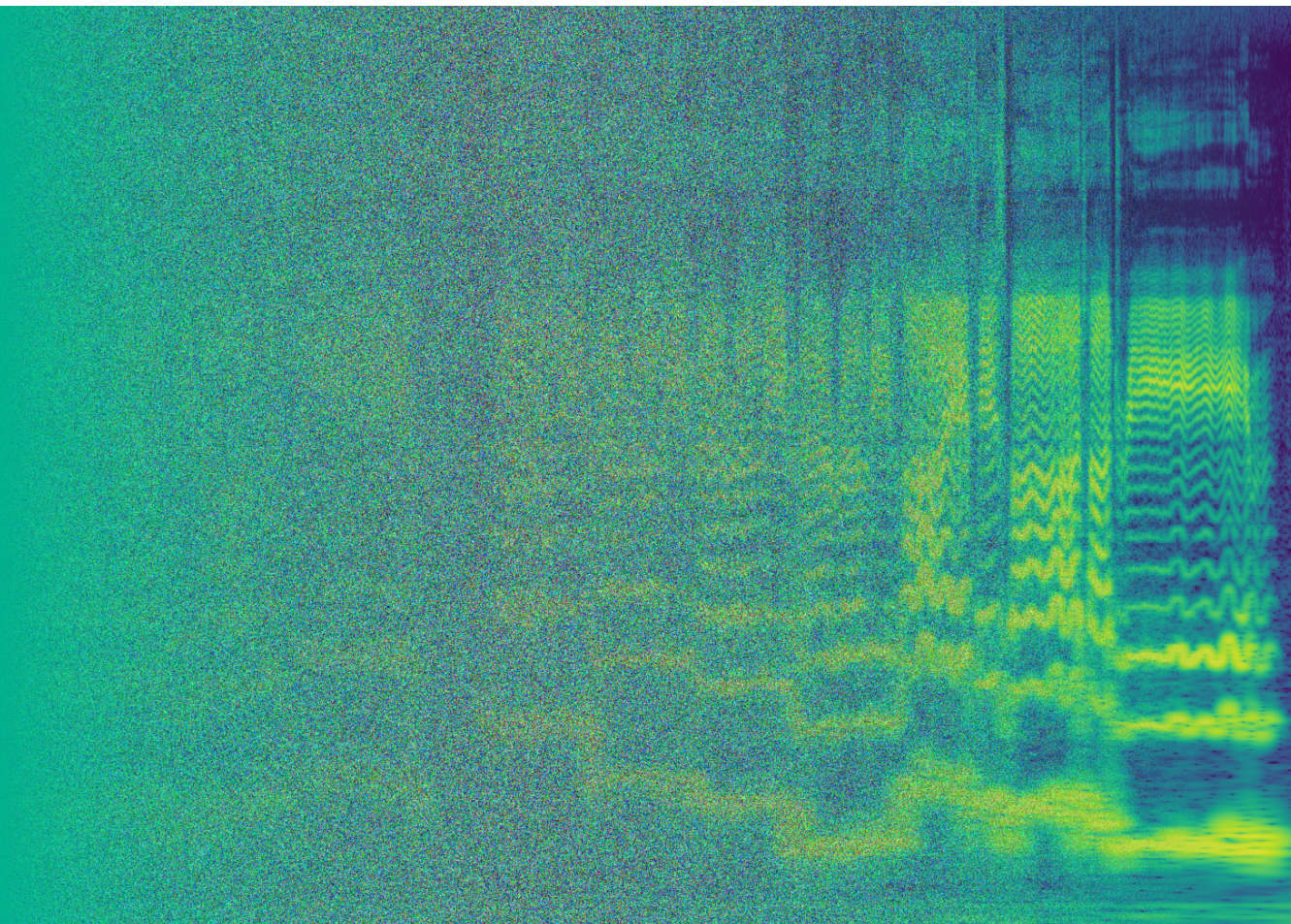
Unsupervised audio enhancement with diffusion-based generative models

Eloi Moliner Juanpere

Aalto University



Aalto University publication series
Doctoral Theses 138/2025

Unsupervised audio enhancement with diffusion-based generative models

Eloi Moliner Juanpere

A doctoral thesis completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall A208d Jeti, of the school on 22 August 2025 at 12:00.

Aalto University
School of Electrical Engineering
Department of Information and Communications Engineering
Aalto Acoustics Lab, Audio Signal Processing Group

Supervising professor
Prof. Vesa Välimäki, Aalto University, Finland

Thesis advisors
Prof. Vesa Välimäki, Aalto University, Finland

Preliminary examiners
Dr. Nicholas J. Bryan, Adobe Research, San Francisco, CA, USA
Prof. Bożena Kostek, Gdansk University of Technology, Poland

Opponent
Prof. Bożena Kostek, Gdansk University of Technology, Poland

Aalto University publication series
Doctoral Theses 138/2025

Image on the cover: Eloi Moliner

ISBN 978-952-64-2646-4 (soft cover)
ISBN 978-952-64-2645-7 (PDF)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (PDF)

http://urn.fi/URN:SBN:978-952-64-2645-7

Unigrafia Oy
Helsinki 2025

Aalto University

ABSTRACT

Author Eloi Moliner Juanpere

Name of the doctoral thesis
Unsupervised audio enhancement with diffusion-based generative models

Article-based thesis

Number of pages 135

Keywords audio restoration, diffusion models,

Audio recordings are often compromised by noise, reverberation, and other distortions, leading to loss of quality. Examples of this include historical music recordings affected by the degradation of analog media or speech recordings where reverberation reduces intelligibility. Audio enhancement and restoration techniques are used to recover and improve the acoustic quality of these recordings. At the time of this thesis, the state-of-the-art audio restoration methods are predominantly data-driven, with deep generative models demonstrating exceptional expressivity. However, most of these approaches rely on supervised learning, which, while successful, comes with inherent limitations. These include a restricted generalization to unseen degradations, as well as the need to train task-specific models for each different restoration scenario.

This thesis explores an alternative unsupervised approach that employs unconditional generative models, specifically diffusion models. In this context, a single generative model, trained without prior knowledge of specific degradation processes, can be adapted to an endless variety of restoration tasks during inference, thus overcoming the limitations of task-specific supervised models. The first and second publications included in this thesis demonstrate the effectiveness of this approach in several known restoration problems, including music bandwidth extension, inpainting, and declipping, supported by objective and subjective evaluations.

This thesis also addresses blind restoration problems, where the characteristics of the degradation are unknown. The third publication presents a blind approach to audio bandwidth extension for historical music restoration, where the lowpass filter degradation is automatically estimated and iteratively refined during the generation process. The fourth publication extends this work to generative equalization, enabling both the correction of spectral coloration and the regeneration of missing content. This method has shown significant improvements in the restoration of historical gramophone recordings, particularly for piano and singing voice performances.

The final two publications focus on single-channel blind speech dereverberation. Here, speech signals affected by room reverberation are enhanced using a diffusion model trained on anechoic speech, combined with a parametric subband filtering model of room impulse responses. This approach allows for simultaneous estimation of both anechoic speech and the room impulse response. The method is evaluated on multiple datasets through objective and subjective experiments, demonstrating performance that matches or surpasses supervised baselines, particularly in conditions that differ from the training data.

Preface

The work presented in this thesis was conducted at the Aalto Acoustics Lab, Espoo, Finland, between May 2021 and May 2025. These four years have been a deeply fulfilling and enriching experience—both professionally and personally. I look back on this journey with sincere gratitude. This work has been shaped and inspired by many people I had the privilege to meet along the way. It would not have been possible without their support, collaboration, and encouragement.

First and foremost, I would like to thank my supervisor, Professor Vesa Välimäki, for making all of this possible. His invaluable guidance, consistent support, and genuine kindness have been instrumental throughout this journey. I feel truly fortunate to have worked under his supervision. I am grateful to the pre-examiners, Dr. Nicholas J. Bryan and Professor Bożena Kostek, for taking the time to review this thesis and for their constructive and insightful feedback. I am especially thankful to Professor Bożena Kostek for kindly agreeing to serve as my opponent in the public defense of this thesis.

I would also like to sincerely thank all my co-authors, whose collaborations greatly contributed to shaping the contents of this thesis. Special thanks go to Dr. Jean-Marie Lemercier for a highly stimulating and fruitful collaboration, from which I learned a great deal about conducting rigorous research. I am also grateful to Professor Jaakko Lehtinen for a brief but impactful exchange that helped refine the focus of this research. My thanks extend to Professor Filip Elvander for his insightful guidance and expertise. Many thanks to Maija Turunen for an inspiring interdisciplinary collaboration that broadened my perspective.

I would further like to acknowledge Michal Svento—it was a pleasure to work together, and I am especially grateful for the collaboration we shared. Although the publications we co-authored are not included in this thesis, they represent work that could very well have been a part of it.

My sincere thanks go to everyone at the Aalto Acoustics Lab over the past four-plus years — those I've had the privilege to work with, exchange research ideas with, share beers and leisure time with, or play foosball

matches with. Thank you for making it such a vibrant, welcoming, and inspiring research community. I truly couldn't have imagined a better environment in which to carry out this work. Special thanks to the past and present members of the Audio Signal Processing team, including all visitors and collaborators. I would also like to thank Ravintola Qvarkki for reliably providing food, and Olarin Panimo for hosting the Thursday after-work gatherings.

I also wish to express my sincere thanks to the teams at Microsoft Research and Sony AI for the opportunity to carry out internships during my doctoral studies. These experiences were both intellectually stimulating and personally enriching, providing valuable insights into diverse research cultures. I am also thankful for the research partnership with Nokia Technologies, which brought practical perspectives and contributed meaningfully to this work. My sincere appreciation also goes to the Nokia Foundation for their financial support.

Finally, I would like to thank my parents for their unwavering love and support, and for encouraging me in all of my life decisions, including moving abroad to pursue this journey. My deepest thanks go to my girlfriend, Sara, for her constant support, understanding, and for helping me stay grounded throughout these four years.

Espoo, January 2025,

Eloi Moliner Juanpere

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Eloi Moliner, Jaakko Lehtinen and Vesa Välimäki. Solving Audio Inverse Problems with a Diffusion Model. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes, Greece, pp. 1-5, June 2023.

II Eloi Moliner and Vesa Välimäki. Diffusion-Based Audio Inpainting. *Journal of the Audio Engineering Society*, Vol. 72, No. 3, pp. 100-113, March 2024.

III Eloi Moliner, Filip Elvander and Vesa Välimäki. Blind Audio Bandwidth Extension: A Diffusion-based Zero-shot Approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 32, pp. 5092-5105, November 2024.

IV Eloi Moliner, Maija Turunen, Filip Elvander, Vesa Välimäki. A Diffusion-based Generative Equalizer for Music Restoration. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pp. 25-32, September 2024.

V Eloi Moliner, Jean-Marie Lemercier, Simon Welker, Timo Gerkmann, Vesa Välimäki. BUDDy: Single-channel Blind Unsupervised Dereverberation with Diffusion Models. In *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 120-124, September 2024.

VI Jean-Marie Lemercier, Eloi Moliner, Simon Welker, Vesa Välimäki, Timo Gerkmann. Unsupervised Blind Joint Dereverberation and Room Acoustics Estimation with Diffusion Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, June 2025.

Author's Contribution

Publication I: "Solving Audio Inverse Problems with a Diffusion Model"

The author planned the study in collaboration with the co-authors. The author implemented the proposed methods, performed the experiments, and served as the principal writer of the manuscript, with contributions from the co-authors. Jaakko Lehtinen assisted in planning the study and contributed to writing the manuscript. Vesa Välimäki provided supervision, helped plan the study, and also contributed to writing the manuscript.

Publication II: "Diffusion-Based Audio Inpainting"

The author planned the study, implemented the proposed algorithms, and conducted the experiments, including the listening tests. The author wrote most of the article, with the input and support of Vesa Välimäki, who provided guidance and supervision throughout the study and assisted in the writing of the manuscript.

Publication III: "Blind Audio Bandwidth Extension: A Diffusion-based Zero-shot Approach"

The author planned the study, implemented the proposed algorithm, and conducted all experiments, including the listening tests. The author wrote most of the article with input and support from the coauthors. Filip Elvander and Vesa Välimäki assisted in presenting the methods, including the design of figures and tables, and contributed by reviewing and editing the manuscript.

Publication IV: "A Diffusion-based Generative Equalizer for Music Restoration"

The author planned the study in collaboration with the co-authors. Maija Turunen proposed the initial idea of using generative models for historical vocal restoration, authored the majority of the discussion presented in Section 5, and assisted in designing the experiments on vocal restoration. The author implemented the proposed method, conducted the experiments, and wrote most of the remaining sections of the paper, with input and support from Filip Elvander and Vesa Välimäki.

Publication V: "BUDDy: Single-channel Blind Unsupervised Dereverberation with Diffusion Models"

The author and Jean-Marie Lemercier contributed equally to the planning and development of the method, as well as to the writing of the manuscript. Together, they co-wrote Sections 1, 2, and 3. The author developed a significant portion of the BUDDy algorithm, while Jean-Marie Lemercier trained all baseline models, conducted several ablation studies, and performed the final speech dereverberation evaluation, authoring Sections 4 and 5. Simon Welker provided valuable feedback through discussions on the methods proposed in the paper. Timo Gerkmann and Vesa Välimäki contributed insights into the experimental validation and mathematical derivations, and reviewed the manuscript.

Publication VI: "Unsupervised Blind Joint Dereverberation and Room Acoustics Estimation with Diffusion Models"

The author and Jean-Marie Lemercier contributed equally to the development of the proposed method and the writing of the manuscript. Together, they co-wrote Sections III and IV. The author conducted the singing voice dereverberation experiments, subjective evaluations, and room impulse response estimation validation, and authored Sections V.B and V.C. Jean-Marie Lemercier carried out the speech dereverberation and robustness experiments and authored Sections I, II, and V.A. Simon Welker provided valuable feedback through discussions on the methods developed in the paper. Timo Gerkmann and Vesa Välimäki contributed insights into experimental validation and mathematical derivations and reviewed the manuscript.

List of Figures

Abbreviations

BABE Blind Audio Bandwidth Extension

BUDDy Blind Unsupervised Dereverberation with Diffusion models

CQT Constant-Q Transform

DNS-MOS Deep Noise Suppression - Mean Opinion Score

DPS Diffusion Posterior Sampling

ESTOI Extended Short Time Objective Intelligibility

FAD Fréchet Audio Distance

FFT Fast Fourier Transform

GAN Generative Adversarial Networks

ICQT Inverse Constant-Q Transform

ISTFT Inverse Short-Time Fourier Transform

LSD Log-Spectral Distance

ODE Ordinary Differential Equation

ODG Objective Difference Grade

PESQ Perceptual Evaluation of Speech Quality

RIR Room Impulse Response

SDE Stochastic Differential Equation

STFT Short-Time Fourier Transform

Symbols

Latin Letters:

$\mathcal{A}(\cdot)$ Forward operator

$\mathcal{A}\psi(\cdot)$ Forward operator with parameters ψ

c Maximum amplitude before clipping

$C(\cdot, \cdot)$ Cost function

d Differential operator

$D_\theta(\cdot)$ Denoiser neural network

$\mathbb{E}(\cdot)$ Expectation operator

$\mathcal{F}(\cdot)$ Time-frequency forward transform

$\mathcal{F}^{-1}(\cdot)$ Time-frequency inverse transform

$\mathbf{f}(\cdot)$ Drift term in a stochastic process

$g(\cdot)$ Diffusion coefficient

$\mathbf{h}$ Linear impulse response of a system

$\mathbf{m}$ Binary mask

$\mathcal{N}(\cdot; \mu, \Sigma)$ Multivariate Gaussian distribution with mean μ and covariance matrix Σ

$p(\cdot)$ Probability distribution

$p(\cdot|\cdot)$ Conditional probability distribution

$s_\theta(\cdot)$ Score neural network

T Maximum diffusion time

$\mathbf{x}_0$ Clean target audio signal

$\hat{\mathbf{x}}_0$ Estimate of the target signal

$\mathbf{x}_\tau$ Noisy state at time τ

$\mathbf{y}$ Degraded measured signal

$\mathbf{z}$ Measurement noise

Greek Letters:

ω Classifier-free guidance scaling hyperparameter

ψ Parameters of the forward operator

$\sigma^2(\cdot)$ Noise variance schedule

τ Diffusion time variable

θ Trainable network parameters

$\zeta(\cdot)$ Weighting parameter for likelihood approximation

Other Symbols:

∇ Gradient operator

$\odot$ Element-wise (Hadamard) product

$*$ Discrete convolution operator

1. Introduction

Audio recordings, whether it is music recorded on a gramophone disk a century ago or speech captured by a modern digital device, are often degraded by factors like noise, reverberation, and distortion, leading to a loss of clarity and detail. Restoring and enhancing these recordings is essential not only to recover lost information but also to preserve their original character and improve the listener's experience. In this thesis, our goal is to design methods that maintain the integrity of the original content while enhancing quality, whether by restoring the frequency range, filling in missing segments, reducing noise, or improving speech intelligibility.

Over the past few decades, extensive research has been conducted on audio restoration [2]. Machine learning, particularly deep neural networks [3], has become the leading approach in this field due to its ability to model complex, high-dimensional relationships in audio data [4–8]. Supervised learning methods dominate these efforts, relying on paired datasets of clean and degraded audio to train deep neural networks that predict clean signals from degraded inputs. These approaches are typically framed as regression tasks, where a similarity criterion—often referred to as the "loss function"—guides the training process. The model's weights are optimized using gradient descent methods to minimize the loss [3].

Most of the popular methods in the literature prior to this thesis [4–7] were predictive, focusing on directly mapping degraded audio to its clean counterpart. While successful in many cases, these methods have limitations, particularly when dealing with ill-conditioned scenarios where multiple solutions are possible. In such cases, predictive models often regress toward an average solution, which may not yield a plausible restoration.

In contrast, there has been a growing interest in generative approaches that aim to learn the underlying distribution of the data, rather than making direct predictions. One well-received approach in this area involves adversarial training, in which a discriminator is trained simultaneously to evaluate the difference between the generated output and the true data distribution [9]. More recently, diffusion models have emerged as a promising class of generative models, offering a robust framework for

addressing the challenges of audio restoration [1].

Another major challenge in audio restoration with supervised methods arises from the reliance on paired datasets for training. In many real-world scenarios, such as historical music restoration, obtaining high-quality paired datasets of degraded and clean audio is difficult or even impossible. A common workaround is to simulate paired data by artificially introducing degradations to clean recordings, despite efforts to design more accurate degradation simulation methods [10, 11]. However, this approach has a significant drawback: simulated degradations often fail to capture the full complexity of real-world conditions, although the design of accurate degradation simulation methods has been explored [10, 11]. As a result, models trained on such synthetic data tend to struggle when faced with unseen or more complex degradation patterns, leading to poor generalization to real-world tasks [12, 13].

In this thesis, we adopt a general perspective based on inverse problems [14], focusing on purely generative, unsupervised approaches to address a range of audio restoration challenges. By framing these tasks within the context of inverse problems, we propose the use of diffusion models as a novel solution that does not rely on paired training data. This approach offers significant advantages over traditional methods, particularly in ill-conditioned cases and where generalization to unseen degradation patterns is crucial.

This thesis includes a set of six publications, each contributing to different aspects of audio restoration. Publication I proposes a general framework for audio restoration that is valid when the degradation process is known. Publication II explores the problem of audio inpainting, which involves filling in lost audio segments of varying lengths. Publications III and IV focus on blind audio bandwidth extension, where the spectral degradation is unknown and is jointly estimated; this work is particularly applicable to historical music restoration. Publications V and VI tackle the problem of speech dereverberation, introducing a novel unsupervised approach to address this challenge.

The introductory portion of this thesis is organized as follows: Chapter 2 introduces the inverse problem formulation used in the publications and offers a comprehensive review of relevant restoration problems. Chapter 3 presents diffusion models, detailing their principles and applications to audio restoration. Chapter 4 provides a summary of the included publications, and Chapter 5 concludes the thesis.

2. Common Inverse Problems in Audio Restoration

In this chapter, audio restoration is presented through the lens of inverse problems. The discussion focuses on specific restoration challenges addressed in this thesis, outlining their unique difficulties and reviewing state-of-the-art approaches to tackle them.

2.1 General Formulation

Consider the following family of inverse problems [14], where the observed measurements $\mathbf{y}$ are related to the unknown signal $\mathbf{x}_0$ through a forward (or degradation) operator $\mathcal{A}(\cdot)$ and an additive noise disturbance $\mathbf{z}$:

$$\mathbf{y} = \mathcal{A}(\mathbf{x}_0) + \mathbf{z}. \tag{2.1}$$

In the context of audio restoration, $\mathbf{y} \in \mathbb{R}^L$ represents the degraded audio signal of L samples, while $\mathbf{x}_0 \in \mathbb{R}^L$ corresponds to the high-quality, original audio signal that we aim to recover. The operator $\mathcal{A}(\cdot) : \mathbb{R}^L \to \mathbb{R}^L$ represents the degradation process, such as masking, distortion, or other artifacts, and $\mathbf{z} \in \mathbb{R}^L$ accounts for any additional additive noise or random perturbations introduced during measurement or transmission. The objective is usually to estimate the original signal $\hat{\mathbf{x}}_0 \approx \mathbf{x}_0$. Depending on how the forward operator $\mathcal{A}(\cdot)$ is defined, this model can generalize to a wide range of problems. Figure 2.1 illustrates distinct cases that hold significant relevance to this thesis.

In many cases, the operator $\mathcal{A}(\cdot)$ leads to an ill-posed problem. An inverse problem is considered ill-posed if it does not possess a unique, well-defined solution. This implies that the operator $\mathcal{A}(\cdot)$ is not invertible. In such cases, there may be multiple solutions $\mathbf{x}_0$ corresponding to the same degraded measurement $\mathbf{y}$. This makes it difficult, if not impossible, to reliably recover the original signal $\mathbf{x}_0$ from the degraded measurements $\mathbf{y}$. To address this, the problem is often reframed from a probabilistic perspective. Instead of estimating a single solution $\mathbf{x}_0$, the goal shifts to approximating the posterior distribution $p(\mathbf{x}_0 \mid \mathbf{y})$, which represents the

range of plausible solutions.

In some cases, the operator $\mathcal{A}(\cdot)$ is not known. In these instances, the problem is referred to as blind, meaning that the degradation process is unknown and must be inferred from the data. A blind inverse problem presents additional challenges, as both the degradation and the original signal must be estimated simultaneously. This contrasts with non-blind problems, where the operator $\mathcal{A}(\cdot)$ is assumed to be known and fixed, making it easier to recover the original signal. Some methods, such as certain speech enhancement techniques [7, 15], focus on recovering the original signal directly without explicitly estimating the degradation operator. In contrast, some of the contributions of this thesis (Publications III, IV, V, and VI) aim to jointly optimize the degradation operator and the recovery of the original signal. This approach is particularly interesting because it allows for a clearer understanding of how the degradation process affects the estimated original signal.

2.2 Audio Inpainting

Audio inpainting, sometimes referred to as audio interpolation [17–19], extrapolation [20] or concealment [21], involves restoring or completing missing or degraded parts of an audio signal [22]. It can be used to remove noise, glitches, or other artifacts and to reconstruct lost segments. Applications include restoring old recordings affected by disturbances [23], recovering audio lost due to CD scratches [17], and compensating for packet loss in communication networks [24]. Additionally, it can be employed in music and audio production for creating effects or manipulating signals [25].

In such cases, the operator $\mathcal{A}()$ is often modeled as a compact binary mask $\mathbf{m}$, defining the forward map as

$$\mathbf{y} = \mathbf{m} \odot \mathbf{x}_0, \tag{2.2}$$

where $\odot$ represents the Hadamard product or element-wise multiplication. The mask $\mathbf{m}$ has values of 0 at locations where samples are missing and 1 otherwise. From the measurement signal $\mathbf{y}$, it is possible to determine the binary mask $\mathbf{m}$ by examining audio portions that are absent or have been suppressed to zero. Thus, it is typically reasonable to regard the mask as a known entity. The task of inpainting is a classic example of an ill-posed inverse problem because the mask $\mathbf{m}$ cannot be inverted due to its zero values. Consequently, there is a large set of possible solutions that could explain the observed data $\mathbf{y}$. To address this ambiguity, additional prior information about the signal $\mathbf{x}_0$ is necessary to guide the reconstruction.

This prior information can be introduced in various ways. Some methods implement autoregressive modeling, which assumes that the signal is

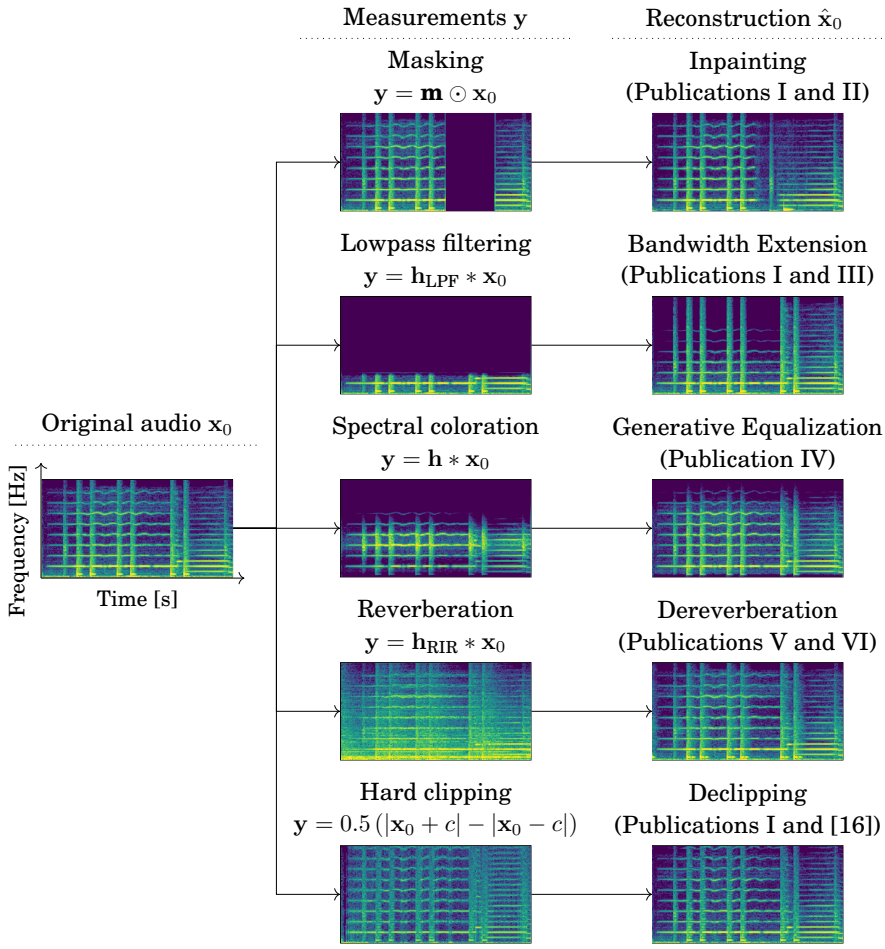


Figure 2.1. Spectrogram representations of the audio restoration inverse problems studied in this thesis. The original audio signal ($\mathbf{x}_0$) is transformed by various forward operators, resulting in degraded measurements ($\mathbf{y}$). Reconstructions ($\hat{\mathbf{x}}_0$) are obtained by solving the respective inverse problems. Figure adapted from [1].

stationary and can be approximated as a linear combination of its previous samples, either in the time domain [17–19], or in the short-time Fourier transform (STFT) [26]. Other successful approaches exploit the sparsity of audio signals when represented in the STFT domain [27–31]. These methods exploit the fact that most of the energy in an audio signal is concentrated in a few significant coefficients, making it possible to reconstruct missing parts efficiently. These methods work well for short gaps (10–100 ms) but struggle with longer gaps where stationarity no longer holds. In such cases, alternative approaches, such as sinusoidal modeling [32–34] or similarity graphs [35], have been explored.

Data-driven approaches have also been increasingly applied to audio inpainting in recent years. For music inpainting, some studies have used deep neural networks with predictive objectives [36], while others have

focused on Generative Adversarial Networks (GANs) to model and generate plausible audio content [37–39]. Additionally, an approach employing untrained deep neural networks, known as deep prior, has also been explored [40]. In the related task of packet loss concealment, which shares similarities with audio inpainting but incorporates real-time constraints, methods have primarily focused on speech signals. Both predictive approaches [21, 41] and GAN-based techniques [42, 43] have been applied in this context.

Publications I and II explore the use of diffusion models for audio inpainting, with Publication II providing a more detailed analysis of the approach. Recently, there has been a growing interest in generative models for audio editing, where inpainting or replacing segments of audio content is a key focus. These methods include conditional latent diffusion models [44–47], inference-time conditioned diffusion models [48, 49], and masked audio modeling [50, 51].

Unlike the techniques introduced in Publications I and II, which concentrate on the audio domain in isolation, many recent approaches use multimodality to constrain the inpainting search space. Examples of multimodal conditioning include symbolic music [52–54], video [55], and text [44, 45, 47, 56]. By incorporating additional modalities, these methods can potentially create more controllable outcomes for audio inpainting.

2.3 Bandwidth Extension

Audio bandwidth extension aims to reconstruct the missing high-frequency content in bandlimited audio signals [57–59]. Applications include audio upsampling or super-resolution [60–62], where higher frequency components are restored to increase the sampling rate, as well as the restoration of historical recordings [63], which often have limited bandwidth due to past technological limitations.

In this problem, the measurements are obtained by applying a low-pass filter to the original signal:

$$\mathbf{y} = \mathbf{h}_{\mathrm{LPF}} * \mathbf{x}_0, \tag{2.3}$$

where $*$ is the discrete convolution operator, and $\mathbf{h}_{\mathrm{LPF}}$ denotes the low-pass filter. When the filter attenuation is high enough, the inverse problem is inherently ill-posed, as the low-pass filter cannot be directly inverted due to numerical limitations or the presence of noise. In the context of audio super-resolution, the measurements are often decimated, with $\mathbf{h}_{\mathrm{LPF}}$ serving as an antialiasing filter.

Early approaches to bandwidth extension relied on techniques such as source-filter models [64, 65], nonlinear devices [57], and spectral band replication [66]. While these methods sometimes produced perceptually ac-

ceptable results, they were primarily heuristic and did not fully account for the statistical properties of audio signals, leading to limited performance. Later efforts explored data-driven approaches, including Gaussian mixture models [67], hidden Markov models [68], and early neural networks [69], though their effectiveness remained limited.

Recently, methods based on deep learning have achieved significant interest in the field, with numerous researchers exploring predictive approaches [60, 70–79]. Some innovations have also integrated principles from source-filter modeling into modern data-driven frameworks to enable efficient processing [80, 81]. Despite these advancements, regression-based methods often introduce oversmoothing artifacts due to their tendency to predict average solutions. To address this issue, generative approaches have been applied to this or related tasks, including GANs [63, 82–90], normalizing flows [61], diffusion models [62, 91–94], latent diffusion models [95, 96], and masked audio modeling [50, 97].

Unlike audio inpainting, where the operator is typically well-defined, audio bandwidth extension often involves uncertainty about the exact shape of the low-pass filter. While the cutoff frequency may be known, either from the sampling rate (in super-resolution) or through analysis of the measurements (in historical recordings), the precise filter characteristics are usually unknown, making the problem effectively blind. Training supervised models under mismatched filter assumptions can result in severe generalization issues [13]. To mitigate this, various data augmentation techniques have been proposed, such as applying a diverse set of filters [74, 86, 98] or adding white Gaussian noise [63]. However, these methods only partially address the generalization challenges.

In Publication III, a joint optimization strategy for both the wideband audio signal $\hat{\mathbf{x}}_0$ and the lowpass filter $\hat{\mathbf{h}}_{\mathrm{LPF}}$ is proposed. This approach utilizes a diffusion model that is exclusively trained on wideband music along with a zero-phase parametric model for the filter. This methodology effectively addresses the filter generalization problem and can be practically employed to restore historical music recordings with unknown degradation.

2.4 Generative Blind Equalization

Publication IV introduces generative equalization, extending the blind bandwidth extension problem studied in Publication III. Scenarios such as historical music restoration, poor-quality microphone recordings, and signals degraded by communication channels often require both bandwidth extension and the correction of spectral coloration. Equalization compensates for spectral coloration modeled as a linear filter, but when specific frequency bands are heavily attenuated, the filter becomes non-invertible.

This makes the problem ill-posed and conceptually similar to bandwidth extension.

While blind equalization has been studied extensively [99, 100], and early works considered combining equalization with bandwidth extension [101], the joint problem of addressing both tasks simultaneously remained unexplored prior to Publication IV. Generative equalization bridges this gap by employing generative models to synthesize plausible content, restoring spectral balance and reconstructing missing frequency bands.

2.5 Dereverberation

Dereverberation refers to the process of mitigating the effects of reverberation in an audio signal. This occurs when sound waves reflect off surfaces within an environment, leading to a blurred or distorted perception of the original sound [102]. Over the past several decades, dereverberation has been extensively studied [102–105] due to its critical role in improving speech intelligibility in applications such as telecommunications and speech recognition. It can also be applied in applications such as virtual and augmented reality, where estimating the anechoic signal allows for simulating the effects of different environments, enhancing user immersion and interaction [106].

The forward map for reverberated audio can be expressed as

$$\mathbf{y} = \mathbf{h}_{\text{RIR}} * \mathbf{x}_0, \tag{2.4}$$

where $\mathbf{h}_{\text{RIR}}$ represents the Room Impulse Response (RIR), a long linear filter that characterizes the reverberation properties of a space, and $\mathbf{x}_0$ is the clean, anechoic audio signal. Within the scope of this thesis, we focus on the settings involving a single microphone, resulting in single-channel RIRs. This approach is more difficult than a multi-channel situation [107].

Informed dereverberation methods assume that the RIR is known, allowing methods based on inverse filtering [108, 109]. Nonetheless, in the single-channel scenario, knowing the RIR does not ensure a stable and causal inverse filter because real-world RIRs are mixed-phase systems [110]. Recent advancements, such as the method introduced by Lemercier *et al.* [111], use diffusion models to improve the performance of informed dereverberation. However, in practical situations, complete knowledge of the RIR is typically unavailable because measuring it is inconvenient, and it can vary depending on factors such as position or temperature. Consequently, dereverberation is often approached as a blind problem.

Model-based methods for blind dereverberation typically rely on specific distributional or structural assumptions about the anechoic and reverberant signals [112–117]. These approaches use priors or physical models to estimate the clean signal. In contrast, data-driven methods rely on

learning signal characteristics directly from data. Predictive methods for single-channel dereverberation have been applied to various signal representations, including time-frequency domains [118–120], raw waveforms [121, 122], and cepstral features [123]. Generative approaches, which are particularly suited to addressing the ill-posed nature of the problem, have also been investigated. Examples include GANs [124–126] and diffusion models [15, 127, 128]. Frequently, dereverberation is performed jointly with denoising as part of speech enhancement benchmarks [7, 15, 124, 129].

The vast majority of data-driven approaches are based on supervised learning, where a set of RIRs is used to generate reverberant data for training. However, these methods often struggle to generalize to diverse acoustic conditions. In contrast, unsupervised methods offer notable advantages, such as improved robustness to unseen acoustic environments without the need for retraining. An early generative model for dereverberation, based on Gaussian mixture models, was introduced by Attias et al. [130]. More recently, other approaches have incorporated anechoic speech or singing voice priors using variational autoencoders [131, 132] or diffusion models [133, 134]. A key feature of these methods is that they do not require room acoustics data during training and can adapt to different acoustic conditions during inference. However, their performance still lags behind that of supervised models. Publications V and VI introduce a diffusion-based unsupervised approach that jointly estimates the RIR and the anechoic signal. This method significantly narrows the performance gap between matched and mismatched acoustic conditions compared to supervised baselines.

A related problem is the blind estimation of acoustic parameters, which involves retrieving acoustic features such as reverberation time, clarity, or the entire RIR from reverberant recordings, often speech [135–139]. The method in Publications V and VI inherently addresses this task by providing a blind RIR estimate. Notably, experiments in Publication VI show that its reverberation time and clarity estimations outperform those of a supervised RIR estimation model [137] under certain conditions.

2.6 Declipping and Nonlinear Restoration

While the previous sections focus on linear degradations, nonlinear degradations are equally significant in audio restoration. One prominent example is clipping, a phenomenon that occurs when the amplitude of an audio signal exceeds a predefined maximum limit, resulting in saturation [140]. This form of distortion is common in scenarios where signal levels are not properly managed, such as during analog-to-digital conversion or in overdriven audio equipment. Hard clipping distortion can be mathematically

expressed as

$$\mathbf{y} = \frac{1}{2}\left(|\mathbf{x}_0 + c| - |\mathbf{x}_0 - c|\right), \tag{2.5}$$

where c represents the clipping threshold. This operation clips signal values exceeding c, leading to a loss of dynamic range. The clipping introduces harmonic distortion, characterized by the generation of additional frequencies that are integer multiples of the original signal frequencies, as well as intermodulation distortion, which produces new frequencies from the interaction of existing ones.

A significant portion of declipping approaches in the literature are model-based and unsupervised. These methods often rely on sparsity-based regularization techniques [141–143], matching pursuits [144], or nonnegative matrix factorization [145], among other strategies. In recent years, data-driven methods utilizing deep neural networks have also been explored for declipping [146–149]. Additionally, hybrid approaches that combine data-driven models with sparsity-based techniques have been proposed [150, 151]. The problem of hard clipping was experimented on in Publication I, using an unsupervised diffusion-based approach.

In addition to hard clipping, soft clipping and other memoryless nonlinearities also play a significant role in audio restoration. Unlike hard clipping, soft clipping [152] introduces gradual compression beyond a threshold, resulting in smoother distortion but complicating the estimation of the underlying operator. Other examples of memoryless nonlinearities include wave-shaping functions, such as those produced by wave rectifiers [153] and wavefolders [154], which introduce nonlinear transformations of the waveform, creating new harmonic content. Unlike hard clipping, where the distortion operator is straightforward to characterize, the nonlinearity in other memoryless distortions is often more complex, making it challenging to estimate the underlying operator from the degraded signal alone. The problem of blind estimation of memoryless nonlinear functions and the restoration of distorted signals has been explored in [16].

2.7 Restoration of Historical Recordings

A significant focus of this thesis was on the restoration and enhancement of historical music recordings, a process that addresses a wide array of challenges inherent to aging audio material. These recordings often suffer from various types of degradations that compromise their quality and intelligibility.

One of the primary concerns is the mitigation of additive noise, such as background hiss, rumble, localized clicks, and low-frequency pulses [2]. Supervised deep learning approaches, trained on datasets with realistic noise profiles, have demonstrated remarkable effectiveness in isolating

music from such disturbances [8, 155, 156]. Yet, denoising represents only a fraction of the broader restoration process. This thesis extends beyond noise reduction to tackle pressing challenges such as bandlimiting effects, the recovery of missing audio segments, and other nonlinear degradations, which are common in historical recordings due to the limitations of early recording technologies and the physical degradation of the media over time.

Bandwidth extension of historical music recordings using generative models was first explored by the author in [63]. Although the approach, which relied on simulated lowpass degradations, demonstrated generalization capabilities, its performance was limited. Speech enhancement systems have also shown promise for historical speech recordings [7, 129], though they similarly depend on simulated degradations. Saeki et al. proposed a self-supervised framework that leverages actual historical data for training, reducing the train/test mismatch [157]. While promising, this method suffers from instabilities and requires significant effort and additional techniques to function effectively. In contrast, the methods proposed in Publications III and IV achieve unprecedented performance in enhancing historical music recordings, utilizing the generative capabilities of diffusion models. A key factor in their success is the unsupervised approach, which enables the model to seamlessly generalize to unseen degradations.

While distinct from the restoration methodologies discussed earlier, the following works are also relevant as they address complementary aspects of the preservation process. For example, Bosi et al. [158] developed methods for detecting speed variations and surface irregularities on tape recordings. Similarly, Ragano et al. [23, 159] employed machine learning techniques to assess and rank audio quality, providing valuable tools for identifying recordings in need of restoration.

3. Audio restoration with Diffusion Models

This chapter introduces diffusion models and explores their application to solving inverse problems, with a particular emphasis on audio. The focus of this chapter is application-agnostic, providing a broad overview of how diffusion models can be utilized across various audio restoration tasks, highlighting their potential and versatility in addressing different types of inverse problems.

3.1 Fundamentals of Diffusion Models

Generative models aim to estimate the underlying probability distribution of observed data, denoted as p_{data}. Diffusion models [160–162] are a specific class of generative models that learn a mapping between a tractable prior distribution, commonly a Gaussian distribution $p_T = \mathcal{N}(\mathbf{0}, \sigma^2(T)\mathbf{I})$, and the empirical data distribution $p_0 = p_{\text{data}}$, which is accessible only through a dataset of realizations $\mathbf{x}_0 \sim p_{\text{data}}$. This mapping is learned progressively over a time interval $\tau \in [0, T]$ that controls the noise variance schedule $\sigma^2(\tau)$. Usually, $\tau = 0$ corresponds to clean data, and $\tau = T$ represents pure noise. The intermediate states between these endpoints facilitate the transformation of data samples into the tractable prior and vice versa.

The earliest contributions to the diffusion model literature used a discretized time parameter $\tau \in \mathbb{Z}$ [161]. In this setting, the transformation between p_T and p_{data} was depicted as a Markov chain $\mathbf{x}_T \to \mathbf{x}_{T-1} \to \cdots \to \mathbf{x}_0$, where the state transition probabilities $p(\mathbf{x}_{i-1} \mid \mathbf{x}_i)$ were parameterized by a trained deep neural network (DNN). Subsequent works proposed defining time as a continuous variable $\tau \in \mathbb{R}$ [162], allowing for greater flexibility while remaining practically equivalent, or as a generalization of the discrete diffusion framework. In the following, we adopt the continuous-time parameterization, particularly following the EDM design proposed by Karras et al. [163].

Under a continuous time parameterization, the transformation between the prior and data distributions can be formalized as a deterministic

mapping through the Probability Flow Ordinary Differential Equation (ODE):

$$\mathrm{d}\mathbf{x} = \Big[\mathbf{f}(\mathbf{x}_\tau, \tau) + \frac{1}{2} g^2(\tau) \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau)\Big] \mathrm{d}\tau, \tag{3.1}$$

where time evolves in reverse, starting from $\tau = T$ (representing pure noise) and ending at $\tau = 0$ (representing clean data). The term $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$, referred to as the score function, indicates the direction of increasing data likelihood within the data space. The drift component $\mathbf{f}(\mathbf{x}_\tau, \tau)$ and the diffusion coefficient $g(\tau)$ are parameters specific to the model. In the works included in this thesis, the parameterization proposed by Karras et al. [163] was adopted, which specifies

$$\mathbf{f}(\mathbf{x}_\tau, \tau) = 0, \quad g(\tau) = \sqrt{2\tau}, \quad \text{and} \quad \sigma(\tau) = \tau. \tag{3.2}$$

Using these parameter choices, the Probability Flow ODE simplifies to

$$\mathrm{d}\mathbf{x} = -\sigma(\tau) \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)\, \mathrm{d}\tau. \tag{3.3}$$

The specified ODE defines a bijective mapping between data and noise distributions. It is worth noting that the reverse process can also be described through other approaches, such as Markov chains [161] or stochastic differential equations [162]. Additionally, frameworks like rectified flows [164–166] provide more streamlined and potentially more generalizable formulations. However, these alternatives are not explored here, as the presented formulation aligns more closely with the methodologies and objectives of this thesis.

The score function $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$ plays a critical role in enabling this mapping. However, this term is intractable in closed form. In the case of Gaussian noise, following Tweedie's formula [167], the score is directly related to the posterior expectation $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_\tau]$, which represents the expected clean data given noisy observations $\mathbf{x}_\tau$ at time τ:

$$\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau) = \frac{\mathbb{E}[\mathbf{x}_0|\mathbf{x}_\tau] - \mathbf{x}_\tau}{\sigma^2(\tau)}. \tag{3.4}$$

This posterior expectation can be approximated using a Gaussian denoiser, typically learned through a neural network $D_\theta(\mathbf{x}_\tau, \tau) = \hat{\mathbf{x}}_0(\mathbf{x}_\tau)$, which is conditioned on the time or diffusion state variable τ. The weights of the neural network, θ, can be optimized by regressing with an L_2 loss, derived from Denoising Score Matching [168]:

$$\mathbb{E}_{\mathbf{x}_0 \sim p_{\text{data}}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\lambda(\tau) \| D_\theta(\mathbf{x}_0 + \tau\boldsymbol{\epsilon}, \tau) - \mathbf{x}_0 \|_2^2 \right], \tag{3.5}$$

where $\lambda(\tau)$ is a time-varying weighting function.

Once the denoising model $D_\theta(\mathbf{x}_\tau, \tau)$ converges, the score can be approximated by substituting it into Eq. (3.4):

$$\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau) \approx \mathbf{s}_\theta(\mathbf{x}_\tau, \tau) = \frac{D_\theta(\mathbf{x}_\tau, \tau) - \mathbf{x}_\tau}{\sigma^2(\tau)}. \tag{3.6}$$

This approximation can then be used to replace the score in Eq. (3.3). By initializing with a sample from the tractable prior $\mathbf{x}_T \sim p_T$, it becomes possible to sample from the approximated data distribution by solving the ODE in Eq. (3.3) from $\tau = T$ to $\tau = 0$.

At inference time, sampling requires discretizing the time variable τ. Discretization strategies can be arbitrarily designed, for instance to prioritize perceptual effects or computational efficiency. The resulting ODE can then be solved using any numerical integration method. A common choice is first-order solvers, such as DDIM [169] or the Euler method [162]. Higher-order solvers, including Runge-Kutta methods [170], allow the use of fewer discretization steps, albeit at the cost of more expensive computations per step. More recently, several numerical solvers have been explicitly developed for efficient diffusion model inference [171, 172].

Most of the contributions in this thesis adopt Heun's second-order method, as proposed by Karras et al. [163], which offers an effective trade-off between efficiency and accuracy. Alternatively, if a stochastic differential equation (SDE) formulation is employed [162], stochastic solvers such as Euler-Maruyama [173] or the predictor-corrector schemes introduced by Song et al. [162] can be used. However, stochastic solvers are generally less efficient than deterministic ones [162]. Nevertheless, stochastic solvers can provide certain advantages; in particular, the injected randomness can sometimes mask discretization errors and thereby improve generation quality [163]. Motivated by this observation, Karras et al. [163] proposed a stochastic variant of Heun's second-order method with controlled noise injection, which has been employed in several contributions of this thesis.

3.2 Diffusion Models in the Audio Domain

Diffusion model frameworks are inherently domain-agnostic. While early works popularized their use for image generation [161, 174, 175], diffusion models (and their variants) have been successfully applied across a broad range of domains, including video [176], molecules [177], text [178, 179], symbolic music [180], and communications [181], among others. Of particular relevance to this thesis, diffusion models have shown great promise in the domains of audio and speech [1, 182, 183].

Designing effective approaches to adapt diffusion models for the audio domain represents a significant aspect of this thesis work. For example, audio waveforms sampled at high rates (typically 44.1 kHz or 48 kHz) exhibit very high dimensionality, presenting a challenge when training diffusion models on such data. This challenge involves not only selecting suitable feature representations for audio but also employing neural network architectures capable of capturing the complexities inherent in audio signals. Additionally, these methods must be designed to function

within the constraints of limited compute resources and data availability. This section reviews recent advancements in the field, highlighting key methodologies, including choices of audio representations and architectural designs.

Early applications of diffusion models for audio generation operated in the raw waveform domain, primarily using one-dimensional convolutional neural networks [182, 183]. These initial efforts were largely dedicated to vocoder tasks, synthesizing waveforms from mel-spectrogram features [182, 183]. However, the unconditional generation capabilities of these models were limited—except in specific works addressing non-tonal audio [11, 184]—due to the high dimensionality of audio data, which made it difficult for a neural network lacking the right inductive biases to learn the underlying structure.

Subsequent approaches adopted a two-stage process in which diffusion models were employed to generate magnitude-only mel-spectrograms, followed by a separate vocoder model to synthesize waveforms from the generated spectrograms [185, 186]. This shift was motivated by the relative simplicity of modeling mel-spectrograms compared to waveforms. Consequently, this methodology became prevalent in conditional generation tasks, such as text-to-speech [185] and MIDI-to-audio synthesis [186]. However, mel-spectrogram-based methods require an additional vocoder, adding complexity and potentially introducing artifacts.

Similarly, latent diffusion models were introduced [187], which operate in the latent space produced by a pre-trained variational autoencoder that compresses audio (or mel-spectrograms) into a much lower-dimensional latent vector. These models have gained popularity and have become the standard in text-to-audio generation [46, 188–195], an emerging task focused on generating audio from descriptive text prompts. These models often utilize diffusion transformers [196] to benefit from large-scale training.

However, adapting these representations to an inverse problem framework, such as the one outlined in Sec. 3.3.1, presents challenges due to the need for flexibility and accurate reconstruction. Typically, measurements are available in the waveform domain, not in the latent space, making it difficult to apply these models effectively. Additionally, artifacts introduced during the decoding stage can significantly degrade quality. A model trained on large-scale music data may also be unsuitable for audio restoration tasks, where clean, high-quality training data is essential.

An alternative approach employs invertible time-frequency representations, specifically the STFT, and applies diffusion in the complex time-frequency domain, adapting diffusion models to handle complex algebra [15, 197]. One implementation option is to design a neural network capable of processing complex operations [197], or alternatively, treat the real and imaginary components as separate channels within a real-valued

neural network [15]. While transforms like the STFT do not reduce dimensionality, time-frequency representations exhibit clear structure, making them particularly well-suited for two-dimensional convolutional neural networks [4]. A notable advantage of using an invertible transform is that it allows the output to be directly converted back into a waveform without the need for an additional model, simplifying the process.

Publication I proposes a slightly different approach. Instead of defining the diffusion process in the time-frequency domain, it is maintained in the waveform domain. However, an invertible time-frequency transform $\mathcal{F}$ and its inverse $\mathcal{F}^{-1}$ are applied at the input and output of the denoising neural network, as shown in the following equation:

$$D_\theta(\mathbf{x}_\tau, \tau) = \mathcal{F}^{-1}\left(D'_\theta\left(\mathcal{F}(\mathbf{x}_\tau), \tau\right)\right), \tag{3.7}$$

where D'_θ represents the neural network layers with trainable weights. This approach is advantageous because it allows the model to operate directly in the waveform domain, making it simpler and more flexible, while still benefiting from the inductive biases provided by the time-frequency transform.

This approach is used throughout all the contributions of this thesis. While, following [15], we use the STFT representation in the experiments involving speech in Publications V and VI, Publications I, II, III and IV use the Constant-Q Transform (CQT) for experiments involving music. The CQT is a time-frequency transform in which the frequency axis is logarithmically warped. In this domain, pitch-shifting corresponds to translation, making it particularly well-suited for convolutional neural networks.

3.3 Conditional Diffusion Models for Inverse Problems

As introduced in Sec. 2.1, a common approach to solving ill-posed inverse problems is to approximate the posterior distribution $p(\mathbf{x}_0 \mid \mathbf{y})$ and use samples drawn from it as potential solutions. However, as explained in Sec. 3.1, diffusion models, by default, approximate unconditional densities. Therefore, these models need to be adapted for the conditional case.

One popular approach is to adapt the unconditional model by incorporating input conditioning. This involves adding the measurement $\mathbf{y}$ as an additional input to the neural network $D_\theta(\mathbf{x}_\tau, \tau, \mathbf{y})$ during both training and inference [182, 198–200]. This conditioning enables the network to estimate the posterior score $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau \mid \mathbf{y})$, analogous to applying Eq. (3.4).

A particularly effective technique for conditioning is Classifier-Free Guidance (CFG) [201], which introduces a scalar hyperparameter ω at inference time to modulate the influence of the conditioning signal $\mathbf{y}$. Specifically,

CFG defines a sharpened posterior distribution [202] as

$$p^{\omega}(\mathbf{x}_\tau \mid \mathbf{y}) \propto p(\mathbf{x}_\tau) p(\mathbf{y} \mid \mathbf{x}_\tau)^{\omega}. \tag{3.8}$$

The score of the modified posterior can be expressed as a weighted combination of the unconditional and conditional score functions:

$$\nabla_{\mathbf{x}_\tau} \log p_\tau^{\omega}(\mathbf{x}_\tau \mid \mathbf{y}) = \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau) + \omega \left(\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau \mid \mathbf{y}) - \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau) \right). \tag{3.9}$$

In this formulation, setting $\omega = 0$ corresponds to unconditional sampling, $\omega = 1$ recovers standard conditional sampling, and $\omega > 1$ amplifies the influence of the conditioning information. To enable this mechanism, the network is trained jointly with conditional and unconditional objectives, typically by randomly omitting the conditioning signal $\mathbf{y}$ with a certain probability during training.

Other approaches have proposed task-adapted diffusion processes [15, 203], such as defining processes that interpolate between clean audio $\mathbf{x}_0$ at $\tau = 0$ and the measurements $\mathbf{y}$ at $\tau = T$, or designing processes that incorporate information beyond simple Gaussian noise at $\tau = T$. A prominent example of the latter is Schrödinger bridges [204–206], which seek optimal probability paths connecting two data distributions.

However, the mentioned strategies typically require retraining for each new task, which can be inconvenient in practice. Moreover, they rely on a supervised setting, necessitating paired clean and degraded data during training. This dependency often leads to generalization issues, as models may underperform when encountering unseen measurement conditions. While recent methods allow training Schrödinger bridges from unpaired data [207], they involve repeated sampling and retraining using generated samples, a procedure that tends to accumulate errors over iterations. Other approaches implicitly construct bridges between two data distributions by connecting them through the Gaussian prior [128, 208, 209], thereby enabling unpaired translation tasks. However, these methods often offer limited control over the resulting mappings.

This thesis adopts an alternative strategy that combines an unconditional diffusion model with an external conditioning mechanism. The conditioning signal is provided during inference based on a Bayesian posterior sampling framework.

3.3.1 Conditional Generation through Posterior Sampling

The probability flow ODE from Sec. 3.3 can be modified for posterior sampling by replacing the score term $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau)$ with the posterior score $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau|\mathbf{y})$ [162]. Using Bayes' rule, the posterior score can be expressed as the sum of two components:

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau|\mathbf{y}) = \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau) + \nabla_{\mathbf{x}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau). \tag{3.10}$$

The first term, $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau)$, represents the prior score, which can be approximated using Eq. (3.6). The second term, $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau)$, corresponds to the likelihood score or measurement-matching term. However, the likelihood $p(\mathbf{y}|\mathbf{x}_\tau)$ is generally intractable for $\tau > 0$. Since $\mathbf{x}_\tau$ is a noisy version of $\mathbf{x}_0$, evaluating the likelihood involves integrating over all possible values of $\mathbf{x}_0$:

$$p(\mathbf{y}|\mathbf{x}_\tau) = \int_{\mathbf{x}_0} p(\mathbf{y}|\mathbf{x}_0) p(\mathbf{x}_0|\mathbf{x}_\tau)\, d\mathbf{x}_0. \tag{3.11}$$

As a result, the likelihood must be approximated to enable practical computation.

Overcoming this intractability has been an active area of research in recent years. For a comprehensive review, see [210]. This explanation focuses on one particular approach, commonly known as Diffusion Posterior Sampling (DPS) [176, 211], which has been employed in all the publications included in this thesis. The approximation assumes $p(\mathbf{y}|\mathbf{x}_\tau) \approx p(\mathbf{y}|\hat{\mathbf{x}}_0(\mathbf{x}_\tau))$, where $\hat{\mathbf{x}}_0(\mathbf{x}_\tau) = D_\theta(\mathbf{x}_\tau, \tau) \approx \mathbb{E}[\mathbf{x}_0|\mathbf{x}_\tau]$. By replacing the noisy variable $\mathbf{x}_\tau$ with the denoised estimate $\hat{\mathbf{x}}_0(\mathbf{x}_\tau)$ at the given time, this method avoids the intractable integral. The denoised estimate $\hat{\mathbf{x}}_0(\mathbf{x}_\tau)$ is directly accessible through the trained diffusion model, making this approximation both practical and effective.

When the likelihood is modeled as a Gaussian distribution

$$p(\mathbf{y}|\mathbf{x}_0) \approx \mathcal{N}(\mathbf{y}; \mathcal{A}(\hat{\mathbf{x}}_0(\mathbf{x}_\tau)), \sigma_{\mathbf{y}}^2 \mathbf{I}), \tag{3.12}$$

where $\sigma_{\mathbf{y}}^2$ represents the measurement noise power, the corresponding likelihood score becomes

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau) \approx -\frac{1}{\sigma_{\mathbf{y}}^2} \nabla_{\mathbf{x}_\tau} \|\mathbf{y} - \mathcal{A}(\hat{\mathbf{x}}_0(\mathbf{x}_\tau))\|_2^2. \tag{3.13}$$

This formulation links the likelihood score to the gradient of an L_2-norm, which measures the discrepancy between the observations $\mathbf{y}$ and their predicted values $\mathcal{A}(\hat{\mathbf{x}}_0(\mathbf{x}_\tau))$. It is important to note that the gradient operator $\nabla_{\mathbf{x}_\tau}$ requires differentiation through both the operator $\mathcal{A}(\cdot)$ and the denoiser $D_\theta(\mathbf{x}_\tau, \tau)$, the latter of which may introduce a significant computational burden.

In practice, the theoretical scaling factor $1/\sigma_{\mathbf{y}}^2$ is often replaced by a heuristic hyperparameter, as it becomes less relevant or even nonsensical in noiseless measurement scenarios. Additionally, replacing the L_2 cost function with alternative cost functions has also been shown to improve performance in certain cases, such as in Publications V and VI. This leads to a generalized likelihood score approximation:

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau) \approx -\zeta(\tau) \nabla_{\mathbf{x}_\tau} C(\mathbf{y}, \mathcal{A}(\hat{\mathbf{x}}_0(\mathbf{x}_\tau))), \tag{3.14}$$

where $C(\cdot, \cdot)$ is a flexible cost function that penalizes dissimilarity between the measurements $\mathbf{y}$ and their estimates, and $\zeta(\tau)$ is a time-dependent

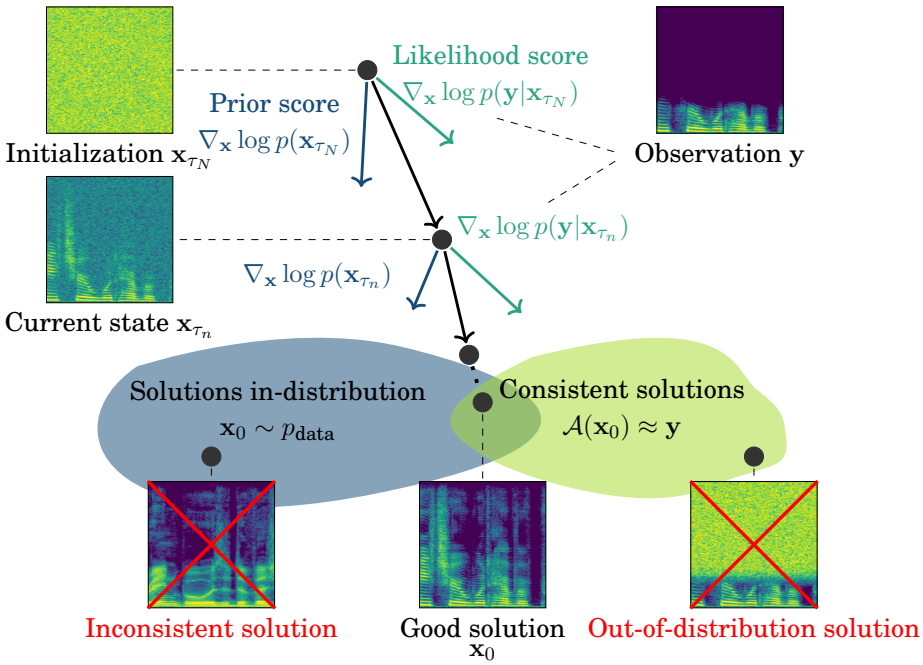


Figure 3.1. Geometric interpretation of posterior sampling in diffusion models. The prior score directs trajectories toward the training data manifold (gray space), while the likelihood score guides them toward the observed data space (light green space). Proper weighting ensures convergence to their intersection, which contains valid solutions to the inverse problem. Adapted from [1].

weighting parameter that balances the contribution of the likelihood term over time.

The interplay between the prior score and the likelihood score is visually illustrated in Fig. 3.1. This figure provides a geometric perspective on posterior sampling with diffusion models. The prior score ensures that sampling remains within the manifold of training data, maintaining consistency with its learned distribution. Meanwhile, the likelihood score adjusts the trajectory toward regions aligned with the observed measurements. The balance of these two components, influenced by the tuning of $\zeta(\tau)$, guides the process to the intersection of the training data manifold and the measurement-consistent space. This intersection represents the solution space to the inverse problem, assuming both score functions are accurately estimated and the solutions lie within the span of the training data manifold.

IEEE/ACM Transactions on Audio, Speech, and Language Processing,

3.3.2 Blind Inverse Problems

Our analysis has thus far assumed that the degradation operator $A(\cdot)$ is known. However, in many scenarios, this operator is often unavailable, making the computation of the posterior $p(\mathbf{x}_0|\mathbf{y})$ a blind inverse problem,

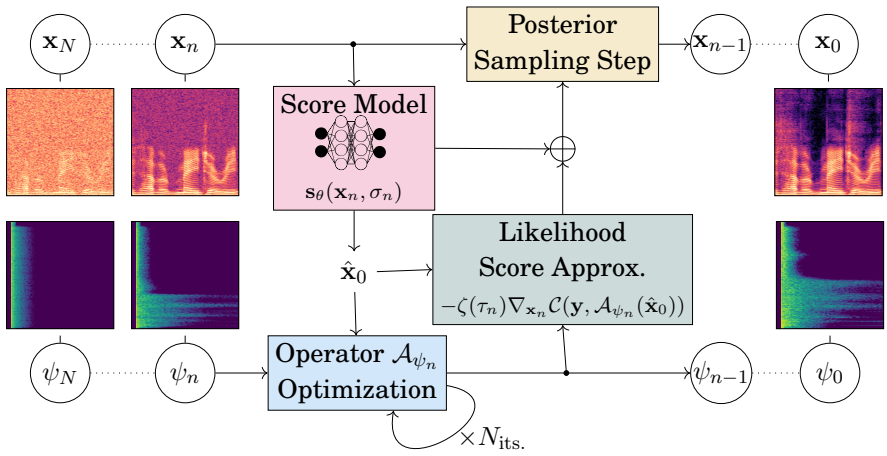


Figure 3.2. Block diagram of the posterior sampling algorithm designed for solving blind inverse problems, with operator optimization integrated into each sampling step. The diagram illustrates an application to blind dereverberation, showcasing the simultaneous generation of an anechoic speech signal and the estimation of the room impulse response. Adapted from Publication V.

where the goal is to estimate both the clean signal and the unknown degradation.

Several studies have proposed methods to tackle blind inverse problems, particularly in the context of image restoration [212–215]. These approaches typically involve parameterizing the degradation operator as $\mathcal{A}_\psi$ with a set of parameters ψ , and estimating the joint posterior $p(\mathbf{x}_0, \psi|\mathbf{y})$ to simultaneously recover the clean image and the degradation operator. This is achieved using separate priors for the clean image and the degradation [134, 212] or a joint prior facilitated by text prompts [214].

Publications III, IV, V, and VI in this thesis delve into this blind restoration scenario across various problems. In these works, the problem is formulated as a joint optimization task with the objective

$$\hat{\mathbf{x}}_0, \hat{\psi} = \arg\min_{\psi, \hat{\mathbf{x}}_0} C(\mathbf{y}, \mathcal{A}_\psi(\hat{\mathbf{x}}_0)), \quad \text{subject to} \quad \hat{\mathbf{x}}_0 \sim p_{\text{data}}, \tag{3.15}$$

where $C(\cdot, \cdot)$ is a cost function that penalizes the dissimilarity between the measurements $\mathbf{y}$ and the predicted $\mathcal{A}_\psi(\hat{\mathbf{x}}_0)$. The constraint $\hat{\mathbf{x}}_0 \sim p_{\text{data}}$ is approximately enforced by using conditional sampling with a diffusion model trained on samples from p_{data}. The parameterization of $\mathcal{A}_\psi$ incorporates prior knowledge into the functional space of the degradation. For instance, in Publications III and IV, a zero-phase filter with a piecewise linear frequency response was employed, whereas in Publications V and VI, a subband filter operator for modeling room impulse responses was used.

To solve Eq. (3.15), this thesis introduces an alternating optimization approach. This method alternates between posterior sampling steps and

parameter optimization iterations. The inference process is illustrated in Fig. 3.2, exemplified for the task of dereverberation. The algorithm builds upon the DPS framework detailed in Sec. 3.3.1, iteratively refining the estimated signal $\hat{\mathbf{x}}_0$ through posterior sampling (yellow box) using a pre-trained score model $s_\theta(\mathbf{x}_\tau, \tau)$ (red box) and a likelihood score approximation (green box). Additionally, at each diffusion step, the operator parameters are refined utilizing gradient-based optimization methods (blue box).

4. Summary of Main Results

This section presents the main results of the featured publications that are related to the author's work.

Publication I - "Solving Audio Inverse Problems with a Diffusion Model"

Publication I introduces CQT-Diff, a diffusion-based generative model designed as a general framework for solving audio inverse problems in a problem-agnostic manner. The model integrates a CQT to incorporate inductive biases compatible with the structure of musical audio. The diffusion model is trained unconditionally and is capable of addressing multiple audio restoration tasks at inference time without the need for retraining, provided the degradation model is known. The paper evaluates the model on three tasks: audio bandwidth extension, inpainting, and declipping. CQT-Diff achieves superior performance in bandwidth extension and competitive results in inpainting and declipping, surpassing traditional baselines in both objective and subjective metrics. Beyond these tasks, the framework is flexible enough to extend to a wider range of linear and nonlinear ill-posed problems, highlighting its potential as the first unsupervised diffusion-based approach for addressing general audio restoration challenges.

Sound examples available at:
research.spa.aalto.fi/publications/papers/icassp23-cqt-diff/

Publication II - "Diffusion-Based Audio Inpainting"

Publication II builds upon the foundational framework introduced in Publication I, focusing specifically on the task of audio inpainting. It presents CQT-Diff+, an improved and more efficient architecture tailored for general diffusion-based audio generation, designed to enhance both performance and scalability. The study evaluates CQT-Diff+ on musical recordings

containing gaps of varying durations, ranging from 25 to 300 ms. Using objective metrics such as log-spectral distance (LSD), Objective Difference Grades (ODG) [216], and Fréchet Audio Distance (FAD) [217], the paper demonstrates significant improvements over traditional baselines that rely on linear prediction or sparsity-based methods. Additionally, subjective listening tests confirm the perceptual quality of the inpainted audio, highlighting the model's ability to produce natural and high-quality reconstructions. These results demonstrate that CQT-Diff+ offers significant improvements in restoration quality and expressivity, establishing it as a promising approach for audio inpainting.

Sound examples available at:
research.spa.aalto.fi/publications/papers/jaes-diffusion-inpainting/

Publication III - "Blind Audio Bandwidth Extension: A Diffusion-Based Zero-Shot Approach"

Publication III focuses on the task of blind audio bandwidth extension, where the lowpass filter characteristics are unknown. The proposed method, BABE (Blind Audio Bandwidth Extension), jointly optimizes the lowpass filter and estimates the wideband audio signal at inference time. It uses a frequency-domain parametric model for a zero-phase lowpass filter with few parameters and constraints. Since the model has not seen filters during training, it operates in a zero-shot setting for previously unseen degradations. The method is evaluated on piano music using both objective and subjective metrics, showing that BABE outperforms state-of-the-art blind bandwidth extension methods and competes well with informed approaches on synthetic data. Additionally, BABE is applied to historical music recordings, enhancing perceived audio quality by restoring lost high-frequency details. When combined with a denoising algorithm [8], BABE demonstrates strong generalization capabilities, effectively reconstructing missing high-frequency content while maintaining the integrity of the original recording, even with out-of-distribution content or artifacts from denoising. The paper reports results on piano, string, woodwind, and brass recordings, highlighting BABE's versatility and its potential as a practical solution for real-world digital audio restoration.

Sound examples available at:
research.spa.aalto.fi/publications/papers/ieee-taslp-babe/

Publication IV - "A Diffusion-Based Generative Equalizer for Music Restoration"

Publication IV introduces BABE-2, an improved version of BABE (Publication III) that generalizes the task of audio bandwidth extension to

generative equalization, with a focus on historical music restoration. Unlike the original method, which is limited to lowpass degradation, BABE-2 can handle a broader range of spectral distortions, offering a more flexible degradation model. Key improvements include a regularization term for filter parameters, noise regularization to enhance generalization, and an initialization strategy based on the long-term average spectrum, which is integrated into the optimization objective. These enhancements contribute to improved optimization stability and performance. BABE-2 is evaluated on historical piano recordings using FAD [217], where it outperforms existing baselines and methods. The paper also explores its effectiveness with singing voice recordings from Enrico Caruso and Nellie Melba, demonstrating its versatility across different audio types. The restoration of historical singing voices is discussed in depth, emphasizing the importance of training data selection and fine-tuning with the VocalSet dataset. Qualitative analysis highlights BABE-2's strengths in reconstructing timbre and handling vocal techniques. In conclusion, BABE-2 significantly advances historical music restoration compared to its predecessor, particularly for singing voices, offering valuable insights for the field of music restoration.

Sound examples available at:
research.spa.aalto.fi/publications/papers/dafx-babe2/

Publication V - "BUDDy: Single-Channel Blind Unsupervised Dereverberation with Diffusion Models"

Publication V addresses the challenge of dereverberation and introduces BUDDy, a novel diffusion-based framework for blind joint dereverberation and RIR estimation. Building on the unsupervised methodology of Publications III and IV, BUDDy models reverberation using an exponentially decaying filter. During the reverse diffusion process, it simultaneously estimates the RIR and refines the dereverberated speech. The performance of BUDDy is evaluated using established speech enhancement metrics, including DNS-MOS [218], PESQ [219], and ESTOI [220]. Results demonstrate that BUDDy significantly outperforms previous unsupervised approaches, including those leveraging diffusion models. While supervised methods perform better under matched training and testing conditions, BUDDy achieves comparable or superior results in mismatched scenarios, highlighting its robustness and generalization capabilities.

Sound examples available at:
www.inf.uni-hamburg.de/en/inst/ab/sp/publications/iwaenc2024-buddy.html

Publication VI - "Unsupervised Blind Joint Dereverberation and Room Acoustics Estimation with Diffusion Models"

Publication VI builds on the work presented in Publication V, further exploring the challenge of blind dereverberation. The work explores the limitations of informed dereverberation methods in partially blind scenarios, revealing their sensitivity to inaccuracies in RIR estimation caused by noise or suboptimal estimators. Beyond instrumental metrics, BUDDy's performance in speech dereverberation is further evaluated through subjective listening tests and ablation studies, providing a comprehensive assessment of its capabilities. Additionally, the study extends BUDDy's scope to singing voice dereverberation at a 44.1 kHz sampling rate, demonstrating its ability to outperform unsupervised baselines and achieve results comparable to supervised methods. Finally, BUDDy's effectiveness in RIR estimation is evaluated using frequency-dependent acoustic descriptors, highlighting its accuracy in estimating reverberation time and clarity relative to state-of-the-art supervised techniques.

5. Conclusions

This thesis has introduced a new paradigm for machine learning-based audio restoration, transitioning from the prevalent supervised approaches to an unsupervised framework that does not require paired clean and degraded training data. Instead, restoration is guided by diffusion-based generative models, which act as learned data priors. Assumptions about the degradation model are only imposed at inference time, while the training process remains fully independent of any specific degradation.

Publication I demonstrates the versatility and potential of this approach to tackle a wide range of audio restoration challenges within a unified framework. Before its release, diffusion-based zero-shot conditioning in the audio domain was largely restricted to specific applications such as text-to-speech synthesis [221] or concurrent explorations for super-resolution [93]. Since then, the publication has motivated additional research efforts, resulting in expansions and improvements [49, 222, 223]. Building on these foundations, Publication II introduced significant improvements, such as a refined network architecture, and conducted a thorough study of the inpainting task.

The contributions on historical recording restoration from Publications III and IV provide valuable opportunities for analyzing and understanding the artistry and unique characteristics of early-20th century performers. In an extended study, the BABE-2 method from Publication IV was applied to the restoration of recordings by Finnish soprano Maikki Järnefelt [224]. This work underscored the critical role of training data, showing that fine-tuning the model using a small, carefully selected set of vocal recordings—approximately five minutes in duration—was essential for producing restorations that were both aesthetically appealing and faithful to the original performances, as confirmed through subjective listening evaluations.

The proposed approach to blind inverse problems extends beyond audio restoration and proves valuable for blind system identification. Publications V and VI introduced BUDDy, a method that, apart from its dereverberation capabilities, also serves as an estimator of room acoustic

properties. This dual functionality highlights its potential for broader applications in acoustic analysis and signal processing. Furthermore, the work on nonlinear system estimation [16, 225] demonstrated the ability of diffusion-based models to accurately estimate and reverse nonlinear distortions. Additional research explored the applicability of diffusion models to head-related transfer function estimation [226].

Despite the advancements of the proposed diffusion models in audio restoration, several limitations remain. At present, diffusion-based audio restoration is not suitable for real-time applications or scenarios where non-accelerated hardware is used. Although methods to enforce causality constraints in diffusion models have been proposed [227], the computational demands remain substantial. To address the computational burden, alternative algorithms that avoid gradient computation via automatic differentiation at inference time have been proposed [228]. However, these alternatives typically do not match the same performance or are not easily generalizable to nonlinear operators.

Another potential improvement involves the use of latent diffusion models, which work with compressed representations [95, 229]. These models are computationally more efficient waveform-domain diffusion models and exhibit superior scaling characteristics, enabling them to learn complex distributions without being restricted to particular types of sources (e.g., speech, piano music, singing voice, etc.). However, latent diffusion models face the challenge of adapting the restoration problem to a compressed domain, which can reduce their flexibility. Additionally, the artifacts introduced by lossy encoding and decoding can affect the quality of restoration, setting an upper bound in quality that depends on the performance of the chosen representation. Finally, distillation techniques have shown potential for improving inference efficiency [230, 231]. However, their relevance and applicability to audio restoration tasks remain unclear, and further investigation is required to determine how they could be effectively integrated.

References

[1] J.-M. Lemercier, J. Richter, S. Welker, E. Moliner, V. Välimäki, and T. Gerkmann. Diffusion models for audio restoration. *IEEE Signal Process. Magazine*, November 2024.

[2] S. J. Godsill and P. J. W. Rayner. *Digital Audio Restoration—A Statistical Model Based Approach*. Springer, 1998.

[3] Y. Bengio, I. Goodfellow, and A. Courville. *Deep Learning*, volume 1. MIT press Cambridge, MA, USA, 2017.

[4] D. S. Williamson, Y. Wang, and D. Wang. Complex ratio masking for monaural speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 24(3):483–492, 2016.

[5] S. Braun, H. Gamper, C.K.A. Reddy, and I. Tashev. Towards efficient models for real-time deep noise suppression. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 656–660, 2021.

[6] A. Defossez, G. Synnaeve, and Y. Adi. Real time speech enhancement in the waveform domain. In *Proc. Interspeech*, 2020.

[7] H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang. VoiceFixer: Toward general speech restoration with neural vocoder. In *Proc. Interspeech*, 2022.

[8] E. Moliner and V. Välimäki. A two-stage U-Net for high-fidelity denoising of historical recordings. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 841–845, Singapore, May 2022.

[9] S. Pascual, A. Bonafonte, and J. Serrà. SEGAN: Speech enhancement generative adversarial network. In *Proc. Interspeech*, page 3642. ISCA, 2017.

[10] V. Välimäki, S. González, O. Kimmelma, and J. Parviainen. Digital audio antiquing—Signal processing methods for imitating the sound quality of historical recordings. *J. Audio Eng. Soc.*, 56(3):115–139, 2008.

[11] E. Moliner and V. Välimäki. Realistic gramophone noise synthesis using a diffusion model. In *Proc. Int. Conf. Digital Audio Effects (DAFX)*, Sep. 2022.

[12] P. Gonzalez, T. Alstrøm, and T. May. Assessing the generalization gap of learning-based speech enhancement systems in noisy and reverberant environments. *IEEE/ACM Trans. Audio Speech Lang. Process.*, September 2023.

[13] S. Sulun and M. E. P. Davies. On filter generalization for music bandwidth extension using deep neural networks. *IEEE J. Sel. Topics Signal Process.*, 15(1):132–142, Nov. 2020.

[14] M. Bertero, P. Boccacci, and C. De Mol. *Introduction to Inverse Problems in Imaging*. CRC press, 2021.

[15] 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.*, 31:2351–2364, 2023.

[16] M. Sventô, E. Moliner, L. Juvela, A. Wright, and V. Välimäki. Estimation and restoration of unknown nonlinear distortion using diffusion. *accepted for publication in the J. Audio Engineering Society*, April 2025.

[17] 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.*, 34(2):317–330, Apr. 1986.

[18] W. Etter. Restoration of a discrete-time signal segment by interpolation based on the left-sided and right-sided autoregressive parameters. *IEEE Trans. Signal Processing*, 44(5):1124–1135, May 1996.

[19] 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)*, pages 08–11, London, UK, Sep. 2003.

[20] I. Kauppinen and K. Roth. Audio signal extrapolation—Theory and applications. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 105–110, Hamburg, Germany, Sep. 2002.

[21] 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.*, 24(2):378–387, Dec. 2015.

[22] A. Adler, V. Emiya, M. G Jafari, M. Elad, R. Gribonval, and M. D Plumbley. Audio inpainting. *IEEE Trans. Audio Speech Lang. Process.*, 20(3):922–932, Mar. 2012.

[23] A. Ragano, E. Benetos, and A. Hines. Automatic quality assessment of digitized and restored sound archives. *J. Audio Eng. Soc.*, 70(4):252–270, Apr. 2022.

[24] D. Goodman, G. Lockhart, O. Wasem, and W-C. Wong. Waveform substitution techniques for recovering missing speech segments in packet voice communications. *IEEE Trans. Audio Speech Lang. Process.*, 34(6):1440–1448, Dec. 1986.

[25] 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, Oct. 2020.

[26] O. Mokrỳ, P. Balušík, and P. Rajmic. Janssen 2.0: Audio inpainting in the time-frequency domain. *arXiv preprint arXiv:2409.06392*, 2024.

[27] F. Lieb and H-G. Stark. Audio inpainting: Evaluation of time-frequency representations and structured sparsity approaches. *Signal Process.*, 153:291–299, Dec. 2018.

[28] Ondřej Mokrỳ and Pavel Rajmic. Audio inpainting: Revisited and reweighted. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 28:2906–2918, 2020.

[29] O. Mokrỳ, P. Záviška, P. Rajmic, and V. Veselỳ. Introducing SPAIN (sparse audio inpainter). In *Proc. Euro. Signal Proc. Conf.*, pages 1–5, Sep. 2019.

[30] G. Tauböck, S. Rajbamshi, and P. Balazs. Dictionary learning for sparse audio inpainting. *IEEE J. Selected Topics Signal Process.*, 15(1):104–119, Jan. 2021.

[31] S. Rajbamshi, G. Tauböck, N. Holighaus, and P. Balazs. Audio inpainting via l1-minimization and dictionary learning. In *Proceedings of the 29th European Signal Processing Conference (EUSIPCO)*, pages 2149–2153, Dublin, Ireland, Aug. 2021.

[32] M. Lagrange, S. Marchand, and J-B. Rault. Long interpolation of audio signals using linear prediction in sinusoidal modeling. *J. Audio Eng. Soc.*, 53(10):891–905, Oct. 2005.

[33] E. Sun and P. Depalle. Hybrid audio inpainting approach with structured sparse decomposition and sinusoidal modeling. In *Proc. Int. Conf. Digital Audio Effects (DAFX)*, 2024.

[34] T. Tanaka, K. Yatabe, and Y. Oikawa. PHAIN: Audio inpainting via phase-aware optimization with instantaneous frequency. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 32:4471–4485, 2024.

[35] N. Perraudin, N. Holighaus, P. Majdak, and P. Balazs. Inpainting of long audio segments with similarity graphs. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 26(6):1083–1094, Jun. 2018.

[36] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak. A context encoder for audio inpainting. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 27(12):2362–2372, Dec. 2019.

[37] P. P Ebner and A Eltelt. Audio inpainting with generative adversarial network. *arXiv preprint*, Mar. 2020.

[38] 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.*, 15(1):120–131, Nov. 2020.

[39] G. Greshler, T. Shaham, and T. Michaeli. Catch-a-Waveform: Learning to generate audio from a single short example. *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 34:20916–20928, 2021.

[40] F. Miotello, M. Pezzoli, L. Comanducci, F. Antonacci, and A. Sarti. Deep prior-based audio inpainting using multi-resolution harmonic convolutional neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2023.

[41] J. Lin, Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen. A time-domain convolutional recurrent network for packet loss concealment. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 7148–7152, May 2021.

[42] S. Pascual, J. Serrà, and J. Pons. Adversarial auto-encoding for packet loss concealment. In *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pages 71–75, Oct. 2021.

[43] 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)*, pages 826–831, Paris, France, May 2022.

[44] Y. Wang, Z. Ju, X. Tan, L. He, Z. Wu, J. Bian, and S. Zhao. AUDIT: Audio editing by following instructions with latent diffusion models. *arXiv preprint arXiv:2304.00830*, Apr. 2023.

[45] M. Xu, C. Li, D. Zhang, D. Su, W. Liang, and D. Yu. Prompt-guided precise audio editing with diffusion models. In *Proc. Int. Conf. Machine Learning*, pages 55126–55143. PMLR, 2024.

[46] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *Proc. Int. Conf. Machine Learning*, 2023.

[47] A. Vyas, B. Shi, M. Le, A. Tjandra, Y-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan, et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023.

[48] Z. Novack, J. McAuley, T. Berg-Kirkpatrick, and N. J. Bryan. DITTO: Diffusion inference-time t-optimization for music generation. In *Proc. Int. Conf. Machine Learning*, 2024.

[49] M. Levy, B. Di Giorgi, F. Weers, A. Katharopoulos, and T. Nickson. Controllable music production with diffusion models and guidance gradients. In *NeurIPS*, 2023.

[50] M. Comunità, Z. Zhong, A. Takahashi, S. Yang, M. Zhao, K. Saito, Y. Ikemiya, T. Shibuya, S. Takahashi, and Y. Mitsufuji. SpecMaskGIT: Masked generative modeling of audio spectrograms for efficient audio synthesis and beyond. *arXiv preprint arXiv:2406.17672*, 2024.

[51] L. Lin, G. Xia, Y. Zhang, and J. Jiang. Arrange, inpaint, and refine: Steerable long-term music audio generation and editing via content-based controls. *arXiv preprint arXiv:2402.09508*, 2024.

[52] C. Aironi, S. Cornell, L. Gabrielli, and S. Squartini. A score-aware generative approach for music signals inpainting. In *2023 4th International Symposium on the Internet of Sounds*, pages 1–7, 2023.

[53] K. Liu, W. Gan, and C. Yuan. MAID: A conditional diffusion model for long music audio inpainting. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1–5, Rhodes, Greece, Jun. 2023.

[54] K. W. Cheuk, R. Sawata, T. Uesaka, N. Murata, N. Takahashi, S. Takahashi, D. Herremans, and Y. Mitsufuji. DiffRoll: Diffusion-based generative music transcription with unsupervised pretraining capability. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1–5, Jun. 2023.

[55] G. Morrone, D. Michelsanti, Z.-H. Tan, and J. Jensen. Audio-visual speech inpainting with deep learning. pages 6653–6657, Jun. 2021.

[56] M. Borsos, Z. Sharifi and M. Tagliasacchi. Speechpainter: Text-conditioned speech inpainting. In *Proc. Interspeech*, Sep. 2022.

[57] E. Larsen and R.M. Aarts. *Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design*. Wiley, 2005.

[58] M Miron and MEP Davies. High frequency magnitude spectrogram reconstruction for music mixtures using convolutional autoencoders. In *Proc. Int. Conf. Digital Audio Effects (DAFX)*, pages 173–180, Aveiro, Portugal, Sep. 2018.

[59] M. Lagrange and F. Gontier. Bandwidth extension of musical audio signals with no side information using dilated convolutional neural networks. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 801–805, Barcelona, Spain, May 2020.

[60] V. Kuleshov, S. Z. Enam, and S. Ermon. Audio super resolution using neural networks. In *Proc. Int. Conf. Learning Repr.*, 2017.

[61] K. Zhang, Y. Ren, C. Xu, and Z. Zhao. WSRGlow: A Glow-based waveform generative model for audio, super-resolution. In *Proc. Interspeech*, pages 1649–1653, Shanghai, China, August 2021.

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

[63] E. Moliner and V. Välimäki. BEHM-GAN: Bandwidth extension of historical music using generative adversarial networks. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 31:943–956, Jul. 2023.

[64] J. Abel, M. Kaniewska, C. Guillaume, W. Tirry, H. Pulakka, V. Myllylä, J. Sjöberg, P. Alku, I. Katsir, D. Malah, et al. A subjective listening test of six different artificial bandwidth extension approaches in english, chinese, german, and korean. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 5915–5919, 2016.

[65] J. Makhoul and M. Berouti. High-frequency regeneration in speech coding systems. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pages 428–431, Washington D.C., USA, April 1979.

[66] Martin Dietz, Lars Liljeryd, Kristofer Kjorling, and Oliver Kunz. Spectral band replication, a novel approach in audio coding. In *Proc. Audio Eng. Soc. 112th Conv.*, Munich, Germany, Apr. 2002.

[67] K.-Y. Park and H. S. Kim. Narrowband to wideband conversion of speech using GMM based transformation. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1843–1846, Istanbul, Turkey, Jun. 2000.

[68] P. Jax and P. Vary. Artificial bandwidth extension of speech signals using mmse estimation based on a hidden Markov model. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, volume 1, 2003.

[69] H. Pulakka and P. Alku. Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum. *IEEE Trans. Audio Speech Lang. Process.*, 19(7):2170–2183, Aug. 2011.

[70] K. Li and C.-H. Lee. A deep neural network approach to speech bandwidth expansion. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 4395–4399, 2015.

[71] K. Li, Z. Huang, Y. Xu, and C.-H. Lee. DNN-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech. In *Proc. Interspeech*, 2015.

[72] S. Birnbaum, V. Kuleshov, Z. Enam, P. W. Koh, and S. Ermon. Temporal FiLM: Capturing long-range sequence dependencies with feature-wise modulations. In *Proc. Neural Inf. Process. Syst.*, 2019.

[73] A. Gupta, B. Shillingford, Y. Assael, and T. C. Walters. Speech bandwidth extension with wavenet. In *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pages 205–208, New Paltz, NY, USA, Oct. 2019.

[74] H. Wang and D. Wang. Towards robust speech super-resolution. *IEEE/ACM transactions on audio, speech, and language processing*, 29:2058–2066, 2021.

[75] T. Y. Lim, R. A. Yeh, Y. Xu, M. N. Do, and M. Hasegawa-Johnson. Time-frequency networks for audio super-resolution. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 646–650, Calgary, Canada, Apr. 2018.

[76] Y. Lin, J. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen. A two-stage approach to speech bandwidth extension. In *Proc. Interspeech*, pages 1689–1693, Brno, Czech Republic, August 2021.

[77] Z.-H. Ling, Y. Ai, Y. Gu, and L.-R. Dai. Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 26(5):883–894, 2018.

[78] H. Liu, W. Choi, X. Liu, Q. Kong, Q. Tian, and D. Wang. Neural vocoder is all you need for speech super-resolution. In *Proc. Interspeech*, Incheon, Korea, August 2022.

[79] L. Wen, L. Wang, Y. Zhang, and K. P. Choi. Multi-stage progressive audio bandwidth extension. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pages 422–427, 2023.

[80] S. Nercessian, A. Lukin, and J. Imort. DSP-Informed bandwidth extension using locally-conditioned excitation and linear time-varying filter subnetworks. In *Proc. Int. Workshop Acoustic Signal Enhancement*, pages 55–59, 2024.

[81] P-A. Grumiaux and M. Lagrange. Efficient bandwidth extension of musical signals using a differentiable harmonic plus noise model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2023(1):51, 2023.

[82] S. E. Eskimez, K. Koishida, and Z. Duan. Adversarial training for speech super-resolution. *IEEE Journal of Selected Topics in Signal Processing*, 13:347–358, 2019.

[83] D. Haws and X. Cui. CycleGAN bandwidth extension acoustic modeling for automatic speech recognition. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 6780–6784, 2019.

[84] S. Li, S. Villette, P. Ramadas, and D. J. Sinder. Speech Bandwidth Extension Using Generative Adversarial Networks. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 5029–5033, April 2018. ISSN: 2379-190X.

[85] J. Su, Y. Wang, A. Finkelstein, and Z. Jin. Bandwidth extension is all you need. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pages 696–700, Toronto, Canada, June 2021.

[86] Y. Li, M. Tagliasacchi, O. Rybakov, V. Ungureanu, and D. Roblek. Real-time speech frequency bandwidth extension. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 691–695, Toronto, Canada, June 2021.

[87] M. Mandel, O. Tal, and Y. Adi. AERO: Audio super resolution in the spectral domain. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Rhodes, Greece, Jun. 2023.

[88] W. Abreu and L. W. P. Biscainho. AEROMamba: An efficient architecture for audio super-resolution using generative adversarial networks and state space models. *arXiv preprint arXiv:2411.07364*, 2024.

[89] J. Hauret, T. Joubaud, V. Zimpfer, and E. Bavu. Configurable EBEN: Extreme bandwidth extension network to enhance body-conducted speech capture. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 31:3499–3512, 2023.

[90] J. Hauret, T. Joubaud, V. Zimpfer, and E. Bavu. EBEN: Extreme bandwidth extension network applied to speech signals captured with noise-resilient body-conduction microphones. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1–5, 2023.

[91] J. Lee and S. Han. NU-Wave: A diffusion probabilistic model for neural audio upsampling. In *Proc. Interspeech*, pages 1634–1638, Brno, Czech Republic, August 2021.

[92] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann. Analysing discriminative versus diffusion generative models for speech restoration tasks. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023.

[93] C.-Y. Yu, S.-L. Yeh, G. Fazekas, and H. Tang. Conditioning and sampling in variational diffusion models for speech super-resolution. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1–5, 2023.

[94] H. Wang, E. W. Healy, and D. Wang. Combined generative and predictive modeling for speech super-resolution. *arXiv preprint arXiv:2401.14269*, 2024.

[95] H. Liu, Q. Chen, K. Tian, W. Wang, and M. D. Plumbley. AudioSR: Versatile audio super-resolution at scale. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1076–1080, 2024.

[96] Y. Fang, J. Bai, J. Wang, and X. Zhang. Vector quantized diffusion model based speech bandwidth extension. *arXiv preprint arXiv:2409.05784*, 2024.

[97] S.-B. Kim, S.-H. Lee, H.-Y. Choi, and S.-W. Lee. Audio super-resolution with robust speech representation learning of masked autoencoder. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 32:1012–1022, 2024.

[98] V.-A. Nguyen, A. H. T. Nguyen, and A. W. H. Khong. TUNet: A block-online bandwidth extension model based on transformers and self-supervised pretraining. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022.

[99] M. Martínez Ramirez and J. Reiss. End-to-end equalization with convolutional neural networks. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, September 2018.

[100] C. J. Steinmetz, N. J. Bryan, and J. D. Reiss. Style transfer of audio effects with differentiable signal processing. *J. Audio Eng. Soc.*, 70(9):708–721, 2022.

[101] Y. Qian and P. Kabal. Combining equalization and estimation for bandwidth extension of narrowband speech. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, volume 1, pages I–713, 2004.

[102] T. Gerkmann and E. Vincent. Spectral masking and filtering. In E. Vincent, T. Virtanen, and S. Gannot, editors, *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018.

[103] J. A. Schell, S. Riter, and R. K. Cavin. Dereverberation by linear systems techniques. *IEEE Transactions on Geoscience Electronics*, 9(1):28–34, 1971.

[104] JB Allen. Speech dereverberation. *The Journal of the Acoustical Society of America*, 53(1_Supplement):322–322, 1973.

[105] P. A. Naylor and N. D. Gaubitch. *Speech Dereverberation*, volume 59. Springer, 2011.

[106] C. Chen, W. Sun, D. Harwath, and K. Grauman. Learning audio-visual dereverberation. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1–5, 2023.

[107] M. Miyoshi and Y. Kaneda. Inverse filtering of room acoustics. *IEEE Trans. Audio Speech Lang. Process.*, 36(2):145–152, 1988.

[108] I. Kodrasi, T. Gerkmann, and S. Doclo. Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2014.

[109] J. Mourjopoulos, P. Clarkson, and J. Hammond. A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1982.

[110] S. T. Neely and J. B. Allen. Invertibility of a room impulse response. *J. Acoust. Soc. Am.*, 66(1):165–169, 1979.

[111] J.M. Lemercier, S. Welker, and T. Gerkmann. Diffusion posterior sampling for informed single-channel dereverberation. In *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2023.

[112] T. Nakatani, B. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi. Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model. *IEEE Trans. Audio Speech Lang. Process.*, 16(8):1512–1527, 2008.

[113] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang. Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 18(7):1717–1731, 2010.

[114] F. Yohena and K. Yatabe. Single-channel blind dereverberation based on rank-1 matrix lifting in time-frequency domain. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2024.

[115] D. Schmid, S. Malik, and G. Enzner. A maximum a posteriori approach to multichannel speech dereverberation and denoising. In *Proc. Int. Workshop Acoustic Signal Enhancement*, 2012.

[116] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo. Speech dereverberation with convolutive transfer function approximation using map and variational deconvolution approaches. In *Proc. Int. Workshop Acoustic Signal Enhancement*, 2014.

[117] E. A. P. Habets. *Speech Dereverberation Using Statistical Reverberation Models*, pages 57–93. Springer, London, 2010.

[118] D. S. Williamson and D. Wang. Speech dereverberation and denoising using complex ratio masks. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 5590–5594, 2017.

[119] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang. Learning spectral mapping for speech dereverberation and denoising. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2015.

[120] V. Kothapally and J. H. L. Hansen. Monaural speech dereverberation using deformable convolutional networks. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 32:1712–1723, 2024.

[121] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger. Speech dereverberation using fully convolutional networks. In *Proc. Euro. Signal Proc. Conf.*, 2019.

[122] Y. Zhao, D. Wang, B. Xu, and T. Zhang. Monaural speech dereverberation using temporal convolutional networks with self attention. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 28:1598–1607, 2020.

[123] X. Liu, S.-J. Chen, and J. Hansen. Dual-path minimum-phase and all-pass decomposition network for single channel speech dereverberation. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2024.

[124] J. Su, Z. Jin, and A. Finkelstein. HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks. In *Proc. Interspeech*, 2020.

[125] V. Kothapally and J. HL Hansen. SkipConvGAN: Monaural speech dereverberation using generative adversarial networks via complex time-frequency masking. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 30:1600–1613, 2022.

[126] J. Su, Z. Jin, and A. Finkelstein. HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features. In *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pages 166–170, 2021.

[127] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann. StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 31:2724–2737, 2023.

[128] E. Moliner, S. Braun, and H. Gamper. Gaussian flow bridges for audio domain transfer with unpaired data. In *Proc. Int. Workshop Acoustic Signal Enhancement*, 2024.

[129] J. Serrà, S. Pascual, J. Pons, R. Araz, and D. Scaini. Universal speech enhancement with score-based diffusion. *arXiv*, 2022.

[130] H. Attias, J. Platt, A. Acero, and L. Deng. Speech denoising and dereverberation using probabilistic models. In *Proc. Neural Inf. Process. Syst.*, 2000.

[131] D. Baby and H. Bourlard. Speech dereverberation using variational autoencoders. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021.

[132] P. Wang and X. Li. RVAE-EM: Generative speech dereverberation based on recurrent variational auto-encoder and convolutive transfer function. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 496–500, 2024.

[133] K. Saito, N. Murata, T. Uesaka, C.-H. Lai, Y. Takida, T. Fukui, and Y. Mitsufuji. Unsupervised vocal dereverberation with diffusion-based generative models. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023.

[134] N. Murata, K. Saito, C.-H. Lai, Y. Takida, T. Uesaka, Y. Mitsufuji, and S. Ermon. GibbsDDRM: A partially collapsed Gibbs sampler for solving blind inverse problems with denoising diffusion restoration. In *Proc. Int. Conf. Machine Learning*, volume 202, pages 25501–25522, 2023.

[135] H. Gamper and I. J. Tashev. Blind reverberation time estimation using a convolutional neural network. In *Proc. Int. Workshop Acoustic Signal Enhancement*, pages 136–140. IEEE, 2018.

[136] N. J. Bryan. Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1–5. IEEE, 2020.

[137] C. J. Steinmetz, V. K. Ithapu, and P. Calamia. Filtered noise shaping for time domain room impulse response estimation from reverberant speech. In *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2021.

[138] P. Götz, C. Tuna, A. Walther, and E. A. P. Habets. Blind reverberation time estimation in dynamic acoustic conditions. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 581–585, 2022.

[139] F. Lluís and N. Meyer-Kahlen. Blind spatial impulse response generation from separate room-and scene-specific information. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2025.

[140] P. Záviška, P. Rajmic, A. Ozerov, and L. Rencker. A survey and an extensive evaluation of popular audio declipping methods. *IEEE J. Selected Topics Signal Process.*, 15(1):5–24, 2020.

[141] S. Kitić, N. Bertin, and R. Gribonval. Sparsity and cosparsity for audio declipping: a flexible non-convex approach. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 243–250. Springer, 2015.

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

[143] C. Gaultier, S. Kitić, R. Gribonval, and N. Bertin. Sparsity-based audio declipping methods: Selected overview, new algorithms, and large-scale evaluation. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 29:1174–1187, 2021.

[144] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley. A constrained matching pursuit approach to audio declipping. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 329–332, 2011.

[145] C. Bilen, A. Ozerov, and P. Pérez. Audio declipping via nonnegative matrix factorization. In *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pages 1–5, 2015.

[146] W. Mack and E. A. P. Habets. Declipping speech using deep filtering. In *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pages 200–204, 2019.

[147] J. Yi, J. Koo, and K. Lee. DDD: A perceptually superior low-response-time DNN-based declipper. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 801–805, 2024.

[148] V. Sechaud, L. Jacques, P. Abry, and J. Tachella. Equivariance-based self-supervised learning for audio signal recovery from clipped measurements. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 852–856, 2024.

[149] Y. Kwon and J.-W. Choi. Speech-declipping transformer with complex spectrogram and learnerble temporal features. *arXiv preprint arXiv:2409.12416*, 2024.

[150] T. Tanaka, K. Yatabe, M. Yasuda, and Y. Oikawa. APPLADE: Adjustable plug-and-play audio declipper combining DNN with sparse optimization. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1011–1015, 2022.

[151] T. Tanaka, K. Yatabe, and Y. Oikawa. UPGLADE: Unplugged plug-and-play audio declipper based on consensus equilibrium of DNN and sparse optimization. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1–5, 2023.

[152] F. R. Ávila, M. P. Tcheou, and L. W. P. Biscainho. Audio soft declipping based on constrained weighted least squares. *IEEE Signal Process. Letters*, 24(9):1348–1352, 2017.

[153] W.-S. Gan and N. Oo. Harmonic and intermodulation analysis of nonlinear devices used in virtual bass systems. In *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

[154] F. Esqueda, H. Poyntinen, J. Parker, and S. Bilbao. Virtual analog model of the lockhart wave folder. In *Proceedings of the 14th Sound and Music Computing Conference*. Aalto University, 2017.

[155] Y. Li, B. Gfeller, M. Tagliasacchi, and D. Roblek. Learning to denoise historical music. In *Proc. 21st ISMIR Conf.*, pages 504–511, Montréal, Canada, October 2020.

[156] I. Irigaray, M. Rocamora, and L. W. P. Biscainho. Noise reduction in analog tape audio recordings with deep learning models. In *AES International Conference on Audio Archiving, Preservation & Restoration*. Audio Engineering Society, 2023.

[157] T. Saeki, S. Takamichi, T. Nakamura, N. Tanji, and H. Saruwatari. SelfRemaster: Self-supervised speech restoration for historical audio resources. *IEEE Access*, 11:144831–144843, 2023.

[158] M. Bosi, S. Canazza, N. Pretto, A. Russo, and M Spanio. From tape to code: An international AI-based standard for audio cultural heritage preservation don't play that song for me (if it's not preserved with ARP!). *IEEE Access*, 2024.

[159] A. Ragano, E. Benetos, and A. Hines. Audio quality assessment of vinyl music collections using self-supervised learning. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1–5, May 2023.

[160] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. Int. Conf. Machine Learning*, 2015.

[161] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Proc. Neural Inf. Process. Syst.*, 2020.

[162] Y. Song, J. Sohl-Dickstein, D. P Kingma, et al. Score-based generative modeling through stochastic differential equations. In *Proc. Int. Conf. Learning Representations (ICLR)*, May 2021.

[163] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. In *Proc. Neural Inf. Process. Syst.*, 2022.

[164] X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proc. Int. Conf. Learning Repr.*, 2023.

[165] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *Proc. Int. Conf. Learning Repr.*, 2023.

[166] A. Tong, K. Fatras, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. Expert Certification.

[167] A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6(4):695–709, Dec. 2005.

[168] P. Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

[169] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *Proc. Int. Conf. Learning Repr.*, 2020.

[170] E. Süli and D. F. Mayers. *An Introduction to Numerical Analysis*. Cambridge university press, 2003.

[171] T. Dockhorn, A. Vahdat, and K. Kreis. GENIE: Higher-order denoising diffusion solvers. In *Proc. Neural Inf. Process. Syst.*, volume 35, pages 30150–30166, 2022.

[172] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

[173] P. E. Kloeden and E. Platen. *Stochastic Differential Equations*. Springer, 1992.

[174] A. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *Proc. Int. Conf. Machine Learning*, 2021.

[175] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2022.

[176] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. *Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2022.

[177] E. Hoogeboom, V. Garcia Satorras, C. Vignac, and M. Welling. Equivariant diffusion for molecule generation in 3d. In *Proc. Int. Conf. Machine Learning*, pages 8867–8887. PMLR, 2022.

[178] Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *Proc. Neural Inf. Process. Syst.*, volume 37, pages 133345–133385, 2024.

[179] S. Dieleman, L. Sartran, A. Roshannai, N. Savinov, Y. Ganin, P. H Richemond, A. Doucet, R. Strudel, C. Dyer, C. Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

[180] G. Mittal, J. Engel, C. Hawthorne, and I. Simon. Symbolic music generation with diffusion models, 2021.

[181] Benedikt Fesl, Michael Baur Florian Strasser, Michael Joham, and Wolfgang Utschick. Diffusion-based generative prior for low-complexity MIMO channel estimation. *IEEE Wireless Communications Letters*, 2024.

[182] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. DiffWave: A versatile diffusion model for audio synthesis. In *Proc. Int. Conf. Learning Repr.*, May 2021.

[183] N. Chen, Y. Zhang, H. Zen, et al. WaveGrad: Estimating gradients for waveform generation. In *Proc. Int. Conf. Learning Representations (ICLR)*, May 2021.

[184] S. Rouard and G. Hadjeres. CRASH: Raw audio score-based generative modeling for controllable high-resolution drum sound synthesis. In *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2021.

[185] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov. Grad-TTS: A diffusion probabilistic model for text-to-speech. In *Proc. Int. Conf. Machine Learning*, pages 8599–8608. PMLR, 2021.

[186] C. Hawthorne, I. Simon, A. Roberts, et al. Multi-instrument music synthesis with spectrogram diffusion. In *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2022.

[187] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[188] Anonymous. Elucidating the design space of text-to-audio models. In *Proc. Int. Conf. Learning Repr.*, 2024. under review.

[189] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *Proc. Int. Conf. Machine Learning*, 2023.

[190] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2024.

[191] F. Schneider, Z. Jin, and B.. Schölkopf. Moûsai: Text-to-music generation with long-context latent diffusion. *arXiv*, 2023.

[192] Z. Evans, C.J. Carr, J. Taylor, S. H. Hawley, and J. Pons. Fast timing-conditioned latent audio diffusion. In *Proc. Int. Conf. Machine Learning*, 2024.

[193] Z. Evans, J. D. Parker, C.J. Carr, Z. Zukowski, J. Taylor, and J. Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024.

[194] M. WY Lam, Q. Tian, T. Li, Z. Yin, S. Feng, M. Tu, Y. Ji, R. Xia, M. Ma, X. Song, et al. Efficient neural music generation. In *Proc. Neural Inf. Process. Syst.*, volume 36, 2024.

[195] G. L. Lan, B. Shi, Z. Ni, S. Srinivasan, A. Kumar, B. Ellis, D. Kant, V. Nagaraja, E. Chang, W.-N. Hsu, et al. High fidelity text-guided music editing via single-stage flow matching. *arXiv preprint arXiv:2407.03648*, 2024.

[196] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[197] S. Welker, J. Richter, and T. Gerkmann. Speech enhancement with score-based generative models in the complex STFT domain. In *Proc. Interspeech*, 2022.

[198] Y.-J. Lu, Y. Tsao, and S. Watanabe. A study on speech enhancement based on diffusion probabilistic model. In *Proc. Asia-Pacific Signal and Information Processing Association (APSIPA)*, 2021.

[199] P. Gonzalez, Z-H. Tan, J. Østergaard, J. Jensen, T. Alstrøm, and T. May. Investigating the design space of diffusion models for speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2024.

[200] Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J Bryan. Music controlnet: Multiple time-varying controls for music generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2692–2703, 2024.

[201] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[202] H. Chung, J. Kim, G. Y. Park, H. Nam, and J. C. Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024.

[203] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao. Conditional diffusion probabilistic model for speech enhancement. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022.

[204] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Proc. Neural Inf. Process. Syst.*, 34:17695–17709, 2021.

[205] A. Jukić, R. Korostik, J. Balam, and B. Ginsburg. Schrodinger bridge for generative speech enhancement. In *Proc. Interspeech*, 2024.

[206] Zhifeng Kong, Kevin J Shih, Weili Nie, Arash Vahdat, Sang-gil Lee, Joao Felipe Santos, Ante Jukic, Rafael Valle, and Bryan Catanzaro. A2SB: Audio-to-audio Schrodinger bridges. *arXiv preprint arXiv:2501.11311*, 2025.

[207] V. De Bortoli, I. Korshunova, A. Mnih, and A. Doucet. Schrodinger bridge flow for unpaired data translation. *Proc. Neural Inf. Process. Syst.*, 37:103384–103441, 2024.

[208] X. Su, J. Song, C. Meng, and S. Ermon. Dual diffusion implicit bridges for image-to-image translation. In *Proc. Int. Conf. Learning Repr.*, 2023.

[209] M. Mancusi, Y. Halychanskyi, K. W. Cheuk, E. Moliner, C.-H. Lai, S. Uhlich, J. Koo, M. A. Martínez-Ramírez, W.-H. Liao, G. Fabbro, and Y. Mitsufuji. Latent diffusion bridges for unsupervised musical audio timbre transfer. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1–5, 2025.

[210] G. Daras, H. Chung, C.-H. Lai, Y. Mitsufuji, J. C. Ye, P. Milanfar, A. G. Dimakis, and M. Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024.

[211] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye. Diffusion posterior sampling for general noisy inverse problems. In *Proc. Int. Conf. Learning Repr.*, May 2023.

[212] H. Chung, J. Kim, S. Kim, and J. C. Ye. Parallel diffusion models of operator and image for blind inverse problems. *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2023.

[213] N. Murata, K. Saito, C.-H. Lai, Y. Takida, T. Uesaka, Y. Mitsufuji, and S. Ermon. GibbsDDRM: A partially collapsed Gibbs sampler for solving blind inverse problems with denoising diffusion restoration. In *Proc. Int. Conf. Machine Learning*, 2023.

[214] M. Dontas, Y. He, N. Murata, Y. Mitsufuji, J Z. Kolter, and R. Salakhutdinov. Blind inverse problem solving made easy by text-to-image latent diffusion. *arXiv preprint arXiv:2412.00557*, 2024.

[215] C. Laroche, A. Almansa, and E. Coupete. Fast diffusion EM: a diffusion model for blind inverse problems with application to deconvolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5271–5281, 2024.

[216] 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.*, 14(6):1902–1911, Nov. 2006.

[217] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proc. Interspeech*, pages 2350–2354, August 2019.

[218] C. K. A. Reddy, V. Gopal, and R. Cutler. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021.

[219] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2001.

[220] J. Jensen and C. Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 24(11):2009–2022, 2016.

[221] S. Kim, H. Kim, and S. Yoon. Guided-TTS 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data. *arXiv preprint arXiv:2205.15370*, 2022.

[222] C. Hernandez-Olivan, K. Saito, N. Murata, C-H. Lai, M. A Martínez-Ramirez, W-H. Liao, and Y. Mitsufuji. VRDMG: Vocal restoration via diffusion posterior sampling with multiple guidance. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 596–600, 2024.

[223] A. Iashchenko, P. Andreev, I. Shchekotov, N. Babaev, and D. Vetrov. UnDiff: Unsupervised voice restoration with unconditional diffusion model. In *Proc. Interspeech*, 2023.

[224] M. Turunen, E. Moliner, and V. Välimäki. AI-based enhancement of gramophone recordings: reconstructing voices from the early 20th century. *unpublished*, 2025.

[225] E. Moliner, M. Švento, A. Wright, L. Juvela, P. Rajmic, and V. Välimäki. Unsupervised estimation of nonlinear audio effects: Comparing diffusion-based and adversarial approaches. *arXiv preprint arXiv:2504.04751*, 2025.

[226] E. Thuillier, J.-M. Lemercier, E. Moliner, T. Gerkmann, and V. Välimäki. HRTF estimation using a score-based prior. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2025.

[227] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann. Causal diffusion models for generalized speech enhancement. *IEEE Open Journal of Signal Processing*, 2024.

[228] M. Švento, P. Rajmic, and O. Mokrỳ. Plug-and-play audio restoration with diffusion denoiser. In *Proc. Int. Workshop Acoustic Signal Enhancement*, pages 115–119, 2024.

[229] T. Dhyani, F. Lux, M. Mancusi, G. Fabbro, F. Hohl, and N. T. Vu. High-resolution speech restoration with latent diffusion model. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2025.

[230] Z. Novack, J. McAuley, T. Berg-Kirkpatrick, and N. J. Bryan. DITTO-2: Distilled diffusion inference-time t-optimization for music generation. In *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2024.

[231] Z. Novack, G. Zhu, J. Casebeer, J. McAuley, T. Berg-Kirkpatrick, and N. J. Bryan. Presto! Distilling steps and layers for accelerating music generation. In *Proc. Int. Conf. Learning Repr.*, 2025.

Errata

Publication I

The norm of the gradient used to normalize the step size $\xi(\tau)$ should not be squared. The correct definition for the step size (or scaling factor) is $\xi(\tau) = \xi' \sqrt{N}/(\tau \|G\|)$.

In Eq. (4), the symbol $\simeq$ was used, but $\approx$ would be more appropriate.

Publication II

The norm of the gradient used to normalize the step size $\xi(\tau)$ in Eq. (8) should not be squared. This is the same error as in Publication I.

Publication III

The norm of the gradient used to normalize the step size $\xi(\tau)$ in Eq. (9) should not be squared. This is the same error as in Publication I.

Publication V

Eq. (3) contains a sign error. The correct equation is:

$$\hat{\mathbf{x}}_0 \stackrel{\Delta}{=} \hat{\mathbf{x}}_0(\mathbf{x}_\tau, \tau) = \mathbf{x}_\tau + \sigma(\tau)^2 \mathbf{s}_\theta(\mathbf{x}_\tau, \tau).$$

Business, Economy
Art, Design, Architecture
Science, Technology
Crossover
Doctoral Theses

Aalto DT 138/2025

ISBN 978-952-64-2646-4
ISBN 978-952-64-2645-7 (pdf)

Aalto University
School of Electrical Engineering
Department of Information and Communications Engineering
aalto.fi