ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095637

SOLVING AUDIO INVERSE PROBLEMS WITH A DIFFUSION MODEL

Eloi Moliner[1] *Jaakko Lehtinen*[2] *Vesa Välimäki*[1*]

[1] Acoustics Lab, Dept. of Information & Communications Eng. [2] Dept. of Computer Science
Aalto University, Espoo, Finland
eloi.moliner@aalto.fi

ABSTRACT

This paper presents CQT-Diff, a data-driven generative audio model that can, once trained, be used for solving various different audio inverse problems in a problem-agnostic setting. CQT-Diff is a neural diffusion model with an architecture that is carefully constructed to exploit pitch-equivariant symmetries in music. This is achieved by preconditioning the model with an invertible Constant-Q Transform (CQT), whose logarithmically-spaced frequency axis represents pitch equivariance as translation equivariance. The proposed method is evaluated with solo piano music, using objective and subjective metrics in three different and varied tasks: audio bandwidth extension, inpainting, and declipping. The results show that CQT-Diff outperforms the compared baselines and ablations in audio bandwidth extension and, without retraining, delivers competitive performance against modern baselines in audio inpainting and declipping. This work represents the first diffusion-based general framework for solving inverse problems in audio processing.

***Index Terms*—** Audio systems, deep learning, inverse problems, signal restoration.

1. INTRODUCTION

Audio restoration tasks, such as bandwidth extension, inpainting, and declipping, are inverse problems with the aim to restore the signal from observations that have suffered a known type of degradation. These problems are ill-posed in the sense that several different restorations may be equally plausible. Algorithms are typically constructed for a particular type of degradation using domain knowledge, such as signal sparsity [1, 2], or the low-rankness of the power spectrum [3]. Data-driven generative models have also been recently proposed [4, 5]. A common shortcoming is that an algorithm engineered for one type of degradation is typically not useful for others.

In this work, we explore whether a single pre-trained generative audio model could be useful for different restoration tasks *without knowledge of which types of degradations it will be applied to at inference time*. We base our work on diffusion models [6, 7], a family of generative models that have shown outstanding performance in different modalities, such as image [8], video [9], speech [10, 11], and audio generation [12, 13]. Particularly relevant to us is that unconditional (purely generative) diffusion models have proven useful also in conditional restoration settings where parts of the result are pre-determined, particularly in the image domain [14, 15, 16]. Crucially, as these approaches do not require paired data for training and the degradation model is only used during inference, the models adapt to a new problem without retraining. To the best of our knowledge, analogous methods have not been employed in audio audio restoration. This paper addresses this research gap.

Training an unconditional audio diffusion model is the first step towards this objective. We argue that this is overly challenging unless suitable inductive biases are built in to the model. While unconditional audio generation has been demonstrated using a raw audio waveform representation, the experiments are limited to small speech datasets [10, 17], or non-tonal audio [12, 18], with performance expected to decrease for more challenging datasets. Building diffusion models using mel-spectrograms or magnitude spectrograms [13] leads to a more expressive generative performance. However, the lack of invertibility of these transforms blocks the required conditioning for solving inverse problems. We design our diffusion model in the time domain, allowing for maximum flexibility, and use an invertible time-frequency transform as a preprocessing step.

Choosing the right transform plays a critical role in the performance of the model. One option is to use a Short-Time Fourier Transform (STFT) and apply a neural network based on 2-D convolutions, adapted from image generation. While this approach succeeded for speech [19], we argue that, contrary to image processing, it is non-ideal for signals with strong harmonics (music), as translation equivariance of harmonics does not hold true in linear frequency. The Constant-Q Transform (CQT), in which the frequency axis is logarithmically spaced, is a possible solution to this problem. Since pitch-shifting corresponds to translation in the CQT spectrogram, applying the convolutional operator is now appropriate. The CQT is widely used in Music Information Retrieval [20], but, despite some exceptions [21], is not common in audio generation tasks.

We first show that CQT preconditioning yields a high-quality diffusion model for audio. We present a versatile framework called CQT-Diff for solving audio restoration problems, from which we evaluate three: bandwidth extension, audio inpainting, and declipping. Pitted against specialized algorithms, our model yields state-of-the art or strongly competitive performance in listening tests.

2. CQT-DIFF FRAMEWORK

2.1. Score-based Generative Modeling

Consider a diffusion process where data $\boldsymbol{x}_0 \sim p_{\text{data}}$ is progressively diffused into Gaussian noise $\boldsymbol{x}_{\tau_{\max}} \sim \mathcal{N}(\mathbf{0}, \sigma_{\max}^2 \mathbf{I})$ over time[1] τ by applying a perturbation kernel $p_t(\boldsymbol{x}_\tau) = \mathcal{N}(\boldsymbol{x}_\tau; \boldsymbol{x}_0, \sigma_\tau^2 \mathbf{I})$. Diffusion models, also known as score-based models, generate data samples by reversing the aforementioned diffusion process. Specifically, diffusion models estimate the gradient of the log probability density with respect to data $\nabla_{\boldsymbol{x}_\tau} \log p_\tau(\boldsymbol{x}_\tau)$, known as the *score* function.

*This research is part of the activities of the Nordic Sound and Music Computing Network—NordicSMC (NordForsk project no. 86892).

[1]The "diffusion time" τ must not be confused with the "audio time" t.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on July 12,2023 at 00:48:46 UTC from IEEE Xplore. Restrictions apply.

Following [22], the reverse diffusion process can be defined by the following Stochastic Differential Equation (SDE):

$$\mathrm{d}\boldsymbol{x} = -\sigma_\tau(\beta(\tau)\sigma_\tau + \dot{\sigma}_\tau)\ \nabla_{\boldsymbol{x}} \log p_\tau(\boldsymbol{x}_\tau)\mathrm{d}\tau + \sqrt{2\beta(\tau)}\sigma_\tau\ d\mathrm{w}, \tag{1}$$

where $\mathrm{d}\tau$ is an infinitesimal negative timestep, $d\mathrm{w}$ is the standard Wiener process when time flows backwards, $\dot{\sigma}_\tau$ is the time derivative of the noise variance schedule, and $\beta(\tau)$ is, intuitively, a function that defines the amount of stochasticity added into the process. Note that stochasticity is, in fact, not needed and (1) could be defined as an ordinary differential equation [7, 22]. However, stochasticity has a role in correcting approximation errors during sampling [22] and is empirically found to improve the generation quality in our model.

The score is intractable but can be approximated with a deep neural network $s_\theta(\boldsymbol{x}_\tau, \sigma_\tau) \simeq \nabla_{\boldsymbol{x}_\tau} \log p_\tau(\boldsymbol{x}_\tau; \sigma_\tau)$. The parameterized score can be written as $s_\theta(\boldsymbol{x}_\tau, \sigma_\tau) = (D_\theta(\boldsymbol{x}_\tau, \sigma_\tau) - \boldsymbol{x}_\tau)/\sigma_\tau^2$, where $D_\theta(\boldsymbol{x}_\tau, \sigma_\tau)$ is a denoiser to be trained with the L2 objective:

$$\mathbb{E}_{\boldsymbol{x_0}\sim p_{\text{data}}, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left[\lambda(\sigma_\tau)\|D_\theta(\boldsymbol{x_0} + \sigma_\tau\boldsymbol{\epsilon}, \sigma_\tau) - \boldsymbol{x}\|_2^2\right], \tag{2}$$

where $\lambda(\sigma_\tau)$ is a weighting function. Once the model has been trained, one can generate new data instances by substituting the estimated score $s_\theta(\boldsymbol{x}_\tau, \sigma_\tau)$ in (1), and solving the SDE with a numerical solver of choice.

2.2. Diffusion Models for Inverse Problems

Consider the goal of retrieving an unknown audio signal $\boldsymbol{x_0}$ from a set of measurements or observations $\boldsymbol{y} = \mathcal{A}(\boldsymbol{x_0}) + \boldsymbol{\epsilon}$, where $\mathcal{A}(\cdot)$ is a known and differentiable degradation function, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_y^2\boldsymbol{I})$ is measurement noise. The objective would be to sample from the posterior distribution given the observations $p(\boldsymbol{x}|\boldsymbol{y})$. In the context of a diffusion model, this would require substituting the score in (1) for the conditional score $\nabla_{\boldsymbol{x}_\tau} \log p_\tau(\boldsymbol{x}_\tau|\boldsymbol{y})$.

One solution is to train a score model $s_\theta(\boldsymbol{x}_\tau, \boldsymbol{y}, \tau)$ that depends explicitly on the observations [23, 11, 8]. However, these approaches lack versatility as they require training with paired data and cannot effectively adapt to unseen degradations. Instead, applying Bayes, the posterior factorizes as $p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})$, which leads to

$$\nabla_{\boldsymbol{x}_\tau} \log p_\tau(\boldsymbol{x}_\tau|\boldsymbol{y}) = \nabla_{\boldsymbol{x}} \log p_\tau(\boldsymbol{x}_\tau) + \nabla_{\boldsymbol{x}_\tau} \log p_\tau(\boldsymbol{y}|\boldsymbol{x}_\tau). \tag{3}$$

Then, an unconditional diffusion model can be used as a generative prior, while the conditioning is applied by some other means. In some cases, the likelihood can be indirectly applied by ensuring data consistency during sampling, replacing parts of the internal predictions $\hat{\boldsymbol{x}}_0 = D_\theta(\boldsymbol{x}_\tau, \sigma_\tau)$ with the measurements $\boldsymbol{y}$ during every discretized step [7, 14, 24, 25]. We refer to this strategy as the *data consistency* (DC) method. Note that this method can only be used in cases where an explicit replacement is available. Degradations such as distortions or nonlinear-phase filtering cannot be treated by replacing parts of the signal in the time domain.

Recent works [15, 16, 9] demonstrated that the aforementioned data consistency trick produces suboptimal results, so they proposed an alternative strategy defining the second term of (3) as

$$\nabla_{\boldsymbol{x}_\tau} \log p_\tau(\boldsymbol{y}|\boldsymbol{x}_\tau) \simeq -\xi(\tau)\ \nabla_{\boldsymbol{x}_\tau} \|\boldsymbol{y} - \mathcal{A}(D_\theta(\boldsymbol{x}_\tau, \sigma_\tau))\|^2. \tag{4}$$

This solution derives from approximating the likelihood as a Gaussian distribution of the form $p_t(\boldsymbol{y}|\boldsymbol{x}_\tau) \sim \mathcal{N}(\mathcal{A}(D_\theta(\boldsymbol{x}_\tau, \sigma_\tau)), \sigma_y^2\mathbf{I})$. Following [9], we refer to this method as *reconstruction guidance* (RG). The gradient operator $\nabla_{\boldsymbol{x}_\tau}$ requires backpropagating through the degradation function $\mathcal{A}(\cdot)$, and through the neural network of the denoiser $D_\theta(\boldsymbol{x}_\tau, \sigma_\tau)$. The scaling function $\xi(\tau)$ defines the amount of guidance that is applied during the reverse diffusion process. Inspired by [26], we empirically found that a robust parameterization of the scaling function is $\xi(\tau) = \xi'\sqrt{N}/(\tau\|G\|^2)$, where $G = \nabla_{\boldsymbol{x}_\tau}\|y - \mathcal{A}(D_\theta(\boldsymbol{x}_\tau, \sigma_\tau))\|^2$ are the guidance gradients and N is the audio segment size in samples. Then, choosing $\xi' = 0$ would lead to an unconditional sampler, but selecting too large a value for ξ' would often result with degenerated solutions. Note that RG and DC can also be combined. However, if the observations are noisy and, as a consequence, unreliable, DC cannot be trivially applied.

2.3. Data Representation and Preprocessing

To minimize error propagation during sampling, we apply a similar preconditioning as Karras *et al.* [22]. The denoiser is defined as

$$D_\theta(\boldsymbol{x}_\tau, \sigma_\tau) = c_{\text{skip}}(\sigma_\tau)\boldsymbol{x}_\tau + c_{\text{out}}(\sigma_\tau)F_\theta(c_{\text{in}}(\sigma_\tau)\boldsymbol{x}_\tau, \tfrac{\ln(\sigma_\tau)}{4}), \tag{5}$$

where $c_{\text{skip}}(\sigma_\tau)$, $c_{\text{out}}(\sigma_\tau)$, and $c_{\text{in}}(\sigma_\tau)$ are scaling parameters optimized in such a way that the input and the output of F_θ have always close-to-unit variance[2]. Following the same approach, the loss weighting in (2) is defined as $\lambda(\sigma_\tau) = 1/c_{\text{out}}(\sigma_\tau)^2$.

Throughout this paper, $\boldsymbol{x}$ and $\boldsymbol{y}$ refer to signals in the raw waveform domain. However, in our early exploration, we found that using the time-domain waveform representation resulted in unsatisfactory results, especially for music bandwidth extension, which requires preserving the harmonic structure. Instead, the data is transformed using an invertible CQT (ICQT), which is applied as a data preprocessing step inside F_θ. As stated in Sec. 1, the motivation for using the CQT is to explore the partial pitch-equivariance symmetry that audio data exhibits in the log-spaced frequency range. Thus, F_θ is composed as $F_\theta = \text{ICQT} \circ F'_\theta \circ \text{CQT}$, where F'_θ refers to the raw neural network layers operating in the transform domain. Using this approach, the gradient is backpropagated through the forward CQT and its inverse, or ICQT, and, thus, the chosen CQT implementation must be both invertible and differentiable.

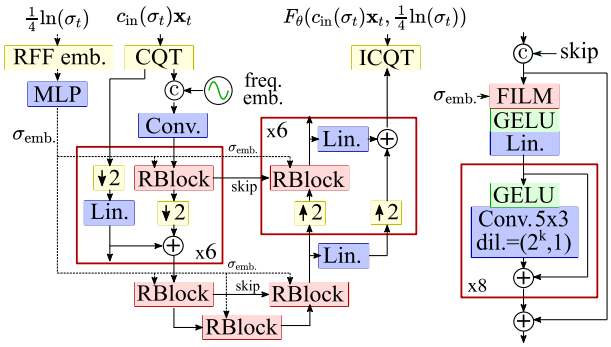
We use the CQT implementation based on non-stationary Gabor frames [27], which guarantees perfect reconstruction and is differentiable, since it is a linear transform implemented by applying FFT-processing. If set up for minimum redundancy, this transform yields a non-uniform sampling grid where the time resolution increases with frequency. Alternatively, we use a rasterized CQT representation where the hop sizes for all frequencies are set to the minimum, hence keeping the time resolution constant. This solution allows CQT processing as a 2D matrix at the cost of adding redundancy to the data. In our experiments, we use a frequency resolution of 64 bins per octave, and a span of seven octaves.

2.4. Architecture Design

Next, we summarize how CQT-Diff fits audio data in the CQT-domain using the neural network architecture shown in Fig. 1. The architecture is a U-Net where only the time resolution is downscaled after each subsequent residual block (*RBlock*), while the frequency resolution is maintained. The non-local dependencies in the frequency domain are tackled with stacks of exponentially-increasing dilated convolutions without bias, accompanied with GELU nonlinearities and residual connections. Inspired by [7], we use anti-aliased downsampling operators and a progressive growing structure for the encoder and decoder sides [28]. The diffusion step is shown to the network as $\frac{1}{4}\ln(\sigma_t)$, and is encoded using Random Fourier Features

[2]These parameters are defined in Table 1 in [22] and depend on the standard deviation of the dataset, which needs to be estimated beforehand.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on July 12,2023 at 00:48:46 UTC from IEEE Xplore. Restrictions apply.



(a) CQT-U-Net architecture (b) RBlock

Fig. 1: Diagram of the neural network architecture adapted for the CQT domain. The design follows a U-Net structure, where only the time resolution is modified. Each intermediate *RBlock* contains a stack of dilated convolutions in the frequency axis.

(RFF) [29] and a multi-layer perceptron. The embeddings are applied as a conditioner in every *RBlock* using FiLM modulations [30]. We concatenate RFF-based absolute frequency positional embeddings to the CQT-transformed input, meant to propagate absolute frequency information in the early layers of the network.

3. EXPERIMENTS AND RESULTS

This section summarizes our experiments on three different audio inverse problems: bandwidth extension, inpainting, and declipping. Audio examples are included in the companion webpage[3], and reproducible code for all the experiments can be found in the official repository[4], to which we refer for further implementation details.

We train and test our model on the MAESTRO dataset [31], which contains 200 hours of classical solo piano recordings. For ease of use, all tracks are resampled to 22.05 kHz. For sampling, we use the second-order stochastic sampler from [22] (Algorithm 2), with the stochasticity parameter $S_{\text{churn}} = 5$. Also following [22], we assign the noise variance as $\sigma_t = t$, with t distributed according to:

$$\tau_{i<T} = \left(\sigma_{\max}^{\frac{1}{\rho}} + \frac{i}{T-1}\left(\sigma_{\min}^{\frac{1}{\rho}} - \sigma_{\max}^{\frac{1}{\rho}}\right)\right)^{\rho}, \tag{6}$$

where T denotes the number of discretized steps and i is the discretization index. In these experiments , we use $\sigma_{\min} = 10^{-4}$, $\sigma_{\max} = 1$, and $\rho = 13$. We found this schedule to be empirically effective for sampling with a reduced number of discretization steps. We use $T = 35$ discretization steps for the bandwidth extension and inpainting experiments, while, for declipping, we raised the number of steps to $T = 140$. Regarding the inference speed, with $T = 35$ steps, CQT-Diff requires 17 s for processing a 2.97-s segment in a GPU. The inference time augments up to 84 s when reconstruction guidance is used, as this feature requires gradient computation.

The neural network F'_θ was trained using the Adam optimizer with a learning rate of $2 \cdot 10^{-4}$. During training, the diffusion time t is randomly sampled following (6), but the parameters are set as $\sigma_{\min} = 10^{-6}$, $\sigma_{\max} = 10$, and $\rho = 10^5$. We set a segment size of 2.97 s (2^{16} samples) and a batch size of 4, but, as the architecture is fully convolutional, the segment size can be freely modified during inference. The model was trained for 320k iterations, which took about three days on a single NVIDIA A100 GPU.

3.1. Bandwidth Extension

The first set of experiments concerns audio bandwidth extension, where the degradation function is a lowpass filter $\boldsymbol{y} = \mathcal{A}(\boldsymbol{x}) = \text{LPF}(\boldsymbol{x})$. In contrast to previous works [4, 23, 32], our method does not suffer from filter generalization as a consequence of its fully generative nature, and, as we demonstrate in the companion webpage, can be applied after many different types of lowpass filters, even when the observations contain Gaussian noise. Nevertheless, we restrict the evaluation to FIR filters designed with a Kaiser window, with order 500 and two cutoff frequencies ($f_c = 1$ and 3 kHz). In this setting, it is possible to apply *data consistency* steps, where the observed low frequencies are replaced with the denoised estimate following $\bar{\boldsymbol{x}}_0 = \boldsymbol{y} + \hat{\boldsymbol{x}}_0 - \text{LPF}(\hat{\boldsymbol{x}}_0)$, in a way similar to [24].

We conduct an objective evaluation using 53 complete recordings (about 6 h of music) from the MAESTRO test set, which were not used during training. As objective metrics, we report in Table 1 the averaged Log-Spectral-Distance (LSD), Frechet Audio Distance (FAD), and the F1 score (onset) of a transformer-based piano transcription model [33] (F1)[6]. For this task, we experiment with applying our method (CQT-Diff) in two conditioning settings: using only *data consistency* (DC) and combining it with *reconstruction guidance* (DC+RG), with the scaling parameter $\xi' = 0.1$. We include two conditions where the CQT-based neural architecture is replaced with one using the STFT (STFT-Diff) and raw audio (SaShiMi [17]). We also compare with BEHM-GAN [4], a recent method for music bandwidth extension which we retrained using the same dataset. The BEHM-GAN was trained using lowpass filters with randomized cutoff frequencies at around 3 kHz, so it could only be used as the baseline in the $f_c = 3$ kHz setting. Table 1 also reports the number of trainable parameters (#prms) as an indicator of the model capacity.

We conducted a subjective listening test following the MUSHRA recommendation. A total of 12 volunteers were asked to rate the different methods in terms of perceptual audio quality, in comparison with a reference. The test was split into two parts, each evaluating the bandwidth extension from the cutoff frequencies of 1 kHz and 3 kHz. Each part of the test contained 4 different audio examples. Regarding the conditions, both parts shared the same low anchor (a lowpass-filtered sample at f_c = 1 kHz), but the other conditions differed. In the 1 kHz part, we included the STFT-Diff and SaShiMi-Diff conditions, as well as the two variants of our method. In the 3 KHz part, the lowpass-filtered sample at f_c = 3 kHz was included, as well as the BEHM-GAN baseline, but the STFT-Diff and SaShiMi-Diff conditions were excluded for length reasons. The results of the subjective experiments are presented in Fig. 2.

After analyzing the results, we reached the following conclusions. (i) The CQT data representation exhibits a more favourable performance against the STFT or SaShiMi, as its MUSHRA and F1 scores are higher. (ii) The use of RG is not critical for this task, as the (DC) and (DC+RG) conditions obtained comparable results. The reason may be that the architecture is explicitly designed to capture long-range data dependencies in the frequency range, thanks to the dilated convolutions. (iii) In the $f_c = 3$ kHz setting, the CQT-Diff (DC) condition obtained significantly better subjective scores than the baseline BEHM-GAN (p-value $1.4 \cdot 10^{-5}$ in a paired t-test).

[3] research.spa.aalto.fi/publications/papers/icassp23-cqt-diff

[4] github.com/eloimoliner/CQTdiff

[5] The distribution of noise levels used during includes a wider range of values than the ones used required during sampling. This allowed for flexibility during the exploration of the optimal sampling parameters without retraining.

[6] The MAESTRO dataset includes the ground-truth MIDI transcriptions.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on July 12,2023 at 00:48:46 UTC from IEEE Xplore. Restrictions apply.

Table 1: Objective metrics for bandwidth extension. The best result in each column is highlighted.

Method	f_c = 1 kHz LSD ↓	FAD ↓	F1 ↑	f_c = 3 kHz LSD ↓	FAD ↓	F1 ↑	#prms
LPF	1.29	12.47	0.67	0.81	3.50	0.73	-
BEHM-GAN	-	-	-	0.68	0.37	0.93	82M
STFT-Diff	0.95	**0.48**	0.68	0.71	0.21	0.93	65M
SaShiMi-Diff	0.97	4.06	0.63	0.70	1.37	0.93	19M
CQT-Diff (DC)	**0.86**	0.51	0.79	0.66	0.22	**0.94**	15M
CQT-Diff (DC+RG)	**0.86**	0.50	**0.80**	**0.65**	**0.17**	**0.94**	15M

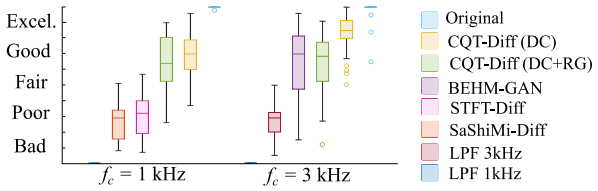


Fig. 2: Results of the bandwidth extension listening test.

3.2. Long Audio Inpainting

We apply our framework for the task of long audio inpainting, where the degradation function is a linear binary mask $\mathcal{A}(\boldsymbol{x}) = (1 - \mathbb{1}_{[t_{\text{start}}, t_{\text{end}}]})\boldsymbol{x}$, which deletes a long signal segment between t_{start} and t_{end}. For this task, the conditioning can be applied by using the DC steps as $\bar{\boldsymbol{x}}_0 = \boldsymbol{y} + \mathbb{1}_{[t_{\text{start}}, t_{\text{end}}]}\hat{\boldsymbol{x}}_0$, and/or using reconstruction guidance, directly applying (4).

We have evaluated the success of our method in long audio inpainting by means of a subjective experiment, which consisted of a multi-stimulus listening test. In the test, an 11-s excerpt of piano music with a long gap, i.e., silence, in the middle was presented to the participants as the reference. The participants were asked to rate the rest of the different conditions in terms of plausibility of the reconstruction. The test included four audio examples with three different gap lengths, forming a total of 12 pages of comparisons. We evaluate CQT-Diff in the same conditioning settings as in Sec. 3.1, relying solely on the DC and combining it with RG, using the scaling parameter $\xi' = 0.35$. We compare our method against two recent baselines: GACELA [5] and Catch-A-Waveform [34] (CAW).

The main conclusions from the results of the subjective experiment reported in Fig. 3 are the following: (i) In contrast with the outcomes from Sec. 3.1, reconstruction guidance is critical for inpainting, given the significant difference in the results. Our explanation is that the neural architecture F_θ has not been explicitly designed to account for long-term time dependencies and, thus, the guidance term is necessary to ensure coherence. (ii) Although the confidence intervals overlap, the MUSHRA scores of CQT-Diff are marginally higher than GACELA. We conclude that the proposed method shows at least comparable performance against the baseline while being more flexible.

3.3. Audio Declipping

Audio declipping refers to canceling the distortion in an audio signal that has been hard-clipped. We treat declipping as a nonlinear inverse problem, where the degradation function is a saturating nonlinearity $\mathcal{A}(\boldsymbol{x}_0) = (|\boldsymbol{x}_0 + c| - |\boldsymbol{x}_0 - c|)/2$, where c is the clipping threshold. In this case, we apply the conditioning on the clipped observations based on, uniquely, the RG strategy.

We conducted an objective evaluation where CQT-Diff was compared against two popular audio declipping methods: Analysis SPArse DEclipper (A-SPADE) [1] and Social Sparsity with Persistent Empirical Wiener (SS-PEW) [2]. We report three objective metrics: FAD [35], PEAQ, and PEMO-Q. The test set consists of 149 segments from different piano recordings of the MAESTRO test set, each of 2.97 s. Table 2 shows the evaluation results on two Signal-to-Distortion Ratio (SDR) settings: 1 dB and 10 dB.

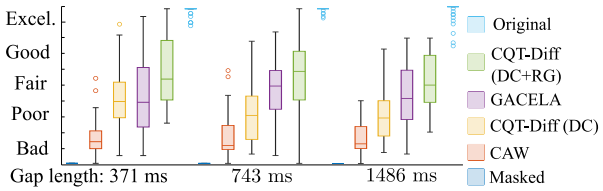


Fig. 3: Results of the audio inpainting listening test.

Table 2: Objective metrics on audio declipping tests.

	SDR = 1 dB FAD ↓	PEAQ ↓	PEMO-Q ↓	SDR = 10 dB FAD ↓	PEAQ ↓	PEMO-Q ↓
Clipped	23.09	-3.88	-3.82	7.85	-3.87	-3.78
CQT-Diff (RG)	**1.84**	**-3.53**	**-3.36**	0.68	-2.99	-2.56
A-SPADE	8.62	-3.87	-3.58	0.59	-2.97	-2.60
SS-PEW	9.69	-3.78	-3.62	**0.46**	**-2.05**	**-0.93**

Thanks to the generative priors, CQT-Diff has an advantage over the other approaches in the most severe circumstances: when SDR = 1 dB, it achieved higher scores than the compared baselines in the three reported metrics in Table 2. However, in moderate cases (SDR = 10 dB), the performance of CQT-Diff is upper-bounded by the still imperfect generation quality of the diffusion model. This explains why CQT-Diff is evidently outperformed by SS-PEW when SDR = 10 dB, although having comparable performance to A-SPADE.

4. CONCLUSIONS

We proposed CQT-Diff, a diffusion-based generative model which can be used as a flexible framework for audio restoration. In objective and subjective experiments, CQT-Diff excels in the task of audio bandwidth extension and shows competitive performance to the compared baselines in long audio inpainting and declipping, without the need of retraining for each task in particular. Apart from the three evaluated tasks, we emphasize that the proposed model could be flexibly applied for solving a much wider range of linear or nonlinear ill-posed problems, as long as the degradation is known. The presented model is designed to benefit from a CQT that leverages data structure as an inductive bias. However, we must admit that the benefits of the CQT are a conjecture that is not empirically substantiated in the current paper and should to be tackled in future work Future work should also target adapting the model to be trained in a large-scale data regime extending beyond piano music, which would allow for general audio restoration applications. This work shows the potential of diffusion models for audio restoration, and how they can benefit from using invertible time-frequency representations.

5. ACKNOWLEDGMENTS

We thank Tero Karras, Miika Aittala, Samuli Laine, and Timo Aila for the helpful discussions and for proposing the idea of using a Constant-Q Transform. We acknowledge the computational resources provided by the Aalto Science-IT project.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on July 12,2023 at 00:48:46 UTC from IEEE Xplore. Restrictions apply.

6. REFERENCES

[1] S. Kitić, N. Bertin, and R. Gribonval, "Sparsity and cosparsity for audio declipping: A flexible non-convex approach," in *Proc. Int. Conf. Latent Variable Analysis Signal Separation*, Aug. 2015, pp. 243–250.

[2] K. Siedenburg, M. Kowalski, and M. Dörfler, "Audio declipping with social sparsity," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2014, pp. 1577–1581.

[3] Ç. Bilen, A. Ozerov, and P. Pérez, "Solving time-domain audio inverse problems using nonnegative tensor factorization," *IEEE Trans. Signal Process.*, vol. 66, no. 21, pp. 5604–5617, Sep. 2018.

[4] 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, Jul. 2022.

[5] A. Marafioti, P. Majdak, N. Holighaus, and N. Perraudin, "GACELA: A generative adversarial context encoder for long audio inpainting of music," *IEEE J. Selected Topics Signal Process.*, vol. 15, no. 1, pp. 120–131, Nov. 2020.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, Dec. 2020, pp. 6840–6851.

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

[8] R. Rombach, A. Blattmann, D. Lorenz, et al., "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. CVPR*, Jun. 2022, pp. 10684–10695.

[9] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *Proc. NeurIPS*, Dec. 2022.

[10] Z. Kong, W. Ping, J. Huang, et al., "DiffWave: A versatile diffusion model for audio synthesis," in *Proc. ICLR*, May 2021.

[11] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv preprint arXiv:2206.03065*, 2022.

[12] S. Rouard and G. Hadjeres, "CRASH: Raw audio score-based generative modeling for controllable high-resolution drum sound synthesis," in *Proc. ISMIR*, Nov. 2021.

[13] C. Hawthorne, I. Simon, A. Roberts, et al., "Multi-instrument music synthesis with spectrogram diffusion," *Proc. ISMIR*, Dec 2022.

[14] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *Proc. ICLR*, May 2022.

[15] H. Chung, B. Sim, D. Ryu, and J. C. Ye, "Improving diffusion models for inverse problems using manifold constraints," in *Proc. NeurIPS*, 2022.

[16] H. Chung, J. Kim, M. T. Mccann, et al., "Diffusion posterior sampling for general noisy inverse problems," in *Proc. ICLR*, 2023.

[17] K. Goel, A. Gu, C. Donahue, and C. Re, "It's raw! Audio generation with state-space models," in *Proc. 39th Int. Conf. Machine Learning*, Mar 2022, pp. 7616–7633.

[18] E. Moliner and V. Välimäki, "Realistic gramophone noise synthesis using a diffusion model," in *Proc. Int. Conf. Digital Audio Effects (DAFX)*, Sep. 2022, pp. 240–247.

[19] J. Richter, S. Welker, J. Lemercier, et al., "Speech enhancement and dereverberation with diffusion-based generative models," *arXiv preprint arXiv:2208.05830*, 2022.

[20] M. Won, S. Chun, O. Nieto, and X. Serra, "Data-driven harmonic filters for audio representation learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2020, pp. 536–540.

[21] S. Huang, Q. Li, C. Anil, et al., "TimbreTron: A WaveNet(CycleGAN(CQT(audio))) pipeline for musical timbre transfer," in *Proc. ICLR*, May 2019.

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

[23] S. Han and J. Lee, "NU-Wave 2: A general neural audio upsampling model for various sampling rates," in *Proc. Interspeech*, Sep. 2022.

[24] J. Choi, S. Kim, Y. Jeong, et al., "ILVR: Conditioning method for denoising diffusion probabilistic models," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, Oct. 2021, pp. 14347–14356.

[25] A. Lugmayr, M. Danelljan, A. Romero, et al., "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proc. IEEE/CVF Conf. CVPR*, Jun. 2022, pp. 11461–11471.

[26] Sungwon Kim, Heeseung Kim, and Sungroh Yoon, "Guided-TTS 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data," *arXiv preprint arXiv:2205.15370*, 2022.

[27] G. A. Velasco, N. Holighaus, M. Dörfler, and T. Grill, "Constructing an invertible constant-Q transform with non-stationary Gabor frames," in *Proc. Int. Conf. Digital Audio Effects (DAFX)*, Paris, France, 2011.

[28] T. Karras, S. Laine, M. Aittala, et al., "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Jun. 2020, pp. 8110–8119.

[29] M. Tancik, P. Srinivasan, B. Mildenhall, et al., "Fourier features let networks learn high frequency functions in low dimensional domains," in *Proc. NeurIPS*, Dec. 2020, vol. 33, pp. 7537–7547.

[30] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artificial Intelligence*, Feb. 2018, vol. 32.

[31] C. Hawthorne, A. Stasyuk, A. Roberts, et al., "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proc. ICLR*, May 2019.

[32] S. Sulun and M. E. P. Davies, "On filter generalization for music bandwidth extension using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 1, pp. 132–142, Nov. 2020.

[33] C. Hawthorne, I. Simon, R. Swavely, et al., "Sequence-to-sequence piano transcription with transformers," in *Proc. ISMIR*, Dec. 2021.

[34] G. Greshler, T. Shaham, and T. Michaeli, "Catch-a-waveform: Learning to generate audio from a single short example," in *Proc. NeurIPS*, 2021, vol. 34, pp. 20916–20928.

[35] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms," in *Proc. Interspeech*, Aug. 2019, pp. 2350–2354.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on July 12,2023 at 00:48:46 UTC from IEEE Xplore. Restrictions apply.