DOI: 10.61782/fa.2025.0517

FORUM ACUSTICUM
EURONOISE 2025

LEARNING MAGNITUDE DISTRIBUTION OF SOUND FIELDS VIA CONDITIONED AUTOENCODER

Shoichi Koyama*
National Institute of Informatics
Tokyo, Japan

Kenji Ishizuka
Yamaha Corporation
Shizuoka, Japan

ABSTRACT

A learning-based method for estimating the magnitude distribution of sound fields from spatially sparse measurements is proposed. Estimating the magnitude distribution of acoustic transfer function (ATF) is useful when phase measurements are unreliable or inaccessible and has a wide range of applications related to spatial audio. We propose a neural-network-based method for the ATF magnitude estimation. The key feature of our network architecture is the input and output layers conditioned on source and receiver positions and frequency and the aggregation module of latent variables, which can be interpreted as an autoencoder-based extension of the basis expansion of the sound field. Numerical simulation results indicated that the ATF magnitude is accurately estimated with a small number of receivers by our proposed method.

Keywords: *acoustic transfer function, deep neural networks, autoencoder, magnitude distribution, spatial interpolation*

1. INTRODUCTION

Sound field estimation, interpolation, capturing, or reconstruction is a fundamental problem in acoustic signal processing and machine learning, which aim to estimate a spatial distribution of acoustic field from a discrete set of sensor observations. The estimation of room impulse response (RIR) and acoustic transfer function (ATF, the frequency-domain representation of RIR) is a special case of the sound field estimation problem, where RIRs or ATFs between unknown sources (loudspeakers) and receivers (microphones) are estimated from those between known sources and receivers in a room. Such estimation allows us to know the acoustic characteristics without measurements and can be applied to various techniques using RIRs and ATFs, such as speech enhancement, visualization, and spatial audio.

There have been many studies on the estimation of RIRs and complex-valued ATFs. One of the most widely used techniques is the basis-expansion-based method, which is based on the expansion of ATF into plane wave functions, spherical wave functions, or equivalent sources [1–3]. The method based on kernel ridge regression with the constraint that the estimated function satisfies the Helmholtz equation is a generalization of the basis-expansion-based method to infinite-dimensional expansion [4,5]. Recently, the learning-based method, especially the neural-network-based (NN-based) method, has attracted attention because it allows us to estimate the sound field with extremely few microphones, compared with the methods not using training data [6–8].

In contrast to many ATF estimation methods, which target the estimation of complex amplitude in the frequency domain, this study deals with the problem of estimating the ATF magnitude, i.e., the absolute value of the complex amplitude in the frequency domain. The ATF magnitude estimation technique will be useful when the reference signal for RIR measurement from the source is unreliable or cannot be output in the first plane. Specifically, this is the case when the signals of each receiver are not synchronized or when estimating the directivity of musical instruments or other vibrating bodies. However, unlike the methods using the complex amplitudes,

**Corresponding author*: koyama.shoichi@ieee.org.

it is difficult for the ATF magnitude estimation to incorporate the physical properties of the sound field because the governing equation of magnitude distribution cannot be well represented.

We propose a learning-based method based on an autoencoder conditioned on the source position, receiver position, and frequency for estimating the ATF magnitude distribution. In many current learning-based methods, ATFs at fixed target positions are estimated from ATFs at observation positions as their subset [6, 9]. Therefore, it is difficult to estimate ATF at arbitrary positions and to combine multiple datasets with different measurement setups. Although there are no such restrictions with the method based on implicit neural representation or neural field (NF) [10, 11], the network needs to be retrained during inference, which is computationally intensive. Our approach, based on the conditioned autoencoder, overcomes these shortcomings. Our proposed method can be considered a nonlinear extension of the basis-expansion-based method and is an extension of the method we proposed for estimating the magnitude of head-related transfer function (HRTF) [12].

2. PROBLEM STATEMENT

ATF from the source position $\boldsymbol{y} \in \mathbb{R}^3$ to the receiver position $\boldsymbol{x} \in \mathbb{R}^3$ at the angular frequency $\omega \in \mathbb{R}$ is denoted as $h(\boldsymbol{x}, \boldsymbol{y}, \omega)$ (i.e., $h : \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R} \to \mathbb{C}$). Suppose that ATF magnitude in logarithmic scale is denoted as $a(\boldsymbol{x}, \boldsymbol{y}, \omega) := 20 \log_{10} |h(\boldsymbol{x}, \boldsymbol{y}, \omega)|$, and it is measured at multiple positions, which are sparsely distributed inside the target region $\Omega \subset \mathbb{R}^3$. Our goal is to estimate a of the given source position $\boldsymbol{y}$ at the target positions $\{\boldsymbol{x}_n^{(\mathrm{t})}\}_{n=1}^N$ inside Ω from its sparse measurements at the measurement positions $\{\boldsymbol{x}_m^{(\mathrm{m})}\}_{m=1}^M$ ($M \ll N$).

In practice, the log-ATF magnitude is obtained for multiple source positions and discrete frequency bins. By defining the indexes of sources and frequencies as $l \in \{1, \dots, L\}$ and $f \in \{1, \dots, F\}$, respectively, the discrete value of the log-ATF magnitude is denoted as $a_{m,l,f}$ for the measurement positions and $a_{n,l,f}$ for the target positions, where L and F are the number of sources and frequency bins, respectively.

3. PRIOR WORK

We here briefly introduce prior work on RIR and ATF estimation. Many previous studies focus on the estimation of the spatial distribution of RIR or complex-valued ATF. There are few studies on the estimation of magnitude distribution, with a few exceptions [6, 9, 13].

3.1 Basis-expansion-based methods

In the basis-expansion-based methods, the ATF complex amplitude $h(\boldsymbol{x}, \boldsymbol{y}, \omega)$ is approximated as the linear combination of basis function φ_j as

$$h(\boldsymbol{x}, \boldsymbol{y}, \omega) \approx \sum_{j=1}^{J} \gamma_j(\boldsymbol{y}, \omega) \varphi_j(\boldsymbol{x}, \boldsymbol{y}, \omega), \tag{1}$$

where j is the index, J is the number of basis functions, and γ_j is the jth expansion coefficient. For the basis function φ_j, plane wave function, spherical wave function, and equivalent source are typically used because these functions are the element solution of the Helmholtz equation [1–3]. For the measurement positions $\{\boldsymbol{x}_m^{(\mathrm{m})}\}_{m=1}^M$, Eq. (1) is rewritten as

$$\boldsymbol{h} \approx \boldsymbol{\Phi} \boldsymbol{\gamma}, \tag{2}$$

where $\boldsymbol{h} \in \mathbb{C}^M$, $\boldsymbol{\Phi} \in \mathbb{C}^{M \times N}$, and $\boldsymbol{\gamma} \in \mathbb{C}^N$ consist of $\{h(\boldsymbol{x}_m^{(\mathrm{m})}, \boldsymbol{y}, \omega)\}_{m=1}^M$, $\{\varphi_j(\boldsymbol{y}, \omega)\}_{j=1}^J$, and $\{\gamma_j(\boldsymbol{y}, \omega)\}_{j=1}^J$, respectively.

The expansion coefficients $\{\gamma_j\}_{j=1}^J$ are generally estimated by solving the following linear regression problem:

$$\underset{\boldsymbol{\gamma} \in \mathbb{C}^J}{\text{minimize}} \, \|\boldsymbol{h} - \boldsymbol{\Phi}\boldsymbol{\gamma}\|^2 + \lambda \|\boldsymbol{\gamma}\|_2^2, \tag{3}$$

where $\lambda \in \mathbb{R}_{\geq 0}$ is the regularization parameter. The expansion coefficients are simply obtained as

$$\hat{\boldsymbol{\gamma}} = \left(\boldsymbol{\Phi}^{\mathsf{H}} \boldsymbol{\Phi} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{\Phi}^{\mathsf{H}} \boldsymbol{h}. \tag{4}$$

The ATF at the target positions can be obtained by substituting the estimated expansion coefficients into Eq. (1).

Kernel ridge regression is also frequently used in the sound field estimation problem, and it can be seen as a generalization of the basis-expansion-based methods to infinite-dimensional expansion. In [4, 5], the kernel ridge regression with the constraint that the estimated function satisfies the Helmholtz equation has been proposed.

3.2 Neural-network-based methods

Deep neural networks (DNNs) have also been applied to the sound field estimation problem because of their high adaptability and representational power [6–8]. There are

several approaches to apply DNNs to the ATF estimation problem. One approach is to set the DNN output to discretized values of ATF at the target positions [6, 14]. The DNN input is the observations at the measurement positions chosen from the target positions. Thus, the DNN parameters are optimized by minimizing a loss function defined for a pair of observation and target ATF data.

The other approach is the use of implicit neural representation or NF, where the continuous function h is represented by DNNs [15, 16]. The input and output of DNN are the argument of h, i.e., $\boldsymbol{x}$, and the scalar function value approximating $h(\boldsymbol{x}) \in \mathbb{C}$. The implicit neural representation is typically used in PINNs by incorporating the regularization term that evaluates the deviation from the governing equation [17, 18].

The main drawback of the first approach is the predefined target positions. Since the DNN model is trained to estimate the ATFs at the fixed target positions from the measurements at their subset positions, it is difficult to combine multiple datasets measured on different setups. The second approach, the NF-based approach, makes it possible to estimate the ATF at an arbitrary target position. However, the NF-based methods are usually applied to estimate the ATF only with a single observation. To apply the NF-based method as a learning-based method, the NF-based DNN model is trained by using the training data, and the DNN parameters are optimized again for the test data, which generally requires high computational cost.

4. PROPOSED METHOD

In the basis-expansion-based methods described in Section 3.1, the sound field is expanded by predefined spatial basis functions, and the expansion coefficients are obtained by linear regression using the measurements. The expansion coefficients are generally independent of the receiver position, whereas the basis functions depend on the receiver position.

We propose a learning-based method for estimating the log-ATF magnitude, which is based on the DNN trained by $a_{n,l,f}$ for multiple source positions. Our proposed method is regarded as a nonlinear extension of the basis-expansion-based method by using an autoencoder consisting of the encoder and decoder conditioned on the receiver positions and the receiver-position-independent latent variables, i.e., the input of the decoder. A similar technique has been applied to the HRTF upsampling [12].

4.1 Model

Figure 1 shows the proposed network architecture, which consists of an encoder, an aggregation module, and a decoder. Based on the insights described above, the encoder and decoder weights depend on receiver positions, whereas the decoder input is independent of receiver positions. In addition, the encoder is conditioned on the source positions, frequency, and the number of observations, and the decoder is conditioned on the source positions and frequency.

The detailed network architecture is almost the same as the one proposed in [12]. The weights and biases of the encoder and decoder are independently generated by NN with the input of the conditioning vector. The conditioning vector is a concatenation of the receiver positions, source positions, angular frequencies, and number of observations (only for the encoder). The elements of the conditioning vector are normalized by their maximum values. Then, Fourier feature mapping (FFM) [19] is applied as a preprocessing step. The weight/bias generators consist of fully connected layers, layer normalization, and Mish nonlinearity.

The encoder transforms the sparse log-ATF magnitudes $a_{m,l,f}$ into latent variables $z_{m,d,l,f}$, which consists of layer normalization, the Mish nonlinearity, and hyper-linear layers [20]. Then, the latent variables, $z_{m,d,l,f}$, are averaged over the measurement positions as

$$\bar{z}_{d,l,f} = \frac{1}{M} \sum_{m=1}^{M} z_{m,d,l,f}, \tag{5}$$

which is referred to as prototypes as used in the prototypical network [21].

The decoder generates the log-ATF magnitudes $\hat{a}_{n,l,f}$ at the target positions from the prototypes $\bar{z}_{d,l,f}$, which consists of layer normalization, the Mish nonlinearity, and hyper-linear layers.

4.2 Loss function

The loss function $\mathcal{L}$ for the training is defined as the mean log-spectral distortion (LSD) of the ATF as

$$\mathcal{L} = \frac{1}{NL} \sum_{n,l} \sqrt{\frac{1}{F} \sum_{f} \left(\hat{a}_{n,l,f} - a_{n,l,f}\right)^2}. \tag{6}$$

5. EXPERIMENTAL EVALUATIONS

We evaluated the proposed method, comparing it with the method based on the kernel ridge regression and NF-based

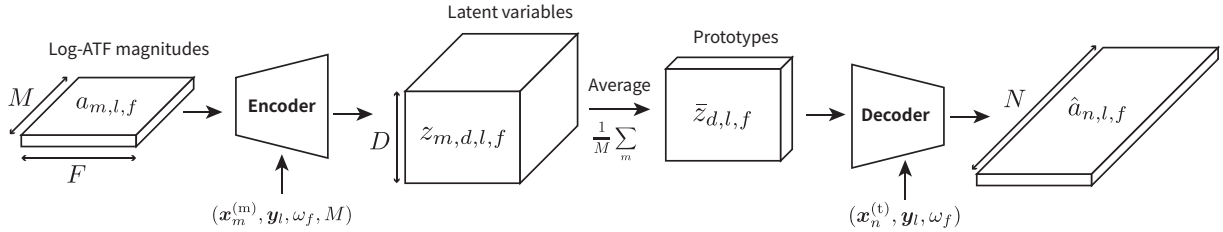


Figure 1: Outline of proposed network architecture.

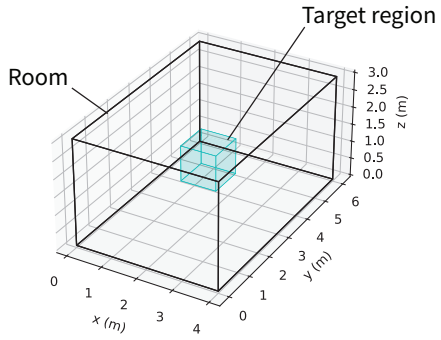


Figure 2: Experimental setting.

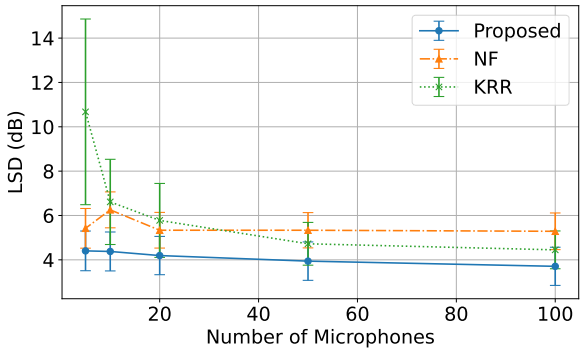


Figure 3: Average LSD with respect to the number of microphones.

method.

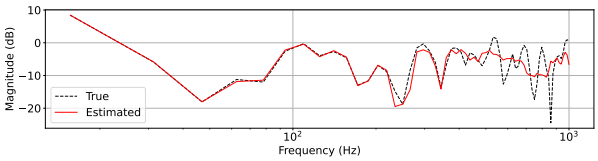
5.1 Experimental setup

We generated an ATF dataset of a single room by using acoustic simulation based on the image source method [22]. As shown in Fig. 2, the shoebox-shape room with the size of 4.0 m × 6.0 m × 3.0 m was assumed. The cuboid target region was set at the center of the room, and its size was 1.0 m × 1.0 m × 1.0 m. The 1331 target positions were obtained by discretizing the target region every 0.1 m. The 1024 source positions were randomly chosen within the room, excluding the target region. The RIRs were generated by setting the target reverberation time to 200 ms by using pyroomacoustics [23]. The sampling frequency was 2000 Hz, so the maximum frequency of the target was 1000 Hz. The log-ATF magnitudes were obtained by truncating the RIR into 128 samples and Fourier transform. The ATF dataset of the 1024 source positions was separated into training, validation, and test data, which contain 820, 122, and 122 log-ATF magnitudes, respectively. We investigated $M = 5$, 10, 20, and 100 for the number of measurements, and the measurement positions were randomly chosen from the target positions.

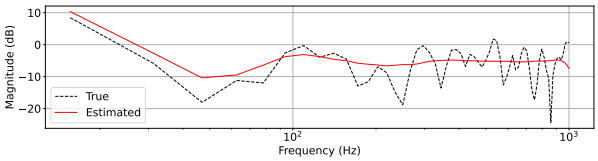
The proposed DNN was trained for 1400 epochs using Adam optimizer [24]. The learning rate was 10^{-3} for the first 800 epochs, 10^{-4} for 801-1200 epochs, and 10^{-5} for 1201-1400 epochs. The number of measurements M used for the training was larger than that for the validation and test. For example, the network for $M = 5$ was trained for $M = 5$, 10, 20, and 100, but that for $M = 100$ was trained only for 100. The used DNN for evaluation was at the epoch with the smallest validation loss.

In the kernel ridge regression (KRR), the log-ATF magnitudes at the target positions were estimated from the observations only in the test data because KRR is not a learning-based method. For the kernel function, we used Gaussian kernel [25] because the constraint of the Helmholtz equation is not valid for the magnitude measurements. The precision (the inverse of the variance) in the Gaussian kernel was set to 10^{-2}. The regularization parameter in KRR was set to 10^{-3}.

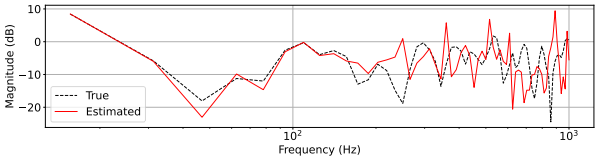
The DNN model for the NF-based method had the input of the source position, receiver position, and angu-



(a) Proposed

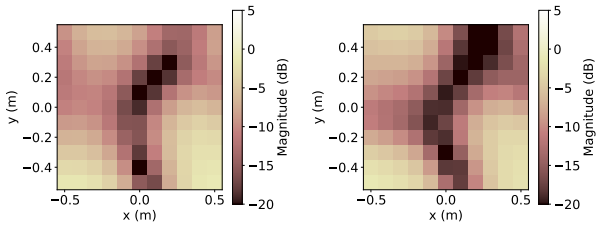


(b) NF

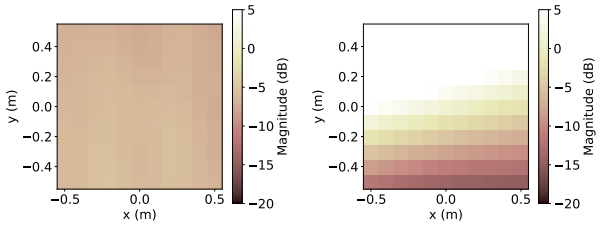


(c) KRR

Figure 4: Estimated ATF magnitude at the center of the target region when $M = 5$.



(a) Ground truth (b) Proposed



(c) NF (d) KRR

Figure 5: Magnitude distribution on x-y-plane at $z = 0$ at 250 Hz when $M = 5$.

lar frequency and the output of the log-ATF magnitude. The input was normalized by their maximum values, and FFM was applied. Then, three fully connected layers of 128 neurons and the ReLu activation function [26] were applied. The loss function was the LSD of the ATF magnitude. The network was trained on the training data with 1400 epochs and validated and tested by fine-tuning the entire network with 10 epochs, using Adam optimizer. The learning rate was 10^{-5}.

5.2 Experimental results

Figure 3 shows the average LSD with the standard deviation on the test data with respect to the number of measurements M. The LSD of KRR deteriorates rapidly as the number of observations decreases because KRR does not use the training data. The NF-based method showed nearly constant LSDs. The LSDs of NF were lower than those of KRR for $M = 5$, 10, and 20, but higher for $M = 50$ and 100. The LSDs of the proposed method were the lowest for all the numbers of observations. Therefore, the learning-based methods are effective when the number of observations is small, owing to the extraction of ATF properties from the training data. Since the proposed method showed high performance in both cases where the number of observations is small and large, the proposed model can successfully learn the characteristics of ATF.

The true and estimated ATF magnitudes at the center of the target region for the case of $M = 5$ are shown in Fig. 4. The estimated ATF of KRR fluctuated to a large extent, especially above 300 Hz. The ATF magnitude of NF was like a smoothed version of the true ATF. Although the LSD was smaller than that of KRR, it did not capture the fine structure of ATF at all. In contrast, the proposed method captured fine structures up to approximately 400 Hz and also captured the general shape at frequencies above that.

Figure 5 shows the spatial distributions of the true and estimated ATF magnitude on the x-y-plane at $z = 0$ in the case of the frequency 250 Hz and $M = 5$. The true distribution shows large peaks and dips, but the estimated distribution of KRR was slanted largely in the positive y direction, while that of NF was almost constant. In the proposed method, although the estimated distribution was somewhat different from the true distribution in the region of $x > 0$ m and $y > 0.2$ m, it showed a magnitude distri-

bution similar to the true one.

6. CONCLUSION

We proposed a method for estimating the ATF magnitude distribution based on autoencoder conditioned on source and receiver positions and frequency. Our proposed method allows us to combine multiple datasets measured by different measurement setups and also keep the computational complexity low during inference. Numerical experiments have shown that the proposed method achieves high estimation accuracy even when there are only a few observations.

7. ACKOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 23K24864.

8. REFERENCES

[1] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. Academic Press, 1999.

[2] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *J. Audio Eng. Soc.*, vol. 53, no. 11, pp. 1004–1025, 2005.

[3] N. Ueno and S. Koyama, "Sound field estimation: Theories and applications," *Foundations and Trends® in Signal Processing*, vol. 19, no. 1, pp. 1–98, 2025.

[4] N. Ueno, S. Koyama, and H. Saruwatari, "Sound field recording using distributed microphones based on harmonic analysis of infinite order," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 135–139, 2018.

[5] N. Ueno, S. Koyama, and H. Saruwatari, "Directionally weighted wave field estimation exploiting prior information on source direction," *IEEE Trans. Signal Process.*, vol. 69, pp. 2383–2395, 2021.

[6] F. Lluís, P. Martínez-Nuevo, M. B. Møller, and S. E. Shepstone, "Sound field reconstruction in rooms: Inpainting meets super-resolution," *J. Acoust. Soc. Amer.*, vol. 148, no. 2, 2020.

[7] A. Luo, Y. Du, M. J. Tarr, J. B. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[8] S. Koyama, J. G. C. Ribeiro, T. Nakamura, N. Ueno, and M. Pezzoli, "Physics-informed machine learning for sound field estimation: Fundamentals, state of the art, and challenges," *IEEE Signal Process. Mag.*, vol. 41, no. 6, pp. 60–71, 2025.

[9] F. Miotello, L. Comanducci, M. Pezzoli, A. Bernardini, F. Antonacci, and A. Sarti, "Reconstruction of sound field through diffusion models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, (Seoul, Republic of Korea), Apr. 2024.

[10] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.

[11] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, p. 99–106, Dec. 2021.

[12] 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 *Proc. Int. Workshop Acoust. Signal Enhancement (IWAENC)*, Sep. 2022.

[13] Z. Liang, W. Zhang, and T. D. Abhayapala, "Sound field reconstruction using neural processes with dynamic kernels," *EURASIP J. Audio, Speech, Music Process.*, vol. 13, 2024.

[14] 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, 2022.

[15] 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," in *Proc. Forum Acusticum*, Sep. 2023.

[16] J. G. C. Ribeiro, S. Koyama, R. Horiuchi, and H. Saruwatari, "Sound field estimation based on physics-constrained kernel interpolation adapted to environment," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 4369–4383, 2024.

[17] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikari, S. Wang, and L. Yang, "Physics-informed machine learning," *Nat. Rev. Phys.*, vol. 3, p. 422–440, 2021.

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

[19] 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," in *Proc. Adv. Neural. Inf. Process. Syst. (NeurIPS)*, vol. 33, p. 7537–7547, 2020.

[20] D. Ha, A. M. Dai, and Q. V. Le, "Hypernetworks," in *Proc. Int. Conf. Learn. Rep. (ICLR)*, 2017.

[21] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for fewshot learning," in *Proc. Int. Conf. Neural Informat. Process. Syst. (NeurIPS)*, p. 4080–4090, 2017.

[22] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[23] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, (Calgary, Canada), pp. 351–355, Jul. 2018.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Rep. (ICLR)*, 2015.

[25] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[26] A. F. Agarap, "Deep learning using rectified linear units (ReLu)." arXiv:1803.08375, 2018.