DIAGONAL COMPLEX-VALUED STATE SPACE MODELS FOR SYSTEM IDENTIFICATION AND MODELING OF METAL PLATE REVERBS

Matthias Bittner, Matthias Wess, Dominik Dallinger, Daniel Schnöll and Axel Jantsch

Christian Doppler Laboratory for Embedded Machine Learning
TU Wien
Vienna, Austria
firstname.lastname@tuwien.ac.at

ABSTRACT

Accurate and interpretable modeling of plate reverbs remains an important challenge in virtual analog modeling of audio effects. While existing neural network-based black-box approaches already achieve high-quality synthesis and strong perceptual quality, they often lack the possibility to identify the underlying physically meaningful complex, long-memory modal behavior. In this work, we address this limitation by proposing a restricted complex-valued diagonal State Space Model (SSM), showing its equivalence to a parallel second-order all-pole filter, also utilizing efficient training via parallel state computation using the parallel scan algorithm. Additionally, we propose a Matrix Pencil (MP) guided eigenvalue initialization, improving synthesis quality and system identification performance. We validate the approach on the DAFx plate reverb benchmark, demonstrating accurate impulse response reconstruction and meaningful recovery of modal parameters.

1. INTRODUCTION

Artificial reverberation remains a central topic in virtual analog modeling. In particular, metal plate reverbs remain an important benchmark, combining characteristic studio coloration with the challenging long-memory dynamics of dense modal resonances [1, 2]. Existing approaches to model reverberation can broadly be categorized into physical or structure-informed emulation and black-box signal generation. While black-box models often achieve good perceptual quality, structure-informed approaches offer the important advantage that physically meaningful parameters remain accessible after training, enabling controllable parameterization and system identification.

Recent advances in deep learning have led to a large body of work on black-box emulation of audio effects, including reverberation, using Convolutional Neural Networks (CNNs), Temporal Convolutional Networks (TCNs), and recurrent architectures such as Long Short-Term Memory Networks (LSTMs) [3, 4, 5]. Within this domain, reverberation presents a particularly challenging problem, as the long exponentially decaying responses with high modal density make plate reverbs a natural long-memory sequence modeling task [1, 6]. The line of work using Differential Digital Signal Processing (DDSP) follows the idea of using deep learning to parameterize Digital Signal Processor (DSP) structures directly [7]. For reverberation and virtual analog audio effects, such DSP-informed neural and modal architectures have demonstrated strong perceptual modeling performance. However, they often remain largely black boxes and do not directly encode meaningful modal parameters such as frequencies, damping factors, and gains [7, 8, 9].

At the same time, recent advances in deep State Space Models (SSMs) have established diagonal continuous-time recurrent architectures as highly effective models for long-range sequence and raw audio processing [10, 11, 12, 13]. Their stable, continuous-time pole parameterization and efficient recurrent inference make them particularly attractive for resonant audio systems. While DDSP-based piano audio synthesis models also already demonstrated the effectiveness of embedding structured resonant signal models into differentiable architectures, recent SSM-based piano synthesis showed that diagonal continuous-time SSMs can directly learn similarly rich resonant decays with efficient causal inference and interpretable pole dynamics [14, 15]. Similarly, DDSP-style reverberation models provide strong structured priors [9], while the present work extends this line of work with a system-identification-aware SSM reverberation model.

Motivated by these developments, this work bridges artificial reverberation, modal system identification, DDSP, and trainable SSMs. In particular, we reinterpret a restricted diagonal complex-valued SSM as a differentiable modal resonator structure and combine it with classical pole-estimation-based initialization for efficient and explainable plate reverb modeling and system identification. The main contributions of this work are as follows:

- We show that restricted discrete-time diagonal complex-valued SSMs are exactly equivalent to parallel second-order all-pole modal filter banks, directly linking trainable SSMs to classical plate reverb synthesis. Based on this equivalence, we propose an SSM layer for metal plate reverb modeling with efficient parallel training via associative scan.
- We propose a Matrix Pencil (MP) [16] guided eigenvalue initialization strategy which improves synthesis quality and modal system identification compared to a standard SSM initialization.
- We validate the proposed method on the DAFx plate reverb benchmark and demonstrate accurate impulse-response fitting together with meaningful modal parameter recovery.

The remainder of this paper is organized as follows. Section 2 reviews related work on deep diagonal SSMs, system identification methods for resonant systems, neural virtual analog audio effect modeling, and classical plate modal synthesis. Section 3 introduces the proposed diagonal complex-valued SSM formulation and its interpretation as a parallel second-order all-pole modal structure. Section 4 evaluates the approach on benchmark plate

reverb system identification tasks, followed by conclusions in Section 5. Audio samples are provided here.[1]

2. RELATED WORK

2.1. Deep State Space Models for Sequence Processing

Modern deep SSMs have emerged as strong alternatives to recurrent neural networks for long-range sequence modeling. S4 [10] introduced structured continuous-time state-space layers with strong long-context memory. Later simplifications such as DSS [17] and S4D [11] showed that diagonal complex-valued state dynamics are already highly expressive, while maintaining stable pole parameterizations. S5 further extends this line to a diagonal MIMO formulation with efficient associative scan training [12]. More recent audio-focused work such as S-Edge [13] adapts diagonal continuous-time SSMs to raw audio tasks and hardware-efficient deployment, while the Linear Recurrent Unit (LRU) further simplifies diagonal complex recurrence [18].

The performance of deep SSMs is highly sensitive to the state matrix initialization. S4 addresses this by proposing HiPPO-based continuous-time dynamics that distribute poles over multiple time scales, while DSS and S4D show that strong diagonal Single Input Single Output (SISO) performance can already be achieved by directly diagonalizing the HiPPO-N matrix. S5 extends this same HiPPO-N initialization to the diagonal Multi Input Multi Output (MIMO) setting and empirically demonstrates strong performance in this more general case.

2.2. System Identification Methods

Classical system identification methods for resonant systems often recover modal parameters by fitting the signal as a sum of exponentially damped sinusoids. Early Prony-type methods provide a parametric formulation for estimating modal frequencies, damping factors, and amplitudes [19]. Subspace-based methods such as ESPRIT further improve high-resolution spectral estimation through rotational invariance of the signal subspace [20]. Closely related, the Matrix Pencil Method estimates frequencies and damping factors by solving a generalized eigenvalue problem on measured data and is known to be more robust to noise than polynomial Prony-type approaches [16].

Related work in structured gray-box state-space identification has shown that good initialization is crucial for non-convex optimization of physically constrained models. Yu et al. [21] improve robustness over classical prediction-error methods through low-rank Hankel factorization. Our work follows a similar philosophy, specializing initialization to resonant modal systems via Matrix Pencil pole estimates for diagonal continuous-time SSM eigenvalues.

2.3. Virtual Analog Modeling for Audio Effects

Neural networks have been widely used for virtual analog modeling of audio effects [3]. While Steinmetz et al. [4] demonstrated that TCN with large receptive fields are highly effective for long-memory dynamic audio effects, their receptive-field-dependent latency can hinder real-time deployment for long-memory effects such as reverberation.

[1] Audio samples https://platereverb.github.io/platereverb/

To improve the modeling of longer temporal dependencies, stateful architectures have been investigated. Simionato et al. [5] showed that stateful models such as LSTM can achieve similar accuracy to CNNs at lower latency and computational cost.

Ramírez et al. [9] combine neural latent processing with a DSP-inspired structure for electromechanical reverberators. While these approaches achieve strong perceptual emulation quality, the learned representations remain largely black-box and do not directly expose modal frequencies, damping factors, or resonant gains.

2.4. Plate Reverb Modal Model

Classical plate reverbs are commonly modeled either through *physical partial differential equations solvers* [22] or efficient *real-time modal resonator formulations* [23], where each mode of the plate is represented as an exact discrete-time second-order all-pole biquad filter with the following discrete-time recurrence,

$$q_{m,k+1} = 2r_m \cos(\Omega_m \mathrm{T_s}) q_{m,k} - r_m^2 q_{m,k-1} + w_m f_k. \quad (1)$$

In this equation $q_{m,k}$ denotes the discrete-time amplitude of the m-th resonant mode, Ω_m its angular frequency, $r_m = e^{-\sigma_m \mathrm{T_s}}$ the decay factor derived from the modal damping σ_m, f_k the excitation signal, and w_m the mode-dependent excitation gain and the sampling-time $\mathrm{T_s}$. Applying the z-transform to Equation 1 and considering the full plate reverb as the sum of M modes, the transfer function of the modal reverb architecture results in,

$$G(z) = \sum_{m=1}^{M} \frac{w_m z}{z^2 - 2z r_m \cos(\Omega_m \mathrm{T_s}) + r_m^2}. \quad (2)$$

In Section 3.1, we show that a restricted diagonal complex-valued state space model can be used as an exact equivalent to this filter structure.

3. STATE SPACE MODELS FOR METAL PLATE REVERB MODELING AND SYSTEM IDENTIFICATION

We propose to use a differentiable SSM based model structure to learn plate reverb impulse responses. Figure 1 highlights the general flow of the method. We send an impulse into the network, and the SSM parameters are updated based on a point-wise loss between the target Impulse Response (IR) and the prediction. The aim of our approach is to present a model architecture usable for both virtual-analog modeling (plate reverb synthesis) and system identification. In Section 3.1, we present a restricted complex-valued diagonal SSM system that can be interpreted as a parallel second-order all-pole filter. The learned SSM parameters are then convertible into the modal information (frequency, decay rate, and gain). Additionally, in Section 3.2, we propose a guided eigenvalue/mode initialization scheme that uses the Matrix Pencil [16] pole estimation method for improving system identification and synthesis quality.

3.1. Viewing Diagonal Complex-Valued State Space Models as Parallel Second-Order All-pole Filters

For the purpose of modeling a parallel second-order all-pole filter structure, within the parallelizable and differentiable SSM model frameworks, such as used by S5 [12], LRU [18], and S-Edge [13],

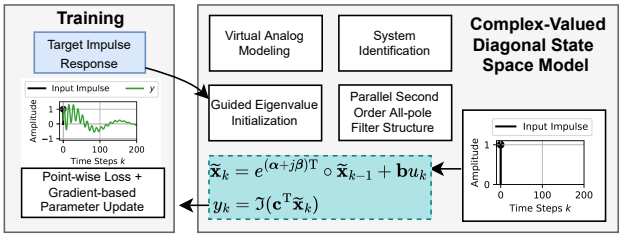

Figure 1: We use a restricted diagonal complex-valued SSM (interpretable as a parallel second-order all-pole filter structure) for system identification and modeling of plate reverbs. The model's parameters are trained to match a target impulse response using a point-wise loss and a guided eigenvalue initialization strategy based on the Matrix Pencil Method.

we define our SSM system by a diagonal complex-valued first-order difference equation,

$$\begin{aligned} \widetilde{\mathbf{x}}_k &= e^{\widetilde{\boldsymbol{\lambda}}\mathrm{T_s}} \circ \widetilde{\mathbf{x}}_{k-1} + \mathbf{b}u_k, \\ y_k &= \Im(\mathbf{c}^{\mathrm{T}}\widetilde{\mathbf{x}}_k). \end{aligned} \tag{3}$$

This SISO system maps a real-valued input $u_k \in \mathbb{R}^1$ to a real-valued output $y_k \in \mathbb{R}^1$, via a complex-valued state $\widetilde{\mathbf{x}}_k \in \mathbb{C}^H$. The dynamics of the state is defined by the continuous-time eigenvalue representation $\widetilde{\boldsymbol{\lambda}} = \boldsymbol{\alpha} + j\boldsymbol{\beta}$. Due to the diagonal form and with Zero Order Hold (ZOH) discretization, $\widetilde{\boldsymbol{\lambda}} \in \mathbb{C}^H$ maps to discrete time by $e^{\widetilde{\boldsymbol{\lambda}}\mathrm{T_s}}$, with $\mathrm{T_s}$ being the sampling rate and $\circ$ the element-wise vector product and H the hidden state size (order) of the system. This diagonal eigenvalue representation allows constraining the system to be stable during training by restricting the real part of the eigenvalues to be strictly negative (which we do by overwriting $\alpha = -\mathrm{abs}(\alpha)$). Additionally, during training, we constrain the eigenvalues to have one-sided poles $\beta = \mathrm{abs}(\beta)$. In contrast to existing SSM layers such as S5, S-Edge, and LRU, we restrict the input weights $\mathbf{b} \in \mathbb{R}^H$ and output weights $\mathbf{c} \in \mathbb{R}^H$ to be real-valued. Additionally, instead of taking the real part of the complex-valued output (which is the common choice for deep SSMs), we take the imaginary part $\Im(\mathbf{c}^{\mathrm{T}}\widetilde{\mathbf{x}}_k)$.

Considering an implementation using the C++ inference framework as proposed in S-Edge [13], the total number of parameters for this SISO system results in Total Parameters $= 4H$ (counting complex-valued parameters as two parameters). The computational demand can be estimated by Multiply ACcumulates (MACs). A complex-valued matrix–vector multiplication requires four times more MACs compared to the real-valued case. However, since our input and output vectors are real-valued, we only have the state update that requires the full complex-valued MACs, resulting in Total MACs $= 6H$ for processing one time step k with our model. If we consider the fact that the weight vectors $\mathbf{b}, \mathbf{c}$ can be fused, see Equation 8, parameter counts and number or MACs reduce to: Total Parameters $= 3H$, Total MACs $= 5H$.

In the next few steps, we show that the proposed specific input-output mapping is equivalent to a parallel second-order all-pole filter structure. Applying the z-transform to the input-output mapping (Equation 3, for now considering the full complex output) allows us to go from the discrete-time domain k into the z-domain,

$$\begin{aligned} \widetilde{\mathbf{x}}_z(z) &= z^{-1}e^{\widetilde{\boldsymbol{\lambda}}\mathrm{T_s}} \circ \widetilde{\mathbf{x}}_z(z) + \mathbf{b}u_z(z), \\ \widetilde{y}_z(z) &= \mathbf{c}^{\mathrm{T}}\widetilde{\mathbf{x}}_z(z), \end{aligned} \tag{4}$$

with the known relationship of $z = e^{j\omega\mathrm{T_s}}$ and the angular frequency ω. Based on this representation and the diagonal structure, we are able to formalize the complex-valued discrete-time transfer function as a sum of first-order complex resonators,

$$\widetilde{G}(z) = \frac{\widetilde{y}_z(z)}{u_z(z)} = \sum_{i=1}^{H} c_i b_i \frac{z}{z - e^{\widetilde{\lambda}_i \mathrm{T_s}}}. \tag{5}$$

Applying the inverse z-transform of this representation and taking the imaginary part as proposed in Equation 3 gives the real-valued discrete-time impulse response,

$$\begin{aligned} g_k = \Im(\widetilde{g}_k) &= \Im\left(\sum_{i=1}^{H} c_i b_i e^{\widetilde{\lambda}_i k\mathrm{T_s}}\right) \\ &= \Im\left(\sum_{i=1}^{H} c_i b_i e^{(\alpha_i + j\beta_i)k\mathrm{T_s}}\right) \\ &= \Im\left(\sum_{i=1}^{H} c_i b_i e^{\alpha_i k\mathrm{T_s}} \left(\cos(\beta_i k\mathrm{T_s}) + j\sin(\beta_i k\mathrm{T_s})\right)\right) \\ &= \sum_{i=1}^{H} c_i b_i e^{\alpha_i k\mathrm{T_s}} \sin(\beta_i k\mathrm{T_s}), \end{aligned} \tag{6}$$

which represents a weighted sum of exponentially decaying sinusoids. The real part of the eigenvalue α_i defines the decay rate, and the imaginary part β_i defines the angular frequency for each individual mode. If we now again apply the z-transform on g_k we get the actual real-valued transfer function representation of our proposed input-output mapping,

$$G(z) = \sum_{i=1}^{H} c_i b_i \frac{z e^{\alpha_i \mathrm{T_s}} \sin(\beta_i \mathrm{T_s})}{z^2 - 2z e^{\alpha_i \mathrm{T_s}} \cos(\beta_i \mathrm{T_s}) + e^{2\alpha_i \mathrm{T_s}}}. \tag{7}$$

Performing a coefficient comparison with the parallel second-order filter structure used to model a plate reverb (Equation 2), we get the following relationship between the SSM parameters and the mode parameters of the reverb model for an individual mode,

$$\Omega_m = \beta_i, \quad \sigma_m = -\alpha_i, \quad w_m = c_i b_i e^{\alpha_i \mathrm{T_s}} \sin(\beta_i \mathrm{T_s}). \tag{8}$$

This specific mapping shows that the learned SSM parameters can be interpreted as modal information for plate reverbs modeled with parallel second-order all-pole filters.

Within our experiments, we empirically evaluate two things. First, the synthesis quality when training the proposed SSM representation to predict target impulse responses (see Section 4.2). Second, we evaluate their performance from a system identification perspective by measuring similarity between the ground truth and estimated/learned modal parameters (see Section 4.3)

3.2. Impulse Response Guided Eigenvalue Initialization with the Matrix Pencil Method

Plate reverbs can be modeled by parallel second-order all-pole filter structures (see Section 2.4), whose impulse responses are con-

structed as the sum of exponentially decaying sinusoids. In Section 3.1, specifically within the Equations 3–6, we show that a restricted complex-valued diagonal SSM structure can have the exact same impulse response/transfer function structure as a second-order all-pole filter.

With the aim of providing the model with a well-suited eigenvalue initialization enabling it to achieve a better optimum, in contrast to common structured eigenvalue initialization, e.g., as used in S5 [12], we propose a guided eigenvalue initialization based on the MP method [16]. Specifically, we perform a two-step approach:

1. Given a target plate reverb impulse response, we estimate a subset of modes $\boldsymbol{\lambda}_{\text{est}} = \boldsymbol{\alpha}_{\text{est}} + j\boldsymbol{\beta}_{\text{est}}$ with the MP method [16].
2. Based on the estimated poles $\boldsymbol{\lambda}_{\text{est}} \in \mathbb{C}^{H_{\text{est}}}$, we fit a polynomial function to the decay rate and frequency relationship and use this function to sample additional eigenvalues by first sampling new frequencies in a log-uniform distribution between $\beta_{\min}$ and $\beta_{\max}$. We then evaluate the decay rates based on the fitted function up to the desired order $H = H_{\text{est}} + H_{\text{sampled}}$.

The matrix pencil method for pole estimation provides a good estimate of modes that are well distinguishable from noise. We therefore generally consider choosing $H_{\text{est}} << H$. The estimated poles $\boldsymbol{\lambda}_d$ are identified in discrete-time and in conjugate-complex pole pairs, since our impulse responses are real-valued and sampled. For making them compatible with the formulation in Equation 3, we just take one-sided poles and convert them back to continuous-time by $\boldsymbol{\lambda} = \ln(\boldsymbol{\lambda}_d)/\mathrm{T_s}$.

In our experiments (Sections 4.2–4.3), we refer to this structured eigenvalue initialization, based on a combination of estimated and sampled eigenvalues, as Matrix Pencil (MP) initialization and demonstrate its benefits on synthesis performance and modal system identification compared to the standard S5 eigenvalue initialization strategy.

4. EXPERIMENTS: SYSTEM IDENTIFICATION FOR METAL PLATE REVERBS

We base our experiments on the Benchmark for Plate Reverb System Identification[2] proposed by Ducceschi and Gabrielli [6]. We use 50 randomly synthesized impulse responses with a sampling rate $\mathrm{f_s} = 1/\mathrm{T_s} = 44.1\,\mathrm{kHz}$ and the default configurations within their benchmark repository. The synthesized IRs are 1 second long. For each of the 50 target IRs we analyze 5 different hidden state dimension $H \in \{128,\ 256,\ 512,\ 1024,\ 2048\}$ and 2 different eigenvalue initialization strategies, S5 [12] and the proposed MP based one. Overall, we trained 500 models, which are the basis for our experimental evaluations.

In Section 4.2, we evaluate the synthesis quality by comparing the target and the predicted IRs in the time and frequency domain. In Section 4.3, we show performance on the system identification task by measuring the similarity between detected and ground truth modal parameters (frequencies, decay rates, gains) that define the plate reverb in its modal representation.

[2]A Benchmark for Plate Reverb System Identification https://github.com/LOGUNIVPM/1st-DAFx-Challenge

4.1. Experimental Setup

Experiments and the code are implemented in Python and PyTorch. The SSM code base was modified and adapted from the open-source S-Edge [13] implementation. For each target impulse response, we train a separate SSM layer (as proposed in Equation 3). In contrast to traditional deep learning, we aim to achieve a perfect fit (overfitting) on the target IR, by fitting our models for 5000 steps using L2-loss. The checkpoint with the lowest loss is considered during evaluation. As an optimizer, we use AdamW and a cosine annealing learning rate scheduler, starting at 0.001. To keep gains within a small range, we apply aggressive weight decay (factor of 10) only to the input weights $\mathbf{b}$. At initialization, we sample the input weights from a normal distribution with zero mean and a standard deviation of $0.001\sqrt{2/3}$. The output weights $\mathbf{c}$ are set to one by default and kept untrained. As it is the common practice during SSM training [12, 13], we make use of the step-scale $\boldsymbol{\Delta}$ parameter, decoupling the learning of the eigenvalues into two parameters $(\boldsymbol{\lambda}', \boldsymbol{\Delta})$, where $\boldsymbol{\Delta}$ is trained and initialized in the log-domain. In our experiments initialized uniformly in the range of $[\ln(0.01),\ \ln(0.0001)]$. So, from the optimization point of view, we learn two parameters that compose the true learned eigenvalue by $\boldsymbol{\lambda} = \boldsymbol{\lambda}' \circ \boldsymbol{\Delta}$. For the MP-based eigenvalue initialization, we always first estimate $H_{\text{est}} = 128$ modes, which are used to fit a second-order polynomial function. We then sample $H_{\text{sample}} = H - H_{\text{est}}$ additional eigenvalues from this function. The frequency range for sampling and extrapolating to higher frequency/decay rate pairs is set to $\beta_{\min}/2\pi = 20\,\mathrm{Hz}$ and $\beta_{\max}/2\pi = 10\,\mathrm{kHz}$. Please note that for the smallest tested model configuration $H = H_{\text{est}} = 128$, we do not sample and take the estimated poles for initialization.

4.2. Synthesis Performance

Given the target impulse response g and the model's prediction $\hat{g}$, we evaluate the performance in the discrete-time domain by

$$\mathrm{MSE}(\hat{g}, g) = \frac{1}{T}\sum_{k=1}^{T}(\hat{g}_k - g_k)^2 \tag{9}$$

$$\text{Normalized L2 Error}(\hat{g}, g) = \frac{\|\hat{g} - g\|_2}{\|g\|_2} \tag{10}$$

with T being the length of the IR. Additionally, we evaluate the performance in the frequency domain by

$$\text{Norm. Spec. Mag. Error} = \frac{1}{K}\sum_{k=1}^{K}\frac{\left|\,|\hat{G}_k| - |G_k|\,\right|}{|G_k| + \epsilon}, \tag{11}$$

while $\mathbf{G} \in \mathbb{C}^{T/2+1}$ depicts the spectrum of the impulse response, calculated with the discrete Fourier transform.

Figure 2 shows an example based on the plate reverb impulse response with the highest number of modes within our investigated IRs. It compares the ground truth with the SSM predictions, trained with either S5 or MP eigenvalue initialization in the time and frequency domain. The absolute errors already indicate the better performance of the MP based eigenvalue initialization strategy. Interestingly, we also observe that the model trained with the S5 eigenvalue initialization has a higher absolute error in the frequency domain at high frequencies, indicating an ill-suited frequency-range initialization.

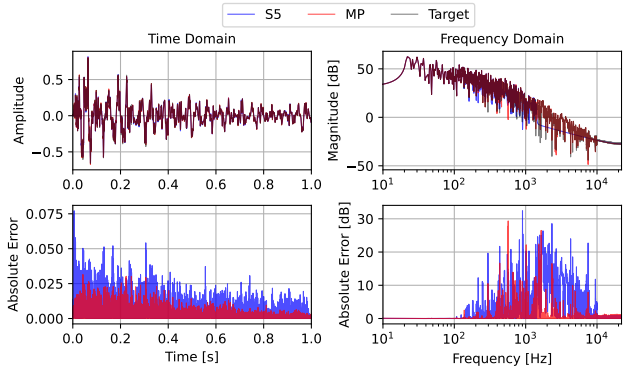

Figure 2: Example impulse response with the highest number of modes within the dataset, comparing the ground truth to SSM predictions ($H = 1024$) either trained with S5 or MP eigenvalue initialization. We present the IR and the absolute errors in the time and frequency domains.

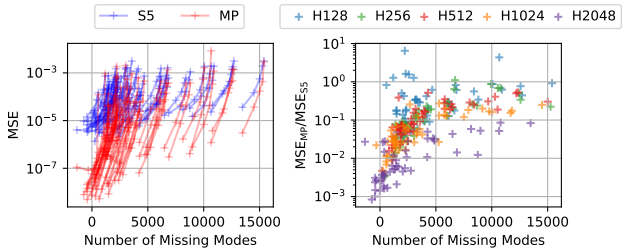

Figure 3: Synthesis performance. Left: Mean Squared Error (MSE) between predicted and ground truth impulse responses versus missing modes, comparing S5 and MP initialization. Fewer missing modes (larger hidden state) consistently reduce the error. Right: Error ratio $\mathrm{MSE}_{\mathrm{MP}}/\mathrm{MSE}_{\mathrm{S5}}$, showing that the proposed MP initialization generally outperforms S5, with minor outliers at small hidden sizes ($H = 128$).

The left subplot in Figure 3 shows the MSE between the predicted and ground truth IRs, also showing the tradeoff between the two eigenvalue initialization strategies S5 and MP. Each line represents the errors for one of the 50 IRs over the number of missing modes (difference between the ground truth number of modes and the hidden state size used for SSM model fitting). Independent of the eigenvalue initialization strategy, we observe a clear trend of a lower MSE with fewer missing modes. Table 1 also supports these findings by reporting the normalized L2 error and the normalized spectral magnitude error on average for the 50 analyzed impulse responses. Both metrics confirm the clear trend of a decreasing error as the hidden state size increases. Please note that for the $H = 2048$ models and some IRs, we already have a higher SSM model order than theoretically necessary (i.e., a negative number of missing modes). The right subplot in Figure 3 shows the ratio $\mathrm{MSE}_{\mathrm{MP}}/\mathrm{MSE}_{\mathrm{S5}}$ between MP and S5 eigenvalue init. Except for some outliers for models with a hidden state size of $H = 128$, we find that the proposed eigenvalue initialization strategy outperforms the S5-init. On average, across the 50 analyzed IRs, we achieve better performance for each state-size configuration with the MP eigenvalue init (see Table 1).

Table 1: Synthesis quality comparison, showing the normalized L2 Error and the normalized spectral magnitude error for the tested SSM hidden state sizes H. We report the average metrics over the 50 analyzed impulse responses and compare our proposed MP eigenvalue init with the common S5-init [12]. Additionally we show the results for the baseline as presented in [6], when taking the estimated modes as initialization for our SSM.

	Norm. L2 Error		Norm. Spec. Mag. Error	
	MP	S5	MP	S5
$H=$ 128	**0.155**	0.259	**0.624**	0.635
$H=$ 256	**0.068**	0.173	**0.409**	0.547
$H=$ 512	**0.039**	0.105	**0.218**	0.513
H=1024	**0.016**	0.060	**0.133**	0.473
H=2048	**0.004**	0.034	**0.103**	0.449
baseline	0.281		0.830	

Ducceschi and Gabrielli propose a baseline for the modal parameter Estimation (Task B) [6]. They use a two-stage procedure. First, they identify modal peaks in the magnitude response to estimate modal frequencies and bandwidths, from which damping (loss) coefficients are derived using the half-power bandwidth rule. Second, modal gains are estimated by evaluating the imaginary part of the frequency response at the detected peak locations. We use their baseline default implementation to estimate the modal parameters for each of our 50 analyzed impulse responses. The identified modal parameters are then converted to our SSM representation, and we again evaluate the synthesis performance by exciting the baseline systems with an impulse and comparing the resulting IRs to the ground truth. The baseline solution searches for modes between 20 Hz–10 kHz and, on average (over the 50 analyzed IRs), detects 139 modes. In Table 1, we refer to this experiment as the baseline, which we are able to outperform with our smallest SSM model configuration only using $H = 128$ states/modes.

4.3. System Identification Performance

To assess the system identification performance, we compare the ground truth modal model parameters, frequency $\boldsymbol{\Omega} \in \mathbb{R}^M$, decay rates $\boldsymbol{\sigma} \in \mathbb{R}^M$ and gains $\mathbf{w} \in \mathbb{R}^M$, used for generating the synthetic plate IR responses with the estimated modal parameters ($\hat{\boldsymbol{\Omega}} \in \mathbb{R}^H$, $\hat{\boldsymbol{\sigma}} \in \mathbb{R}^H$, $\hat{\mathbf{w}} \in \mathbb{R}^H$) derived from the learned SSM parameters. (For the specific mapping, see Section 3.1). Specifically, as also similarly proposed in [6], we evaluate the following similarity metrics for the modal parameters,

$$R_\Omega = \frac{1}{M} \sum_{m=1}^{M} \frac{|\hat{\Omega}_m - \Omega_m|}{|\Omega_m|}, \tag{12}$$

$$R_\sigma = \frac{1}{M} \sum_{m=1}^{\bar{M}} \frac{|\hat{\sigma}_m - \sigma_m|}{\sigma_m}, \tag{13}$$

$$R_w = \frac{1}{M} \sum_{m=1}^{M} \frac{|\hat{w}_m - w_m|}{|w_m|}. \tag{14}$$

We determine the correspondence between the estimated and ground truth parameters by frequency, sorting the estimated parameters by their closest frequency match to the ground truth data. In most cases, the number of ground truth modal parameters will

Table 2: System Identification metric comparison, showing relative errors with respect to the estimated modal parameters (frequency Ω, decay rates σ, gain w). We report the average across the 50 analyzed impulse responses for each individual state-size H configuration, and compare the two initialization strategies, S5 and MP. We also report baseline results using the DAFx challenge implementation [6].

	R_Ω		R_σ		R_w	
	MP	S5	MP	S5	MP	S5
H= 128	**0.957**	0.986	**0.970**	1.534	**1.124**	1.307
H= 256	**0.934**	0.981	**0.974**	1.778	3.031	**1.537**
H= 512	**0.897**	0.974	**0.956**	1.976	8.290	**2.273**
H=1024	**0.837**	0.967	**0.904**	2.292	10.850	**3.303**
H=2048	**0.740**	0.955	**0.829**	2.377	14.149	**7.203**
baseline	0.965		1.047		1.133	

differ from the defined hidden state size ($M \neq H$). We therefore assign a value of 0 to each unidentified modal parameter.

Table 2 summarizes the average system identification metrics over the 50 analyzed plate IRs. We compare the two different eigenvalue initialization schemes over increasing SSM hidden state sizes H. Aligned with the results in Section 4.2 of achieving a better point-wise match between ground truth and predicted IR when using MP eigenvalue init instead of the naive S5 init, we also achieve better system identification performance for the frequency and the decay rates. However, we observe a worse gain match. We also report the metrics for the baseline (generated with the two-stage approach, as described in [6], Task B).

With our smallest SSM model configuration ($H = 128$) and the proposed MP eigenvalue-init strategy, we slightly outperform the baseline on all metrics. While increasing the state size yields significant improvements in both frequency and decay rate similarity, we face issues in matching the correct gains. Though we already applied weight regularization during training to prevent gains from growing, we still face certain outliers that worsen the gain match. Future work will focus on improving the gain similarity metric through more effective regularization strategies. Additionally, post-processing approaches, such as outlier filtering or model-order-reduction-based pruning techniques [24, 25, 26, 27], represent a promising direction, as they have already proven effective in removing redundant states in deep SSMs.

5. CONCLUSION

This work introduces a restricted diagonal complex-valued SSM layer and shows its equivalence to a parallel second-order all-pole filter structure. Leveraging this formulation, we effectively model plate reverbs by fitting impulse responses, while benefiting from the efficient, parallelizable training of modern diagonal SSMs. Furthermore, a guided eigenvalue initialization strategy based on Matrix Pencil pole estimation is shown to improve both synthesis quality and modal parameter identification compared to a standard SSM initialization. Results are demonstrated on the DAFx plate reverb system identification benchmark. Extending the presented linear approach, future work will focus on modeling nonlinear audio effects and optimizing system identification performance by automatically determining the optimal number of modal parameters.

6. ACKNOWLEDGMENTS

This work is supported by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology, and Development, the Christian Doppler Research Association.

7. REFERENCES

[1] V. Välimäki, J. D. Parker, L. Savioja, J. O. S. III, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Trans. Speech Audio Process.*, vol. 20, no. 5, pp. 1421–1448, 2012.

[2] K. Arcas and A. Chaigne, "On the quality of plate reverberation," *Applied Acoustics*, vol. 71, no. 2, pp. 147–156, 2010.

[3] E. Damskägg, L. Juvela, E. Thuillier, and V. Välimäki, "Deep learning for tube amplifier emulation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*. IEEE, 2019, pp. 471–475.

[4] C. J. Steinmetz and J. D. Reiss, "Efficient neural networks for real-time analog audio effect modeling," in *152nd Audio Engineering Society Convention*, 2022.

[5] R. Simionato and S. Fasciani, "Comparative study of state-based neural networks for virtual analog audio effects modeling," *EURASIP J. Audio Speech Music. Process.*, vol. 2025, no. 1, p. 30, 2025.

[6] M. Ducceschi and L. Gabrielli, "The 1st dafx parameter estimation challenge: A benchmark for plate reverb system identification," 2026.

[7] B. Hayes, J. Shier, G. Fazekas, A. P. McPherson, and C. Saitis, "A review of differentiable digital signal processing for music & speech synthesis," *CoRR*, vol. abs/2308.15422, 2023.

[8] R. Diaz, B. Hayes, C. Saitis, G. Fazekas, and M. B. Sandler, "Rigid-body sound synthesis with differentiable modal resonators," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*. IEEE, 2023, pp. 1–5.

[9] M. A. M. Ramírez, E. Benetos, and J. D. Reiss, "Modeling plate and spring reverberation using A dsp-informed deep neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*. IEEE, 2020, pp. 241–245.

[10] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *The Tenth International Conference on Learning Representations, ICLR*, 2022.

[11] A. Gu, K. Goel, A. Gupta, and C. Ré, "On the parameterization and initialization of diagonal state space models," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[12] J. T. Smith, A. Warrington, and S. Linderman, "Simplified state space layers for sequence modeling," in *The Eleventh International Conference on Learning Representations*, 2023.

[13] M. Bittner, D. Schnöll, M. Wess, and A. Jantsch, "Efficient and interpretable raw audio classification with diagonal state space models," *Machine Learning*, vol. 114, no. 8, Jun. 2025.

[14] L. Renault, R. Mignot, and A. Roebel, "Ddsp-piano: A neural sound synthesizer informed by instrument knowledge," *Journal of the Audio Engineering Society*, vol. 71, no. 9, pp. 552–565, 2023.

[15] D. Dallinger, M. Bittner, D. Schnöll, M. Wess, and A. Jantsch, "Piano-ssm: Diagonal state space models for efficient midi-to-raw audio synthesis," in *Proc. Intl. Conf. Digital Audio Effects (DAFx25)*, Ancona, Italy, Sept. 2–5, 2025.

[16] Y. Hua and T. Sarkar, "Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 5, pp. 814–824, 1990.

[17] A. Gupta, A. Gu, and J. Berant, "Diagonal state spaces are as effective as structured state spaces," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22, 2022.

[18] A. Orvieto, S. L. Smith, A. Gu, A. Fernando, C. Gulcehre, R. Pascanu, and S. De, "Resurrecting recurrent neural networks for long sequences," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[19] R. Kumaresan and D. Tufts, "Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 6, pp. 833–840, 1982.

[20] R. H. R. III and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 37, no. 7, pp. 984–995, 1989.

[21] C. Yu, L. Ljung, and M. Verhaegen, "Identification of structured state-space models," *Autom.*, vol. 90, pp. 54–61, 2018.

[22] S. Bilbao, *Numerical sound synthesis: finite difference schemes and simulation in musical acoustics*. John Wiley & Sons, 2009.

[23] M. Ducceschi, C. J. Webb, and G. Fratoni, "Efficient synthesis of large room impulse responses in the modal domain," in *2025 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, 2025, pp. 1–8.

[24] M. Bittner, D. Schnöll, D. Dallinger, M. Wess, and A. Jantsch, "Pruning state space models with model order reduction for efficient raw audio classification," in *2025 33rd European Signal Processing Conference (EUSIPCO)*, 2025, pp. 271–275.

[25] M. Forgione, M. Mejari, and D. Piga, "Model order reduction of deep structured state-space models: A system-theoretic approach," *CoRR*, vol. abs/2403.14833, 2024.

[26] M. Gwak, S. Moon, J. Ko, and P. Park, "Layer-adaptive state pruning for deep state space models," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[27] P. Schwerdtner, J. Berman, and B. Peherstorfer, "Hankel singular value regularization for highly compressible state space models," in *39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025.