DIFFERENTIABLE FEEDBACK DELAY NETWORK FOR COLORLESS REVERBERATION

Gloria Dal Santo, Karolina Prawda, Sebastian J. Schlecht, and Vesa Välimäki*

Acoustics Lab,
Department of Information and Communications Engineering,
Aalto University,
FI-02150 Espoo, Finland
{gloria.dalsanto, karolina.prawda, sebastian.schlecht, vesa.valimaki}@aalto.fi

ABSTRACT

Artificial reverberation algorithms often suffer from spectral coloration, usually in the form of metallic ringing, which impairs the perceived quality of sound. This paper proposes a method to reduce the coloration in the feedback delay network (FDN), a popular artificial reverberation algorithm. An optimization framework is employed entailing a differentiable FDN to learn a set of parameters decreasing coloration. The optimization objective is to minimize the spectral loss to obtain a flat magnitude response, with an additional temporal loss term to control the sparseness of the impulse response. The objective evaluation of the method shows a favorable narrower distribution of modal excitation while retaining the impulse response density. The subjective evaluation demonstrates that the proposed method lowers perceptual coloration of late reverberation, and also shows that the suggested optimization improves sound quality for small FDN sizes. The method proposed in this work constitutes an improvement in the design of accurate and high-quality artificial reverberation, simultaneously offering computational savings.

1. INTRODUCTION

Since the pioneering work of Schroeder and Logan [1], delay-based digital recursive structures have been used in reverberation synthesis [2]. Nowadays, one of the most widely used approaches in artificial reverberation is the feedback delay network (FDN), a system that generalizes the parallel comb-filter structure by interconnecting delays via a feedback matrix [3, 4, 5]. In FDNs, a commonly used approach is to first design a lossless prototype [6] to then achieve the desired frequency-dependent decay with attenuation filters [7, 8]. However, a common bane of systems utilizing comb filters is sound coloration [1]. Strong coloration is undesirable in artificial reverberation since it impairs the perceived sound quality.

Recent research suggests using modal decomposition to study the properties of the FDN in more detail [9]. The modal decomposition showed that the coloration in an FDN is related to the wide distribution of modal excitation values. In particular, modes with strong excitations are perceived as metallic ringing [10]. The modal excitation depends on all FDN parameters, and directly improving the coloration remains challenging. Recently, Schlecht proposed a method to achieve a uniform magnitude response and found the necessary conditions for an allpass FDN [11]. However, this approach suffers from temporal buildup of echoes [10], thus leaving the need for a more versatile method to design colorless FDNs.

Although many well-known reverb topologies, such as the Moorer-Schroeder [12], can be translated into FDN designs, the design of FDNs still presents several unresolved challenges. These arise from the inherent trade-off between computational complexity, mode density, and echo density. The cost of implementing the matrix-vector-multiplication for a single time step in an FDN increases with the number of delay lines and varies depending on the type of feedback matrix. However, the number of delay lines cannot be arbitrarily low, as there are certain dependencies between the delay lengths that become more severe as the number of delays decreases. In addition, a smaller number of delays decreases both the modal and the echo density, which leads to metallic sounding artifacts [13, 14].

Automatic tuning of FDN parameters has been previously explored in the literature, with genetic algorithms being widely used [15, 16, 17]. More recently, a multi-stage approach was employed to optimize FDN parameters to match a target room impulse response (IR)[18]. The input, output, direct gains, and delay lengths were optimized using a genetic algorithm. However, differences between the model and the target IR were revealed in the listening tests. To circumvent the challenge of optimizing infinite-impulse-response filters with differentiable machine-learning techniques, frequency sampling was used to implement a differentiable approximation of delay networks. An end-to-end deep-learning model was presented for the estimation of parameters, although only the absorption filters and input and output filters were estimated [19].

In this study, we present a novel approach to design FDNs for colorless artificial reverberation. To this end, we use a differentiable FDN (DiffFDN) in an optimization framework to learn a set of FDN parameters leading to less coloration. Specifically, we show that a narrower modal excitation distribution can be achieved without requiring the allpass property, offering more flexibility since the reverberation time (RT) values can be arbitrarily set after designing the prototype FDN. The perceptual evaluation against several common FDN designs shows that the proposed method successfully decreased perceived coloration.

The paper is organized as follows. Section 2 offers background information about FDNs and their modal decomposition. Section 3 introduces the proposed method of designing colorless FDNs. The results of the objective evaluation are presented in Section 4, and Section 5 shows the results of the listening test. Section 6 offers concluding remarks.

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

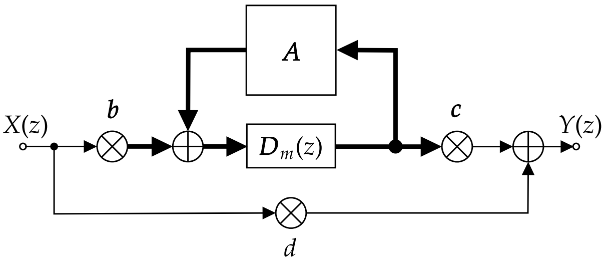

Figure 1: *Block diagram of a SISO FDN. Thin and thick lines indicate single- and multichannel connections, respectively.*

2. BACKGROUND

This section gives some background information about FDNs and presents related concepts that are relevant to the proposed method, such as modal decomposition and homogeneous decay in FDNs.

2.1. Feedback Delay Network

An FDN is a recursive system consisting of delay lines, a set of gains, and a scalar feedback matrix through which the delay outputs are coupled to the delay inputs. An example of a simple single-input single-output (SISO) FDN architecture is presented in Fig. 1. The transfer function of the FDN is

$$H(z) = \boldsymbol{c}^\top \left[\boldsymbol{D}_{\boldsymbol{m}}(z)^{-1} - \boldsymbol{A}\right]^{-1} \boldsymbol{b} + d \,, \tag{1}$$

where $\boldsymbol{A}$ is the $N \times N$ feedback matrix, N being the number of delay lines. The $N \times 1$ column vectors $\boldsymbol{b}$ and $\boldsymbol{c}$ and the scalar coefficient d respectively represent the input, output, and direct gains. The operator $(\cdot)^\top$ denotes the transpose. The vector $\boldsymbol{m} = [m_1, \dots, m_N]$ defines the lengths of delays in samples. The corresponding delay matrix $\boldsymbol{D}_{\boldsymbol{m}}(z)$ is created by taking a diagonal matrix with entries given by $[z^{-m_1}, \dots, z^{-m_N}]$.

The system poles λ_i are the roots of the generalized characteristic polynomial $p(z)$ of the system, which is fully characterized by $\boldsymbol{m}$ and $\boldsymbol{A}$:

$$p(z) = \det(\boldsymbol{D}_{\boldsymbol{m}}(z)^{-1} - \boldsymbol{A}) \,. \tag{2}$$

The sum of the delays gives the order of the system, i.e., $\mathcal{M} = \sum_{i=1}^{N} m_i$ [20].

2.2. Modal Decomposition

The IR of the FDN can be represented as the sum of complex one-pole modes, or resonators, in the time domain [9]:

$$h(n) = \sum_{i=1}^{\mathcal{M}} h_i(n) \,. \tag{3}$$

Each mode $h_i(n)$ is defined by the pole λ_i and the residue ρ_i:

$$h_i(n) = |\rho_i||\lambda_i|^n e^{\jmath(n\angle\lambda_i + \angle\rho_i)} \,, \tag{4}$$

where $|\cdot|$ is the magnitude, $\angle$ indicates the argument of a complex number in radians, $\jmath = \sqrt{-1}$, and n indicates the discrete time index. The sum of the delay-line lengths $\mathcal{M}$ coincides with the number of poles.

The transfer function of the FDN (1) can be represented in terms of its poles and residues from its partial fraction decomposition as

$$H(z) = d + \sum_{i=1}^{\mathcal{M}} \frac{\rho_i}{1 - \lambda_i z^{-1}} \,, \tag{5}$$

which is often referred to as the modal decomposition of the FDN [9]. The excitation and initial phase of the i^{th} mode are determined by the magnitude $|\rho_i|$ and phase $\angle\rho_i$, respectively, of its corresponding residue, whereas the magnitude and the phase of the i^{th} pole, $|\lambda_i|$ and $\angle\lambda_i$, respectively, determine its decay rate and frequency.

2.3. Homogeneous FDN

For the design of an artificial reverberator, starting with a lossless prototype is beneficial. The FDN is said to be lossless if the roots of $p(z)$ have magnitude equal to one, i.e., $|\lambda_i| = 1$ for all i [21]. Frequency-dependent RT, here also denoted as T_{60}, is then easily achieved by extending the delay lines with a frequency-dependent attenuation filter [4].

In this study, we focus only on the specific case of frequency-independent homogeneous decay. This refers to the case where all modes experience the same rate of decay, i.e., $|\lambda_i| = \gamma$ for all i. Homogeneous decay is achieved with a feedback matrix $\boldsymbol{A}$ being the product of a unilossless matrix $\boldsymbol{U}$ and a diagonal matrix $\boldsymbol{\Gamma}$, whose entries are delay-proportional absorption coefficients, $\boldsymbol{\Gamma} = \text{diag}(\gamma^{\boldsymbol{m}})$. The feedback matrix can be expressed as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Gamma} \,. \tag{6}$$

A matrix $\boldsymbol{U}$ is unilossless if, regardless of the choice of delays $\boldsymbol{m}$, its eigenvalues are unimodular and its eigenvectors are linearly independent. A matrix $\boldsymbol{U}$ satisfying the unitary condition, $\boldsymbol{U}\boldsymbol{U}^H = \boldsymbol{I}$, is also unilossless [22, 23]. As $\boldsymbol{U}$ is unilossless, the modal decay is controlled entirely by gain-per-sample parameter γ, where $0 \leqslant \gamma \leqslant 1$. The gain-per-sample in dB is

$$\gamma_{\text{dB}} = \frac{-60}{f_{\text{s}} T_{60}} \,, \tag{7}$$

where f_{s} is the sampling rate in Hz and T_{60} is the reverberation time defined as the time required for the sound level to decay by 60 dB from the initial steady-state value.

2.4. Coloration in FDN

In artificial reverberation, the properties of the resonating modes have direct implications on coloration. A flat magnitude response, implicitly achieved by the allpass property, is often desirable.

Schroeder and Logan [1] made the initial attempt to produce colorless artificial reverberation by establishing specific requirements for the reverberators in addition to a flat frequency response. Overlapping normal modes across all frequencies, equal RTs for each mode, sufficient echo density, lack of periodicity in the time domain, and no periodic or comb-like frequency responses were deemed necessary to achieve colorlessness [1]. Despite fulfilling the aforementioned conditions, however, the Schoreder series allpass did not attain complete colorlessness.

A recent study was conducted to further understand the role of modal excitation in late reverberation coloration [10]. Listening

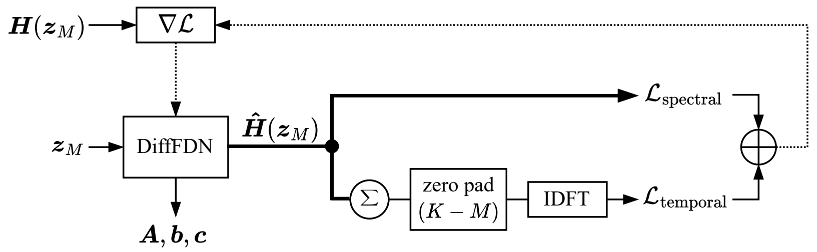
$\boldsymbol{H}(\boldsymbol{z}_M)$
$\nabla\mathcal{L}$
$\boldsymbol{z}_M$
DiffFDN
$\hat{\boldsymbol{H}}(\boldsymbol{z}_M)$
$\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}$
$\mathcal{L}_{\text{spectral}}$
Σ
zero pad $(K - M)$
IDFT
$\mathcal{L}_{\text{temporal}}$

Figure 2: *Architecture of the proposed optimization workflow. Dotted lines indicate the stochastic gradient descent method of optimizing the parameters. Thin and thick lines indicate single- and multichannel connections, respectively.*

test results suggest that a narrow distribution of the modal excitation values $|\rho_i|$ tends to result in a flatter magnitude response. For large values of $|\rho_i|$, coloration starts to become noticeable. In agreement with [13], the study found that the perception of colorlessness correlates with the number of modes, and that more than 6000 modes are needed for an IR to be perceived as rather colorless.

The literature also shows that, for large values of $\mathcal{M}$, the modes of the FDNs are uniformly distributed [9], preventing additional coloration that usually results from clusters of modes. Nonetheless, a flat magnitude response and a uniform modal frequency distribution are insufficient to achieve colorlessness.

When the feedback matrix $\boldsymbol{A}$ is diagonal, the FDN takes the form of a parallel comb-filter structure. If the FDN is homogeneous, the transfer function in (1) is equivalent to a combination of comb filters, where each filter has the transfer function

$$H_i(z) = \frac{1}{1 + \gamma^{m_i} z^{-m_i}} \, . \tag{8}$$

The contribution of each filter to the total energy of the response can be calculated as

$$\|H_i(z)\|^2 = \int_0^{2\pi} |H_i(e^{\imath\omega})|^2 \mathrm{d}\omega \tag{9}$$

$$= \frac{1}{1 - \gamma^{2m_i}} \, . \tag{10}$$

Fundamentally, shorter delays m_i contribute more energy and produce strong, audible metallic-sounding comb peaks, whereas longer delays m_i contribute less energy and tend to be masked by the more dominant comb filters. In order to achieve colorless FDNs, we aim to avoid strongly recirculating short delays and encourage strongly exciting long delays.

2.5. Problem Statement

In this paper, we aim to optimize the feedback delay matrix $\boldsymbol{A}$, and input and output gains $\boldsymbol{b}$ and $\boldsymbol{c}$ such that the resulting IR is colorless. In this study, we keep the number and lengths of the delays fixed.

From previous studies, we know that coloration is little impacted by the choice of the frequency-dependent attenuation [10]. Thus, the optimization is performed on a long-ringing frequency-independent prototype FDN.

The proposed method utilizes two losses to improve coloration and temporal density. A stochastic gradient descent scheme is used to avoid convergence at spurious local minima. A parameter remapping guarantees a lossless FDN prototype at each optimization step.

3. FDN OPTIMIZATION

In the following, we present a method to reduce coloration in an FDN response for arbitrary RTs. Stochastic gradient descent is used to optimize the parameters of a differentiable FDN.

3.1. Differentiable FDN

This work applies the frequency-sampling method to approximate an FDN as a finite-impulse-response (FIR) filter. This is done by evaluating the delay matrix $\boldsymbol{D_m}(\boldsymbol{z_M})$ at the discrete frequency points in the vector

$$\boldsymbol{z}_M = [e^{\jmath\pi\frac{0}{M}}, e^{\jmath\pi\frac{1}{M}}, \dots, e^{\jmath\pi\frac{M-1}{M}}], \tag{11}$$

where M indicates the total number of frequency bins equally distributed on the unit circle. The discrete-frequency transfer function of the FDN thus becomes

$$H(\boldsymbol{z}_M) = \boldsymbol{c}^\top \big[\boldsymbol{D_m}(\boldsymbol{z}_M)^{-1} - \boldsymbol{A}\big]^{-1} \boldsymbol{b} + d \, . \tag{12}$$

The diagram of the proposed architecture is shown in Fig. 2. We integrated $H(\boldsymbol{z}_M)$ into an optimization framework to estimate the set of FDN parameters based on a spectral and a temporal loss by gradient descent. The learnable parameters are the feedback matrix $\boldsymbol{A}$ and the input and output gain vectors $\boldsymbol{b}$ and $\boldsymbol{c}$, respectively. The delay lengths $\boldsymbol{m}$ are set at initialization, and kept constant during training. The direct gain d is set to zero. The FDN is set to have a homogeneous decay by forcing $\boldsymbol{A}$ to satisfy (6) for a given γ.

At each training step the estimated channel-wise transfer function $\hat{\boldsymbol{H}}(\boldsymbol{z}_M)$ is computed at M frequency bins. The input to the network is $\boldsymbol{z}_M$, where the value of M is sampled from the uniform distribution around values that ensure oversampling. This allows training the model at different sample rates, which proved to help avoiding narrow local minima and to improve convergence. To allow batch processing, the length of $\hat{\boldsymbol{H}}(\boldsymbol{z}_M)$ has to be constant for all values of M. This is achieved by zero-padding $\hat{\boldsymbol{H}}(\boldsymbol{z}_M)$ to length K. The network's output is evaluated in both the spectral and temporal domains. The IR of the system is computed from the K-point inverse discrete Fourier transform, $\hat{h} = \text{IDFT}(\hat{H}(\boldsymbol{z}_M))$, where $\hat{H}$ is the system transfer function computed from the sum of the N channels. The process of zero-padding in the frequency domain results in zero-phase rate conversion [24], and allows evaluating the IR at different timestamps.

3.2. Feedback Matrix Parametrization

The unilossless matrix $\boldsymbol{U}$ is computed from the weights $\boldsymbol{W}$ of a parameterized linear layer. Matrix $\boldsymbol{U}$ is limited to the class of

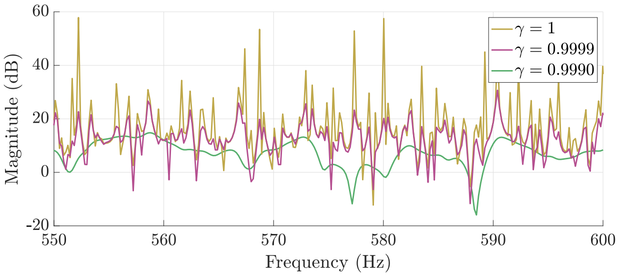

Figure 3: *Magnitude response of an FDN with random orthogonal feedback matrix and unitary input/output gains at different values of gain-per-sample value γ. For high values of γ, the resonances are better separated due to a smaller half-width.*

orthogonal matrices, satisfying the unitary condition for unilosslessness. To ensure orthogonality, at each optimization step $\boldsymbol{W}$ is mapped to a skew-symmetric matrix, and the matrix exponential is computed,

$$\boldsymbol{U} = e^{\boldsymbol{W}_{\mathrm{Tr}} - \boldsymbol{W}_{\mathrm{Tr}}^{\top}}\,, \tag{13}$$

where $\boldsymbol{W}_{\mathrm{Tr}}$ is the upper triangular part of $\boldsymbol{W}$ and the operator $e^{(\cdot)}$ denotes the matrix exponential. The mapping in (13) implicitly ensures orthogonality of $\boldsymbol{U}$ and can be used in regular gradient descent optimizers without creating spurious minima [25].

3.3. Gain-per-sample

When $\boldsymbol{A}$ is lossless, i.e., $\gamma = 1$, the modulus of all system eigenvalues is equal to one: $|\lambda_i| = 1$. However, under this condition, evaluating $H(\boldsymbol{z}_M)$ on the unitary circle becomes unfeasible, as the discrete generalized characteristic polynomial $p(\boldsymbol{z}_M) = \det(\boldsymbol{D}_{\boldsymbol{m}}(\boldsymbol{z}_M)^{-1} - \boldsymbol{A})$ becomes singular and non-invertible. To avoid instabilities, we use a homogeneous FDN where $\boldsymbol{A}$ is parameterized according to (6), and γ is set at initialization to a value lower than one and kept constant during optimization.

The value of γ used during optimization is chosen by examining the connection between the mean damping factor $\overline{\delta}$, used in room acoustics, and the mean spacing of resonance frequencies $\overline{\Delta f}$. To guarantee that the modes are well separated, the mean spacing of resonance frequencies should be larger than the average resonance half-width [26]

$$\overline{\Delta f} \gg \frac{\overline{\delta}}{\pi}\,. \tag{14}$$

In room acoustics, the limiting frequency below which the modes are well-separated is called Schroeder frequency, indicated here as $f_{\text{Schroeder}}$ [27]. This frequency marks the threshold above which an average of at least three modes falls within one resonance half-width. Using the fact that in FDNs the modal frequencies are nearly euqally distributed [9], we can derive the limiting average resonance half-width

$$\overline{\Delta f}_{|f=f_{\text{Schroeder}}} = 3\frac{f_{\text{s}}}{\mathcal{M}}\,. \tag{15}$$

We can use the above conditions to determine the minimum value for T_{60} to be used during training

$$T_{60} \gg \frac{\mathcal{M}\ln(10)}{\pi f_{\text{s}}}\,. \tag{16}$$

Increasing the value of T_{60} leads to modes with lower half-widths and greater separation between them. For a target T_{60}, the value of γ can be derived from (7). However, as γ approaches 1, the resonance peaks in the magnitude response become narrow, making obtaining a flat magnitude response by combining the resonances impossible. Fig. 3 shows the effect of increasing γ on the resonance width in a short section of the magnitude response. The sharp peaks visible when $\gamma = 1$ are significantly smoothed when $\gamma = 0.9990$.

Experiments showed good convergence of the loss used in the optimization when $T_{60} \leqslant 10\,\text{s}$. During inference γ is a free parameter, allowing to generate reverberation with any desired T_{60} value.

3.4. Parameters Initialization

We initialize the values of $\boldsymbol{W}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ by drawing from the normal distribution $\mathcal{N}(0, N^{-1})$.

The design of the delays is a rather non-trivial task that requires further constraints. To maximize the echo density, the delay lengths should be co-prime [28]. However, concentration of delays around a certain value may lead to perceivable strong fluctuation of energy over time. Moreover, low-order dependencies, which are integer linear combinations of delays that coincide with other integer linear combinations of delays with small coefficients, can also contribute negatively to the smoothness of the response [22]. To avoid degenerative patterns and ensure a smooth-sounding reverb, we choose delays that are logarithmically distributed coprime numbers leading to $\mathcal{M} \geqslant 6000$.

3.5. Loss Function

The network is trained on two losses, $\mathcal{L}_{\text{spectral}}$ and $\mathcal{L}_{\text{temporal}}$, respectively, in the frequency and time domains. The spectral loss aims to minimize the frequency-domain mean-squared error between the absolute value of the predicted magnitude response for each channel and the target flat magnitude response. The temporal loss penalizes sparseness in the time domain. The total loss function is

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{spectral}}(\hat{\boldsymbol{H}}(\boldsymbol{z}_M)) + \alpha\mathcal{L}_{\text{temporal}}(\hat{h}) \\ &= \frac{1}{K}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\left|\hat{H}_i(\boldsymbol{z}_M[k])\right| - 1\right)^p + \alpha\frac{\left\|\hat{h}\right\|_2}{\left\|\hat{h}\right\|_1}\,,\end{aligned} \tag{17}$$

where $\hat{H}_i(\boldsymbol{z}_M)$ is the output of the network's i^{th} channel computed from the output of the i^{th} delay line and scaled by c_i. The operators $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the ℓ_1 and ℓ_2 norm, respectively. The value of the scaling factor α depends on the FDN size and is chosen during initialization to ensure that the temporal and spectral losses have similar magnitudes.

Audibility of a resonant frequency depends on its loudness and on the presence of neighbouring masker tones [29]. To account for tone masking effects, we adjust the exponent p in $\mathcal{L}_{\text{spectral}}(\hat{\boldsymbol{H}})$ based on the sign of the magnitude difference. Specifically:

$$p = \begin{cases} 2 & \text{for } \left|\hat{H}_i(\boldsymbol{z}_M)\right| - 1 \leqslant 0\,, \\ 4 & \text{for } \left|\hat{H}_i(\boldsymbol{z}_M)\right| - 1 > 0\,. \end{cases} \tag{18}$$

This adjustment ensures that higher loss values are assigned when the predicted magnitude response exceeds one. For negative differences, $\mathcal{L}_{\text{spectral}}(\hat{\boldsymbol{H}})$ corresponds to the mean squared error.

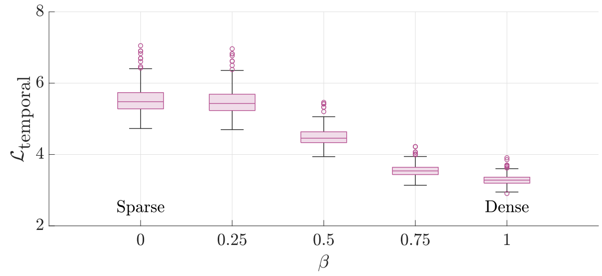

Figure 4: *Progression of temporal loss for different values of the interpolation parameter β. Density of the feedback matrix increases from left to right.*

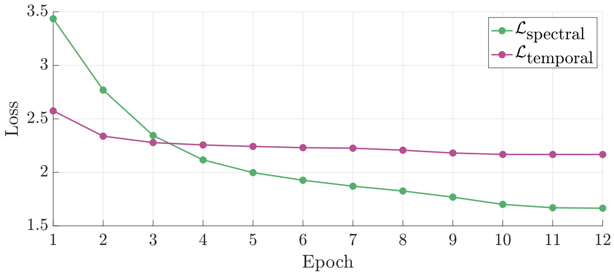

Figure 5: *Progression of spectral and temporal components of the loss function during optimization.*

The temporal loss $\mathcal{L}_{\text{temporal}}(\hat{h})$ is computed as the ratio of the ℓ_2 norm to the ℓ_1 norm of the estimated IR $\hat{h}$. We found that the absence of this term may lead to sparsity in the learnable parameters and cause the matrix $\boldsymbol{U}$ to converge towards either a diagonal matrix or its permutation. In this configuration, the magnitude response is periodic, with the spacing between peaks and troughs determined by the delay lengths, and the height of the peaks and the depth of the troughs depending on the gains. In time domain, this yields a sparse sequence of impulses whose sound is far from the intended Gaussian noise-like reverb.

To visualize the impact of the matrix on $\mathcal{L}_{\text{temporal}}(\hat{h})$, Fig. 4 summarizes the distribution of the loss values computed from an FDN with five different feedback matrices. The feedback matrix is interpolated between the values at initialization $\boldsymbol{U}$ and the identity matrix $\boldsymbol{I}$:

$$\boldsymbol{A}_\beta = e^{(1-\beta)\,\log(\boldsymbol{I})+\beta\,\log(\boldsymbol{U})}\,, \tag{19}$$

where β is the interpolation parameter $0 \leqslant \beta \leqslant 1$. Operator $\log(\cdot)$ represents the matrix logarithm. For $\beta = 1$, the feedback matrix corresponds to the initial configuration $\boldsymbol{A}_1 = \boldsymbol{U}$, whereas for $\beta = 0$ matrix $\boldsymbol{A}_0$ coincides with the identity matrix $\boldsymbol{I}$. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers stretch to include the most extreme data points that are not classified as outliers, and any outliers are plotted separately. The parameters of the FDN are initialized as described in Sec. 3.1, and the temporal loss is evaluated at 256 different values of M. The numbers in Fig. 4 show that the temporal loss $\mathcal{L}_{\text{temporal}}(\hat{h})$ grows for sparser feedback matrices, thus actively preventing convergence towards sparse matrices.

The evolution of losses at each epoch is shown in Fig. 5. Although $\mathcal{L}_{\text{spectral}}$ decreases at all displayed epochs, a near-steady value is attained by $\mathcal{L}_{\text{temporal}}$ after a few iterations. This controls the FDN and prevents convergence towards a set of comb filters.

4. OBJECTIVE EVALUATION

The following section presents the FDN configuration and the objective evaluation of the proposed method. The objective assessment is based on the modal excitation distribution.

4.1. Analyzed FDN Configurations

We evaluate a total of six FDN configurations, two sets of delay lengths for each of the three FDN sizes of $N = 4, 6, 8$. The values of the delay-line lengths are presented in Table 1. In the first delay set, the delay lengths were prime numbers distributed logarithmically. In the second delay set, only the second half of the delay lengths were logarithmically distributed, and the first half consisted of prime numbers with similar values. In all configurations, the total number of modes is $6000 < \mathcal{M} < 9000$.

Table 1: *Values of the delay-line lengths for each size N of the analyzed FDNs. In the delay set #1, all the delay lengths are logarithmically distributed prime numbers. For the delay set #2, half of the delay lengths are prime numbers with similar low values, and half are logarithmically distributed.*

N	**Delay Set #1**
4	[1499, 1889, 2381, 2999]
6	[997, 1153, 1327, 1559, 1801, 2099]
8	[809, 877, 937, 1049, 1151, 1249, 1373, 1499]
	Delay Set #2
4	[797, 839, 2381, 2999]
6	[887, 911, 941, 1699, 1951, 2053]
8	[241, 263, 281, 293, 1193, 1319, 1453, 1597]

During the training process, we used a sampling rate of $f_s = 48\,\text{kHz}$, and an inverse discrete Fourier transform of length $K = 480000$. The dataset consists of integer values M randomly selected from a uniform distribution ranging between $M_{min} = 0.8K$ and $M_{max} = K$. To train our model, we randomly selected 80% of the data from the dataset, and the remaining 20% was used for validation. The dataset size is 256 values of M. We set the batch size to 4, and employed the Adam optimizer [30] with a learning rate of $\eta = 10^{-3}$. Training was stopped after 15 epochs, as experiments showed no further improvement with extended training.

The choice of the gain-per-sample value γ is crucial when optimizing the feedback matrix. To satisfy (16) during training, we set $\gamma = 0.9999$, which implies $T_{60} = 1.439\,\text{s}$.

Configuration details and audio examples are available online [1]. The PyTorch implementation of the proposed method can be found in the dedicated repository [2]. A set of optimized FDN parameter values is readily available in the FDN Toolbox [5].

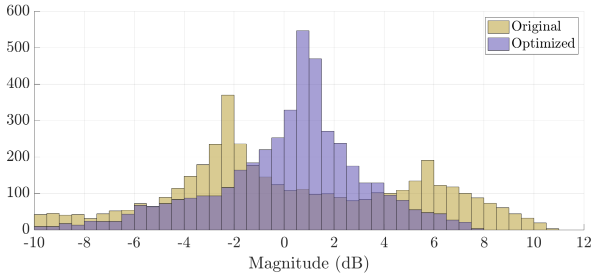

Figure 6: *Distribution of the modal excitation of an FDN with size $N = 4$ at the beginning (Original) and at the end of optimization (Optimized), which has led to a decrease of the loudest modal excitation by about 3 dB.*

4.2. Modal Excitation Distribution

We compare the FDN parameters after optimization with the corresponding FDN configurations at initialization. All the compared FDNs are homogeneous and have equal delays and gain-per-sample. We compute the modal decomposition (5) to analyze the modal excitation distribution of $|\rho_i|$.

The histograms in Fig. 6 show the distribution of the modal excitation at the beginning and at the end of the optimization processes for an FDN of size $N = 4$. The modal excitation values have been centered around 0 dB. At initialization, the distribution appears bimodal with the highest concentration of values around 6 dB and -2.5 dB. After optimization, the peak of the distribution is centered around 1 dB. The rightmost part of the distribution, which represents the modes with the highest excitation values, is important for coloration. In Fig. 6, the optimization attenuates the loudest modes by around 3 dB. The change toward narrower excitation distribution indicates an improvement in the coloration, which we further evaluate with a subjective test.

5. PERCEPTUAL EVALUATION

In the following, we describe a listening test conducted to evaluate the perceived coloration in the IRs of the differentiable FDN optimized with the proposed method.

5.1. Listening Test Procedure

The test followed the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) recommendation [31], and it was carried out using the web audio API-based experiment software webMUSHRA developed by International Audio Laboratories Erlangen [32].

On each page, the listening test compared four sets of FDN parameters against a reference. The test items included six configurations, i.e., three FDN sizes $N = 4, 6, 8$ with two sets of delays. The sounds were compared for two different RTs. In total, there were 12 listening test pages with five stimuli each.

At the beginning of the test, a training page was presented to familiarize the subjects with the sound samples and to adjust the overall loudness. The loudness was kept unchanged during the rest of the test. The reference was a white Gaussian-noise sequence due to its ideal reverberation tail [12], and since it has a flat magnitude response by definition. During the test, the subjects were asked to rate the similarity between each of the presented items and the reference sound on a scale from 0 to 100. On each page, six sounds were assessed, including an anchor and the hidden reference. The hidden reference was an instance of white Gaussian noise different from the reference, to encourage the subject to compare samples based on their coloration rather than any possible subtle temporal features.

The test evaluated the coloration of the DiffFDN IRs for the configurations presented in Sec. 4.1. Each configuration was tested on a separate page where the number and lengths of the delays were constant, and only the feedback matrix, input and output gains were altered. In particular, the FDN implementation of the Schroeder-Moorer reverberator (SM) with N delay lines was used as the anchor, whereas for the remaining conditions, the random orthogonal feedback matrix (RO), the proposed optimized FDN (DiffFDN) and the Householder (HH) feedback matrix were used. The RO condition were the initial values of optimization of the DiffFDN. Unitary input and output gains were used for the HH condition. The direct gain d was set to zero in all cases. Each individual IR was normalized to ensure a constant root-mean-square value across conditions.

The experiment was conducted in a sound-insulated booth at the Aalto Acoustics Lab, with participants wearing Sennheiser HD650 headphones. The final items were presented to 12 listeners. One participant was excluded from the analysis as they failed to correctly identify the anchor more than four times in their responses. The average age of the participants whose results were analyzed was 28.6 years with standard deviation of 4.1, and none of them reported any hearing impairments. All the participants were either students or employees of the Aalto University Acoustics Lab, and had previous experience with the MUSHRA test.

The IRs presented in the first part of the test (expDE) had an exponential decaying envelope corresponding to $T_{60} = 2.5$ s and $\gamma = 0.99994$. The subjects were asked to compare the coloration of the FDN responses against that of decaying white Gaussian noise. To ease the grading process, the slider was labeled with 0 - *certainly colored*, 25 - *rather colored*, 50 - *fairly colored / colorless*, 75 - *rather colorless*, and 100 - *certainly colorless*.

The second part of the test (LL) focused on the coloration of the late reverberation part. It compared the non-decaying IRs with $T_{60} = \infty$ and $\gamma = 1$. In order to exclude the echo build up of the early reflection from the comparison, the test items started after the mixing time, i.e., at 6 s. Each audio example was 10 s long. The slider labels were the same as in the first part of the test.

5.2. Listening Test Results

The results of the listening test are shown in the box charts in Fig. 7 and Fig. 8 for the expDE and LL cases, respectively. The meaning of marks and whiskers on the chart is the same as in Fig. 4 (*cf.* Sec. 3.5). The shaded regions around the medians help comparing the sample medians across different box charts. Shaded regions that do not overlap indicate that the compared box charts have different medians at the 5% significance level based on a normal-distribution assumption.

Conducting the Shapiro-Wilk test [33] showed that even when excluding the reference and anchor conditions, the data did not follow a normal distribution. In addition, the Wilcoxon signed-

[1] `http://research.spa.aalto.fi/publications/papers/dafx23-colorless-fdn/`

[2] `https://github.com/gdalsanto/diff-fdn-colorless/`

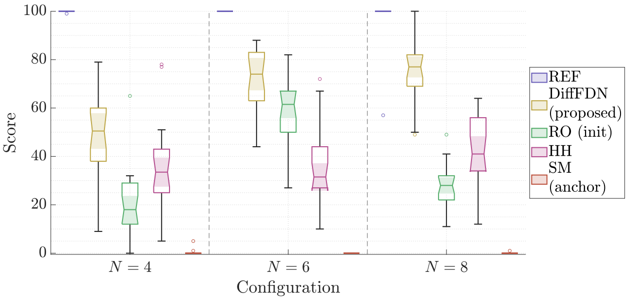

Figure 7: *Results of the listening test on exponential decaying IRs (expDE), showing that the proposed DiffFDN has the highest median score of colorlessness in all cases.*

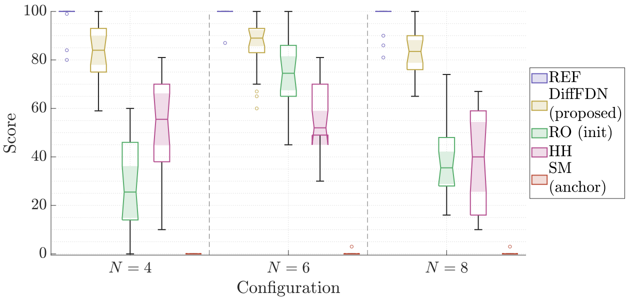

Figure 8: *Results of the listening test on the late reverberation, employing lossless FDNs (LL), showing that the proposed DiffFDN has the highest median score of colorlessness in all cases.*

rank test [34] was used to compare the distribution of the scores given to each pair of conditions within each page. To account for multiple comparisons (10 hypotheses per page), we applied the Bonferroni method to adjust the alpha level.

The p-values for all combinations of paired conditions suggest that all pairs of results are significantly different, with exception of the lossless case with $N = 8$ for RO and HH FDNs ($p = 0.68$). These results are indicated by the overlapping shaded regions of the corresponding box charts in Fig. 8. This may be due to the lack of early reflections in the lossless case, which makes differentiating between conditions difficult. Additionally, the configuration with a higher number of delays ($N = 8$) produces a denser output, which might result in a more challenging test.

The results presented in Figs. 7 and 8 show that the hidden reference and anchor signals were easily detected by most subjects, with few outliers. The median ratings for the proposed method were consistently higher than those for the remaining conditions, indicating that the optimization method was successful in improving colorlessness from the initial values.

In the first part of the test (expDE), increasing FDN sizes resulted in higher ratings for DiffFDN, with median values of 50.5, 74, and 77. The results for lossless FDNs (LL) reported a similar trend, with overall higher ratings primarily due to the elimination of the temporal build up. The proposed method was deemed rather colorless, with median ratings of 84, 89, and 83.5, respectively, for increasing FDN size. The RO matrix was rated more colorless than the HH matrix for the configuration with $N = 6$ delay lines, while it was rated more colored in the remaining configurations. This inconsistency may be attributed to the random sampling of the orthogonal matrix, which is performed without any preselection based on perceptual factors.

To emphasize the ratings relative to the proposed method, the box charts in Figs. 9 and 10 were calculated based on the difference between the DiffFDN and the remaining conditions. The ratings assigned to the reference are not displayed. The results show that in the majority of test questions, proposed method was rated higher than the remaining stimuli. Significant improvements are observed in the lossless case for $N = 4$. Specifically, the median value of the RO configuration was 59 lower than its optimized version (DiffFDN). Moreover, in both conditions, the median of the score differences and their 75th quartiles are consistently negative. The confidence intervals in Fig. 9 are noticeably narrower compared to those in Fig. 10, suggesting that the test on lossless FDNs was more challenging.

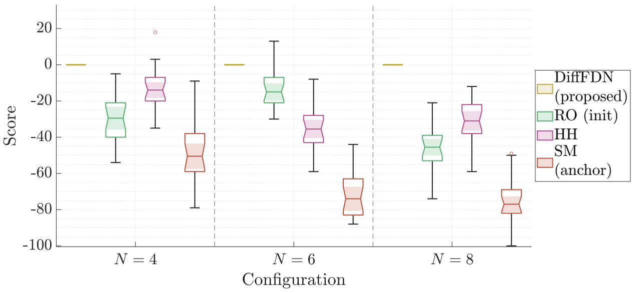

Figure 9: *Relative difference of the results of Fig. 7 from the results of the proposed DiffFDN method (expDE case).*

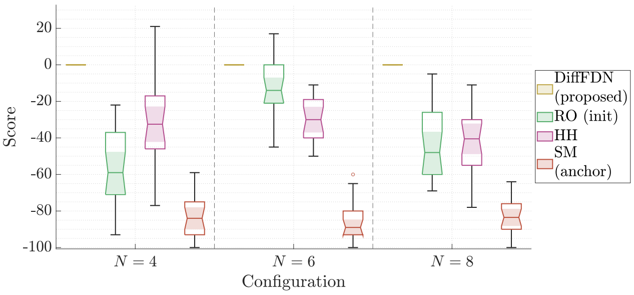

Figure 10: *Relative difference of the results of Fig. 8 from the results of the proposed DiffFDN method (LL case).*

6. CONCLUSIONS

This work presents a method for designing colorless artificial reverberation using a differentiable feedback delay network (DiffFDN). The technique optimizes elements of the DiffFDN architecture—the feedback matrix as well as the input and output gains—to achieve a flat magnitude response. In addition, the temporal properties of the synthesized reverb are taken into account to avoid overly sparse results.

In the objective evaluation, we showed that the proposed method reduces the width of the modal excitation distribution, decreasing the number of loudest modes. This indicates that the DiffFDN achieves more colorless sound and flatter magnitude response of the produced reverb.

The results of the listening test show that, compared to other popular FDN designs, DiffFDN showed a significant improvement in reverberation quality. Reverberation obtained with DiffFDN was consistently graded as the most colorless among several conditions, placing it perceptually closer to white Gaussian noise than the other evaluated methods. This further confirmed the results of the objective assessment and proved that the proposed method successfully synthesizes colorless sound.

7. ACKNOWLEDGMENTS

This research belongs to the activities of the Nordic Sound and Music Computing Network (NordForsk project number 86892). The work of the first author was funded by the Aalto University School of Electrical Engineering. Special thanks go to Luis Costa for proofreading this paper.

8. REFERENCES

[1] M. R. Schroeder and B. F. Logan, “"Colorless" artificial reverberation,” *IRE Trans. Audio*, vol. AU-9, no. 6, pp. 209–214, July 1961.

[2] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, “Fifty years of artificial reverberation,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 5, pp. 1421–1448, July 2012.

[3] M. A. Gerzon, “Synthetic stereo reverberation: Part one,” *Studio Sound*, vol. 13, pp. 632–635, Dec. 1971.

[4] J.-M. Jot and A. Chaigne, “Digital delay networks for designing artificial reverberators,” in *Proc. 90th AES Conv.*, Feb. 1991, pp. 1–12.

[5] S. J. Schlecht, “FDNTB: The feedback delay network toolbox,” in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Sept. 2020, pp. 211–218.

[6] S. J. Schlecht and E. A. P. Habets, “On lossless feedback delay networks,” *IEEE Trans. Signal Process.*, vol. 65, no. 6, pp. 1554–1564, Mar. 2017.

[7] S. J. Schlecht and E. A. P. Habets, “Accurate reverberation time control in feedback delay networks,” in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Aug. 2017, pp. 337–344.

[8] K. Prawda, S. J. Schlecht, and V. Välimäki, “Improved reverberation time control for feedback delay networks,” in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Sept. 2019, pp. 299–306.

[9] S. J. Schlecht and E. A. P. Habets, “Modal decomposition of feedback delay networks,” *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5340–5351, Aug. 2019.

[10] J. Heldmann and S. J. Schlecht, “The role of modal excitation in colorless reverberation,” in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Sept. 2021, pp. 206–213.

[11] S. J. Schlecht, “Allpass feedback delay networks,” *IEEE Trans. Signal Process.*, vol. 69, pp. 1028–1038, Jan. 2021.

[12] J. A. Moorer, “About this reverberation business,” *Computer Music J.*, vol. 3, no. 2, pp. 13–28, June 1979.

[13] M. Karjalainen and H. Järveläinen, “More about this reverberation science: Perceptually good late reverberation,” in *Proc. AES Conv.*, Nov. 2001, pp. 1–8.

[14] N. Agus, H. Anderson, J.-M. Chen, S. Lui, and D. Herremans, “Perceptual evaluation of measures of spectral variance,” *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3300–3311, June 2018.

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

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

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

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

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

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

[21] M. A. Gerzon, “Unitary (energy-preserving) multichannel networks with feedback,” *Electronics Letters*, vol. 12, no. 11, pp. 278–279, May 1976.

[22] S. J. Schlecht and E. A. P. Habets, “Feedback delay networks: Echo density and mixing Time,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 2, pp. 374–383, Dec. 2017.

[23] J. Stautner and M. Puckette, “Designing multi-channel reverberators,” *Comput. Music J.*, vol. 6, no. 1, pp. 52–65, Spring 1982.

[24] V. Välimäki and S. Bilbao, “Giant FFTs for sample-rate conversion,” *J. Audio Eng. Soc.*, vol. 71, no. 3, pp. 88–99, Mar. 2023.

[25] M. Lezcano-Casado and D. Martınez-Rubio, “Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group,” in *Proc. Int. Conf. Machine Learning*, May 2019, pp. 3794–3803.

[26] H. Kuttruff, *Room Acoustics, Fifth Edition*, CRC Press, June 2009.

[27] H. Kuttruff, “Eigenschaften und Auswertung von Nachhallkurven,” *Acta Acust. United Acust.*, vol. 8, no. 4, pp. 273–280, Jan. 1958.

[28] S. J. Schlecht, *Feedback Delay Networks in Artificial Reverberation and Reverberation Enhancement*, Ph.D. thesis, Oct. 2017.

[29] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, vol. 22, Springer Science & Business Media, Mar. 2006.

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

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

[32] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, “WebMUSHRA—A comprehensive framework for web-based listening tests,” *J. Open Research Software*, vol. 6, no. 1, 2018.

[33] 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.

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