Švento, Michal; Moliner Juanpere, Eloi; Juvela, Lauri; Wright, Alec; Välimäki, Vesa

Estimation and Restoration of Unknown Nonlinear Distortion Using Diffusion

Published in:
AES: Journal of the Audio Engineering Society

DOI:
10.17743/jaes.2022.0221

Published: 01/09/2025

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Švento, M., Moliner Juanpere, E., Juvela, L., Wright, A., & Välimäki, V. (2025). Estimation and Restoration of Unknown Nonlinear Distortion Using Diffusion. *AES: Journal of the Audio Engineering Society*, *73*(9), 519-532. https://doi.org/10.17743/jaes.2022.0221

OPEN ACCESS Freely available online

M. Švento, E. Moliner, L. Juvela, A. Wright, and V. Välimäki,
"Estimation and Restoration of Unknown Nonlinear Distortion Using Diffusion,"
J. Audio Eng. Soc., vol. 73, no. 9, pp. 519–532 (2025 Sep.).
https://doi.org/10.17743/jaes.2022.0221.

Estimation and Restoration of Unknown Nonlinear Distortion Using Diffusion

MICHAL ŠVENTO,[1,*,2] *AES Student Member*, **ELOI MOLINER,**[1] *AES Student Member*, **LAURI JUVELA,**[1]
(michal.svento@vut.cz) (eloi.moliner@aalto.fi) (lauri.juvela@aalto.fi)

ALEC WRIGHT,[3] *AES Student Member* **AND VESA VÄLIMÄKI,**[1] *AES Fellow*
(alec.wright@ed.ac.uk) (vesa.valimaki@aalto.fi)

[1]*Acoustics Laboratory, Department of Information and Communications Engineering, Aalto University, Espoo, Finland*
[2]*Department of Telecommunications, Faculty of Electrical Engineering and Communication, Brno University of Technology, Brno, Czech Republic*
[3]*Acoustics and Audio Group, University of Edinburgh, Edinburgh, UK*

The restoration of nonlinearly distorted audio signals, alongside the identification of the applied memoryless nonlinear operation, is studied. The paper focuses on the difficult but practically important case in which both the nonlinearity and the original input signal are unknown. The proposed method uses a generative diffusion model trained unconditionally on guitar or speech signals to jointly model and invert the nonlinear system at inference time. Both the memoryless nonlinear function model and the restored audio signal are obtained as output. Examples of successful blind estimation of hard and soft-clipping, digital quantization, half-wave rectification, and wavefolding nonlinearities are presented. The results suggest that, out of the nonlinear functions tested here, the Cubic Catmull-Rom spline is best suited to approximating these nonlinearities. In the case of guitar recordings, comparisons with informed and supervised restoration methods show that the proposed blind method is at least as good as they are in terms of objective metrics. Experiments on distorted speech show that the proposed blind method outperforms general-purpose speech enhancement techniques and restores the original voice quality. The proposed method can be applied to memoryless audio effects modeling, restoration of music and speech recordings, and characterization of analog recording media.

0 INTRODUCTION

Nonlinear distortions are a common artifact in audio signals, arising in various scenarios such as hardware limitations, intentional effects in music production, or environmental interactions during signal capture. The restoration of such distorted signals, alongside the estimation of the nonlinear function that caused the distortion, is a challenging yet crucial task and the topic of this study.

A well-known example is clipping, a phenomenon that occurs when the amplitude of the audio signal exceeds the maximum limit, causing saturation [1]. In the case of hard clipping, the nonlinearity is easy to characterize by analyzing the distorted measurement, and the known operator can be used to solve the declipping problem [1–3]. However, there is an infinite range of other nonlinear distortions that can occur, and estimating the operator in those cases is often impossible without access to a clean reference signal. The identification and modeling of nonlinear systems can be useful not only from a restoration perspective but also for tasks such as audio effects modeling [4] and characterizing or modeling analog recording media [5, 6].

Traditionally, characterizing nonlinear systems using data-driven methods necessitated pairs of aligned input and output audio recordings. This includes Wiener-Hammerstein models [4], Volterra series [7–10], Chebyshev models [11], neural networks [12–14], or other specialized signal processing pipelines [15]. However, in many practical cases, such paired data is often unavailable. The challenge of estimating the nonlinearity based solely on the observed distorted output without paired input data is called a blind estimation problem.

Recent studies have addressed this or similar tasks using different approaches. For example, Wright et al. proposed using adversarial training to model guitar distortion

*To whom correspondence should be addressed: michal.svento@vut.cz.

effects with unpaired data [16]. Diffusion bridges have also been explored for blind effects transformation [17]. Alternatively, when a sufficiently large pretraining database is available, it has been shown recently that feature representations capturing relevant audio effect information can be learned [18–20]. These representations have proven useful for modeling unseen effects, although their success depends heavily on the quality and diversity of the pretraining data.

Various strategies for distortion or audio effects inversion, or restoration, have been explored, ranging from sparsity-based approaches, such as A-SPADE for declipping [21], to compensating nonlinear function estimation for soft declipping [22]. Recently, neural network–based methods have become the focus, including supervised learning for guitar distortion removal and speech declipping [23, 24], vocoder-based methods for distortion removal and speech restoration [25, 26], and compositional approaches for the inversion of audio effects [27].

A strategy combining estimation and restoration is the approach by Hinrichs et al. [28]. Their method determines the nonlinear system by using a blind estimate of the clean signal obtained via a supervised restoration model rather than the clean input itself. However, the method heavily relies on the accuracy of the restoration model. Although effective in some cases, it has limitations that can hinder system identification. A major limitation is the need for paired input-output data during training. Although these pairs can be synthetically generated through randomized distortions, this can lead to generalization issues, especially with unseen distortions. Furthermore, due to the supervised framework, the model often struggles with ill-conditioned distortions, causing oversmoothing or regression to the mean. Errors in the restoration process can also propagate, degrading performance in the system identification stage.

A promising alternative to address these limitations is using generative models that jointly perform restoration and distortion estimation. Recent research into diffusion models for inverse problems shows promise in solving both linear and nonlinear problems when the distortion operator is known [29–31]. Notably, the framework used in these works, popularly known as diffusion posterior sampling (DPS) [32], only requires training a diffusion model on clean audio files. Previous works have adapted this framework to blind linear inverse problems, where a linear filter or room impulse response is iteratively estimated alongside the restored audio or speech signal [33, 34]. Although these prior studies have addressed similar challenges, none has effectively tackled blind nonlinear distortion estimation in the context of audio.

This paper focuses on the case of memoryless nonlinear distortions, which include common effects such as clipping, rectification, or quantization. This provides an ideal experimental setup for evaluating this method because these distortions are challenging yet relatively easy to model compared with more complex audio effects. The goal in this study is to investigate whether the DPS, previously successful for linear cases, can be applied for nonlinear distortion identification and restoration in the audio domain. To explore this, a few different models for parameterizing the memoryless nonlinearity are tested. In this study, it is demonstrated that the proposed method accurately estimates a variety of memoryless nonlinear distortions. Additionally, the distortion restoration performance of the proposed method is evaluated on commonly studied tasks, specifically hard and soft declipping, using both guitar and speech data.

The rest of this paper is organized as follows. SEC. 1 discusses the diffusion model formalism and its use in solving inverse problems in audio. SEC. 2 describes nonlinear function models used in this work, two of which are inherited from previous studies, and the third, the Cubic Catmull-Rom (CCR) spline, is tested here for this application for the first time. SEC. 3 presents experimental results and reports the outcome of listening tests. SEC. 4 summarizes the research and suggests ideas for future work.

1 SOLVING BLIND INVERSE PROBLEMS WITH DIFFUSION

The following inverse problem is considered:

$$y(n) = f(x(n)), \quad n \in [0, L-1], \tag{1}$$

where $y(n)$ represents the observed distorted measurement, $x(n)$ is an unknown clean audio signal, $n \in \mathbb{Z}$ is the discrete sample index, L is the length of the signal in samples, and $f(\cdot) : \mathbb{R} \to \mathbb{R}$ is an unknown forward operator. For convenience in subsequent sections, vector notation is introduced, where the signals are represented as $\mathbf{y} = [y(0), \ldots, y(L-1)] \in \mathbb{R}^L$ and $\mathbf{x}_0 = [x(0), \ldots, x(L-1)] \in \mathbb{R}^L$. The goal is to estimate both the clean audio signal, denoted as $\hat{\mathbf{x}}_0 \approx \mathbf{x}_0$, and the forward operator $f(\cdot)$, given only the distorted measurements $\mathbf{y}$.

Here, $f(\cdot)$ represents a memoryless nonlinear transfer function, meaning its output at any given time index n depends only on the input at that instant without influence from past or future inputs. it is further assumed that the distortion introduced by $f(\cdot)$ is time invariant over intervals of a few seconds, such that its characteristics remain constant within these segments. Under this assumption, both $f(\cdot)$ and $\mathbf{x}_0$ are estimated using observations from these intervals.

To address the ill-posed nature of this problem, it is assumed that $\mathbf{x}_0$ is drawn from an underlying data distribution $\mathbf{x}_0 \sim p_{\text{data}}$. Additionally, the authors have access to a training data set containing realizations from this distribution, where all realizations are normalized to have equal power. Using this data set, a diffusion model is trained, a type of generative model, that serves as a data-driven prior of the clean signal distribution. Diffusion models are particularly well-suited for capturing the underlying distribution of complex data, such as audio signals [29, 35, 36], and for generating high-quality samples consistent with this distribution.

Initially, this section offers a summary of diffusion models followed by details on applying them to audio restoration. The algorithm utilizing diffusion model priors to resolve blind inverse problems is further described.

1.1 Diffusion Model Formalism

A diffusion model iteratively transforms Gaussian noise samples $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \sigma^2(T)\mathbf{I})$ to samples drawn from the training data distribution $\mathbf{x}_0 \sim p_{\text{data}}$. The generative process is defined in a time interval $\tau \in [0, T]$, with the audio signal at each intermediate state denoted as $\mathbf{x}_\tau \sim \mathcal{N}(\mathbf{x}_0, \sigma^2(\tau)\mathbf{I})$.

The generative process can be identified by means of a probability flow ordinary differential equation (PF-ODE):

$$\mathrm{d}\mathbf{x}_\tau = \left[\mathbf{f}(\mathbf{x}_\tau, \tau) - \frac{1}{2} g^2(\tau) \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau)\right] \mathrm{d}\tau. \tag{2}$$

By convention, the generative process is formulated to evolve from $\tau = T$ to $\tau = 0$, meaning that τ decreases over time. The term $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$ is the score function, a vector field that points to areas of higher data likelihood. The drift $\mathbf{f}(\mathbf{x}_\tau, \tau)$ and diffusion $g(\tau)$ are design parameters. The authors adopt the choices from Karras et al. [37], which use $\mathbf{f}(\mathbf{x}_\tau, \tau) = 0$, $g(\tau) = \sqrt{2\tau}$, and $\sigma(\tau) = \tau$. That yields the simplified ODE

$$\mathrm{d}\mathbf{x}_\tau = -\tau \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau) \mathrm{d}\tau. \tag{3}$$

The score function $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$ is generally intractable and must be approximated in practice. A key result states that under Gaussian noise, the score can be expressed as [38]

$$\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau) = (\mathbb{E}[\mathbf{x}_0|\mathbf{x}_\tau] - \mathbf{x}_\tau)/\sigma^2(\tau), \tag{4}$$

where $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_\tau]$ represents the posterior expectation of the terminal state, or clean signal, $\mathbf{x}_0$ given the noisy $\mathbf{x}_\tau$. This expectation serves as a one-step denoised estimate of $\mathbf{x}_0$.

In practice, this posterior expectation is approximated using a deep neural network $D_\theta(\mathbf{x}_\tau, \tau)$ trained to predict $\mathbf{x}_0$ from $\mathbf{x}_\tau$. The model parameters θ are optimized via an L^2-based denoising objective

$$\mathbb{E}_{\mathbf{x}_0 \sim p_{\text{data}}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\lambda(\tau) \| D_\theta(\mathbf{x}_0 + \tau\boldsymbol{\epsilon}, \tau) - \mathbf{x}_0 \|_2^2\right], \tag{5}$$

where the expectation is taken over a data set of clean examples $\{\mathbf{x}_{0_i} \sim p_{\text{data}}\}_{i=0}^{N_{\text{data}}}$, and $\lambda(\tau)$ is a time-varying weighting function optimized to ensure the initial loss approximates 1 for all τ, as detailed by Karras et al. [37]. For brevity, the notation $D_\theta(\mathbf{x}_\tau, \tau) = \hat{\mathbf{x}}_0(\mathbf{x}_\tau) \approx \mathbb{E}[\mathbf{x}_0|\mathbf{x}_\tau]$ is adopted to refer to the one-step estimate $\mathbf{x}_0$ from which the approximate score can be derived using Eq. (4).

1.2 Diffusion Models for Inverse Problems

By framing the inverse problem introduced within a probabilistic context, the objective becomes approximating the posterior distribution $p(\mathbf{x}_0|\mathbf{y})$, given the observations $\mathbf{y}$. Solving the ODE in Eq. (3) from $\tau = T$ to $\tau = 0$ allows for sampling from the learned unconditional distribution. However, in the context of inverse problems, the goal shifts to sampling from the posterior distribution $p(\mathbf{x}_0|\mathbf{y})$.

This can be achieved by replacing the score in Eq. (3) by the score of the posterior $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau|\mathbf{y})$. Applying Bayes' theorem, this term factorizes as

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau|\mathbf{y}) = \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau) + \nabla_{\mathbf{x}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau). \tag{6}$$

The first term, or *prior score*, is approximated with the trained denoiser $D_\theta(\mathbf{x}_\tau, \tau)$ via Eq. (4). The second term, or *likelihood score*, is intractable in closed form [32] and must also be approximated.

Following the DPS framework [32] and recent extensions [34, 39], the likelihood score is approximated as

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau) \approx -\zeta(\tau) \nabla_{\mathbf{x}_\tau} C(\mathbf{y}, f(\hat{\mathbf{x}}_0(\mathbf{x}_\tau))), \tag{7}$$

where $C(\cdot, \cdot) : \mathbb{R}^L \times \mathbb{R}^L \to \mathbb{R}$ is a cost function defined as

$$C(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} |S_{\text{comp}}(\mathbf{y})_{m,k} - S_{\text{comp}}(\hat{\mathbf{y}})_{m,k}|^2, \tag{8}$$

and $S_{\text{comp}}(\mathbf{y}) = |\text{STFT}(\mathbf{y})|^{2/3} \exp j\angle \text{STFT}(\mathbf{y})$ is the short-time Fourier transform (STFT) spectrogram with compressed magnitudes, M is the number of STFT frames, and K is the number of frequency bins. The weighting parameter $\zeta(\tau)$ is a time-dependent scaling factor, following [29]:

$$\zeta(\tau) = \frac{\sqrt{L}\,\tilde{\zeta}}{\sigma(\tau) \| \nabla_{\mathbf{x}_\tau} C(\mathbf{y}, f(\hat{\mathbf{x}}_0(\mathbf{x}_\tau)))\|_2}, \tag{9}$$

where $\tilde{\zeta}$ is a fixed hyperparameter, typically chosen between 0 and 1. A higher value of $\tilde{\zeta}$ places greater emphasis on enforcing consistency with the measurements $\mathbf{y}$, ensuring that the reconstructed signal aligns closely with the observed data. However, this may come at the cost of reduced perceptual audio quality because excessive weighting can overly constrain the generative process.

1.3 Joint Estimation and Restoration Algorithm

The framework for solving inverse problems introduced in SEC. 1.2 typically requires knowledge of the operator $f(\cdot)$. However, in the problem setting here, this operator is unknown. To address this, similar to [33, 34], a parametric function $\hat{f}(\cdot; \psi)$, where ψ is a set of optimizable parameters. Practical examples of the studied operators are presented in SEC. 2.

The authors jointly estimate the clean audio signal $\tilde{\mathbf{x}}_0 \approx \mathbf{x}_0$ and the transfer function parameters ψ as proposed recently [33, 39]. The aim is to minimize the following objective:

$$\arg\min_{\psi, \tilde{\mathbf{x}}_0} C(\mathbf{y}, \hat{f}(\tilde{\mathbf{x}}_0; \psi)), \quad \text{so that} \quad \tilde{\mathbf{x}}_0 \sim p_{\text{data}}, \tag{10}$$

where $C(\cdot, \cdot)$ is the cost function defined in Eq. (8). The constraint $\tilde{\mathbf{x}}_0 \sim p_{\text{data}}$ is weakly enforced by using a diffusion model $D_\theta(\mathbf{x}_\tau, \tau)$ trained using realizations of p_{data} as the score-based prior.

The inference algorithm, summarized in Algorithm 1, implements block-based optimization by alternating state updates of the audio signal $\mathbf{x}_\tau$ with parameter optimization iterations. The algorithm combines a Euler sampler for solving the posterior probability flow ODE introduced earlier in this section. For optimizing the operator $\hat{f}(\cdot; \psi)$ at a given time step τ, the first step is calculation of the one-step denoised estimate $\hat{\mathbf{x}}_0(\mathbf{x}_\tau)$ using $D_\theta(\mathbf{x}_\tau, \tau)$, then evaluate the cost function and perform several gradient descent iterations on the parameters ψ. In practice, the Adam optimizer

Algorithm 1. Diffusion Inference/operator optimization.

Require: distorted measurements $\mathbf{y}$
  Sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{y}, \sigma^2(T)\mathbf{I})$ ▷ *Warm initialization*
  Initialize ψ_T ▷ *Operator parameters initialization*
  for $i \leftarrow T, \ldots, 1$ **do** ▷ *Discrete step backwards*
    $\hat{\mathbf{x}}_0(\mathbf{x}_i) \leftarrow D_\theta(\mathbf{x}_i, \tau_i)$ ▷ *Evaluate denoiser model*
    $\psi_i^0 \leftarrow \psi_{i+1}$
    for $j \leftarrow 0, \ldots, N_{\text{its.}}$ **do** ▷ *Operator optim. loop*
      $\hat{\mathbf{y}} \leftarrow \hat{f}(\hat{\mathbf{x}}_0(\mathbf{x}_i); \psi_i^j)$ ▷ *Evaluate operator*
      $\psi_i^{j+1} \leftarrow \psi_i^j - \text{Adam}\left(\nabla_\psi C(\mathbf{y}, \hat{\mathbf{y}})\right)$ ▷ *optim. step*
    $\hat{\mathbf{y}} \leftarrow \hat{f}(\hat{\mathbf{x}}_0; \psi_i)$
    $\mathbf{l}_\mathbf{x} = -\zeta(\tau_i)\nabla_{\mathbf{x}_i} C(\mathbf{y}, \hat{\mathbf{y}})$ ▷ *Likelihood score estimate*
    $\mathbf{s}_i \leftarrow (\hat{\mathbf{x}}_0(\mathbf{x}_i) - \mathbf{x}_i)/\sigma^2(\tau_i)$ ▷ *Tweedie's formula (4)*
    $\mathbf{x}_{i-1} \leftarrow \mathbf{x}_i - \tau_i(\tau_{i-1} - \tau_i)(\mathbf{s}_i + \mathbf{l}_\mathbf{x})$ ▷ *Update step*
  return $\mathbf{x}_0$, ψ_0

is employed, which accounts for second-order momentum and empirically stabilizes convergence.

Following that procedure, the estimates $\hat{\mathbf{x}}_0(\mathbf{x}_\tau)$ at the beginning are coarse and contain mostly low-frequency components, leading to suboptimal minimizers for the operator optimization objective. Nevertheless, as the process continues, high-frequency details will gradually appear, enabling a more accurate fitting of the operator in the later stages of inference.

To accelerate and stabilize inference, a warm initialization strategy similar to [33, 34] is employed. This approach involves adding noise to the distorted measurement $\mathbf{y}$ to begin the process, formulated as $\mathbf{x}_T \sim \mathcal{N}(\mathbf{y}, \sigma^2(T)\mathbf{I})$, rather than initializing from pure Gaussian noise. The algorithm, outlined in Algorithm 1, is based on a first-order Euler sampler. Although a deterministic sampler for brevity is presented, the practical implementation incorporates stochasticity following the approach described in [37]. This stochasticity, which is consistent with prior works [33, 34], helps mitigate errors introduced during sampling and provides a modest but significant improvement in results.

2 NONLINEAR FUNCTION APPROXIMATIONS

The primary focus of this work is the approximation of an unknown nonlinear degradation operator, defined as $\hat{f}(\cdot; \psi)$. As explained in SEC. 1, the degradation is assumed to be a memoryless nonlinear function. Consequently, the search space of $\hat{f}(\cdot; \psi)$ is restricted to this class of functions. Importantly, $\hat{f}(\cdot; \psi)$ must be differentiable with respect to ψ almost everywhere, enabling gradient-based optimization methods such as Adam. This section presents the three nonlinear function models used in the experiments of this study.

2.1 Sum of Hyperbolic Tangent Functions

Colonel et al. [40] proposed a variety of parametric models to reverse engineer nonlinear distortion effects in recording mixes. Unlike their approach, which benefited from having input/output pairs, this work does not have such access, making the task more challenging. After conducting preliminary experiments, it was observed that the harmonic sum of hyperbolic tangent functions (SumTanh) demonstrated the most stability under these conditions, which motivated its selection for further study.

The mapping using the SumTanh function is defined as

$$y(n) = f(x(n)) = \sum_{q=1}^{Q} a(q) \tanh(q\, x(n)), \tag{11}$$

where Q is the order of the approximation. This means that a vector $\mathbf{a} = [a(1), \ldots, a(Q)]$ has Q learnable parameters. Based on author observations, order $Q = 8$ was selected. The composition of only smooth monotonic tanh functions is the main disadvantage of this approach. It does not allow reproducing nonlinear functions with sharp edges or nonmonotonic shapes.

2.2 Multilayer Perceptron

A multilayer perceptron (MLP) is a universal function approximator [41], making it a natural choice for modeling nonlinearities. For example, MLPs have been employed as the static nonlinearity component within a Wiener-Hammerstein model to simulate distortion audio effects [42]. Furthermore, Hayes et al. [43] proposed the use of MLPs as differentiable waveshapers for sound synthesis.

In this study, a small four-layer neural network with 20 neurons in both hidden layers was used. Rectified linear unit activations, which provide piecewise linear functions, are applied to all layers except the output layer. The weights are initialized using the Kaiming-normal initialization, and biases are initialized with zeroes.

The main advantages of MLPs are their high expressivity and efficient inference, enabling a large number of iterations to be performed during operator optimization. However, the network functions as a black box, making it difficult to design appropriate initialization strategies based on the characteristics of the observed distorted data.

2.3 Cubic Catmull-Rom Splines

Splines are parametric curves that are widely used in computer graphics. They can fit a wide variety of shapes with a small number of control points [44]. One of the most commonly used is a cubic, or third-order spline with four control points and a curve parameter $s \in [0, 1)$. Given N control points, a curve can be constructed by composing partial cubic splines $\mathbf{c}_i$, defined as follows:

$$\mathbf{c}_i(s) = \sum_{j=0}^{3} b_j(s)\mathbf{p}_{i+j}, \quad i = 0, 1, \ldots, N-3, \tag{12}$$

where b_j are elements of a basis vector $\mathbf{b}$ that defines the shape of the curves $\mathbf{c}_i$ between control points $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$. In particular, the basis vector of a CCR spline is defined as [44]

$$\mathbf{b}(s) = \begin{cases} b_0(s) = \frac{1}{2}(-s + 2s^2 - s^3), \\ b_1(s) = \frac{1}{2}(2 - 5s^2 + 3s^3), \\ b_2(s) = \frac{1}{2}(s + 4s^2 - 3s^3), \\ b_3(s) = \frac{1}{2}(-s^2 + s^3). \end{cases} \tag{13}$$

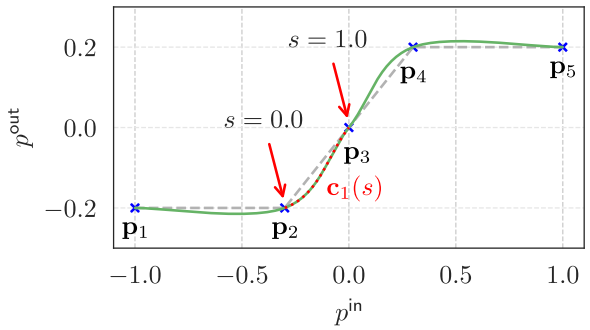

Fig. 1. Graphic representation of a CCR spline (solid green line) approximating a hard-clipping function (dashed line). The red line highlights one curve segment, $\mathbf{c}_1(s)$, that is controlled by the points $\mathbf{p}_1$ to $\mathbf{p}_4$.

An example of CCR is represented in Fig. 1. Notably, as shown in the figure, the curve $\mathbf{c}_1(s)$ connects two of the inner control points $\mathbf{p}_2$ and $\mathbf{p}_3$. This property allows accurate control across points that facilitates initialization strategies.

To describe the distortion operator $\hat{f}(\cdot, \psi)$, the authors use $N+2$ control points $\mathbf{p}_0, \ldots \mathbf{p}_{N+1}$ are used. Each control point $\mathbf{p}_i = [p_i^{\text{in}}, p_i^{\text{out}}]$ is composed of a grid point (input value) p_i^{in} and an output value p_i^{out}. The set of output values are the optimizable parameters $\psi = \{p_0^{\text{out}}, \ldots, p_{N+1}^{\text{out}}\}$. The grid points $p_1^{\text{in}}, \ldots, p_N^{\text{in}}$ are fixed and defined in the range $p_i^{\text{in}} \in (-1, 1)$. After initializing the grid points p_i^{in} uniformly in the range $(-1, 1)$, the following transformation is applied to each point to map it according to the μ-law:

$$p_i^{\text{in}} \leftarrow \text{sgn}(p_i^{\text{in}}) \frac{(1+\mu)^{|p_i^{\text{in}}|}}{\mu}. \tag{14}$$

This transformation increases the density of control points near zero with higher μ, aligning with the natural clustering of audio data, which often exhibits a higher concentration of values close to zero. The extremal points are set to $p_0^{\text{in}} < -1$ and $p_{N+1}^{\text{in}} > 1$ to ensure control over the curve throughout the whole range $(-1, 1)$. An odd number of control points is used to have one point at zero. In the experiments here, $N = 41$ points and $\mu = 20$ are used. The point p_i^{out} is initialized by mapping the distorted signal values to the nearest grid point p_i^{in} using the minimum distance between the signal points and the grid point.

To evaluate the distortion operator $\hat{f}(\cdot, \psi)$ during runtime, the input values $x(n)$ are discretized into bins defined by the grid points p_i^{in}. For each bin, the curve parameter $s \in [0, 1)$ is calculated based on the relative position of $x(n)$ within the interval. The corresponding spline segment $\mathbf{c}_i(s)$ is then evaluated by determining the basis vector $\mathbf{b}(s)$ and applying Eq. (12). This operation is fully differentiable with respect to the parameters ψ, allowing for gradient-based optimization.

3 EXPERIMENTS AND RESULTS

Because two tasks are solved in parallel in this study, experiments and results are presented in separate parts as well. The first part, SEC. 3.1, discusses how accurate the method is in estimating the unseen degradation operator. The second, SEC. 3.2, compares the proposed method for generating plausible restored guitar audio signals. Finally, SEC. 3.3 presents results on speech restoration.

The proposed method is designed to restore audio signals affected by any memoryless nonlinear distortion. The starting point is assessing its ability to estimate a varied range of memoryless nonlinear functions. Next, its effectiveness in restoring distorted audio signals is evaluated, with a particular focus on hard and soft clipping. This emphasis is motivated by the practical significance and widespread occurrence of these distortions and the availability of well-established baselines for comparison.

3.1 Operator Estimation

The first experiment evaluates how effectively the proposed approach generalizes across various memoryless nonlinear distortions and compares the performance of the nonlinear function approximations introduced in SEC. 2. Fig. 2 depicts the ramp responses of the evaluated nonlinear functions along with their blind estimates obtained using the different waveshaper methods. The specific nonlinear distortions studied are as follows:

a) Hard clipping: A classic distortion that applies a threshold, resulting in a sharply truncated waveform. In this study, hard clipping is evaluated with an input signal-to-distortion ratio (SDR) of 3 dB.
b) Soft clipping: Similar to hard clipping but with smooth transitions at the thresholds, making it less trivial to identify. Soft clipping is evaluated with the same input SDR of 3 dB.
c) Wavefolding: A nonmonotonic function where portions of a waveform are "folded back" on itself [45].
d) Quantization: This distortion occurs when signal values are quantized to a finite number of levels [46, 47]. Here, an extreme case with only three quantization levels that can be achieved with two bits is studied.
e) Half-wave rectifier (HWR): An asymmetric nonlinearity where only the positive half of the signal is preserved and the negative values are set to zero.

Additionally, although not included in the evaluation, the authors experiment with a carbon microphone distortion, a nonlinearity characteristic of early recording technologies, which introduces asymmetric waveshaping [48, 49]. This distortion is represented in Fig. 2(f). All details and audio examples are available on the project website.[1]

3.1.1 *Experimental Details*

A diffusion model was trained on direct input clean guitar data. This study utilizes the IDMT-SMT-GUITAR data set [50], specifically its fourth subset, which includes 64 brief musical pieces performed at varying tempos. These

[1] https://michalsvento.github.io/NLDistortionDiff/

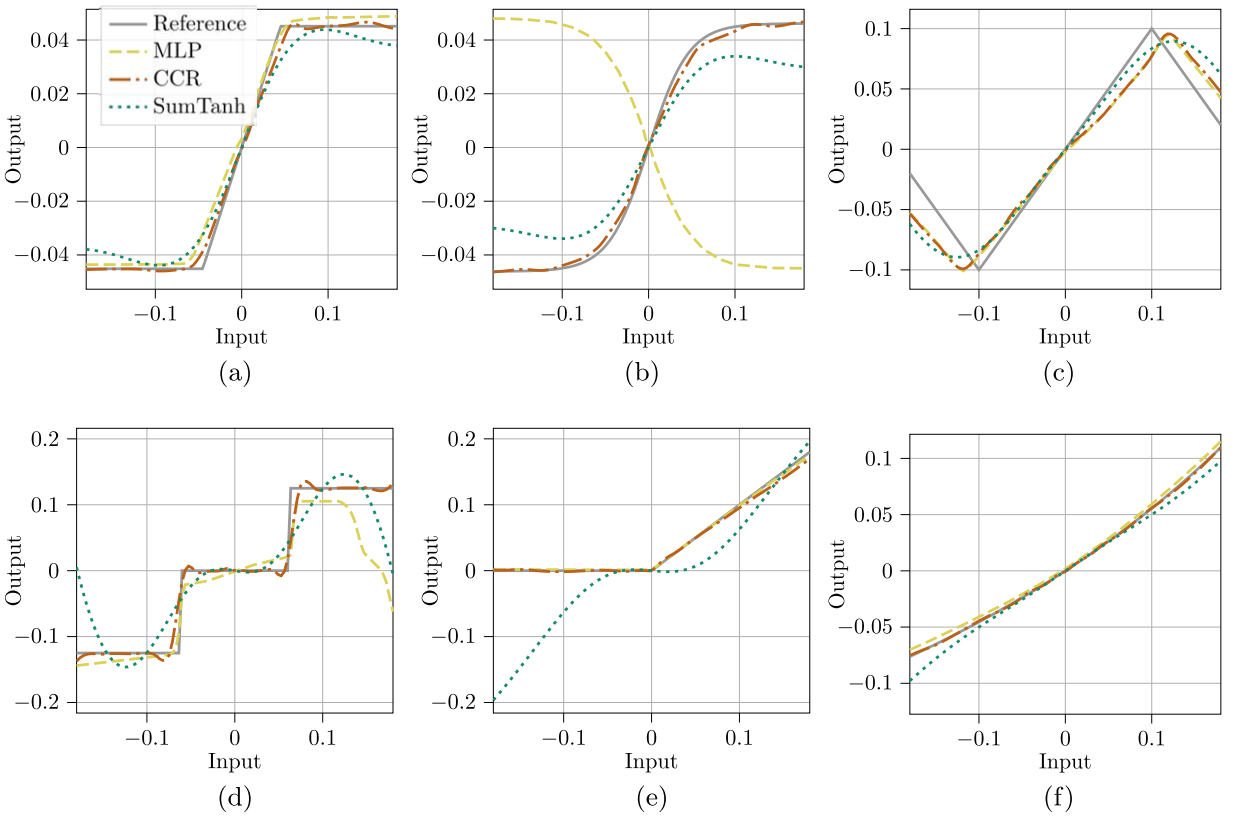

Fig. 2. Memoryless nonlinearities and the corresponding blind estimates with the three different models: (a) hard clipping, (b) soft clipping, (c) wavefolding, (d) quantization, (e) HWR, and (f) carbon microphone. The proposed method CCR is the most accurate one in most cases.

compositions were captured using two distinct electric guitars; however, the Career SG guitar is used exclusively in this study. Five minutes were allocated for testing, 5 min were allocated for validation, and the remaining 26.5 min were allocated for training. The evaluation data set was split into 50 segments, each approximately 6 s long, that were employed for objective evaluation. The data from this experiment were sampled at 44.1 kHz.

Given the limited volume of data available for training, it was observed that exclusive training of the diffusion model on this data set resulted in significant overfitting, evidenced by the model's tendency to memorize the data. To minimize this issue, the authors opted for pretraining the diffusion model with a related larger data set and then fine-tuned the model for only a small number of iterations. Specifically, the direct input recordings from the EGDB data set [51] were used, which contains around 2 h of audio.

The model architecture is based on a constant-Q transform–based U-Net from [52], with approximately 40 million parameters. The AdamW optimizer with weight decay (0.01) was used, as was a learning rate of 2e-5, and a batch size of four segments, each 6 s long. The training process involved 290,000 iterations of pretraining followed by 26,000 iterations of fine-tuning. Exponential moving average weights were tracked with a rate of 0.9999 and applied during evaluation.

For inference, $T = 50$ steps of reverse diffusion were performed with a DPS step size $\zeta' = 0.3$ and 20 optimization steps for operator estimation, using a learning rate of 0.02. For additional details, refer to the source code.[2] The authors aimed to ensure a fair comparison by selecting hyperparameters that worked well across all of the nonlinear function approximation models.

3.1.2 Evaluation Metrics and Results

In this evaluation, log-spectral distance (LSD) and ramp response mean squared error (RR-MSE) were used. Given clean guitar test samples $\mathbf{x}_0$, LSD quantifies the similarity between the blindly predicted distorted signal $\hat{f}(\mathbf{x}_0, \psi)$ and the reference distorted signal $f(\mathbf{x}_0)$ in the log-spectral domain. Given $Y \in \mathbb{C}^{M\times K}$ the STFT of the reference distorted audio, and $\hat{Y} \in \mathbb{C}^{M\times K}$ the STFT of the predicted one, LSD is computed as follows:

$$\mathrm{LSD} = \frac{1}{M \times K} \sum_{m=1,k=1}^{M,K} \Big(\log_{10}(|Y_{m,k}|^2 + \varepsilon) - \log_{10}(|\hat{Y}_{m,k}|^2 + \varepsilon)\Big)^2. \quad (15)$$

To calculate RR-MSE, a ramp signal $\mathbf{r} \in \mathbb{R}^P$ with $P = 1000$ linearly-spaced points in the range $[-3\sigma_{\mathrm{data}}, 3\sigma_{\mathrm{data}}]$

[2] https://github.com/michalsvento/NLDistortionDiff

Table 1. Comparison of three nonlinear function models in approximating five nonlinearities.

Nonlinearity	Method	LSD (dB)↓	RR-MSE (dB)↓
Hard clipping	SumTanh	15.19	−46.25
	MLP	3.77	−47.32
	CCR	**2.51**	**−54.82**
Soft clipping	SumTanh	3.27	−49.67
	MLP	6.71	−47.27
	CCR	**5.94**	**−56.81**
Wavefolding	SumTanh	13.41	−36.57
	MLP	4.09	−35.60
	CCR	**3.74**	**−39.29**
HWR	SumTanh	13.00	−23.98
	MLP	4.07	−39.45
	CCR	**2.94**	**−50.61**
Quantization	SumTanh	68.66	−27.33
	MLP	**4.72**	−32.62
	CCR	15.64	**−35.86**

Note. The metrics LSD and RR-MSE are reported in decibel scale. The best results are bolded.

was used as input, where σ_{data} represents the standard deviation of the clean data computed from the data set. With RR-MSE the accuracy of the learned operator was evaluated with respect to the reference operator within the range of prior data. The metric is defined as

$$\text{RR-MSE} = \frac{1}{P}\sum_{p=1}^{P}(f(\mathbf{r}_p) - \hat{f}(\mathbf{r}_p, \psi))^2. \tag{16}$$

In some cases, it was observed that the operator estimations converge to a result with a sign flip [see Fig. 2(b) for an example]. Although this may seem unexpected, the authors argue that this phenomenon should not be penalized because it does not affect the audio quality. This sign flip is an expected consequence of the blind estimation paradigm, where an audio signal estimate with a flipped sign is equally likely according to the probability distribution inferred from the diffusion model. Although this does not impact the LSD, it does significantly influence RR-MSE. To account for this, RR-MSE was also evaluated using a sign-flipped operator and the result that minimizes the error was selected.

Table 1 presents the average results of the experiment across the different evaluated nonlinearities. The CCR consistently outperforms the other methods in nearly all cases except for the quantization scenario, in which the MLP achieves a lower LSD, indicating superior performance in this specific instance. However, this trend does not hold for the RR-MSE case, where CCR maintained its advantage.

Fig. 2 exemplifies blind transfer function estimates generated by the three nonlinear function models. A single randomly selected estimate is shown for each example, which may not fully represent the performance across the entire data set. The figures show that the SumTanh model struggles to fit asymmetric nonlinearities, as observed in the HWR case [Fig. 2(e)]. Additionally, the SumTanh model has difficulty handling sharp transitions, making it most effective for soft clipping [Fig. 2(b)] compared with the others. The MLP, however, is good at fitting sharp edges and may outperform CCR in this regard because CCR tends to introduce small errors at these transitions[see Figs. 2(a) and 2(b)]. Nevertheless, CCR generally provides more robust estimates than MLP, as indicated by the metrics given in Table 1. It is speculated that the advantage of CCR lies in its optimization stability, partly due to easy initialization.

3.2 Distorted Guitar Restoration

The restoration of guitar signals was evaluated, focusing on hard and soft declipping, because other baselines for declipping are available. The same data set and implementation details as in SEC. 3.1 are used for consistency. Three degrees of distortion are analyzed: input SDRs of 1, 3, and 7 dB, where lower means higher distortion.

3.2.1 Compared Methods

As a baseline for guitar restoration, the Demucs network[3] was used, trained in a supervised manner over 100,000 iterations, following the approach outlined in [23]. Specialized models were trained separately for hard and soft clipping tasks to address their distinct characteristics.

For the hard clipping task, the input SDR was sampled from the range [0.1, 10], and the clipping threshold was adjusted accordingly. For the soft-clipping task, the model was trained on the same SDR range. However, the degradation operator for soft clipping was parameterized as described in [53]. This parameterization employs an adjustable hardness level $r \in [0, 1]$ sampled uniformly at random during training. At the extremes, when $r = 1$, the distortion closely resembles hard clipping, whereas $r = 0$ produces a smooth, tanh-like degradation. This setup aims to model a sufficiently diverse range of soft-clipping distortions.

Also included as a baseline is an informed version of the authors proposed algorithm. In this setup, the ground-truth degradation operator $f(\cdot)$ is utilized instead of blindly optimizing it. This corresponds to the method described in SEC. 1.2 and previously investigated for hard clipping in prior works [29, 30].

3.2.2 Evaluation Metrics and Results

Two reference-based metrics are reported: LSD and Oldenburg Perception Model quality assessment (PEMO-Q) [54]. LSD, as purely objective metric, quantifies the spectral distortion between the restored signal and the target. Its formula is the same as in (15), but applied to the restored and target signals. A value of zero indicates the best result, meaning the signals are identical. PEMO-Q is designed to estimate the perceptual quality of audio, focusing on aspects that are important for human perception. In Fig. 3, the objective difference grade score is shown, which ranges from −4 to 0, where higher values indicate better quality. The results are presented as violin plots for both metrics. Violin plots show the distribution of audio metrics, with

[3] https://github.com/facebookresearch/denoiser

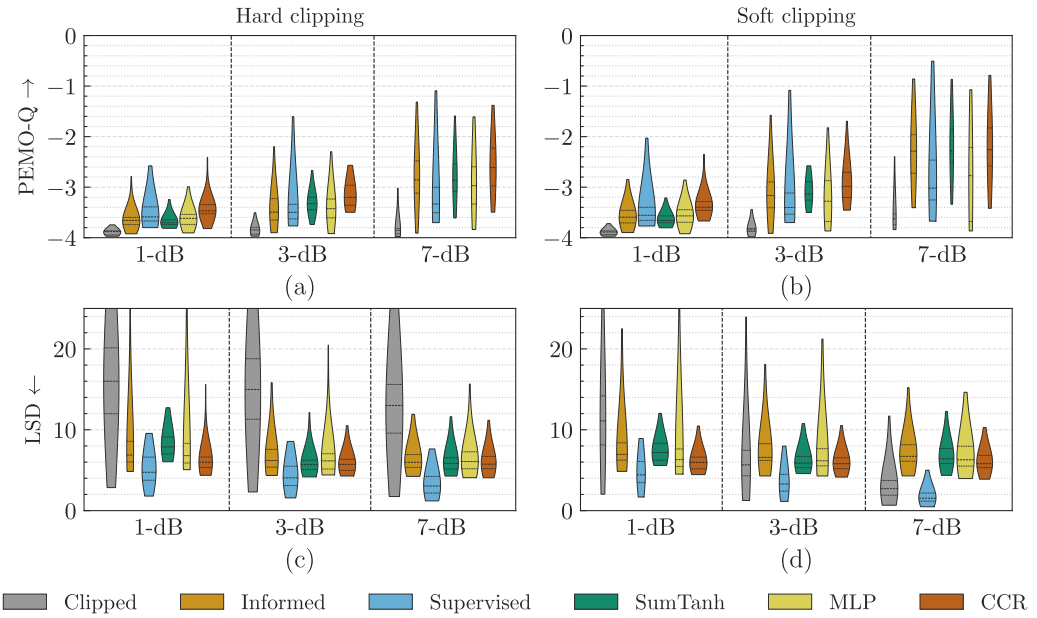

Fig. 3. Violin plots of PEMO-Q and LSD results evaluated on 50 selected samples from the IDMT-SMT-GUITAR test set. Every bin represents defined input SDR. Plots (a) and (b) show PEMO-Q results for hard and soft clipping, respectively. Plots (c) and (d) show the LSD metric for hard and soft clipping, respectively.

width representing data density. The outer dashed lines indicate the interquartile range, and the whiskers represent the data range. A symmetric, centered distribution suggests stable performance, whereas wider or skewed plots may indicate variability or instability.

The supervised baseline outperforms the proposed methods in terms of LSD. However, some of the proposed methods achieve better performance on PEMO-Q. This aligns with expectations because LSD is not well suited for evaluating generative models, whereas PEMO-Q, being a perceptual metric, provides a more relevant assessment. In the 1 dB SDR case for both hard and soft clipping, the CCR method demonstrates superior performance compared with all other approaches. For the 3 and 7 dB SDR cases, CCR shows marginally better results. Interestingly, CCR surpasses the informed setting in certain scenarios, particularly for the 1 dB case [see Figs. 3(a) and (b)].

Additionally, the Fréchet audio distance (FAD) [55], which quantifies the similarity between the distributions of real and generated audio embeddings, is reported. FAD is computed by comparing embeddings of the original and restored signals. Using the fadtk toolkit [56], FAD is computed with both the original VGGish embeddings and the embeddings from the audio compression model EnCodec [57]. FAD evaluates the quality of the generated audio in terms of both timbral and temporal aspects but does not directly capture hallucinations or instances where the restored audio may not match the content of the distorted measurement.

The results for FAD are shown in Fig. 4. When computed using VGGish embeddings, FAD heavily penalizes the supervised baseline at 1 and 3 dB SDR. Most methods achieve comparable scores for the remaining conditions, including the informed setting, except SumTanh, which performs worse at 1 dB SDR. When computed with EnCodec embeddings, the supervised baseline is less penalized and even achieves the best performance at 1 dB SDR. However, it is slightly outperformed by the blind methods in hard-clipping scenarios at 3 and 7 dB SDR. For soft clipping, the scores for the supervised baseline and CCR are comparable. Interestingly, in the EnCodec-based evaluation, the informed setting performs worse than both the supervised and blind methods for both hard and soft clipping at 1 and 3 dB SDR.

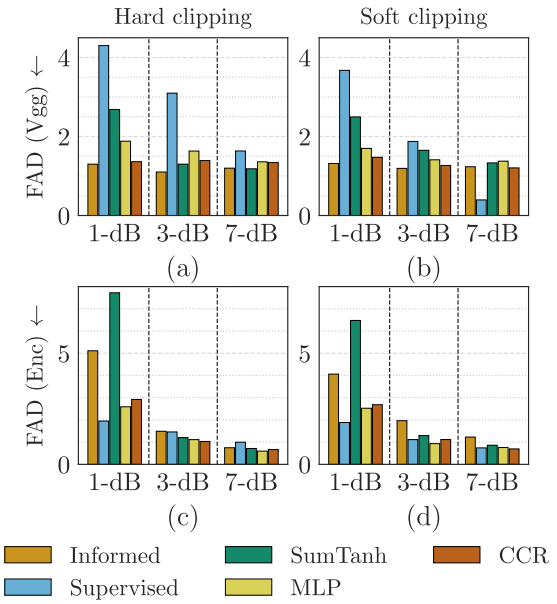

Fig. 4. FAD results on the IDMT-SMT-GUITAR test set using VGGish (Vgg) (a, b) and EnCodec (c, d) embeddings for hard (a, c) and soft (b, d) clipping.

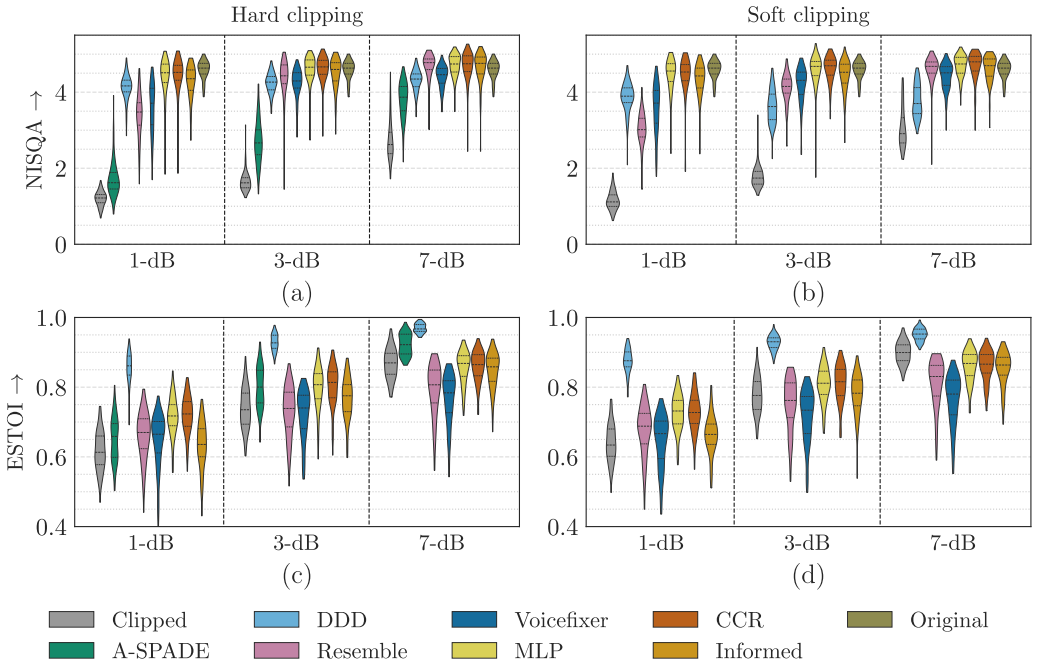

Fig. 5. Violin plots of NISQA and ESTOI results evaluated on 100 selected samples from the Voice Cloning Toolkit test set. Every bin represents defined input SDR. Panels (a) and (c) show results for hard clipping, while (b) and (d) show results for soft clipping.

3.3 Speech Restoration

In this section, the proposed method is evaluated on speech hard/soft declipping, a task with established relevance. A comparison with recent baselines is presented.

3.3.1 Experimental Details

The authors experiment with the Voice Cloning Toolkit data set [58], which contains studio-quality speech recordings. The same train split as Yi et al. [24] is employed to facilitate the comparison. Additionally, the data set is resampled to 16 kHz. The evaluation set contains 100 utterances from the two speakers in the test split: "p232" and "p257".

The authors use an STFT-based NCSN++M architecture that has proven successful for speech diffusion models [39, 59], whereas the original NCSN++ was proposed for image generation [60]. The model contains 27.8 million parameters. The training setup is similar to that described in SEC. 2 for the guitar, with a batch size of 8. The model undergoes a single-stage training process without fine-tuning for approximately 300,000 iterations. The inference details remain similar, but the authors opted for $T = 100$ sampling steps. For further specifications, please refer to the source code.[4]

3.3.2 Compared Methods and Baselines

The proposed method is evaluated for speech hard and soft declipping using the MLP and CCR models, comparing it against several representative baselines. The baselines include A-SPADE [21], a sparsity-based approach designed for hard-clipping restoration, and Demucs-Discriminator-Declipper (DDD) [24], a supervised model specifically trained for hard-clipping scenarios. The A-SPADE implementation provided by Záviška et al. [1][5] was used. Additionally, general-purpose speech enhancement tools are considered, such as VoiceFixer [26], an open-source model for audio restoration, and Resemble Enhance,[6] another recent speech enhancement tool. Finally, the informed setting is included, where the conditioning provides knowledge of the true clipping function.

3.3.3 Evaluation Metrics and Results

The performance of the proposed method is evaluated using two metrics. The Nonintrusive objective speech quality assessment (NISQA) [61] is a nonintrusive metric that estimates speech quality without requiring a reference signal. It ranges from 0 to 5, with higher values indicating better quality. Extended short-time objective intelligibility (ESTOI) [62], on the other hand, is a reference-based metric designed to evaluate speech intelligibility. It ranges from 0 to 1, where higher values indicate better intelligibility.

The results are presented as violin plots in Fig. 5. For NISQA, the proposed blind methods consistently achieve the best performance and, in some cases, even surpass the scores of the original signal [Fig. 5(a) and Fig. 5(b)]. The informed method performs slightly worse than the blind approaches, although the differences are not always statistically significant.

For ESTOI, the supervised baseline DDD significantly outperforms all other methods [(Figs. 5(c) and 5(d)]. Inter-

[4]https://github.com/michalsvento/NLDistortionDiff

[5]https://github.com/rajmic/declipping2020_codes

[6]https://github.com/resemble-ai/resemble-enhance

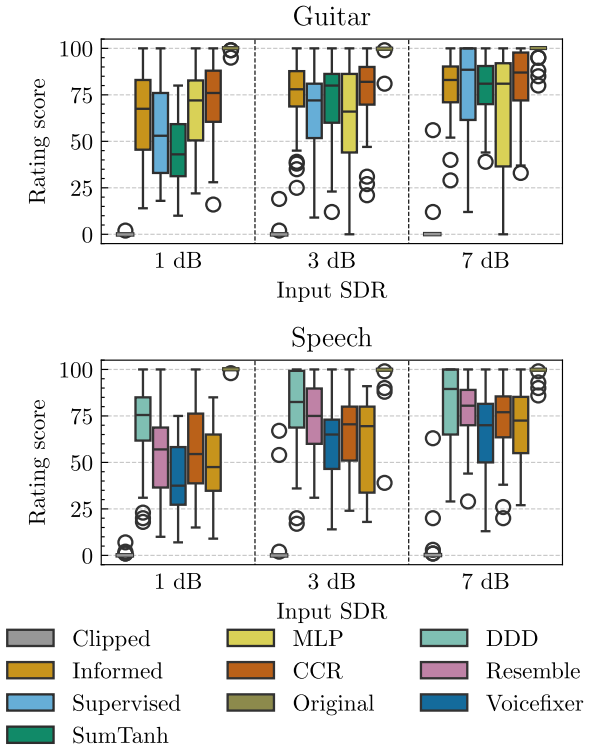

Fig. 6. Box plot representations of the listening test results categorized by input SDR.

estingly, A-SPADE and the clipped condition, which performed poorly on NISQA, achieve relatively high scores on ESTOI. This is attributed to phase differences that, although minimally perceptible, can significantly influence the ESTOI metric. Notably, the proposed blind methods surpass the speech enhancement baselines, VoiceFixer and Resemble Enhance. Overall, the results demonstrate that the proposed method achieves state-of-the-art performance for blind distorted speech restoration, delivering a strong balance between speech quality and intelligibility.

3.4 Subjective Evaluation

The authors conducted a listening test to evaluate the restoration performance for both guitar and speech signals. The test followed the multiple stimuli with hidden reference and anchor methodology [63], which includes a hidden reference and a low-quality anchor. Participants rated the audio excerpts on a scale from 0 to 100 based on perceived audio quality. They were instructed to identify the hidden reference and low anchor, assigning them scores of 100 and 0, respectively.

The test covered various distortion scenarios, including hard clipping and soft clipping, at input SDR levels of 1, 3, and 7 dB. For each condition, two examples were included for both guitar and speech signals, resulting in a total of 24 test pages. The same two guitar and speech excerpts were used across all distortion conditions to ensure consistency. Most of the baselines from SEC. 3.2 and SEC. 3.3 were included; however, some speech restoration baselines and conditions were omitted to reduce the cognitive load on participants and improve test reliability. For instance, the MLP and SumTanh models were not included in the speech listening test pages.

Table 2. The p values from pairwise comparisons of listening test results using the Wilcoxon signed-rank test.

Type		1 dB	3 dB	7 dB
Guitar	Informed vs. SumTanh	$<$ **0.001**	0.171	0.938
	Informed vs. MLP	0.380	**0.012**	0.184
	Informed vs. CCR	0.068	0.095	0.290
	Supervised vs. SumTanh	**0.019**	0.052	0.908
	Supervised vs. MLP	**0.004**	0.801	0.255
	Supervised vs. CCR	$<$ **0.001**	$<$ **0.001**	0.456
	SumTanh vs. CCR	$<$ **0.001**	**0.003**	0.281
	SumTanh vs. MLP	$<$ **0.001**	0.106	0.223
	MLP vs. CCR	0.164	**0.002**	0.221
Speech	DDD vs. CCR	$<$ **0.001**	**0.005**	**0.009**
	Resemble vs. CCR	0.942	0.052	0.189
	VoiceFixer vs. CCR	$<$ **0.001**	**0.010**	**0.006**
	Informed vs. CCR	0.147	**0.016**	0.328

Note. Statistically significant values ($p < 0.05$) are highlighted in bold.

A total of 13 participants completed the test. Following the guidelines in [63], participants whose reference ratings fell below 90 on more than 15% of the test pages were excluded from the analysis, resulting in a final set of ten valid participants.

3.4.1 Listening Test Results

The subjective evaluation results, summarized in Fig. 6, reveal distinct restoration patterns between guitar and speech signals that align with the objective findings from SECS. 3.2 and 3.3. Results aggregate both hard and soft-clipping conditions because preliminary analysis showed no systematic quality preference between distortion types.

Analysis of guitar signals at 1 dB input SDR shows the CCR method significantly outperforming the supervised baseline ($p < 0.001$) and demonstrating strong advantages over both SumTanh ($p < 0.001$) and MLP ($p = 0.004$) approaches (Table 2). This superiority persists at 3 dB SDR, although with reduced effect sizes (supervised baseline: $p < 0.001$, SumTanh: $p = 0.003$, and MLP: $p = 0.002$). At 7 dB SDR, no method shows statistically significant dominance ($p > 0.05$ for all comparisons), suggesting listeners found restoration quality comparable for mild distortions.

For speech signals, the specialized DDD baseline consistently outperforms CCR across all SDR levels ($p < 0.01$). However, CCR demonstrates competitive performance against general-purpose approaches, matching the informed baseline at 1 ($p = 0.147$) and 7 dB SDR ($p = 0.328$) and outperforming it at 3 dB ($p = 0.016$), whereas significantly surpassing VoiceFixer at all distortion levels ($p < 0.05$). CCR achieves parity with the Resemble enhancement system ($p > 0.05$ at all SDR).

4 CONCLUSION

The authors introduced a novel approach for the blind estimation and inversion of memoryless nonlinearities in mu-

sic and speech recordings. This method represents the first general application of diffusion models and CCR splines for modeling nonlinear audio distortions. The approach operates in challenging scenarios where both the original input signal and the applied nonlinearity are unknown. Additionally, the method is unsupervised, utilizing only distorted measurements and a diffusion model that is trained on clean audio files.

The proposed method demonstrated strong performance in estimating a diverse set of memoryless nonlinearities and restoring clipped music and speech signals. Among the nonlinear function approximation models, CCR splines emerged as the most effective overall. Although, in the task of audio declipping, the method underperformed supervised baselines in reference-based metrics, it excelled in nonintrusive metrics, such as NISQA, FAD, and the perceptually motivated PEMO-Q metric. Additionally, multiple stimuli with hidden reference and anchor listening tests were conducted, further supporting the method's perceptual quality improvements.

Memoryless systems, although limited in scope, provide a valuable academic exercise for understanding and addressing audio distortions in difficult blind and unsupervised settings. However, in real-world applications, distortion effects are typically more complex and often involve memory. Previous work has demonstrated promising results in modeling linear systems using similar approaches [33, 34] but not with the same level of accuracy achieved for memoryless distortions in this study. Future work should focus on extending this method to address more complex nonlinear effects with intricate dynamics.

5 ACKNOWLEDGMENT

This study was conducted during a 4-month Erasmus+ traineeship of the first author at the Aalto Acoustics Lab from in August to December 2024. The work of the first author was also supported by the Czech Science Foundation (GAČR) Project No. 23-07294S. The authors acknowledge the computational resources provided by the Aalto Science-IT project. The authors are grateful to Professor Pavel Rajmic for his support and helpful discussions.

6 REFERENCES

[1] R. Huber, and B. Kollmeier, "Pemo-Q—A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," in *Proceedings of the IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911 (2006 Dec.). https://doi.org/10.1109/TASL.2006.883259.

[2] J. S. Abel and J. O. Smith, "Restoring a Clipped Signal," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3, pp. 1745–1748 (Toronto, Candada) (1991 Apr.). https://doi.org/10.1109/ICASSP.1991.150655.

[3] K. Siedenburg, M. Kowalski, and M. Dörfler, "Audio Declipping With Social Sparsity," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1577–1581 (Florence, Italy) (2014 Jul.). https://doi.org/10.1109/ICASSP.2014.6853863.

[4] F. Eichas and U. Zölzer, "Black-Box Modeling of Distortion Circuits with Block-Oriented Models," in *Proceedings of the International Conference on Digital Audio Effects (DAFx-16)*, pp. 5–9 (Brno, Czech Republic) (2016 Sep.).

[5] V. Välimäki, S. González, O. Kimmelma, and J. Parviainen, "Digital Audio Antiquing—Signal Processing Methods for Imitating the Sound Quality of Historical Recordings," *J. Audio Eng. Soc.*, vol. 56, no. 3, pp. 115–139 (2008 Mar.).

[6] O. Mikkonen, A. Wright, E. Moliner, and V. Välimäki, "Neural Modeling of Magnetic Tape Recorders," in *Proceedings of the International Conference on Digital Audio Effects (DAFx-23)*, pp. 196–203 (Copenhagen, Denmark) (2023 Sep.).

[7] L. Tronchin, "The Emulation of Nonlinear Time-Invariant Audio Systems With Memory by Means of Volterra Series," *J. Audio Eng. Soc.*, vol. 60, no. 12, pp. 984–996 (2013 Dec.).

[8] D. Bouvier, T. Hélie, and D. Roze, "Nonlinear Homogeneous Order Separation for Volterra Series Identification," in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*, pp. 3–10 (Edinburgh, UK) (2017 Sep.).

[9] S. Orcioni, A. Terenzi, S. Cecchi, F. Piazza, and A. Carini, "Identification of Volterra Models of Tube Audio Devices Using Multiple-Variance Method," *J. Audio Eng. Soc.*, vol. 66, no. 10, pp. 823–838 (2018 Oct.). https://doi.org/10.17743/jaes.2018.0046.

[10] C. Bennett and S. Hopman, "Antialiasing for Simplified Nonlinear Volterra Models," *J. Audio Eng. Soc.*, vol. 70, no. 9, pp. 690–697 (2022 Sep.). https://doi.org/10.17743/jaes.2022.0033.

[11] A. Novak, L. Simon, P. Lotton, and J. Gilbert, "Chebyshev Model and Synchronized Swept Sine Method in Nonlinear Audio Effect Modeling," in *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, p. 15 (Graz, Austria) (2010 Sep.).

[12] J. D. Parker, F. Esqueda, and A. Bergner, "Modelling of Nonlinear State-Space Systems Using a Deep Neural Network," in *Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx-19)*, pp. 2–6 (Birmingham, UK) (2019 Sep).

[13] A. Wright, E.-P. Damskägg, L. Juvela, and V. Välimäki, "Real-Time Guitar Amplifier Emulation With Deep Learning," *Appl. Sci.*, vol. 10, no. 3, p. 766 (2020 Jan.). https://doi.org/10.3390/app10030766.

[14] T. Vanhatalo, P. Legrand, M. Desainte-Catherine, P. Hanna, and G. Pille, "Evaluation of Real-Time Aliasing Reduction Methods in Neural Networks for Nonlinear Audio Effects Modelling," *J. Audio Eng. Soc.*, vol. 72, no. 3, pp. 114–122 (2024 Mar.). https://doi.org/10.17743/jaes.2022.0122.

[15] Y.-T. Yeh, Y.-H. Chen, Y.-C. Cheng, et al., "DDSP Guitar Amp: Interpretable Guitar Amplifier Modeling," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal*

Processing, pp 1–5 (Hyderabad, India) (2025 Apr.). https://doi.org/10.1109/ICASSP49660.2025.10889306.

[16] A. Wright, V. Välimäki, and L. Juvela, "Adversarial Guitar Amplifier Modelling with Unpaired Data," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1–5 (Rhodes Island, Greece) (2023 Jun.). https://doi.org/10.1109/ICASSP49357.2023.10094600.

[17] E. Moliner, S. Braun, and H. Gamper, "Gaussian Flow Bridges for Audio Domain Transfer With Unpaired Data," in *Proceedings of the 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 374–378 (Aalborg, Denmark) (2024 Sep.). https://doi.org/10.1109/IWAENC61483.2024.10693958.

[18] 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," in *Proceedings of the 25th International Society for Music Information Retrieval Confrerence* (San Francisco, CA) (2024 Nov.). https://doi.org/10.48550/arXiv.2407.10646.

[19] A. Wright, A. Carson, and L. Juvela, "OpenAmp: Synthetic Data Framework for Audio Effect Foundation Models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5 (Hyderabad, India) (2025 Apr.). https://doi.org/10.1109/ICASSP49660.2025.10888232.

[20] C. J. Steinmetz, S. Singh, M. Comunità, et al., "ST-ITO: Controlling Audio Effects for Style Transfer With Inference-Time Optimization," in *Proceedings of the 25th International Society for Music Information Retrieval* (San Francisco, CA) (2024 Nov.).

[21] S. Kitić, N. Bertin, and R. Gribonval, "Sparsity and Cosparsity for Audio Declipping: A Flexible Non-Convex Approach," in *Proceedings of the 12th International Conference on Latent Variable Analysis and Signal Separation*, vol. 9237, pp. 243–250 (2015 Aug.). https://doi.org.10.1007/978-3-319-22482-4_28.

[22] F. R. Ávila, M. P. Tcheou, and L. W. Biscainho, "Audio Soft Declipping Based on Constrained Weighted Least Squares," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1348–1352 (2017 Sep.). https://doi.org/10.1109/LSP.2017.2727964.

[23] J. Imort, G. Fabbro, M. A. M. Ramírez, S. Uhlich, Y. Koyama, and Y. Mitsufuji, "Distortion Audio Effects: Learning How to Recover the Clean Signal," *arXiv preprint arXiv:2202.01664* (2022).

[24] J. Yi, J. Koo, and K. Lee, "DDD: A Perceptually Superior Low-Response-Time DNN-Based Declipper," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 801–805 (Seoul, South Korea) (2024 Apr). https://doi.org/10.1109/ICASSP48485.2024.10446235.

[25] Y.-S. Lee, Y.-P. Peng, J.-T. Wu, M. Cheng, L. Su, and Y.-H. Yang, "Distortion Recovery: A Two-Stage Method for Guitar Effect Removal," in *Proceedings of the 27th International Conference on Digital Audio Effects (DAFx-24)*, pp. 177–184 (Guildford, UK) (2024 Sep.).

[26] H. Liu, X. Liu, Q. Kong, et al., "VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration," in *Proceedings of Interspeech*, pp. 4232–4236 (Incheon, South Korea) (2022 Sep.). https://doi.org/10.21437/Interspeech.2022-11026.

[27] M. Rice, C. J. Steinmetz, G. Fazekas, and J. D. Reiss, "General Purpose Audio Effect Removal," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5 (New Paltz, NY) (2023 Oct.). https://doi.org/10.1109/WASPAA58266.2023.10248157.

[28] 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, p. 9 (2024 Feb.). https://doi.org/10.1186/s13636-024-00330-0.

[29] E. Moliner, J. Lehtinen, and V. Välimäki, "Solving Audio Inverse Problems With a Diffusion Models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5 (Rhodes Island, Greece) (2023 Jun.). https://doi.org/10.1109/ICASSP49357.2023.10095637.

[30] C. Hernandez-Olivan, K. Saito, N. Murata, et al., "VRDMG: Vocal Restoration via Diffusion Posterior Sampling With Multiple Guidance," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 596–600 (Seoul, South Korea) (2024 Apr.). https://doi.org/10.1109/ICASSP48485.2024.10446423.

[31] M. Levy, B. D. Giorgi, F. Weers, A. Katharopoulos, and T. Nickson, "Controllable Music Production With Diffusion Models and Guidance Gradients," *CoRR*, vol, abs/2311.00613 (2023 Dec.). https://doi.org/10.48550/arXiv.2311.00613.

[32] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion Posterior Sampling for General Noisy Inverse Problems," in *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, p. 1619 (Kigali, Rwanda) (2023 May).

[33] E. Moliner, F. Elvander, and V. Välimäki, "Blind Audio Bandwidth Extension: A Diffusion-Based Zero-Shot Approach," *IEEE Trans. Audio Speech Lang. Process*, vol. 32, pp. 5092–5105 (2024 Nov.). https://doi.org/10.1109/TASLP.2024.3507566.

[34] J.-M. Lemercier, E. Moliner, S. Welker, V. Välimäki, and T. Gerkmann, "Unsupervised Blind Joint Dereverberation and Room Acoustics Estimation With Diffusion Models," *arXiv preprint arXiv:2408.07472* (2024).

[35] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating Gradients for Waveform Generation," in *Proceedings of the Ninth International Conference on Learning Representations (ICLR)* (Virtual) (2021 May).

[36] H. Liu, Z. Chen, Y. Yuan, et al., "AudioLDM: Text-to-Audio Generation With Latent Diffusion Models," in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, 21450–21474 (Honolulu, HI) (2023 Jul.).

[37] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the Design Space of Diffusion-Based Gen-

erative Models," in *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 26565–26577 (New Orleans, LA) (2022 Dec.).

[38] A. Hyvärinen, "Estimation of Non-Normalized Statistical Models by Score Matching," *J. Mach. Learn. Res.*, vol. 6, no. 4, pp. 695–709 (2005 Apr.).

[39] E. Moliner, J.-M. Lemercier, S. Welker, T. Gerkmann, and V. Välimäki, "BUDDy: Single-Channel Blind Unsupervised Dereverberation With Diffusion Models," in *Proceedings of the 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, p. 120–124 (Aalborg, Denmark) (2024 Sep.). https://doi.org/10.1109/IWAENC61483.2024.10694254.

[40] J. T. Colonel, M. Comunità, and J. Reiss, "Reverse Engineering Memoryless Distortion Effects With Differentiable Waveshapers," presented at the *153rd Convention of the Audio Engineering Society Convention* (2022 Oct.), paper 10626.

[41] K. Hornik, M. Stinchcombe, and H. White, "Multilayer Feedforward Networks Are Universal Approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366 (1989 Jan.). https://doi.org/10.1016/0893-6080(89)90020-8.

[42] B. Kuznetsov, J. D. Parker, and F. Esqueda, "Differentiable IIR Filters for Machine Learning Applications," in *Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx-20)*, pp. 297–303 (Vienna, Austria) (2020 Sep.).

[43] B. Hayes, C. Saitis, and G. Fazekas, "Neural Waveshaping Synthesis," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, pp. 254–261 (Online) (2021 Nov.).

[44] R. H. Bartels, J. C. Beatty, and B. A. Barsky, *An Introduction to Splines for Use in Computer Graphics and Geometric Modeling* (Morgan Kaufmann, San Francisco, CA, 1995).

[45] F. Esqueda, H. Poyntinen, J. Parker, and S. Bilbao, "Virtual Analog Model of the Lockhart Wave Folder," in *Proceedings of the 14th Sound and Music Computing Conference* , pp 336–342 (Espoo, Finland) (2017 Jul.).

[46] P. Záviška, P. Rajmic, and O. Mokrý, "Audio Dequantization Using (Co)Sparse (Non)Convex Methods," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 701–705 (Toronto, Canada) (2021 Jun). https://doi.org/10.1109/ICASSP39728.2021.9414637.

[47] O. Mokrý, P. Rajmic, and P. Záviška, "Flexible Framework for Audio Reconstruction," in *Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx-20)*, pp. 117–124 (Vienna, Austria) (2020 Sep.).

[48] T. Quatieri, D. Reynolds, and G. O'Leary, "Estimation of Handset Nonlinearity with Application to Speaker Recognition," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 5, pp. 567–584 (2000 Sep.). https://doi.org/10.1109/89.861376.

[49] S. Oksanen and V. Välimäki, "Digital Modeling of the Vintage Telephone Sound," in *Proceeding of the International Computer Music Conference*, p. 636–639 (Huddersfield, UK) (2011 Jul.-Aug.).

[50] 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 *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx-14)*, pp. 219–226 (Erlangen, Germany) (2014 Sep.).

[51] Y.-H. Chen, W.-Y. Hsiao, T.-K. Hsieh, 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 786–790 (Singapore, Singapore) (2022 May). https://doi.org/10.1109/ICASSP43922.2022.9747697.

[52] E. Moliner, M. Turunen, F. Elvander, and V. Välimäki, "A Diffusion-Based Generative Equalizer for Music Restoration," in *Proceedings of the 27th International Conference on Digital Audio Effects (DAFx-24)*, pp. 25–32 (Guildford, UK) (2024 Sep.).

[53] B. Defraene, T. van Waterschoot, H. J. Ferreau, M. Diehl, and M. Moonen, "Real-Time Perception-Based Clipping of Audio Signals Using Convex Optimization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 10, pp. 2657–2671 (2012 Dec.). https://doi.org/10.1109/TASL.2012.2210875.

[54] R. Huber and B. Kollmeier, "PEMO-Q—A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911 (2006 Nov.). https://doi.org/10.1109/TASL.2006.883259.

[55] D. Roblek, K. Kilgour, M. Sharifi, and M. Zuluaga, "Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms," in *Proceedings of Interspeech*, pp. 2350–2354 (Graz, Austria) (2019 Sep.). https://doi.org/10.21437/Interspeech.2019-2219.

[56] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou, "Adapting Frechet Audio Distance for Generative Music Evaluation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1331–1335 (Seoul, South Korea) (2024 Apr.). https://doi.org/10.1109/ICASSP48485.2024.10446663.

[57] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," *Trans. Mach. Learn. Res.* (2023 Sep.). https://openreview.net/forum?id=ivCd8z8zR2.

[58] C. Valentini-Botinhao, "Reverberant Speech Database for Training Speech Dereverberation Algorithms and TTS Models," University of Edinburgh (2016 Jun.) https://doi.org/10.7488/ds/1425.

[59] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech Enhancement and Dereverberation With Diffusion-Based Generative Models," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol., 31, pp. 2351–2364 (2023 Jun.). https://doi.org/10.1109/TASLP.2023.3285241.

[60] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models," *arXiv preprint arXiv:2010.02502arXiv:2010.02502* (2020).

[61] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction With Crowdsourced Datasets," in *Proceedings of Interspeech*, pp. 2127–2131 (Brno, Czech Republic) (2021 Aug.-Sep.). https://doi.org/10.21437/Interspeech.2021-299.

[62] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 11, pp. 2009–2022 (2016 Nov.). https://doi.org/10.1109/TASLP.2016.2585878.

[63] ITU-R, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," *Recommendation ITU-R BS.1534-3*, (2015 Oct.).

THE AUTHORS


Michal Švento


Eloi Moliner


Lauri Juvela


Alec Wright


Vesa Välimäki

Michal Švento completed his B.Sc. and M.Sc. degrees in audio engineering at Brno University of Technology, Czechia, in 2021 and 2023, respectively. He is currently a doctoral student in the telecommunications program at the university's Department of Telecommunications. His research focuses on solving audio inverse problems by combining machine learning methods with traditional approaches.

•

Eloi Moliner received his B.Sc. degree in telecommunications technologies and services engineering from the Polytechnic University of Catalonia, Spain, in 2018 and his M.Sc. degree in telecommunications engineering from the same university in 2021. He is currently a doctoral candidate at the Acoustics Lab of Aalto University in Espoo, Finland. His research interests include digital audio restoration and audio applications of machine learning.

•

Lauri Juvela is an assistant professor in machine learning for speech and language technology at Aalto University, Finland. He earned his D.Sc. degree in speech synthesis at Aalto University in 2020 and worked as a machine learning researcher at Neural DSP Technologies from 2019 to 2023. His research interests include deep generative models for speech and audio.

•

Alec Wright is a Chancellor's Fellow in audio machine learning at the University of Edinburgh. He received his D.Sc. degree from Aalto University, Finland, in 2023. His research interests include virtual analog modeling of guitar amplifiers and other audio effects and machine learning for audio applications.

•

Vesa Välimäki is a full professor of audio signal processing and vice dean for research at Aalto University, Espoo, Finland. He received his D.Sc. degree from the Helsinki University of Technology in 1995. In 2008, he was the chair of the 11th International Conference on Digital Audio Effects, DAFx-08. In 2017, he was the chair of the 14th International Sound and Music Computing Conference, SMC-17. Prof. Välimäki is the editor-in-chief of the *Journal of the Audio Engineering Society*.