Moliner Juanpere, Eloi; Välimäki, Vesa

Diffusion-Based Audio Inpainting

Published in:
AES: Journal of the Audio Engineering Society

DOI:
10.17743/jaes.2022.0129

Published: 01/03/2024

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

Please cite the original version:
Moliner Juanpere, E., & Välimäki, V. (2024). Diffusion-Based Audio Inpainting. *AES: Journal of the Audio Engineering Society*, *72*(3), 100-113. https://doi.org/10.17743/jaes.2022.0129

OPEN ACCESS Freely available online
E. Moliner and V. Välimäki,
"Diffusion-Based Audio Inpainting,"
J. Audio Eng. Soc., vol. 72, no. 3, pp. 100–113 (2024 Mar.).
https://doi.org/10.17743/jaes.2022.0129.

Diffusion-Based Audio Inpainting

ELOI MOLINER,* *AES Student Member,* **AND VESA VÄLIMÄKI,** *AES Fellow*
(eloi.moliner@aalto.fi) (vesa.valimaki@aalto.fi)

Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Espoo, Finland

Audio inpainting aims to reconstruct missing segments in corrupted recordings. Most existing methods produce plausible reconstructions when the gap lengths are short but struggle to reconstruct gaps larger than about 100 ms. This paper explores diffusion models, a recent class of deep learning models, for the task of audio inpainting. The proposed method uses an unconditionally trained generative model, which can be conditioned in a zero-shot fashion for audio inpainting and is able to regenerate gaps of any size. An improved deep neural network architecture based on the constant-Q transform that allows the model to exploit pitch-equivariant symmetries in audio is also presented. The performance of the proposed algorithm is evaluated through objective and subjective metrics for the task of reconstructing short to mid-sized gaps, up to 300 ms. The results of a formal listening test indicate that, for short gaps in the range of 50 ms, the proposed method delivers performance comparable to the baselines. For wider gaps up to 300 ms long, the authors' method outperforms the baselines and retains good or fair audio quality. The method presented in this paper can be applied to restoring sound recordings that suffer from severe local disturbances or dropouts.

0 INTRODUCTION

Audio inpainting refers to repairing or filling in missing or degraded parts of an audio signal [1]. Inpainting can be used to remove noise, glitches, or other unwanted artifacts from an audio recording or to fill in missing sections of audio that have been lost or damaged. Application examples include the restoration of old recordings corrupted by local disturbances [2, 3], the reconstruction of missing audio samples caused by scratches in CDs [4], or compensation for audio packet loss in communication networks [5]. In addition audio inpainting can be used in music and audio production to create special effects or to manipulate audio signals in creative ways [6]. This paper presents a novel audio inpainting method based on diffusion models [7, 8], a recently proposed generative deep-learning technique.

The task of audio inpainting is an ill-posed inverse problem, characterized by a non-unique set of solutions. Audio inpainting has been widely studied in the literature [1, 9–12]. The methods employed in audio inpainting are primarily distinguished by the way the observed signal samples are used as a prior or how they incorporate pre-existing assumptions about the signal. For instance, some techniques are based on autoregression [4] or signal sparsity [13]. However most of these techniques demonstrate strong performance only when applied to gaps of less than 100 ms in duration. Such techniques tend to encounter challenges with longer gaps or in cases where the assumption of stationarity, explicitly required by autoregressive methods and implicitly relied upon by sparsity-based methods, does not hold.

In this work, the authors use generative priors, learned from a diffusion probabilistic model, assuming that the solution belongs to the same probability distribution as the dataset used for training. Inpainting methods based on deep generative models can reach new levels of expressivity, since they are not grounded by the stationarity condition and can generate new events in the inpainted gap [14–16]. In particular diffusion models have a strong potential to excel at this task as they possess a great versatility for solving inverse problems [17, 18, 16].

In the authors' previous study [16], the invertible Constant-Q Transform (CQT) was used with a diffusion model to solve inverse problems in audio. This paper revisits the use of the CQT, proposing an improved neural network architecture operating in the transform domain using a small amount of signal redundancy. A diffusion model, built with a deep neural network, is first trained with audio material as an unconditional generator. During inference, the model is conditioned in a zero-shot manner to generate a plausible reconstruction of the missing segment. In contrast to existing audio inpainting methods [10–12], the proposed diffusion model can restore gaps of arbitrary length, retaining high quality for longer gaps.

This paper addresses the inpainting of compact gaps in an audio signal without any accompanying side informa-

*To whom correspondence should be addressed, email: eloi.moliner@aalto.fi. Last updated: February 12, 2024

tion. Specifically the authors focus on gaps in the short-to-medium size range, ranging from 25 to 300 ms. Note that this differs from the goal of the authors' previous work [16], in which the model was tested on larger gaps up to 1.5 s in length. It was observed that when the gap was very long, the model had to generate new events. Although these generated events were often statistically plausible, they did not align with the musical context and were deemed musically incorrect, which is undesirable. This led to the conclusion that a practical inpainting method for large gaps would require a high-level understanding of the music structure or the ability to be conditioned with a guiding signal, as proposed in recent research [19]. However such considerations fall outside the scope of this paper. As a result, the authors limit the evaluation to gaps no longer than 300 ms. Within this range, they assume that the content to be filled can be anticipated by a human listener, ensuring a reliable evaluation of the inpainting performance.

The rest of this paper is organized as follows. SEC. 1 reviews the relevant audio and image inpainting literature. SEC. 2 explains the basics of diffusion models and the conditioning method for the inpainting task. SEC. 3 introduces the new diffusion-model architecture, which employs the invertible CQT. SEC. 4 compares the proposed method with previous inpainting methods in terms of objective and subjective metrics. SEC. 5 concludes the paper.

1 OVERVIEW OF INPAINTING METHODS

This section reviews some relevant methods in the audio inpainting literature. In addition the authors summarize some recent work on image inpainting with diffusion models that inspired this work.

1.1 Existing Audio Inpainting Methods

Adler et al. first used the term "audio inpainting" to describe the restoration of gaps in audio signals [1], adopting the name from the image inpainting literature. However this is an old problem in audio processing, and the same task has been previously referred to in the literature as audio interpolation [4, 20, 21], audio extrapolation [22, 23], reconstruction of missing samples [24, 25], waveform substitution [5], and imputation [26], among other things. The first methods used interpolation techniques based on the observed samples surrounding the gap [4]. A family of successful methods uses autoregressive modeling based on the assumption that the signal is stationary and can be approximated by a linear combination of past samples [4, 20, 21].

A more recent family of methods takes advantage of sparse signal representations [9, 11, 13], such as the short-time Fourier transform (STFT). These methods try to find the sparse representation of the missing part of the signal that best fits the surrounding, uncorrupted signal. An established method to enhance sparsity-based audio inpainting is to learn the dictionary of basis functions [12, 27]. Another recent work uses non-negative matrix factorization, exploiting the low-rankness of the magnitude spectra as a prior [28].

The methods mentioned above only perform well for inpainting short gaps, roughly in the range from 10 to 100 ms. For longer gaps, these methods tend to fail to produce plausible reconstructions since the stationarity condition does not hold true. Some inpainting attempts for long gaps are based on strong assumptions about the underlying structure of the gap, including sinusoidal modeling [29] or similarity graphs [30].

1.2 Deep-Learning–Based Audio Inpainting

During the last few years, a new trend has emerged using deep-learning–based techniques for audio restoration, including the task of inpainting. Most of these studies use generative models as the prior for inpainting. This allows for methods that are able to generate new content in the gaps to be filled. For instance generative adversarial networks have been explored for this task [14, 15, 31]. Most of these methods are based on a supervised problem-specialized setting, where a dataset of masked/reconstructed audio signals needs to be built to train the model. A shortcoming of this approach is that a model trained with a certain set of degradations does often not generalize to unseen degradations and, as a consequence, lacks the versatility to be applied for restoring gaps of arbitrary length.

Some other closely related studies fall under the category of packet-loss concealment, which is a similar problem to audio inpainting but with real-time constraints and usually targeting speech signals. Within this context, predictive methods based on convolutional and recurrent neural networks [32, 33] as well as generative adversarial networks [34, 35] have been proposed.

Also worth mentioning are other recent works that have applied multi-modal side information as a conditioner for the inpainting algorithm, including video frames [36], symbolic music [37, 19], or text [38, 39]. Although this idea falls outside the scope of this paper, exploiting multi-modal information may turn out to be beneficial to inpaint large gaps, where the context of the gap does not contain enough information to reconstruct the missing segment.

1.3 Diffusion Models for Image Inpainting

Deep generative models have tremendously impacted image processing research, not least their application to the image inpainting problem. Relevant to this paper are recent papers applying diffusion models for image inpainting. There are two main strategies in the literature to solve inverse problems with diffusion models, including inpainting. The first one consists of sequentially replacing the observed part of the signal in the reverse diffusion process [8, 40]. This idea ensures data consistency and is conceptually simple but, in practice, struggles to generate consistent content. Other work refined this approach by incorporating ideas that benefited its versatility and performance, such as using singular-value decomposition [17] or multiple resampling during each sampling step [41]. The other strategy builds on a Bayesian interpretation of posterior sampling and estimates the gradients of the log-likelihood function [18, 42], as elaborated in SEC. 2.2. These methods allow for

a good approximation of the posterior distribution, which leads to enhanced inpainting results, at the expense of a higher computational cost.

2 A DIFFUSION MODEL FOR AUDIO INPAINTING

Diffusion models are a class of generative models that have gained interest during recent years for a wide range of modalities, such as images [7, 43, 44], audio [45, 46, 16], video [47], and symbolic music [37], among others. These models generate new data instances by reversing the diffusion process, by which data $\boldsymbol{x}_0 \sim p_{\text{data}}$ is progressively diffused into Gaussian noise $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{max}}^2\mathbf{I})$ over time τ [7].[1]

The present authors follow the parameterization by Karras et al. [48], who define the reverse diffusion process with the following *probability flow ordinary differential equation (ODE)*:

$$\mathrm{d}\boldsymbol{x} = -\dot{\sigma}(\tau)\sigma(\tau)\nabla_{\boldsymbol{x}} \log p_\tau(\boldsymbol{x}_\tau)\mathrm{d}\tau, \tag{1}$$

where $\mathrm{d}\tau$ is an infinitesimal negative timestep, the noise level is defined as $\sigma(\tau) = \tau$, and its first derivative as $\dot{\sigma}(\tau) = 1$. The ODE is governed by the gradient of the log probability density $\nabla_{\boldsymbol{x}} \log p_\tau(\boldsymbol{x}_\tau)$, formally known as the *score* function [49].

The score is analytically intractable but can be approximated as

$$\nabla_{\boldsymbol{x}} \log p_\tau(\boldsymbol{x}_\tau) \approx (D_\theta(\boldsymbol{x}_\tau, \tau) - \boldsymbol{x}_\tau)/\sigma(\tau)^2, \tag{2}$$

where $D_\theta(\boldsymbol{x}_\tau, \tau) = \hat{\boldsymbol{x}}_0$, which refers to $\hat{\mathbf{x}}_\tau$ at $\tau = 0$, is a deep neural network with weights θ, optimized with a denoising Euclidean objective:

$$\mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\lambda(\tau)\| D_\theta(\boldsymbol{x}_0 + \sigma(\tau)\boldsymbol{\epsilon}, \tau) - \boldsymbol{x}_0\|_2^2\right], \tag{3}$$

where $\lambda(\tau)$ is a time-dependent weighting parameter. Furthermore, the preconditioning strategy proposed by Karras et al. [48] is used:

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

where F_θ represents an optimizable deep neural network, and $c_{\text{skip}}(\tau)$, $c_{\text{out}}(\tau)$, and $c_{\text{in}}(\tau)$ are weighting parameters optimized in such a way that the input and output of F_θ always have close-to-unit variance, a well-known good practice when training deep neural networks.

For more comprehensive details on the diffusion model formalism and optimization, as well as the optimal weighting parameters, the authors refer to [48]. In the rest of this section, the authors elaborate on the audio inpainting problem and required changes that are applied to the inference process of a diffusion model to solve this task.

2.1 Inverse Problem Formulation

The audio inpainting task can be formulated as a linear inverse problem [1]. Consider an audio signal $\boldsymbol{x}_0$ and its observed version $\boldsymbol{y}$ with missing samples. Their relation can be written as

$$\boldsymbol{y} = \mathbf{m} \odot \boldsymbol{x}_0, \tag{5}$$

where $\mathbf{m}$ is a binary mask operator and $\odot$ represents the Hadamard product, or element-wise multiplication. In this work, the operator $\mathbf{m}$ is considered as a known compact binary mask, having the value 0 at locations where samples are missing and 1 otherwise. The goal is to recover the original signal $\boldsymbol{x}_0$ when the observed measurements $\boldsymbol{y}$ and mask $\mathbf{m}$ are known.

2.2 Audio Inpainting via Posterior Sampling

The iterative nature of diffusion models offers great flexibility for solving inverse problems [18, 16]. All that is needed is to substitute the score in Eq. (1) with the *posterior score* $\nabla_{\boldsymbol{x}} \log p_\tau(\boldsymbol{x}_\tau|\boldsymbol{y})$ [8].

Applying Bayes' rule, the posterior factorizes as $p_\tau(\boldsymbol{x}_\tau|\boldsymbol{y}) \propto p_\tau(\boldsymbol{x}_\tau)p_\tau(\boldsymbol{y}|\boldsymbol{x}_\tau)$, which leads to

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

The authors refer to the term $\nabla_{\boldsymbol{x}_\tau} \log p_\tau(\boldsymbol{y}|\boldsymbol{x}_\tau)$ as the *noise-perturbed likelihood score*. Note that this term cannot be derived in closed form due to its dependence on the noise level $\sigma(\tau)$, because $\boldsymbol{x}_\tau$ represents the noise-perturbed signal $\boldsymbol{x}_\tau = \boldsymbol{x}_0 + \sigma(\tau)\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. However, Chung et al. [18] propose to approximate the *noise-perturbed likelihood* with $p_\tau(\boldsymbol{y}|\boldsymbol{x}_\tau) \simeq p(\boldsymbol{y}|\hat{\boldsymbol{x}}_0)$, where $\hat{\boldsymbol{x}}_0$ is the denoised estimate at an intermediate noise level.

Modeling the likelihood as a normal distribution, the *noise-perturbed likelihood score* is approximated as

$$\nabla_{\boldsymbol{x}} \log p_\tau(\boldsymbol{y}|\boldsymbol{x}_\tau) \simeq -\xi(\tau)\ \nabla_{\boldsymbol{x}} \|\boldsymbol{y} - \mathbf{m} \odot \hat{\boldsymbol{x}}_0\|^2. \tag{7}$$

This strategy can be understood as a sort of guidance [47], in analogy with classifier guidance [43]. Importantly, note that the gradient computation requires differentiating through the neural network F_θ, which is responsible for the estimation of $\hat{\boldsymbol{x}}_0$, resulting in computational overhead.

The variable $\xi(\tau)$ in Eq. (7) is a scaling function that defines the amount of guidance that is applied during sampling or, in other words, how strongly the conditioning affects the sampling trajectories. The authors parameterize the scaling function as [16]

$$\xi(\tau) = \xi'\sqrt{N}/(\sigma(\tau)\|\nabla_{\boldsymbol{x}} \|\boldsymbol{y} - \mathbf{m} \odot \hat{\boldsymbol{x}}_0\|^2\|^2), \tag{8}$$

where N is the length of the audio signal in samples and ξ' is a scalar hyperparameter. Choosing $\xi' = 0$ leads to an unconditional sampler, but selecting too large a value for ξ' results in a degenerate solution. This parameterization scales the likelihood gradient by its norm in a way similar to [50], regularizing the influence of the likelihood throughout the inference process. The authors empirically observed through qualitative analysis that this strategy allows for robust results.

However the above conditioning method does not ensure data consistency with the observed samples. When the observed samples $\boldsymbol{y}$ are noiseless and reliable, as is assumed in this work, the preferred solution is to keep them unchanged

[1] The "diffusion time" variable τ must not be confused with the "audio time" t. The authors use this formulation for notational consistency.

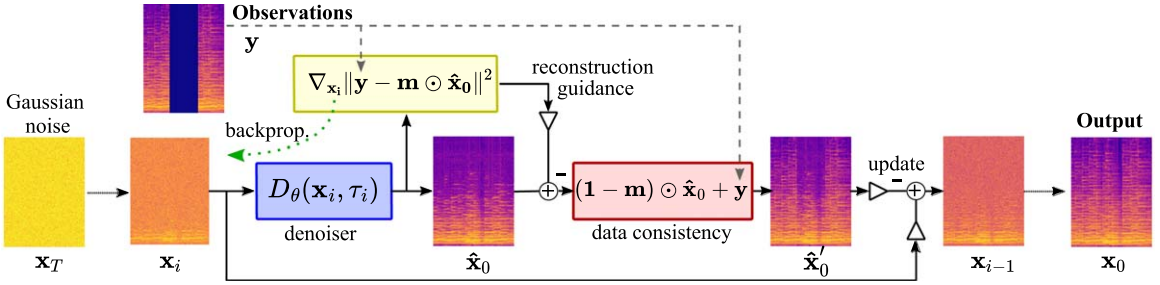

Fig. 1. Inference block diagram for audio inpainting, where all straight lines represent a feedforward signal flow in the time domain. The deep neural network is included in the denoiser block. The computation of the reconstruction gradient requires differentiating through the mask and denoiser block by means of backpropagation, denoted as "backprop." above, requiring a backward pass through the deep neural network, illustrated here with a curved dotted line. The spectrograms are shown for illustrative purposes.

in the final output. A straightforward way to avoid changing the existing samples is replacing the reliable samples from the intermediate estimates $\hat{\boldsymbol{x}}_0$ using the inpainting mask. To keep the observed samples, the following data consistency step at each sampling iteration can be applied:

$$\hat{\boldsymbol{x}}'_0 = \boldsymbol{y} + (\mathbf{1} - \mathbf{m}) \odot \hat{\boldsymbol{x}}_0. \tag{9}$$

Although some studies have proved the data consistency step suboptimal [18, 51], others rely solely on data consistency as a method to condition the diffusion model to solve inverse problems [8, 41]. The authors observe that applying data consistency steps usually produces discontinuity effects at the boundaries of the mask. To mitigate this effect, the authors apply a smoothed version of the mask $\mathbf{m}$ for the data consistency step of Eq. (9), which is implemented by fading 1 ms of the reliable signal at the edges of each gap with a raised cosine function.

Algorithm 1. Inference conditioned for audio inpainting.

Require: observations $\boldsymbol{y}$, inpainting mask $\boldsymbol{m}$, number of iterations T, noise schedule τ_i, stochasticity S_{churn}
Sample $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\text{max}}\mathbf{I})$ ▷ Initial noise is $\boldsymbol{x}_T$
$\gamma = \min(S_{\text{churn}}/T, \sqrt{2} - 1)$ ▷ Amount of stochasticity
for $i = T, \ldots, 1$ **do** ▷ Step backwards
  $\tilde{\tau}_i = (1 + \gamma)\tau_i$ ▷ Increased noise level
  Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i + \sqrt{\sigma(\tilde{\tau}_i)^2 - \sigma(\tau_i)^2}\boldsymbol{\epsilon}$ ▷ Add extra noise
  $\hat{\boldsymbol{x}}_0 = D_\theta(\tilde{\boldsymbol{x}}_i, \tilde{\tau}_i)$ ▷ Denoiser
  $\hat{\boldsymbol{x}}_0 = H_{\text{post}}(\hat{\boldsymbol{x}}_0)$ ▷ Post-processing filter
  $\hat{\boldsymbol{x}}_0 = \hat{\boldsymbol{x}}_0 - \sigma(\tilde{\tau}_i)^2 \xi(\tilde{\tau}_i)\, \nabla_{\tilde{\boldsymbol{x}}} \|\boldsymbol{y} - \mathbf{m} \odot \hat{\boldsymbol{x}}_0\|^2$ ▷ Eq. (7)
  $\hat{\boldsymbol{x}}'_0 = \boldsymbol{y} + (\mathbf{1} - \mathbf{m}) \odot \hat{\boldsymbol{x}}_0$ ▷ Data consistency
  $\boldsymbol{x}_{i-1} = \tilde{\boldsymbol{x}}_i - (\sigma(\tau_{i-1}) - \sigma(\tilde{\tau}_i)) \left(\frac{\hat{\boldsymbol{x}}'_0 - \tilde{\boldsymbol{x}}_i}{\sigma(\tilde{\tau}_i)}\right)$ ▷ Update step
end for
return $\hat{\mathbf{x}}_0$ ▷ Output is the reconstructed signal

2.3 Inference

Having defined the probability flow ODE, Eq. (1), and the posterior sampling mechanism, Eq. (7), the next tasks are to discretize and solve the reverse diffusion process, using a trained diffusion model. In this work, the authors use the second-order stochastic sampler proposed by Karras et al. [48], offering a good tradeoff between algorithmic complexity and accuracy. This sampler also adds controllable stochasticity into the process, which is intended to regularize approximation errors. The sampling algorithm, specific for inpainting, is described in Algorithm 1. Fig. 1 summarizes graphically the sampling process, omitting the second-order correction for brevity. It is important to note that implementing the second-order correction would necessitate an additional denoising forward pass and the guidance computation.

As presented on the left-hand side of Fig. 1, the audio signal $\boldsymbol{x}_T$ is initialized with Gaussian noise, and the noise is iteratively removed throughout the inference process. During each discretization step, the authors acquire a denoised estimate denoted as $\hat{\boldsymbol{x}}_0$. They then condition this estimate with the input observations shown at the top of Fig. 1 by incorporating the reconstruction gradient from Eq. (7) and implementing the data consistency step described in Eq. (9). Each update step is a weighted sum of the noisy signal at the given step $\boldsymbol{x}_i$ and the modified denoised estimate $\hat{\boldsymbol{x}}'_0$. At the end of the process, shown on the right-hand side of Fig. 1, the noise level becomes imperceptible, and thus the reconstructed output signal $\hat{\boldsymbol{x}}_0$ is obtained.

The noise schedule represents one of the most critical design choices. Also following [48], given a number of discretization steps T, the authors define the noise levels as

$$\tau_i = \left(\sigma_{\text{max}}^{\frac{1}{\rho}} + \frac{i}{T-1}\left(\sigma_{\text{min}}^{\frac{1}{\rho}} - \sigma_{\text{max}}^{\frac{1}{\rho}}\right)\right)^{\rho}, \tag{10}$$

where $0 \leq i \leq T - 1$ is the discretization index; σ_{min} and σ_{max} are, respectively, the minimum and maximum noise levels; and ρ is a parameter controlling the warping of the schedule, with higher values of ρ representing more steps at lower noise levels. As in [16], the authors choose $\sigma_{\text{min}} = 10^{-4}$, $\sigma_{\text{max}} = 1$, and $\rho = 13$. The number of steps T exhibits a clear tradeoff between sample quality and speed. The authors use the value $T = 70$ in these experiments.

The amount of stochasticity injected into the process is controlled with the parameter S_{churn} [48]. Empirically, the authors have observed that adding a certain amount of stochasticity helps to produce clean outputs. In these experiments, they choose $S_{\text{churn}} = 10$.

3 IMPROVED CQT-BASED ARCHITECTURE

Diffusion models are architecture agnostic, imposing no constraints on how the denoiser backbone architecture F_θ is designed. However, although the choice of architecture does not have theoretical implications, in the best case it can accelerate the convergence of the diffusion, producing perceptually satisfying samples efficiently.

In their previous study [16], the authors used an invertible CQT [52] to leverage structure from the audio signal and to exploit the pitch-equivariant symmetry that harmonic signals exhibit when they are represented in a logarithmically spaced time-frequency domain. The most interesting property of the CQT is that a translation on the frequency axis is equivalent to pitch transposition. This symmetry motivates the usage of two-dimensional convolutional neural networks (CNNs), considering that the convolutional operator, which CNNs are composed of, is translation equivariant. In this section, the authors elaborate the usage of a CQT and introduce an improved neural network architecture that processes audio signals as CQT spectrograms. The authors call the improved diffusion model based on this architecture the CQT-Diff+ algorithm.

3.1 Using a CQT Representation

The diffusion process described in SEC. 2 is developed in the time domain. However, as part of the computation inside the deep neural network F_θ, the input waveform is represented with an invertible CQT. Concisely, the neural network F_θ is composed as

$$F_\theta = \mathrm{ICQT} \circ F'_\theta \circ \mathrm{CQT}, \tag{11}$$

where CQT and ICQT are the constant-Q-transform operation and its inverse, respectively; $\circ$ is the function composition operation; and F'_θ refers to the neural network layers with trainable weights. This approach takes advantage of the structure imposed by the CQT while maintaining maximum versatility. Applying the neural network weights in the transform domain does not impact the optimization, because both the forward transform and its inverse are differentiable.

The authors use the CQT proposed by Velasco et al. [52] and by Holighaus et al. [53]. Briefly, this transform is built on a set of K bandpass filters g_k with an equal Q-factor and logarithmically spaced center frequencies, defined as

$$f_k = f_{\min} 2^{\frac{k-1}{B}}, \quad \text{for } k = 1, 2, 3..., K, \tag{12}$$

where B is the number of bins per octave band (when the number of octave bands is $N_{\mathrm{oct}} = K/B$) and $f_{\min} = f_1$ is the lowest center frequency. The maximum center frequency can be designed to be placed at the Nyquist limit $f_K = f_s/2$. The CQT is applied using the fast Fourier transform–based processing, as introduced in [52], which allows for a computationally efficient implementation. The authors refer to [52, 53] for further details on the CQT transform, as well as to their publicly available implementation.[2]

[2]https://github.com/eloimoliner/CQT_pytorch.

3.1.1 Discarding the DC Component

An obvious inconvenience caused by the logarithmic frequency resolution is that there is no DC bin at 0 Hz. If perfect reconstruction is required, one solution is to encode the DC component with a low-pass filter g_{DC}. In the authors' prior work [16], the DC component was included in the model input by concatenating it to the time-frequency matrix. However the authors observed the presence of low-frequency artifacts in the generated outputs, which they attribute to the disruption of the logarithmically uniform frequency resolution at the DC component. Therefore the authors made the decision to discard the DC component. By excluding the DC component, the completeness of the CQT as a transform is compromised. The model would now train on only a subset of the transformed space, producing an irreducible error in the frequency bands that it is blind to. However this issue does not represent a problem in practice, because the audio signals are assumed to be bandlimited, and very little relevant information exists below 43 Hz, which corresponds to the lowest frequency band in the authors' specific case.

Nevertheless this irreducible error must be accounted for when propagating the loss during training as well as during the sampling stage. This compensation can be implemented by applying a post-processing filter to the denoiser output:

$$\hat{\boldsymbol{x}}_0 = H_{\mathrm{post}}(D_\theta(\boldsymbol{x}_\sigma, \sigma)), \tag{13}$$

where H_{post} is a DC notch filter, designed to suppress the frequency range that is not covered by the CQT bandpass filters g_k. This filter is applied during both training and inference after each forward evaluation of $D_\theta(\,\cdot\,)$.

3.1.2 Optimizing Redundancy

In CQTs the receptive field of the filters decreases geometrically with frequency. To guarantee invertibility, the decimation factors of the frequency bands also need to decrease geometrically, producing a non-uniform sampling grid. This feature is not only impractical for constructing a parallelizable and GPU-efficient implementation but also complicates the architecture of the neural network.

An easy way to overcome this problem is to use instead a CQT with a uniform sampling grid, where the decimation factors remain constant across the frequency range [54]. Previously this approach allowed treating the CQT as a 2D matrix and thus directly applying 2D CNNs [16]. A major drawback of this strategy is its overcompleteness, which leads to a suboptimal consumption of memory and computational requirements, as the 2D CNN is forced to process a substantial amount of signal redundancy. This considerably slowed down the training and inference processes in the authors' previous study [16] and limited the potential scalability of the model.

Schörkhuber et al. [55] proposed splitting the CQT into a sequence of octave-wise sub-transforms as a way to reduce redundancy. The present authors adopt this strategy and apply different sub-transforms with a constant decimation factor for each octave band, in this case, each of them having 64 frequency bins. A strong advantage of separating

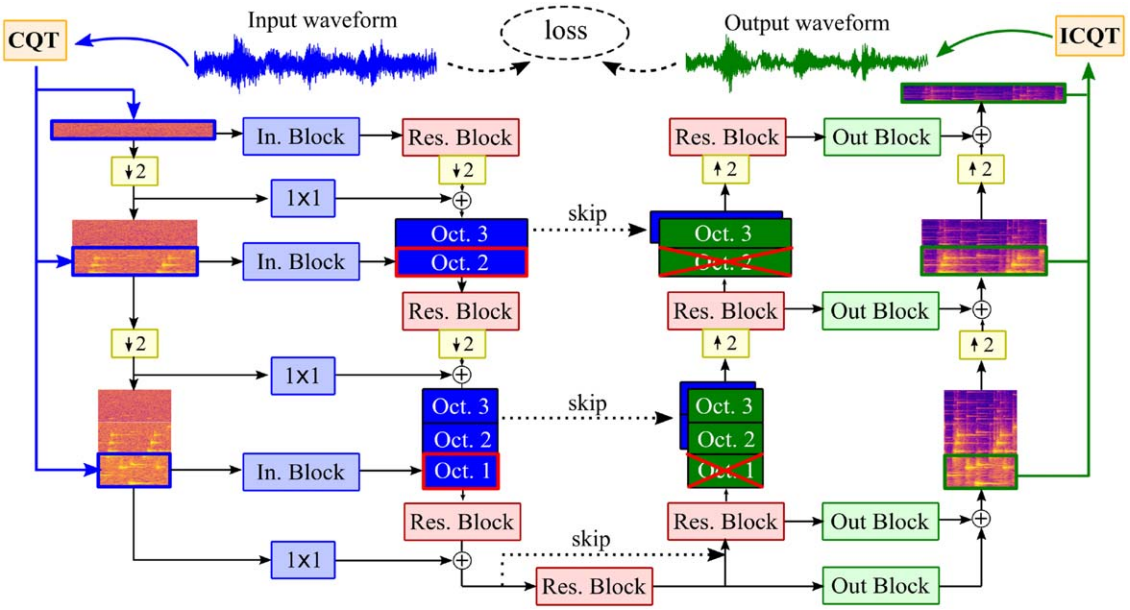

Fig. 2. Main diagram of the CQT-U-Net deep neural network architecture. In the diagram, only three octaves of eight are shown for clarity. The sizes of the spectrograms are not proportional to the real signals.

the CQT into octave bands is that, when powers of two are used as the sequence length, the time resolution decreases exactly by a factor of two between two consecutive octaves. This choice leads to a hierarchical representation that is suitable for processing with a U-Net architecture [56].

3.2 Architecture Design

A U-Net has a hierarchical encoder/decoder structure, as shown in Fig. 2, where the left-hand side is the encoder and right-hand side is the decoder. The center part of Fig. 2 is called the "bottleneck," which corresponds to the lowest point of the letter "U." The temporal resolution is progressively reduced by a factor of two between consecutive layers in Fig. 2, while the frequency resolution remains unchanged [16]. The authors make use of this hierarchy by concatenating features from each CQT octave at the U-Net layers where the time resolutions match, as Fig. 2 also illustrates.

The proposed architecture utilizes a double real representation of the complex CQT features, where the real and imaginary parts are stacked as two separate channels. Thus the real and imaginary parts are freely merged in the channel dimension of the network, where the number of channels is further increased, but the synchrony between real and imaginary parts in the time-frequency space is conserved. The chosen strategy aims to circumvent the computational complexity associated with complex-valued layers, since they generally lack empirical performance advantages compared to their real-valued counterparts while imposing higher computational costs [57]. However, even though the neural network views the features as real, the underlying data is complex, and one must be cautious with how the features are processed.

In particular it was observed that shift-based operations, such as biases in convolutional layers or mean normalizations, introduced perceptual artifacts to the generated output and, as a consequence, they should be avoided. The intuition behind this lies in the unique nature of complex numbers and the way they interact during computations. In a complex number, the real and imaginary parts represent different dimensions of information. If shift-based operations were applied, they could introduce imbalances between the real and imaginary components, leading to inconsistent phase relationships and distorted information. Thus bias terms in all the layers are set to zero. However this does not apply to additive residual connections, since they are designed to add the activations of one layer to another without altering their phase relationships. Note that this does not represent a practical limitation to the model because all the signals are approximately zero mean.

In accordance with typical U-Net architectures, concatenative skip connections bridge the intermediate resolutions of the encoder and decoder as shown in Fig. 2. An antialiasing filter was used in the downsampling and upsampling layers in the encoder and decoder stages, respectively. At each resolution of both the encoder and decoder stages and at the bottleneck, a residual block (referred to as "Res. Block") is applied, which constitutes the primary building block of the architecture.

In the encoder, the left-hand side of Fig. 2, the input coefficients are divided into octave-specific "pieces" and processed separately using "In. Blocks." The features from each octave are then concatenated along the frequency dimension to the corresponding latent vectors of the U-Net at corresponding time resolutions, as represented graphically in Fig. 2. Additionally residual connections are applied between the (resized) input features and the corresponding latent vectors of the U-Net to facilitate information flow through all layers of the encoder.

The decoder, or the right-hand side of Fig. 2, comprises the main signal path containing "Res. Blocks" and the outer

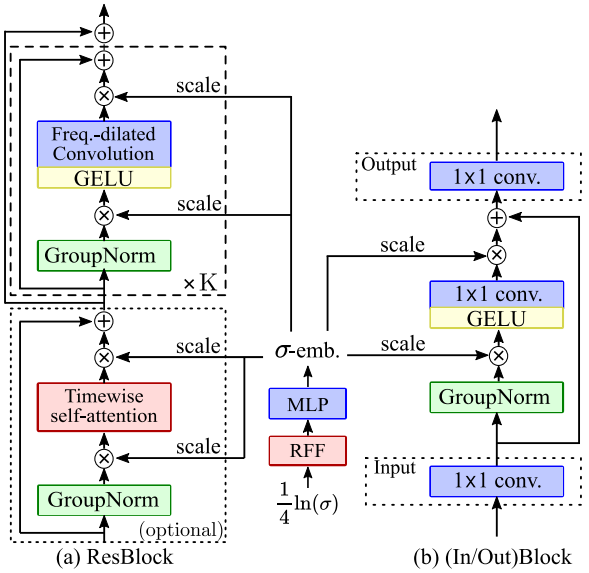

Fig. 3. Building blocks of the backbone U-Net architecture, cf. Fig. 2.

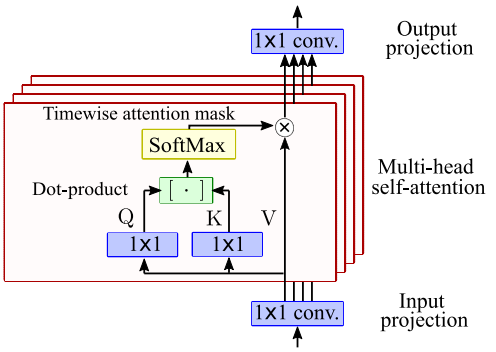

Fig. 4. Timewise self-attention block used in Fig. 3.

path with residual connections. As the temporal resolution is upsampled in the main path, the features from the lowest octaves at each layer are discarded and projected to the "outer" path via "Out. Blocks." In the outer path, at each resolution, the lowest octave is extracted and sent to the ICQT block, as indicated at the right part of Fig. 2 with green lines. This dual-path strategy is inspired by Karras et al. [44], and its purpose is to improve the gradient flow during the optimization process.

3.2.1 Building Blocks of the Architecture

The building blocks of Fig. 2 are presented in detail in Fig. 3. They are all conditioned with the noise-level embedding σ-emb, which is built with random Fourier features (RFF) [58] followed by a multi-layer perceptron (MLP) having three layers. The conditioning is realized with feature-wise linear modulation [59], without shifts. The "In. Block" shown in Fig. 3(b) applies a 1×1 convolution to expand the channel size from two (real and imaginary) to the required number of latent features at every layer. They are followed by Group Normalization (without shift operations), a Gaussian-error-linear-unit ("GELU") non-linearity, and a linear layer. The "Out. Blocks" have a similar form, but with the 1×1 convolution placed at the end, mapping the latent vector to a channel size of two.

Fig. 3(a) shows that each residual block, "Res. Block," contains a stack of shift-free Group Normalization layers, followed by a "GELU" non-linearity and convolutions in both time and frequency, but with exponentially-increasing dilations in the frequency dimension, meant to provide a wide receptive field while exploiting the symmetry of pitch-equivariance. The authors additionally include a timewise self-attention layer in the deeper "Res. Block" layers, as explained in SEC. 3.2.2.

In contrast to the authors' former work [16], frequency-positional embeddings designed to encode absolute frequency positional information are not used. The reason is that, while the absolute frequencies cannot, in principle, be retrieved with a CNN, they can, in practice, be spuriously learned through the use of zero padding [60]. With this modified architecture, zero padding is used in the convolutional layers at each intermediate stage, also in the frequency dimension, propagating absolute positional information throughout the network, even at the shallower layers. The authors observed that, in this setting, the use of frequency-positional embeddings provided no significant benefit.

3.2.2 Timewise Self-Attention

The motivation behind using timewise self-attention is to allow the model to learn global features in the time dimension, overcoming the locality of CNNs. The use of attention would allow the model to analyze similarities between different segment pairs in time, a feature that could intuitively be highly beneficial for the task of inpainting. The latent features are two-dimensional (time and frequency), but since the idea is to apply attention only through time and not in frequency, some modifications are required with respect to the basic self-attention mechanism [61].

Fig. 4 presents the functionality of the timewise self-attention block. In order to reduce the otherwise unfeasible computational complexity, the authors introduce a 1×1 convolution before the self-attention mechanism. This reduces the number of channels, which can be quite large (up to 256), to only a few attention heads (set to eight in this work). For each head, the queries Q and keys K are computed with a linear layer that sees the frequency dimension as the feature dimension. A timewise attention mask is computed via standard dot-product attention and is later applied to the values V. In order to preserve the structure in the frequency dimension, the authors do not apply any processing to the values V, apart from the timewise attention. Finally, another 1×1 convolution is applied at the output to expand the reduced number of heads to the original channel size.

Note that, as implemented in this work, the use of timewise self-attention breaks the translation-equivariant property of fully convolutional networks, rendering the model unsuitable for processing sequences of different lengths than the one used during training. If one wishes to process

a longer sequence, a segment-by-segment approach with a fixed segment length can be applied.

3.2.3 Hyperparameter Specification

The architecture of Fig. 2 is designed to work at a sampling frequency of $f_s = 44.1$ kHz. The authors use a CQT with $B = 64$ bins per octave and $N_{oct} = 8$ octaves. The depth of the U-Net matches the number of octaves, and the feature sizes range from 64 features at the shallower U-Net layers to 256 features at the bottleneck. The number of stacked dilated convolutions on each "Res. Block" ranges from two to eight, with fewer dilations at the shallower layers since they need to cover fewer frequency bins and, thus, a large receptive field is not needed. Because of their quadratic complexity, timewise self-attention is only used at the three deepest layers, where the time resolution has been significantly reduced. The total parameter count is 242 million parameters. The authors refer to the public repository for further specifications.[3]

4 EVALUATION

The performance of the proposed method, which the authors refer to as CQT-Diff+, is evaluated for inpainting short to middle-sized gaps in musical recordings, ranging from 25 to 300 ms. For comparison, two baselines are considered:

- **LPC**: A method based on signal extrapolation using linear predictive coding [23].
- **A-SPAIN-L:** A sparsity-based method for audio inpainting with dictionary learning [12], which is regarded as a state-of-the-art method for short-gap inpainting.

Results of both objective and subjective experiments are reported. All experiments use the sample rate of 44.1 kHz. In the objective evaluation, the authors analyze various gap lengths within intervals of 25 ms.

In the subjective evaluation, the authors examine four different gap sizes: 50, 100, 200, and 300 ms. In all instances, they intentionally introduce four gaps uniformly across audio segments that have a duration of 4.17 s. These gaps are applied simultaneously at predetermined time locations. Other machine-learning–based methods could not be included in the evaluation, since they either were not designed for wideband music audio [14, 34, 16] or do not allow enough flexibility to be tested with gaps of different length [10, 15]. For instance, the authors' previous diffusion model [16], which corresponds to a prior iteration of the proposed method, could not be included as a baseline either, being unsuitable to work at sample rates higher than 22.05 kHz due to memory constraints.

The initial hypothesis is that the authors' method provides no advantage against the baselines when the gap is very short, as, in this case, stationary conditions can be safely assumed. However, as the duration of the gap increases, the problem gets more challenging, and the performance of the baselines is likely to degrade. On the other hand, a diffusion-based generative model should not suffer from this limitation and should be capable of generating audio content regardless of the gap length. Thus, the question the authors want to resolve is the following: *How does the performance of CQT-Diff+ compare to the baselines in terms of reconstruction quality as the gap length increases?*

4.1 Training

The authors train their model with the MusicNet dataset, a collection of 330 freely licensed classical music recordings sampled at 44.1 kHz. MusicNet is a multi-instrument dataset containing recordings from a wide variety of acoustical environments and recording conditions, representing a challenging and realistic scenario. The authors use a segment length of 4.17 s, limited by memory requirements. The training is performed using the Adam optimizer, with a learning rate of 2×10^{-4} and a batch size of four. The model is trained for roughly 500,000 iterations, taking approximately 5 d on a single NVIDIA A100 GPU. During training, the authors track an exponential moving average of the weights, which corresponds to the one used during testing.

4.2 Objective Evaluation

The authors first conduct an objective evaluation where they report three metrics. The first one is *log-spectral distance* (LSD) [62], a reference-based metric specified as

$$\mathrm{LSD} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\log|X_{t,k}|^2 - \log|\hat{X}_{t,k}|^2\right)^2}, \quad (14)$$

where $X_{t,k} = \mathrm{STFT}(\mathbf{x}_0)$ and $\hat{X}_{t,k} = \mathrm{STFT}(\hat{\mathbf{x}}_0)$ are the STFTs of the reference $\mathbf{x}_0$ and the restored audio signal $\hat{\mathbf{x}}_0$, respectively. For the STFT computation, an analysis window of $K = 2{,}048$ samples and a hop length of 512 samples is used. The LSD provides information about the reconstruction performance, with respect to the original signal.

The authors also use the *Objective Difference Grades* (ODG), estimated using the PEMO-Q auditory model [63]. This metric is also reference-based and aims to replicate the subjective difference grades that are obtained through a subjective listening test. The last metric is the reference-free *Fréchet Audio Distance* (FAD) that compares the statistics of a set of generated data against those of a reference dataset [64]. This metric has been demonstrated to correlate with perceptual audio quality [64]. In this case, the authors compare the distribution of inpainted audio signals with that of the original ones.

The results for different gap lengths are plotted in Fig. 5. In their evaluation, the authors used a subset of the MusicNet test set [65] comprising 60 randomly selected 4.17-s samples, each of them containing four equally spaced gaps. The test samples were not seen during the training of the proposed method. The authors deliberately chose a smaller test set due to computational limitations. They believe that

[3] https://github.com/eloimoliner/audio-inpainting-diffusion/tree/main/conf.

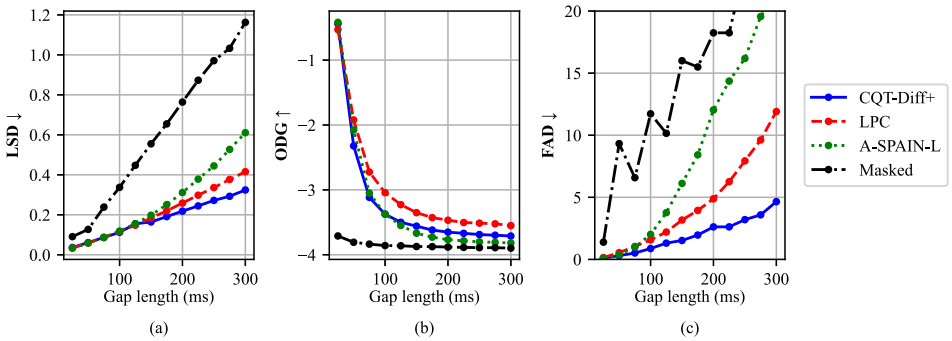

Fig. 5. Average objective metrics, including (a) LSD, (b) ODG, and (c) FAD, computed for various gap lengths from 25 to 300 ms. Lower is better for LSD and FAD, whereas higher is better for ODG. The proposed method (CQT-Diff+) obtained competitive results against the baselines in the reference-based metrics LSD and ODG, while being superior in terms of FAD.

expanding it would not notably alter the results, considering the added computational load.

Fig. 5(a) shows that, according to the LSD metric, for gaps smaller or equal to 100 ms, the proposed method yields a performance similar to the baselines and marginally outperforms them for longer gaps. The results of the ODG metric are presented in Fig. 5(b), showing how all methods obtain similar values for small gap lengths, with LPC performing marginally better. Above 100 ms, all the ODG values are below -3, which refers to annoying or very annoying sound quality [66]. Finally Fig. 5(c) shows the FAD results, where lower values (<5) indicate that the distribution of inpainted audio is statistically similar to the reference. The proposed method consistently achieves lower FAD values than the compared baselines, meaning that the inpainted audio is in-distribution with the rest. On the other hand, the baselines show a strong decline in terms of FAD as the gap size increases.

4.3 Subjective Evaluation

Since there is no guarantee that objective metrics provide reliable information about the perceived quality of the inpainting methods, the authors conducted a subjective listening experiment. The listening test was designed in accordance with the MUSHRA recommendation [67], using the webMUSHRA evaluation tool [68]. The participants were asked to rate, on a scale from 0 to 100, the perceptual similarity of each item with respect to the reference, which was the original audio sample (without gaps).

The conditions included a low anchor (a masked version of the reference with four gaps); three reconstructed versions of the low-anchor, produced using the LPC, A-SPAIN-L, and the proposed CQT-Diff+; and a hidden reference (the original unprocessed signal). Fig. 6 shows an example of all five conditions with two gaps. The listeners were allowed to loop and focus in detail on the gap locations. The items represented four gap lengths (50, 100, 200, and 300 ms) and 12 randomly picked 4.17-s examples from the MusicNet test set. The test contained a total of 48 pages of the five items above to be evaluated. In order to reduce the duration of the experiment and avoid listening fatigue, the test was split into two equal-length parts of 24 pages that were alternatively assigned to the participants. A total of 15 volunteers, all of whom reported no hearing loss, participated in the experiment. The average age of the listeners was 28 years. The audio examples used for the listening test are available at the companion webpage.[4]

The results of the listening test are presented in Fig. 7. Except for the 50-ms case where A-SPAIN-L was superior, LPC obtained higher scores than A-SPAIN-L. For the gap length of 50 ms, the proposed method obtained scores similar to the compared baselines, all of them close to 100. For the remaining evaluated gaps longer than 50 ms, the proposed method outperformed the baselines. The authors studied the statistical significance of the score differences between CQT-Diff+ and LPC through a Wilcoxon signed-rank test that gave a p value of 1.2×10^{-4}, 1×10^{-9}, and 3.5×10^{-9} for the gap lengths 100, 200, and 300 ms, respectively. The authors conclude that the differences are significant since the p values are way below 0.05. With the exception of gap lengths of 50 ms, the findings depicted in Fig. 7 reveal that the proposed approach consistently achieves median scores ranging from 50 to 80. These scores exhibit a proportional decrease as the gap length increases. In the case of the shortest 50-ms gaps, the median score for the CQT-Diff+ method reaches 100, indicating that it was difficult for listeners to find discerning differences in this particular test scenario.

Considering the test question, the listening test result can be interpreted so that the proposed diffusion model performs at least as well as the compared baselines for all gap lengths. The minimum gap length for which the reconstruction using the proposed method is better than the baselines is 100 ms; above that CQT-Diff+ consistently outperforms the baselines. The authors can conclude that, up to the length of 200 ms, the proposed CQT-Diff+ algorithm produces perceptually "good" audio inpainting (median scores above 60), although distinguishable from the reference in pairwise comparison. For the gap length of 300 ms, the proposed method offers "fair" sound quality.

To gain a deeper understanding of the subjective test results, the authors qualitatively analyze a specific exam-

[4] http://research.spa.aalto.fi/publications/papers/jaes-diffusion-inpainting/.

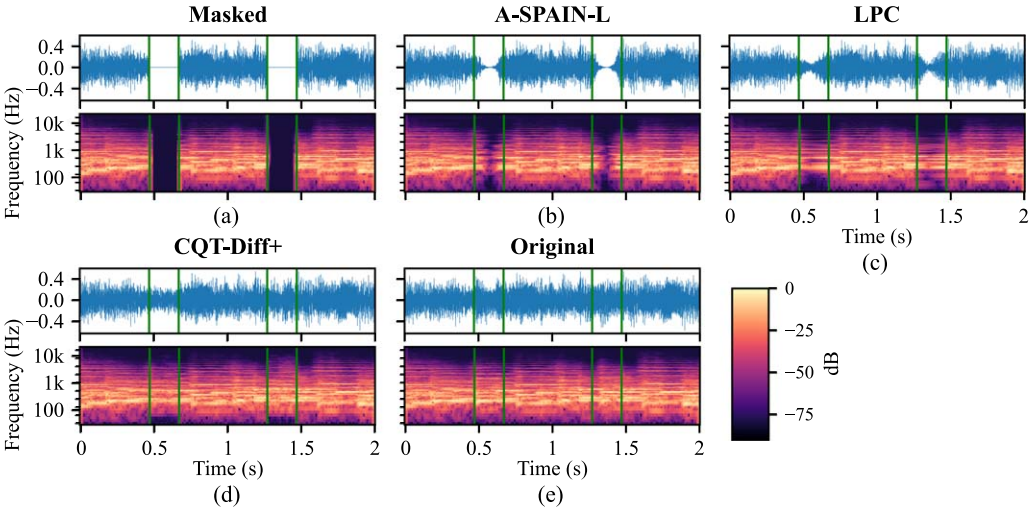

Fig. 6. Audio inpainting examples with two gaps of 200 ms, for all five signal types included in the listening test, including waveform and spectrogram representations. The start and end timestamps of the gaps are highlighted with vertical lines. The masked waveform (a) and the original waveform (e) are included for comparison. The proposed method, CQT-Diff+ (d), produces more coherent and realistic reconstructions than the compared baselines (b) and (c).

ple depicted in Fig. 6. This figure exhibits waveform and spectrogram representations of a masked music signal in Fig. 6(a), along with three reconstructed versions, Figs. 6(b), (c), and (d), and the original signal for reference in Fig. 6(e), all with a gap length of 200 ms. Upon examination, A-SPAIN-L evidently generates an attenuated reconstruction that fades out toward the middle of the gap and fades in again before reaching the gap's end. In practice the method shortens the dropout but cannot fill it. The decay observed in the reconstruction arises due to the sparsity penalty that restricts the generation of content further into the gap. On the other hand, relying on extrapolation, LPC suffers less from this issue. Nevertheless this method is only capable of extending stationary sounds and cannot create new attacks and events. Consequently the reconstructions produced by LPC often sound artificial.

Visually the reconstruction generated with the proposed CQT-Diff+ algorithm fills the gap in a credible manner in Fig. 6(d). However the comparison with the original signal shown in Fig. 6(e) reveals that the reconstruction of the proposed methods is not exact, since the two waveforms look different.

5 CONCLUSION

This paper presents a novel audio inpainting method CQT-Diff+ that is based on recent diffusion models. For the reconstruction of short gaps of 50 ms or less, the proposed method works as well as a previous high-quality inpainting method. For longer gaps, from 100 to 300 ms, the CQT-Diff+ method outperforms the baseline algorithms and retains good or fair quality.

One limitation of the method presented in this paper is that the performance is limited to the types of audio recordings seen during the training, in this case, classical music. To demonstrate the versatility of the proposed approach, the authors also report, in the form of audio examples in the companion webpage, results obtained with a model trained on a large variety of sound effects. However these results were excluded from the evaluation in this paper. In the future, the generalizability of the diffusion-based audio inpainting technique should be evaluated by considering models trained on a wider variety of audio recordings.

Another promising direction for future work could involve providing the CQT-Diff+ model with conditional information. In the results of the authors' tests reported else-

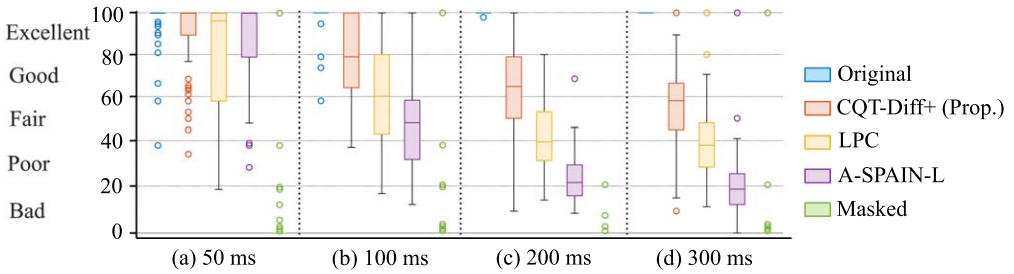

Fig. 7. Boxplot diagram representing the results of the subjective listening experiment for evaluated gap lengths of 50 ms (a), 100 ms (b), 200 ms (c), and 300 ms (d).

where, a generative method based on a diffusion model created plausible content and new events, when the gaps were 1.5 s long [16]. However, due to the lack of contextual information, the results were difficult to control. A conditional diffusion approach may offer a higher degree of control over the results, especially when dealing with extremely long gaps.

6 ACKNOWLEDGMENT

This research is part of the activities of the Nordic Sound and Music Computing Network (NordForsk project no. 86892). The authors acknowledge the computational resources provided by the Aalto Science-IT project. The authors thank the volunteers who participated in the listening test and are grateful to Luis Costa for proofreading.

7 REFERENCES

[1] A. Adler, V. Emiya, M. G. Jafari, et al., "Audio Inpainting," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 3, pp. 922–932 (2012 Mar.). https://doi.org/10.1109/TASL.2011.2168211.

[2] S. J. Godsill and P. J. Rayner, *Digital Audio Restoration* (Springer, Berlin, Germany, 1998).

[3] A. Ragano, E. Benetos, and A. Hines, "Automatic Quality Assessment of Digitized and Restored Sound Archives," *J. Audio Eng. Soc.*, vol. 70, no. 4, pp. 252–270 (2022 Apr.). https://doi.org/10.17743/jaes.2022.0002.

[4] A. J. E. M. Janssen, R. N. J. Veldhuis, and L. B. Vries, "Adaptive Interpolation of Discrete-Time Signals That Can Be Modeled as Autoregressive Processes," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, no. 2, pp. 317–330 (1986 Apr.). https://doi.org/10.1109/TASSP.1986.1164824.

[5] D. Goodman, G. Lockhart, O. Wasem, and W.-C. Wong, "Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, no. 6, pp. 1440–1448 (1986 Dec.). https://doi.org/10.1109/TASSP.1986.1164984.

[6] T. Bazin, G. Hadjeres, P. Esling, and M. Malt, "Spectrogram Inpainting for Interactive Generation of Instrument Sounds," in *Proceedings of the Joint Conference on AI Music Creativity* (Stockholm, Sweden) (2020 Oct.).

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Proceedings of theConference on Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 6840–6851 (Vancouver, Canada) (2020 Dec.).

[8] Y. Song, J. Sohl-Dickstein, D. P. Kingma, et al., "Score-Based Generative Modeling Through Stochastic Differential Equations," in *Proceedings of theInternational Conference on Learning Representations (ICLR)* (Vienna, Austria) (2021 May).

[9] F. Lieb and H.-G. Stark, "Audio Inpainting: Evaluation of Time-Frequency Representations and Structured Sparsity Approaches," *Signal Process.*, vol. 153, pp. 291–299 (2018 Dec.). https://doi.org/10.1016/j.sigpro.2018.07.012.

[10] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak, "A Context Encoder for Audio Inpainting," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 12, pp. 2362–2372 (2019 Dec.). https://doi.org/10.1109/TASLP.2019.2947232.

[11] O. Mokrý and P. Rajmic, "Audio Inpainting: Revisited and Reweighted," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2906–2918 (2020 Oct.). https://doi.org/10.1109/TASLP.2020.3030486.

[12] G. Tauböck, S. Rajbamshi, and P. Balazs, "Dictionary Learning for Sparse Audio Inpainting," *IEEE J. Sel. Top. Signal Process.*, vol. 15, no. 1, pp. 104–119 (2021 Jan.). https://doi.org/10.1109/JSTSP.2020.3046422.

[13] O. Mokrý, P. Záviška, P. Rajmic, and V. Veselý, "Introducing SPAIN (Sparse Audio Inpainter)," in *Proceedings of the 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5 (A Coruña, Spain) (2019 Sep.). https://doi.org/10.23919/EUSIPCO.2019.8902560.

[14] P. P. Ebner and A. Eltelt, "Audio Inpainting With Generative Adversarial Network," *arXiv preprint arXiv:2003.07704* (2020 Mar.). https://doi.org/10.48550/arXiv.2003.07704.

[15] 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 Jan.). https://doi.org/10.1109/JSTSP.2020.3037506.

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

[17] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising Diffusion Restoration Models," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 23593–23606 (New Orleans, LA) (2022 Dec.).

[18] 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)* (Kigali, Rwanda) (2023 May).

[19] K. Liu, W. Gan, and C. Yuan, "MAID: A Conditional Diffusion Model for Long Music Audio Inpainting," 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/ICASSP49357.2023.10095769.

[20] W. Etter, "Restoration of a Discrete-Time Signal Segment by Interpolation Based on the Left-Sided and Right-Sided Autoregressive Parameters," *IEEE Trans. Signal Process.*, vol. 44, no. 5, pp. 1124–1135 (1996 May). https://doi.org/10.1109/78.502326.

[21] P. A. A. Esquef, V. Välimäki, K. Roth, and I. Kauppinen, "Interpolation of Long Gaps in Audio Signals Using the Warped Burg's Method," in *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx)*, pp. 8–11 (London, UK) (2003 Sep.).

[22] I. Kauppinen, J. Kauppinen, and P. Saarinen, "A Method for Long Extrapolation of Audio Signals," *J. Audio Eng. Soc.*, vol. 49, no. 12, pp. 1167–1180 (2001 Dec.).

[23] I. Kauppinen and K. Roth, "Audio Signal Extrapolation—Theory and Applications," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 105–110 (Hamburg, Germany) (2002 Sep.).

[24] I. Kauppinen and J. Kauppinen, "Reconstruction Method for Missing or Damaged Long Portions in Audio Signal," *J. Audio Eng. Soc.*, vol. 50, no. 7/8, pp. 594–602 (2002 Jul.).

[25] P. A. A. Esquef and L. W. P. Biscainho, "An Efficient Model-Based Multirate Method for Reconstruction of Audio Signals Across Long Gaps," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1391–1400 (2006 Jul.). https://doi.org/10.1109/TSA.2005.858018.

[26] P. Smaragdis, B. Raj, and M. Shashanka, "Missing Data Imputation for Spectral Audio Signals," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6 (Grenoble, France) (2009 Sep.). https://doi.org/10.1109/MLSP.2009.5306194.

[27] S. Rajbamshi, G. Tauböck, N. Holighaus, and P. Balazs, "Audio Inpainting via ℓ_1-Minimization and Dictionary Learning," in *Proceedings of the 29th European Signal Processing Conference (EUSIPCO)*, pp. 2149–2153 (Dublin, Ireland) (2021 Aug.). https://doi.org/10.23919/EUSIPCO54536.2021.9616132.

[28] O. Mokrý, P. Magron, T. Oberlin, and C. Févotte, "Algorithms for Audio Inpainting Based on Probabilistic Nonnegative Matrix Factorization," *Signal Process.*, vol. 206, paper 108905 (2023 May). https://doi.org/10.1016/j.sigpro.2022.108905.

[29] M. Lagrange, S. Marchand, and J.-B. Rault, "Long Interpolation of Audio Signals Using Linear Prediction in Sinusoidal Modeling," *J. Audio Eng. Soc.*, vol. 53, no. 10, pp. 891–905 (2005 Oct.).

[30] N. Perraudin, N. Holighaus, P. Majdak, and P. Balazs, "Inpainting of Long Audio Segments With Similarity Graphs," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 6, pp. 1083–1094 (2018 Jun.). https://doi.org/10.1109/TASLP.2018.2809864.

[31] G. Greshler, T. Shaham, and T. Michaeli, "Catch-a-Waveform: Learning to Generate Audio From a Single Short Example," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 20916–20928 (Online) (2021 Dec.).

[32] B.-K. Lee and J.-H. Chang, "Packet Loss Concealment Based on Deep Neural Networks for Digital Speech Transmission," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 2, pp. 378–387 (2015 Dec.). https://doi.org/10.1109/TASLP.2015.2509780.

[33] J. Lin, Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen, "A Time-Domain Convolutional Recurrent Network for Packet Loss Concealment," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7148–7152 (Toronto, Canada) (2021 May). https://doi.org/10.1109/ICASSP39728.2021.9413595.

[34] S. Pascual, J. Serrà, and J. Pons, "Adversarial Auto-Encoding for Packet Loss Concealment," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 71–75 (New Paltz, NY) (2021 Oct.). https://doi.org/10.1109/WASPAA52581.2021.9632730.

[35] L. Ou and Y. Chen, "Concealing Audio Packet Loss Using Frequency-Consistent Generative Adversarial Networks," in *Proceedings of the 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, pp. 826–831 (Paris, France) (2022 May).

[36] G. Morrone, D. Michelsanti, Z.-H. Tan, and J. Jensen, "Audio-Visual Speech Inpainting With Deep Learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6653–6657 (Toronto, Canada) (2021 Jun.). https://doi.org/10.1109/ICASSP39728.2021.9413488.

[37] K. W. Cheuk, R. Sawata, T. Uesaka, et al., "DiffRoll: Diffusion-Based Generative Music Transcription With Unsupervised Pretraining Capability," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (Rhodes Island, Greece) (2023 Jun.). https://doi.org/10.1109/ICASSP49357.2023.10095935.

[38] Z. Borsos, M. Sharifi, and M. Tagliasacchi, "SpeechPainter: Text-Conditioned Speech Inpainting," in *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 431–435 (Incheon, South Korea) (2022 Sep.).

[39] Y. Wang, Z. Ju, X. Tan, et al., "AUDIT: Audio Editing by Following Instructions With Latent Diffusion Models," *arXiv preprint arXiv:2304.00830* (2023 Apr.).

[40] Y. Wang, J. Yu, and J. Zhang, "Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model," in *Proceedings of the International Conference on Learning Representations (ICLR)* (Online) (2022 May).

[41] A. Lugmayr, M. Danelljan, A. Romero, et al., "Repaint: Inpainting Using Denoising Diffusion Probabilistic Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471 (New Orleans, LA) (2022 Jun.).

[42] J. Song, A. Vahdat, M. Mardani, and J. Kautz, "Pseudoinverse-Guided Diffusion Models for Inverse Problems," in *Proceedings of the International Conference on Learning Representations (ICLR)* (Online) (2022 May).

[43] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 8780–8794 (Online) (2021 Dec.).

[44] T. Karras, S. Laine, M. Aittala, et al., "Analyzing and Improving the Image Quality of StyleGAN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119 (Seattle, WA) (2020 Jun.).

[45] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A Versatile Diffusion Model for Audio Synthesis," in *Proceedings of the International Conference on Learning Representations (ICLR)* (Vienna, Austria) (2021 May).

[46] 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.

[47] J. Ho, T. Salimans, A. Gritsenko, et al., "Video Diffusion Models," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 8633–8646 (New Orleans, LA) (2022 Dec.).

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

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

[50] H. Kim, S. Kim, and S. Yoon, "Guided-TTS: A Diffusion Model for Text-to-Speech via Classifier Guidance," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, pp. 11119–11133 (Baltimore, MD) (2022 Jul.).

[51] H. Chung, B. Sim, D. Ryu, and J. C. Ye, "Improving Diffusion Models for Inverse Problems Using Manifold Constraints," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 25683–25696 (New Orleans, LA) (2022 Dec.).

[52] G. A. Velasco, N. Holighaus, M. Dörfler, and T. Grill, "Constructing an Invertible Constant-Q Transform With Non-Stationary Gabor Frames," in *Proceedings of the InternationalConference on Digital Audio Effects (DAFX)*, pp. 93–99 (Paris, France) (2011 Sep.).

[53] N. Holighaus, M. Dörfler, G. A. Velasco, and T. Grill, "A Framework for Invertible, Real-Time Constant-Q Transforms," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 4, pp. 775–785 (2012 Apr.). https://doi.org/10.1109/TASL.2012.2234114.

[54] C. Schörkhuber and A. Klapuri, "Constant-Q Transform Toolbox for Music Processing," in *Proceedings of the 7th Sound and Music Computing Conference*, pp. 3–64 (Barcelona, Spain) (2010 Jul.).

[55] C. Schörkhuber, A. Klapuri, and A. Sontacchi, "Pitch Shifting of Audio Signals Using the Constant-Q Transform," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (York, UK) (2012 Jul.).

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

[57] H. Wu, K. Tan, B. Xu, A. Kumar, and D. Wong, "Rethinking Complex-Valued Deep Neural Networks for Monaural Speech Enhancement," in *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3889–3893 (York, UK) (2023 Jan.).

[58] M. Tancik, P. Srinivasan, B. Mildenhall, et al., "Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 7537–7547 (Vancouver, Canada) (2020 Dec.).

[59] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual Reasoning With a General Conditioning Layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, pp. 3942–3951 (New Orleans, LA) (2018 Mar.).

[60] M. A. Islam, S. Jia, and N. D. Bruce, "How Much Position Information Do Convolutional Neural Networks Encode?" in *Proceedings of the International Conference on Learning Representations (ICLR)* (Addis Ababa, Ethiopia) (2020 May).

[61] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is All You Need," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 6000–6010 (Long Beach, CA) (2017 Dec.).

[62] A. Gray and J. Markel, "Distance Measures for Speech Processing," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 24, no. 5, pp. 380–391 (1976 Oct.). https://doi.org/10.1109/TASSP.1976.1162849.

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

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

[65] J. Thickstun, Z. Harchaoui, and S. Kakade, "Learning Features of Music From Scratch," in *Proceedings International Conference on Learning Representations (ICLR)* (San Juan, Puerto Rico) (2016 May).

[66] ITU, "Method for Objective Measurements of Perceived Audio Quality," *ITU Recommendation BS.1387-2* (2023 May).

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

[68] M. Schoeffler, S. Bartoschek, F.-R. Stöter, et al., "WebMUSHRA—A Comprehensive Framework for Web-Based Listening Tests," *J. Open Res. Softw.*, vol. 6, no. 1, paper 8 (2018 Feb.). https://doi.org/10.5334/jors.187.

THE AUTHORS

Eloi Moliner

Vesa Välimäki

Eloi Moliner received his B.Sc. degree in Telecommunications Technologies and Services Engineering from the Polytechnic University of Catalonia, Spain, in 2018 and his M.Sc. degree in Telecommunications Engineering from the same university in 2021. He is currently a doctoral candidate at the Acoustics Lab of Aalto University in Espoo, Finland. His research interests include digital audio restoration and audio applications of machine learning. He is the winner of the Best Student Paper Award of the 2023 IEEE ICASSP conference.

•

Vesa Välimäki is Full Professor of audio signal processing and Vice Dean for Research at Aalto University, Espoo, Finland. He received his D.Sc. degree from the Helsinki University of Technology in 1995. In 1996 he was a Postdoctoral Research Fellow at the University of Westminster, London, UK. In 2008–2009 he was a visiting scholar at Stanford University. He is a Fellow of the AES, IEEE, and Asia-Pacific Artificial Intelligence Association. Prof. Välimäki is the Editor-in-Chief of the *Journal of the Audio Engineering Society*.