UNSUPERVISED ESTIMATION OF NONLINEAR AUDIO EFFECTS: COMPARING DIFFUSION-BASED AND ADVERSARIAL APPROACHES

Eloi Moliner[*1], *Michal Švento*[*2], *Alec Wright*[3], *Lauri Juvela*[1], *Pavel Rajmic*[2] *and Vesa Välimäki*[1]

[1]Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Espoo, Finland
[2]Department of Telecommunications, FEEC, Brno University of Technology, Brno, Czech Republic
[3]Acoustics and Audio Group, University of Edinburgh, Edinburgh, UK
[*]Equal contribution eloi.moliner@aalto.fi michal.svento@vut.cz

ABSTRACT

Accurately estimating nonlinear audio effects without access to paired input-output signals remains a challenging problem. This work studies unsupervised probabilistic approaches for solving this task. We introduce a method, novel for this application, based on diffusion generative models for blind system identification, enabling the estimation of unknown nonlinear effects using black- and gray-box models. This study compares this method with a previously proposed adversarial approach, analyzing the performance of both methods under different parameterizations of the effect operator and varying lengths of available effected recordings. Through experiments on guitar distortion effects, we show that the diffusion-based approach provides more stable results and is less sensitive to data availability, while the adversarial approach is superior at estimating more pronounced distortion effects. Our findings contribute to the robust unsupervised blind estimation of audio effects, demonstrating the potential of diffusion models for system identification in music technology.

1. INTRODUCTION

Audio effects play a pivotal role in shaping the timbral characteristics of music and audio signals. Systems such as guitar amplifiers, dynamic range compressors, and fuzz pedals introduce complex nonlinear transformations that define their sonic signature. Emulating these transformations with software has important applications in music production. When paired input-output recordings are available, supervised learning methods can effectively approximate these transformations [1, 2]. However, in many practical scenarios, obtaining paired data is challenging or infeasible. This occurs when attempting to reproduce the effects processing used in a recording or when the target device is unavailable. Such a situation necessitates blind estimation, where the effect must be inferred without access to the dry input signal, making it a highly underdetermined problem due to the unknown source.

Existing approaches are primarily data-driven, with most relying on supervised learning and pre-training deep neural networks on diverse effect populations, assuming they will generalize to unseen systems. Since paired data can often be generated on-the-fly by randomizing effect parameters, some of these methods are classified as self-supervised. For example, prior works trained models to infer effect parameters from processed audio [3, 4] or predict signal chains [5, 6, 7]. Other approaches use contrastive learning to determine representations that extract effect information [8, 9, 10, 11], which can enable conditional one-to-many-effect modeling [8, 9] or serve as optimization objectives for inference-time effect matching [10]. Another strategy estimates the dry signal with an effect removal model, trained on a diverse effect population, before applying supervised effect estimation [12]. While these methods succeed in modeling unseen effects, their performance depends heavily on the quality and diversity of pre-training data. Generalization requires large, well-curated effect populations, and many methods suffer from flexibility constraints, being limited to simple audio effect implementations with few controllable parameters [10, 3, 4, 6] or fixed architectures dictated by conditional modeling paradigms [9, 8].

This work explores unsupervised approaches that rely only on unpaired examples from input and output data distributions, avoiding reliance on predefined effect populations and their associated generalization challenges. We focus on methods that are agnostic to the functional form of the effect model, allowing them to optimize arbitrary operator models, including black-box and gray-box models, as long as they are suitable for modeling the system. This flexibility is particularly advantageous for emulating real-world effects with unknown or highly complex behaviors, and the optimized operator can serve as a computationally-cheap and real-time capable audio effect. Specifically, we investigate two probabilistic frameworks that fulfill these criteria: one based on generative diffusion models, which is novel in this context, and another based on adversarial training, which has been proposed previously [13].

Diffusion models [14, 15] have emerged as state-of-the-art generative modeling techniques across various domains, including audio generation [16, 17] and restoration [18, 19]. However, their utility in this work does not lie in their generative capabilities but in their potential to serve as data-driven priors for unsupervised system identification, a relatively unexplored application. Recent studies have applied diffusion models to blind inverse problems [20, 21], where the unknown degradation operator is optimized jointly with the clean signal estimate, given only distorted measurements and a diffusion model trained on examples from the reference signal distribution. Such an approach has been explored in historical music restoration [18, 22] and speech dereverberation [23, 24], where *linear* degradation operators were optimized as a byproduct of the restoration process. Building on previous findings [25], which demonstrated that diffusion-based methods can estimate a memoryless nonlinearity, we now extend this methodology to general classes of *nonlinear* operators in audio.

Methods based on adversarial training aim to align the output of a learned effect model with the target distribution using a discriminator model, which is trained with an adversarial objective to the effect model. Wright et al. [13] first proposed an adver-

Copyright: © 2025 Eloi Moliner et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, adaptation, and reproduction in any medium, provided the original author and source are credited.

sarial approach for unsupervised guitar amplifier modeling with unpaired data, employing discriminators designed in the spectrogram domains. Chen et al. [26] later extended this framework by experimenting with alternative discriminator architectures. Recently, Park et al. [27] applied adversarial training to optimize a larger family of audio effects and degradation operators, though their method conditions on unpaired effected measurements and requires pre-training with a population of known audio effects, as in supervised approaches.

While adversarial approaches have demonstrated success in unsupervised effect estimation, they are known to suffer from training instabilities. In particular, discriminator collapse can occur when the discriminator's training dynamics become unbalanced relative to the operator, a risk exacerbated by limited or unbalanced data availability [28, 29]. In many practical cases, one cannot assume access to a sufficiently large and diverse set of target-domain data, increasing the likelihood of instability and poor optimization outcomes. By contrast, the diffusion-based approaches do not rely on adversarial training and are potentially less susceptible to mode collapse or training instabilities. This suggests that diffusion models may provide a more reliable solution for unsupervised nonlinear system identification, particularly in data-scarce scenarios.

This paper is structured as follows. Sec. 2 formalizes the problem from a probabilistic perspective. Sec. 3 introduces the two approaches under comparison, describing their key principles. Sec. 4 describes the operator models used in our experiments, including black-box models parameterized with neural networks, as well as a gray-box model based on a Wiener-Hammerstein (W-H) structure with a novel design. Sec. 5 evaluates both methods in guitar distortion modeling, analyzing their behavior when a limited amount of target-domain data is available. The range of operator models that each method can optimize is explored, and their suitability for blind system identification in guitar distortion effects is assessed. Additionally, we compare a W-H model to black-box models in this setting. Finally, Sec. 6 summarizes our findings.

2. PROBLEM FORMULATION

Let $\mathcal{X} = \{\mathbf{x}^{(m)}\}_{m=1}^{M}$ denote a dataset of source audio signals, where each $\mathbf{x} \in \mathbb{R}^L$ is assumed to be drawn from a prior distribution p_x. Additionally, let $\mathcal{Z}_0 = \{\mathbf{z}_0^{(n)}\}_{n=1}^{N}$ represent an independent dataset of signals also drawn from the same distribution p_x. Importantly, the dataset $\mathcal{Z}_0$ is unobserved; instead, we are provided with a dataset $\mathcal{Y} = \{\mathbf{y}^{(n)}\}_{n=1}^{N}$ of effected measurements. Each observed signal $\mathbf{y} \in \mathbb{R}^L$ is assumed to be generated from a corresponding clean source signal $\mathbf{z}_0$ through an unknown distortion process, described by a function

$$\mathbf{y} = f(\mathbf{z}_0), \tag{1}$$

where $f : \mathbb{R}^L \to \mathbb{R}^L$ represents the distortion operator. The function f is assumed to be deterministic, time-invariant, and otherwise unknown. Our goal is to estimate it.

Since $\mathbf{z}_0$ follows the prior distribution p_x, the distribution of $\mathbf{y}$ is induced through the transformation f. Moreover, since f is deterministic, the conditional distribution of the distorted measurements $\mathbf{y}$ given the source signal $\mathbf{z}_0$ is expressed as a Dirac delta:

$$p(\mathbf{y}|\mathbf{z}_0) = \delta(\mathbf{y} - f(\mathbf{z}_0)). \tag{2}$$

This means that p_y is implicitly defined as

$$p_y(\mathbf{y}) = \int_{\mathbb{R}^L} \delta(\mathbf{y} - f(\mathbf{z}_0))p_x(\mathbf{z}_0)\,\mathrm{d}\mathbf{z}_0. \tag{3}$$

Since f may not be invertible, different values of $\mathbf{z}_0$ can map to the same $\mathbf{y}$, leading to potential density transformations that are difficult to express in closed form.

3. UNSUPERVISED OPERATOR ESTIMATION

To approximate the unknown function f, we introduce a parametric model $\hat{f}(\cdot\,;\psi)$, where ψ represents the learnable parameters. This model can take various functional forms, ranging from black box approaches, such as neural networks including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or state space models, to gray box approaches, such as W-H models, which incorporate stronger inductive biases that reflect prior knowledge about the distortion process. Our objective is to optimize ψ using the unpaired datasets $\mathcal{X}$ and $\mathcal{Y}$, i.e., in the absence of explicit supervision through paired samples.

In this section, we explore two distinct approaches for tackling the unpaired estimation task. First, we introduce a novel method based on an Expectation-Maximization (EM) objective using diffusion models, inspired by recent advances in image inverse problems [21] and speech dereverberation [24]. Second, we examine the adversarial approach proposed recently [13], which formulates the problem as a generative adversarial learning task. These approaches use different strategies for estimating $\hat{f}(\cdot\,;\psi)$, and are summarized in Fig. 1.

In the diffusion-based approach of Fig. 1(a), the source dataset $\mathcal{X}$ is used to train a score model s_θ. Diffusion posterior sampling is applied to estimate the unseen variables in the set $\mathcal{Z}_0$, from which the distorted dataset $\mathcal{Y}$ originates. The operator parameters ψ are optimized through EM updates, while the estimates of the elements from $\mathcal{Z}_0$ are refined.

In the adversarial approach of Fig. 1(b), the source dataset $\mathcal{X}$ is processed by the operator $\hat{f}(\cdot\,;\psi)$, producing the estimated output $\tilde{\mathcal{Y}}$. A discriminator D_ϕ is trained to distinguish $\tilde{\mathcal{Y}}$ from the distorted dataset $\mathcal{Y}$, allowing the training of $\hat{f}(\cdot\,;\psi)$ through adversarial learning [13].

3.1. Proposed Diffusion-Based Approach

The first approach builds on recent advancements in diffusion models for blind inverse problems in music [30, 22], speech [24, 31], and image [21] restoration. These methods jointly estimate the degradation operator alongside the restored signal at inference time.

Following [24, 31, 21], we formulate the operator estimation problem as an EM objective over an observed dataset $\mathcal{Y} = \{\mathbf{y}^{(n)}\}_{n=1}^{N}$:

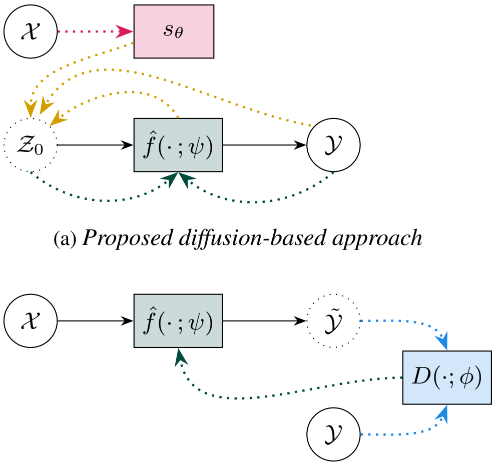
$$\max_{\psi} \mathbb{E}_{\mathcal{Z}_0 \sim p(\mathcal{Z}_0|\mathcal{Y})} \log p(\mathcal{Y}|\mathcal{Z}_0;\psi)\,, \tag{4}$$

where $\mathcal{Z}_0 = \{\mathbf{z}_0^{(n)}\}_{n=1}^{N}$ represents the latent clean source signals corresponding to $\mathcal{Y}$, inferred during optimization.

Instead of the theoretical Dirac likelihood in Eq. (2), we introduce a convex surrogate:

$$p(\mathcal{Y}|\mathcal{Z}_0;\psi) \propto \exp\left(-\zeta \sum_{n=1}^{N} \mathcal{C}(\mathbf{y}^{(n)}, \hat{f}(\mathbf{z}_0^{(n)};\psi))\right), \tag{5}$$

where $\mathcal{C}(\cdot,\cdot)$ denotes a convex cost function that quantifies the discrepancy between the observed and estimated signals, aggregated across the dataset, and ζ is a scaling hyperparameter.



(a) *Proposed diffusion-based approach*

(b) *Adversarial approach [13]*

Figure 1: *High-level diagrams of the two studied unpaired operator estimation methods, (a) the diffusion-based approach and (b) the adversarial approach. Solid lines indicate the signal processing flow, while dashed lines represent optimization dependencies between components.*

The key challenge in optimizing Eq. (4) is evaluating the expectation, which requires drawing samples from the posterior distribution $p(\mathcal{Z}_0|\mathcal{Y})$. Since the dataset $\mathcal{Y}$ consists of N distorted signals $\mathbf{y}^{(n)}$, the algorithm requires estimating the corresponding source input signals $\mathbf{z}_0^{(n)}$ for each $n \in \{1, \ldots, N\}$. For clarity, we omit the index i in the following discussion, but the following procedure must be repeated for each observation in the dataset.

The estimation of each $\mathbf{z}_0$ is performed using approximate posterior sampling with a diffusion model trained exclusively on the clean dataset $\mathcal{X}$. The optimization procedure alternates between two steps, also visualized in Fig. 1(a). In the E-step, given fixed parameters ψ and the trained diffusion model, we estimate the source samples $\mathbf{z}_0$. In the M-step, using the estimated $\mathbf{z}_0$, we update the parameters ψ of the degradation model to maximize the likelihood of the observations.

3.1.1. Diffusion models for posterior sampling

A diffusion model provides a powerful framework for sampling from complex distributions, such as p_x, and for sampling from approximate posterior distributions, which factorize as $p(\mathbf{z}_0|\mathbf{y};\psi) \propto p(\mathbf{y}|\mathbf{z}_0;\psi)p_x(\mathbf{z}_0)$, useful for solving inverse problems [20, 32]. Diffusion models approach the generation problem by breaking it into a sequence of denoising steps. The process starts from a Gaussian prior $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \sigma_T^2\mathbf{I})$ and gradually refines the sample until it follows the target data distribution $\mathbf{z}_0 \sim p_x$. This transformation is governed by the probability flow ordinary differential equation (ODE):

$$\mathrm{d}\mathbf{z}_\tau = -\tau \nabla_{\mathbf{z}_\tau} \log p(\mathbf{z}_\tau) \mathrm{d}\tau, \tag{6}$$

where time τ evolves backward from T to 0. Since the score function $\nabla_{\mathbf{z}_\tau} \log p(\mathbf{z}_\tau)$ is generally intractable, it is approximated using a time-conditional deep neural network $s_\theta(\mathbf{z}_\tau, \tau) \approx \nabla_{\mathbf{z}_\tau} \log p(\mathbf{z}_\tau)$, trained via denoising score matching [33] on the available dataset of clean signals $\mathcal{X}$. We adopt the same diffusion parameterization as in previous works [25, 24] and refer to [25] for a more detailed introduction.

To sample from the posterior, we replace the score function in Eq. (6) with the posterior score [14]:

$$\nabla_{\mathbf{z}_\tau} \log p(\mathbf{z}_\tau|\mathbf{y};\psi) = \nabla_{\mathbf{z}_\tau} \log p(\mathbf{z}_\tau) + \nabla_{\mathbf{z}_\tau} \log p(\mathbf{y}|\mathbf{z}_\tau;\psi). \tag{7}$$

Following Chung et al. [32], we approximate this as

$$\nabla_{\mathbf{z}_\tau} \log p(\mathbf{z}_\tau|\mathbf{y};\psi) \approx s_\theta(\mathbf{z}_\tau, \tau) + \zeta(\tau) \nabla_{\mathbf{z}_\tau} \log p(\mathbf{y}|\hat{\mathbf{z}}_0(\mathbf{z}_\tau, \tau);\psi), \tag{8}$$

where the prior score is replaced by $s_\theta(\mathbf{z}_\tau, \tau)$, and the likelihood score is estimated using the one-step denoised estimate $\hat{\mathbf{z}}_0(\mathbf{z}_\tau) = \mathbb{E}[\mathbf{z}_0|\mathbf{z}_\tau]$, which can be obtained without extra cost via Tweedie's formula:

$$\hat{\mathbf{z}}_0(\mathbf{z}_\tau, \tau) = \tau^2 s_\theta(\mathbf{z}_\tau, \tau) + \mathbf{z}_\tau. \tag{9}$$

3.1.2. E-Step

By discretizing and solving the posterior-replaced ODE introduced in Section 3.1.1, from $\tau = T$ to $\tau = 0$, we could obtain samples from the approximate posterior $p(\mathbf{z}_0|\mathbf{y};\psi)$, from which we could evaluate the expectation in Eq. (4). However, simulating this conditional reverse diffusion process is computationally expensive due to the need for T evaluations of s_θ per EM iteration. To reduce this cost, we integrate EM updates with the ODE discretization, following [21, 24]. Specifically, we discretize the time variable τ into K steps,

$$\{\tau_1, \ldots, \tau_{k-1}, \tau_k, \ldots, \tau_K\}, \quad \text{with} \quad \tau_1 = T \quad \text{and} \quad \tau_K \approx 0.$$

At each step k, we approximate the expectation in Eq. (4) using an intermediate latent variable $\mathbf{z}_{\tau_k}$:

$$\mathbb{E}_{\mathbf{z}_0 \sim p(\mathbf{z}_0|\mathbf{y};\psi)} \log p(\mathbf{y}|\mathbf{z}_0;\psi) \approx \mathbb{E}_{\mathbf{z}_0 \sim p(\mathbf{z}_0|\mathbf{z}_{\tau_k})} \log p(\mathbf{y}|\mathbf{z}_0;\psi). \tag{10}$$

This approximation is motivated by the fact that the reverse diffusion process gradually refines samples toward the posterior mode. Since $\mathbf{z}_{\tau_k}$ already encodes significant information about $\mathbf{z}_0$, it serves as a useful intermediate representation. Rather than explicitly conditioning on $\mathbf{y}$ at every EM iteration, we rely on the structure of the conditional reverse diffusion process to implicitly incorporate the data constraint.

Following [32], we further approximate $p(\mathbf{z}_0|\mathbf{z}_\tau)$ as a Dirac delta located at the one-step denoised estimate: $p(\mathbf{z}_0|\mathbf{z}_\tau) \approx \delta(\hat{\mathbf{z}}_0(\mathbf{z}_\tau, \tau))$. This results in the following one-sample Monte Carlo estimate of the expectation from Eq. (4):

$$\mathbb{E}_{\mathcal{Z}_0 \sim p(\mathcal{Z}_0|\mathcal{Y})} \log p(\mathcal{Y}|\mathcal{Z}_0;\psi) \approx \log p(\mathcal{Y}|\hat{\mathcal{Z}}_0^k;\psi)\,, \tag{11}$$

where $\hat{\mathcal{Z}}_0^k = \{\hat{\mathbf{z}}_0(\mathbf{z}_{\tau_k}^{(n)}, \tau)\}_{n=1}^N$. The latent variables $\mathbf{z}_{\tau_{k+1}}^{(n)}$ are then updated following the process explained in Sec. 3.1.1.

3.1.3. M-Step

In the M-step, we update the operator parameters ψ to maximize the expected log-likelihood. By integrating Eqs. (5) and (11) into Eq. (4), we obtain the M-step objective:

$$\psi \leftarrow \arg\min_{\psi} \sum_{n=1}^{N} \mathcal{C}\left(\mathbf{y}^{(n)}, \hat{f}(\hat{\mathbf{z}}_0(\mathbf{z}_{\tau_k}^{(n)}, \tau);\psi)\right), \tag{12}$$

In practice, this objective is optimized via stochastic gradient descent. Optimization is performed by sampling random batches of pairs $\{\mathbf{y}, \hat{\mathbf{z}}_0\}$ and updating the parameters accordingly.

3.2. Adversarial Approach

The second approach, sketched in Fig. 1(b), is inspired by Generative Adversarial Networks (GANs) [34] and was proposed for guitar amplifier modeling by Wright et al. [13]. It consists of the following optimization objective [29]:

$$\psi = \arg\min_{\psi} \mathcal{D}(p_y, \hat{p}_y^{\psi}), \tag{13}$$

where $\hat{p}_y^{\psi}$ denotes the distribution induced by p_x through $\hat{f}(\cdot\,;\psi)$, defined analogously to Eq. 3 for p_y, and $\mathcal{D}(\cdot,\cdot)$ denotes a distributional distance or divergence. $\mathcal{D}(\cdot,\cdot)$ is designed such that it attains its minimum when $p_y = \hat{p}_y^{\psi}$.

The distributional distance $\mathcal{D}(\cdot,\cdot)$ is parameterized by a discriminator or critic $D(\cdot\,;\phi) : \mathbb{R}^L \to \mathbb{R}$, typically a neural network with parameters ϕ, such that

$$\mathcal{D}(p_y, \hat{p}_y^{\psi}) = \arg\max_{\phi} \Big\{ \mathbb{E}_{\mathbf{y}\sim p_y}\big[-\max(0, 1 - D(\mathbf{y};\phi))\big] + \mathbb{E}_{\mathbf{x}\sim p_x}\big[-\max(0, 1 + D(\hat{f}(\mathbf{x};\psi);\phi))\big] \Big\} \tag{14}$$

which corresponds to the hinge loss objective [35], as implemented in [13]. In practice, the expectations in the objective function are approximated using Monte Carlo estimates. The dataset of clean signals $\mathcal{X}$ provides realizations of p_x, while the dataset of distorted signals $\mathcal{Y}$ serves as realizations of p_y.

This formulation results in a minimax optimization game, where the discriminator $D(\cdot\,;\phi)$ is trained adversarially against the operator $\hat{f}(\cdot\,;\psi)$. The operator aims to transform p_x such that its output distribution aligns with p_y, while the discriminator attempts to distinguish real samples from transformed ones. One common issue in this adversarial framework is the discriminator collapse, where $D(\cdot\,;\phi)$ becomes too strong and provides poor gradient information to $\hat{f}(\cdot\,;\psi)$, or conversely, too weak, failing to guide the operator's learning process.

4. OPERATOR MODELS

The parametric model $\hat{f}(\cdot\,;\psi)$ can take various forms. In this paper, we explore different black-box models parameterized by neural networks, as well as a gray-box approach, specifically a W-H model.

4.1. Black-Box Operator Models

Black-box models are data-driven and do not require explicit knowledge of the underlying system in their design. Deep neural networks, as universal function approximators, are commonly used to approximate the input-output behavior of an effect. In this study, we focus on two specific neural architectures that have proven successful at modeling audio effects: a Gated Convolution Network (GCN) and S4, a state-space model [36].

The GCN consists of a stack of temporally dilated convolutional operations combined with gated activation functions [37]. This architecture was used for unsupervised guitar effect estimation with an adversarial strategy [13], and in this study, we use the same architecture as in [13], consisting of 31k parameters.

S4, an architecture based on state-space models [36], has been applied to nonlinear audio effects like dynamic range compression [38] and virtual analog effects [39], often outperforming other architectures [40]. We use its implementation from ∇Fx [41][1], with a 19k-parameter non-conditional configuration [40].

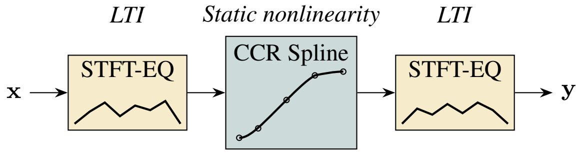


Figure 2: *Structure of the employed Wiener-Hammerstein model.*

4.2. Wiener-Hammerstein Model

One of the most widely used approaches for modeling nonlinear systems is the Wiener-Hammerstein model, which represents the system as a serial connection of a Linear-Time Invariant (LTI) block, followed by a static nonlinearity, and another LTI block [42, 43, 44, 40]. Unlike black-box models, gray-box approaches such as W-H structures offer better interpretability of the learned transfer function, allowing a detailed analysis of each block after the optimization process. In the context of unsupervised optimization, we hypothesize that a more targeted model with a constrained parameter space, such as a W-H model, can be particularly beneficial. The model's inductive biases, derived from prior knowledge of the unknown operator, can potentially help guide the optimization, improving stability and robustness. However, this advantage may come at the cost of reduced expressivity, as the model structure may limit its ability to fit highly complex systems that cannot be predicted with the utilized parametric structure.

The structure employed in this work is shown in Fig. 2. The LTI blocks are designed as equalizers implemented in the frequency domain, similar to the approach in [45]. The magnitude responses are optimized at a reduced grid, spaced according to third-octave bands. These values are then linearly interpolated to cover the entire frequency range. This is implemented by computing the Short-time Fourier Transform (STFT) of the signal and multiplying each column (i.e., each time frame) element-wise with the interpolated frequency response. Additionally, we optimize the phase values of the frequency-domain filter across the entire frequency range, allowing the operator to adapt to unknown phase responses. For the STFT, a Hann window with 2048 samples and 75% overlap is used, along with zero-padding to double the window length, preventing temporal aliasing. The inverse STFT is posteriorly applied to recover the filtered waveform.

The static nonlinearity is parameterized with a Cubic Catmull-Rom (CCR) spline, as proposed in previous work [25], since it outperformed alternative parameterizations such as Multilayer Perceptrons [46] or parametric tanh structures [44] for modeling memoryless nonlinear distortion using a similar diffusion-based framework. The spline is parameterized with 41 control points.

5. GUITAR DISTORTION MODELING EXPERIMENTS

We evaluate the performance of the diffusion-based and adversarial methods in unsupervised guitar distortion modeling. To do so, we use a similar experimental framework as [13]. We investigate the effect of reducing the amount of available effected data in the

[1] https://github.com/mcomunita/nablafx

results, reducing the size of $\mathcal{Y}$, and how the two methods perform when different operator models are used. Audio examples are available on the webpage[2].

5.1. Data

Following [13], we utilize the fourth subset of the IDMT-SMT Guitar dataset [47], which comprises 64 short musical excerpts spanning various styles and tempos. While the dataset includes recordings from two different guitars, we exclusively use those from the Career SG guitar. Each recording is processed with three distinct distortion effects of increasing severity: 'Clean Distortion,' 'Light Distortion,' and 'Heavy Distortion.' The distortion effects were applied using commercial audio plugins, consistent with the procedure in [13]. All the material is sampled at 44.1 kHz.

Our experiments require the following distinct, non-overlapping dataset splits: two unpaired datasets representing the input and output distributions, denoted as $\mathcal{X}$ and $\mathcal{Y}$, respectively, and a paired test set ($\mathcal{X}_{\text{test}}$, $\mathcal{Y}_{\text{test}}$) for evaluation. Of the 37 min of available recordings, we allocate 16 min for $\mathcal{X}$, and 16 min for $\mathcal{Y}$. Additionally, we construct three reduced versions of $\mathcal{Y}$ containing 18 s, 1 min, and 4 min of distorted recordings. The remaining 5 min are used for the test set, which is segmented into chunks of 6 s, resulting in 50 segments. Our dataset split differs from the one used in [13], as this configuration enables a more systematic evaluation of both approaches. Consequently, our results may not be directly comparable to those reported in [13].

5.2. Experimental Details: Diffusion-Based Approach

Our diffusion-based method closely follows the configuration of previous work [25], using the same hyperparameter choices unless otherwise specified. The diffusion model operates in the waveform domain, while its score network is built on an architecture based on the Constant-Q Transform (CQT), as proposed in [22]. This design leverages the invertibility and differentiability of the CQT to introduce inductive biases tailored to musical signals, while retaining the flexibility of waveform-domain modeling. The model is trained on 6-second audio segments. The model was pre-trained for 80k iterations using the EGDB dataset [48], followed by 14k iterations of training on the dataset split $\mathcal{X}$, with a batch size of 4. The training phase (excluding pre-training) took 70 min[3]. While the impact of pre-training remains to be formally evaluated, preliminary experiments suggest that it may have only a minimal effect.

We use a consistent hyperparameter configuration for all the operator optimization experiments, setting $T = 101$ steps, which corresponds to both the reverse diffusion discretization and the number of EM iterations. The likelihood scaling parameter ζ in Eq. (5) is defined following [25] as a function of τ controlled by $\tilde{\zeta} = 0.2$. As the cost function $C(\cdot, \cdot)$, we employ a ℓ_2-norm in a compressed STFT representation [23, 24], using a compression factor of 0.5. This compression equalizes the spectral energy distribution and enhances high-frequency content, which typically has lower energy but is perceptually important. Each M-step includes 20 operator optimization steps. We employ the AdamW optimizer with random batches of 4 examples per iteration, a learning rate of 0.001, momentum parameters ($\beta_1 = 0.9$, $\beta_2 = 0.99$), and a weight decay of 0.01.

Operator optimization times varied based on dataset $\mathcal{Y}$ size: approximately 1 h for 16 min of data, 18 min for 4 min of data, 6 min for 1 min, and 3 min for 18 s. Exact times may vary depending on operator efficiency and implementation-specific details such as logging. Certain operations during the E-step and latent variable updates, particularly the forward and backward score model evaluations, demand substantial memory if processed naively in a single evaluation. To mitigate memory overhead, we process all dataset elements in small batches of 1, 2, or 4, depending on available GPU memory.

5.3. Experimental Details: Adversarial Approach

Following [13], we adopt the set of three log-mel spectrogram discriminators, each operating on 160 mel bands but with varying window sizes of 512, 1024, and 2048 samples. This configuration was chosen as it demonstrated superior performance in the experiments reported in that study. All models were trained for exactly 100k iterations with batches of 5 segments of 1.5 s each, requiring approximately 2 h. We follow the remaining hyperparameter settings from [13]. Experiments using the W-H model proved unstable with this approach, suffering from severe discriminator collapse. Consequently, they were excluded from the evaluation, and only the black-box models, GCN and S4, were studied.

We acknowledge the extensive body of work aimed at improving the stability of adversarial training [49, 50], and recognize that some of these techniques could be relevant to our setting, particularly in imbalanced scenarios. However, exploring and properly implementing these approaches is considered beyond the scope of the present work. Therefore, we limit ourselves to the configuration proposed in [13].

5.4. Experimental Details: Supervised Baseline

In addition to the unsupervised methods that are the focus of this study, we incorporate a supervised baseline to serve as an upper performance bound for the unsupervised approaches. This baseline is trained using the dataset $\mathcal{Y}$, with reduced size when applicable, along with the paired input signal, which is unavailable in the unsupervised setting. The supervised operator models were trained for 5k iterations using the Adam optimizer with a learning rate of 0.0001 and 6-s-long audio segments.

5.5. Evaluation

To systematically evaluate the performance of the different methods, we apply each optimized operator to all instances in the test set $\mathbf{x}_{\text{test}} \in \mathcal{X}_{\text{test}}$, obtaining the corresponding signal estimates: $\hat{\mathbf{y}} = \hat{f}(\mathbf{x}_{\text{test}}; \psi)$. We then assess the quality of these estimates by comparing them to the paired ground-truth targets $\mathbf{y} \in \mathcal{Y}_{\text{test}}$ using three objective evaluation metrics.

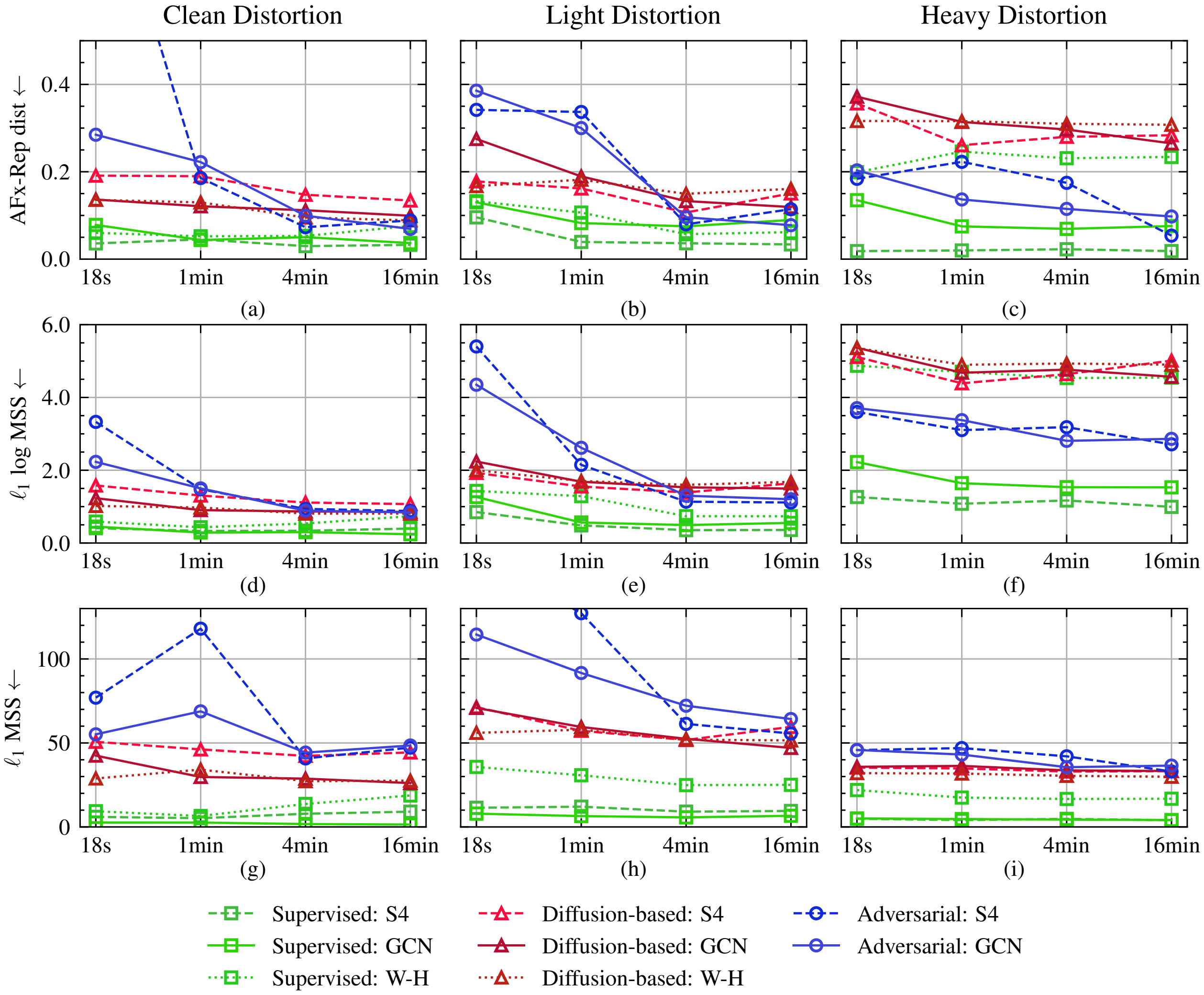
The first metric is based on the AFx-Rep representation [10], which was specifically designed to capture information related to audio production style. To quantify the similarity between the reference signal $\mathbf{y}$ and its estimate $\hat{\mathbf{y}}$, we compute the cosine distance between their respective embeddings:

$$\text{dist}(\hat{\mathbf{y}}, \mathbf{y}) = 1 - \frac{g(\hat{\mathbf{y}}) \cdot g(\mathbf{y})}{\max(\|g(\hat{\mathbf{y}})\| \|g(\mathbf{y})\|, \epsilon)}, \tag{15}$$

The operator g is the AFx-Rep trained encoder, $\cdot$ denotes the dot product, the norm used is the ℓ_2 norm, and ϵ is a small constant for numerical stability.

[2] https://michalsvento.github.io/UnNAFx/

[3] All computation times here were measured on an NVIDIA H200 GPU.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3: Objective metrics AFx-Rep, $\ell_1 \log \mathrm{MSS}$, and $\ell_1\mathrm{MSS}$ calculated on the test set, based on the amount of observed effected data.

Next, we compute the ℓ_1 distance between Multi-Scale Spectrograms (MSS), which measures the difference between STFT magnitudes across multiple analysis settings. The STFT is computed using predefined window lengths $W = \{2048, 1024, 512, 256, 128, 64\}$. For each $w \in W$, the number of FFT bins is set to w, and the hop size is chosen as $w/4$. The ℓ_1 MSS metric is then defined as:

$$\ell_1\mathrm{MSS}(\hat{\mathbf{y}}, \mathbf{y}) = \sum_{w \in W} \| \, |\mathrm{STFT}_w(\hat{\mathbf{y}})| - |\mathrm{STFT}_w(\mathbf{y})| \, \|_1, \quad (16)$$

where $\mathrm{STFT}_w(\cdot)$ is the STFT operator with an analysis window of size w. To obtain values on a more interpretable scale, we normalize the ℓ_1-norm by taking its mean.

As the third metric, we define a logarithmic variant of the previously mentioned metric, $\ell_1 \log \mathrm{MSS}$, where $\log_{10}$ is applied to the STFT magnitudes before computing the distance.

5.6. Results

The results of the objective metrics are shown in Fig. 3. As expected, the supervised experiments (green markers) consistently achieve the best performance across all scenarios, establishing a lower bound on the error that can be expected from the unsupervised methods.

When evaluating the robustness of unsupervised methods in data-scarce scenarios, it can be observed that the adversarial method (blue markers in Fig. 3) exhibits a notable drop in performance as the duration of available data decreases. In contrast, the diffusion-based approach (red markers in Fig. 3) maintains a more stable performance, even with as little as 18 s of data, although a slight decline is still noticeable.

The experiments with the diffusion-based method outperform the adversarial ones in the 'Clean' and 'Light' distortion scenarios when only 18 s or 1 min of effected recordings are available, and achieves a comparable performance when 4 or 16 min are used. In terms of AFx-Rep dist. (Fig. 3a,b) and $\ell_1 \log \mathrm{MSS}$ (Fig. 3d,e), the adversarial approach yields slightly better results when the full dataset is available. Conversely, the diffusion-based method achieves consistently lower errors on the linear $\ell_1\mathrm{MSS}$ metric (Fig. 3g,h). These differences are potentially attributed to the alignment between each method's training objective and the evaluation metrics. The adversarial model was trained using logscaled mel-spectrogram features, which are more closely related to the $\ell_1 \log \mathrm{MSS}$ metric. In contrast, the diffusion-based model was optimized using magnitude-compressed STFT features, which may be more aligned with linear-scale ℓ_1 MSS, though the correspondence is not exact.

These trends differ in the 'Heavy' distortion setting, where the adversarial method experiments consistently obtained better results than diffusion-based ones in terms of AFx-Rep dist (Fig. 3(c)) and ℓ_1 log MSS (Fig. 3(f)). Interestingly, this is not the case in terms of ℓ_1 MSS, where the diffusion-based approach still obtains lower values. Also in this case, for both unsupervised approaches, increased data availability almost consistently leads to improved operator estimation.

In the Heavy distortion setting, the trend shifts: the adversarial method consistently outperforms the diffusion-based approach in AFx-Rep distance (Fig. 3c) and $\ell_1 \log \mathrm{MSS}$ (Fig. 3f). The diffusion-based method still performs better in terms of linear ℓ_1 *MSS*. Also in this case, for both unsupervised approaches, performance generally improves with increased data availability.

We do not observe a clear trend regarding the most suitable black-box model architecture: both GCN and S4 yield comparable performance when used with either of the unsupervised methods. While one may outperform the other in specific cases, the difference is not consistent across settings. Within the diffusion-based framework, a comparison between black-box and grey-box operators shows that the W-H model performs on par with both the GCN and S4 models (see Fig. 3), and slightly better in some cases. Attempts to train the W-H model using the adversarial approach were unsuccessful, and the corresponding results are therefore not included in the figure. These findings suggest that diffusion-based methods offer stable performance across different operator architectures, whereas the adversarial approach appears more sensitive to the choice of model.

6. CONCLUSIONS

This study addressed unsupervised operator estimation, with experiments on guitar distortion modeling, comparing diffusion-based and adversarial approaches. While adversarial methods performed well under heavy distortion, they lack consistency across scenarios. In contrast, diffusion-based methods show strong robustness to data scarcity and operator choice, making them a more reliable option overall. Additionally, diffusion methods offer the benefit of jointly estimating the clean guitar signal, although reconstruction quality has not been explored in this work.

Both approaches—diffusion-based and adversarial—require substantial training time and computational resources, which can limit their accessibility for applications that require training with low-power or time constraints. A key limitation of the diffusion-based method is the need for a separately pre-trained model on clean guitar signals, and obtaining such dry data can be challenging depending on the context. Future work should investigate how performance varies with different amounts of available dry data and assess the trade-offs between robustness and computational efficiency. Although all evaluated models are technically capable of real-time operation [1], the practical computational demands remain an open area for further study. Finally, although this study focused specifically on distortion effects, we believe the proposed framework could extend to other nonlinear audio effects, such as dynamic range compression or modulation effects, given the generality of the black-box operators used in training.

7. ACKNOWLEDGMENTS

Part of this work was conducted during E. Moliner's research visit to Brno University of Technology in March 2025. M. Švento completed a four-month Erasmus+ traineeship at the Aalto Acoustics Lab in fall 2024. The authors acknowledge the computational resources provided by the Aalto Science-IT project. The work carried out at the Brno University of Technology was supported by the Czech Science Foundation (GAČR) Project No. 23-07294S.

8. REFERENCES

[1] A. Wright, E.-P. Damskägg, L. Juvela, and V. Välimäki, "Real-time guitar amplifier emulation with deep learning," *Appl. Sci.*, vol. 10, no. 3, pp. 766, 2020.

[2] M. A. Martínez Ramírez, E. Benetos, and J. Reiss, "Deep learning for black-box modeling of audio effects," *Appl. Sci.*, vol. 10, no. 2, pp. 638, 2020.

[3] C. J. Steinmetz, N. J. Bryan, and J. D. Reiss, "Style transfer of audio effects with differentiable signal processing," *J. Audio Eng. Soc.*, vol. 70, no. 9, 2022.

[4] C. Peladeau and G. Peeters, "Blind estimation of audio effects using an auto-encoder approach and differentiable digital signal processing," in *Proc. IEEE ICASSP*, 2024.

[5] S. Lee, J. Park, S. Paik, and K. Lee, "Blind estimation of audio processing graph," in *Proc. IEEE ICASSP*, 2023.

[6] O. Take, T. Cheng, T. Nakano, et al., "Audio effect chain estimation and dry signal recovery from multi-effect-processed musical signals," in *Proc. Int. Conf. DAFx*, 2024.

[7] M. Rice, Ch. J. Steinmetz, G. Fazekas, and J. Reiss, "General purpose audio effect removal," in *Proc. IEEE WASPAA*, 2023.

[8] Y.-H. Chen, Y.-T. Yeh, Y.-C. Cheng, et al., "Towards zero-shot amplifier modeling: One-to-many amplifier modeling via tone embedding control," *arXiv preprint arXiv:2407.10646*, 2024.

[9] A. Wright, A. Carson, and L. Juvela, "Open-amp: Synthetic data framework for audio effect foundation models," *Proc. IEEE ICASSP*, 2025.

[10] C. J. Steinmetz, S. Singh, M. Comunità, et al., "ST-ITO: Controlling audio effects for style transfer with inference-time optimization," in *Proc. ISMIR*, 2024.

[11] J. Koo, S. Uhlich, K. Lee, and Y. Mitsufuji, "Music mixing style transfer: A contrastive learning approach to disentangle audio effects," in *Proc. IEEE ICASSP*, 2023.

[12] R. Hinrichs, K. Gerkens, A. Lange, and J. Ostermann, "Blind extraction of guitar effects through blind system inversion and neural guitar effect modeling," *EURASIP J. Audio Speech Music Process.*, vol. 2024, no. 1, pp. 9, Feb. 2024.

[13] A. Wright, V. Välimäki, and L. Juvela, "Adversarial guitar amplifier modelling with unpaired data," in *Proc. IEEE ICASSP*, 2023.

[14] Y. Song, J. Sohl-Dickstein, D. P Kingma, et al., "Score-based generative modeling through stochastic differential equations," in *Proc. ICLR*, 2021.

[15] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Proc. Advances in NeurIPS*, 2022.

[16] H. Liu, X. Mei, X. Liu, et al., "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proc. 40th Int. Conf. Machine Learning*, 2023.

[17] Z. Evans, J. D. Parker, CJ Carr, et al., "Stable audio open," in *Proc. IEEE ICASSP*, 2025.

[18] E. Moliner, F. Elvander, and V. Välimäki, "Blind audio bandwidth extension: A diffusion-based zero-shot approach," *IEEE/ACM Trans. Audio, Speech, Lang. Proc.*, vol. 32, Dec. 2024.

[19] J.-M. Lemercier, J. Richter, S. Welker, et al., "Diffusion models for audio restoration: A review," *IEEE Signal Processing Mag.*, vol. 41, no. 6, 2024.

[20] E. Moliner, J. Lehtinen, and V. Välimäki, "Solving audio inverse problems with a diffusion model," in *Proc. IEEE ICASSP*, 2023.

[21] C. Laroche, A. Almansa, and E Coupete, "Fast diffusion EM: a diffusion model for blind inverse problems with application to deconvolution," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[22] E. Moliner, M. Turunen, F. Elvander, and V. Välimäki, "A diffusion-based generative equalizer for music restoration," in *Proc. Int. Conf. DAFx*, 2024.

[23] E. Moliner, J.-M. Lemercier, S. Welker, et al., "BUDDy: Single-channel blind unsupervised dereverberation with diffusion models," in *Proc. IWAENC*, Aalborg, Denmark, 2024.

[24] J.-M. Lemercier, E. Moliner, S. Welker, et al., "Unsupervised blind joint dereverberation and room acoustics estimation with diffusion models," *IEEE/ACM Trans. Audio, Speech, Lang. Proc.*, June 2025.

[25] M. Švento, E. Moliner, L. Juvela, et al., "Estimation and restoration of unknown nonlinear distortion using diffusion," *accepted for publication in J. Audio. Eng. Soc.*, 2025.

[26] Y.-H. Chen, W. Choi, W.-H. Liao, et al., "Improving unsupervised clean-to-rendered guitar tone transformation using GANs and integrated unaligned clean data," in *Proc. Int. Conf. DAFx*, 2024.

[27] J. Park, J. W. Lee, D. Lee, and K. Lee, "Solving blind non-linear forward and inverse problem for audio applications," 2025, `https://openreview.net/forum?id=mlPTNEIsgb`.

[28] H. Thanh-Tung and T. Tran, "Catastrophic forgetting and mode collapse in GANs," in *Proc. IJCNN*, 2020.

[29] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*, MIT Press, 2023.

[30] E. Moliner, F. Elvander, and V. Välimäki, "Blind audio bandwidth extension: A diffusion-based zero-shot approach," *IEEE/ACM Trans. Audio, Speech, Lang. Proc.*, 2024.

[31] B. Nortier, M. Sadeghi, and R. Serizel, "Unsupervised speech enhancement with diffusion-based generative models," in *Proc. IEEE ICASSP*, 2024.

[32] H. Chung, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," *Proc. ICLR*, 2023.

[33] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, 2011.

[34] I. J. Goodfellow, B. Xu, D. Warde-Farley, et al., "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[35] J. H. Lim and J. C. Ye, "Geometric GAN," *arXiv preprint arXiv:1705.02894*, 2017.

[36] A. Gu, K. Goel, and C. Re, "Efficiently modeling long sequences with structured state spaces," in *Proc. ICLR*, 2022.

[37] E.-P. Damskägg, L. Juvela, E. Thuillier, and V. Välimäki, "Deep learning for tube amplifier emulation," in *Proc. IEEE ICASSP*, 2019.

[38] H. Yin, G. Cheng, C. J. Steinmetz, et al., "Modeling analog dynamic range compressors using deep learning and state-space models," *arXiv preprint arXiv:2403.16331*, 2024.

[39] R. Simionato and S. Fasciani, "Comparative study of state-based neural networks for virtual analog audio effects modeling," *arXiv preprint arXiv:2405.04124*, 2024.

[40] M. Comunità, C. J. Steinmetz, and J. D. Reiss, "Differentiable black-box and gray-box modeling of nonlinear audio effects," *arXiv preprint arXiv:2502.14405*, 2025.

[41] M. Comunità, C. J. Steinmetz, and J. D. Reiss, "NablAFx: A framework for differentiable black-box and gray-box modeling of audio effects," *arXiv preprint arXiv:2502.11668*, 2025.

[42] F. Eichas and U. Zölzer, "Gray-box modeling of guitar amplifiers," *J. Audio Eng. Soc.*, vol. 66, no. 12, 2018.

[43] S. Nercessian, A. Sarroff, and K. J. Werner, "Lightweight and interpretable neural modeling of an audio distortion effect using hyperconditioned differentiable biquads," in *Proc. IEEE ICASSP*, 2021.

[44] J. T. Colonel, M. Comunità, and J. Reiss, "Reverse engineering memoryless distortion effects with differentiable waveshaper," in *Proc. AES Conv. 153*, 2022.

[45] J. T. Colonel and J. Reiss, "Reverse engineering of a recording mix with differentiable digital signal processing," *J. Acoust. Soc. Amer.*, vol. 150, no. 1, 2021.

[46] B. Kuznetsov, J. Parker, and F. Esqueda, "Differentiable IIR filters for machine learning applications," in *Proc. Int. Conf. DAFx*, 2020.

[47] C. Kehling, J. Abeßer, C. Dittmar, and G. Schuller, "Automatic tablature transcription of electric guitar recordings by estimation of score- and instrument-related parameters," in *Proc. Int. Conf. DAFx*, 2014.

[48] Y.-H. Chen, J.-S. R. Jang, and Y.-H. Yang, "Towards automatic transcription of polyphonic electric guitar music: A new dataset and a multi-loss transformer model," in *Proc. IEEE ICASSP*, 2022.

[49] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. ICLR*, 2018.

[50] I. Gulrajani, F. Ahmed, M. Arjovsky, et al., "Improved training of wasserstein GANs," in *Proc. Advances in NeurIPS*, 2017, vol. 30.