbf{z}_i\}_{n=1}^N; \boldsymbol{\theta})\right) + \mathcal{R}(\boldsymbol{\theta}) \tag{8}$$

where $\mathbf{f}(\{\mathbf{z}_i\}_{n=1}^N; \boldsymbol{\theta}) = [f(\mathbf{z}_1; \boldsymbol{\theta}), \dots, f(\mathbf{z}_N; \boldsymbol{\theta})]^\mathsf{T} \in \mathbb{K}^N$ is the vector of the discretized function f represented by $\boldsymbol{\theta}$. $\mathcal{L}$ is a loss function evaluating the distance between $\mathbf{x} andf$ at $\{\mathbf{z}_i\}_{n=1}^N$, and $\mathcal{R}$ is a regulation term for $\boldsymbol{\theta}$ to prevent overfitting.

A. Regression with basis expansion

A common approach in regression is to represent f as a linear combination of basis functions, as

$$f(\mathbf{z}; \gamma) = \sum_{l=1}^{L} \gamma_l \psi_l(\mathbf{z}), \tag{9}$$

where $\boldsymbol{\gamma} = [\gamma_1, \dots, \gamma_L]^\mathsf{T} \in \mathbb{K}^L$ and $\psi(\mathbf{z}) = [\psi_1, \dots, \psi_L]^\mathsf{T} \in \mathbb{K}^L$, for instance the spherical wave function expansion. If the squared error loss function and a ℓ_2 penalty are used, eq. (8) yield the following closed-form solution

$$\hat{\boldsymbol{\gamma}} = \operatorname*{arg\,min}_{\boldsymbol{\gamma}} \|\mathbf{y} - \boldsymbol{\Psi}\boldsymbol{\gamma}\|_2^2 + \lambda\|\boldsymbol{\gamma}\|_2^2 \tag{10}$$

$$= \left(\boldsymbol{\Psi}^\mathsf{H}\boldsymbol{\Psi} + \lambda\mathbf{I}\right)^{-1} \boldsymbol{\Psi}^\mathsf{H}\mathbf{y}, \tag{11}$$

where $\boldsymbol{\Psi} = [\psi(\mathbf{x}_1), \dots, \psi(\mathbf{x}_N) \in \mathbb{K}^{N \times L}$, $\mathbf{I}$ is the identity matrix and $\cdot^\mathsf{H}$ denotes Hermitian transposition. Then, f is a linear combination of the basis function $\{\psi_l\}_l$ by construction. The problem of recovering the plane wave coefficients is typically ill-posed, rank deficient and typically under-determined as the measured pressure positions are significantly less than the number of plane waves used to reconstruct the sound field. To obtain a unique and stable solution, additional measurements or regularization techniques are needed. Regularization methods involve adding a penalty term to the optimization problem, which encourages certain properties of the solution, such as smoothness or sparsity [99].

B. Kernel Ridge regression

On the basis of the representer theorem [157], f is represented by a weighted sum of kernel function k as

$$f(\mathbf{z}; \boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n k(\mathbf{z}, \mathbf{z}_n) \tag{12}$$

$$= \mathbf{k}(\mathbf{z})^\mathsf{T}\boldsymbol{\alpha}, \tag{13}$$

where $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_N]^\mathsf{T} \in \mathbb{K}^N$ and the weight coefficient are $\mathbf{k}(\mathbf{z}) = [k(\mathbf{z}, \mathbf{z}_1), \dots, k(\mathbf{z}, \mathbf{z}_N)]^\mathsf{T} \in \mathbb{K}^N$ is the vector of kernel functions. In the kernel ridge regression [157], the estimated of $\boldsymbol{\alpha}$ is compute as

$$\hat{\boldsymbol{\alpha}} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}, \tag{14}$$

with the Gram matrix $\mathbf{K} \in \mathbb{K}^{N \times N}$ defined as

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{z}_1, \mathbf{z}_1) & \dots & k(\mathbf{z}_1, \mathbf{z}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{z}_N, \mathbf{z}_1) & \dots & k(\mathbf{z}_N, \mathbf{z}_N) \end{bmatrix} \tag{15}$$

Then f is interpolated by substituting $\hat{\boldsymbol{\alpha}}$ in eq. (13) A common way of constructing kernel is through inner products in Hilbert spaces. ... Ask Koyama-sensei to fill this part. ... The design of kernel functions is not limited to inner products, but also combination of existing kernels, such as summation and multiplication [156].

C. Gaussian process interpolation

A Gaussian process (GP) [156] is a collection of random variables, any finite number of which have a joint Gaussian distribution. Given any finite set of n input $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ and the corresponding set of latent function values $\mathbf{y} = \{y(\mathbf{z}_1), \dots, y(\mathbf{z}_n)\}$, the relationship between the input data $\mathbf{x}_n$ and the observed noisy target y_n are given by

$$y_n = f(\mathbf{z}_n) + \varepsilon_n, \quad \varepsilon_n \sim \mathcal{N}(0, \sigma^2) \tag{16}$$

where ε is the zero-mean Gaussian noise with variance σ^2.

The prior distribution over the latent function can be written as

$$\mathrm{p}(\mathbf{f} \mid \mathbf{Z}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}) \tag{17}$$

where the $\boldsymbol{\mu} = m_{\boldsymbol{\theta}}(\mathbf{Z}) \in \mathbb{K}^N$ is the mean vector computed with the mean function $m(\cdot)$ and $\mathbf{K}$ is the Gram matrix whose element $\mathbf{K}_{nn'} = k_{\boldsymbol{\theta}}(\mathbf{z}_n, \mathbf{z}_{n'}) \in \mathbb{K}^{N \times N}$ is based on the kernel function. Both the mean and the kernel function may have hyper-parameters, denoted here as $\boldsymbol{\theta}$. The kernel function $k(\cdot, \cdot)$ model the correlation between points of the GP.

For the predictive distribution, also known as posterior distribution, of the function values $\mathbf{f}_\star$ at the test set $\mathbf{z}_\star$ is

$$\mathrm{p}(\mathbf{f}_\star \mid \mathbf{z}_\star, \mathbf{y}, \mathbf{Z}) \sim \mathcal{N}(\boldsymbol{\mu}_\star, \boldsymbol{\Sigma}_\star) \tag{18}$$

where the mean $\boldsymbol{\mu}_\star$ and the covariance $\boldsymbol{\Sigma}_\star$ are calculated as

$$\boldsymbol{\mu}_\star = m(\mathbf{z}_\star) + \mathbf{k}_\star^\mathsf{T}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \boldsymbol{\mu}) \tag{19}$$

$$\boldsymbol{\Sigma}_\star = k(\mathbf{z}_\star, \mathbf{z}_\star) - \mathbf{k}_\star^\mathsf{T}(\mathbf{K} + \sigma_n^2\mathbf{I})\mathbf{k}_\star \tag{20}$$

where $\mathbf{k}_\star = [k(\mathbf{z}_\star, \mathbf{z}_1), \dots, k(\mathbf{z}_\star, \mathbf{z}_n)]$ is the vector of covariances between the test point and the N training points.

Kernel functions for sound field reproduction: The GP prior covariance function encodes the assumed constraints on the latent function h. In particular the correlation between any subset of points, that is, the smoothness of the interpolation, is fully specified by the GP prior as function over the input domain and hyper-parameters. The works of [96], [105] discusses the use two spatial kernel functions based on the plane wave decomposition: one an anisotropic (i.e., direction-dependent) stationary kernel to model directional components of the sound field (e.g., direct path and early echoes) and a stationary and isotropic kernel to model the late reverberation. In [125], a zero-mean GP process is assumed to model HRTF both in space and frequency. The joint spatial-frequency covariance function is specified through single GP covariance as the product of a *Ornstein-Uhlenbec* (OU) density (suitable to model exponentially-decaying process, or the continuous-time analogue of the discrete-time auto-regressive AR1 process) and a stationary covariance of based on the Matérn $3/2$ function [156] of the chordal distance (See [158] for a comparison of different kernel function for HRTF interpolation). The associated Gram matrix for a measurement set as a Cartesian outer-product $\mathbf{Z} = \mathbf{Z}^{(\omega)} \times \mathbf{Z}^{(\mathbf{q})}$ reads $\mathbf{K} = \mathbf{K}^{(\omega)} \otimes \mathbf{K}^{(\mathbf{q})}$, that is,

$$k = k_\omega(\omega_f, \omega_{f'}) k_{\mathbf{q}}(\Omega_j, \Omega_{j'}) \tag{21}$$

$$k_\omega(\omega_f, \omega_{f'}) = \frac{\alpha_f}{\ell_\omega^2 + (\omega_f - \omega_{f'})^2} \tag{22}$$

$$k_{\mathbf{q}}(\Omega_j, \Omega_{j'}) = \left(1 + \frac{\sqrt{3}C_{jj'}}{\ell_{\mathbf{q}}}\right) \exp\left(-\frac{\sqrt{3}C_{jj'}}{\ell_{\mathbf{q}}}\right) \tag{23}$$

$$C_{jj'} = 2\sqrt{\sin^2\left(\frac{\vartheta_j - \vartheta_i}{2}\right) + \sin\vartheta_i \sin\vartheta_j \sin^2\left(\frac{\varphi_i - \varphi_j}{2}\right)}, \tag{24}$$

where $\Omega_j = (\varphi_j, \vartheta_j)$ is the j-th DOA, the length-scale parameters ℓ_d is the distance for function values to become uncorrelated in the d-th dimension. α and λ are the global scale factor and the mean drift rate to 0 in the OU process, respectively.

D. Regression with Neural Fields

A neural field (NF) $\mathcal{F}_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}, \mathbf{z} \mapsto y$ is a *coordinate*-based neural network that maps points $\mathbf{z}$ to the function value y [121]. Let be $\hat{y}(\cdot; \boldsymbol{\theta}) = \mathcal{F}_{\boldsymbol{\theta}}(\cdot)$ the network returned value. The network parameters $\boldsymbol{\theta}$ are commonly optimized by minimizing the loss function

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \left(y_n - \hat{y}(\mathbf{z}_n; \boldsymbol{\theta})\right)^2 + \lambda \mathcal{R}(\boldsymbol{\theta}) \tag{25}$$

where the first term is the empirical risk function and $\mathcal{R}$ is a regularization function that prevents over-fitting, scaled by λ. As the network size grows to infinity, the function learned by these neural network can be viewed as the solution of kernel regression, with kernel function being the Neural Tangent Kernel (NTK) [159]. The study of the empirical NTK of a NF helps to better design and train of these model [138].

The fundamental property of NF (actually, coordinate-based neural networks) is to be grid-free (mesh-less) model: although the training set is discrete, $\{\mathbf{z}_n\}_n \subset \mathbb{R}^d$, the model can evaluate any point in $\mathbb{R}^d$. Meaning that a NF can perform continuous evaluation at inference, but also during training if non-intrusive metrics are used for regularization. This property enable novel view synthesis for NeRF and PINNs, and modeling signals on manifolds [121]. Finally, Neural fields can compactly store complicated shapes without spatial discretization [160]; and, being generally infinitely differentiable, allows them to be optimized for objectives that involve higher-order derivatives.

Natural signals, such as shapes, images and sounds, contains rich high-frequency content. Due to spectral bias, standard neural network (e.g., MLP architectures), fails to learn high-frequency function from low dimensional data [120], [138], and generate blurry or over-smooth version of the target quantity. To address this issues two main approaches have been proposed:

a) Random Fourier Features: the composition $\mathcal{F}_{\boldsymbol{\theta}} \circ \gamma$ of a neural field and a Random Fourier Features (RFF) [120] encoding γ helps to overcome the spectral bias, enabling the neural network of represent signals with high-frequency components. The RFF encoding $\gamma : \mathbb{R}^P \to \mathbb{R}^{2D}$ with $D \gg P$ is defined as

$$\gamma(\mathbf{z}) = [\sin(2\pi\mathbf{B}\mathbf{z}), \cos(2\pi\mathbf{B}\mathbf{z})], \tag{26}$$

where $\mathbf{B} \in \mathbb{R}^{D \times P}$ whose elements are randomly drawn from the normal distribution $\mathcal{N}(0, \sigma^2_{\text{RFF}})$. RFF features are generally not optimized during training and $\sigma^2_{\mathbf{RFF}}$ is an hyper-parameters that balancing under/overfitting, resampling the characteristic length-scale in GP.

b) SIREN network: Sinusoidal activation functions have been recently proposed as an effective alternative way to overcome the spectral bias [161]. The proposed model share the same fundamental structure of a MLP, for which the r-th layer reads

$$\phi_r(\mathbf{z}_r) = \sin\left(g_r \mathbf{W}_r \mathbf{z}_r + \mathbf{b}_r\right) \tag{27}$$

where $\mathbf{z}_r, \mathbf{W}_r, \mathbf{b}_r, g_r$ are the input vector, the weight, the biases and a hyper-parameter of the the r-th layer, respectively. Finally, the SIREN architecture is a composition of R layers,

$$\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{z}) = (\phi_R \circ \phi_{R-1} \circ \cdots \circ \phi_1)(\mathbf{z}). \tag{28}$$

This architectural design became popular for learning implicit representation and as backbones for PINNs models thanks to the property of representing high-order signal derivatives effectively [121]. Yet, this model forces periodicity of the underlying estimated function yielding high-frequency artifacts (resampling Gibbs artifacts for truncated Fourier series) [162]. To overcome this limitation, extensive parameter tuning, parametrization and meta-learning training strategies have been proposed [162], [163].

c) Sinusoidal feature (SF):

$$\gamma(\mathbf{z}) = [\sin(2\pi\mathbf{W}_1\mathbf{z} + \mathbf{b}_1)], \tag{29}$$

In both the approaches, the hyper-parameter offers a trade-off between reconstruction fidelity and over-fitting. It has been shown that it is related to the bandwidth of the target function and it must be tuned accordingly.

Given the set $\{h_n\}_{n=1}^N$, with $h_n \in \mathbb{C}$, measurements of the sound field at angular frequency and location $\{(\omega_n, \mathbf{q}_n)\}_{n=1}^N \subset$

$\mathbb{R} \times \mathbb{R}^3$, a NF, $\mathcal{F}_{\boldsymbol{\theta}_{\text{NF}}}$, can be used for regression problem. The loss function use to train the network reads

$$\mathcal{L}_{\boldsymbol{\theta}} = \frac{1}{N} \sum_{n=1}^{N} \left| h_n - \hat{h}(\omega_n, \mathbf{q}_n; \boldsymbol{\theta}) \right|^2 . \tag{30}$$

In general, at test time, the networks can always evaluate continuous any new coordinate $(\omega_\star, \mathbf{q}_\star) \in \mathbb{R} \times \mathbb{R}^3$.

d) Neural Fields for sound field interpolation: NF have been recently applied for continuous representation of sounds [161], [164], acoustics impulse response [123], [151], [165], [166] and HRTF upsampling and personalization [27], [80], [167]; their use for sound field reconstruction leveraging on the wave equation is discussed below. The interest in regression in accompanied by the goal of condensing multiple observation sets (e.g., different HRTF measurements) into a single learned la fully differentiable latent representation. These work focusing on spatial upsampling investigates different strategies to achieve the task: some focus on using geometric properties of far-field prorogation to guide the training, also called *geometric wrapping (GW)* [58], [123], [165], or relying on parametric model of the HRTF filters [80]. Besides, each work proposed different MLP configuration (positional encoding, activation functions), output encoding (real and imaginary or log-magnitude and phase), and training objectives (magnitude-only loss, combination of magnitude- and phase-bases loss terms, multi-resolution STFT, etc.) whose performances depends on the specific task and data.

Geometric wrapping: it can be seen an adaptation of interaural phase difference compensation when processing minimum phase HRTF [38], [168]. Following the model we presented in our previous work [123], the propagation effects at the i-th microphone in $\mathbf{m}_i$ attending a j-the source in $\mathbf{s}_j$ at frequency ω_f with respect to the reference point $\mathbf{q}$ writes

$$d(\underbrace{\omega_f, \mathbf{m}_i \,|\, \mathbf{s}_j}_{\mathbf{z}_n}, \mathbf{q}) = \frac{d_{ij}}{d_j} \exp(-\jmath \omega_f (d_{ij} - d_j)/c) \tag{31}$$

where $d_{ij} = \|\mathbf{s}_j - \mathbf{m}_i\|_2$ and $d_j = \|\mathbf{s}_j - \mathbf{q}\|_2)$ is the distance from source to the microphone and references, respectively, and c is the speed of sound. Equation (31) is commonly known as relative steering vectors in beamforming techniques [14], [17].

The output of the NF featuring GW as follows

$$\hat{h}_{\text{GW}}(\mathbf{z}_n) = d(\mathbf{z}_n, \mathbf{q}) f_{\theta_{\text{NF}}}(\mathbf{z}_n), \tag{32}$$

where $f(\mathbf{z})_{\theta_{\text{NF}}}$ is the output of the neural network and $\mathbf{z} = [\omega_f, \mathbf{m}_i, \mathbf{s}_j]$ extend the input coordinate with the microphone position. Thanks to the differentiability of the model 31, the training can be performed end-to-end. This simple approach share principles with differentiable digital signal processing [126].

Our previous contribution [123] showed the effectiveness of this model to interpolate steering vectors with respect to frequency and source-mic positions. To the best knowledge of the authors, there are no other works that study sound field upsampling with respect these three axis.

Output encoding: Neural Fields models are generally implemented with real-valued neural networks. While complex-value neural networks are currently being investigated, a common approach is it estimate real and imaginary part separately in the field of sound field reconstruction and log-magnitude and phase for HRTF upsampling. The latter is subject to instability due to phase wrapping. While it is common to focus on the log-magnitude of minimum-phase HRTF [27], phase components are essential for correct processing in analysis and synthesis problem [36]. In our previous work of [123] we proposed the following output encoding to deal with phase wrapping enhancing stability and convergence:

$$h_n = \exp(a_1) \exp\left(-\jmath 2\pi \arctan\left(\frac{a_2}{a_3}\right)\right) \in \mathbb{C} \tag{33}$$

where $[a_1, a_2, a_3] \in \mathbb{R}^3$ are the actual outputs of the real-valued NF.

E. Physics-informed Neural Networks (PINNs)

A common PINN [11] consider a feed-forward MLP network for modeling a dynamical function u of a physical system in a space $\mathbf{z} \in \Omega \subset \mathbb{R}^d$, with networks parameter $\boldsymbol{\theta}$ to be optimized. In general, u mathematically obeys known priors, such as PDE of the general form

$$\mathcal{M}_z[u(z)] = 0, \quad z \in \Omega \tag{34}$$

$$\mathcal{B}[u(z)] = d(z), \quad z \in \partial\Omega \tag{35}$$

where $\mathcal{M}_z[\cdot]$ is a general combination of nonlinear differential operator, which can include any combination of derivative with respect the input variable z, such as the first- and second-order derivative $\frac{\partial u}{\partial z}$ and $\frac{\partial^2 u}{\partial z^2}$, respectively. The boundary operator $\mathcal{B}[\cdot]$ enforces the desired condition $d(z)$ at the boundary $\delta\Omega$. In case of the wave equation in the frequency domain $\mathcal{M}_{\mathbf{q},\omega} = \nabla^2_{\mathbf{q}} - \frac{\omega^2}{c^2}$.

The training loss function of a PINN extends the regular loss function in (25) with PDE-based regularization

$$\mathcal{R}(\boldsymbol{\theta}) = \lambda_{\text{PDE}} \mathcal{L}_{\text{PDE}} + \lambda_{\text{IC}} \mathcal{L}_{\text{IC}} \tag{36}$$

$$\mathcal{L}_{\text{PDE}} = \|\mathcal{M}_z[\hat{u}(z; \boldsymbol{\theta})]\|^2 \qquad z \in \Omega \tag{37}$$

$$\mathcal{L}_{\text{IC}} = \|\mathcal{B}_z[\mathcal{F}_{\boldsymbol{\theta}}(z)] - d(z)\|^2 \qquad z \in \partial\Omega. \tag{38}$$

The relative weight, $\lambda_{(\cdot)}$ control the trade-off between different components and need to be scaled depending on the problem at hand. The physical loss components $\mathcal{L}_{\text{PDE}}$ and $\mathcal{L}_{\text{IC}}$ are defined over the continuous domain Ω, but in practice, they are computed over a finite set of collocation points that must be sampled, for example, on a uniform grid. The computation of differential operators is conveniently computed via automatic differentiation.

Training PINNs is known to be challenging due to possible noise in the data [169]. To mitigate this challenge several techniques have been proposed, spanning from meta-learning for multi-objective optimization and curriculum learning [138], [170], optimizer-switching [171], strategic sampling to evaluate the residuals during training [172], and practical design choice (similar to the one discussed for NF) [138], [148]. Reader can refer to [173] for a practical introduction.

Self-adaptive learning rate annealing proposed in [173] can be used to automatically balance the losses during training. The idea is constrain the the norm of gradients of each weighted loss to be equal to each other, preventing our model from

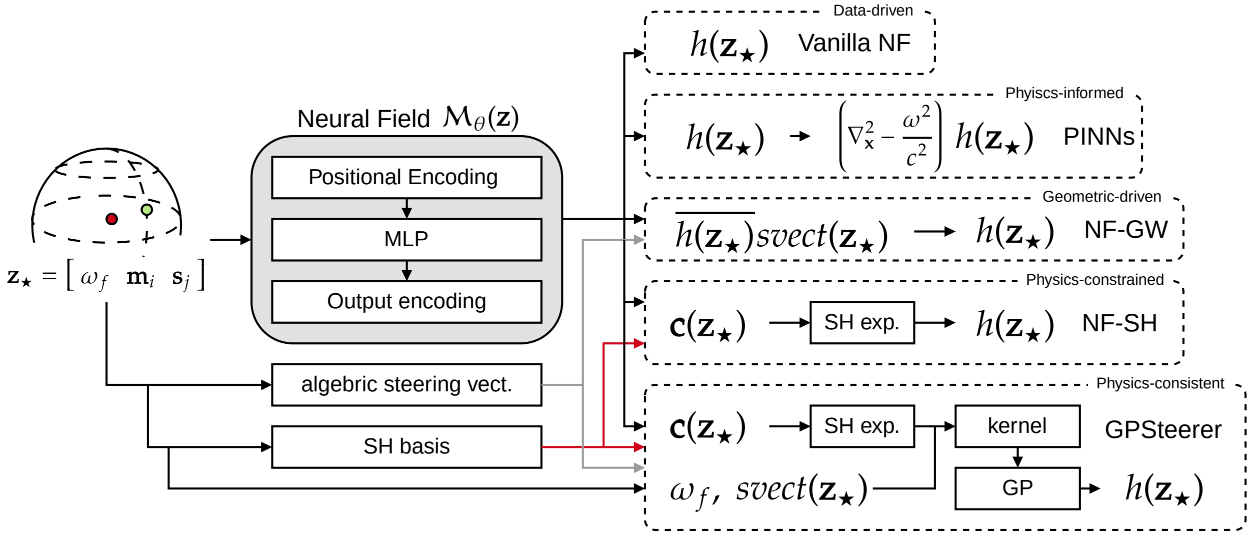
Data-driven
$h(\mathbf{z}_\star)$ Vanilla NF
Phyiscs-informed
$h(\mathbf{z}_\star)$ → $\left(\nabla^2_{\mathbf{x}} - \frac{\omega^2}{c^2}\right) h(\mathbf{z}_\star)$ PINNs
Geometric-driven
$\overline{h(\mathbf{z}_\star)}svect(\mathbf{z}_\star)$ → $h(\mathbf{z}_\star)$ NF-GW
Physics-constrained
$\mathbf{c}(\mathbf{z}_\star)$ → SH exp. → $h(\mathbf{z}_\star)$ NF-SH
Physics-consistent
$\mathbf{c}(\mathbf{z}_\star)$ → SH exp. → kernel GPSteerer
$\omega_f,\ svect(\mathbf{z}_\star)$ GP → $h(\mathbf{z}_\star)$
Neural Field $\mathcal{M}_\theta(\mathbf{z})$
Positional Encoding
MLP
Output encoding
$\mathbf{z}_\star = [\,\omega_f \ \ \mathbf{m}_i \ \ \mathbf{s}_j\,]$
algebric steering vect.
SH basis

Fig. 5: Steering vector upsampling models

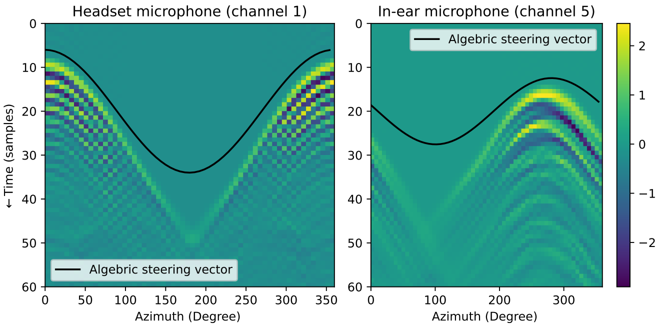

Fig. 6: Comparison between measured and algebraic steering vectors available in the SPEAR Challenge data.

being biased towards minimizing certain terms during training. Adapted for Equation (36), the regularization parameters λ_{PDE} and λ_{IC} are computed as

$$\lambda_{\text{PDE}} = \frac{\|\nabla_\theta \mathcal{L}_{\text{PDE}}\|_2 + \|\nabla_\theta \mathcal{L}_{\text{IC}}\|_2}{\|\nabla_\theta \mathcal{L}_{\text{PDE}}\|_2}, \tag{39}$$

$$\lambda_{\text{IC}} = \frac{\|\nabla_\theta \mathcal{L}_{\text{PDE}}\|_2 + \|\nabla_\theta \mathcal{L}_{\text{IC}}\|_2}{\|\nabla_\theta \mathcal{L}_{\text{IC}}\|_2} \tag{40}$$

obtaining $\|\nabla_\theta \mathcal{L}_{\text{PDE}}\|_2 = \|\nabla_\theta \mathcal{L}_{\text{IC}}\|_2 = \|\nabla_\theta \mathcal{L}_{\text{PDE}}\|_2 + \|\nabla_\theta \mathcal{L}_{\text{IC}}\|_2$. Note that $\nabla_\theta(\cdot)$ can be easily computed via automatic differentiation.

At every training iteration, the evaluation of the PDE residual require sampling the continuous input domain. The location and distribution of these residual points impact the training stability and the performance of PINNs as the model's gradient may vary significantly over the input domain. To address this issue, different strategies have been proposed from fixed grid-based to adaptive non-uniform samplings, which nevertheless depends on the problem under investigation. A simple, yet effective approach consists is to use residual points that are uniformly resampled every certain number of iteration, that a value that becomes a hyper-parameter of the model. The work in [172] presents a recent review of more advanced strategic sampling methods. Our previous contribution [174] discuss the possibility to use statistical description of the field to sample residual point at different scales, while [123] uses the causal learning paradigm presented in [173].

a) PINNs for sound field reconstruction: In case of sound field reconstruction regression in the frequency domain, the loss function use to optimize the internal parameters $\boldsymbol{\theta}_{\text{PINN}}$ of the PINN reads [12], [115]

$$\mathcal{L}_{\boldsymbol{\theta}} = \frac{1}{N}\sum_{n=1}^{N}\left(h_n - \hat{h}(\omega_n, \mathbf{q}_n; \boldsymbol{\theta})\right)^2 + \lambda_{\text{PDE}}\frac{1}{M}\left(\nabla^2_\theta \hat{h}(\omega_n, \mathbf{q}_n; \boldsymbol{\theta}) + \frac{\omega_n^2}{c^2}\hat{h}(\omega_n, \mathbf{q}_n; \boldsymbol{\theta})\right)^2. \tag{41}$$

Similarly to the NFs, a PINNs can evaluate any continuous coordinate at test time. In case of HRTF upsampling, authors of [12] propose a rearrangement of the Helmholtz equation used in eq. (41), where the ω_n^2/c^2 is used as denominator of the Laplacian ∇^2 , making the magnitude of the PDE loss comparable to the data-fit term, leading to a more balanced and simplified training. Besides, the authors propose to use different PINNs model to spatially upsampling single frequency independently, using different small independent architectures (less than 100 trainable parameters) that only evaluate spatial coordinates. This idea was motivated by the common narrow-band processing spatial audio pipelines. When dealing with spherical data (e.g., HRTF), it is common to adapt polar coordinate systems. However, the PDE becomes unstable due to $\sin\vartheta$ term in the denominator [12]. Adopting a Cartesian system simplifies the PDE, automatically handle wrapping of the spherical data, and avoid coding error due to different system representation convention[1].

The works in [13], [175] proposed a similar approach in the

[1] This article adopted ISO 80000-2:2019 convention as depicted in **??**.

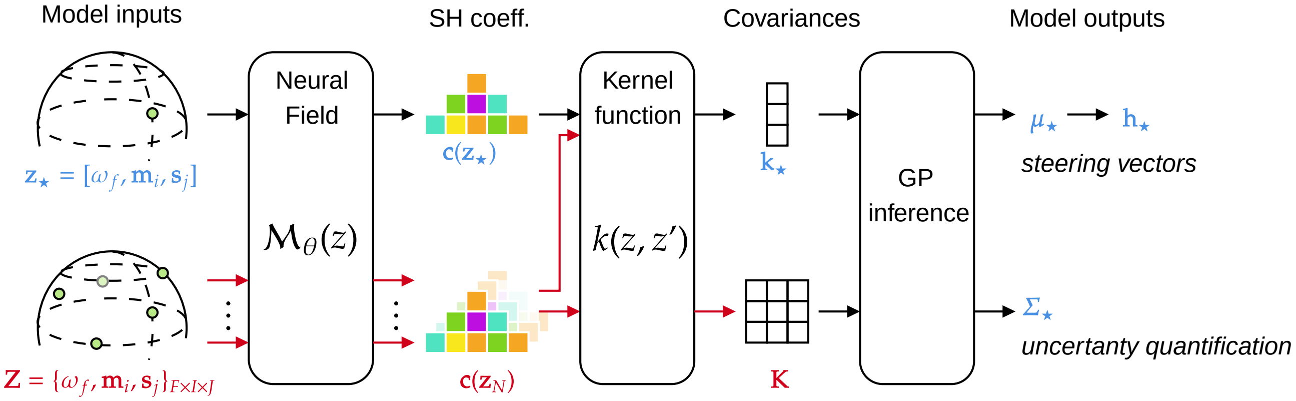
Model inputs
SH coeff.
Covariances
Model outputs
Neural Field
$\mathcal{M}_\theta(z)$
Kernel function
$k(z, z')$
GP inference
$\mathbf{z}_\star = [\omega_f, \mathbf{m}_i, \mathbf{s}_j]$
$\mathbf{c}(\mathbf{z}_\star)$
$\mathbf{k}_\star$
$\mu_\star$
$\mathbf{h}_\star$
steering vectors
$\mathbf{Z} = \{\omega_f, \mathbf{m}_i, \mathbf{s}_j\}_{F\times I\times J}$
$\mathbf{c}(\mathbf{z}_N)$
$\mathbf{K}$
$\Sigma_\star$
uncertanty quantification

Fig. 7: Pipeline of the proposed model

time domain for RIR interpolation, which requires to compute the gradient with respect to spatial and temporal coordinates. The benefits of a time (resp. freq) domain modeling depends on the application and downstream tasks as the independent variable t (f) can be used for further processing. In our previous work [123], the frequency-base modeling was found useful to compute algebraic steering vectors in closed form and to evaluate a non-intrusive regularization term. One disadvantage of frequency domain processing is dealing with complex values.

V. Proposed model

In this work, we aim at upsampling the steering vectors $\mathbf{h}(\omega_f, \mathbf{s}_j) = [h_1(\omega_f, \mathbf{s}_j), \ldots, h_I(\omega_f, \mathbf{s}_j)]^\top \in \mathbb{C}^{I\times 1}$ of a set of I microphones capturing the scattering sound field on the sphere around the head emitted by a sound source at position $\mathbf{s}_j$ at angular frequency ω_f. The setup is illustrated in **??**. We propose to model the quantity h_i with a continuous field $\mathcal{M}_\theta$ with parameter θ over frequency, microphone-source pair positions, that is

$$h_i(\omega_f, \mathbf{s}_j) = \mathcal{M}_\theta(\omega_f, \mathbf{s}_j, \mathbf{m}_i), \tag{42}$$

where $\mathbf{m}_i$ is the position of the i-th microphone in Cartesian coordinates.

We propose to overcome the limits of the physics-informed and physics constrained approaches introducing the following *physically coherent machine learning* framework. The core idea is to extend physics-constrained neural networks within a Bayesian framework for regression. In a nutshell, we use NF to encode the manifold of the coefficient of the spherical harmonics expansions which is used as prior for a GP-based regression. While the former provide physically plausible prediction, the latter enable model flexibility to external noise, and uncertainty quantification.

Following the formulation presented in **??**, we model the underlying continuous sound field h, we place a GP prior distribution over the latent function h. Specifically we assume that h following a zero-mean GP whose prior covariance $\mathbf{K}$ decomposes into separable covariance evaluations in the spectral, "source-spatial", and "microphone-spatial" domain. The associated kernel function for $\mathbf{K}$ decomposes into

$$k(\mathbf{z}_{fij}, \mathbf{z}'_{fij}) = k_\omega(\omega_f, \omega_{f'}) k_\mathbf{s}(\mathbf{z}_{fij}, \mathbf{z}'_{fij}) k_\mathbf{m}(\mathbf{z}_{fij}, \mathbf{z}'_{fij}) \tag{43}$$

where $\mathbf{z}_{fij} = [\omega_f, \mathbf{m}_i, \mathbf{s}_j]$ and $\mathbf{z}'_{fji} := [\omega_{f'}, \mathbf{m}_{i'}, \mathbf{s}_{j'}]$ are shorthand for readability. The spectral kernel is modeled by the inverse-quadratic kernel given by Equation (22), which ensure an exponentially decays temporal response and some degree of smoothness in the spectral response. The spatial-source kernel is derived by the spherical harmonics expansions [102] as,

$$k_\mathbf{s}(\mathbf{z}_{fij}, \mathbf{z}'_{fij}) = \Psi(\mathbf{z}_{fij}) \Psi^*(\mathbf{z}'_{fij}), \tag{44}$$

$$\Psi(\mathbf{z}'_{fij}) = \sum_{l\geq 1}^{N} \sum_{m=-l}^{l} c_{lm}(\mathbf{z}_{fij}) \bar{Y}_{lm}(f, \mathbf{s}_j - \mathbf{q}_0), \tag{45}$$

where $()^*$ denotes complex conjugate and $\mathbf{q}_0$ is the reference point of the system at the center of the head. $\bar{Y}_{lm}(f, \mathbf{s}_j - \mathbf{q}_0)$ is computed as in Equation (5).

The spatial-microphone kernel is derived by the directional plane wave kernel [105] using the notation of steering vectors used in the definition of rank-1 Spatial Covariance Matrices (SCM) in speech enhancement [14], that is,

$$k_\mathbf{m}(\mathbf{z}_{fij}, \mathbf{z}'_{fij}) = \bar{d}(\mathbf{z}_{fij}) \bar{d}^*(\mathbf{z}_{fij}), \tag{46}$$

where $\bar{d}(\mathbf{z}_{fij}) := d(\omega_f, \mathbf{m}_i \parallel \mathbf{s}_j, \mathbf{q}_0)$ is the algebraic steering vector of Equation (31).

Note that Equation (44) and Equation (46) can be thought as the source and the spatial model in the Local Gaussian Modeling used in speech enchantment. One could think of our model as using the spherical harmonics expansion to describe the sources and their spatial distribution on the sphere, and the a SCM based on the anechoic propagation to encode the sound impinging an multichannel array.

A. Parameter estimation

The parameter of the proposed model consist in the parameter of the kernel function k used to compute the prior covariance matrix of the GP: the characteristic length-scale ℓ_f, the global scale α, the complex coefficients $\mathbf{c}(\mathbf{z}_{fij}) = [c_{00}(\mathbf{z}_{fij}), \ldots, c_{L(L+1)}(\mathbf{z}_{fij})]^\top \in \mathbb{C}^{L(L+1)\times 1}$ of the spherical

Model name	Regressor	Basis functions	Mean function	Covariance function	Soft constraint	Hard constraint	Require training	Independen
LRR	Linear Regression	Spherical harmonics	N.A.	N.A.	Tickonov	Basis functions	No	freqs, mics
SP	Linear Regression	Spherical splines	N.A.	N.A.	Tickonov	Basis functions	No	freqs, mics
GP	GP	Spherical harmonics	0	k	Smoothness	Basis functions	Yes*	
	GP	Spherical harmonics	svect	k	Smoothness	Basis functions	Yes*	
NF	Neural Field	N.A.	N.A.	N.A.	Smoothness	N.A.	Yes	
PINN	Neural Field	N.A.	N.A.	N.A.	PDE	N.A.	Yes	
Inwards	Neural Field + GP	Spherical harmonics	0	k	Smoothness	Basis functions	Yes	
	Neural Field + GP	Spherical harmonics	svect	k	Smoothness	Basis functions	Yes	
Outwards	Neural Field + GP	Spherical harmonics	0	k	Smoothness	Basis functions	Yes	
			svect	k	Smoothness	Basis functions	Yes	

harmonics expansion and the noise variance ε.

We propose to use a NF with parameters θ to estimate the mapping for the problem input to the coefficients $\mathbf{c}$, that is,

$$\mathbf{c}(\mathbf{z}_{fij}) = \mathrm{NF}_\theta(\mathbf{z}_{fij}). \tag{47}$$

The architecture, depicted in **??**, features the sinusoidal positional encoding proposed with a MLP using hyperbolic tangent activation function as in [148].

a) Loss function: To optimize the model parameters we maximize following objective function,

$$\log p(\mathbf{c}, \ell_f, \alpha, \varepsilon \,|\, \mathbf{h}) = \log \mathcal{N}_\mathbb{C}\left(\mathbf{h}|\mathbf{z}, \mathbf{K}_\theta\right) \tag{48}$$

where $\mathbf{K}_\theta$ is the prior covariance matrix computed with the kernel k of Equation (43). The spherical harmonics spectrum (SHS) [176] indicates the average contribution made by harmonics of increasing order. The structure of SHS of HRTF shows downward trend [84], which correspond a smooth response in the spatial-source domain. Meanwhile the coefficients for both degree and order are generally sparse. To conference these behaviors we propose the following regularization term:

$$\mathcal{L}_\text{reg} = \lambda_\text{L1} \sum_{n,l} |\mathbf{c}_{n,l}| + \lambda_\text{exp} \sum_{n,l} \text{ReLU}(\mathbf{c}_{n,l+1} - \mathbf{c}_{n,l}), \tag{49}$$

where the l-th order SHS coefficient $\mathbf{c}_{n,l}$ is computed as [176]

$$\mathbf{c}_{n,l} = \sqrt{\textstyle\sum_m \|\mathbf{c}_{lm}(\mathbf{z}_n)\|_2^2/(2l+1)}. \tag{50}$$

The final loss function results in

$$\mathcal{L} = -\log p(\mathbf{c}, \ell_f, \alpha, \varepsilon \,|\, \mathbf{h}) + \mathcal{L}_\text{reg} \tag{51}$$

b) Initialization: The number and location of available measurement limits the order L of the spherical harmonic expansion, which lead to negligible spatial aliasing if $\kappa_f r \leq L$ [136]. A good *rule of thumb* state that a minimum of $D = (L+1)^2$ direction per frequency is need to resolve the expansion of order L. We use this empirical rule to initialize the SH coefficients c_{lm}. In practice, given D measurement, we pre-compute c_{lm}^SH for $l < \lfloor\sqrt{D} - 1\rfloor$ and we sum them to the estimation of the NF as in a residual connection,

$$c_{lm}(\mathbf{z}_n) = \begin{cases} c_{\theta,lm}(\mathbf{z}_n) + c_{lm}^\text{SH}(\mathbf{z}_n) & \text{if} \quad l < \lfloor\sqrt{D} - 1\rfloor \\ c_{\theta,lm}(\mathbf{z}_n) & \text{otherwise.} \end{cases} \tag{52}$$

where $c_{\theta,lm}(\mathbf{z}_n)$ are the output of the NF in Equation (47). To make the model continuous also in the frequency domain, a linear interpolation is used to interpolate the pre-computed coefficients c_{lm}^SH over unseen frequencies.

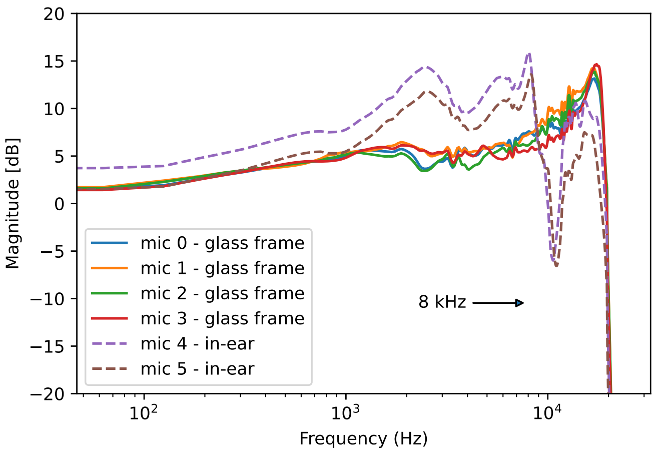

Fig. 8: Average magnitude spectrum (in dB) of the steering vectors for each microphone in the SPEAR array. Dashed-lines are used for in-ear microphones.

VI. Experimental Evaluation

This section reports the evaluation of the PC-NeuralSteerer for steering vector upsampling and downstream speech enhancement tasks. Qualitative results are available online.[2]

A. Datasets

We studied the performances of the proposed methods using the evaluation environment and data of the SPEAR Challenge [8], an extension of the EASYCOM dataset [177]. The latter contains real-world egocentric audio-visual data from several groups in a dynamic, noisy, and reverberant environment. It was recorded using AR glasses with four microphones, a camera, and binaural microphones. Clock-synchronized head-pose information and close-talking speech recordings are provided for all participants in the conversation. The main limitation is the lack of ground-truth binaural signals for objective evaluation. The SPEAR challenge overcomes this by digitally recreating of the EASYCOM environment, extending it to more acoustics environments and different conversation setups. Also, it provides ATFs of the microphone arrays and binaural microphones, which are measured on a manikin in an anechoic room for 1020 directions on a sphere. Hereafter, we will consider the six microphones as a single calibrated array for simplicity. Besides, while the geometrical position of the smart-glass microphones is provided, the calibration of the two in-ear microphones was

[2] PC-NeuralSterer demo webpage: https://diegodicarlo.com/demo/pcnsteerer.

done manually by the authors of this proposed work.

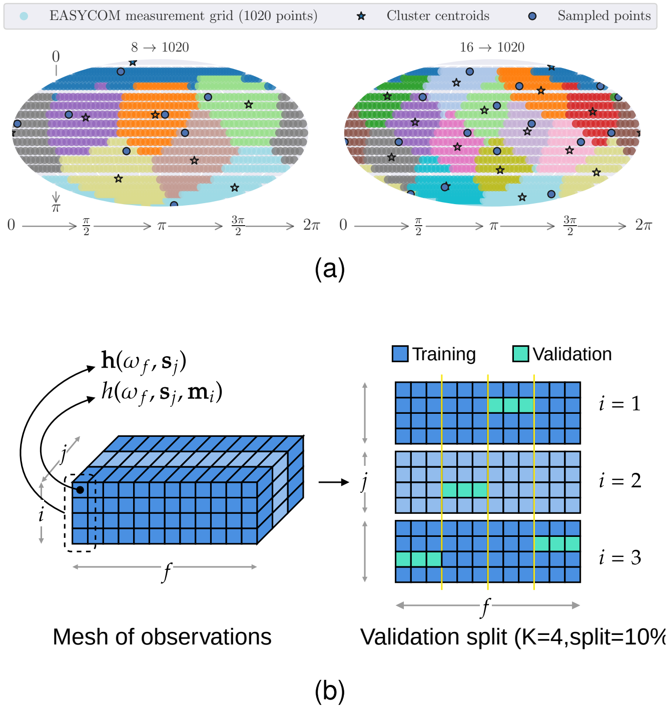

Fig. 9: Figure 9a shows the Mollweide projection of observed coordinates for two upsampling factors against the observation coordinates in the EASYCOM dataset. Stars and colors denote clusters and centroids for a quasi-uniform sampling. Figure 9b illustrates the sampling strategy to select the validation data.

B. Steering vector upsampling

a) Task and Data: Given a set of steering vectors observed at $N_{\text{obs}} \in \{8, 16, 32, 64, 128\}$ locations (dataset *train*), we aim to estimate the steering vectors for all the directions on the sphere, here represented by the 1020 measurements available in the SPEAR challenge dataset (dataset *test*). As common in sound field reconstruction, we consider the following performance metrics. The normalized mean squared error in decibels per frequency,

$$\text{nMSE}(f) = \frac{1}{IJ} \sum_{n=1}^{N} 10 \log_{10} \frac{|h_{fij} - \hat{h}_{fij}|^2}{|h_{fij}|^2} \quad [\text{dB}], \qquad (53)$$

captures the reconstruction error between the target and reconstructed steering vectors. To better quantify the phase reconstruction in the time domain and the spatial similarity of the filters, we also consider the cosine similarity between estimated and target filters in the time domain for the direction j,

$$\text{CSIM}(\Omega_j) = \frac{1}{TI} \sum_{ij} \frac{\sum_t h_{tji} \hat{h}_{tji}}{\sum_t h_{tji}^2 \sum_t \hat{h}_{tji}^2} \quad \in [-1, 1], \qquad (54)$$

where h_{tji} and $\hat{h}_{tij}$ are the time domain representation (*i.e.*, inverse Fourier transform) of h_{fij} and $\hat{h}_{fij}$, respectively.

The train observations were sampled randomly on the unit sphere with the following protocol: first, the 1020 test data were clustered using N_{obs} centroids corresponding to the N_{obs} points of Spherical Fibonacci mapping; then, for each centroid, one point is sampled uniformly among the one in the cluster. One exception is made for the cluster of the point corresponding to the frontal-axis direction (azimuth 0, elevation 0), for which the closest point to this direction in the datasets was always used. This approach could better represent a real user's and practitioner's sampling of a spherical space; besides, it avoids introducing bias in the downstream task of spatial filter in which users often point at their interlocutors. In a scenario where only a few directional measurements are available, 10% of the observation in a held-out fashion for model selection is restrictive, while k-fold cross-validation is demanding for iterative algorithms such as NF. Moreover, in the NF-based models, the training dataset is reshaped to match the continuous regression model over frequency, microphone, and source position. This introduced an imbalance in the validation dataset, with the frequency axis being more dense than the spatial coordinates. To address these issues, rather than sampling the frequency axis uniformly, we reduce the number of frequencies by 2 by taking every other frequency. We subdivide the resulting set into 8 equal portions for each direction and channel as illustrated in Figure 9b, and sample the 10% of these points. The number of clusters was empirically tuned.

b) Compared Methods: The nearest neighbor (NN) from the Scipy library, the regularized spherical cubic spline (SP) [124] adapted from the MNE library, and the regularized spherical harmonics (SH) interpolation method were used as baseline methods. The hyper-parameters for SP (smoothness coefficient 10^{-5}, number of Legendre term of 50, stiffness of 3) and SH (smoothing coefficient of 10^{-5}) were tuned on the held-out validation. In this work, we used a differentiable implementation of (complex) SH [178]. The order of spherical harmonics was automatically set according to the *rule-of-thumb* $L = \lfloor \sqrt{N} - 1 \rfloor$ where N is the number of available observed directions. For DNN-based techniques, the basic neural field (NF), the PINN and our previously proposed neural field with geometric warping (NF+GW) [123] shares the same backbone architecture of *sf*-PINN [163]: an MLP with 3 layers of 128 nodes and hyperbolic tangent ($\tanh$) as activation function; a sinusoidal positioning encoding with 128 features was used to project the input coordinates before the MLP. We applied non-dimensionalization to the input coordinates and pre-multiply each dimension by a gain factor to balance its resolution along the corresponding axis, $\mathbf{g} = [g_f, g_{\mathbf{s}}, g_{\mathbf{m}}] = [g_f, g_{\mathbf{s}_x}, g_{\mathbf{s}_y}, g_{\mathbf{s}_z}, g_{\mathbf{m}_x}, g_{\mathbf{m}_y}, g_{\mathbf{m}_z}] = [10, 1, 1, 1, 1, 1, 1]$. As for the optimization, we used the ADAM optimizers with a learning rate of 10^{-3} with an initial linear warm-up starting at 10^{-4} for 1000 steps and an exponential decay with a rate of 0.9 every 1000 steps and end-value of 10^{-5} were found to produce better results in initial empirical investigation. Previous investigations indicated that clipping the gradient norm to 1 and no weight decay regularization yielded better results, in line with the discussion in [173]. The same splitting of the frequency axis described in the paragraph above was applied to these models. A batch size of $B = 1024$ samples was used. All the models are implemented in the JAX library [179].

c) Proposed model configuration: The proposed GP-NSteerer was configured similarly to the NF-based models dis-

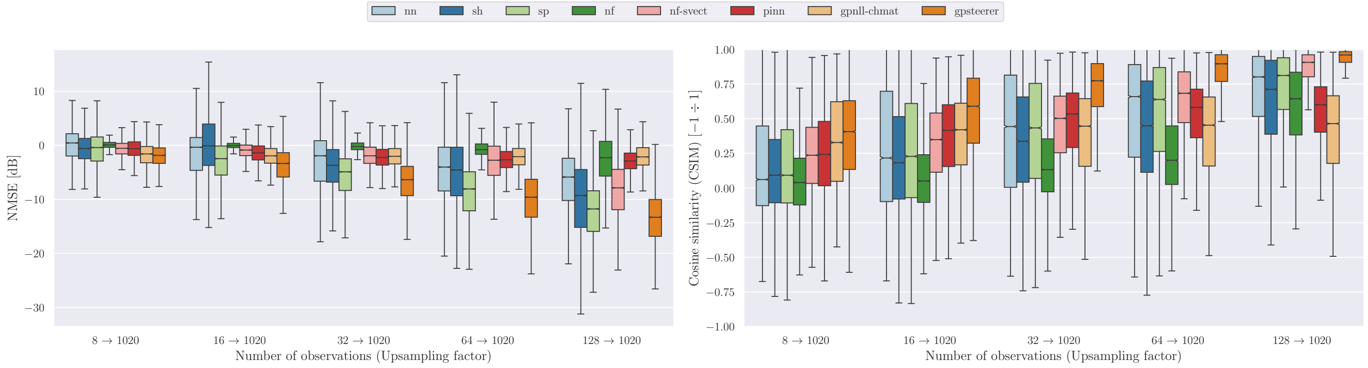

Fig. 10: Interpolation results: normalized Mean Squared Error (left) and cosine similarity per number of observed directions.

cussed above. While the latter performs a direct regression in the steering vector space, the GP-NSteerer predicts a parameterization of the kernel function, namely the spherical harmonics coefficients. Therefore, the positional encoding must match this space's underlying structure. We found the following parameterization to perform the best: learning rate starting and peak values are set to 10^{-3} and 10^{-2}, respectively; the number of layers of the MLP is 2; the gain for each input coordinates is $\mathbf{g}_f = \mathbf{g}_\mathbf{s} = 1, \mathbf{g}_\mathbf{m} = 100$. Prior investigations show that the performances do not change significantly by modulating the gain parameter for the frequencies and source position axes. This is probably due to the kernel formulation, which allows the control of the spectral bias along these axes. By contrast, the microphone position coordinates require higher input gain, probably due to *??? I don't have an explanation for this actually*. To compute the SH-based kernel, we used the differentiable complex-valued implementation available in [178] and the asymptotic expansion for large arguments of first-order Hankle function[3].

d) Experimental Evaluation:: This paragraph will compare the proposed method `NeuralGPSteerer` for steering vector spatial upsampling against the three classical baselines described earlier (`NN`, `SP`, `SH`), three NF-based approaches (`NF`, `NF-GW`, `PINN`) and one data-driven GP regression-based methods (`GP-Chmat`). The upsampling level includes 8, 16, 32, 64 to 1020 source positions. For each upsampling level, the results aggregate 3 different sampling of the source position.

The nMSE metric, defined in eq. (53), can be calculated for every frequency index and then averaged; similarly, the CSIM metric defined in eq. (54). Figure 10 shows the average results for these two metrics over the 3 configurations. As expected, the performances of all the methods decrease with the sparsity of the observations. Nonetheless, it is clear to see the benefit of the proposed approach over the compared methods, both in low and higher spatial sampling regimes. Among the baseline methods, `SP` yields better spatial interpolation results as the performance of SH-based interpolation is affected by the measurements' location. As a purely data-driven method with a limited dataset, a vanilla NF is not able to produce a reliable approximation even if one could notice the interpolation ability of a basic MLP as relative performances increase with the number of observations. As expected, the introduction of prior knowledge on the problem improves the results: thanks to the geometric warping [123] the NF is able to produce better results in terms of spatial coherence (CSIM), but poor performances in terms of nMSE. While baseline methods are designed to produce optimum results in MSE terms, the NF-based approaches using (physical or geometrical) regularization do not significantly improve nMSE but are physically feasible solutions. The results of the PINNs show a similar behavior. While the PDE-based regularization helps the spatial coherences and in low data regimes, it performs poorly in terms of nMSE compared with the baselines. Interestingly, the PINNs approach and the NF-GW have comparable average performances for interpolation factors lower than 64 → 1020. Compared to NF-CW, PINN performances saturate with more measurements, leading to unsatisfactory performances. This is probably due to the difficulty of balancing data-driven and task-driven losses in multi-objective optimization and exploring the hyperparameter space of the model. In fact, the PDE in crefeq:helmholtz used as a regularization term ineq. (41) is agnostic to any boundary effects or initial condition, meaning that the optimization could also erroneously tend to the anechoic solution of Equation (2) The GP-based regression using a purely data-driven covariance matrix shares the same trend behavior as the PINNs. The native GP's smoothness property, together with the physical constraints of the steering vectors, helps in the low measurement regimes but saturates when more measurements are available.

To better understand the results, illustrative examples of one configuration are given in Figure 11, where the real part of the frequency-domain sound field related to one channel is reported for different upsampling factors. It can be seen that both SP and SH interpolation schemes produce a smooth spatial field, while the introduction of physics (and geometric) knowledge produces solutions that align better with the algebraic solution of the problem. Interestingly, the GP-based regression provides a qualitatively closer solution to the algebraic steering vectors, which is used as a kernel function to generate the covariances. However, a simple kernel based on the chordal distance (see Equation (23)) is not able to capture the complex distribution of the scattering effects. The proposed method GP-Steerer overcomes this limitation with a kernel based on a mixture of SH. This approach produces a good balance between a data-driven and physical-driven solution that attempts to dis-

[3] https://dlmf.nist.gov/10.17

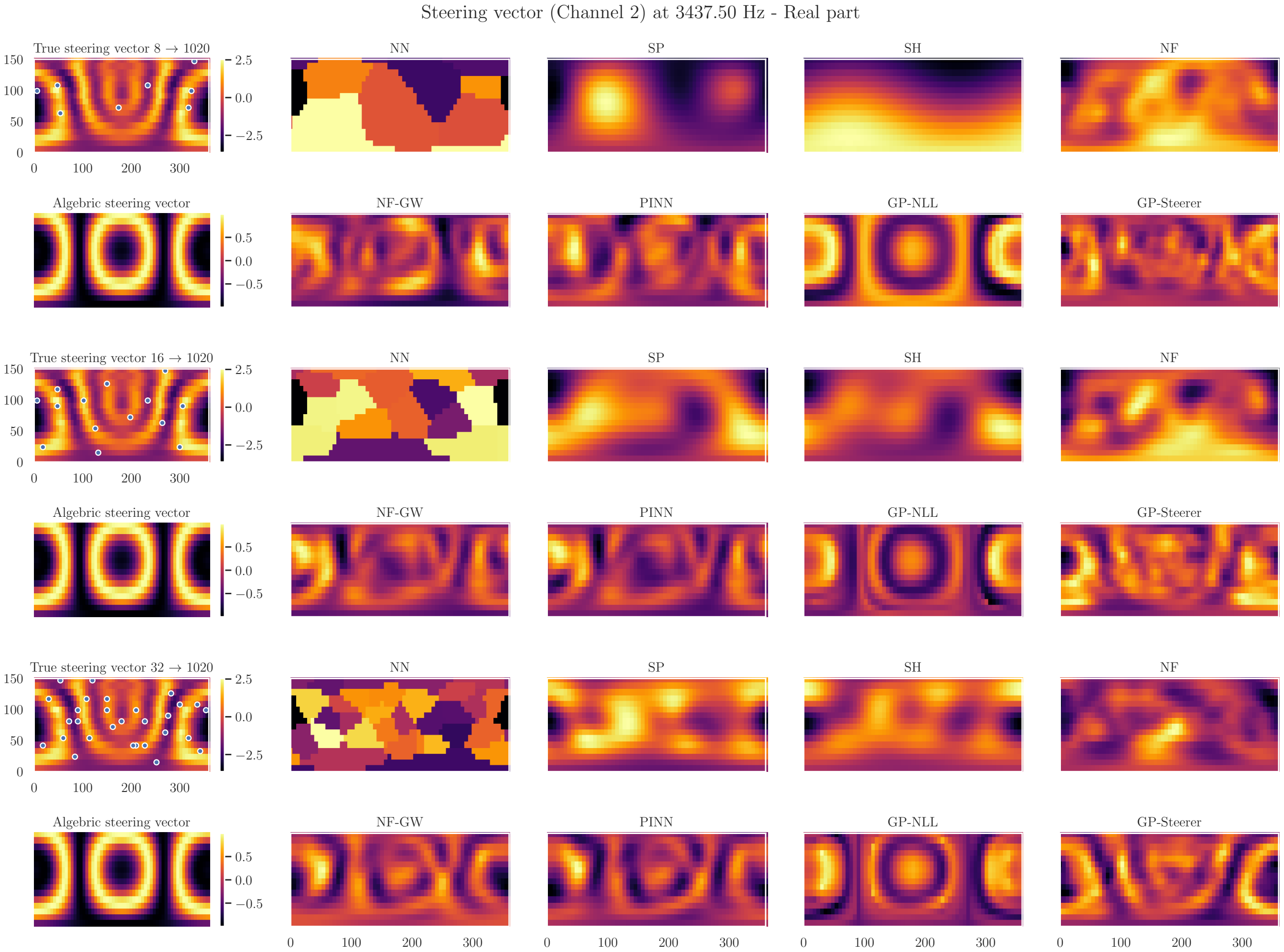

Fig. 11: Qualitative results of interpolation

tribute the modeled field correctly and the very challenging scenario of only 8 measurements.

More quantitative insights on the interpolation capabilities of the proposed methods are provided in Figure 12 (top). Here, the CSIM is plotted against angular distance computed as the central angle $\Delta\alpha_{jj'} = |\alpha_j - \alpha_{j'}| = 2\arcsin\left(\frac{C_{jj'}}{2}\right)$, where $C_{jj'}$ is the chordal distance as in eq. (24). One can identify two behaviors: the baselines NN, SH, and SP are good local interpolators, while knowledge-driven learning-based methods (NF-GW, PINNs, GP-Chmat) outperform at bigger distances. The baselines do not alter the observed data and provide a good local interpolation retaining 0.8 of CSIM for angular distance $\Delta\alpha < 10°$. This fact is particularly interesting in beamforming downstream applications when the frontal angle is predominantly used in face-to-face conversation and a single steering vector at elevation 0° and azimuth 0° could cover a roster spanning 20° size. Besides, the performances drop rather quickly with the angular distances. In contrast, knowledge-driven learning-based methods (NF-GW, PINNs, and GP-Chmat) produce solutions that outperform the baselines at greater distances, indicating the benefits of prior knowledge could “inpaint” the missing data. However, as a major drawback, these learning methods alter the observed data. Especially for deep learning-based models such as NF, it is very difficult to maintain unaltered observations. The proposed approach produces results that are the best of two worlds: a principled GP regression framework can retain the distorted observed measurements, leading to a local interpolation that outperforms the baseline methods; secondly, thanks to the data-driven approach supported by physical constraints, the best (or at least comparable) results are obtained when reconstructing unknown distant locations.

The nMSE metrics are averaged per configuration and reported as a function of frequency $F = 385$ positive frequency bins[4] in Figure 12 (bottom). The performance generally deteriorates with increasing frequencies, consistent with the known behavior that relates the number of spatial observations to the maximum resolved frequencies found in spatial acoustics [136]. The SH baseline provides good interpolation of the low frequencies. In particular, with 64 observations, these baseline methods can effectively interpolate frequencies below 2 kHz with less than −15 dB of nMSE. However, this effect is countered by a poor reconstruction in high frequencies; a positive nMSE in dB

[4] This choice of this value is due to the downstream task

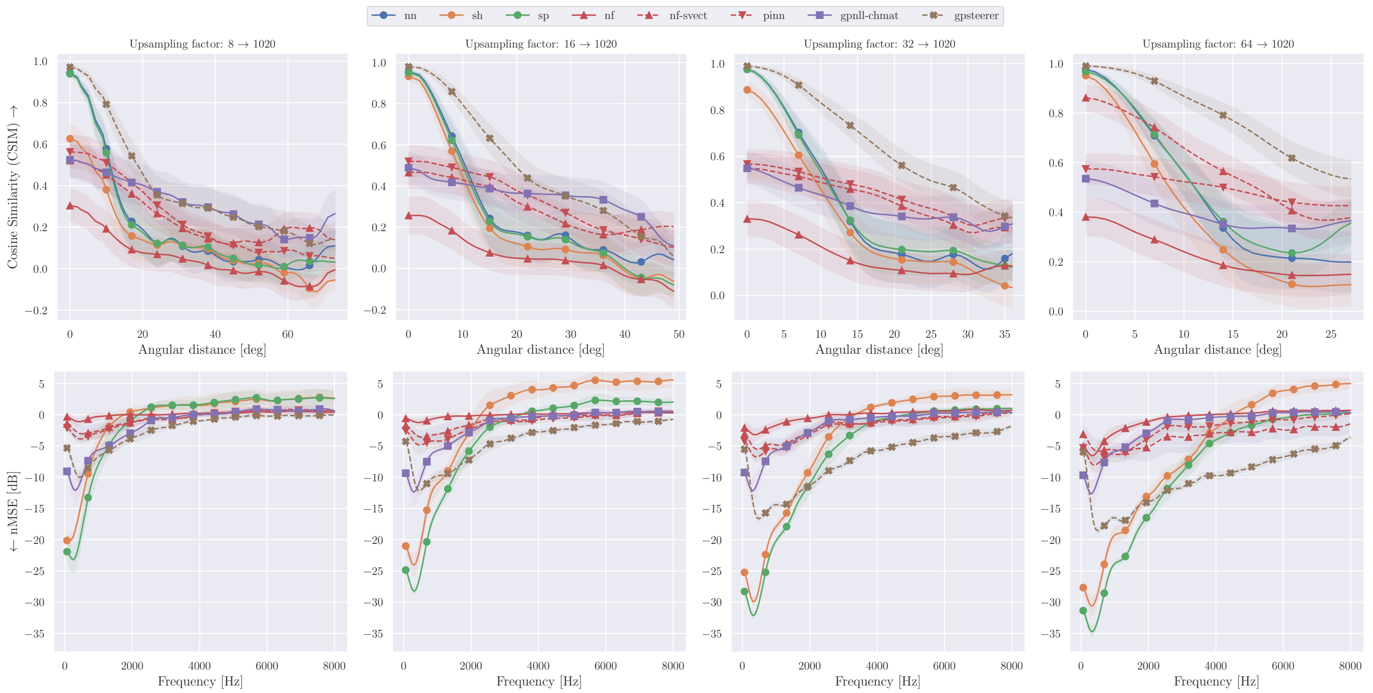

Fig. 12: Interpolation results: (top) Cosine similarity (CSIM) versus (angular) distance from an observed sample. (bottom) Normalized Mean Squared Error (nMSE) in dB versus frequency.

scale at high scales indicates the introduction of noisy artifacts components at small scales, suggesting overfitting. SP interpolation method also provides good interpolation at low frequencies without introducing spurious components. The other compared methods report performances higher that do not go below $-10\,\mathrm{dB}$. Besides the poor relative results, these values are in line with the one reported in sound-field reconstruction works: The comparative studies in [10] reported performances in the above this value for frequencies above $1.4\,\mathrm{kHz}$ and authors of [105] report value values is $[-10, 5]$ dB for reconstruction of random sound field from sparse measurements. The proposed model outperforms the baselines for higher frequencies than $2\,\mathrm{kHz}$ confirming the approach's effectiveness. For frequencies below this value, the average reconstruction error is below $-10\,\mathrm{dB}$, demonstrating the usefulness of interpolating precomputed low SH coefficients. Curiously, all the methods exhibit poor reconstruction for frequencies below $20\,\mathrm{kHz}$. Notice that, while the baseline methods perform regression for each frequency independently with $F = 385$. The NF, NF-GW, and PINN methods model the frequency axis continuously and use portions of this axis for validation. In contrast, due to the computational complexity of constructing covariance matrices of the training set, GP-based methods (GP-Chmat and GP-NeuralSteeer) use $F = 127$. One can then notice the effectiveness of the interpolation over the frequency axis using the kernel function proposed by [68] and used in the proposed methods.

C. Downstream task: Speech enhancement

We conducted experiments using the SPEAR challenge dataset to study steering vector interpolation methods in terms of speech enhancement performances.

...

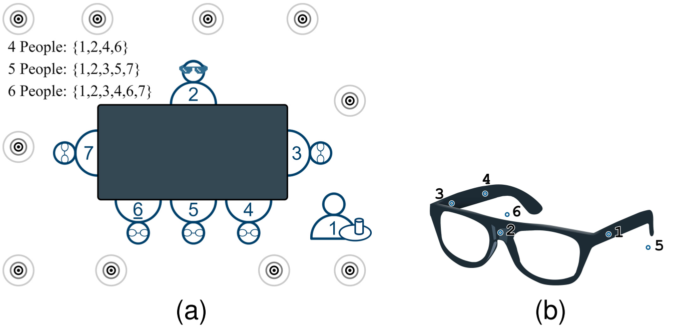

Fig. 13: The SPEAR Challenge (a) scene setup and (b) the microphone array worn by person number 2 [8], [177].

a) Beamformer design: Let $\mathbf{x}(t, f) = [x_1(f, t), \ldots, x_I(f, t)]^{\mathsf{T}} \in \mathbb{C}^{I \times 1}$ denotes the vector of observed signal $x_i(f, t)$ at time frame t, frequency index t, microphone index i for total I microphones. The beamformer output is

$$r(t, f) = \mathbf{w}^{\mathsf{H}}(t, f)\mathbf{x}(t, f), \tag{55}$$

where $\mathbf{w} \in \mathbb{C}^{I \times 1}$ is the beamformer weights. For notation simplicity, (t, f) will be omitted for the remaining of the paper unless specified.

The weights of the MVDR beamformer can be derived as [14],

$$\mathbf{w} = \left(\mathbf{d}^{\mathsf{H}}\mathbf{R}^{-1}\mathbf{d}\right)^{-1}\mathbf{R}^{-1}\mathbf{d}, \tag{56}$$

where $d = \mathbf{h}(\Omega_s) \in \mathbb{C}^{I \times 1}$ is the steering vector for the target DOA Ω_s, $\mathbf{R} \in \mathbb{C}^{I \times I}$ is the noise covariance matrix.

The Isotropic-MVDR (Iso-MVDR), also known as a superdirective beamformer, assumes stationary spherically isotropic noise covariance matrix [14] and is the baseline method in

Method	SI-SDR [dB]					fwSegSNR [dB]					PESQ [$0 \div 5$]					MBSTOI [$0 \div 1$]				
N. Obs.	8	16	32	64	128	8	16	32	64	128	8	16	32	64	128	8	16	32	64	128
Passthrough	-12.198					4.136					1.084					0.428				
NN	-9.353	-8.884	-8.400	-7.644	-7.455	**4.913**	**4.901**	5.071	5.342	5.337	1.131	1.136	1.133	1.137	1.141	0.503	0.509	0.513	0.531	0.546
SP	**-8.297**	**-7.786**	**-7.303**	**-7.074**	**-7.068**	4.873	4.845	**5.061**	**5.355**	**5.454**	1.138	1.138	1.141	1.144	1.142	**0.530**	**0.542**	**0.547**	**0.559**	**0.564**
SH	-10.692	-7.903	-7.393	-7.210	-7.120	4.526	4.687	4.978	5.262	5.389	1.113	1.135	1.137	1.142	1.143	0.487	0.531	0.544	0.553	0.559
NF	-16.781	-18.065	-15.761	-15.955	-13.843	3.266	3.164	3.220	3.223	3.633	1.118	1.115	1.118	1.129	1.145	0.421	0.391	0.422	0.417	0.425
NF+GW [123]	-10.554	-10.817	-10.295	-11.992	-10.216	4.386	4.374	4.542	4.218	4.326	1.123	1.119	1.120	1.133	1.148	0.518	0.506	0.504	0.460	0.489
PINN	-10.891	-10.770	-10.730	-10.680	-10.480	4.330	4.349	4.454	4.371	4.421	1.124	1.123	1.126	1.125	1.124	0.505	0.509	0.500	0.499	0.512
GP-Steerer	-9.455	-9.801	-8.381	-7.614	-7.349	4.695	4.413	4.761	5.054	5.290	**1.148**	**1.167**	**1.177**	**1.183**	**1.177**	0.495	0.505	0.519	0.530	0.539
Baseline [8]	-5.146					4.982					1.111					0.603				

TABLE II: Enhancement results: Median values for signal-based (SI-SDR, fwSegSNR) and perceptual (PESQ, MBSTOI) objective metrics available in the SPEAR Challenge evaluation environment.

the SPEAR challenge. The associated noise covariances matrix writes [9, Sect. 2.1])

$$\mathbf{R} = \sum_{j \in \mathcal{J}} w_j \mathbf{a}(\Omega_j)\mathbf{a}^{\mathsf{H}}(\Omega_j), \tag{57}$$

where w_j is the quadrature weight for each sample point given by

$$w_j = \frac{2 \sin \vartheta}{N_\varphi N_\vartheta} \sum_{m=0}^{N_\vartheta/2-1} \frac{\sin((2m+1)\vartheta_j)}{2m+1} \tag{58}$$

in which ϑ_j is the inclination of the j-th direction, and the number of directions in azimuth and inclination is N_φ and N_ϑ, respectively.

The SPEAR challenge baseline uses all the 1020 available steering vectors to compute the $\mathbf{R}$. Given a method for steering vector upsampling, such method is used to evaluate $\mathbf{a}(\Omega_j)$ on the target DOA Ω_j and to construct $\mathbf{R}$ by evaluating $\mathbf{a}$ at the same 1020 location as the baseline. Note that $\mathbf{R}$ is invariant to the source position and time index, so it can be precomputed beforehand.

b) Data: Our research utilizes the simulated datasets from the SPEAR Challenge (D2, D3, D4), which offer binaural reference recordings to compute objective metrics. Among all the metrics implemented in the evaluation environment, we selected the following. As reported in [8], the perceptual metric MBSTOI and the signal level frequency-wise SNR demonstrated the best correlation with user-subjective evaluations. Additionally, we included the ISR in our analysis to quantitatively assess the spatial coherence of the outputs from the proposed methods.

D. Discussion

E. Conclusion

F. Future work

RTF vs ATF, frequency as latent variables

VII. Conclusion

The conclusion goes here.

Acknowledgments

This should be a simple paragraph before the References to thank those individuals and institutions who have supported your work on this article.

Appendix
Complex Spherical Harmonics

Maybe I write here the complex spherical harmonics and differentiable implementation pseudo-code?

References

References

[1] R. M. Corey, "Microphone array processing for augmented listening," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2019.

[2] T. Betlehem, W. Zhang, M. A. Poletti, and T. D. Abhayapala, "Personal sound zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 81–91, 2015.

[3] M. Cobos, J. Ahrens, K. Kowalczyk, and A. Politis, "An overview of machine learning and other data-based methods for spatial audio capture, processing, and reproduction," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, p. 10, 2022.

[4] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith *et al.*, "Project aria: A new tool for egocentric multi-modal ai research," *arXiv preprint arXiv:2308.13561*, 2023.

[5] E. H. A. De Haas and L.-H. Lee, "Deceiving audio design in augmented environments: a systematic review of audio effects in augmented reality," in *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 2022, pp. 36–43.

[6] G. Kailas and N. Tiwari, "Design for immersive experience: Role of spatial audio in extended reality applications," in *Design for Tomorrow—Volume 2: Proceedings of ICoRD 2021*. Springer, 2021, pp. 853–863.

[7] J. Herskovitz, J. Wu, S. White, A. Pavel, G. Reyes, A. Guo, and J. P. Bigham, "Making mobile augmented reality applications accessible," in *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 2020, pp. 1–14.

[8] V. Tourbabin, P. Guiraud, S. Hafezi, P. A. Naylor, A. H. Moore, J. Donley, and T. Lunner, "The spear challenge-review of results," in *Proc Forum Acusticum*, 2023.

[9] S. Hafezi, A. H. Moore, P. Guiraud, P. A. Naylor, J. Donley, V. Tourbabin, and T. Lunner, "Subspace hybrid beamforming for head-worn microphone arrays," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] S. Koyama, J. G. Ribeiro, T. Nakamura, N. Ueno, and M. Pezzoli, "Physics-informed machine learning for sound field estimation," *arXiv preprint arXiv:2408.14731*, 2024.

[11] 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.

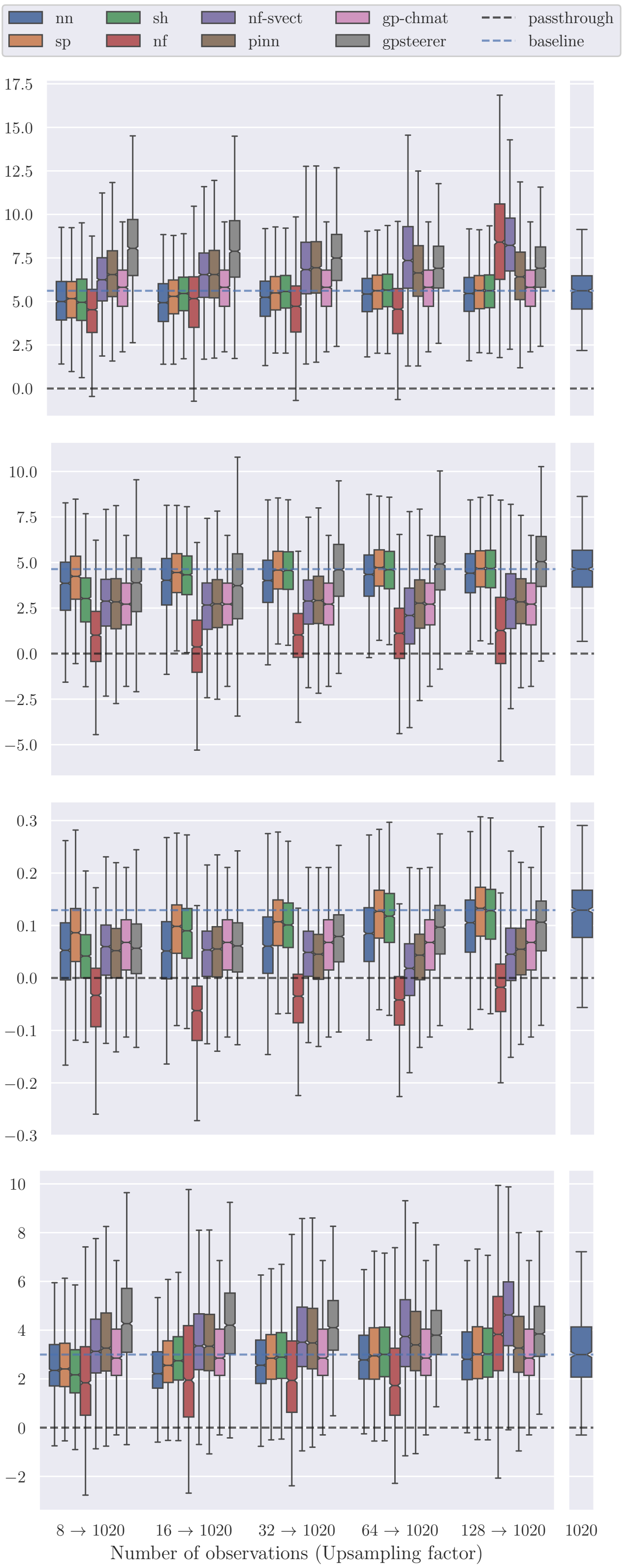

Fig. 14: Enhancement results: SDR, ISR, and SAR improvements in dB per number of observed directions relative to the passthrough.

[12] F. Ma, T. D. Abhayapala, P. N. Samarasinghe, and X. Chen, "Physics informed neural network for head-related transfer function upsampling," *arXiv preprint arXiv:2307.14650*, 2023.

[13] M. Pezzoli, F. Antonacci, and A. Sarti, "Implicit neural representation with physics-informed neural networks for the reconstruction of the early part of room impulse responses," *arXiv preprint arXiv:2306.11509*, 2023.

[14] 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.

[15] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[16] K. Sekiguchi, A. A. Nugraha, Y. Du, Y. Bando, M. Fontaine, and K. Yoshii, "Direction-aware adaptive online neural speech enhancement with an augmented reality headset in real noisy conversational environments," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 9266–9273.

[17] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2002.

[18] G. Chardon, "Theoretical analysis of beamforming steering vector formulations for acoustic source localization," *Journal of Sound and Vibration*, vol. 517, p. 116544, 2022.

[19] W. Zhang, P. N. Samarasinghe, H. Chen, and T. D. Abhayapala, "Surround by sound: A review of spatial audio recording and reproduction," *Applied Sciences*, vol. 7, no. 5, p. 532, 2017.

[20] R. O. Schmidt, "Multilinear array manifold interpolation," *IEEE transactions on signal processing*, vol. 40, no. 4, pp. 857–866, 1992.

[21] X. Karakonstantis and E. Fernandez-Grande, "Generative adversarial networks with physical sound field priors," *The Journal of the Acoustical Society of America*, vol. 154, no. 2, pp. 1226–1238, 2023.

[22] A. Levi and H. F. Silverman, "An alternate approach to adaptive beamforming using srp-phat," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 2726–2729.

[23] V. Valimaki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1421–1448, 2012.

[24] M. Jälmby, F. Elvander, and T. Van Waterschoot, "Low-rank room impulse response estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 957–969, 2023.

[25] A. Ratnarajah, I. Ananthabhotla, V. K. Ithapu, P. Hoffmann, D. Manocha, and P. Calamia, "Towards improved room impulse response estimation for speech recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[26] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Transactions on speech and audio processing*, vol. 13, no. 1, pp. 135–143, 2004.

[27] Y. Zhang, Y. Wang, and Z. Duan, "Hrtf field: Unifying measured hrtf magnitude representation with neural fields," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[28] C. Pörschmann, J. M. Arend, D. Bau, and T. Lübeck, "Comparison of spherical harmonics and nearest-neighbor based interpolation of head-related transfer functions," in *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2020.

[29] V. Bruschi, L. Grossi, N. A. Dourou, A. Quattrini, A. Vancheri, T. Leidi, and S. Cecchi, "A review on head-related transfer function generation for spatial audio," *Applied Sciences*, vol. 14, no. 23, p. 11242, 2024.

[30] 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.

[31] K. Hartung, J. Braasch, and S. J. Sterbing, "Comparison of different methods for the interpolation of head-related transfer functions," in *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*. Audio Engineering Society, 1999.

[32] D. R. Begault and L. J. Trejo, *3-D sound for virtual reality and multimedia*. San Diego, CA, USA: Academic Press Professional, Inc., 2000.

[33] H. Gamper, "Head-related transfer function interpolation in azimuth, elevation, and distance," *The Journal of the Acoustical Society of America*, vol. 134, no. 6, pp. EL547–EL553, 2013.

[34] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, and A. Reyes-Lecuona, "3d tune-in toolkit: An open-source library for real-time binaural spatialisation," *PloS one*, vol. 14, no. 3, p. e0211899, 2019.

[35] Z. Ben-Hur, D. Alon, P. W. Robinson, and R. Mehra, "Localization of virtual sounds in dynamic listening using sparse hrtfs," in *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2020.

[36] A. Srivastava, G. Routray, and R. M. Hegde, "Spatial hrtf interpolation using spectral phase constraints," in *2020 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2020, pp. 1–5.

[37] A. Acosta, F. Grijalva, R. Álvarez, and B. Acuña, "Bilinear and triangular spherical head-related transfer functions interpolation on non-uniform meshes," in *2020 IEEE ANDESCON*. IEEE, 2020, pp. 1–6.

[38] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1637–1647, 1992.

[39] J. Chen, B. D. Van Veen, and K. E. Hecox, "A spatial feature extraction and regularization model for the head-related transfer function," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 439–452, 1995.

[40] V. Larcher, O. Warusfel, J.-M. Jot, and J. Guyard, "Study and comparison of efficient methods for 3-D audio spatialization based on linear decomposition of hrtf data," in *Audio Engineering Society Convention 108*. Audio Engineering Society, 2000.

[41] F. P. Freeland, L. W. Biscainho, and P. S. Diniz, "Efficient hrtf interpolation in 3d moving sound," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society, 2002.

[42] L. Wang, F. Yin, and Z. Chen, "Head-related transfer function interpolation through multivariate polynomial fitting of principal component weights," *Acoustical Science and Technology*, vol. 30, no. 6, pp. 395–403, 2009.

[43] B.-S. Xie, "Recovery of individual head-related transfer functions from a small set of measurements," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 282–294, 2012.

[44] M. Zhang, Z. Ge, T. Liu, X. Wu, and T. Qu, "Modeling of individual hrtfs based on spatial principal component analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 785–797, 2020.

[45] J. C. Torres and M. R. Petraglia, "Hrtf interpolation in the wavelet transform domain," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2009, pp. 293–296.

[46] Z. Jiang, J. Sang, C. Zheng, A. Li, and X. Li, "Modeling individual head-related transfer functions from sparse measurements using a convolutional neural network," *The Journal of the Acoustical Society of America*, vol. 153, no. 1, pp. 248–259, 2023.

[47] I. D. Gebru, D. Marković, A. Richard, S. Krenn, G. A. Butler, F. De la Torre, and Y. Sheikh, "Implicit hrtf modeling using temporal convolutional networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3385–3389.

[48] T.-Y. Chen, T.-H. Kuo, and T.-S. Chi, "Autoencoding hrtfs for dnn based hrtf personalization using anthropometric features," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 271–275.

[49] J. W. Lee, S. Lee, and K. Lee, "Global hrtf interpolation via learned affine transformation of hyper-conditioned features," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[50] R. Miccini and S. Spagnol, "Hrtf individualization using deep learning," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2020, pp. 390–395.

[51] B. Zhi, D. N. Zotkin, and R. Duraiswami, "Towards fast and convenient end-to-end hrtf personalization," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 441–445.

[52] P. Siripornpitak, I. Engel, I. Squires, S. J. Cooper, and L. Picinali, "Spatial up-sampling of hrtf sets using generative adversarial networks: A pilot study," *Frontiers in Signal Processing*, vol. 2, p. 904398, 2022.

[53] A. Hogg, H. Liu, M. Jenkins, and L. Picinali, "Exploring the impact of transfer learning on gan-based hrtf upsampling," in *Proc. EAA Forum Acusticum, Eur. Congress on Acoust*, 2023.

[54] F. Lluis, P. Martinez-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.

[55] M. S. Kristoffersen, M. B. Møller, P. Martínez-Nuevo, and J. Østergaard, "Deep sound field reconstruction in real rooms: introducing the isobel sound field dataset," *arXiv preprint arXiv:2102.06455*, 2021.

[56] 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.

[57] F. Miotello, L. Comanducci, M. Pezzoli, A. Bernardini, F. Antonacci, and A. Sarti, "Reconstruction of sound field through diffusion models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1476–1480.

[58] J. W. Lee and K. Lee, "Neural fourier shift for binaural speech rendering," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[59] R. Duraiswami and V. C. Raykar, "The manifolds of spatial hearing," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 3. IEEE, 2005, pp. iii–285.

[60] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *International journal of neural systems*, vol. 25, no. 01, p. 1440003, 2015.

[61] F. Grijalva, L. C. Martini, D. Florencio, and S. Goldenstein, "Interpolation of head-related transfer functions using manifold learning," *IEEE Signal Processing Letters*, vol. 24, no. 2, pp. 221–225, 2017.

[62] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the audio engineering society*, vol. 45, no. 6, pp. 456–466, 1997.

[63] A. Franck, W. Wang, and F. M. Fazi, "Sparse ℓ_1-optimal multiloudspeaker panning and its relation to vector base amplitude panning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 996–1010, 2017.

[64] J. Chen, B. D. Van Veen, and K. E. Hecox, "Synthesis of 3d virtual auditory space via a spatial feature extraction and regularization model," in *Proceedings of IEEE Virtual Reality Annual International Symposium*. IEEE, 1993, pp. 188–193.

[65] T. Nishino, S. Kajita, K. Takeda, and F. Itakura, "Interpolating head related transfer functions in the median plane," in *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA'99 (Cat. No. 99TH8452)*. IEEE, 1999, pp. 167–170.

[66] S. Carlile, C. Jin, and J. Leung, "Performance measures of the spatial fidelity of virtual auditory space: Effects of filter compression and spatial sampling," in *Proceedings of the International Conference on Auditory Display (ICAD 2002)*, 2002.

[67] R. L. Jenison and K. Fissell, "A spherical basis function neural network for modeling auditory space," *Neural computation*, vol. 8, no. 1, pp. 115–128, 1996.

[68] Y. Luo, D. N. Zotkin, H. Daume, and R. Duraiswami, "Kernel regression for head-related transfer function interpolation and spectral extrema extraction," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 256–260.

[69] E. Thuillier, C. Jin, and V. Välimäki, "Hrtf interpolation using a spherical neural process meta-learner," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[70] T. Lübeck, J. M. Arend, and C. Pörschmann, "Spatial upsampling of sparse spherical microphone array signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1163–1174, 2023.

[71] G. Del Galdo, O. Thiergart, T. Weller, and E. A. Habets, "Generating virtual microphone signals using geometrical information gathered by distributed arrays," in *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays*. IEEE, 2011, pp. 185–190.

[72] 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 *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1537–1541.

[73] Y. Haneda, S. Makino, Y. Kaneda, and N. Koizumi, "Arma modeling of a room transfer function at low frequencies," *Journal of the Acoustical Society of Japan (E)*, vol. 15, no. 5, pp. 353–355, 1994.

[74] P. Runkle, M. Blommer, and G. Wakefield, "A comparison of head related transfer function interpolation methods," in *Proceedings of 1995 workshop on applications of signal processing to audio and accoustics*. IEEE, 1995, pp. 88–91.

[75] K. Watanabe, S. Takane, and Y. Suzuki, "Interpolation of head-related transfer functions based on the common-acoustical-pole and residue model," *Acoustical science and technology*, vol. 24, no. 5, pp. 335–337, 2003.

[76] G. Ramos and M. Cobos, "Parametric head-related transfer function modeling and interpolation for cost-efficient binaural sound applications," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 1735–1738, 2013.

[77] B. Al-Sheikh, M. A. Matin, and D. J. Tollin, "Head related transfer function interpolation based on finite impulse response models," in *2019 Seventh International Conference on Digital Information Processing and Communications (ICDIPC)*. IEEE, 2019, pp. 8–11.

[78] M. gebru2021implicitroz and G. H. De Sousa, "Structured iir models for hrtf interpolation." in *ICMC*, 2010.

[79] P. Nowak and U. Zölzer, "Spatial interpolation of hrtfs approximated by parametric iir filters," in *Proc. DAGA*, 2022.

[80] Y. Masuyama, G. Wichern, F. G. Germain, Z. Pan, S. Khurana, C. Hori, and J. Le Roux, "Niirf: Neural iir filter field for hrtf upsampling and personalization," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1016–1020.

[81] N. Murata, S. Koyama, H. Kameoka, N. Takamune, and H. Saruwatari, "Sparse sound field decomposition with multichannel extension of complex nmf," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 345–349.

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

[83] T. Ajdler, C. Faller, L. Sbaiz, and M. Vetterli, "Sound field analysis along a circle and its application to hrtf interpolation," *Journal of the Audio Engineering Society*, vol. 56, no. 3, pp. 156–175, 2008.

[84] M. J. Evans, J. A. Angus, and A. I. Tew, "Analyzing head-related transfer function measurements using surface spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 104, no. 4, pp. 2400–2411, 1998.

[85] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and range extrapolation of hrtfs [head related transfer functions]," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. IEEE, 2004, pp. iv–iv.

[86] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, "Regularized hrtf fitting using spherical harmonics," in *2009 IEEE workshop on applications of signal processing to audio and acoustics*. IEEE, 2009, pp. 257–260.

[87] J. Ahrens, M. R. Thomas, and I. Tashev, "Hrtf magnitude modeling using a non-regularized least-squares fit of spherical harmonics coefficients on incomplete data," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–5.

[88] C. Pörschmann, J. M. Arend, and F. Brinkmann, "Directional equalization of sparse head-related transfer function sets for spatial upsampling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1060–1071, 2019.

[89] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3616–3627, 2018.

[90] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely, "Efficient representation and sparse sampling of head-related transfer functions using phase-correction based on ear alignment," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2249–2262, 2019.

[91] J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing spherical harmonics interpolation of time-aligned head-related transfer functions," *Journal of the Audio Engineering Society*, vol. 69, no. 1/2, pp. 104–117, 2021.

[92] P. Samarasinghe, T. Abhayapala, M. Poletti, and T. Betlehem, "An efficient parameterization of the room transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2217–2227, 2015.

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

[94] 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.

[95] S. Damiano, F. Borra, A. Bernardini, F. Antonacci, and A. Sarti, "Soundfield reconstruction in reverberant rooms based on compressive sensing and image-source models of early reflections," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 366–370.

[96] N. Ueno, S. Koyama, and H. Saruwatari, "Kernel ridge regression with constraint of helmholtz equation for sound field interpolation," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 1–440.

[97] J. G. Ribeiro, S. Koyama, R. Horiuchi, and H. Saruwatari, "Sound field estimation based on physics-constrained kernel interpolation adapted to environment," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[98] R. Mignot, G. Chardon, and L. Daudet, "Low frequency interpolation of room impulse responses using compressed sensing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 205–216, 2013.

[99] N. Bertin, L. Daudet, V. Emiya, and R. Gribonval, "Compressive sensing in acoustic imaging," in *Compressed Sensing and its Applications: MATHEON Workshop 2013*. Springer, 2015, pp. 169–192.

[100] 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.

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

[102] Y. Ito, T. Nakamura, S. Koyama, and H. Saruwatari, "Head-related transfer function interpolation from spatially sparse measurements using autoencoder with source position conditioning," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2022, pp. 1–5.

[103] H. Bi and T. D. Abhayapala, "Point neuron learning: a new physics-informed neural network architecture," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 56, 2024.

[104] G. D. Romigh, R. M. Stern, D. S. Brungart, and B. D. Simpson, "A bayesian framework for the estimation of head-related transfer functions," *Journal of the Acoustical Society of America*, vol. 137, no. 4_Supplement, pp. 2323–2323, 2015.

[105] D. Caviedes-Nozal, N. A. Riis, F. M. Heuchel, J. Brunskog, P. Gerstoft, and E. Fernandez-Grande, "Gaussian processes for sound field reconstruction," *The Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 1107–1119, 2021.

[106] D. Caviedes-Nozal and E. Fernandez-Grande, "Spatio-temporal bayesian regression for room impulse response reconstruction with spherical waves," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[107] X. Feng, J. Cheng, S. Chen, and Y. Shen, "Room impulse response reconstruction using pattern-coupled sparse bayesian learning with spherical waves," *IEEE Signal Processing Letters*, 2024.

[108] F. Ma, S. Zhao, and I. S. Burnett, "Sound field reconstruction using a compact acoustics-informed neural network," *The Journal of the Acoustical Society of America*, vol. 156, no. 3, pp. 2009–2021, 2024.

[109] 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)*. IEEE, 2022, pp. 1–5.

[110] X. Chen, F. Ma, A. Bastine, P. Samarasinghe, and H. Sun, "Sound field estimation around a rigid sphere with physics-informed neural network," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 1984–1989.

[111] M. Olivieri, X. Karakonstantis, M. Pezzoli, F. Antonacci, A. Sarti, and E. Fernandez-Grande, "Physics-informed neural network for volumetric sound field reconstruction of speech signals," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 42, 2024.

[112] X. Karakonstantis, D. Caviedes-Nozal, A. Richard, and E. Fernandez-Grande, "Room impulse response reconstruction with physics-informed deep learning," *The Journal of the Acoustical Society of America*, vol. 155, no. 2, pp. 1048–1059, 2024.

[113] M. Middleton, D. T. Murphy, and L. Savioja, "The application of fourier neural operator networks for solving the 2d linear acoustic wave equation," in *Proceedings of Forum Acusticum, European Acoustics Association, Turin, Italy*, 2023, pp. 1–8.

[114] N. Borrel-Jensen, S. Goswami, A. P. Engsig-Karup, G. E. Karniadakis, and C.-H. Jeong, "Sound propagation in realistic interactive 3d scenes with parameterized sources using deep neural operators," *Proceedings of the National Academy of Sciences*, vol. 121, no. 2, p. e2312159120, 2024.

[115] F. Miotello, F. Terminiello, M. Pezzoli, A. Bernardini, F. Antonacci, and A. Sarti, "A physics-informed neural network-based approach for the spatial upsampling of spherical microphone arrays," in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2024, pp. 215–219.

[116] A. O. Hogg, M. Jenkins, H. Liu, I. Squires, S. J. Cooper, and L. Picinali, "Hrtf upsampling with a generative adversarial network using a gnomonic

equiangular projection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[117] T. Carpentier, H. Bahu, M. Noisternig, and O. Warusfel, "Measurement of a head-related transfer function database with high spatial resolution," in *7th forum acusticum (EAA)*, 2014.

[118] S. Li and J. Peissig, "Measurement of Head-Related Transfer Functions: A Review," *Applied Sciences*, vol. 10, no. 14, p. 5014, Jan. 2020.

[119] B. Xie, *Head-related transfer function and virtual auditory display*. J. Ross Publishing, 2013.

[120] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in neural information processing systems*, vol. 33, pp. 7537–7547, 2020.

[121] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar, "Neural fields in visual computing and beyond," in *Computer Graphics Forum*, vol. 41, no. 2. Wiley Online Library, 2022, pp. 641–676.

[122] X. Chen, F. Ma, Y. Zhang, A. Bastine, and P. N. Samarasinghe, "Head-related transfer function interpolation with a spherical cnn," *arXiv preprint arXiv:2309.08290*, 2023.

[123] D. Di Carlo, A. A. Nugraha, M. Fontaine, Y. Bando, and K. Yoshii, "Neural steerer: Novel steering vector synthesis with a causal neural field over frequency and direction," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 740–744.

[124] F. Perrin, J. Pernier, O. Bertrand, and J. F. Echallier, "Spherical splines for scalp potential and current density mapping," *Electroencephalography and clinical neurophysiology*, vol. 72, no. 2, pp. 184–187, 1989.

[125] Y. Luo, D. N. Zotkin, and R. Duraiswami, "Gaussian process data fusion for heterogeneous hrtf datasets," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[126] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "Ddsp: Differentiable digital signal processing," *arXiv preprint arXiv:2001.04643*, 2020.

[127] B. Bernschütz, A. V. Giner, C. Pörschmann, and J. Arend, "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 972–983, 2014.

[128] Z. Ben-Hur, D. L. Alon, B. Rafaely, and R. Mehra, "Loudness stability of binaural sound with spherical harmonic representation of sparse head-related transfer functions," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, pp. 1–14, 2019.

[129] J. M. Arend and C. Pörschmann, *Spatial upsampling of sparse head-related transfer function sets by directional equalization-influence of the spherical sampling scheme*. Universitätsbibliothek der RWTH Aachen Aachen, Germany, 2019.

[130] D. Bau, J. M. Arend, and C. Pörschmann, "Estimation of the optimal spherical harmonics order for the interpolation of head-related transfer functions sampled on sparse irregular grids," *Frontiers in Signal Processing*, vol. 2, p. 884541, 2022.

[131] F. Brinkmann and S. Weinzierl, "Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition," in *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2018.

[132] E. F. Grande, "Sound field reconstruction in a room from spatially distributed measurements," in *23rd International Congress on Acoustics*. German Acoustical Society (DEGA), 2019, pp. 4961–68.

[133] M. Hahmann and E. Fernandez-Grande, "A convolutional plane wave model for sound field reconstruction," *The Journal of the Acoustical Society of America*, vol. 152, no. 5, pp. 3059–3068, 2022.

[134] J. M. Schmid, E. Fernandez-Grande, M. Hahmann, C. Gurbuz, M. Eser, and S. Marburg, "Spatial reconstruction of the sound field in a room in the modal frequency range using bayesian inference," *The Journal of the Acoustical Society of America*, vol. 150, no. 6, pp. 4385–4394, 2021.

[135] R. Sibson, "A brief description of natural neighbour interpolation," *Interpreting multivariate data*, pp. 21–36, 1981.

[136] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. Academic press, 1999.

[137] D. Veerababu and P. K. Ghosh, "Neural network-based approach for solving problems in plane wave duct acoustics," *Journal of Sound and Vibration*, vol. 585, p. 118476, 2024.

[138] S. Wang, X. Yu, and P. Perdikaris, "When and why pinns fail to train: A neural tangent kernel perspective," *Journal of Computational Physics*, vol. 449, p. 110768, 2022.

[139] R. A. Kennedy, P. Sadeghi, T. D. Abhayapala, and H. M. Jones, "Intrinsic limits of dimensionality and richness in random multipath fields," *IEEE Transactions on Signal processing*, vol. 55, no. 6, pp. 2542–2556, 2007.

[140] C. Pörschmann and J. M. Arend, "A method for spatial upsampling of directivity patterns of human speakers by directional equalization," *Proceedings of the 45th DAGA*, pp. 1458–1461, 2019.

[141] Z. Chen, I. D. Gebru, C. Richardt, A. Kumar, W. Laney, A. Owens, and A. Richard, "Real acoustic fields: An audio-visual room acoustics dataset and benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 886–21 896.

[142] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ace challenge—corpus description and performance evaluation," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

[143] A. Ratnarajah, Z. Tang, R. Aralikatti, and D. Manocha, "Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 924–933.

[144] A. Ratnarajah, S.-X. Zhang, M. Yu, Z. Tang, D. Manocha, and D. Yu, "Fast-rir: Fast neural diffuse room impulse response generator," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 571–575.

[145] S. Lee, H.-S. Choi, and K. Lee, "Yet another generative model for room impulse response estimation," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2023, pp. 1–5.

[146] D. Sundström, F. Elvander, and A. Jakobsson, "Estimation of impulse responses for a moving source using optimal transport regularization," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 921–925.

[147] L. Kelley, D. Di Carlo, A. A. Nugraha, M. Fontaine, Y. Bando, and K. Yoshii, "Rir-in-a-box: Estimating room acoustics from 3d mesh data through shoebox approximation," in *INTERSPEECH*, 2024.

[148] A. Luo, Y. Du, M. Tarr, J. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3165–3177, 2022.

[149] Z. Lan, C. Zheng, Z. Zheng, and M. Zhao, "Acoustic volume rendering for neural impulse response fields," *arXiv preprint arXiv:2411.06307*, 2024.

[150] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu, "Av-nerf: Learning neural fields for real-world audio-visual scene synthesis," *Advances in Neural Information Processing Systems*, vol. 36, pp. 37 472–37 490, 2023.

[151] K. Su, M. Chen, and E. Shlizerman, "Inras: Implicit neural representation for audio scenes," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8144–8158, 2022.

[152] A. Manikas, *Differential geometry in array processing*. Imperial College Press, 2004.

[153] L. Qiong, G. Long, and Y. Zhongfu, "An overview of self-calibration in sensor array processing," in *6th International SYmposium on Antennas, Propagation and EM Theory, 2003. Proceedings. 2003*. IEEE, 2003, pp. 279–282.

[154] B. Friedlander, "Antenna array manifolds for high-resolution direction finding," *IEEE Transactions on Signal Processing*, vol. 66, no. 4, pp. 923–932, 2017.

[155] N. A. Gumerov and R. Duraiswami, *Fast multipole methods for the Helmholtz equation in three dimensions*. Elsevier, 2005.

[156] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 2, no. 3.

[157] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[158] Y. Luo, D. N. Zotkin, and R. Duraiswami, "Gaussian process models for hrtf based 3d sound localization," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2858–2862.

[159] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[160] G. Yang, S. Belongie, B. Hariharan, and V. Koltun, "Geometry processing with neural fields," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 483–22 497, 2021.

[161] 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.

[162] N. Benbarka, T. Höfer, A. Zell *et al.*, "Seeing implicit neural representations as fourier series," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2041–2050.

[163] J. C. Wong, C. C. Ooi, A. Gupta, and Y.-S. Ong, “Learning in sinusoidal spaces with physics-informed neural networks,” *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 3, pp. 985–1000, 2022.

[164] F. Szatkowski, K. J. Piczak, P. Spurek, J. Tabor, and T. Trzciński, “Hypernetworks build implicit neural representations of sounds,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2023, pp. 661–676.

[165] A. Richard, D. Markovic, I. D. Gebru, S. Krenn, G. A. Butler, F. Torre, and Y. Sheikh, “Neural synthesis of binaural speech from mono audio,” in *International Conference on Learning Representations*, 2021.

[166] A. Richard, P. Dodds, and V. K. Ithapu, “Deep impulse responses: Estimating and parameterizing filters with deep networks,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3209–3213.

[167] T. Lobato and R. Sottek, “A process for calibrating hrtfs based on differentiable implicit representations and domain adversarial learning,” in *2024 32nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2024, pp. 271–275.

[168] J. Nam, J. S. Abel, and J. O. Smith III, “A method for estimating interaural time difference for binaural synthesis,” in *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

[169] F. M. Rohrhofer, S. Posch, C. Gößnitzer, and B. C. Geiger, “On the apparent pareto front of physics-informed neural networks,” *IEEE Access*, 2023.

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

[171] V. Dharanalakota and P. K. Ghosh, “Loss-based optimizer switching to solve 1-d helmholtz equation using neural networks,” *The Journal of the Acoustical Society of America*, vol. 154, no. 4_supplement, pp. A98–A98, 2023.

[172] C. Wu, M. Zhu, Q. Tan, Y. Kartha, and L. Lu, “A comprehensive study of non-adaptive and residual-based adaptive sampling for physics-informed neural networks,” *Computer Methods in Applied Mechanics and Engineering*, vol. 403, p. 115671, 2023.

[173] S. Wang, S. Sankaran, H. Wang, and P. Perdikaris, “An expert’s guide to training physics-informed neural networks,” *arXiv preprint arXiv:2308.08468*, 2023.

[174] D. Di Carlo, D. Heitz, and T. Corpetti, “Post processing sparse and instantaneous 2d velocity fields using physics-informed neural networks,” in *Proceedings of the 20th International Symposium on Application of Laser and Imaging Techniques to Fluid Mechanics*, 2022.

[175] X. Karakonstantis and E. F. Grande, “Room impulse response reconstruction using physics-constrained neural networks,” in *10th Convention of the European Acoustics Association*. European Acoustics Association, 2023.

[176] H. N. Pollack, S. J. Hurter, and J. R. Johnson, “Heat flow from the earth’s interior: analysis of the global data set,” *Reviews of Geophysics*, vol. 31, no. 3, pp. 267–280, 1993.

[177] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, “Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments,” *arXiv preprint arXiv:2107.04174*, 2021.

[178] F. Bigi, G. Fraux, N. J. Browning, and M. Ceriotti, “Fast evaluation of spherical harmonics with sphericart,” *J. Chem. Phys.*, no. 159, p. 064802, 2023.

[179] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, “JAX: composable transformations of Python+NumPy programs,” 2018. [Online]. Available: http://github.com/jax-ml/jax

Biography Section

If you have an EPS/PDF photo (graphicx package needed), extra braces are needed around the contents of the optional argument to biography to prevent the LaTeX parser from getting confused when it sees the complicated `\includegraphics` command within an optional argument. (You can create your own custom macro containing the `\includegraphics` command to make things simpler here.)

If you will not include a photo:

John Doe Use `\begin{IEEEbiographynophoto}` and the author name as the argument followed by the biography text.