Diffusion-Based Denoising of Historical Recordings

BERNARDO V. MIRANDA,[1]* RAFAEL A. DESLANDES,[1]
(bvm@poli.ufrj.br) (rafael.deslandes@smt.ufrj.br)
IGNACIO IRIGARAY,[2] *AES Associate Member* **AND LUIZ W. P. BISCAINHO,[1]** *AES Member*
(irigaray@fing.edu.uy) (wagner@smt.ufrj.br)

[1]Signals, Multimedia, and Telecommunications Laboratory, Department of Electronic and Computer Engineering, Polytechnic School (DEL/POLI), and Electrical Engineering Program, Alberto Luiz Coimbra Institute of Graduate Studies and Research in Engineering (PEE/COPPE), Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[2]Grupo de Procesamiento de Audio, Instituto de Ingeniería Eléctrica (IIE), Facultad de Ingeniería (FING), Universidad de la República, Montevideo, Uruguay

In the context of audio restoration, the need to remove background noise from historical music recordings is a recurring problem, for which traditional signal processing and supervised deep learning methods have been previously applied. In this work, a generative approach that adapts conditional diffusion sampling for removing perceptually distributed noise is investigated, using the particular case of background noise removal from solo classical piano recordings as a proof of concept. The proposed method uses a set of noise examples to simulate perceptually distributed noise with specific characteristics throughout conditional diffusion sampling. Experiments with real historical 78 RPM recordings and clean recordings with added 78 RPM noise and tape hiss demonstrate that diffusion-based audio denoising performs comparably to state-of-the-art deep learning methods.

0 INTRODUCTION

Recorded audio has a long history dating back to the late 19th century. From wax cylinders and vulcanite discs, to vinyl long plays and magnetic tapes, to CDs and digital audio formats, technological advances in audio recording have shaped the way people listen to music today.

All recorded sound has some level of degradation associated with the recording, storage, or reproduction stages of its production. The study of techniques to correct defects in previously recorded audio defines the domain of audio restoration, in which this work is inserted.

Historically speaking, audio restoration works classified artifacts as local (e.g., clicks, thumps) or global (e.g., tape hiss) and dealt with these two classes of defects using different strategies. For example, traditional methods for restoring distributed additive noise, such as [1–3], rely mainly on assumptions on the spectrum of the noise to be removed and apply processing to the whole degraded signal. In contrast, typical methods for removing localized artifacts, such as the click removal procedure of [4], apply processing to a few samples of the signal and rely mainly on statistical detection schemes to determine signal samples that must be removed and then interpolated.

In [5] and [6], novel approaches were introduced to jointly remove local and global artifacts from musical recordings using supervised neural networks. In both cases, a strategy was devised to artificially degrade a dataset of clean digital recordings using realistic noise from recordings of the acoustical and electrical eras, thus generating many pairs of <clean,degraded> signals. Having these pairs, the authors of both papers were able to use supervised learning to train deep neural networks capable of denoising historical recordings. In [6], a dataset of 78 RPM disc noise was created to prepare a model specialized in removing this type of additive background noise. The same procedure was later adapted with a tape hiss dataset in [7] to create a model specialized in dehissing. However, due to their supervised learning strategy, models trained within the framework of [5–7] are limited to a specific noise family and have no guarantee of performance when faced with other additive defects.

*To whom correspondence should be addressed, email: bvm@poli.ufrj.br.

One way to overcome this limitation is to use diffusion models. Diffusion models are probabilistic generative models that have been applied most frequently in the (different, but correlated) domain of speech enhancement [8, 9] but also achieve remarkable results for musical audio restoration [10, 11]. Although they can be used within a supervised learning framework, as done in [8] for speech dereverberation, they also have the capability of being used as flexible zero-shot inverse problem solvers.

The sampling procedure of a pretrained diffusion model can be conditioned with the aid of a likelihood term. By changing this term accordingly, the same diffusion model can be used to restore different audio defects. For example, in [9], a diffusion model pretrained in clean speech data is used to correct artifacts introduced by four different speech enhancement modules; in [11], diffusion models pretrained on piano and singing voice data are used to equalize historical recordings with different band limitations; and in [10], a single diffusion model pretrained on piano data is used for bandwidth extension, inpainting, and declipping. A detailed review on diffusion models applied to general audio restoration can be found in [12].

This paper proposes a novel method to remove additive background noise from historical music recordings using conditional sampling with a diffusion-based generative model. The proposed method is zero-shot in a similar fashion to [9–11]; once the unconditional diffusion model is trained for a data distribution, different types of additive noise can be removed without the need to train a specialized model. However, unlike [9], it does not require any additional restoration modules, and unlike [10] and [11], it does not require an explicit model for the degradation that should be restored.

The remainder of this paper is organized as follows. SEC. 1 gives the theoretical background required to understand the proposed method, with an overview of existing diffusion frameworks and an explanation of conditional sampling for inverse audio problems. Following this, SEC. 2 gives a detailed description of the proposed method, and SEC. 3 outlines the model used for denoising in this work. SEC. 4 discusses experiments evaluating the proposed method for 78 RPM noise and tape hiss removal. The model presented in SEC. 3 is used in both experiments, without further training, to showcase the zero-shot nature of the method. Finally, SEC. 5 closes this work with remarks on the results obtained and possible research directions.

1 THEORETICAL BACKGROUND

1.1 Diffusion Models

In physics, the phenomenon of diffusion refers to the spread of particles in a medium due to the action of a random external force. During physical diffusion, the distribution of the position of the particles changes from $p_0(\mathbf{x})$ at time $t = 0$ to $p_T(\mathbf{x})$ at time $t = T$. Diffusion generative models borrow this idea and learn the *reverse diffusion* process, that is, the process that maps a sample from a known distribution $p_T(\mathbf{x})$ (usually Gaussian) back to a sample from a data distribution $p_0(\mathbf{x})$ [13].

Diffusion generative models first appeared in [13] but were popularized in [14], where forward diffusion is formulated as the gradual conversion of $p_0(\mathbf{x})$ into a normal distribution by iterative addition of Gaussian noise over many steps. In [14], reverse diffusion is performed using a neural network that is trained to estimate the amount of noise added to a data sample at a given step i. Then, starting from a sample of a Gaussian distribution, it is possible to gradually remove the noise until a sample of the desired data distribution is obtained.

An alternative formulation for diffusion appeared in [15], in which the authors suggested using a Langevin Markov chain Monte Carlo (MCMC) [16] procedure to sample data points from an arbitrary distribution. In Langevin MCMC, a recurrence based on the *score function* $\nabla_{\mathbf{x}} \log p(\mathbf{x})$—where $p(\mathbf{x})$ is the distribution to be modeled—is used to transform a sample from an arbitrary distribution into a sample from $p(\mathbf{x})$. In [15], the authors propose using a neural network to estimate the score function using a *denoising score matching* [17] loss function.

Recent articles on diffusion models use the framework of [18] and [19], which unifies the formulations of [14] and [15]. In [18], the authors showed that the loss functions of the previous two diffusion formulations could be cast as denoising score matching losses. More formally, the loss of the neural networks in [14] and [15] can be written as

$$L(\boldsymbol{\theta}) = \mathbb{E}_i \left[\lambda_i \, \mathbb{E}_{f_{0,i}} \left[\left|\left| s^{\boldsymbol{\theta}}(\mathbf{x}_i, i) - \nabla_{\mathbf{x}_i} \log p_i(\mathbf{x}_i|\mathbf{x}_0) \right|\right|^2 \right] \right], \tag{1}$$

where vector $\boldsymbol{\theta}$ represents the neural network parameters, i indexes a forward diffusion step, λ_i is a weighting constant that varies according to the noise level, $s^{\boldsymbol{\theta}}(\cdot, \cdot)$ represents the neural network to be optimized, and $p_i(\mathbf{x}_i|\mathbf{x}_0)$ represents the distribution of data sample $\mathbf{x}_0$ corrupted with noise up until diffusion step i.

Starting from this observation, the authors of [18] postulated that diffusion could be represented in continuous time by a pair of stochastic differential equations (SDEs), where a forward equation defines the forward diffusion process and a backward equation, dependent on the score function, defines the reverse diffusion process. Using this formalism, the corrupted version $\mathbf{x}(t)$ at continuous-time t of an initial data sample $\mathbf{x}(0) \sim p_0(\mathbf{x})$ can be obtained via the forward equation, and the initial data sample $\mathbf{x}(0)$ can be obtained from $\mathbf{x}(t)$ via the backward equation. In this setting, the terminal diffusion time T can be considered equal to one without loss of generality. Moreover, the terminal distribution $p_T(\mathbf{x})$ can have the general form $p_T(\mathbf{x}) = N(0, \sigma_{\max}^2 \mathbf{I})$, where $\sigma_{\max}$ denotes the maximum noise level of the forward process [19]. The continuous-time version of the loss in Eq. (1) is

$$L(\boldsymbol{\theta}) = \mathbb{E}_t \left[\lambda(t) \, \mathbb{E}_{f_{0,t}} \left[\left|\left| s^{\boldsymbol{\theta}}(\mathbf{x}(t), t) - \nabla_{\mathbf{x}(t)} \log p_t(\mathbf{x}(t)|\mathbf{x}(0)) \right|\right|^2 \right] \right], \tag{2}$$

where t is a randomly sampled time between the start and end of the forward diffusion process.

In [19], the authors give empirical arguments to choose

$$\mathrm{d}\mathbf{x} = \sqrt{2t}\,\mathrm{d}\mathbf{w} \tag{3}$$

as the forward SDE and

$$\mathrm{d}\mathbf{x} = -2t\nabla_{\mathbf{x}(t)} \log p_t(\mathbf{x}(t))\mathrm{d}t + \sqrt{2t}\,\mathrm{d}\mathbf{w} \tag{4}$$

as the backward SDE, where $\mathrm{d}\mathbf{w}$ is the differential of a standard Wiener process [20]. They also show that, for this specific choice of SDEs, $p_t(\mathbf{x}(t)|\mathbf{x}(0)) = \mathcal{N}(\mathbf{x}(0), t^2\mathbf{I})$, implying

$$\nabla_{\mathbf{x}(t)} \log p_t(\mathbf{x}(t)|\mathbf{x}(0)) = \frac{\mathbf{x}(0) - \mathbf{x}(t)}{t^2}. \tag{5}$$

With Eq. (5) in mind, the authors of [19] replaced the loss of Eq. (2) by the loss function

$$L(\boldsymbol{\theta}) = \mathbb{E}_t \left[\lambda(t)\, \mathbb{E}_{f_{0,t}} \left[\left|\left| \mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t) - \mathbf{x}(0) \right|\right|^2 \right] \right], \tag{6}$$

where $\mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t)$ is a neural network approximation for the data sample $\mathbf{x}(0)$ calculated from $\mathbf{x}(t)$. This version of the loss function comes from applying Eq. (5) to Eq. (2) and from adopting

$$s^{\boldsymbol{\theta}}(\mathbf{x}(t), t) = \frac{\mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t) - \mathbf{x}(t)}{t^2} \tag{7}$$

as the score approximation. Note that the denominators of Eqs. (5) and (7) were incorporated into $\lambda(t)$.

Inference with diffusion models within the framework of [18, 19] can be made with a numerical SDE solver [20, 19], given a score approximation. The diffusion model used for audio denoising in this work is based on [10], which follows the formalism of [19] and adopts Eqs. (3) and (4) for the forward and backward processes and Eq. (6) for the training loss. More details on training and inference with this model, including the choice of training weighting function $\lambda(t)$ and the SDE solving procedure for inference, are given in SEC. 3.

1.2 Diffusion Models for Inverse Problems

Diffusion models can be used for conditional sampling (i.e., sampling new data according to a constraint) with small modifications and no need to retrain the neural network approximating the score. Supposing that one wishes to sample from a distribution $p_0(\mathbf{x}(0)|\mathbf{y})$—with $\mathbf{y}$ being a vector representing the degraded audio—it is enough to replace $\nabla_{\mathbf{x}(t)} \log p_t(\mathbf{x}(t))$ in Eq. (4) by $\nabla_{\mathbf{x}(t)} \log p_t(\mathbf{x}(t)|\mathbf{y})$ [18].

In addition, applying Bayes' rule and the multiplication property of the logarithm, it is possible to write that

$$\begin{aligned}\nabla_{\mathbf{x}(t)} \log p_t(\mathbf{x}(t)|\mathbf{y}) = \nabla_{\mathbf{x}(t)} \log p_t(\mathbf{y}|\mathbf{x}(t)) \\ + \nabla_{\mathbf{x}(t)} \log p_t(\mathbf{x}(t)).\end{aligned} \tag{8}$$

Therefore, conditional sampling given a degraded sample is simply a matter of replacing the pure unconditional score by a sum of the unconditional score with a likelihood term, $\nabla_{\mathbf{x}}(t) \log p_t(\mathbf{y}|\mathbf{x}(t))$. As the unconditional score is already approximated by a pretrained neural network, the challenge of conditional sampling with diffusion is that of correctly estimating the likelihood term.

In this work, the likelihood term was estimated using diffusion posterior sampling (DPS) [21, 10], a method successfully applied with conditional diffusion for general inverse problems in the audio [10] and image [21] domains.[1] In DPS, $\nabla_{\mathbf{x}}(t) \log p_t(\mathbf{y}|\mathbf{x}(t))$ is given by

$$\nabla_{\mathbf{x}(t)} \log p_t(\mathbf{y}|\mathbf{x}(t)) = \xi(t) \nabla_{\mathbf{x}(t)} ||\mathbf{y} - \mathcal{A}(\mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t))||^2, \tag{9}$$

where $\xi(t)$ is a function of t chosen by the user, $\mathcal{A}(\cdot)$ is a model of the degradation to be restored, and the gradient on $\mathbf{x}(t)$ is calculated numerically. Intuitively, DPS is equivalent to assuming that $p_t(\mathbf{y}|\mathbf{x}(t))$ is Gaussian around a degraded version of the estimate of $\mathbf{x}(0)$; the term $\xi(t)$ is used to represent the variance of this Gaussian distribution.

In [10], the authors empirically show that, with the SDEs of [19], a reasonable choice for $\xi(t)$ is

$$\xi(t) = \frac{-\xi' \sqrt{N}}{t ||\nabla_{\mathbf{x}(t)} ||\mathbf{y} - \mathcal{A}(\mathbf{x}^{\boldsymbol{\theta}})||^2 ||^2}, \tag{10}$$

where N is the audio length in samples, and ξ' is a constant to be set by the user. Making this choice of $\xi(t)$ is advantageous because it forces the norm of the likelihood term to be $\frac{\xi' \sqrt{N}}{t}$. This makes conditioning stronger near the end of the reverse diffusion process due to the diminishing values of t. At the same time, it makes conditioning controllable through the choice of the parameter ξ'.

The experiments of [10] evaluate conditional sampling with diffusion in the tasks of bandwidth extension, gap filling, and audio declipping. In all three cases, conditional sampling with diffusion yields competitive results. Audio bandwidth extension with diffusion is evaluated objectively and subjectively for signals low-passed at 1 and 3 kHz, and the proposed method outperforms the models of [22] and [23]. Gap filling is subjectively evaluated in a test asking volunteers to assess the plausibility of the fills generated by different models, and conditional diffusion sampling obtains results comparable to [24] and [25]. For declipping, psychoacoustically inspired objective audio quality metrics show that the results of their diffusion model are comparable to the sparsity-based methods of [26] and [27].

2 PROPOSED METHOD

In this work, a heuristic adaptation of the framework of [10] is proposed to remove additive background noise from classical piano recordings. The idea is that, although an example of additive background noise is likely unique to a particular recording (e.g., 78 RPM noise in a specific record, tape hiss from a specific recorder), it is possible to use $\mathcal{A}(\mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t))$ to simulate its degradation effect using an inference set (i.e., a set of noise samples that are statistically similar to the background noise to be removed).

For this simulation to work, $\mathcal{A}(\mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t))$ can be defined as the addition of a randomly selected noise sample from the inference set to $\mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t)$. It was heuristically found

[1] In [10], DPS is called reconstruction guidance.

that randomly selecting a sample from the inference set at each reverse diffusion step allows conditioning the diffusion output with a specific type of additive degradation without targeting a particular noise example. The random selection of an inference set sample can intuitively be seen as a form of regularization, similar to the "noise regularization" procedure of [11]. Another possible interpretation is to understand the overall effect of $\mathcal{A}(\cdot)$ throughout reverse diffusion as similar to that of a generic noise of the same family as the one that needs to be removed.

More formally, $\mathcal{A}(\mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t))$ is defined here as

$$\mathcal{A}(\mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t)) = \mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t) + G\mathbf{n}, \tag{11}$$

where $\mathbf{n}$ is a unit power vector of the same size as $\mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t)$ representing a randomly selected, power-normalized sample from the inference set and G is a factor used to regulate the added noise power as explained further below. Adjustments were made in this implementation of $\mathcal{A}(\cdot)$ to ensure that $\mathcal{A}(\mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t))$ would result in a vector with values in $[-1, 1]$, avoiding overflow. The reader is invited to look at this project's GitHub repository[2] for further details.

To properly simulate additive background noise degradation with $\mathcal{A}(\mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t))$, it is important to ensure that $\mathbf{n}$ is added to $\mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t)$ with the same power as the noise in the degraded signal. Assuming $\mathbf{n}$ has unit power, G^2 is the power of the noise added to $\mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t)$, and therefore, G can be used to regulate the added noise power. To estimate G according to the background noise in the degraded signal, a minimum statistics approach similar to the one mentioned in [28] was used. A sliding window was used to calculate the power $P_{\text{noisy}}[n]$ of the degraded signal over time, and G was estimated heuristically as $\sqrt{\min P_{\text{noisy}}}$.

Intuitively, assuming that the sliding window is small enough, $\min P_{\text{noisy}}$ occurs in noise-only segments of the degraded audio, and therefore, G^2 approximates the noise power. However, it is important to note that the minimum operator necessarily introduces a downward bias. The sliding window should be short enough to capture noise-only segments of the target signal, but large enough so that these segments will be representative of the noise.

A final point that was addressed by the method proposed in this paper is block processing for conditional diffusion sampling. Diffusion models work from the first iteration with vectors of the desired output size. Since high-fidelity music has sampling rates of at least 44.1 kHz, the length of the audio signals restored by conditional sampling is limited to a few seconds. To overcome this, in all the examples in this work, restoration was applied to overlapping blocks of the degraded signals. The authors used 1-s blocks with 0.25-s overlap, selected using rectangular windows. The blocks were treated sequentially in all experiments, and the overlap-and-add algorithm [29, 30] was used to reconstruct the complete audio signals later. Hann windows of the same size as the rectangular analysis windows were used for synthesis.

[2]https://github.com/bvm810/diffusion-audio-restoration.

3 MODEL SETUP

The unconditional diffusion model in this work uses the CQT-Diff neural network architecture as an estimator for $\mathbf{x}^{\boldsymbol{\theta}}(\mathbf{x}(t), t)$, exactly as described in [10]. This architecture can be broadly understood as a U-Net [31] that is preceded by a differentiable constant-Q transform (CQT) [32] block and succeeded by a differentiable inverse CQT block.

Constant-Q time-frequency representations allow for logarithmic frequency resolution, which in turn enables representing bass and treble notes sharply while retaining good time resolution in the upper frequencies. Furthermore, constant-Q representations can be designed so that pitch shifting operations in the input audio correspond to simple translations, which allows exploring the full potential of translation-equivariant convolution operations in the backbone U-Net.

Through the use of CQT and inverse CQT blocks, the CQT-Diff architecture is able to explore a traditional U-Net architecture in the context of audio. Although U-Nets were originally developed for image tasks, the use of differentiable constant-Q time-frequency representations ensures proper adaptation for the audio domain due to the equivariance of pitch translation.

In [10], the CQT-Diff model is trained using classical piano recordings with 22.05-kHz sampling rates. As this sampling rate is half of the one used in CD quality recordings, their proposed model was retrained for this work. Training was carried out on high-quality recordings extracted from the MIDI and Audio Edited for Synchronous Tracks and Organization (MAESTRO) dataset [33], a collection of about 200 hours of classical piano recordings played by different performers on Yamaha Disklavier pianos. MAESTRO has recordings sampled at 44.1-kHz and 48-kHz rates, but for convenience, the whole training set was resampled to 44.1 kHz using the *sox* command line tool [34]. MAESTRO is originally divided in train, validation, and test splits, but they were all mixed for model training, as there was no superposition between MAESTRO and the test excerpts used in the experiments.

Following [19], the training employed the forward SDE of Eq. (3) and the loss of Eq. (6) using

$$\lambda(t) = \frac{\sigma_{\text{data}}^2 + t^2}{\sigma_{\text{data}}^2 t^2}, \tag{12}$$

where $\sigma_{\text{data}} = 0.057$ is the estimated data distribution variance. Additionally, following [10], the time t in the loss of Eq. (6) was distributed following

$$t = \left(\sigma_{\min}^{1/\rho} + \tau\left(\sigma_{\max}^{1/\rho} - \sigma_{\min}^{1/\rho}\right)\right)^{\rho}, \tag{13}$$

where $\sigma_{\max}$ and $\sigma_{\min}$ are the maximum and minimum noise levels of the diffusion process, and where $\tau \sim \mathcal{U}(0, 1)$. Intuitively, this choice of distribution for t emphasizes lower noise levels, which are responsible for adding detail to the model outputs.

This training procedure used $\sigma_{\min} = 10^{-6}$, $\sigma_{\max} = 10$, and $\rho = 10$. The model was trained using 1.5-s audio samples

with a batch size of one[3] for 2,304,000 iterations. The Adam optimizer was used with a learning rate of 2×10^{-4} and a learning rate decay factor of 0.8 every 60,000 iterations. The final network weights were averaged using an exponential moving average of rate 0.9999. Training took about 1 month with an NVIDIA GeForce RTX 2080 GPU and required 11 GB of VRAM. The model weights are available for download in the GitHub repository for this work.

For inference in all experiments, the heuristic second-order solver of [19] was applied to the backward SDE of Eq. (4) with parameters $S_{\text{tmin}} = 0$, $S_{\text{tmax}} = 50$, $S_{\text{noise}} = 1$, and $S_{\text{churn}} = 5$. The solver time steps were chosen following Eq. (13), with τ replaced by $\frac{i}{S-1}$, where $S = 140$ denotes the total number of solver time steps and $i = 0, \ldots, S-1$ indexes them. Using geometrically spaced time steps with the solver allows more steps (and less error) in lower noise levels, where more detail is needed [19]. Inference time steps were established using $\sigma_{\text{min}} = 10^{-6}$, $\sigma_{\text{max}} = 10$, and $\rho = 13$. Note that different values of σ_{min}, σ_{max}, and ρ can be used during training and inference, as long as the time steps remain within the distribution defined for t.

With the setup above, the inference time for a 5-s signal was around 20 min using an NVIDIA GeForce RTX 3090 GPU. Taking into account a single 1-s block, the processing time was around 170 s using the same hardware. For comparison, [6] infers on a 1-s audio signal in seconds, and [10] does so in around a minute. However, it should be taken into account that due to memory constraints, the authors were forced to load one inference set noise sample at a time during reverse diffusion. This generated a high number of input/output operations and had a considerably negative impact on processing time.

The gain of the degradation model G was estimated using the sliding window method mentioned in SEC. 2, with a window of 8,194 samples. DPS was used for conditional sampling in all experiments, and ξ' was set on a case-by-case basis, as explained in the sections describing the experiments.

Manual adjustment of ξ' is required for the proposed method, which is not necessarily a drawback. Occasionally, denoising can excessively remove high-frequency signal content, causing muffling. Adjusting ξ' allows the user to change the conditioning strength and consequently how much content is removed from the target signal during restoration. Therefore, manually controlling ξ' can allow users to choose their desired balance between noise removal and signal preservation.

In informal listening tests with artificially degraded recordings, it was found that the power of the background noise to be removed appears to be somewhat correlated with the ξ' values that were used in the most pleasant restored outputs. This can be investigated in future contributions so that a starting value for ξ' can be automatically determined prior to human-assisted fine-tuning.

[3]Due to memory constraints, the authors were unable to use longer audio samples or a larger batch size. This had a considerable impact on training time.

4 EXPERIMENTS

4.1 78 RPM Noise Removal

The proposed method was objectively and subjectively evaluated for the removal of 78 RPM disc noise. This denoising task is particularly interesting because 78 RPM noise is statistically nonstationary and also because its characteristic sound is created by a composition of common artifacts in historical recordings, such as clicks, thumps, and motor rumble. Despite the fact that some of these artifacts can be modeled as local degradations, the perceived result is that of a distributed noise.

4.1.1 Objective Evaluation

For objective testing, 16 excerpts from clean digital solo piano recordings were artificially degraded using 78 RPM disc noise samples. The noise samples were power-normalized, extended through repetition, and then added to the clean signals with various SNRs to create a set of noisy test recordings. This test set was then restored using diffusion conditional sampling and a benchmark method, and the quality improvement in the restored outputs was measured using Virtual Speech Quality Objective Listener (ViSQOL) Audio [35], an objective audio quality metric.

The clean excerpts had lengths between 9 and 23 s and were extracted from the *Gyorgi Ligeti – Works* [36], *American Classics – Samuel Barber* [37], and *Schumann – Piano Works* [38] albums. Ligeti's piano pieces were recorded by Pierre-Laurent Aimard in 1995, Barber's pieces by Daniel Pollack in 1995, and Schumann's pieces by Bernd Glemser in 1993. Their complete metadata can be found in the GitHub repository for this work.

The noise samples were taken from the gramophone noise dataset of [6], which consists of noise-only segments extracted from publicly available digitized 78 RPM recordings [39]. The noise dataset was divided into train, validation, and test sets, but only the test split was used to degrade the clean excerpts. Each of the 16 excerpts was degraded with SNRs of 10, 20, 30, and 40 dB, making a total of 64 test signals.

A retrained version of the neural denoiser of [6] was used as the benchmark model in this test. Retraining also used the strategy proposed in [6], using pairs of clean and artificially degraded signals. However, the degraded signals used to train the model of [6] had SNRs between 2 and 20 dB, while those used here had SNRs between 10 and 40 dB. Using higher SNRs in training prepared the benchmark to handle higher SNR test signals, making the comparison with diffusion denoising more reasonable. In addition, very low SNRs might not accurately represent late electric-era gramophone recordings, which make up a significant part of 78 RPM recordings.

Aside from this, the retraining procedure was similar to [6], employing the same clean recording [40] and 78 RPM noise datasets. The benchmark was trained for 2,400,000 iterations using 5-s audio segments and a batch size of one. The weights of the neural network were updated using the Adam optimizer with a learning rate of 10^{-4} and a decay

Table 1. Average ΔMOS for the signals artificially contaminated with 78 RPM noise.

Restoration Method	Average ΔMOS
Retrained Moliner et al. [6]	**2.267 ± 0.261**
Diffusion	1.952 ± 0.271

factor of 0.1 every 800,000 steps, while the rest of the training setup was exactly as described in [6].

The diffusion denoiser in this test used the train split of the gramophone noise dataset as the inference set for conditional sampling. The ξ' values used for DPS were set according to a heuristic based on the SNRs of the test signals. Ideally, choosing a custom value of ξ' through informal listening for each of the 64 test signals would be necessary. However, as this would be very time-consuming, the authors opted for using the heuristic of Eq. (14), in which S represents the SNR of the test signal in decibels.

$$\xi' = \begin{cases} 0.35, & \text{if } S < 20; \\ 1.4, & \text{if } 20 \leq S < 30; \\ 2.45, & \text{if } 30 \leq S < 40; \\ 3.5, & \text{if } S \geq 40. \end{cases} \tag{14}$$

It is important to note that this heuristic is likely suboptimal. As mentioned in SEC. 3, the most pleasant restored outputs appear to be obtained using ξ' related to the power of the noise to be removed from the recording. Since the SNR also depends on the power of the signal, using this rule to determine ξ' could result in using the same ξ' to restore signals where the noise power varies significantly. However, this heuristic proved to be useful in obtaining satisfactory results without performing many inference operations for a large set of test signals.

A final limitation of this experiment is the fact that ViSQOL Audio was developed to evaluate quality loss in compressed audio formats [35]. Audio restoration is outside the original scope of ViSQOL, and therefore, it could be biased in some sense in this domain (e.g., by favoring complete noise removal over signal preservation).

ViSQOL compares a degraded signal with a reference in order to calculate a mean opinion score (MOS) that ranges from 1.0 (worst, very annoying impairment) to 5.0 (best, imperceptible impairment). To measure the performance of each method, ΔMOS, the improvement in ViSQOL MOS after restoration, is calculated. Table 1 (best result in bold) shows the average ΔMOS with Gaussian 95% confidence intervals for the two methods.

In this test, the benchmark method was consistently better; in a one-tailed Student t test comparing the two methods, the null hypothesis that diffusion restoration had ΔMOS greater or equal to the benchmark was rejected with over 99% confidence. However, the considerable overlap in the ΔMOS ranges shows that diffusion denoising achieves results that are comparable to the benchmark. Furthermore, it is important to remember that ξ' was not optimized for the test signals, which means that the diffusion results could potentially be improved.

Qualitative (informal) listening of the restored excerpts appears to show that the proposed method allows more residual noise to appear in the restored outputs. In exchange, it also preserves more high-frequency content in the original signals. The benchmark method, on the other hand, completely removes the noise, at the cost of potentially eliminating musical content in the signals being restored.

4.1.2 Subjective Evaluation

In addition to the objective test, a subjective test was designed to evaluate the restoration performance of the proposed method in denoising historical 78 RPM recordings. A group of volunteers rated the quality of signals restored with diffusion conditional sampling and with the benchmark method of the previous test. The results were compared to assess the performance of the methods in a realistic restoration scenario.

This test used six excerpts of 23 s extracted from six different historical recordings. Five excerpts were extracted from tracks of the companion CD of *The Art of the Piano* piano encyclopedia [41], and one excerpt was taken from a recording of Friedrich Gulda's *Complete Decca Recordings* [42] collection. To the best of the authors' knowledge, no preprocessing had been applied to any of the signals before restoration.

The tracks from [41] chosen for this test were Eugene D'Albert's 1920 recording of Franz Lizst's "Au Bord d'une Source" (excerpt 1); Arthur De Greef's 1929 recording of Edvard Grieg's "Arietta" Op. 12 No. 1 (excerpt 2); Walter Gieseking's 1938 recording of Claude Debussy's "Mouvement," from the first book of "Images" (excerpt 3); Bela Bartok's 1945 recording of his own piece "Evening in Transylvania" (excerpt 4); and Ignaz Friedman's 1930 recording of Felix Mendelssohn's "Song Without Words" Op. 102 No. 5 (excerpt 5). Friedrich Gulda's 1949 recording of the first movement of Ludwig van Beethoven's Sonata No. 31 (excerpt 6) was selected from [42].

The choice of recordings was made to cover important technological innovations in recording techniques, and by listening to the excerpts, it is possible to see that the background noise gradually diminishes for more recent excerpts. Avoiding as much as possible the occurrence of nonadditive defects, such as hard-clipping, was also a selection criterion.

In the test, 13 volunteers answered six questions, one per excerpt, using the interface of [43]. In each question, the historical recording excerpt was presented as a reference and volunteers were asked to rate the quality of three restored test signals from zero to 100. An introductory text explaining that quality was to be evaluated both in terms of absence of background noise and integrity of the musical content in the signal was presented to all volunteers.

Two of the test signals were created using the proposed method with different values of ξ'; one of them used a fixed value ξ' of 0.35, while the other used ξ' customized through informal listening by the authors. The third signal was the benchmark for this test and was created using the same retrained neural denoiser as the objective test.

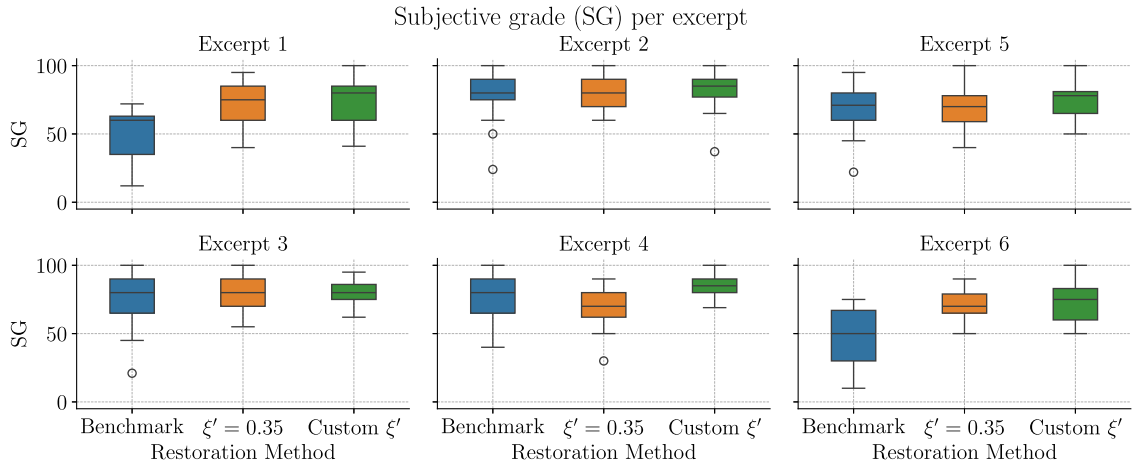


Fig. 1. Boxplots with the SGs of the six restored 78 RPM recordings, for the three methods.

Table 2. Average SGs for the restored 78 RPM recordings.

Restoration Method	Average SG
Benchmark	65.436 ± 5.257
$\xi' = 0.35$	72.974 ± 3.453
Custom ξ'	$\mathbf{77.833 \pm 3.257}$

Evaluating with customized values of ξ' mirrors real-life scenarios in which experimenting with various values of ξ' would be possible to achieve optimal restoration results. Custom $\xi' = 0.3$ was used for all excerpts taken from [41], and $\xi' = 1.4$ for Friedrich Gulda's 1949 recording. In all cases, the training split of the 78 RPM noise dataset of [6] was used as the inference set for conditional sampling.

The volunteers took the test in a dedicated listening room using Sennheiser HD265 linear headphones. The output volume of the testing equipment was defined in advance for each question and volunteers were instructed to respect it during the test. The individual grades given by each volunteer for each restored signal can be found in this work's GitHub page.

Table 2 (best result in bold) shows the average subjective grades (SGs) for the benchmark, $\xi' = 0.35$, and custom ξ' restored excerpts with 95% confidence intervals set using Gaussian distributions. One-tailed Student's *t* tests with null hypothesis that the mean benchmark results were greater or equal to each of the mean diffusion results were used to assess the differences between the benchmark and the two diffusion mean results; in both cases, the null hypothesis was rejected with p values below 1%. In addition to this, it is possible to see that there is no overlap in the confidence intervals of the benchmark and custom ξ' methods.

Fig. 1 shows the boxplots for the SGs of each of the six excerpts, in chronological order. It is possible to see that the benchmark method had a particularly poor performance in the first and last excerpts. In the case of excerpt 6, the results could be explained by the presence of a strong motor rumble, which made it difficult for the benchmark method to distinguish musical content in the bass region from the noise component of the signal. The output of the benchmark method in this case was noticeably noisier than the outputs of the other two methods. For excerpt 1, a possible explanation lies in the fact that the reference excerpt was very limited in bandwidth. In addition to removing the noise, the benchmark method also removed the (little) high-frequency content that was originally present in the recording, making it sound particularly muffled.

There are a few possible explanations for the different results in the objective and subjective tests. First, given the fact that the benchmark was trained with artificial data, a decrease in performance when restoring historical recordings was not entirely unexpected. Second, as mentioned earlier, the objective test might have underestimated the performance of the proposed method as a consequence of the suboptimal choice of ξ' and of possible biases in ViSQOL. Finally, the blind setup of the subjective test, without clean references, might have made the volunteers mistake small amounts of residual noise for high-frequency musical content, favoring the proposed method in the subjective test.

Despite these limitations, the results of the benchmark and the proposed method in both tests are quite close. This suggests that diffusion denoising can achieve results comparable to the state-of-the-art for 78 RPM noise removal.

4.2 Tape Hiss Removal

The proposed method was also objectively evaluated for the removal of tape hiss. Analog tapes dominated audio recording technology for much of the 20th century and are a significant part of global sound archives. Because of the importance of analog tape recordings, tape hiss removal in historical recordings is well studied, making it suitable for demonstrating the zero-shot nature of diffusion denoising on another kind of perceptually distributed noise.

4.2.1 Objective Evaluation

This experiment replicated the setup from SEC. 4.1.1 using analog tape hiss in place of 78 RPM noise to degrade the same 16 solo piano excerpts as before. Like in the previous experiment, ViSQOL Audio was chosen as the audio quality metric for this test.

The tape hiss dataset of [7] was used to additively degrade the clean recordings. This dataset consists of tape hiss snippets created by reproducing blank tapes in six different

(open reel and cassette) tape recorders at different speeds. Revox, Uher, and Technics tape recorders were used to play the blank tapes, and digitization was performed at a 44.1-kHz sampling rate. More details on the tape recorder models, their speeds, and the hiss dataset can be found in [7].

Similarly to SEC. 4.1.1, this experiment also degraded the clean excerpts with SNRs of 10, 20, 30, and 40 dB. This choice of SNR for the test signals was meant to extend the recording settings of the objective experiment of [7]. In their work, only 10-dB and 16-dB test signals (simulating adverse recording conditions) were used. All excerpts were corrupted with all SNR levels, resulting in 64 test signals in total, but only Uher noise samples were used during this procedure.

Unlike the test of SEC. 4.1.1, here the proposed method was evaluated against two different benchmarks. The first was a retrained version of the neural denoiser from [7], which is an adaptation of [6] for tape hiss removal. The second was an improved implementation of the Wiener filter, following the heuristics recommended in [4].

As in the test benchmark of SEC. 4.1.1, the neural denoiser of [7] was retrained here using higher SNRs than in its original implementation. The model was originally trained in [7] using artificially degraded signals with SNRs between 6 and 32 dB. Here, this range was changed to [10, 40] dB to cover all possible SNRs in the test signals.

Retraining used the dataset of [40] for the clean recordings and Revox snippets from the tape hiss dataset for the noise samples. The model was retrained for 340,000 iterations using 5-s audio segments and a batch size of one. It was optimized with the Adam optimizer using a learning rate of 10^{-4} and a decay factor of 0.1 every 50,000 iterations. The remainder of the training setup was exactly as described in [7].

The Wiener filter acts as a frequency-dependent gain based on an estimate of the noise power spectral density (PSD) in the signal being restored and might suppress low-amplitude frequency components of the noise while failing to attenuate high-amplitude ones if the noise PSD estimate is inaccurate. Unsuppressed components are typically spread across the spectrum and change randomly across frames, which makes the residual noise in Wiener-filtered musical signals acquire a tonal characteristic known as "musical noise" or "birdies."

To reduce the occurrence of such artifacts, this implementation of the Wiener filter was improved following [4] by including a lower bound on the Wiener gain, smoothed estimates of the PSD of the signal being restored, and an additional gain $\alpha \geq 1$ applied to the noise PSD. Applying a lower bound to the Wiener gain leaves a noise floor in the restored signal, reducing the tonal aspect of the residual noise. Smoothing out the PSD of the signal that is being restored makes the Wiener gain estimate more stable, which in turn reduces random tonal fluctuations in the residual noise. Finally, using $\alpha > 1$ overestimates the PSD of the noise to ensure its complete suppression, although it tends to also remove musical content from the signal.

Table 3. Average ΔMOS for the signals artificially contaminated with tape hiss.

Restoration Method	Average ΔMOS
Retrained Irigaray et al. [7]	**2.298 ± 0.223**
Wiener	1.781 ± 0.287
Diffusion	2.217 ± 0.239

For the restoration of the test signals with Wiener filtering, frames of 1,024 samples with 50% overlap were used. The lower bound of the Wiener gain was set to 0.05, and the PSD of the signal being restored was estimated using the median of the squared spectrum over five frames. Because tape hiss is a broadband signal, the noise PSD was considered constant for all frequencies and was estimated by the median of the last 256 frequency bins of the smoothed PSD of the test signal. All noisy excerpts were restored with six different values of α ($\{1, 2, 3, 4, 5, 6\}$), but the aggregate results for this experiment only consider the best α for each test signal, according to ViSQOL.

In this experiment, only Revox samples from the hiss dataset were used in the inference set for diffusion denoising. As in SEC. 4.1.1, the heuristic of Eq. (14) was used to determine the values of ξ' for the test signals. Once again, it is important to remember that this rule was used for convenience and could negatively impact the results of the proposed method.

Table 3 (best result in bold) shows the average ΔMOS with Gaussian 95% confidence intervals for the three methods. Using pairwise one-tailed Student t tests, the null hypotheses that diffusion denoising and Wiener filtering had ΔMOS results greater than or equal to the first benchmark were rejected with $p = 0.026$ and $p < 0.01$, respectively, while the hypothesis that Wiener filtering performed better than the proposed method was rejected with confidence above 99%.

There is significant overlap in the confidence intervals of the retrained denoiser of [7] and the proposed method, indicating that the—zero-shot—proposed method surpasses traditional tape hiss removal techniques and performs comparably, though slightly less effectively, to the current state-of-the-art neural method. This should be considered along with the suboptimal heuristic for ξ' and the potential limitations of ViSQOL as an evaluation metric for restoration.

Once again qualitatively listening to the restored signals, it appears that the proposed method has a tendency to allow more residual noise while preserving better the musical content in the test signals. In contrast, the method of [7] (based on [6]) appears to remove more hiss at the cost of potentially removing musical content from the original signal. This analysis is consistent with what was observed informally in the outputs of both 78 RPM tests.

5 CONCLUSION

This paper investigated a zero-shot approach for background noise removal inspired by diffusion models. The

concept of the method was proven for digitized classical solo piano recordings in three experiments evaluating the restoration of recordings degraded by two distinct types of perceptually distributed noise.

The proposed method uses the continuous-time SDE formulation of diffusion models and DPS to remove additive background noise. A degradation model $\mathcal{A}(\cdot)$ randomly selects noise samples from an inference set in each reverse diffusion step, simulating specific additive degradations during conditional sampling. By changing the inference set, the same unconditional model can be used to remove different types of noise without retraining.

Diffusion denoising was evaluated in the removal of 78 RPM disc noise and analog tape hiss. In both cases, objective experiments comparing the proposed method with state-of-the-art benchmarks showed that diffusion denoising was competitive, despite obtaining slightly inferior results. For removal of 78 RPM noise, an additional subjective test was also performed using historical gramophone recordings, and the proposed method outperformed the benchmark.

The diffusion denoising results in the objective experiments may have been underestimated due to the chosen heuristic for ξ' and potential limitations of the audio quality metric. At the same time, the proposed method might have gained an advantage from the absence of clean references in the subjective test with historical recordings. Nevertheless, despite these limitations across the three experiments, the similar performance of the benchmarks and the proposed method indicates that it is comparable to the current state-of-the-art in different denoising tasks.

Diffusion denoising was adapted for all experiments only through the selection of a new inference set, without any retraining of the base unconditional model. This is an advantage over current neural denoisers, which require some strategy to artificially create pairs of clean and degraded signals.

Future improvements to this method might be made both in the conditional sampling procedure and in the inference set used for restoration. Using automatic algorithms based on noise power estimates to determine a starting value for ξ' is likely to improve the user experience. Similarly, a part of the fine-tuning procedure for ξ' could be performed using automatic adaptive strategies, such as that in [44]. In addition to this, using augmentation strategies for the inference set, such as enhancing it with noise-only excerpts of the target signal, can potentially ameliorate the degradation model $\mathcal{A}(\cdot)$ and consequently improve the overall restoration results.

6 ACKNOWLEDGMENT

This study was financed in part by the Coordination of Superior Level Staff Improvement (CAPES) - Finance Code 001 and the National Council for Scientific and Technological Development (CNPq), under Grant 311146/2021-0.

7 REFERENCES

[1] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications* (MIT Press, Cambridge, MA, 1949). https://doi.org/10.7551/mitpress/2946.001.0001.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum-Mean Square Error Log Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 33, no. 2, pp. 443–445 (1985 Apr.). https://doi.org/10.1109/TASSP.1985.1164550.

[3] L. W. P. Biscainho, F. P. Freeland, P. A. A. Esquef, and P. S. R. Diniz, "Wavelet Shrinkage Denoising Applied to Real Audio Signals Under Perceptual Evaluation," in *Proceedings of the European Signal Processing Conf. (EUSIPCO)*, pp. 2061–2064 (Tampere, Finland) (2000 Sep.).

[4] S. J. Godsill and P. J. Rayner, *Digital Audio Restoration* (Springer London, Cambridge, UK, 1998). https://doi.org/10.1007/978-1-4471-1561-8.

[5] Y. Li, B. Gfeller, M. Tagliasacchi, and D. Roblek, "Learning to Denoise Historical Music," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 504–511 (Montreal, Canada) (2020 Oct.).

[6] E. Moliner and V. Välimäki, "A Two-Stage U-Net for High-Fidelity Denoising of Historical Recordings," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 841–845 (Singapore) (2022 May). https://doi.org/10.1109/ICASSP43922.2022.9746977.

[7] I. Irigaray, M. Rocamora, and L. W. P. Biscainho, "Noise Reduction in Analog Tape Audio Recordings with Deep Learning Models," in *Proceedings of the AES International Conference on Audio Archiving, Preservation and Restoration* (2023 Jun.), paper 1.

[8] 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.

[9] R. Sawata, N. Murata, Y. Takida, et al., "Diffiner: A Versatile Diffusion-Based Generative Refiner for Speech Enhancement," in *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3824–3828 (Dublin, Ireland) (2023 Aug.). http://doi.org/10.21437/Interspeech.2023-1547.

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

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

[12] J.-M. Lemercier, J. Richter, S. Welker, et al., "Diffusion Models for Audio Restoration: A Review," *IEEE*

Signal Process. Mag., vol. 41, no. 6, pp. 72–84 (2025 Jan.). https://doi.org/10.1109/MSP.2024.3445871.

[13] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep Unsupervised Learning Using Nonequilibrium Thermodynamics," in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2256–2265 (Lille, France) (2015 Jul.).

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Proceedings of the Advances in Neural Information Processing Systems Conference (NeurIPS)*, pp. 6840–6851 (Montreal, Canada) (2020 Dec.).

[15] Y. Song and S. Ermon, "Generative Modeling by Estimating Gradients of the Data Distribution," in *Proceedings of the Advances in Neural Information Processing Systems Conference (NeurIPS)*, pp. 11886–11898 (Vancouver, Canada) (2019 Dec.).

[16] R. Neal, "MCMC Using Hamiltonian Dynamics," in S. Brooks, A. Gelman, and G. Jones (Eds.), *Handbook of Markov Chain Monte Carlo*, pp. 113–162 (CRC Press, Boca Raton, FL, 2011). http://doi.org/10.1201/b10905-6.

[17] P. Vincent, "A Connection Between Score Matching and Denoising Autoencoders," *Neural Comput.*, vol. 23, no. 7, pp. 1661–1674 (2011 Jul.). https://doi.org/10.1162/NECO_a_00142.

[18] Y. Song, J. Sohl-Dickstein, D. Kingma, et al., "Score-Based Generative Modeling Through Stochastic Differential Equations," in *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–36 (Vienna, Austria) (2021 May).

[19] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the Design Space of Diffusion-Based Generative Models," in *Proceedings of the Advances in Neural Information Processing Systems Conference (NeurIPS)*, pp. 26565–26577 (New Orleans, LA) (2022 Dec.).

[20] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations* (Cambridge University Press, Cambridge, UK, 2019). https://doi.org/10.1017/9781108186735.

[21] 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 International Conference on Learning Representations (ICLR)*, pp. 1–30 (Kigali, Rwanda) (2023 May).

[22] E. Moliner and V. Välimäki, "BEHM-GAN: Bandwidth Extension of Historical Music Using Generative Adversarial Networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 943–956 (2023 Jul.). https://doi.org/10.1109/TASLP.2022.3190726.

[23] K. Goel, A. Gu, C. Donahue, and C. Re, "It's Raw! Audio Generation With State-Space Models," in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 7616–7633 (Honolulu, HI) (2022 Jul.).

[24] A. Marafioti, P. Majdak, N. Holighaus, and N. Perraudin, "GACELA: A Generative Adversarial Context Encoder for Long Audio Inpainting of Music," *IEEE J. Sel. Top. Signal Process.*, vol. 15, no. 1, pp. 120–131 (2021 Nov.). https://doi.org/10.1109/JSTSP.2020.3037506.

[25] G. Greshler, T. Shaham, and T. Michaeli, "Catch-A-Waveform: Learning to Generate Audio From a Single Short Example," in *Proceedings of the Advances in Neural Information Processing Systems Conference (NeurIPS)*, pp. 20916–20928 (Online) (2021 Dec.).

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

[27] 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 (ICASSP)*, pp. 1577–1581 (Florence, Italy) (2014 May). https://doi.org/10.1109/ICASSP.2014.6853863.

[28] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512 (2001 Jul.). https://doi.org/10.1109/89.928915.

[29] P. S. R. Diniz, E. A. B. da Silva, and S. L. Netto, *Digital Signal Processing: System Analysis and Design* (Cambridge University Press, Rio de Janeiro, Brazil, 2010), 2nd ed. https://doi.org/10.1017/CBO9780511781667.

[30] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards* (Springer, Palo Alto, CA, 2003). https://doi.org/10.1007/978-1-4615-0327-9.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention Society Conf. (MICCAI)*, pp. 234–241 (Munich, Germany) (2015 Oct.). https://doi.org/10.1007/978-3-319-24574-4_28.

[32] N. Holighaus, M. Dörfler, G. A. Velasco, and T. Grill, "A Framework for Invertible, Real-Time Constant-Q Transforms," *IEEE Trans. Audio, Speech, Language Processing*, vol. 21, no. 4, pp. 775–785 (2013 Dec.). https://doi.org/10.1109/TASL.2012.2234114.

[33] C. Hawthorne, A. Stasyuk, A. Roberts, et al., "Enabling Factorized Piano Music Modeling and Generation With the MAESTRO Dataset," in *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–12 (New Orleans, LA) (2019 May).

[34] C. Bagwell, "SoX Arch Linux Man Page," https://man.archlinux.org/man/sox.1.en (accessed Oct. 14, 2024).

[35] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric," in *Proceedings of the International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6 (Athlone, Ireland) (2020 May). https://doi.org/10.1109/QoMEX48832.2020.9123150.

[36] G. Ligeti, *György Ligeti – Works*, P.-L. Aimard, Sony Classical (1995). https://www.discogs.com/release/3226060-Gy.

[37] S. Barber, *American Classics – Samuel Barber – Solo Piano Music*, D. Pollack, Naxos Mu-

sic Group (1999 Dec.). https://www.naxos.com/CatalogueDetail/?id=8.559015.

[38] R. Schumann, *Schumann – Piano Works*, B. Glemser, Naxos Music Group (1994 Apr.). https://www.naxos.com/CatalogueDetail/?id=8.550715.

[39] Internet Archive, "The Great 78 Project," https://great78.archive.org/ (accessed Oct. 14, 2024).

[40] J. Thickstun, Z. Harchaoui, and S. Kakade, "Learning Features of Music From Scratch," in *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–14 (Toulon, France) (2017 Apr.).

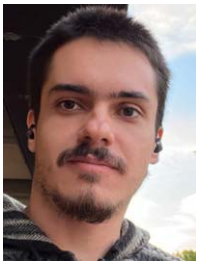
[41] D. Dubal, *The Art of the Piano: Its Performers, Literature, and Recordings Revised* (Amadeus Press, Milwaukee, WI, 2005), 3rd ed.

[42] F. Gulda, *Friedrich Gulda: Complete Decca Recordings*, Decca Classics (2021 Jul.). https://www.deccaclassics.com/en/catalogue/products/friedrich-gulda-complete-decca-recordings-12420.

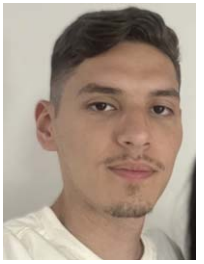
[43] P. H. L. Leite, "Audio Subjective Tests Web Interface," https://github.com/pedrohlopes/web-audio-subjective-tests (accessed Oct. 14, 2024).

[44] B. Nortier, M. Sadeghi, and R. Serizel, "Unsupervised Speech Enhancement With Diffusion-Based Generative Models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12481–12485 (Seoul, South Korea) (2024 Apr.). https://doi.org/10.1109/ICASSP48485.2024.10447736.

THE AUTHORS


Bernardo V. Miranda


Ignacio Irigaray


Rafael A. Deslandes


Luiz W. P. Biscainho

Bernardo Vieira de Miranda was born in 1996 in Curitiba, Brazil. He holds an M.Sc. in Electrical Engineering from the Electrical Engineering Program of the Alberto Luiz Coimbra Institute of Graduate Studies and Research in Engineering of the Federal University of Rio de Janeiro (UFRJ) (2024), with an emphasis on signal processing. His research interests include machine learning, statistics, and applications of generative models for audio. He also earned a Diplôme d'Ingénieur from CentraleSupélec (2022) and a B.Sc. in Electronic Engineering from the Department of Electronic and Computer Engineering of the Polytechnic School of the Federal University of Rio de Janeiro (UFRJ) (2021). Bernardo has served as a teaching assistant for courses in signal processing and probability, and he has previous professional experience as a software engineer.

•

Ignacio Irigaray was born in Montevideo, Uruguay, in 1981. He received his B.Sc. (2005) and M.Sc. (2014) in Electrical Engineering from the Universidad de la República, Uruguay. He is currently pursuing a Ph.D. in signal processing, specializing in audio restoration. Since 2004, he has been an assistant professor in the Signal Processing Department at the Universidad de la República. His research focuses on the processing of music signals and the preservation of musical heritage. As a musician, Ignacio plays both guitar and bandoneon. He currently performs with the ensemble Cuarteto Colibriyo (https://colibriyo.uy/), which specializes in milongas, valses, and tangos. From 2011 to 2021, he played with the folkloric group Los Extranjeros (https://losextranjeros-uy.bandcamp.com/).

•

Rafael Antonioli Deslandes, born in Rio de Janeiro in 2002, completed his secondary education at the Fernando Rodrigues da Silveira Application Institute (CAp-UERJ). He is currently pursuing a B.Sc. degree in Electronic and Computer Engineering at the Department of Electronic and Computer Engineering of the Polytechnic School of the Federal University of Rio de Janeiro (UFRJ), where he also serves as a teaching assistant for the department's Signals and Systems course. His undergraduate research focuses on signal processing, with a particular emphasis on automatic evaluation of audio quality.

•

Luiz W. P. Biscainho was born in Rio de Janeiro, Brazil, in 1962. He received an Electronics Engineering degree (magna cum laude) from the Engineering School (EE), now Polytechnic School (POLI) of the Federal University of Rio de Janeiro (UFRJ), Brazil, in 1985, and M.Sc. and D.Sc. degrees in Electrical Engineering from the Alberto Luiz Coimbra Institute of Graduate Studies and Research in Engineering (COPPE) at UFRJ in 1990 and 2000, respectively. Having worked in the telecommunication industry between 1985 and 1993, Dr. Biscainho is now Associate Professor at the Department of Electronic and Computer Engineering of POLI and COPPE Electrical Engineering Program at UFRJ. His research area is digital audio processing. He is currently a member of the Institute of Electrical and Electronics Engineers (IEEE), AES, Brazilian Telecommunications Society (SBrT), and Brazilian Computer Society (SBC).