DIFFERENTIABLE SCATTERING DELAY NETWORKS FOR ARTIFICIAL REVERBERATION

Alessandro Ilic Mezza [1], *Riccardo Giampiccolo* [1], *Enzo De Sena* [2,*], *and Alberto Bernardini* [1,†]

[1] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, IT
[2] Institute of Sound Recording, University of Surrey, Guildford, UK
alessandroilic.mezza@polimi.it | riccardo.giampiccolo@polimi.it | e.desena@surrey.ac.uk
alberto.bernardini@polimi.it

ABSTRACT

Scattering delay networks (SDNs) provide a flexible and efficient framework for artificial reverberation and room acoustic modeling. In this work, we introduce a differentiable SDN, enabling gradient-based optimization of its parameters to better approximate the acoustics of real-world environments. By formulating key parameters such as scattering matrices and absorption filters as differentiable functions, we employ gradient descent to optimize an SDN based on a target room impulse response. Our approach minimizes discrepancies in perceptually relevant acoustic features, such as energy decay and frequency-dependent reverberation times. Experimental results demonstrate that the learned SDN configurations significantly improve the accuracy of synthetic reverberation, highlighting the potential of data-driven room acoustic modeling.

1. INTRODUCTION

Room acoustic modeling plays an important role in interactive applications such as video games, virtual reality (VR), and augmented reality (AR), where perceptually plausible reverberation has been shown to significantly enhance the sense of immersion [1] and externalization of virtual sounds [2], among others.

A number of room acoustic models have been proposed over the past 60 years [3]. Within the context of gaming/VR/AR, the most suitable models tend to be delay-network artificial reverberators, due to their low computational complexity, which can be two-to-three orders of magnitude lower than (fast) convolution alone. Among the most used artificial reverberators are feedback delay networks (FDNs), which consist of parallel delay lines connected recursively through a unitary feedback matrix [3]. FDNs are not directly linked to the geometric properties of any given room, but rather aim to render certain high-level acoustical features such as a given reverberation time. For this reason, in gaming/VR/AR applications, they are most often used in conjunction with geometric-acoustic models such as the image source method (ISM) [4], with FDN rendering the late reverberation and ISM the early reflections.

Scattering delay networks (SDNs) [5, 6] are artificial reverberators that render both late reverberation and early reflections within the same design. They consist of recursive networks of delay lines that reproduce first-order reflections exactly, while making progressively coarser approximations of higher-order reflections [5]. All parameters are derived directly from the physical properties of the target room, enabling simulation of unequal and frequency-dependent wall absorption, as well as directional sources and microphones [5]. SDNs have been shown to achieve high perceived naturalness [7], immersion [1] and externalization [2]. SDNs have also been recently extended to higher orders [8, 9], and coupled volumes [10]. They have been incorporated in real-time binaural rendering [2, 11] and AR audio applications [12], and heavily inspired the acoustic model of “Grand Theft Audio V” [13].

In many applications, the input into the acoustic renderer is the room geometry and wall materials, e.g., obtained from the 3D mesh and textures of a video game scene. At the same time, when measured data, such as room impulse responses (RIRs), are available, there is an opportunity to render the acoustics of real-world environments and physical spaces with greater fidelity. In this context, one may wish to combine the accuracy of convolutional models with the low computational complexity of artificial reverberators. This involves optimization of the reverberator’s parameters, a task made challenging by the nonconvex relation between parameters and reverberator output, in addition to the difficulty of computing gradients. Significant strides have been made recently in this sense thanks to differentiable artificial reverberators [14]. Most of the work, however, focused on FDNs [15, 16, 17, 18, 19, 20].

In this paper, we shift attention to SDNs, which bring two key advantages. First, they enable principled parameter initialization, using (approximate) prior knowledge of the room characteristics, when available. Second, they provide a physically-inspired design, enabling explicit interpretation of all the optimized parameters and post-optimization manipulation of the simulated environment, such as repositioning of source and receiver [21].

Extending SDNs to model real-world conditions requires relaxing several of its constraints. SDN is derived from an assumed room geometry, and any errors in the room dimensions or source/ receiver positions will degrade its accuracy. Furthermore, standard SDNs rely on simplifying assumptions such as isotropic scattering, whereas actual rooms exhibit direction-dependent reflection characteristics [22]. When it comes to modeling wall absorption characteristics, SDNs, like most other models, use frequency-dependent absorption values tabulated for different materials [22], which are often an approximation of real-world conditions. Finally, rooms are rarely completely empty, i.e., with no furniture or other absorbent surfaces, which is not directly modeled by SDNs.

To address the complexity of real-world acoustic environments,

[*] This work was partially supported by EPSRC under grant EP/V002554/1 and EP/X032914/1.

[†] This work was partially supported by the European Union – Next Generation EU under the Italian National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.3, CUP D43C22003080001, partnership on “Telecommunications of the Future” (PE00000001 – program “RESTART”).

we propose to design a differentiable SDN (DSDN) in which all parameters are optimized via gradient descent, leveraging automatic differentiation, just like in standard backpropagation [23]. Our experiments show that optimized DSDNs model the energy decay behavior of real-life rooms better than an SDN that was initialized based solely on (approximate) room geometry and wall material properties.

2. DIFFERENTIABLE SCATTERING DELAY NETWORKS

Scattering delay networks (SDNs) [5, 6] are artificial reverberator models inspired by digital waveguide meshes (DWM) [3] that aim to reduce the number of nodes in the mesh to the bare minimum needed to reproduce first-order reflections accurately. This results in a network of bidirectional delay lines connected at scattering nodes, each approximating a specularly reflecting wall (see Figure 1). For an in-depth overview of the SDN architecture, we refer the reader to [5].

In the following, we outline our DSDN implementation. Section 2.1 discusses the geometrical prior that underpins the SDN. Section 2.2 focuses on scattering junctions. Section 2.3 presents the differentiable delay lines. Section 2.4 concerns learnable pressure extraction weights. Section 2.5 proposes a method to account for inaccurate room dimensions and distance measurements. Section 2.6 describes the reparameterization functions used to enforce constraints on the SDN parameters. Finally, the proposed method is summarized in Section 2.7. Notice that the initialization of the parameters described below follows the original geometry-based heuristics [5]. For the sake of simplicity, we disregard source and microphone directivity and assume both to be omnidirectional. Learning directivity patterns is left for future work.

2.1. Geometrical Prior

Let the SDN have N scattering junctions (*wall nodes*). Given the source at location $\mathbf{r}_\mathrm{S} \in \mathbb{R}^3$ and the microphone at location $\mathbf{r}_\mathrm{M} \in \mathbb{R}^3$, let $r_{\mathrm{S},\mathrm{M}} = \|\mathbf{r}_\mathrm{S} - \mathbf{r}_\mathrm{M}\|$ be the source-to-microphone distance. Likewise, given nodes at location $\mathbf{r}_k \in \mathbb{R}^3$, $k = 1, ..., N$, let $r_{\mathrm{S},k} = \|\mathbf{r}_\mathrm{S} - \mathbf{r}_k\|$ be the source-to-node distances, $r_{k,m} = \|\mathbf{r}_m - \mathbf{r}_k\|$ the node-to-node distances, and $r_{k,\mathrm{M}} = \|\mathbf{r}_k - \mathbf{r}_\mathrm{M}\|$ the node-to-microphone distances. The length of these connections, derived from the room geometry, determines the corresponding delay lines as well as the respective attenuation gains.

With c the speed of sound, F_s the sampling frequency, and $G = c/F_\mathrm{s}$, the time-of-flight between the source and the kth node introduces a delay $D_{\mathrm{S},k} = r_{\mathrm{S},k}/G$. Similarly, the delay between the kth node and the microphone is $D_{k,\mathrm{M}} = r_{k,\mathrm{M}}/G$, and the delay between the kth and mth nodes is $D_{k,m} = r_{k,m}/G$.

The attenuation coefficients also depend on the Euclidean distances according to the spherical spreading law:

$$g_{\mathrm{S},k} = \frac{G}{r_{\mathrm{S},k}}, \quad g_{k,\mathrm{M}} = \frac{1}{1 + r_{k,\mathrm{M}}/r_{\mathrm{S},k}}, \quad g_{\mathrm{S},\mathrm{M}} = \frac{1}{r_{\mathrm{S},\mathrm{M}}}. \tag{1}$$

Let

$$\mathbf{p}^+ = [p^+_{1,2}, \ldots, p^+_{k,m}, \cdots p^+_{N,N-1}]^T, \tag{2}$$

$$\mathbf{p}^- = [p^-_{1,2}, \cdots, p^-_{k,m}, \cdots p^-_{N,N-1}]^T, \tag{3}$$

with $k, m = 1, \ldots, N$, $k \neq m$, be "global" vectors of incident and reflected pressure waves, respectively, traveling from/to all scattering nodes at any given time. To route reflected waves back to the junctions onto which they impinge, we use a (fixed) permutation matrix $\mathbf{P}$ depending on the network topology, such that, assuming no superimposing source pressure,

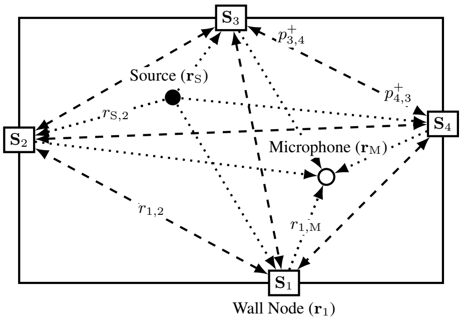

S3
$p^+_{3,4}$
Source ($\mathbf{r}_\mathrm{S}$)
$p^+_{4,3}$
$r_{\mathrm{S},2}$
S4
S2
Microphone ($\mathbf{r}_\mathrm{M}$)
$r_{1,2}$
$r_{1,\mathrm{M}}$
S1
Wall Node ($\mathbf{r}_1$)

Figure 1: *SDN reverberator. Dashed lines represent bidirectional delay lines between wall nodes. Dotted lines represent unidirectional source-to-node and node-to-microphone delay lines. Direct component and floor/ceiling wall nodes are also present, but not shown here for clarity.*

$$\mathbf{p}^+ = \mathbf{P}\mathbf{D}(z)\mathbf{p}^-, \tag{4}$$

where $\mathbf{D}(z)$ is a diagonal matrix of node-to-node time-of-flight delays. For a shoebox room shape, $\mathbf{P} = \delta_{\imath, f(\jmath)}$ [5], where $\delta_{\imath,\jmath}$ is the Kronecker delta, $f(\jmath) = ((6\jmath - ((\jmath - 1))_N - 1))_{N(N-1)} + 1$, and $((\cdot))_N$ is the modulo-N operation. Respectively, assuming the source pressure $p_\mathrm{S}[n]$ is nonzero, (4) can be rewritten as

$$p^+_{k,m} = p^-_{m,k} z^{-D_{m,k}} + \frac{1}{2} g_{\mathrm{S},k} p_\mathrm{S}[n] z^{-D_{\mathrm{S},k}}. \tag{5}$$

By defining pressure vectors *globally*, we are able to parallelize the delay operations for all node-to-node connections (but also source-to-node and node-to-microphone connection lines, see Section 2.3) by applying a delay filter bank $\mathbf{D}(z)$ to the outgoing wave variables in $\mathbf{p}^-$, all while modeling reflections *locally* within each scattering junction.

2.2. Scattering Junctions

Let us consider the kth junction. A junction scatters incoming wave variables $\mathbf{p}^+_k = [p^+_{k,1}, \ldots, p^+_{k,K}]^T$ to produce outgoing wave variables $\mathbf{p}^-_k = [p^-_{k,1}, \ldots, p^-_{k,K}]^T$ such that $\mathbf{p}^-_k = \mathbf{S}_k \mathbf{p}^+_k$, with $\mathbf{S}_k$ a $K \times K$ scattering matrix. Shoebox rooms have $K = 5$.

Let $\mathbf{S}_k = \mathbf{H}_k(z)\mathbf{A}_k$, where

$$\mathbf{H}_k(z) = \mathrm{diag}\{H_k(z), \ldots, H_k(z)\}, \tag{6}$$

$H_k(z)$ is the absorption filter associated to the kth wall, and $\mathbf{A}_k$ is a lossless matrix. In the simplest case, frequency-independent absorption is modeled by $H_k(z) = \beta_k = \sqrt{1 - \alpha_k}$, where $\alpha_k \in [0, 1]$ is the kth wall absorption coefficient. In the following, we optimize two kinds of scattering matrices, learning either N scalars β_k or the taps $\beta_k[\ell]$ of L-order FIR wall filters

$$H_k(z) = \sum_{\ell=0}^{L} \beta_k[\ell] z^{-\ell}, \quad k = 1, \ldots, N. \tag{7}$$

In the former case ($L = 0$), we initialize α_k based on the tabulated random-incidence absorption values associated to the wall material. In the latter case ($L > 0$), we initialize all taps to 0 except for the zero-lag coefficient that is set to $\beta_k[0] = \sqrt{1-\alpha_k}$, making the two methods equivalent at the very first iteration.

In the case of an isotropic medium, $\mathbf{A}_k$ is typically defined as

$$\bar{\mathbf{A}} = \frac{2}{K}\mathbf{1}\mathbf{1}^T - \mathbf{I}_K\,, \tag{8}$$

where $\mathbf{1} = [1, \ldots, 1]^T$ and $\mathbf{I}_K$ is the $K \times K$ identity matrix.

Here, we augment the expressive power of the SDN by introducing a learnable vector of admittances $\mathbf{y}_k = [y_{k,1}, \ldots, y_{k,K}]^T$ at each junction, such that

$$\mathbf{A}_k = \frac{2}{\langle \mathbf{1}, \mathbf{y}_k \rangle}\mathbf{1}\mathbf{y}_k^T - \mathbf{I}_K \tag{9}$$

is a parametric Householder matrix that can be optimized for each junction independently of the others. We initialize $\mathbf{y}_k = \mathbf{1}, \forall k$, so as to have, at the first iteration, the isotropic case $\mathbf{A}_k = \bar{\mathbf{A}}$.

According to the definition of the global pressure vectors $\mathbf{p}^+$ and $\mathbf{p}^-$ given in Section 2.1, we obtain $\mathbf{p}_k^+ = \mathbf{R}_k\mathbf{p}^+$, where $\mathbf{R}_k$ is a $K \times K(K+1)$ selection matrix

$$\mathbf{R}_k = \left[\mathbf{0}_{\{K,kK\}},\, \mathbf{I}_K,\, \mathbf{0}_{\{K,K(K-k)\}}\right], \tag{10}$$

where $\mathbf{0}_{\{R,C\}}$ is a zero matrix with R rows and C columns. Thus, wall scattering can be written as

$$\mathbf{p}^- = \sum_{k=1}^{N}\mathbf{R}_k^T\mathbf{p}_k^- = \sum_{k=1}^{N}\mathbf{R}_k^T\mathbf{S}_k\mathbf{R}_k\mathbf{p}^+. \tag{11}$$

2.3. Differentiable Delay Lines

Integer delays can be efficiently implemented as a reading operation from a buffer that accumulates past samples. Unfortunately, this approach is not differentiable. In [24], the authors implemented a differentiable lookup table mechanisms that linearly interpolates between the two closest discrete-time samples via weighted sum. In [15], we implemented differentiable delay lines based on the closed-form variable fractional time delay filter by Pei and Lai [25]. In the following, we adopt the latter approach.

Each unidirectional delay line contains a B-sample first-in first-out buffer that stores the signal $x[n]$. To apply a delay of D fractional samples, we first zero-pad $x[n]$ to minimize artifacts from circular convolution. We compute its Q-point Fast Fourier Transform (FFT), with $Q = 2B$. In the frequency domain, the delay is thus performed via the Hadamard product of the transformed signal $X[\omega_n]$ and the conjugate symmetric frequency response $Z[\omega_n]$, which corresponds to a windowed-sinc finite impulse response in the time domain [25]. Finally, we go back to the time domain by computing the inverse FFT, which yields

$$x[n-D] = \text{IFFT}\{Z[\omega_n] \odot \text{FFT}\{x[n]\}\}. \tag{12}$$

With (12) easily parallelizable, we define three filter banks for the source-to-node, node-to-node, and node-to-microphone connection lines, respectively, plus a unidirectional delay line for the source-to-microphone connection. The node-to-node filter bank has $N(N-1)$ filters, wheres the remaining ones have N. All delay operations in a filter bank are performed in parallel.

To reduce computational costs, the time-domain wall absorption filters $H_k(z)$ (Section 2.2) are applied to the output of every delay line departing from the kth node. This approach allows us to use a single buffer for both the FFT-based fractional delay and the time-domain absorption filters. This is made possible because all operations within the SDN are linear.

2.4. Pressure Extraction Weights

The signal impinging on the microphone is taken as the node's pressure extracted from the kth junction as a linear combination of outgoing wave variables, i.e., $\check{p}_k[n] = \mathbf{w}_k^T\mathbf{p}_k^-$, such that the microphone signal is given by

$$p_\text{M}[n] = \sum_{k=1}^{N} g_{k,\text{M}} \cdot \check{p}_k[n] \cdot z^{-D_{k,\text{M}}}, \tag{13}$$

We optimize the real-valued vector $\mathbf{w} = [\mathbf{w}_1^T, \ldots, \mathbf{w}_N^T]^T$ of size $N(N-1)$ containing one unconstrained pressure extraction weight $w_{k,m} \in \mathbb{R}$ for each outgoing connection line in the SDN. Since $\mathbf{A}_k$ is initialized as in (9) with $\mathbf{y}_k = \mathbf{1}$, following [5], we initialize $w_{k,m} = 2/\langle \mathbf{1}, \mathbf{y}_k \rangle = 2/5,\ \forall k, m$.

2.5. Accounting for Uncertainty in the Geometrical Prior

Sometimes room geometry measurements are unreliable. In turn, this turns out to affect the performance of SDNs, as they heavily rely on such a prior. We account for this problem by introducing a learnable correction term $\Delta r_{k,m}$ for each Euclidean distance $r_{k,m}$ in the delay network. Namely, we use the following proxy distance

$$r'_{k,m} = r_{k,m} + \Delta r_{k,m} \tag{14}$$

in place of $r_{k,m}$ in every computation relative to delays and absorption coefficients (see Section 2.1).

The correction term is defined as

$$\Delta r_{k,m} = \Delta r_\text{max} \cdot \tanh(s \cdot \delta_{k,m}), \tag{15}$$

with $\delta_{m,k}$ an unconstrained learnable parameter initialized at zero, and $s \in \mathbb{R}_{\geq 0}$ a dilation hyperparameter. The hyperbolic tangent function ensures that the learnable parameter remains bounded within $[-1, 1]$, regardless of the values assumed by $\delta_{m,k}$. Consequently, Δr_max controls the extent to which the correction term can modify the geometric prior.

While it is worth emphasizing that learned distances may not necessarily define a physically realizable enclosure, Δr_max is chosen to be small compared to the room size in order to preserve, at least to some extent, the geometric interpretation of the SDN.

2.6. Parameter Constraints

SDNs require minimal parameter constraints. We enforce them by reparameterizing the unconstrained learnable parameters as the arguments of differentiable (almost everywhere) functions whose codomain satisfies the given constraints. In particular, we make sure that the admittances in $\mathbf{y}_k$ are strictly positive by taking $y_{k,i} = |\tilde{y}_{k,i}| + \epsilon$, with $\epsilon = 10^{-12}$, where $\tilde{y}_{k,i} \in \mathbb{R}$ is an unconstrained learnable parameter. Similarly, connection lengths $r_{k,m}, \forall k, m$, are forced to be nonnegative by taking the absolute value on the unconstrained values, i.e., $r_{k,m} = |\tilde{r}_{k,m}|$. In the case of frequency-independent wall absorption, $H_k(z) = \beta_k$, instead of learning β_k directly, we learn the argument $\tilde{\beta}_k$ of a logistic function that maps the unconstrained parameter onto $[0, 1]$.

In case of very large rooms, it is also possible to learn distances in decameters or hectometers instead of meters. The rationale here is that variables several orders of magnitude larger than the learning rate, which is typically chosen to be small, turn out to be insensitive to gradient-based updates. From a computational point of view, this just entails applying a constant multiplicative factor in all distance-related SDN computations.

Table 1: **MeshRIR:** *Reverberation time* (T_{30})*, Clarity* (C_{80})*, Definition* (D_{50})*, Center time* (t_{s})*, and Early Decay Time* (EDT).

		T_{30}	C_{80}	D_{50}	t_{s}	EDT
	Reference	0.395	15.6596	92.3548	17.7154	0.0437
DSDN	Pre-optim	0.808	8.0633	78.1428	35.1	0.0982
	$L = 0$	**0.404**	16.2073	93.9937	16.2225	0.0322
	$L = 6$	0.452	**15.9434**	**93.5167**	**16.6381**	**0.0355**

Table 2: **HOMULA-RIR:** *Reverberation time* (T_{30})*, Clarity* (C_{80})*, Definition* (D_{50})*, Center time* (t_{s})*, and Early Decay Time* (EDT)*. With* $\{\Delta r\}$*, we denote distance correction.*

		$\{\Delta r\}$	T_{30}	C_{80}	D_{50}	t_{s}	EDT
	Reference	N/A	0.5818	18.8547	97.8042	6.7726	0.0041
DSDN	Pre-optim	N/A	0.8927	12.7132	90.0311	19.3109	0.05
	$L = 0$	×	0.4885	20.8139	98.2969	7.6645	0.0087
	$L = 6$	×	0.5498	**19.0282**	**97.7611**	8.0415	0.0051
	$L = 0$	✓	0.5222	20.0003	97.8991	6.5627	0.0040
	$L = 6$	✓	**0.5508**	19.3521	97.7457	**6.6504**	**0.0041**

2.7. Summary

To recap, we optimize the following parameters:

- (Section 2.2) N vectors $\tilde{\mathbf{y}}_k \in \mathbb{R}^K$ such that $\mathbf{y}_k = |\tilde{\mathbf{y}}_k| + \epsilon$ (with the absolute value applied element-wise) are the characteristic admittances parameterizing the Householder scattering matrices;
- (Section 2.2) N scalars $\tilde{\beta}_k$ parameterizing zero-order wall filters $H_k(z) = \beta_k$, where $\beta_k = 1/(1 + \exp(-\tilde{\beta}_k))$;
- (Section 2.2) N vectors $\boldsymbol{\beta}_k = [\beta_k[0], ..., \beta_k[L]]^T$ parameterizing L-order wall filters as in (7);
- (Section 2.4) $N(N-1)$ unconstrained pressure extraction weights $w_{k,m} \in \mathbb{R}$, one for each outgoing wave variable;
- (Section 2.5) $(N^2 + N + 1)$ correction terms $\Delta r_{*,\star}$, each adjusting the length of the corresponding connection line by up to $\pm\Delta r_{\mathrm{max}}$ m.

2.8. Learning Objective

We minimize the loss function from [16], which was previously used to optimize time-domain FDNs as to match spectro-temporal features of a target RIR:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{EDC}} + \lambda_2 \mathcal{L}_{\mathrm{EDP}} + \lambda_3 \mathcal{L}_{\mathrm{EDR}}, \tag{16}$$

where $\mathcal{L}_{\mathrm{EDC}}$ is a normalized L^2-loss between full-band Energy Decay Curves (EDC) obtained via Schroeder's backward integration [26], $\mathcal{L}_{\mathrm{EDR}}$ is a normalized L^1-loss between mel-frequency Energy Decay Relief (EDR) features expressed in dB, and $\mathcal{L}_{\mathrm{EDP}}$ is a L^2-loss between Soft EDP functions, i.e., a differentiable approximation of Abel and Huang's Echo Density Profile (EDP) [27] obtained by substituting the non-differentiable indicator function with a scaled logistic function [15]. Weights λ_1, λ_2, and λ_3 are positive scalars that balance the magnitude of the three loss terms.

As in [15, 16, 17], the loss function is evaluated on the first portion of the RIR, obtained by truncating the full response at the estimated reverberation time. The minimization is carried out for 300 iterations using Adam optimizer [28], with a learning rate of 0.01 and default moving average hyperparameters.

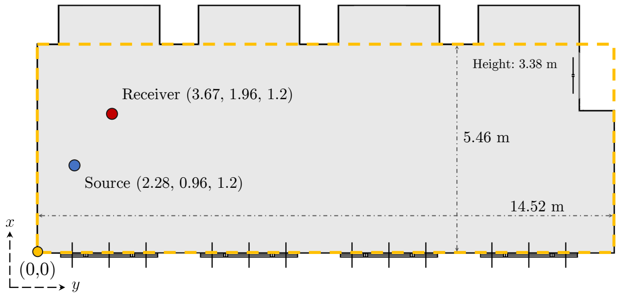


Figure 2: **HOMULA-RIR:** *Floor plan of the "Schiavoni" room. The source (in blue) is a Genelec 8020D loudspeaker; the receiver (in red) is an omnidirectional eStick V3 microphone by Eventide Inc. and Politecnico di Milano* [31]*. Overlaid in yellow, we depict the approximating shoebox geometry used to initialize the SDNs.*

3. EVALUATION

First, we select a RIR from MeshRIR [29], which was measured in an empty cuboid room, with source and microphone positions precisely controlled via a Cartesian robot [29]. Second, we focus on a sparsely furnished irregular shaped room from HOMULA-RIR [30], where distances were measured manually.

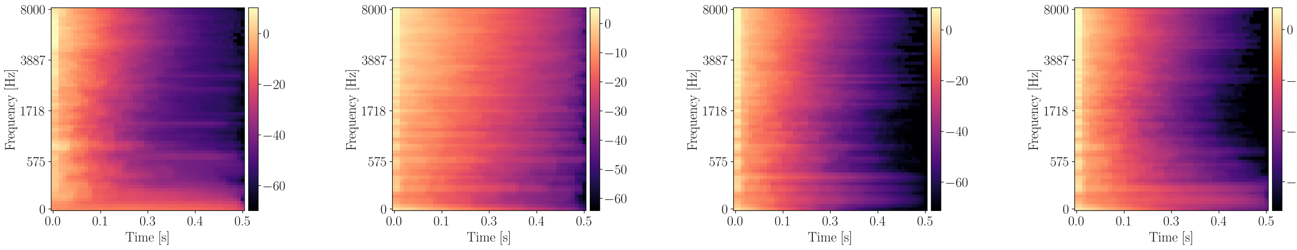
In our experiments, we set $F_{\mathrm{s}} = 16$ kHz and $c = 343$ m/s. We do not carefully tune random-incidence absorption coefficients, e.g., through a grid search. Instead, we assume that all walls are plastered. Hence, we set $\alpha_k = 0.02, \forall k$, based on the value listed for the 250 Hz octave band in [22]. This deliberate (albeit arguably naïve) choice is meant to highlight the advantage of automatic parameter estimation via gradient descent over laborious manual tuning. Finally, we set the order of the wall filters to $L = 6$ in the frequency-dependent absorption case.

All DSDNs are implemented in PyTorch and optimized on a single NVIDIA Titan RTX with 24 GB of RAM. The source code is publicly available online.[1]

Test I: MeshRIR. From MeshRIR [29], we select a RIR from subset `S1-M3969` associated to microphone index `1984`. The cuboid room has approximate dimensions 7.0 m × 6.4 m × 2.7 m. The source, a DIATONE DS-7 closed loudspeaker, is located at coordinates $(2.0, 1.5, 0.0)$ m. We discard any leading silence that exceeds the expected time-of-flight delay associated with the line-of-sight distance $r_{\mathrm{S,M}} = 2.5$ m. The RIR is resampled from 48 kHz to 16 kHz and scaled to unit norm. DSDN training involves discrete-time simulations where the SDN is fed a unit pulse $\delta[n]$ and run for 0.38 s, equal to the average T_{60} reported in [29].

Test II: HOMULA-RIR. HOMULA-RIR [30] comprises multichannel RIRs recorded in the "Schiavoni" seminar room at Politecnico di Milano. The room, whose floor plan in shown in Figure 2, is irregularly shaped and features four large windows that, at the time of recording, were covered with heavy curtains. Furthermore, the walls were partially covered with decorative boards, and the room contained several tables and chairs. We select the RIR from source `S1`, a Genelec 8020D loudspeaker located at $(2.28, 0.96, 1.2)$ m, to the furthest eStick V3 microphone located at $(3.67, 1.96, 1.2)$ m with respect to the origin located in the bottom-left corner of the room. The RIR is resampled to 16 kHz

[1] `https://github.com/ilic-mezza/differentiable-sdn/`



(a) *Reference* (b) *Pre-optimization* (c) *DSDN* ($L = 0$) (d) *DSDN* ($L = 6$)

Figure 3: **MeshRIR:** *Mel-scale Energy Decay Relief (EDR) without distance correction.*

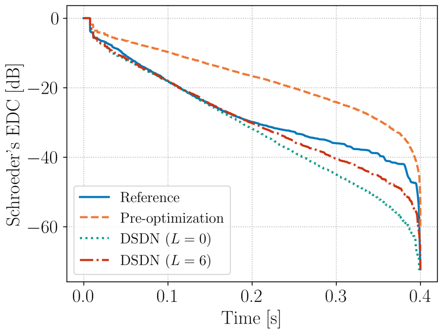


Figure 4: **MeshRIR:** *Schroeder's Energy Decay Curves (EDCs).*

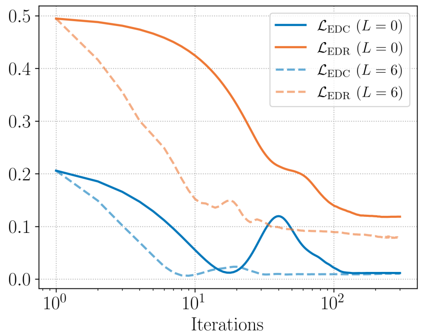


Figure 5: **MeshRIR:** *Losses as a function of the iteration index.*

and scaled to unit norm. To instantiate the SDN distance parameters, we approximate the irregular room shape with a shoebox enclosure of size 5.46 m × 14.52 m × 3.38 m, depicted in yellow in Figure 2. At each training iteration, the DSDN is run for 0.58 s, equal to the T_{30} estimated from the target RIR using `pyroomacoustics`.

Test III: HOMULA-RIR With Distance Correction. Given the uncertainty coming from real-world distance measurements and to account for the many nonidealities of the "Schiavoni" room, we experiment with learning distance correction terms. Note that in Test I and Test II, distances were kept fixed. Here, we optimize $\Delta r_{*,\star}$ by setting $s = 10$ and $\Delta r_{\text{max}} = 0.5$ m (see Section 2.5). Apart from this, the experimental setup follows that of Test II.

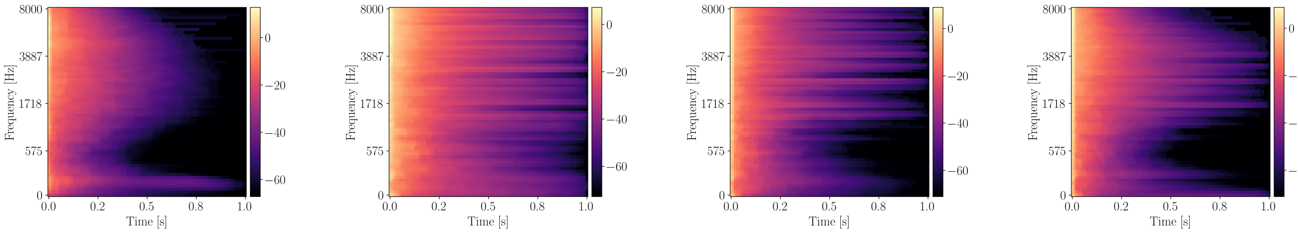
4. RESULTS

Figure 3 shows the 64-bin mel-scale EDRs of the MeshRIR test case. Left to right, we depict the EDR of (a) the RIR (*Reference*), (b) the SDN initialized as in [5] based exclusively on geometrical assumptions (*Pre-optimization*), (c) the optimized DSDN implementing frequency-independent wall absorption, $L = 0$, and (d) the optimized DSDN implementing frequency-dependent absorption via learnable FIR filters, $L = 6$. Figure 4 shows the corresponding EDCs in dB. Table 1 reports several ISO 3382 metrics [32]. Figure 5 shows the evolution of $\mathcal{L}_{\text{EDC}}$ and $\mathcal{L}_{\text{EDR}}$ over 300 training iterations, both for the case of frequency-independent wall absorption (solid lines) and frequency-dependent wall absorption (dashed lines).

The EDRs in Figure 3 reveal that the SDN, before optimization, exhibits nearly uniform decay across all mel bands, with longer reverberation times at just about every frequency compared to the target RIR. The optimized DSDNs more closely match the reference EDR, though the choice of absorption model affects the decay characteristics. When wall absorption is modeled with zero-order FIR filters, the response at low frequencies is characterized by a faster decay. In contrast, sixth-order FIR filters render low frequencies with higher fidelity (cf. Figure 3a and Figure 3d).

The EDCs in Figure 4 demonstrate that both DSDNs significantly improve upon the pre-optimization model (orange dashed line). Relative to the target $T_{30} = 0.395$ s, the estimated reverberation times deviate by no more than 10 ms for $L = 0$, and by just under 57 ms for $L = 6$. Up to approximately -30 dB, indeed, the optimized DSDNs align with the reference EDC (blue solid line). However, beyond this point, the RIR exhibits a change in slope that the models fail to reproduce. While SDNs are perfectly capable of modeling double-slope decay profiles [10], the observed discrepancy is likely due to the choice of $\mathcal{L}_{\text{EDC}}$ [15, 16, 17], which, being a linear-scale L^2-loss function, prioritizes the early portion of the decay over the reverberation tail. Future improvements could involve integrating a log-scale EDC loss term, similar to how spectral losses are dealt with in [33], to better account for the decay of low-amplitude late reverberation.

The loss curves in Figure 5 illustrate how the inclusion of FIR wall absorption filters affects optimization. The addition of 36 extra parameters to the SDN increases its expressive power, though using sixth-order FIR filters has little impact on the local minimum reached by $\mathcal{L}_{\text{EDC}}$. Conversely, a notable improvement is observed for $\mathcal{L}_{\text{EDR}}$. Moreover, this approach accelerates convergence of both energy decay losses by a factor of 10, with results previously obtained in 100 iterations being achieved after only 10 when learning frequency-dependent absorption filters.



(a) *Reference* (b) *Pre-optimization* (c) *DSDN* ($L = 0$) (d) *DSDN* ($L = 6$)

Figure 6: **HOMULA-RIR:** *Mel-scale Energy Decay Relief (EDR) without distance correction.*

Table 1 indicates that this also leads to small yet consistent improvements in Clarity (C_{80}), Definition (D_{50}), Center time (t_s), and Early Decay Time (EDT).

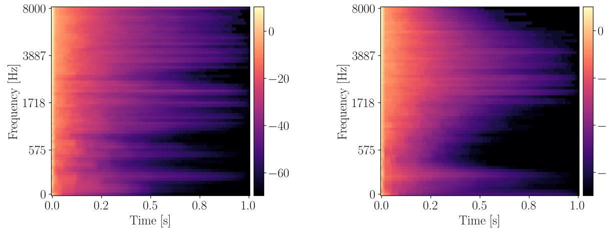
As for the HOMULA-RIR test case, Figure 6 shows the mel-scale EDRs when distances are kept frozen, whereas Figure 7 illustrates the effect of learning distance correction terms. Figure 8 shows the corresponding EDCs, Table 2 reports ISO 3382 metrics, and Figure 9 shows $\mathcal{L}_\mathrm{EDC}$ and $\mathcal{L}_\mathrm{EDR}$ as a function of the iteration index. Finally, Figure 10 depicts the evolution of the distance correction terms throughout the optimization process; the line-of-sight $\Delta r_\mathrm{S,M}$ is depicted in black, while the terms associated to the first reflection bouncing off the kth wall, i.e., $\Delta r_{\mathrm{S},k,\mathrm{M}} = \Delta r_{\mathrm{S},k} + \Delta r_{k,\mathrm{M}}$, are shown for $k = 1, \ldots, N$.

In Figure 6 and Figure 7, whereas the pre-optimization model exhibits a largely frequency-independent behavior, the optimized DSDNs demonstrate notable improvements in capturing the time-frequency characteristics of the target, particularly at very high and low frequencies when learning sixth-order wall filters (Figures 6d and 7b). SDN models though, either pre- and post-optimization, appear to be characterized by few strongly excited resonant frequencies, resulting in the comb-like behavior noticeable in the respective EDRs. This is a well-documented issue with recursive delay-network based artificial reverberators, which, in turn, results in a metallic sound quality [18, 19, 34]. Notably, learning distances along with FIR absorption filters appears to somewhat mitigate this effect (see Figure 7b).

As previously noted for Figure 5, Figure 9 shows that filter order affects the convergence of $\mathcal{L}_\mathrm{EDR}$ while having minimal impact on the final value of $\mathcal{L}_\mathrm{EDC}$. However, comparing Figure 5 with Figure 9a, we notice that $\mathcal{L}_\mathrm{EDC}$ is significantly larger in the latter case than for MeshRIR, suggesting inaccuracies in the geometrical prior when it comes to HOMULA-RIR.

Figure 9 also reveals that learning distances has little effect on the trajectory of $\mathcal{L}_\mathrm{EDR}$, aside from introducing jitter in the later stages of training (see Figure 9b). In contrast, for $\mathcal{L}_\mathrm{EDC}$, learning distance correction terms dramatically improves the optimization results, decreasing the error by more than two orders of magnitude.

The reason for this is to be found in Figure 8, where zoomed-in regions of the plots show that errors in the line-of-sight and early reflections are successfully compensated. Figure 10 confirms this, showing that the correction terms use the entire available range, with $\Delta r_\mathrm{S,M}$ stabilizing at a negative value within the first few iterations. The early portion of the RIR is, in fact, where most of the error is concentrated for DSNDs without distance correction. Indeed, optimized DSDNs show an otherwise strong fit throughout the EDC, regardless of the order of the wall filters. In turn, this leads to remarkable improvements across all metrics in Table 2 compared to the baseline.



(a) *DSDN* ($L = 0$) (b) *DSDN* ($L = 6$)

Figure 7: **HOMULA-RIR:** *Mel-EDR with distance correction.*

5. CONCLUSIONS

In this paper, we introduced the first differentiable implementation of SDNs enabling gradient computation through automatic differentiation. We demonstrated that optimizing delay-network parameters such as scattering matrices, wall absorption filters, and pressure extraction weights via gradient descent can improve the modeling capabilities of SDNs over relying solely on geometrical and physical priors. We showed that learning frequency-dependent wall absorption filters enhances the energy decay relief of the output, and accelerates convergence compared to optimizing full-band random-incidence absorption coefficients. Moreover, we demonstrated that it is possible to compensate for uncertainties in real-world distance measurements by learning bounded correction terms for the length of the SDN connection lines.

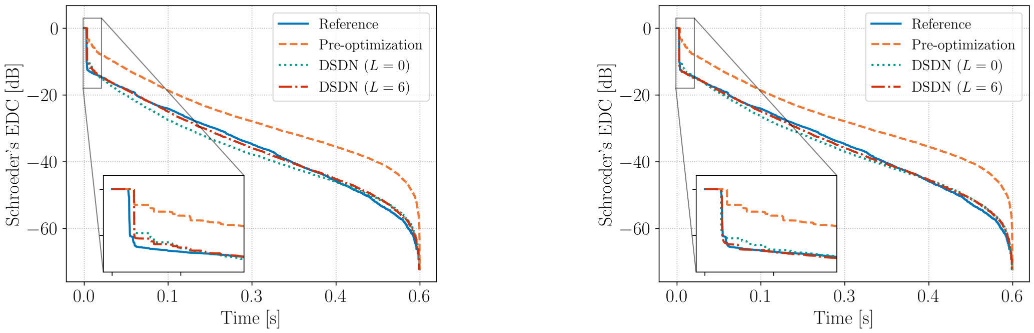
Beyond parameter estimation, differentiable SDNs can also be integrated into deep learning models, either as layers of a neural network or as components of a loss function. Furthermore, future work will explore their use in VR/AR applications, thus taking into account dynamic source–receiver positioning.

6. REFERENCES

[1] Thomas Potter, Zoran Cvetković, and Enzo De Sena, “On the relative importance of visual and spatial audio rendering on VR immersion,” *Front. Signal Process.*, vol. 2, 2022.

[2] Michele Geronazzo, Jason Yves Tissieres, and Stefania Serafin, “A minimal personalization of dynamic binaural synthesis with mixed structural modeling and scattering delay networks,” in *2020 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 411–415.

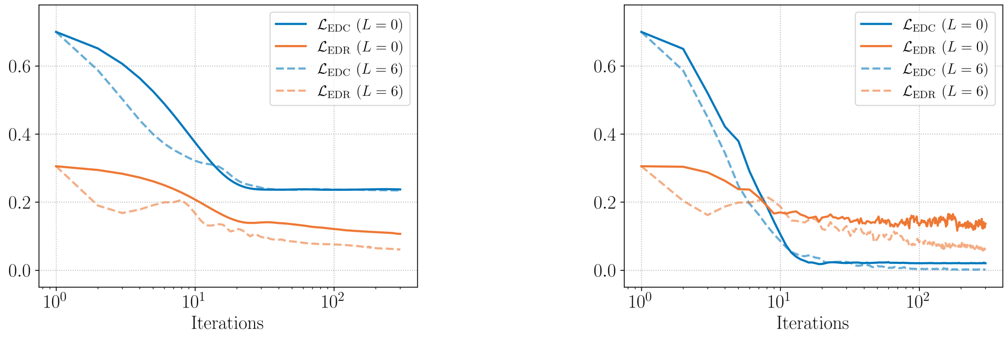
[3] Vesa Välimäki, Julian D. Parker, Lauri Savioja, Julius O. Smith, and Jonathan S. Abel, “Fifty years of artificial re-



(a) *Without distance correction*

(b) *With distance correction*

Figure 8: **HOMULA-RIR:** *Schroeder's Energy Decay Curves (EDCs).*



(a) *Without distance correction*

(b) *With distance correction*

Figure 9: **HOMULA-RIR:** *Losses as a function of the iteration index.*

verberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 5, pp. 1421–1448, 2012.

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

[5] Enzo De Sena, Hüseyin Hacıhabiboğlu, Zoran Cvetković, and Julius O. Smith, "Efficient synthesis of room acoustics via scattering delay networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 9, pp. 1478–1492, 2015.

[6] Enzo De Sena, Zoran Cvetković, and Hüseyin Hacıhabiboğlu, "Electronic device with digital reverberator and method," Dec. 2014, USPTO Patent 8,908,875.

[7] Stojan Djordjevic, Hüseyin Hacıhabiboğlu, Zoran Cvetković, and Enzo De Sena, "Evaluation of the perceived naturalness of artificial reverberation algorithms," in *148th AES Conv.*, 2020, Online.

[8] Matteo Scerbo, Orchisama Das, Patrick Friend, and Enzo De Sena, "Higher-order scattering delay networks for artificial reverberation," in *Proc. 25th Int. Conf. Digital Audio Effects (DAFx-22)*, 2022.

[9] Leny Vinceslas, Matteo Scerbo, Hüseyin Hacıhabiboğlu, Zoran Cvetković, and Enzo De Sena, "Low-complexity higher order scattering delay networks," in *2023 IEEE Workshop Appl. Signal Process. Audio Acoustics (WASPAA)*, 2023, pp. 1–5.

[10] Timuçin Berk Atalay, Zühre Sü Gül, Enzo De Sena, Zoran Cvetković, and Hüseyin Hacıhabiboğlu, "Scattering delay network simulator of coupled volume acoustics," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 582–593, 2022.

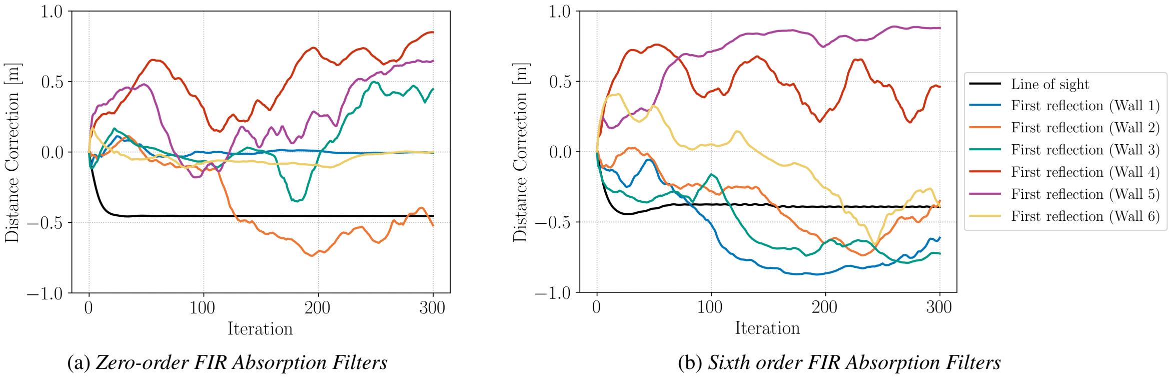
[11] Christopher Yeoward, Rishi Shukla, Rebecca Stewart, Mark Sandler, and Joshua D. Reiss, "Real-time binaural room modelling for augmented reality applications," *J. Audio Eng. Soc.*, vol. 69, no. 11, pp. 818–833, 2021.

[12] Alex Baldwin, Stefania Serafin, and Cumhur Erkut, "Towards the design and evaluation of delay-based modeling of acoustic scenes in mobile augmented reality," in *Proc. 4th VR Workshop on Sonic Interactions for Virtual Environments (SIVE-2018)*, 2018.

[13] Alastair MacGregor, Vice President of Audio at Rockstar North, personal communication, Mar. 2023.

[14] Sungho Lee, Hyeong-Seok Choi, and Kyogu Lee, "Differentiable artificial reverberation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 2541–2556, 2022.

[15] Alessandro Ilic Mezza, Riccardo Giampiccolo, Enzo De Sena, and Alberto Bernardini, "Data-driven room acoustic modeling via differentiable feedback delay networks with learnable delay lines," *EURASIP J. Audio Speech Music Process.*, vol. 2024, no. 51, 2024.



(a) *Zero-order FIR Absorption Filters*

(b) *Sixth order FIR Absorption Filters*

Figure 10: **HOMULA-RIR:** *Distance correction as a function of the iteration index;* $\Delta r_{\mathrm{S,M}}$ *(black) and* $\Delta r_{\mathrm{S},k,\mathrm{M}}$ *(walls one through six).*

[16] Alessandro Ilic Mezza, Riccardo Giampiccolo, and Alberto Bernardini, "Modeling the frequency-dependent sound energy decay of acoustic environments with differentiable feedback delay networks," in *Proc. 27th Int. Conf. Digital Audio Effects (DAFx24)*, 2024, pp. 238–245.

[17] Riccardo Giampiccolo, Alessandro Ilic Mezza, and Alberto Bernardini, "Differentiable mimo feedback delay networks for multichannel room impulse response modeling," in *Proc. 27th Int. Conf. Digital Audio Effects (DAFx24)*, 2024, pp. 278–285.

[18] Gloria Dal Santo, Karolina Prawda, Sebastian Schlecht, and Vesa Välimäki, "Differentiable feedback delay network for colorless reverberation," in *Proc. 26th Int. Conf. Digital Audio Effects (DAFx23)*, 2023, pp. 244–251.

[19] Gloria Dal Santo, Karolina Prawda, Sebastian Schlecht, and Vesa Välimäki, "Optimizing tiny colorless feedback delay networks," *EURASIP J. Audio Speech Music Process.*, vol. 2025, no. 1, 2025.

[20] Riccardo Giampiccolo, Alessandro Ilic Mezza, Mirco Pezzoli, Shoichi Koyama, Alberto Bernardini, and Fabio Antonacci, "Modeling the impulse response of higher-order microphone arrays using differentiable feedback delay networks," in *Proc. 28th Int. Conf. Digital Audio Effects (DAFx25)*, 2025.

[21] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi, "Novel-view acoustic synthesis," in *Proc. of the IEEE/CVF Conv. on Computer Vision and Pattern Recogn. (CVPR)*, 2023, pp. 6409–6419.

[22] Michael Vorländer, *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*, Springer Verlag, 2008.

[23] Alessandro Ilic Mezza, Riccardo Giampiccolo, and Alberto Bernardini, "Data-driven parameter estimation of lumped-element models via automatic differentiation," *IEEE Access*, vol. 11, pp. 143601–143615, 2023.

[24] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts, "DDSP: Differentiable digital signal processing," in *Int. Conf. Learning Representations*, 2020.

[25] Soo-Chang Pei and Yun-Chiu Lai, "Closed form variable fractional time delay using FFT," *IEEE Signal Process. Lett.*, vol. 19, no. 5, pp. 299–302, 2012.

[26] Manfred R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am.*, vol. 37, no. 3, pp. 409–412, 1965.

[27] Jonathan S. Abel and Patty Huang, "A simple, robust measure of reverberation echo density," in *121st AES Conv.*, 2006.

[28] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learning Representations*, 2015.

[29] Shoichi Koyama, Tomoya Nishida, Keisuke Kimura, Takumi Abe, Natsuki Ueno, and Jesper Brunnström, "MeshRIR: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods," in *2021 IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2021, pp. 1–5.

[30] Federico Miotello, Paolo Ostan, Mirco Pezzoli, Luca Comanducci, Alberto Bernardini, Fabio Antonacci, and Augusto Sarti, "HOMULA-RIR: A room impulse response dataset for teleconferencing and spatial audio applications acquired through higher-order microphones and uniform linear microphone arrays," in *2024 IEEE Int. Conf. Acoust. Speech Signal Process. Workshops (ICASSPW)*, 2024, pp. 795–799.

[31] Mirco Pezzoli, Luca Comanducci, Joe Waltz, Anthony Agnello, Luca Bondi, Antonio Canclini, and Augusto Sarti, "A Dante powered modular microphone array system," *J. Audio Eng. Soc.*, , no. 479, 2018.

[32] "Acoustics — Measurement of room acoustic parameters. Part 1: Performance spaces," ISO 3382-1:2009, International Organization for Standardization, Geneva, Switzerland, June 2009.

[33] Alessandro Ilic Mezza, Matteo Amerena, Alberto Bernardini, and Augusto Sarti, "Hybrid packet loss concealment for real-time networked music applications," *IEEE Open J. Signal Process.*, vol. 5, pp. 266–273, 2024.

[34] Manfred R. Schroeder and Benjamin F. Logan, ""Colorless" artificial reverberation," *IRE Trans. Audio*, vol. AU-9, no. 6, pp. 209–214, 1961.