RIR2FDN: An Improved Room Impulse Response Analysis and Synthesis

Gloria Dal Santo, Benoit Alary, Karolina Prawda, Sebastian J Schlecht, Vesa Välimäki

RIR2FDN: AN IMPROVED ROOM IMPULSE RESPONSE ANALYSIS AND SYNTHESIS

Gloria Dal Santo,
Acoustics Lab,
Department of Information and Communications Engineering,
Aalto University,
FI-02150 Espoo, Finland
gloria.dalsanto@aalto.fi

Benoit Alary
STMS,
IRCAM,
Sorbonne Université, CNRS,
Ministère de la Culture,
FR-75004 Paris, France
benoit.alary@ircam.fr

Karolina Prawda, Sebastian J. Schlecht and Vesa Välimäki*
Acoustics Lab,
Department of Information and Communications Engineering,
Aalto University,
FI-02150 Espoo, Finland

ABSTRACT

This paper seeks to improve the state-of-the-art in delay-network-based analysis-synthesis of measured room impulse responses (RIRs). We propose an informed method incorporating improved energy decay estimation and synthesis with an optimized feedback delay network. The performance of the presented method is compared against an end-to-end deep-learning approach. A formal listening test was conducted where participants assessed the similarity of reverberated material across seven distinct RIRs and three different sound sources. The results reveal that the performance of these methods is influenced by both the excitation sounds and the reverberation conditions. Nonetheless, the proposed method consistently demonstrates higher similarity ratings compared to the end-to-end approach across most conditions. However, achieving an indistinguishable synthesis of measured RIRs remains a persistent challenge, underscoring the complexity of this problem. Overall, this work helps improve the sound quality of analysis-based artificial reverberation.

1. INTRODUCTION

Artificial reverberation is commonly used to synthesize the acoustics of physical spaces [1]. A typical approach is to measure a room impulse response (RIR) and analyze it to obtain parameters used to specify a reverberator [2]. This analysis-synthesis technique offers several benefits over direct convolution, such as minimizing data storage and computational costs while offering parametric tuning capabilities [1]. However, it relies on the accuracy of both the underlying analysis and synthesis methods.

Early artificial reverberators, based on interconnected delay lines, required subjective parameter tuning to obtain satisfactory results [3, 4]. In the early 1990s, a generalized model for feedback delay networks (FDNs) introduced the use of reverberation time T_{60} as a design parameter [5]. Attenuation filters were inserted in the feedback paths of the FDN, with their coefficients informed by the frequency-dependent T_{60} of the measured RIRs [2]. Moreover, to keep the frequency-response envelope unaffected by the attenuation filters, a tone-correction filter was placed in series [2].

Although the T_{60} may be analyzed from a measured RIR, other delay network parameters, such as delay-line lengths and the mixing matrix, cannot be easily estimated using an analytic approach and may benefit from an optimization process to best match the reverberator output to a measured RIR. In [6], a genetic algorithm is used to optimize the mixing matrix in a delay network based on [7]. In [8] and more recently in [9], a genetic algorithm is used to optimize the delay lengths as well as the input and output gains of an FDN, whereas the decay of the system relies on a traditional analysis-synthesis approach. In [10], the gains and delay lengths are estimated using perceptual room descriptors. In [11], an autoregressive model is used to match the early reflections of a measured RIR, and a mel-spectrum analysis is used as a loss function during an optimization phase to tune the parameters of two plugin implementations of delay-based reverberators.

While most of these optimization methods rely on traditional reverberator methods, recent advancements, such as the use of velvet noise to reproduce RIRs [12] and to improve modal density in lower-order FDNs [13], may prove beneficial. Other improvements in FDN-based reverberators include a method to find the optimal mixing matrix to achieve a colorless prototype [14, 15] and a two-stage filter design to improve the accuracy of attenuation filters in the feedback path [16].

Another aspect to consider is the analysis method used to estimate the T_{60} from a measured RIR. If performed incorrectly, the estimation of the RIR energy decay can be hindered by background noise and multi-slope decay [17]. As such, Bayesian analysis proved suitable to model more complex behaviors in RIRs [18]. More recently, a machine-learning (ML) approach showed similar results [19]. Finally, following advancements in differentiable audio signal processing, a complete end-to-end ML approach, matching an artificial reverberator to an RIR, was proposed [20], employing both a delay-based and a velvet-noise reverberator [12].

This paper proposes to use recently developed methods to estimate and reproduce a target energy decay in conjunction with an optimized FDN-based reverberator for RIR synthesis. This design is evaluated perceptually in a formal listening test, wherein it is compared to an end-to-end deep-learning approach based on [20].

The paper is organized as follows. Section 2 provides background regarding RIR synthesis with FDNs and recent methods for FDN parameter estimation. Section 3 introduces the proposed method for the analysis-synthesis of RIRs. The evaluation and results are described in Sections 4 and 5, followed by a discussion on the outcomes in Section 6. Section 7 offers concluding remarks.

2. BACKGROUND

This section describes the relevant background on FDNs, along with the techniques utilized to achieve control over T_{60} and the echo density. This is followed by an overview of recent approaches to FDN parameter estimation.

* Also at: Media Lab, Department of Art and Media, Aalto University, FI-02150 Espoo, Finland

2.1. Feedback delay network

An FDN is a recursive system consisting of delay lines, a set of gains, and a feedback matrix through which the delay-line outputs are coupled with the delay-line inputs. The transfer function of a single-input, single-output FDN can be written as

$$H(z) = \frac{Y(z)}{X(z)} = T(z)\boldsymbol{c}^\top \left[\boldsymbol{D_m}(z)^{-1} - \boldsymbol{A}(z)\right]^{-1}\boldsymbol{b} + d\,, \quad (1)$$

where $\boldsymbol{A}(z)$ is a filter feedback matrix (FFM) consisting of $N \times N$ mixing matrices and filtering stages, where N is the number of delay lines, $\boldsymbol{D_m}(z)$ is the $N \times N$ delay matrix, vectors $\boldsymbol{b}$ and $\boldsymbol{c}$ are $N \times 1$ column vectors of input and output gains, respectively, the scalar coefficient d is the direct gain, $T(z)$ is the tone-correction filter, and the operator $(\cdot)^\top$ denotes the matrix transpose. The vector $\boldsymbol{m} = [m_1, \ldots, m_N]$ defines the delay lengths in samples. The corresponding delay matrix $\boldsymbol{D_m}(z)$ is the diagonal matrix with entries $[z^{-m_1}, \ldots, z^{-m_N}]$. The sum of the delays is the order of the system, i.e., $\mathcal{M} = \sum_{i=1}^{N} m_i$ [21].

2.2. Problem statement

This paper explores recent advancements in designing an artificial reverberator using the FDN structure for synthesizing RIRs. Given a target RIR $\tilde{h}(t)$, our aim is to find a mapping f to FDN parameters $\theta = \{\boldsymbol{m}, \boldsymbol{A}(z), T(z), \boldsymbol{c}, \boldsymbol{b}, d\}$, i.e. $\theta = f(\tilde{h}(t))$, such that for an FDN-based artificial reverberator with transfer function $H_\theta(z)$ its impulse response $h_\theta(t)$ is perceptually similar to $\tilde{h}(t)$. Throughout this paper, we refer to this task as RIR2FDN.

In this study, f is optimized with respect to a metric $\mathcal{L}$ so that

$$\min_f \mathcal{L}(h_{f(\tilde{h})}(t), \tilde{h}(t)). \quad (2)$$

While, in principle, all parameters can be derived from the target RIR, some parameters can be pre-determined using heuristic criteria. For instance, both methods in this work use a set of pre-selected delays $\boldsymbol{m}$. Perceptual similarity is further evaluated through formal listening tests.

2.3. Energy decay control in FDNs

Designing an FDN often starts by creating a lossless prototype with an energy-preserving feedback loop [22, 23]. This can be achieved by using an orthogonal feedback matrix, since it meets the condition for losslessness [24]. The advantage of initially designing a lossless FDN lies in the straightforward implementation of frequency-dependent decay that equally influences all system poles. This is realized by introducing an attenuation filter associated with each delay line in the feedback loop [5]. Such a filter is designed to approximate a frequency-dependent T_{60} by achieving a target gain-per-sample $\gamma(\omega)$ [23, 25]

$$\gamma(\omega) = 10^{-\frac{3}{f_s T_{60}(\omega)}}\,, \quad (3)$$

where ω denotes the normalized frequency in radian per second and f_s is the sampling frequency in Hz.

The magnitude response in (3) is adjusted to compensate for the delay m_i introduced by the delay line:

$$|\Gamma_i(e^{\jmath\omega})| = \gamma(\omega)^{m_i}\,, \quad (4)$$

where Γ_i is the response of the attenuation filter relative to the i^{th} delay line, $\jmath = \sqrt{-1}$, and m_i is the delay length, with $i = 1, ..., N$. Attenuation filters can be placed in the feedback matrix, i.e., $\boldsymbol{A}(z) = \boldsymbol{U}\boldsymbol{\Gamma}(z)$, where $\boldsymbol{U}$ is the orthogonal mixing matrix, and $\boldsymbol{\Gamma}(z)$ is the diagonal attenuation matrix whose diagonal entries are the delay-line-specific attenuation filters $\Gamma_i(z)$.

2.4. Scattering feedback matrix

A main challenge in designing FDNs is to generate a sufficient echo density in the RIR while maintaining computational efficiency [26]. To accelerate the echo density growth over time and reproduce a scattering-like effect, the mixing matrix $\boldsymbol{U}$ can be generalized to a filter matrix [27], where each entry is a finite impulse response (FIR) filter. The FFM is

$$\boldsymbol{A}(z) = \boldsymbol{U}(z)\boldsymbol{\Gamma}(z). \quad (5)$$

To satisfy the losslessness condition, $\boldsymbol{U}(z)$ can be realized as a paraunitary FIR filter [27] using the following factorization:

$$\boldsymbol{U}(z) = \boldsymbol{D}_{\boldsymbol{m}_K}(z)\boldsymbol{U}_K \cdots \boldsymbol{U}_2\boldsymbol{D}_{\boldsymbol{m}_1}(z)\boldsymbol{U}_1\boldsymbol{D}_{\boldsymbol{m}_0}(z), \quad (6)$$

where $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_K$ are $N \times N$ orthogonal matrices and $\boldsymbol{m}_0, \ldots, \boldsymbol{m}_K$ are vectors of N integer delays [27]. In this arrangement, the FFM incorporates $K + 1$ delays and K mixing stages into the feedback loop. To compensate for the delay introduced by $\boldsymbol{U}(z)$, we approximate the average delay as half of the maximum filter order of $\boldsymbol{U}(z)$ and add it to m_i to compute (4).

2.5. Artificial reverberator parameter estimation network

In [20], Lee et al. presented a deep-learning approach to the RIR2FDN task where an artificial reverberator parameter estimation network (ARP-net) is used as the mapping $\theta = f(\tilde{h}(t))$ to determine the FDN parameters from a target RIR $\tilde{h}(t)$ in an end-to-end manner. The ARP-net employs an encoder to convert audio spectrograms into a latent vector followed by ARP-groupwise layers for FDN parameter projection. The FDN for which the ARP-net estimates the parameters is depicted in Fig. 1. The blocks highlighted in blue represent the components whose parameters are being estimated. The structure consists of constant input and output gain vectors $\boldsymbol{b}$ and $\boldsymbol{c}$, respectively, a Householder feedback matrix $\boldsymbol{U}$, attenuation filters with common response $\Gamma(z)$, a common tone-correction filter $T(z)$, and a cascade of four Schroeder allpass (SAP) filter sections in each feedback path, denoted as $Q(z)$. The serial SAPs provide a faster echo density build-up that otherwise would be impractical to achieve in small FDNs.

To synthesize the energy decay of the reference RIR, a common absorption filter $\Gamma(z)$ was utilized [20]. It was defined as an eight-stage parametric equalizer using the state-variable filter (SVF) parameters and consists of one low-shelving, six peaking, and one high-shelving filter. The ARP-net was trained to estimate the resonance and cutoff frequencies, and the gain of each band in the filter [20]. The ability to change the cutoff frequency at each RIR can increase the generalization of the network. The tone-correction filter $T(z)$ is designed as a series of eight SVF filters each with learnable cutoff frequency, resonance, and mixing coefficients [20]. The ARP-net minimizes a multi-scale spectral loss [28] using the ℓ_1 distance between the magnitudes of the short-time Fourier transforms of $h_\theta(t)$ and $\tilde{h}(t)$ at five FFT sizes.

The structure in Fig. 1 deviates from the general FDN structure outlined in the previous sections because the attenuation filter does

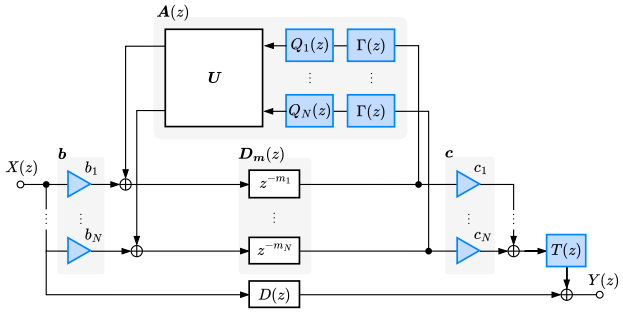

Figure 1: *Structure of the FDN [20] used in the comparison in this study. The blocks highlighted in blue represent components whose parameters are estimated by the ARP-net.*

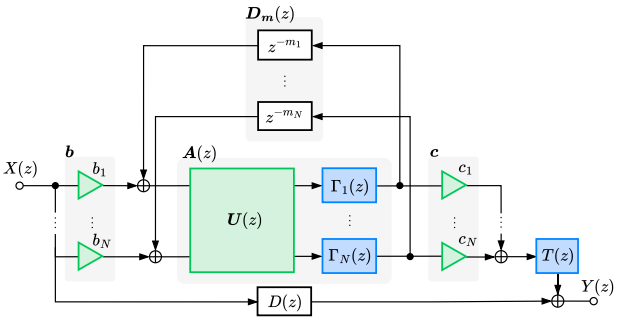

Figure 2: *Proposed DiffFDN structure [15]. The blocks highlighted in green are optimized for perceptual colorlessness. The blue blocks have RIR-dependent coefficients.*

not fulfill (4), causing varying decay rates for each feedback path. Despite this, we still denote it as an FDN throughout the paper.

2.6. Colorless FDN optimization

Motivated by the finding that coloration is minimally affected by the choice of the frequency-dependent attenuation [29], the authors presented an optimization technique for tuning homogeneous FDNs, i.e., FDNs where modes decay at the same rate, to achieve a flatter magnitude response [14]. Similarly to the ARP-net, frequency sampling was used to approximate the recursive structure of the FDN to that of an FIR and implement it in a differentiable manner. The feedback matrix, input, and output gains of a differentiable FDN were optimized using stochastic gradient descent. This optimization narrows the distribution of modal excitation, reducing the prominence of the loudest modes [29]. Listening tests confirmed its effectiveness in attenuating coloration artifacts, particularly in small FDN configurations with as few as 4 delay lines [14]. We improved the training speed and naturalness of the synthesized sound of the RIRs by including attenuation filters in the FDN structure and optimized the scattering feedback matrix to improve temporal density [15]. Moreover, the authors present a parameterization that allows for the optimization of the Householder feedback matrices, reducing computational costs both during training and operation [15].

3. PROPOSED METHOD

This section proposes an informed approach to the RIR2FDN task. The target RIR is analyzed with a neural network to estimate its energy decay [19]. Subsequently, this information is synthesized using a recently proposed attenuation filter [16] within an optimized lossless FDN [15]. The optimized FDN structure and the method used for accurate energy decay analysis and synthesis are outlined.

3.1. Differentiable temporally dense FDN

We work upon the framework presented in [15] in which the gain parameters and feedback matrix of an FDN are optimized through the Adam optimizer [30] to minimize perceptual coloration of the produced RIR by maximizing its spectral flatness. The structure of the differentiable FDN (DiffFDN) is depicted in Fig. 2, where the blocks highlighted in green represent components whose parameters are optimized for perceptual colorlessness. As opposed to Fig. 1, to increase the echo density build-up, we use a scattering feedback matrix $\boldsymbol{U}(z)$ with four stages, i.e., $K = 4$, as specified in (6). For the orthogonal matrices, $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_K$, K different optimized Householder matrices are used. The minimum number of required delay elements to implement $\boldsymbol{U}(z)$ contributes to the effective delay lengths and thus to the system order. To accelerate the echo build-up, we used the transposed configuration in which the delay lines $\boldsymbol{D_m}(z)$ are in the feedback path, and the feedback matrix $\boldsymbol{A}(z)$ is placed in the feedforward path. The transfer function of the system in Fig. 2 is

$$H(z) = T(z)\boldsymbol{c}^\top \left[\boldsymbol{I} - \boldsymbol{U}(z)\boldsymbol{\Gamma}(z)\boldsymbol{D_m}(z)\right]^{-1}\boldsymbol{U}(z)\boldsymbol{\Gamma}(z)\boldsymbol{b} + D(z),$$

where $\boldsymbol{I}$ is the identity matrix and $D(z)$ is an FIR filter used to process the direct sound.

With this configuration, the input passes immediately through $\boldsymbol{U}(z)$ where each channel undergoes mixing and convolution with the FIR filters constituting $\boldsymbol{U}(z)$. This arrangement prevents temporal gaps that are typical in systems where the feedback matrix is in the feedback path, similarly to Fig. 1.

The colorless optimization aims to minimize the mean-squared error between the magnitude response of the FDN and a target flat magnitude response using a spectral loss $\mathcal{L}_{\text{spectral}}$ while disabling the frequency-dependent filtering, i.e., $T(e^{\jmath\omega}) = 1, \boldsymbol{\Gamma}(e^{\jmath\omega}) = \gamma\boldsymbol{I}$. Additionally, the optimization encourages density in the time domain by penalizing sparseness in the coefficients of the orthogonal matrices, using a sparsity loss term $\mathcal{L}_{\text{sparsity}}$. The total loss function is

$$\mathcal{L} = \mathcal{L}_{\text{spectral}}(\boldsymbol{H}(e^{\jmath\omega})) + \mathcal{L}_{\text{sparsity}}(\boldsymbol{U})$$

and the individual loss terms are

$$\mathcal{L}_{\text{spectral}}(\boldsymbol{H}(e^{\jmath\omega})) = \sum_{i=1}^{N} \left(|H_i(e^{\jmath\omega})| - 1\right)^2 + \left(|H(e^{\jmath\omega})| - 1\right)^2,$$

$$\mathcal{L}_{\text{sparsity}}(\boldsymbol{U}) = \frac{1}{K}\sum_{k=1}^{K} \frac{N\sqrt{N} - \sum_{i,j}|U_k^{ij}|}{N(\sqrt{N}-1)},$$

where U_k^{ij} denotes the entry at coordinates i and j of the matrix at the k^{th} mixing stage, $\boldsymbol{H}(e^{\jmath\omega})$ is the multiple output FDN transfer function, evaluated on the unit circle, and $H_i(e^{\jmath\omega})$ that of the FDN's i^{th} channel, i.e.,

$$H_i(e^{\jmath\omega}) = \left(\left[\boldsymbol{I} - \boldsymbol{U}(e^{\jmath\omega})\boldsymbol{\Gamma}(e^{\jmath\omega})\boldsymbol{D_m}(e^{\jmath\omega})\right]^{-1}\boldsymbol{U}(e^{\jmath\omega})\boldsymbol{\Gamma}(e^{\jmath\omega})\boldsymbol{b}\right)_i.$$

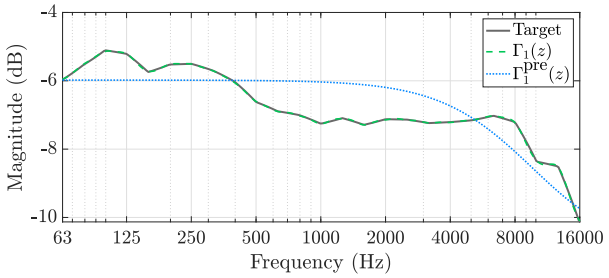

Figure 3: *Magnitude response of a given target attenuation filter, the implemented two-stage attenuation filter* $\Gamma_1(z)$*, and its first-order pre-filter* $\Gamma_1^{\text{pre}}(z)$ *for the first feedback loop delay in Fig. 2.*

3.2. Design of attenuation and tone-correction filters

In the DiffFDN, we devise an attenuation filter to satisfy (3). In doing so, it is crucial to ensure accurate estimation of $T_{60}(\omega)$. To achieve this, we compute the target magnitude response as per (4), leveraging estimates derived from DecayFitNet [19]. DecayFitNet is a neural-network-based approach to estimate RIR decay parameters from energy decay curves, which are modeled as a combination of multiple exponential decays, each characterized by an amplitude, decay time, and a noise term. Although the model offers detection of up to three slopes, the FDN structure in Fig. 2 can model only one slope, and hence we limited the parameter estimation accordingly.

After estimating $T_{60}(\omega)$, we design $\Gamma_i(z)$ utilizing the two-stage design [16] in which a pre-filter and a graphic equalizer (GEQ) are cascaded to increase the accuracy of the T_{60} approximation in each band. In the first stage, we use a first-order shelving filter, denoted as $\Gamma_i^{\text{pre}}(z)$, to approximate the general shape of the target magnitude response and to set the attenuation filter gains at dc and at the Nyquist limit. The pre-filter also shifts the command gains for the GEQ, placing them toward the optimal range of ± 12 dB [16]. For the second-stage filter, the choice of the GEQ allows replication of the details in the target magnitude response. Thus, we use the one-third-octave GEQ proposed in [31], denoted as $\Gamma_i^{\text{GEQ}}(z)$, in which the interaction between different band filters is optimized using least-squares optimization with one iteration. The transfer function of the attenuation filter then is

$$\Gamma_i(z) = \Gamma_i^{\text{pre}}(z)\Gamma_i^{\text{GEQ}}(z)\,. \tag{7}$$

Fig. 3 presents an example of $\Gamma_i(z)$ and $\Gamma_i^{\text{pre}}(z)$ for a target magnitude response and a feedback loop delay of 25.8 ms, showing an excellent match. The same GEQ structure, albeit without the pre-filter, was used to implement the tone-correction filter $T(z)$. The attenuation filters control the energy decay from $T_{60}(\omega)$ relative to the initial energy, while the tone-correction filter is used to set the actual initial energy level from which the impulse response begins to decay [2]. The desired magnitude response of $T(z)$ is derived from the amplitude of each decay component, which is estimated by DecayFitNet. To ensure the correct initial energy level across all bands, the amplitude values are normalized by the cumulative energy of the corresponding exponential decay.

4. EVALUATION SETUP

To study the performance of the presented RIR2FDN methods, we conducted a formal listening test on the perceptual similarity with target (measured) RIRs. This section presents the compared methodologies and their computational complexity.

4.1. Test configurations

The RIR2FDN task was evaluated using three different configurations of the structures presented in Sec. 3, and by comparing them to a reference RIR and anchor stimuli. The first two models were based on the DiffFDN structure presented in Fig. 2, one optimized to minimize coloration and a second initialized using random values. As such, the first model used the full DiffFDN optimization procedure described in Sec. 2.6. In the second method, referred to as RandFDN, the same reverberation structure was used, but the optimization step was replaced with random gain values and a random orthogonal feedback matrix with $K = 4$ scattering stages. In both models, the number of delay lines in $\boldsymbol{D_m}$ is set to $N = 6$, and the sampling frequency is fixed at $f_s = 48$ kHz. Any delay introduced by the filters in the feedback loop contributes to the system order. In DiffFDN and RandFDN (Fig. 2), the lengths of the delay-lines are $\boldsymbol{m} = [593, 743, 929, 1153, 1399, 1699]$. The values in $\boldsymbol{m}$ are coprime numbers distributed logarithmically, aiming to maximize the echo density [32] and avoid degenerative patterns. The scattering matrix contributes to the effective delay lengths, giving a system order of $\mathcal{M} = 8457$.

For the RIR2FDN tasks, both the optimized DiffFDN and the RandFDN structures were used as lossless prototypes, and the attenuation and tone-correction parameters were tuned according to the target RIRs. As described in 3.2, the filter designs were based on energy decay estimation from the DecayFitNet [19].

The third method evaluated is the ARP-net structure [20] (Fig. 1), which synthesizes a target RIR by inferring from a trained network without any additional parameter tuning. We replicated the model based on the specifications in [20] and additional clarifications provided by the authors. The network has about 7.39M parameters and was trained for the analysis-synthesis task using a dataset of 50k RIRs generated using shoebox room simulations based on the image-source method [33, 34]. Different conditions were simulated by varying parameters such as room size, frequency-dependent wall absorption coefficients, and source/microphone positions. In the ARP-net structure (Fig. 1), the effective delay lengths need to be computed by considering both delays in $\boldsymbol{D_m}(z)$ and those introduced by the SAP filters. In total the delays are $[1205, 1291, 1399, 1437, 1547, 1583]$ giving a system order $\mathcal{M} = 8462$.

The anchor was synthesized according to Fig. 2, with parameters designed to simulate various types of degradation resulting from poor analysis-synthesis practices. To simulate errors in the filter design, the target magnitude of the attenuation and tone correction was perturbed by a random variation of up to 25% of its original value. The scattering matrix $\boldsymbol{U}(z)$ was designed to enhance sparseness and repetitiveness in the response. To further degrade its response, we applied a bandpass filter with cut-off frequencies 100 Hz and 1 kHz and 12-dB roll-off. For the reference sound, direct convolution of the measured RIR was used.

To synthesize the direct sound, approximately 2 ms from the onset of the reference RIR were used to design the FIR filter $D(z)$ in Figs. 1 and 2. Its amplitude remained unchanged from the reference RIR, while the energy of the remaining synthetic RIR was adjusted to match the root mean square of the corresponding section of the reference RIR. Audio examples and configuration details are available online [1]. The PyTorch implementations of the

Table 1: *Number of operations for the FDN structures used by the DiffFDN and ARP-net. For the tested configuration, i.e. $N = 6$ delay lines and $K = 4$ mixing stages, the total operation count is 132, excluding the filters. With the inclusion of filters, the operation count rises to 2808 for DiffFDN and 790 for ARP-net.*

FDN Type	$\boldsymbol{b}, \boldsymbol{c}$	$\boldsymbol{D_m}(z)$	$\boldsymbol{U}(z)$	$\Gamma(z)$	$T(z)$
DiffFDN	$2N$	$2N$	$2N(2K+1)$	$384N$	372
ARP-net	$2N$	$2N$	$18N$	$94N$	94

DiffFDN and ARP-net are offered in the dedicated repository [2].

4.2. Computational complexity

The computational complexity during operation of the DiffFDN and ARP-net, when filters are excluded from the calculation, is the same for their respective FDNs of equal size $N = 6$, and $K = 4$. However, the filters used by the DiffFDN increase its computational complexity compared to the ARP-net. Table 1 shows the number of multiply-and-add operations performed during operation. The operation count for RandFDN is not reported since it shares the same structure and number of operations as DiffFDN. While the delay line itself does not involve any multiplication or addition operations, it requires write and read operations, each counted as one. The necessary operations add up to $2N$ for the input and output gains $\boldsymbol{b}$ and $\boldsymbol{c}$, $2N$ for the main delay lines $\boldsymbol{D_m}(z)$, and $2N$ for the Householder matrix multiplication. The scattering matrix multiplication, involving K mixing stages and $K + 1$ delays, requires $2N(2K + 1)$ operations. Each section in the SAP filters requires $4N$ operations. The DiffFDN employs a one-third-octave band GEQ for both the attenuation and the tone-correction filters, alongside a shelving filter for modeling the global energy decay. The shelving filter and each band of the GEQ are implemented as biquads, necessitating 12 operations each. In total, the DiffFDN employs 32 biquads for $\Gamma_i(z)$ and 31 for $T(z)$. The equalizers used by the ARP-net for both $\Gamma(z)$ and $T(z)$ comprise eight bands, resulting in 94 operations each.

5. SUBJECTIVE EVALUATION

We conducted a formal listening test to assess the perceptual similarity of the RIR2FDN task presented in the previous section. In the following, the RIRs selected for the test are presented, followed by a description of the listening test setup and results.

5.1. Measured RIRs

We evaluated the models on the RIR2FDN task using seven RIRs selected from the MIT RIR survey collection of real-world RIRs [35]. These RIRs were recorded in real-world environments with a 1.5-m distance between the sound source and the receiver, utilizing the same speaker and microphone across measurements.

The selected RIRs encompass various spaces, including a lobby, dining room, hallway, meeting room, classroom, bathroom, and bedroom. Of these, the hallway is the most reverberant, with a T_{60} value of 2.02 s. This value is calculated as the mean of the T_{60} values across the one-octave bands from 200 Hz to 2 kHz. The meeting room has the shortest reverb ($T_{60} = 0.24$ s). The initial energy level, corresponding to the energy at onset time, shows similar behavior across spaces, mostly due to the measurement loudspeaker magnitude response.

5.2. Listening test procedure

The test was based on the principles of the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) recommendation [36], although it was not strictly adhered to. The test was performed using the web audio API-based experiment software webMUSHRA developed by International Audio Laboratories Erlangen [37]. The experiment was conducted in sound-insulated booths at the Aalto Acoustics Lab, with participants wearing Sennheiser HD 650 headphones. The final items were presented to a group of 16 listeners comprising 12 men and 4 women. The participants' mean age was 28.6 years, with a standard deviation of 3.8, and none of them reported any hearing impairments. All but two participants were either students or employees of the Aalto University Acoustics Lab. All participants had previous experience with listening tests.

On each page of the listening test, five reverberated sounds were compared against a reference. We selected three anechoic sounds with different time-frequency characteristics: a speech signal, a drum loop, and a saxophone sound. In total, 21 reverberation conditions were assessed, seven for each sound source. On each page, a hidden reference and an anchor were present, and the subjects were instructed to rate them as 100 and 0, respectively, upon detection. The other tested conditions, as detailed in Sec. 4.1, included DiffFDN, RandFDN, and ARP-net. We adopted an anchor design approach different from the standard MUSHRA recommendation to ensure the differences were on the same scale. Before the test, two training pages were presented to familiarize the subjects with the sound samples. Adjustments to the overall loudness were allowed during the initial training page but remained constant throughout the test. During the evaluation, participants rated the overall similarity between the reference sound and each presented item using a scale ranging from 0 to 100.

5.3. Results

For the following analysis, the results of two subjects were excluded. One subject was excluded for failing to detect the anchor, and the other for failing to detect the reference, in more than 15% of the tested reverb conditions. The outcomes of the listening test are illustrated in Fig. 4, across sound sources, and in Fig. 5 across RIRs. In each box, the median is represented by the central mark, whereas the lower and upper edges indicate the 25th and 75th percentiles, respectively. The whiskers extend to encompass the most extreme data points not identified as outliers, with any outliers plotted separately. The shaded regions surrounding the medians facilitate the comparison of sample medians across various box charts. Non-overlapping shaded regions signify differing medians between the compared box charts at the 5% significance level, assuming a normal distribution.

A Shapiro-Wilk test [38] confirmed that the data deviates from the normal distribution, even when reference and anchor are excluded. Additionally, we used the Wilcoxon signed-rank test [39] to assess the score distributions for each pair of conditions. To address multiple comparisons (10 hypotheses per page), we applied the Bonferroni method to adjust the alpha level, i.e., the threshold of the p-value for statistical significance, to 0.005. The p-value

[1] `http://research.spa.aalto.fi/publications/papers/dafx24-rir2fdn/`

[2] `https://github.com/gdalsanto/rir2fdn`

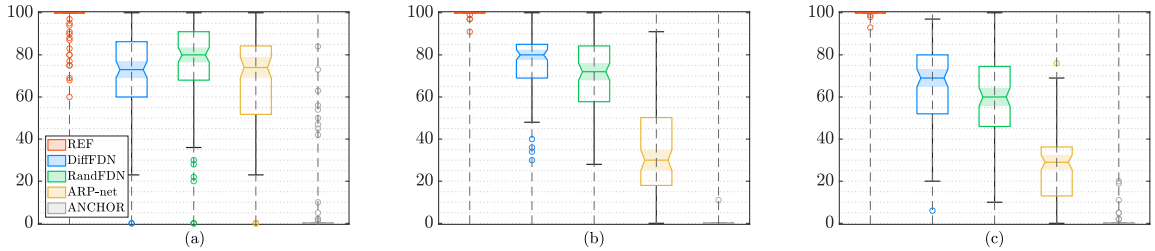

Figure 4: *Results of the listening test across sound sources:* (a) *saxophone,* (b) *speech, and* (c) *drum loop. The scores given to the saxophone signal show no statistically significant difference. On the other hand, for the speech and drum loop, the subjects could distinguish between DiffFDN and ARP-net, rating higher similarity with the reference in the former.*

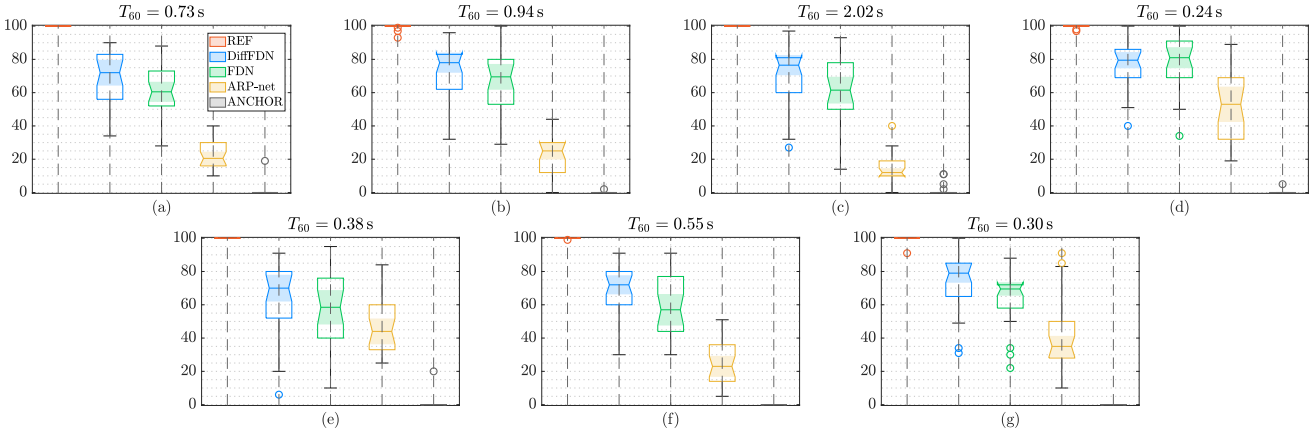

Figure 5: *Results of the listening test across RIRs:* (a) *lobby,* (b) *dining room,* (c) *hallway,* (d) *meeting room,* (e) *classroom,* (f) *bathroom, and* (g) *bedroom. The scores assigned to the items contain the most critical test signals, excluding the saxophone. The proposed DiffFDN is rated consistently higher than both RandFDN and ARP-net, apart from (d). The T_{60} values are stated in the title of each pane.*

represents the likelihood of obtaining the observed difference between the paired samples if there were no true differences between the populations from which the samples were drawn. When the results are analyzed on isolated sound sources (Fig. 4), the p-values indicate statistically significant differences among all pairs of results only when the drum loop is used as the source signal.

For speech signals, the scores given to DiffFDN and RandFDN converge, leading to a p-value above the alpha level ($p = 0.006$). The scores assigned to the three tested conditions for the saxophone show no statistically significant differences ($p = 0.022$ for DiffFDN–RandFDN, $p = 0.472$ for DiffFDN–ARP-net, and $p = 0.068$ for RandFDN–ARP-net). This suggests that the saxophone is not a critical test signal for this task. Consequently, scores assigned to the pages containing the saxophone as a sound source have been excluded from Fig. 5, where the listening test results are shown across RIRs. When the results are analyzed within each RIR, the DiffFDN and RandFDN conditions lead to a p-value above the alpha level of all tested rooms. Only for the classroom, the scores given to RandFDN and ARP-net show no statistically significant difference ($p = 0.073$).

6. DISCUSSION

The selection of the source sound may impact the artificial reverberator's performance, as shown in Fig. 4. Transient sounds, similar to the drum loop, represent a broadband excitation, allowing for a comprehensive assessment of the spectrum and facilitating the distinction between different RIRs. The low scores received by the ARP-net in Fig. 4(c) suggest some challenges in effectively suppressing the strong temporal repetitions inherent in small FDNs, resulting in metallic-sounding RIRs. Conversely, a harmonic sound with a smooth temporal envelope, such as that of a saxophone, lacks a percussive part to trigger a broadband reverb effect. In Fig. 4 (a), several outliers are observed in the results for both reference and anchor conditions, confirming that, for harmonic signals with a sparse spectrum, precise reverberation might be less crucial. The speech signal in Fig. 4 (b) falls somewhere in between, as it comprises both harmonic and noise-like elements.

Fig. 5 shows the listening test results across the different RIRs. For cases (a), (b), and (c), the results are more separated when compared to the remaining RIRs, probably due to the extended release time that can be used to perceive and distinguish the reverberation characteristics. The classroom's T_{60} is 0.38 s, which is one-fourth of that of the hallway, and exhibits the least noticeable difference in the scores assigned to DiffFDN, RandFDN, and ARP-net. Except for the meeting room, the DiffFDN was consistently rated higher than the other conditions.

To further investigate these results, Figs. 6 and 7 show the $T_{60}(\omega)$ and initial energy level synthesized by the DiffFDN and ARP-net for the (a) lobby, (b) hallway, and (c) meeting room. The values were computed using DecayFitNet on both the full synthesized RIRs and reference RIRs. It is clear from Fig. 6 that none of the methods achieved an accurate decay rate. The values are outside the just noticeable difference (JND) of 5% [40], indicated by the shade around the target $T_{60}(\omega)$. Surprisingly, the attenuation filter utilized by the ARP-net demonstrates the capability to converge towards the global energy decay despite not conforming

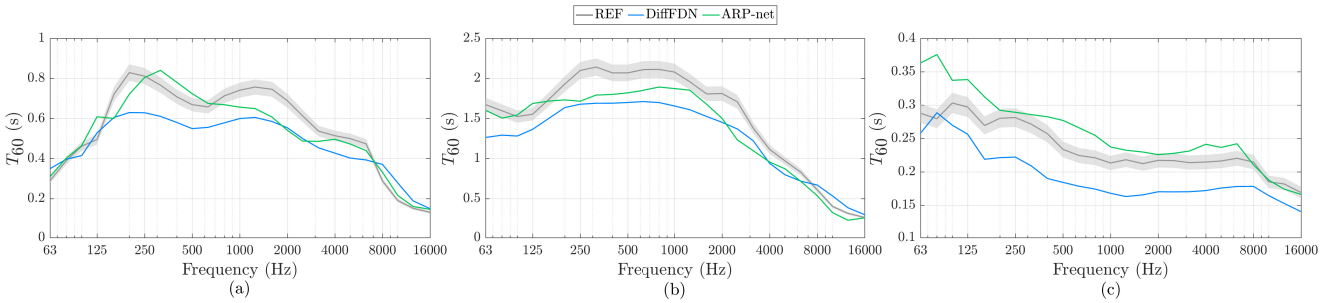

Figure 6: *Reverberation times T_{60} of the reference and synthesized RIRs corresponding to* (a) *lobby,* (b) *hallway, and* (c) *meeting room. The T_{60} values were estimated using DecayFitNet on one-third-octave bands and linearly interpolated. Reverberation times T_{60} of both DiffFDN and ARP-net exhibit errors greater than the JND, as indicated by the shading around the target curve.*

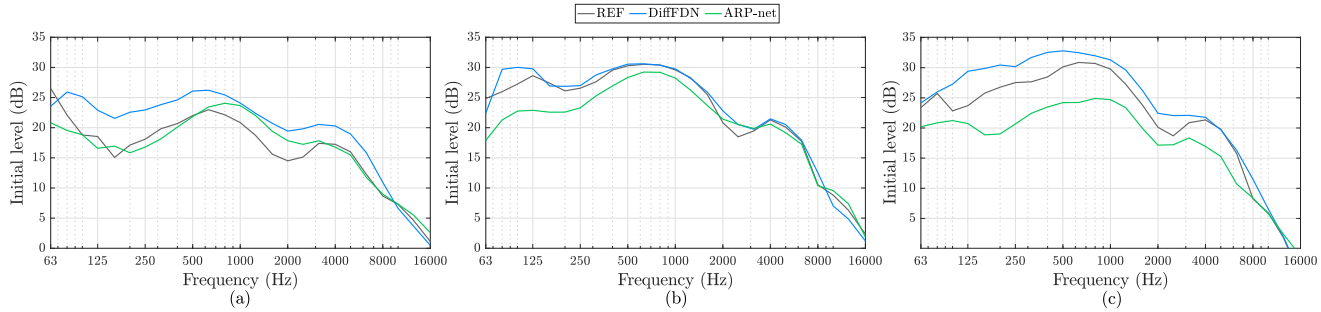

Figure 7: *Initial energy level of the reference and synthesized RIRs corresponding to* (a) *lobby,* (b) *hallway, and* (c) *meeting room. The values were calculated using the DecayFitNet on one-third-octave bands and linearly interpolated.*

to (4). The initial energy levels also exhibit errors, except for the mid-high frequencies of the hallway. The errors observed could potentially be attributed to neglecting the background noise and the multi-slope decay in the models. However, further studies are necessary to understand the cause of this discrepancy, which persists despite the use of accurate filters. This analysis, along with the results of the listening test, suggests that perceptual differences are only partially influenced by variations in coloration and T_{60}. Time artifacts can drastically affect the performance of the FDN. The structure employed in both DiffFDN and RandFDN (Fig. 2) appears to be more successful in concealing these artifacts.

The ARP-net, being a neural network, relies on its training data. Since gathering a large number of measured RIRs is challenging and time-consuming, it prompts a transition to synthetic datasets, although they may contain unnatural artifacts. In this study, we tested the DiffFDN and the non-optimized FDN (RandFDN) to assess colorless optimization benefits. The DiffFDN outperformed RandFDN in most situations, indicating optimization advantages. Although the median was lower in the RandFDN scenario in all but one case, the confidence intervals overlapped. This could be attributed to the presence of ARP-net, which possesses attributes different from those of DiffFDN and RandFDN. The similarity between the scores of DiffFDN and RandFDN may be due to both systems using attenuation and tone-correction filters based on DecayFitNet estimations and sharing the same scattering FDN core structure (Fig. 2).

7. CONCLUSIONS

This work presents advancements in artificial reverberator design using the FDN structure for synthesizing measured RIRs. The objective is to establish a mapping between RIR and FDN parameters optimized for perceptual similarity. We introduce an optimized method for designing artificial reverberation utilizing a recently developed differentiable FDN (DiffFDN) employing as few as six delay lines. This method incorporates a neural-network-based energy decay estimator for accurate estimation of T_{60} and a recently developed two-stage attenuation filter design.

Listening-test results demonstrate that the proposed method surpasses the state-of-the-art end-to-end approach. Moreover, the outcome of the listening test is highly dependent on the synthesized RIR and the excitation signal. Further analysis indicates that accurate reproduction remains a challenge. In this regard, this study provides insight into where future efforts should be focused.

8. ACKNOWLEDGMENTS

The work of the first author was funded by the Aalto University School of Electrical Engineering. Part of the research was conducted at IRCAM, Paris, France, where the first author was a visitor from September to December 2023, supported by a Foundation for Aalto University Science and Technology grant.

9. REFERENCES

[1] V. Välimäki, J. D. Parker, L. Savioja, et al., "Fifty years of artificial reverberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 5, pp. 1421–1448, Jul. 2012.

[2] J.-M. Jot, "An analysis/synthesis approach to real-time artificial reverberation," in *Proc. ICASSP*, 1992, pp. 221–224.

[3] M. R. Schroeder, "Natural sounding artificial reverberation," *J. Audio Eng. Soc.*, vol. 10, no. 3, pp. 219–213, Jul. 1962.

[4] J. A. Moorer, "About this reverberation business," *Computer Music J.*, vol. 3, no. 2, pp. 13–28, Jan. 1979.

[5] J.-M. Jot and A. Chaigne, "Digital delay networks for designing artificial reverberators," in *Proc. 90th AES Conv.*, Paris, France, Feb. 1991.

[6] M. Chemistruck, K. Marcolini, and W. Pirkle, "Generating matrix coefficients for feedback delay networks using genetic algorithm," in *Proc. 133rd AES Conv.*, Oct. 2012.

[7] J. Stautner and M. Puckette, "Designing multi-channel reverberators," *Computer Music J.*, vol. 6, no. 1, pp. 52–65, Spring 1982.

[8] J. Coggin and W. Pirkle, "Automatic design of feedback delay network reverb parameters for impulse response matching," in *Proc. 141st AES Conv.*, Sep. 2016.

[9] I. Ibnyahya and J. D. Reiss, "A method for matching room impulse responses with feedback delay networks," in *Proc. 153rd AES Conv.*, 2022.

[10] J. Shen and R. Duraiswami, "Data-driven feedback delay network construction for real-time virtual room acoustics," in *Proc. 15th Int. Audio Mostly Conf.*, 2020, pp. 46–52.

[11] R. Bona, D. Fantini, G. Presti, et al., "Automatic parameters tuning of late reverberation algorithms for audio augmented reality," in *Proc. 17th Int. Audio Mostly Conf.*, New York, NY, USA, Sep. 2022, p. 36–43.

[12] V. Välimäki, B. Holm-Rasmussen, B. Alary, and H.-M. Lehtonen, "Late reverberation synthesis using filtered velvet noise," *Appl. Sci.*, vol. 7, no. 483, May 2017.

[13] J. Fagerström, B. Alary, S. Schlecht, and V. Välimäki, "Velvet-noise feedback delay network," in *Proc. DAFx*, Austria, Sep. 2020, pp. 219–226.

[14] G. Dal Santo, K. Prawda, S. J. Schlecht, and V. Välimäki, "Differentiable feedback delay network for colorless reverberation," in *Proc. DAFx*, Sep. 2023, pp. 244–251.

[15] G. Dal Santo, K. Prawda, S. J. Schlecht, and V. Välimäki, "Feedback delay network optimization," *arXiv preprint arXiv:2402.11216*, 2024.

[16] V. Välimäki, K. Prawda, and S. J. Schlecht, "Two-stage attenuation filter for artificial reverberation," *IEEE Signal Process. Lett.*, vol. 31, pp. 391–395, Jan. 2024.

[17] M. Karjalainen, P. Antsalo, A. Mäkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy response measurements," *J. Audio Eng. Soc.*, vol. 50, no. 11, pp. 867–878, Nov. 2002.

[18] N. Xiang, P. Goggans, T. Jasa, and P. Robinson, "Bayesian characterization of multiple-slope sound energy decays in coupled-volume systems," *J. Acoust. Soc. Am.*, vol. 129, no. 2, pp. 741–752, Feb. 2011.

[19] G. Götz, R. Falcón Pérez, S. J. Schlecht, et al., "Neural network for multi-exponential sound energy decay analysis," *J. Acoust. Soc. Am.*, vol. 152, no. 2, pp. 942–953, Aug. 2022.

[20] S. Lee, H.-S. Choi, and K. Lee, "Differentiable artificial reverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2541–2556, Jul. 2022.

[21] D. Rocchesso and J. O. Smith, "Circulant and elliptic feedback delay networks for artificial reverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 5, no. 1, pp. 51–63, Jan. 1997.

[22] J.-M. Jot, "Proportional parametric equalizers—Application to digital reverberation and environmental audio processing," in *Proc. 139th AES Conv.*, Oct. 2015.

[23] S. J. Schlecht and E. A. Habets, "Accurate reverberation time control in feedback delay networks," in *Proc. DAFx*, Edinburgh, UK, Sep. 2017, pp. 337–344.

[24] S. J. Schlecht and E. A. Habets, "On lossless feedback delay networks," *IEEE Trans. Signal Process.*, vol. 65, no. 6, pp. 1554–1564, Jun. 2016.

[25] K. Prawda, S. J. Schlecht, and V. Välimäki, "Improved reverberation time control for feedback delay networks," in *Proc. DAFx*, Sep. 2019, pp. 299–306.

[26] S. J. Schlecht and E. A. Habets, "Dense reverberation with delay feedback matrices," in *Proc. IEEE WASPAA*, Oct. 2019, pp. 150–154.

[27] S. J. Schlecht and E. A. Habets, "Scattering in feedback delay networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1915–1924, Oct. 2020.

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

[29] J. Heldmann and S. J. Schlecht, "The role of modal excitation in colorless reverberation," in *Proc. DAFx*, Sep. 2021, pp. 206–213.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] J. Liski and V. Välimäki, "The quest for the best graphic equalizer," in *Proc. DAFx*, Sep. 2017, pp. 95–102.

[32] S. J. Schlecht and E. A. Habets, "Feedback delay networks: Echo density and mixing time," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 2, pp. 374–383, 2016.

[33] P. Svensson and U. R. Kristiansen, "Computational modelling and simulation of acoustic spaces," in *Proc. AES 22nd Int. Conf. Virt. Synth. Ent. Audio*, Jun. 2002.

[34] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE ICASSP*, Apr. 2018, pp. 351–355.

[35] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proc. Natl. Acad. Sci*, vol. 113, no. 48, pp. E7856–E7865, Nov. 2016.

[36] ITU, "Method for the subjective assessment of intermediate quality level of audio systems," Recommendation ITU-R BS.1534-3, Oct. 2015.

[37] M. Schoeffler, S. Bartoschek, F.-R. Stöter, et al., "Web MUSHRA—A comprehensive framework for web-based listening tests," *J. Open Res. Softwr.*, vol. 6, no. 1, 2018.

[38] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, Dec. 1965.

[39] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[40] ISO, "Acoustics — measurement of room acoustic parameters — part 1: Performance spaces," Standard, International Organization for Standardization, Geneva, CH, Jun. 2009.