Automatic Parameters Tuning of Late Reverberation Algorithms for Audio Augmented Reality

Riccardo Bona
Lab. of Music Informatics (LIM),
Department of Computer Science,
University of Milan,
Milan, Italy
riccardo.bona@unimi.it

Davide Fantini
Lab. of Music Informatics (LIM),
Department of Computer Science,
University of Milan,
Milan, Italy
davide.fantini@unimi.it

Giorgio Presti
Lab. of Music Informatics (LIM),
Department of Computer Science,
University of Milan,
Milan, Italy
giorgio.presti@unimi.it

Marco Tiraboschi
Lab. of Music Informatics (LIM),
Department of Computer Science,
University of Milan,
Milan, Italy
marco.tiraboschi@unimi.it

Isaac Engel
Audio Experience Design (AXD),
Dyson School of Design Engineering,
Imperial College London,
London, United Kingdom
isaac.engel@imperial.ac.uk

Federico Avanzini
Lab. of Music Informatics (LIM),
Department of Computer Science,
University of Milan,
Milan, Italy
federico.avanzini@unimi.it

ABSTRACT

The matching of reverberation features between real sound sources and virtual ones is a key task in Audio Augmented Reality. An adequate matching provides a proper auditory immersion to the user. In this paper, we propose a method for reverb matching. The method automatically optimizes the parameters of an artificial reverberator to match a target Room Impulse Response (RIR). We used a Bayesian optimization procedure using a Gaussian Process as a prior distribution. This procedure iteratively tunes the artificial reverberator parameters to match its output, i.e. the approximated RIR, with the target one. The matching between the approximation and the target is implemented with a perceptually motivated loss function. Before this parameters optimization, the target early reflections are approximated through an autoregressive model. The method has been implemented with two artificial reverberation algorithms: a Feedback Delay Network (FDN) and an implementation of the Schroeder-Moorer reverberator (Freeverb). We evaluated the method with a listening test to assess the similarity with target reverberation. Our method yields overall statistically significant higher scores with respect to other anchor conditions. Further, subjective differences between FDN and Freeverb are not significant.

CCS CONCEPTS

• **Human-centered computing → Mixed / augmented reality**; *Sound-based input / output*; **Auditory feedback**; • **Computing methodologies → Gaussian processes**; • **Applied computing → Sound and music computing**.

KEYWORDS

Spatial audio, Audio Augmented Reality, Reverberation, Artificial reverberation, Reverb matching, Feedback Delay Networks

ACM Reference Format:
Riccardo Bona, Davide Fantini, Giorgio Presti, Marco Tiraboschi, Isaac Engel, and Federico Avanzini. 2022. Automatic Parameters Tuning of Late Reverberation Algorithms for Audio Augmented Reality. In *AudioMostly 2022 (AM '22), September 6–9, 2022, St. Pölten, Austria.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3561212.3561236

1 INTRODUCTION

1.1 Problem overview

In Audio Augmented Reality (AAR), virtual sound sources are blended in an existing acoustic environment with real sound sources and the user is unable to distinguish them. To provide a proper immersion in the soundscape, the virtual sources should share the same acoustic features of the real ones. The virtual and real blending is achieved by simulating the reverberation of the real environment, a task known as *reverb matching*.

A reverberant environment can be modeled as Linear and Time-Invariant System (LTI), thus it is described by its *Room Impulse Response* (RIR). So, the convolution between the RIR and an input audio signal applies the environment reverberation to the input. Convolution techniques are popular in such fields as music production or film audio post-production, as they allow for the simulation of a specific listening space, but are computationally expensive for long RIRs, and lack flexibility in control [30]. Physical room modeling techniques (e.g. image-source and ray tracing) provide a highly controllable solution and accurate results [23]. However, physical approaches rely on the knowledge of the environment properties (e.g. shape, size, objects presence, materials). Thus, physical approaches are impractical in real-world application and in AAR scenarios where real-time is needed.

In real-time contexts, *Digital Artificial Reverberators* (DARs hereafter) based on delay networks and low-order filters represent an efficient and controllable solution for reverb matching [30]. The DAR's reverberation effect is controlled by a set of *Digital Artificial*

AM '22, September 6–9, 2022, St. Pölten, Austria
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9701-8/22/09…$15.00
https://doi.org/10.1145/3561212.3561236

Reverberator Parameters (DARPs). So, the DARPs can be tuned in order to match the reverberation of a target environment. The DARPs tuning can be manually performed by an expert audio engineer, but this is a time-consuming task. Therefore, several algorithmic solutions have been studied, typically relying on genetic algorithms or deep learning. In general, the reverb matching is performed by optimizing the DARPs to minimize a loss function between a target reverberation and the one produced by the DAR.

1.2 Method overview

We propose a reverb matching method that automatically tunes the parameters of a DAR to match a target RIR. The method's core is a Bayesian optimization procedure using a Gaussian process as a prior. We define a perceptual loss function between the target RIR and the DAR's output. The approximation procedure iteratively optimizes the DARPs to minimize the perceptual loss function.

The employed DARs are limited to late reverberation only. Thus, before the DARPs tuning, we applied a separate method to match the target RIR early reflections. First, the early reflections are separated from the late reverberation using a custom procedure based on the spectral difference between time frames extracted from the target RIR. Then an autoregressive model is fitted on the early reflections and it is applied on the input audio signal to be reverberated.

We evaluated our method through a MUSHRA test. In the test, we compared the matching procedure performed with two DARs: a Feedback Delay Network (FDN) and Freeverb. Further, we included an hidden reference and two anchors in the comparison. The test was composed of 18 trials (six target RIRs and three audio stimuli).

We provide supplementary materials in a dedicated web page[1] about the MUSHRA test and the objective evaluation.

2 RELATED WORK

Several reverb matching methods have been proposed in the literature. The automatic tuning of DARPs is a common approach. We provide here an overview of the main works in this field.

Heise et al. [16] proposed an early work of automatic DARPs adjustment. The authors tuned the parameters of a reverb plug-in to fit a target RIR. They used a genetic algorithm comparing four optimization strategies. The genetic algorithm minimized as loss function the Euclidean distance between Mel-Frequency Cepstral Coefficients (MFCCs) vectors. In a MUSHRA-type test, their method yields comparable results with other reverberation conditions.

Some works exploited the internal structure of a specific DAR in order to perform the DARPs tuning. The most used DARs are from the class of FDNs. The FDN parameters are tuned to match the late reverberation part of a target RIR (a genetic algorithm is commonly used). Then, the direct path and early reflections are rendered separately. Coggin and Pirkle [6] tuned the DARPs of a FDN using a fitness function based on power envelope [5]. The early reflections were rendered through convolution with the target RIR. The method was evaluated through a listening test. Convincing results were found for small room reverb, whereas performances decreased with increasing room size. Another example with FDN is proposed in [26]. The authors generated a dataset of shoebox RIRs, and tuned the FDN parameters to match the target RIR through a genetic algorithm employing a loss function based on reverberation time and clarity index matching. Finally, a SVM regression model was trained to predict the FDN parameters given the room parameters. However, no listening test is reported.

Recent solutions make use of deep learning. In [29], two neural networks are proposed. The first one maps noisy reverberant recordings into a low-dimensional embedding vector characterizing the acoustic environment. Then, a waveform-to-waveform network, conditioned on the embeddings, transforms the input audio to match the acoustic features of the given embeddings. Listening tests showed improvements over baseline methods, especially in noisy cases. Another example based on deep learning is proposed in [21] where a neural network involving Bidirectional Gated Recurrent Units (BGRUs) was designed to perform two tasks, regression of DARPs and classification of reverberation presets, using as input any reverberated audio signal. A MUSHRA test compared the models (DARPs regression and preset classification), DARPs manually tuned by an audio engineer, and random DARPs. DARPs regression yielded results comparable with the audio engineer ones, while the classification performed slightly worse.

The *Differentiable Digital Signal Processing* (DDSP) [8] introduction has been a significant landmark in this field. DDSP allows the integration of classic DSP elements in deep learning frameworks. That means audio effects parameters are directly tuned via backpropagation. DDSP approaches are sometimes referred to as differentiable artificial reverberation when applied to DARs. In [18], a DARPs estimation network is connected to a differentiable artificial reverberator to allow the end-to-end training. So, the DARPs are tuned to match the input reverberation. The input can be either a RIR or a reverberated speech track. While any DAR could be used in their network, the authors selectively derived differentiable version of FDN and Filtered Velvet Noise (FVN). They showed that their model can capture the target reverberation accurately in terms of reverberation time, direct-to-reverberant ratio and clarity.

3 DATA

For the sake of clarity all the data employed in our work are described in this section. In particular, we present the data used in the reverb matching method (matching audio signal) and in the evaluation phase (target RIRs, test audio stimuli). The sampling frequency f_s is 44.1 kHz for all audio signals.

3.1 Target RIRs

We evaluated the method proposed in this work on six target RIRs to be matched. Target RIRs were chosen to be both recorded and artificial responses, in order to assess the generality of the proposed approach. Specifically, four of the chosen RIRs are real binaural recordings from four different locations (auditorium hall, outdoors courtyard, recording studio and small room). The RIRs have been recorded with a Neumann KU-100 dummy head positioned in front of the emitter. Another RIR is generated with CATT-Acoustic software [4] simulating a living room. The remaining RIR has been generated feeding the *REVelation* VST3 plug-in by Steinberg[2] with an impulse. We set a plate reverb preset on this plug-in.

[1] https://zenodo.org/record/6854824

[2] https://steinberg.help/cubase_plugin_reference/v9/en/_shared/topics/plug_ref/revelation_r.html

3.2 Test audio stimuli

Three audio stimuli were selected to be used in a subsequent listening test. They were chosen to have different temporal and spectral characteristics, in order to test the DARs in a variety of conditions. All stimuli are monophonic dry recordings. Two were retrieved from a performance of "Take Five" by Paul Desmond: one is a drums track (kick, snare and ride) while the second one is a saxophone solo track, both 11 s long. The remaining stimulus is an English speech recording of a female speaker [14], and is 12 s long.

3.3 Matching phase audio signal

During the reverb matching phase, the DAR generates a reverberated signal that is compared with the target. To generate this signal, we fed the DAR with a sweep s instead of an impulse. We also convolved the target with s to allow the comparison with the reverberated sweep. In particular,s is an ascending logarithmic sine sweep ranging from 20 Hz to 20 kHz. The sweep s was 3 s long.

We selected a sweep in order to achieve better performances in the matching of the dry-wet ratio parameter. Since the matching loss function splits the signal in time frames, using an impulse, most of the dry-wet information is in the first frame. Given that the loss function is averaged over the frames, this information would exert a limited influence on the loss. Conversely, using a sweep, dry-wet information is present in all time frames. As a consequence, this information has more influence on the loss function.

4 AUTOMATIC REVERB MATCHING

4.1 Method workflow

An overview on the method workflow is shown in Fig. 1.

1) The target RIR early reflections are so modelled:
   a) The boundary point between early reflections and late reverberation is automatically estimated.
   b) An autoregressive model is fitted on the mid-side encoded early reflections giving the equalization coefficients $A^{m,s}$.
2) The target RIR r is convolved with the sweep s giving s_r.
3) The sweep s is equalized with the mid component equalization coefficients A^m computed at step 1 giving s_A.
4) In the DARPs optimization, the DAR reverberates s_A as equalized at step 3 to match the target RIR s_r as convolved at step 2. The optimized DARPs $\hat{P}$ are obtained.
5) The input dry audio mid component x^m is equalized with the coefficients A^m giving x_A.
6) The signal x_A is reverberated with the DARPs $\hat{P}$ giving $\tilde{x}_A$.
7) The reverberated audio side component $\tilde{x}^s_A$ is equalized with the side component coefficients A^s giving y.
8) The audio signal y is the final output simulating the reverberation effect of r on the input signal x.

4.2 Early reflections equalization

Since the employed DARs simulate only late reverberation, we decided to model early reflections separately. We propose an early reflection modelling approach based on equalization matching through an autoregressive model. This approach only considers spectral effects and not the temporal structure of early reflections.

In order to mitigate the difference given by the overall coloration of the target RIR early reflections, we decided to equalize the input signal x with an estimate of such coloration, represented by the mid component coefficients A^m. Furthermore, the side component of the reverberated signal $\tilde{x}_A$ was equalized with the side component coefficients A^s. This was done to mimic also the stereophonic image coloration provided by the early reflections.

Before this matching, we designed a custom method to estimate the boundary point between early reflections and late reverberation from the target RIR r. The description of this method follows.

The mid-side encoded target RIR r is split in multiple time frames with increasing size. The n-th frame ranges from $r[0]$ to $r[n\delta - 1]$, where $\delta = 512$ is the base frame size. A cosine fade-out is applied at the end of each frame. For each frame, we fit an autoregressive model with the Yule-Walker method. We set the model order p to $2\lfloor(2 + f_s/1000)\rfloor$. So, the fitting procedure for the frame n returns the corresponding $p + 1$ coefficients A_n. For each frame n, we also compute the Power Spectral Density PSD_n as follows:

$$PSD_n(f) = \left|\frac{1}{A_n(f)}\right|^2. \tag{1}$$

Then, we compute the Logarithmic Spectral Distance LSD between PSD_n and PSD_{n+1} for each pair of successive frames. Only the mid component PSD^m is involved in this operation because the amount of energy in the side component is negligible. LSD is defined as the distance in dB between two power spectra:

$$LSD^m_n = \sqrt{\frac{1}{F}\sum_{f=0}^{F}\left(10\log_{10}\frac{PSD^m_{n+1}(f)}{PSD^m_n(f)}\right)^2} \quad [\text{dB}], \tag{2}$$

where F is the length of the PSD frames.

Then, the boundary point between early reflections and late reverberation is estimated with a knee detection algorithm [22] performed on LSD values. This criterion is chosen because higher differences in PSD values will be generated by frames containing stand-alone early reflections, with respect to those containing diffuse reverberation too. So, the frame $\hat{n}$ is the one containing the early reflections as detected by the described method. The autoregressive coefficients $A_{\hat{n}}$ of the frame $\hat{n}$ are used to filter the dry audio signals in the following steps.

4.3 Late reverberation matching

The core of the proposed method is the automatic tuning of DARPs to match the late reverberation of the target RIR r. We perform this tuning iteratively: at each iteration the DARPs are adjusted to minimize the difference between the DAR output and target signal. In the following, we describe first the DARPs tuning procedure, then we present the employed kernel and loss function. The DARPs optimization has been implemented with the Python library *Scikit-Optimize* [15]. The employed DARs in VST3 format have been manipulated with the Python library *Pedalboard* [1].

4.3.1 DARPs tuning. Before the DARPs tuning, we pre-processed the target RIR r and the sweep s, the DAR input. We computed s_A, used as DAR input, by filtering the sweep s with the mid component coefficients $A^m_{\hat{n}}$. Then, in order to compare r with the DAR output $\tilde{s}_A$, we compute s_r as the convolution between r and the sweep

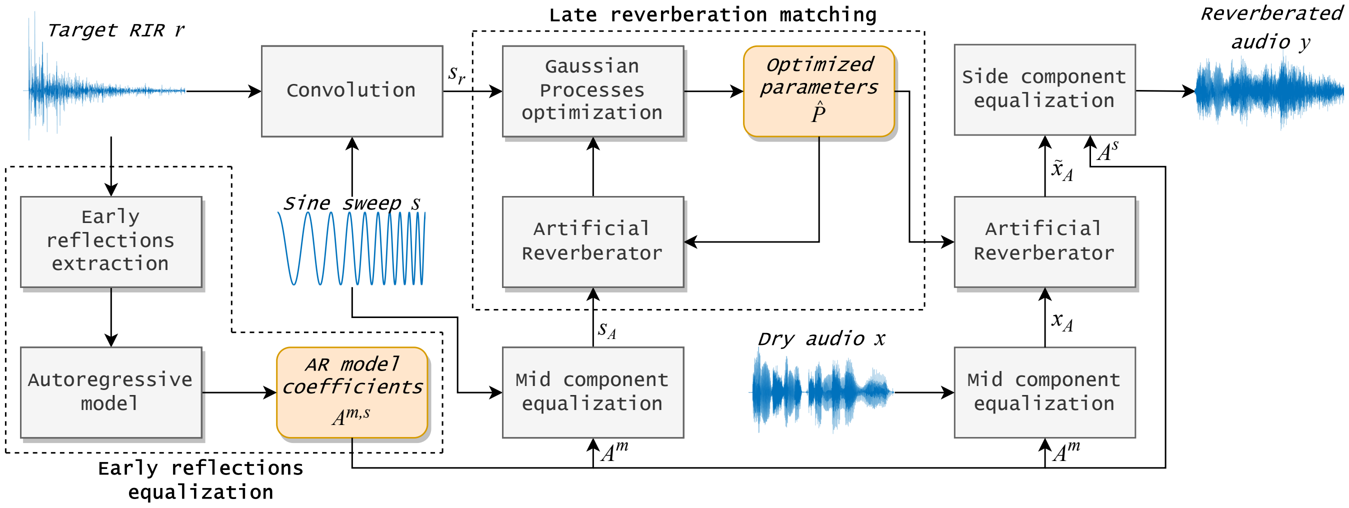
Target RIR r
Convolution
s_r
Late reverberation matching
Gaussian Processes optimization
Optimized parameters $\hat{P}$
Side component equalization
Reverberated audio y
$\tilde{x}_A$
A^s
Early reflections extraction
Sine sweep s
Artificial Reverberator
Artificial Reverberator
s_A
Dry audio x
x_A
Autoregressive model
AR model coefficients $A^{m,s}$
Mid component equalization
Mid component equalization
A^m
A^m
Early reflections equalization

Figure 1: Overall workflow of the proposed reverb matching method.

s. The DARPs tuning is based on a Bayesian optimization using a Gaussian process as a prior [28] to minimize the function $\mathcal{F}$:

$$\mathcal{F}(s_r, s_A) = \ell(s_r, D(s_A, P)), \tag{3}$$

where P are the DARPs and D is the artificial reverberator. The function $\mathcal{F}$ is assumed to follow a multivariate Gaussian distribution. At iteration i, the equalized sweep s_A is processed with the DAR given the current DARPs P_i. The DAR returns the reverberated sweep $\tilde{s}_A$. Then, the loss function $\ell(s_r, \tilde{s}_A)$ between the target s_r and the estimate $\tilde{s}_A$ is computed. Since the mid and side components have unbalanced energy content, we opted to compute ℓ with left-right encoding. Then, the left and right loss values are averaged.

An acquisition function $\mathcal{A}$ chooses the DARPs P_{i+1} for the next iteration within a range of values. A kernel function C, describing the covariance of $\mathcal{F}$, compares the current parameters P_i with the new candidate ones. In our work, $\mathcal{A}$ is a choice between three acquisition functions: Lower Confidence Bound (LCB), negative Expected Improvement (EI) and negative Probability of Improvement (PI). The choice is based on a gain assigned to each acquisition function and updated every iteration.

4.3.2 *Matérn kernel.*

To model the prior Gaussian distribution of $\mathcal{F}$, we used the Matérn kernel C [31, Ch. 4, Sec. 4.2] as covariance function between the parameters:

$$C\left(P_j, P_k\right) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{l}d(P_j, P_k)\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}}{l}d(P_j, P_k)\right), \tag{4}$$

where $d(P_j, P_k)$ is the Euclidean distance between the points P_j and P_k (the DARPs in our case), $K_\nu(\cdot)$ is a modified Bessel function and $\Gamma(\cdot)$ is the gamma function. Further, the kernel is controlled by the hyperparameters l and ν, where $l > 0$ is the *length-scale parameter* and ν controls the smoothness of the resulting function. We set l as a vector of ones (one element for each DARP) and $\nu = \frac{5}{2}$.

4.3.3 *Perceptual loss function.*

We defined the loss function ℓ as a modified version of the ones used in [8, 18]. The loss $\ell(h, \hat{h})$ is the mean absolute difference between the multi-resolution spectrograms of two signals h and $\hat{h}$. In particular, to define a perceptually motivated loss, we compute the mel-spectrogram in dB: the spectrogram frequencies are mapped to the Mel scale through a mel filter-bank, then, the values are converted in dB. Thus, the mel-spectrogram $\mathcal{M}_f$ in dB for a signal h is defined as:

$$\mathcal{M}_f(h) = 10\log_{10}\left[\mathrm{mel}(\mathrm{STFT}_f(h))\right] \quad [\mathrm{dB}], \tag{5}$$

where STFT_k is the Short-Time Fourier Transform with frame size k and mel is the filter-bank mapping to the mel scale. In the STFT computation, we employed an Hanning window of size equals to k and an overlap equals to 25% of k. All values of the spectrum are truncated to the lower threshold of −60 dB.

Finally, the loss function ℓ is computed as:

$$\ell(h, \hat{h}) = \sum_{f \in F} \frac{1}{T} \sum_{t=1}^{T} \left|\mathcal{M}_f^t(h) - \mathcal{M}_f^t(\hat{h})\right|, \tag{6}$$

where t is the time frame index, T is the number of time frames and $F = \{256, 512, 1024, 2048, 4096\}$ are the selected spectrum frame sizes. Before the loss function computation, a third-order high-pass filter at 20 Hz is applied to the target RIR. This prevents infrasonic components to influence the loss function.

4.3.4 *Implementation details.*

In the DARPs tuning procedure, we considered all the parameters with few exceptions. We tuned all the FDN parameters (delay length, reverberation time, fade-in time, high/low cutoff, high/low Q, high/low gain, dry-wet ratio) except for the feedback matrix size, fixed to 64. We tuned all the Freeverb parameters (room size, damping, wet and dry levels, stereo image width) except for the freeze mode parameter, fixed to 0 to avoid a continuous feedback loop state. For the detailed parameters behavior, refer to the related FDN[3] and Freeverb[4] repositories.

[3] https://git.iem.at/audioplugins/IEMPluginSuite/-/tree/master/FdnReverb
[4] https://github.com/juce-framework/JUCE/blob/master/modules/juce_audio_basics/utilities/juce_Reverb.h

We decided to limit the matching procedure iterations to 180 since this was found to be enough to reach the loss function's minimum. We evaluated the DARPs tuning on a computer with a Ryzen 5 @ 3.6 GHz CPU and a 24 GB RAM. The procedure took 13 and 7.3 minutes on average for FDN and Freeverb, respectively.

5 EVALUATION

The devised reverb matching method has been implemented and evaluated with two DARs: FDN and Freeverb. FDN [11] is a DAR based on an orthogonal matrix feedback reverberation unit. We employed the *FdnReverb* implementation from the IEM Plug-in suite.[5] Freeverb is made of four Schroeder allpass filters in series and eight Schroeder-Moorer filtered-feedback comb-filters in parallel [27].

We evaluated the method both with objective metrics and a listening tests. In this section, we describe the metrics and the test protocol. Then, the obtained results are reported and discussed.

5.1 Objective evaluation

We computed a set of reverberation features from the target RIRs r and the corresponding RIRs approximated with our method. The computed features are: reverberation time T_{20}, early decay time EDT, centre time T_S, strength index G, clarity index C_{80}, lateral energy fraction LF_{80} and spectral centroid SC [12, 13]. The features have been computed as averages according to ISO 3382 standard [9]. The strength index G is based on the first 5 ms of the RIR as reference for the loudness. We computed LF_{80} comparing the mid and side components of the RIRs. Then, we computed SC with an Hanning window 2048 samples long and with 25% of overlap.

Table 1 reports the features values for each RIR and for each reverberation condition (target, matched with FDN and matched with Freeverb). From the table, we notice that T_{20} and C_{80} are the best matched features values, i.e. with the lower Mean Absolute Percentage Error (MAPE) values. The good match of T_{20} suggests that the DARPs optimization provides a reasonable match of the reverberation tail length. Additional details (T_{20} per octave) are given in the supplementary materials. While the good match of C_{80} is probably due to the early reflections equalization procedure. The remaining parameters have higher MAPE values. In particular, LF_{80} has the lowest performances, probably because both DARs lack of parameters to consistently control the stereo image and the recorded RIRs are binaural.

Comparing the performances of the two DARs we notice mixed results. Freeverb has better performances for the majority of the features: EDT, T_S, C_{80} and SC. However, Freeverb has a significant higher MAPE for LF_{80} which is mostly influenced by its value for the Small Room RIR. So, for LF_{80} and the remaining features (T_{20} and G), FDN performs better than Freeverb.

Finally, in Fig. 2 the mel-spectrogram in dB of an example target RIR is compared with the corresponding versions matched with FDN and Freeverb. Note that the reverberation tail is very similar between the target and the matched versions. This is an example of the aforementioned good match of T_{20}. However, we notice also that there are some inconsistencies in the mel-spectrograms. For example, there are some notches in the high frequencies for FDN that do not exist in the target mel-spectrogram.

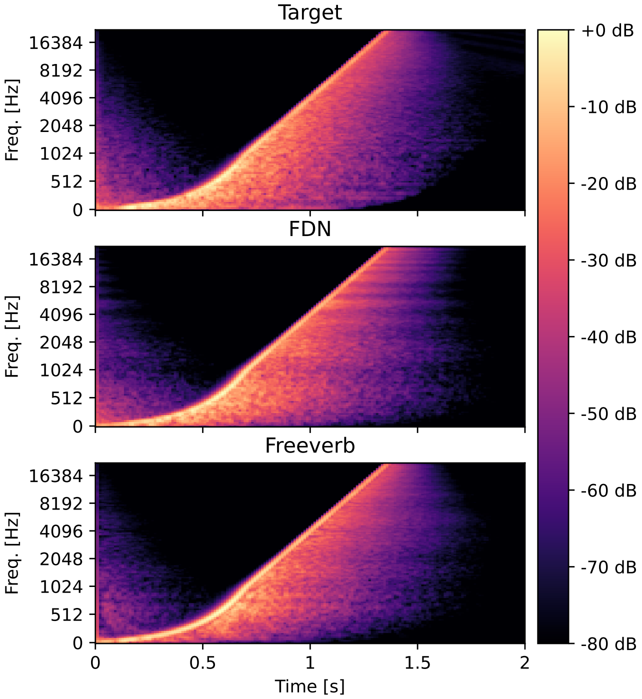

Figure 2: Comparison between an example target RIR (Auditorium Hall) and the versions matched with FDN and Freeverb. The mel-spectrograms in dB of the left channel is shown for each RIR convolved with the logarithmic sine sweep *s*.

5.2 MUSHRA test

We performed a MUSHRA test [3] to assess the subjective performances of our method. We employed *webMUSHRA* [24], a web-based implementation of the test. All the stimuli and the matched RIRs used in the test are provided in the supplementary materials.

5.2.1 *Test design.* In each trial we asked subjects to rate the similarity between the reverberation of a "Reference" audio track and five conditions. The "Reference" is an audio stimulus reverberated with one of the target RIRs. Two conditions are the reverberation matched with our method using FDN and Freeverb. Two conditions are the "Standard Anchor" (SA) and the "Mid-quality Anchor" (MA), i.e. the "Reference" processed with a low-pass filter at 3.5 kHz and at 7 kHz, respectively. The remaining condition is the same as the "Reference", i.e. the "Hidden Reference" (HR). Subjects rate the similarity between the "Reference" and each of the conditions with a scale from 0 to 100. According to MUSHRA guidelines, we informed the subjects about the "Hidden Reference" among the conditions and we encouraged them to give it the highest score, if identified.

The test was composed by a total of 18 trials given by the combination of 6 target RIRs (see Section 3.1) and 3 audio stimuli (see Section 3.2). The test lasted about 20 minutes and we recommended the subject to take a break halfway. In a post-test interview we asked the following information about the subject: age range, gender, degree of experience in audio reverberation (None, Intermediate, Expert),

[5] https://plugins.iem.at/docs/plugindescriptions/#fdnreverb

Table 1: Reverberation features for each target RIR and for each reverberation condition. The last row represent the Mean Absolute Percentage Error (MAPE) between the target RIRs features and the ones of the two DARs.

RIR	Reverb	T_{20} [ms]	EDT [ms]	T_S [ms]	G [dB]	C_{80}	LF_{80} [dB]	SC [Hz]
Auditorium Hall	Reference	1557.5	1101.5	30.7	1.5	8.6	-9.1	6647.9
	FDN	1459	1186.5	42.4	2.3	6.5	-5.8	1335.5
	Freeverb	1405	1507.5	37.0	1.4	6.6	-9.9	3158.9
Outdoors Courtyard	Reference	1005	92.5	8.2	0.4	15.4	-10.6	7136.9
	FDN	927	339	11.6	0.6	13.0	-10.0	2989.9
	Freeverb	971	41	10.2	0.4	12.5	-18.7	4460.7
Recording Studio	Reference	427.5	115.1	11.8	2.2	17.1	-1.0	9268.5
	FDN	410	541.5	10.4	1.3	16.0	-6.5	1673.0
	Freeverb	577.5	5	3.7	0.2	18.0	-15.8	6319.7
Small Room	Reference	488	456	32.1	8.2	10.3	-0.3	8427.5
	FDN	474	504	39.2	8.2	8.9	3.1	2311.3
	Freeverb	597	475	12.4	0.7	11.9	-162.7	7799.9
Living Room	Reference	803	487.5	22.6	2.8	11.3	-15.2	9616.2
	FDN	823	819	41.2	3.4	6.5	-4.2	849.4
	Freeverb	784.5	908.5	15.9	0.8	10.6	-11.4	3890.1
Plate Reverb	Reference	525.5	446.5	4.5	0.7	18.1	-14.0	7239.2
	FDN	518	645	10.5	1.1	14.3	-7.7	2102.6
	Freeverb	702	35	3.2	0.2	18.0	-17.5	4730.0
MAPE (%)	FDN	**3.3**	128.3	53.3	**36.6**	20.0	**306.7**	75.0
	Freeverb	16.7	**63.3**	**38.3**	55.0	**13.3**	9290.0	**36.7**

type of used headphones (Sennheiser HD650, Other headphones model, In-Ear/Earbuds) and auditory impairment (Yes/No).

5.2.2 Test execution. The tests were performed both in a controlled laboratory environment and online via a web page. In the laboratory tests, Sennheiser HD650 headphones were employed. In the online version, we advised subjects to use these headphones if possible (hence the post-test question mentioned above).

We collected 16 subjects (15 male and 1 female) to perform the test. MUSHRA guidelines defines as an outlier any subject rating the hidden reference below 90 for more than 15% of the trials. In our test, only one subject was found to be an outlier and was discarded from subsequent analysis. With regard to age, 12 subjects were in the 18-27 range, two in the 28-37 range and one in the 48-57 range. Only one subject reported to be an expert of audio reverberation while the remaining ones had intermediate experience. Only one subject reported to have an auditory impairment. However, the subject correctly identified all hidden references, and was therefore kept for subsequent analysis. All subjects used Sennheiser HD650 headphones, except for one who used a different model.

5.2.3 Test results. The scores for each condition are reported in Fig. 3, which shows that our method achieves higher median scores with respect to the anchors. Median scores are about 70 for both FDN and Freeverb. We investigated the statistical significance of the distribution differences. MUSHRA guidelines recommends to use Analysis of Variance (ANOVA). However, parametric statistical tests like ANOVA require that data satisfy certain assumptions. The study in [19] claims that MUSHRA data typically violates these assumptions and suggests to use non-parametric tests, instead.

We tested the normality of the distributions with the Shapiro-Wilk test [25], and found that the null hypothesis that a population is normal was always rejected for at least one or more of the populations. Thus, we resorted to using the non-parametric Friedman test [10] as omnibus test and the post-hoc Nemenyi test [20]. All statistical tests were performed with the Python library *Autorank* [17].

We set the family-wise significance level of the tests to $\alpha = 0.05$. We decided to investigate the statistical significance of the distributions differences for the overall scores and grouped by audio stimuli (3 groups) and target RIR (6 groups). Therefore, we balanced these multiple comparisons applying the Bonferroni correction [2]. For all groups, the null-hypothesis of the Friedman test that there is no difference in the central tendency of all populations is rejected. Thus, we can apply the Nemenyi post-hoc test. For the overall group, the difference in median scores between our method and the two anchors is statistically significant for both FDN and Freeverb. The difference between FDN and Freeverb is not significant. These results are summarized in Fig. 3. Results of per stimulus and per target RIR analyses are shown in Tables 2 and 3, respectively. Here, the comparison with the hidden reference is left out since its scores are always significantly higher than the other conditions.

Then, we performed the statistical analysis on scores grouped by audio stimuli (drums, sax and speech). The analysis results are shown in Table 2 where for each stimulus we report if the differences between conditions pairs distributions are not statistically significant. For drums and speech stimuli, we found results similar

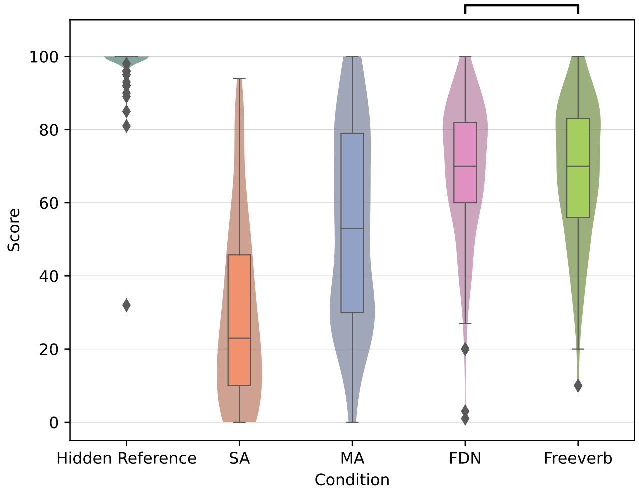

Figure 3: Violin plot of the overall MUSHRA scores distributions for each condition. Lines between conditions means no statistical significant difference.

Table 2: Results of the statistical analysis on the comparison between the score distributions for each condition pair and for each audio stimulus. The black dot • means no significant difference. The last column represent the median score.

		SA	MA	FDN	Freeverb	Med. Score
Drums	SA					14.5
	MA					34.0
	FDN				•	**63.0**
	Freeverb			•		62.0
Sax	SA					39.5
	MA			•	•	**80.0**
	FDN		•		•	**80.0**
	Freeverb		•	•		77.5
Speech	SA					25.0
	MA					47.5
	FDN				•	**70.5**
	Freeverb			•		70.0

to the overall group scenario where FDN and Freeverb scores are significantly higher than the anchors, while their difference it is not significant. On the contrary, for the sax stimulus FDN and Freeverb distributions are not significantly different from the Mid-quality anchor (MA) one. We hypothesize that the low-pass filtered version of the sax stimulus is still very similar to the reference due to the faint high-frequency content of the stimulus. This could explain the high median score of MA for the sax stimulus.

Finally, we analyzed the scores grouping by the six target RIRs. Table 3 shows the comparisons between distributions for each condition pair and for each target RIR. Except for Plate Reverb, the difference with MA is not significant for both FDN and Freeverb for all RIRs. For the Plate Reverb RIR, instead, Freeverb scores distribution is significantly different from the MA one. However, Freeverb scores for the Small Room RIR case is the only case where our method distribution is not significantly different from the SA one. The Small Room is also the only case where the difference between FDN and Freeverb is not significant. The bad performances of Freeverb with the Small Room RIR are probably due to the more perceivable coloration effects of the RIR. So, with the target RIR grouping, the difference between our method and the anchors is more often not significant. This is true even though the median scores of our method is always greater or equal than the anchors' ones with both FDN and Freeverb. This could be explained by the small number of observations for each group in the RIR grouping with respect to the overall group and the stimuli groups.

Table 3: Results of the statistical analysis on the comparison between the score distributions for each condition pair and for each target RIR. The black dot • means no significant difference. The last column represent the median score.

		SA	MA	FDN	Freeverb	Med. Score
Auditorium hall	SA					27.0
	MA			•	•	56.0
	FDN		•		•	69.0
	Freeverb		•	•		**71.0**
Outdoors courtyard	SA		•			29.0
	MA	•		•	•	54.0
	FDN		•		•	**78.0**
	Freeverb		•	•		72.0
Recording studio	SA					21.0
	MA			•	•	60.0
	FDN		•		•	**65.0**
	Freeverb		•	•		61.0
Small room	SA				•	27.0
	MA			•	•	55.0
	FDN		•			**80.0**
	Freeverb	•	•			55.0
Living room (CATT)	SA					27.0
	MA			•	•	51.0
	FDN		•		•	68.0
	Freeverb		•	•		**70.0**
Plate reverb (REVelation)	SA		•			18.0
	MA	•		•		41.0
	FDN		•		•	80.0
	Freeverb			•		**83.0**

6 CONCLUSION

In this paper, we proposed a reverb matching method based on the automatic optimization of DARPs to match a target RIR. We evaluated the method through a MUSHRA test employing two DARs: FDN and Freeverb. Analyzing the test results, both DARs provide an overall good match of the target RIR. Our method has always a median score greater or equal than the anchors' ones although grouping by target RIR the statistical significance of this difference is less frequent. However, this could be explained by a lower number of observations in the RIR groups.

While it is difficult to draw a direct comparison with other existing approaches, in terms objective and subjective evaluations, a number of methodological and computational advantages can be mentioned. First, it is suitable for any DAR in VST3 format (besides, in the MUSHRA we showed that FDN and Freeverb achieve comparable results). Second, no assumption is made on the type of target reverberation, that can be both natural and artificial. Third, the method is not based on deep learning; thus, we perform the reverb matching task without the need of large datasets and expensive training procedures. Lastly, the method is suitable for real-time rendering. The DARPs optimization procedure cannot be run in real-time; however, once the DARPs has been matched to the target RIR, the DAR can process any audio signal in real-time.

We plan to improve several aspects of our method. The modeling of early reflection could be improved. We designed the early reflection extraction method to get a reliable estimation on the employed RIRs. In future, we plan to improve the method and make it suitable for any RIR. Then, early reflections modelling could be performed using DARs that can accurately render them. A potential candidate DAR is Scattering Delay Network (SDN) [7]. Further, about the listening test, we can select more meaningful anchors than those defined in the MUSHRA recommendation. As an example, a possible anchor could be a random setting of DARPs. An additional condition that could be evaluated in the test is the DARPs tuning performed by an expert audio engineer. A further improvement would be the possibility to tune the DARPs given any reverberated signal as input. In fact the need of a RIR in input could be a substantial limitation in end user applications.

The MUSHRA test reported in this paper was aimed at assessing the "authenticity" of the matched reverberation, i.e. the perceived similarity between a real sound and its virtual approximation. Our results suggest that the method does not provide authentic reverberation matching. However, a more meaningful evaluation of the method for AAR scenarios should be aimed at assessing the "transfer plausibility" of the scene [32]. Therefore, we plan to perform a different test where various sound sources with real and virtual reverberation are rendered together in a complex auditory scene.

ACKNOWLEDGMENTS

This work is part of SONICOM, a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017743.

REFERENCES

[1] Spotify AB. 2021. Pedalboard. https://github.com/spotify/pedalboard.

[2] Carlo Emilio Bonferroni. 1936. *Teoria statistica delle classi e calcolo delle probabilita*. Vol. 8. 3–62 pages.

[3] Recommendation ITU-R BS.1534-3. 2015. Method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union* (2015).

[4] CATT. 2020. CATT-Acoustic / The FIReverb Suite / ReflPhinder / GratisVolver Pro. [Online]. Available: https://www.catt.se/, Last accessed on 2022-04-05.

[5] Michael Chemistruck, Kyle Marcolini, and Will Pirkle. 2012. Generating matrix coefficients for feedback delay networks using genetic algorithm. In *Audio Engineering Society Convention 133*. Audio Engineering Society.

[6] Jay Coggin and Will Pirkle. 2016. Automatic design of feedback delay network reverb parameters for impulse response matching. In *Audio Engineering Society Convention 141*. Audio Engineering Society.

[7] Enzo De Sena, Hüseyin Hacıhabiboğlu, Zoran Cvetković, and Julius O Smith. 2015. Efficient synthesis of room acoustics via scattering delay networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 9 (2015), 1478–1492.

[8] Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, and Adam Roberts. 2020. DDSP: Differentiable Digital Signal Processing. In *International Conference on Learning Representations*. https://openreview.net/forum?id=B1x1ma4tDr

[9] International Organization for Standardization. 2009. *ISO 3382-1: International Standard ISO/DIS 3382-1: Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces*. International Organization for Standardization.

[10] Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association* 32, 200 (1937), 675–701.

[11] M. A. Gerzon. 1971(I),1972(II). Synthetic stereo reverberation, parts i and ii. *Studio Sound* 13(I), 14(II) (1971(I),1972(II)), 632–635(I), 24–28(II).

[12] Theodoros Giannakopoulos and Aggelos Pikrakis. 2014. Chapter 4 - Audio Features. In *Introduction to Audio Analysis*. Academic Press, Oxford, 59–103. https://doi.org/10.1016/B978-0-08-099388-1.00004-2

[13] Constant CJM Hak, Remy HC Wenmaekers, and LCJ Van Luxemburg. 2012. Measuring room impulse responses: Impact of the decay range on derived room acoustic parameters. *Acta Acustica united with Acustica* 98, 6 (2012), 907–915.

[14] Villy Hansen and Gert Munch. 1991. Making recordings for simulation tests in the Archimedes project. *Journal of the Audio Engineering Society* 39, 10 (1991), 768–774.

[15] Tim Head, Manoj Kumar, Holger Nahrstaedt, Gilles Louppe, and Iaroslav Shcherbatyi. 2021. Scikit-Optimize. https://doi.org/10.5281/zenodo.5565057

[16] Sebastian Heise, Michael Hlatky, and Jörn Loviscach. 2009. Automatic adjustment of off-the-shelf reverberation effects. In *Audio Engineering Society Convention 126*. Audio Engineering Society.

[17] Steffen Herbold. 2020. Autorank: A Python package for automated ranking of classifiers. *Journal of Open Source Software* 5, 48 (2020), 2173. https://doi.org/10.21105/joss.02173

[18] Sungho Lee, Hyeong-Seok Choi, and Kyogu Lee. 2021. Differentiable Artificial Reverberation. *arXiv preprint arXiv:2105.13940* (2021). https://arxiv.org/abs/2105.13940

[19] Catarina Mendonça and Symeon Delikaris-Manias. 2018. Statistical tests with MUSHRA data. In *Audio Engineering Society Convention 144*. Audio Engineering Society.

[20] Peter Bjorn Nemenyi. 1963. *Distribution-free multiple comparisons*. Ph. D. Dissertation.

[21] Andy Sarroff and Roth Michaels. 2020. Blind arbitrary reverb matching. In *Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx-2020)*.

[22] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*. IEEE, 166–171.

[23] Lauri Savioja and U Peter Svensson. 2015. Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America* 138, 2 (2015), 708–730.

[24] Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre. 2018. webMUSHRA — A Comprehensive Framework for Web-based Listening Tests. *Journal of Open Research Software* 6, 1 (2018). https://doi.org/10.5334/jors.187

[25] Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52, 3/4 (1965), 591–611.

[26] Justin Shen and Ramani Duraiswami. 2020. Data-driven feedback delay network construction for real-time virtual room acoustics. In *Proceedings of the 15th International Conference on Audio Mostly*. 46–52.

[27] Julius O. Smith. accessed 2022-04-27. *Physical Audio Signal Processing*. https://ccrma.stanford.edu/~jos/pasp/ online book, 2010 edition.

[28] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems* 25 (2012).

[29] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. 2020. Acoustic matching by embedding impulse responses. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 426–430.

[30] Vesa Valimaki, Julian D Parker, Lauri Savioja, Julius O Smith, and Jonathan S Abel. 2012. Fifty years of artificial reverberation. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 5 (2012), 1421–1448.

[31] Christopher K Williams and Carl Edward Rasmussen. 2006. *Gaussian processes for machine learning*. Vol. 2. MIT press Cambridge, MA.

[32] Stefan A Wirler, Nils Meyer-Kahlen, and Sebastian J Schlecht. 2020. Towards Transfer-Plausibility for Evaluating Mixed Reality Audio in Complex Scenes. In *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society.