A DIFFUSION-BASED GENERATIVE EQUALIZER FOR MUSIC RESTORATION

Eloi Moliner[1], Maija Turunen[2], Filip Elvander[1] and Vesa Välimäki[1]

[1]Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Espoo, Finland
[2]Sibelius Academy, University of the Arts Helsinki, Helsinki, Finland
`eloi.moliner@aalto.fi`

ABSTRACT

This paper presents a novel approach to audio restoration, focusing on the enhancement of low-quality music recordings, and in particular historical ones. Building upon a previous algorithm called BABE, or Blind Audio Bandwidth Extension, we introduce BABE-2, which presents a series of significant improvements. This research broadens the concept of bandwidth extension to *generative equalization*, a novel task that, to the best of our knowledge, has not been explicitly addressed in previous studies. BABE-2 is built around an optimization algorithm utilizing priors from diffusion models, which are trained or fine-tuned using a curated set of high-quality music tracks. The algorithm simultaneously performs two critical tasks: estimation of the filter degradation magnitude response and hallucination of the restored audio. The proposed method is objectively evaluated on historical piano recordings, showing a marked enhancement over the prior version. The method yields similarly impressive results in rejuvenating the works of renowned vocalists Enrico Caruso and Nellie Melba. This research represents an advancement in the practical restoration of historical music. Historical music restoration examples are available at: research.spa.aalto.fi/publications/papers/dafx-babe2/.

1. INTRODUCTION

Historical music recordings suffer from severe impairment due to limitations of the physical recording media, alongside the wear and tear from playback and storage [1, 2]. These degradations can take the form of additive disturbances, decreased bandwidth, coloration, and nonlinear effects [3]. The ultimate goal is to enhance the quality of historical music recordings as to meet the audio standards of today, making it more accessible for people used to modern sound quality. Restoring these recordings is challenging for several reasons. First, the problem is ill-posed, meaning there are many potential solutions, which complicates the search for the correct one. Additionally, the degradation process is unknown, making this a blind inverse problem for which the solution must be found without definite information on the necessary repairs.

In the 1970s, Stockham et al. began groundbreaking efforts in digitally restoring gramophone recordings, notably those of the celebrated singer Enrico Caruso [4]. At the time, the approach was limited by technological constraints, relying on the estimation of an equalization curve derived from the spectral average of a contemporary recording. Since then, the field of audio restoration has evolved significantly [1]. Today, data-driven techniques employing deep neural networks can remove all additive disturbances

simultaneously [5]. Additionally, deep generative models, such as diffusion models [6], address complex restoration tasks, such as bandwidth extension [7], with unprecedented success. This study investigates a critical question: "*Can deep generative models elevate the quality of historical music recordings to modern standards?*"

To aid in this investigation, we introduce and explore the concept of *generative equalization*, a task designed to overcome the limitations of traditional equalization. Traditional equalizers face limitations on ill-conditioned scenarios, i.e., when some frequency bands are not present in the original recording. In such cases, the equalization curve must be cropped, otherwise, it will only amplify background noise. In contrast, a generative equalizer can synthesize the missing spectral components as to achieve the target spectral profile, as it is a generalization of bandwidth extension.

This work builds upon BABE, the authors' prior research in audio bandwidth extension [7], advancing it within the framework of generative equalization. While this method showed promise for addressing lowpass degradation, its application to real-world historical recordings revealed significant limitations, primarily due to an oversimplified degradation model. The model inadequately described the complex coloration effects often present in historical recordings, effects that are typically attributed to the recording technology, such as the recording horn [3, 2].

This paper introduces BABE-2, an improvement and extension of the original BABE model, bringing a set of technical contributions designed to refine its efficacy for the restoration of historical music recordings. This includes incorporating a novel parametrization of the degradation filter model as a zero-phase frequency response estimate, allowing for a more flexible and richer filter structure extending beyond the previously constrained lowpass characteristic. Additionally, we implement a regularization strategy to counteract the breakpoint-collapse problem encountered in BABE, introduce noise regularization to foster stable convergence, and propose an initialization scheme based on the long-term average spectra (LTAS), alongside other critical implementation details.

The remaining of this paper is structured as follows. Sec. 2 presents sufficient information to follow the description of the new algorithm. There, the basics of diffusion models, diffusion posterior sampling, and the previous BABE algorithm are described. Following this, Sec. 3 delves into the enhancements made in the BABE-2 method, expanding the bandwidth extension techniques introduced in [7] to encompass the broader concept of generative equalization. Sec. 4 reports on experiments conducted with historical recordings of piano and singing voice. The discussion in Sec. 5 describes our procedure for selecting the training data used for restoring specific historical singers, and elucidates on some of the capabilities and limitations of BABE-2 for singing voice restoration. Finally, Sec. 6 concludes the paper.

2. BACKGROUND

2.1. Diffusion Models

Diffusion models have emerged as a powerful class of generative models, demonstrating remarkable capabilities in various domains, including image and audio processing [8, 9, 6]. These models operate by gradually transforming data through a process of adding and removing noise. In the context of audio, we define time as the variable τ, with $\mathbf{x}_\tau$ representing the state of the audio data at time τ. At $\tau = T$, $\mathbf{x}_T$ is distributed as Gaussian noise, while $\mathbf{x}_0$ represents the clean audio signal. Following the formulation by Karras et al. [9], the underlying Ordinary Differential Equation governing this transformation is written as

$$d\mathbf{x} = -\tau \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau) d\tau. \tag{1}$$

The *score* function $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau) d\tau$, which guides the model to reverse the noise addition process, is approximated using a denoiser $D_\theta(\mathbf{x}_\tau, \tau)$, a deep neural network with parameters θ:

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

where $\sigma(\tau)^2$ is the noise variance at timestep τ[1]. For training these models, Denoising Score Matching is often employed, which aims to minimize the difference between the denoised and the original clean audio. This objective is encapsulated as follows:

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

In the subsequent sections, the term $\hat{\mathbf{x}}_0$ denotes the denoised estimate obtained through the application of neural network D_θ.

2.2. Diffusion Posterior Sampling

Many restoration tasks can be understood as inverse problems, where an observed measurement signal is obtained by applying a certain degradation operator $H(\cdot)$ to an unknown signal $\mathbf{x}_0$, yielding the observation $\mathbf{y} = H(\mathbf{x}_0)$. Diffusion models are highly effective as data-driven priors for solving inverse problems, where the goal is to estimate the original signal $\mathbf{x}_0$ from the degraded observations $\mathbf{y}$ [10]. When the degradation operator $H(\cdot)$ is known, these models can be used to approximate the posterior distribution $p(\mathbf{x}_0|\mathbf{y})$. Ideally, to solve an inverse problem using diffusion models, one can modify the ODE in Eq. (1), replacing the score with $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau|\mathbf{y})$, also known as the *posterior score*.

We adopt the "Diffusion Posterior Sampling" method introduced by Chung et al. [11], which utilizes a Bayesian framework, decomposing the posterior score into two parts:

$$\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau|\mathbf{y}) = \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau) + \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{y}|\mathbf{x}_\tau). \tag{4}$$

The first term, the *prior score*, is derived from Eq. (2), while the second, the *likelihood score*, is estimated through the gradients of an application-specific cost function:

$$\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{y}|\mathbf{x}_\tau) \approx -\xi(\tau) \nabla_{\mathbf{x}_\tau} C_{\text{audio}}(\mathbf{y}, H(\hat{\mathbf{x}}_0)), \tag{5}$$

where $\xi(\tau)$ serves as a gradient normalization factor [10].

Within this framework, $\hat{\mathbf{x}}_0 = D_\theta(\mathbf{x}_\tau, \tau)$ represents the Minimum Mean Squared Error (MMSE) estimate of the clean audio $\mathbf{x}_0$ given the noisy or transformed state $\mathbf{x}_\tau$. It is important to note that implementing this method necessitates differentiating through the

---

[1]Following Karras et al. [9], we employ $\sigma(\tau) = \tau$.

denoiser, which is a deep neural network, and adds a layer of complexity to the process. Among other applications, this approach has been successfully used for solving audio restoration tasks, such as bandwidth extension, inpainting, and declipping [10].

2.3. Application to Blind Inverse Problems

In scenarios where the degradation operator H is unknown, the challenges in solving the inverse problem increase significantly. These are referred to as blind inverse problems. Some recent works offer solutions for solving blind inverse problems with diffusion models by jointly optimizing the degradation operator alongside the data during the sampling process [12]. The methodology relies on alternating between updates from diffusion sampling and optimization steps for finding the unknown operator.

Our recent work BABE [7], focusing on the blind bandwidth extension of music signals without knowledge of the lowpass filter characteristics, fits in this framework. In this approach, the degradation operator H_ϕ is modeled as a zero-phase filter in the frequency domain, with a piecewise linear magnitude response defined by a set of parameters ϕ. The parameters are iteratively optimized through stochastic gradient descent while sampling from the diffusion model in a coarse-to-fine manner.

In BABE, the filter design is constrained to yield a lowpass filter by projecting the parameters to ensure they describe a strictly decaying function [7]. This choice stabilizes the optimization process, as the primary focus is on responses attenuating the high frequencies. Such a parameterization, however, does not accommodate spectral coloration or resonances, which are common in historical recordings and can significantly impact the audio quality. We hypothesize that the constrained modeling capability degrades the performance of posterior sampling in real historical recordings due to unaddressed spectral properties, pushing the samples towards out-of-distribution regions compared to the training data and thus reducing the reliability of diffusion model predictions.

3. BABE-2: UNIQUE CONTRIBUTIONS

3.1. Filter Parameterization

In addressing blind inverse problems within a posterior sampling framework, it is essential to specify a set of pertinent degradation operators. Following our earlier approach [7], the focus is on zero-phase frequency-domain forward filters denoted as H_ϕ. The relationship between the observations $\mathbf{y}$ and the reconstructed estimate $\hat{\mathbf{x}}_0$ through these filters is given by:

$$\mathbf{y} \approx \mathcal{F}^{-1}(H_\phi \odot \mathcal{F}(\hat{\mathbf{x}}_0)), \tag{6}$$

where $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ represent the Fourier transform and its inverse, respectively. We use the FFT length of 4096 samples. For simplicity, we use the notation $H_\phi(\hat{\mathbf{x}}_0)$ interchangeably to describe this operation throughout the paper.

The lowpass filter operator, initially applied in the BABE framework [7], is extended to a parametric frequency-response equalizer. The equalizer configuration, depicted in Fig. 1, incorporates an adjustable anchor cutoff frequency (f_0) maintained at $0\,\text{dB}$ to ensure stability in the filter's gain across its operational bandwidth. The design includes a set of adjustable slopes and breakpoints above this anchor frequency (f_i and A_i), similar to those in BABE but with the flexibility of non-strictly decaying slopes. Beyond a certain point, marked by the last optimizable breakpoint and a
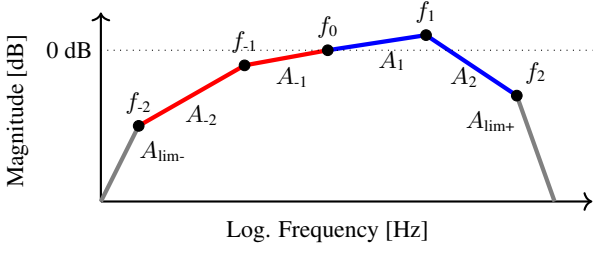
Figure 1: *Proposed frequency-response equalizer model consists of breakpoints creating a piecewise linear magnitude response.*

fixed steep slope set to $A_{\text{lim+}} = -80\,$dB, it is assumed that the band has no significant spectral content. This structure is symmetrically mirrored below the anchor cutoff, featuring adjustable slopes and breakpoints (f_{-i} and A_{-i}), culminating in a fixed steep slope $A_{\text{lim-}} = 80\,dB$ at the lowest frequency band to delineate the effective spectral range. The filter is piecewise differentiable with respect to the cutoff frequencies and slope parameters.

Choosing the right number of filter stages is crucial, especially under the assumption that the magnitude response should exhibit smoothness. Employing a filter with more stages provides greater flexibility during optimization but risks overfitting to minor spectral details. This overfitting can lead to an unrealistic magnitude response that could negatively impact the optimization. With this in mind, a smaller number of stages is favored to maintain smoothness, and we have chosen a configuration with 5 breakpoints.

To mitigate potential training instabilities, we impose restrictions on the slopes, ensuring that they do not exceed predefined minimum and maximum values of -40 dB and 40 dB, respectively. Additionally, the cutoff frequencies are confined within the range $(f_{\text{min}}, f_{\text{max}})$, which is determined by the FFT size employed in the filter parameterization. Furthermore, we enforce that a sequential ordering of the cutoff frequencies, requiring each frequency to be less than the subsequent one. These constraints are implemented through the application of projections within the specified ranges during the optimization process. Ideally, the constraint on the ordering of cutoff frequencies would be unnecessary, as the regularization scheme introduced in the next subsection aims to prevent parameter configurations that could lead to such scenarios.

3.2. Breakpoint-Collapse Regularization

We identify a previously unrecognized problem in the BABE algorithm [7], which we term *breakpoint collapse*. The problem occurs when, during inference, two or more cutoff frequency breakpoints converge to very close values, receiving nearly identical gradients. Consequently, the breakpoints cannot separate from each other for the rest of the process, reducing the flexibility of the degradation model by losing these breakpoint and slope parameters.

Interestingly, this problem was not prevalent in simulated lowpass filtered recordings as tested in BABE. However, it became significantly more frequent and problematic when the model was applied to real historical recordings, particularly those that differed much from the training data distribution. To address this problem, we introduce a novel regularization term, the Breakpoint-Collapse Regularization (BCR). The BCR term is expressed as

$$\mathcal{C}_{\text{BCR}}(\phi) = e^{-\beta(f_{\text{min}} - f_{-S'})} + \sum_{i=-S'}^{S-1} e^{-\beta(f_i - f_{(i+1)})} + e^{-\beta(f_S - f_{\text{max}})}.$$

$$(7)$$

This function imposes a minimal cost when breakpoints are adequately spaced but enforces an exponentially increasing cost, when they approach each other. The parameter β, set to 0.1 in our implementation, modulates this behavior.

To integrate this into our framework, the total cost function is modified to include the BCR term, weighted appropriately. The revised total cost equation is

$$\mathcal{C}_{\text{total}} = \mathcal{C}_{\text{filter}}(\mathbf{y}, H_\phi(\hat{\mathbf{x}}_0)) + \gamma_{\text{BCR}} \mathcal{C}_{\text{BCR}}(\phi),$$

$$(8)$$

where γ_{BCR}, the BCR term weight, is set to 10 in our experiments.

3.3. Noise Regularization

Our parametric model, designed to fit linear magnitude responses, faces challenges in addressing nonlinear artifacts within historical audio recordings. These artifacts, often made worse by preprocessing steps like denoising, can bias the optimization of magnitude response parameters, leading to convergence at local minima and suboptimal restoration outcomes. This problem was overlooked in the previous iteration of the method [7]. As we show in Sec. 4.1, the performance of BABE on historical recordings is suboptimal.

To solve these problems, we suggest adding a certain amount of stochasticity, or "noise," into the optimization objective. We do this by adding random noise to our observations, which we describe as $\bar{\mathbf{y}} = \mathbf{y} + \gamma\boldsymbol{\epsilon}$, where γ is the noise scale and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a vector of Gaussian noise. Technically, this approach smooths the likelihood distribution by convolving it with a Gaussian kernel. This trick enables the optimization process to navigate the cost function landscape more effectively than without it. Each iteration's unique noise ensures the optimizer does not prematurely converge to local minima, thereby encouraging a thorough search for optimal solutions. Furthermore, the additive noise potentially masks some of the artifacts that may be present in the observed data, preventing these artifacts from negatively affecting the optimization. This approach aligns with techniques used by Sadat et al. [13], which aimed to enhance diversity in image generation. In Sec. 4.1, we conduct an ablation study where this technique is not applied, which is there denoted as "BABE-2 w/o noise reg."

3.4. Long-term Average Spectrum-based Initialization

A known blind equalization technique designs a filter to align the input LTAS with that off a reference set [4]. To achieve this, we calculate the reference LTAS, denoted as $\tilde{\mathbf{R}}_{\text{LTAS}}$, by applying the Short-Time Fourier Transform (STFT) to the reference set and average the power spectra across frames. In the STFT, we use 4096-sample windows at the sampling rate of $f_s = 44.1\,$kHz, or 2048-sample windows at $f_s = 22.05\,$kHz. We employ a hop size of 1/4 in both cases. One-third-octave band Gaussian smoothing is applied to the averaged spectra. The LTAS of the input $\tilde{\mathbf{Y}}_{\text{LTAS}}$ is determined in the same way, based on the original recording.

By computing the ratio between the LTAS spectra, the equalization filter is obtained as

$$|H_{\text{LTAS}}| = |\tilde{\mathbf{Y}}_{\text{LTAS}}| / |\tilde{\mathbf{R}}_{\text{LTAS}}|,$$

$$(9)$$

where H_{LTAS} is an estimate of the forward degradation. The inverse filter H_{LTAS}^{-1} can be used as a frequency-domain equalizer:

$$\tilde{\mathbf{y}} = \mathcal{F}^{-1}(H_{\text{LTAS}}^{-1} \odot \mathcal{F}(\mathbf{y})),$$

$$(10)$$

where $\odot$ is the Hadamard product.

From now on, this equalization method is referred to as "LTAS-EQ." While this approach proved useful for correcting significant colorations [4], it has strong limitations. Firstly, if the observations are bandlimited, the ratio in Eq. (9) is ill-conditioned, as it approaches zero for some frequency bands, causing the inverse filter to amplify to extremely high values. To mitigate this, we limit H_{LTAS} at $-20\,\text{dB}$. Secondly, an estimate involving time averages assumes the measured signal is ergodic, which may often not be true. The LTAS method overlooks the dynamics of the audio, such as the unique sound of a piano when played at different intensities. Consequently, during LTAS computation, louder sections may disproportionately influence the average spectrum, leading to an imbalance. Due to these limitations, relying solely on this approach for audio restoration is not recommended.

Considering these limitations, we adopt this method as a preliminary step that BABE-2 will refine. Our goal is to facilitate the equalization process, aiming for a more stable optimization of the filter at the initial stages and achieving convergence with fewer iterations. We explore two approaches to accomplish this. Firstly, we utilize H_{LTAS} for an improved initial setup of the audio signal at starting time T, $\mathbf{x}_T$. Unlike in BABE [7], where $\mathbf{x}_T$ began as a noisy version of the observations $\mathbf{x}_T \sim \mathcal{N}(\mathbf{y}, \sigma\text{start}^2\mathbf{I})$, we now introduce a modified approach. We suggest beginning with a noisy version of the equalized observations $\mathbf{x}_T \sim \mathcal{N}(\tilde{\mathbf{y}}, \sigma\text{start}^2\mathbf{I})$, where $\tilde{\mathbf{y}}$ is derived using Eq. (10). This version of our method is hereon referred to as "BABE-2 w/ LTAS-EQ init."

The second strategy extends the use of LTAS-EQ beyond initial setup by incorporating it into the optimization objective. This involves replacing the original recording $\mathbf{y}$ with $\tilde{\mathbf{y}}$ in the reconstruction cost functions from Eqs. (12) and (??). It is crucial to recognize that in this context, the estimated frequency response H_ϕ no longer corresponds directly to the original recording response $\mathbf{y}$. Instead, the connection between the original recording and the estimated restored version $\hat{\mathbf{x}}_0$ becomes a composite of $H\phi$ and H_{LTAS} as formalized below:

$$\mathbf{y} \approx \mathcal{F}^{-1}\left((H_\phi \odot H_{\text{LTAS}}) \odot \mathcal{F}(\hat{\mathbf{x}}_0)\right). \tag{11}$$

We denote this variant as "BABE-2 w/ LTAS-EQ obj."

3.5. Inference Algorithm

The inference algorithm of BABE-2 shares strong similarities with the one employed in BABE [7]. The pseudocode in Algorithm 1 summarizes the inference process, highlighting the the new contributions in blue. Despite, for brevity reasons, Algorithm 1 shows an adaptation of a 1st order Euler sampler, we use the 2nd order stochastic sampler introduced in [9]. One relevant addition is the use of the Adam optimizer [14] to optimize the filter parameters, instead of a simpler stochastic gradient descent as used in [7]. Adam applies adaptive per-parameter learning rates and second-order moments of the gradients, which stabilizes the performance for non-convex optimization problems with sparse gradients.

As the reconstruction cost function for the audio signal, we employ the same standard L_2 norm as used in BABE:

$$C_{\text{audio}}(\mathbf{y}, H_\phi(\hat{\mathbf{x}}_0)) = \|\mathbf{y} - H_\phi(\hat{\mathbf{x}}_0)\|_2^2. \tag{12}$$

We also use the same reconstruction cost function as BABE for optimizing the filter parameters, which consists of a frequency-weighted L_2 norm between spectral magnitudes, denoted as:

$$C_{\text{filter}}(\mathbf{y}, H_\phi(\hat{\mathbf{x}}_0)) = \|\mathbf{W}(|\mathcal{F}(\mathbf{y})| - |\mathcal{F}(H_\phi(\hat{\mathbf{x}}_0))|)\|_2^2, \tag{13}$$

Algorithm 1 Inference phase of the BABE-2 method

Require: observations $\mathbf{y}$
$\tilde{\mathbf{y}} = \mathcal{F}^{-1}(H_{\text{LTAS}}^{-1} \odot \mathcal{F}(\mathbf{y}))$ ▷ LTAS-based init. (Sec. 3.4)
if use LTAS-EQ in objective **then**
    $\mathbf{y} \leftarrow \tilde{\mathbf{y}}$
end if
Sample $\mathbf{x}_T \sim \mathcal{N}(\tilde{\mathbf{y}}, \sigma_{\text{start}}^2\mathbf{I})$ ▷ apply warm initialization
Initialize ϕ_T ▷ initialize the filter parameters
for $i \leftarrow T, \ldots, 1$ **do**   ▷ discrete step backwards
    $\hat{\mathbf{x}}_0 \leftarrow D_\theta(\mathbf{x}_i, \sigma_i)$ ▷ evaluate denoiser
    $\phi_i^0 \leftarrow \phi_{i+1}$ ▷ use the filter from last step
    for $j \leftarrow 0, \ldots, M_{\text{max its.}}$ **do**   ▷ filter optimization
       $\hat{\mathbf{y}}_{\phi_i^j} \leftarrow \mathcal{F}^{-1}(H_{\phi_i^j} \odot \mathcal{F}(\hat{\mathbf{x}}_0))$ ▷ apply filter
       $C_{\text{filter}}(\phi_i^j) \leftarrow C_{\text{rec}}(H_{\phi_i^j}(\hat{\mathbf{x}}_0), \mathbf{y}) + \gamma_{\text{BCR}}C_{\text{BCR}}(\phi_i^j)$
       $\phi_i^{j+1} \leftarrow \phi_i^j - \text{AdamGrad}(C_{\text{filter}}(\phi_i^j))$ ▷ optim. step
       $\phi_i^{j+1} \leftarrow \text{project}(\phi_i^{j+1})$ ▷ project the filter params.
    end for
    $\phi_i \leftarrow \phi_i^M$
    $\hat{\mathbf{y}}_{\phi_i} \leftarrow \mathcal{F}^{-1}(H_{\phi_i} \odot \mathcal{F}(\hat{\mathbf{x}}_0))$ ▷ apply filter
    $\bar{\mathbf{y}} = \mathbf{y} + \gamma\epsilon$ ▷ noise regularization (Sec. 3.3)
    $\mathbf{g}_i \leftarrow -\xi(\sigma_i) \nabla_{\mathbf{x}_i} C_{\text{audio}}(\bar{\mathbf{y}}, \hat{\mathbf{y}}_{\phi_i})$ ▷ rec. guidance
    $\mathbf{s}_i \leftarrow \frac{\hat{\mathbf{x}}_0 - \mathbf{x}_i}{\sigma_i^2}$ ▷ prior score
    $\mathbf{x}_{i-1} \leftarrow \mathbf{x}_i - \sigma_i(\sigma_{i-1} - \sigma_i)(\mathbf{s}_i + \mathbf{g}_i)$ ▷ update step
end for
return $\mathbf{x}_0$ ▷ reconstructed audio signal

---

where the matrix $\mathbf{W}$ denotes a frequency-dependent weighting function. Further implementation details are omitted from the main paper, but they are given in the Appendix A [2] and in the source code [3].

4. EXPERIMENTS AND EVALUATION

We experiment with the restoration of solo piano and singing recording. The conducted objective evaluation compares the performance of BABE-2 against comprehensive baselines for enhancing the quality of historical 78-RPM gramophone recordings.

4.1. Piano Recordings Evaluation

Here we assess the effectiveness of our proposed method for restoring historical piano recordings. For ease of comparison, we used the same diffusion model as the one used in BABE [7]. This model is based on the CQT-Diff+ architecture [16], and was trained on the MAESTRO dataset [17], resampled at 22 kHz.

The test set comprises 54 gramophone recordings of solo piano performances from 1910 to 1930, all of them collected from "the Internet Archive"[4]. Each recording was trimmed to a duration of 1 min. These recordings were preprocessed using our denoiser [5] to eliminate additive noise disturbances, such as clicks, hisses, and pops. The processed evaluation set is made available as part of the supplementary materials[5].

---

[2] The appendix is available in the pre-print: (link to arxiv here)
[3] github.com/eloimoliner/BABE2
[4] archive.org/details/78rpm
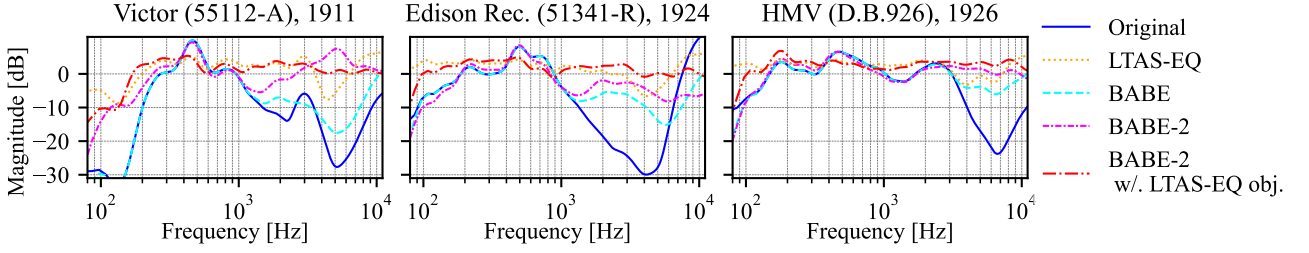[5] research.spa.aalto.fi/publications/papers/dafx-babe2/

Figure 2: *Comparative LTAS analysis of original and restored piano recordings using different methods.*

The objective evaluation is based on the Fréchet Audio Distance (FAD), a reference-free metric comparing the statistics between two sets of embeddings, as detailed by Kilgour et al. [18]. The efficacy of FAD is influenced by the specific embeddings employed and the chosen reference set [19]. To provide a thorough analysis, we computed FAD across various embeddings utilizing the 'fadtk' library [19]. We began with the VGGish classifier' embeddings, widely used in this context [18], and additionally assessed the model using CLAP, a joint audio–text representation, [20], and Encodec, designed for low-rate audio codecs [21]. As a reference test set, we used the MAESTRO dataset test and validation splits from the year 2018.

Furthermore, we examined the distance between the LTAS of the restored recordings and the LTAS of the reference test set. We acknowledge that a distance based on LTAS is not a definitive metric for the same reasons as outlined in Sec. 3.4. However, we include it here as it can shed light on significant differences in coloration with respect to the reference set. The LTAS distance is measured as follows:

$$\text{LTAS dist.} = 10\log_{10}\left(\frac{1}{K}\sum_{f}\frac{|\tilde{\mathbf{X}}_{\text{LTAS}} - \tilde{\mathbf{R}}_{\text{LTAS}}|}{\tilde{\mathbf{R}}_{\text{LTAS}}}\right), \quad (14)$$

where K denotes the number of frequency bins, and $\tilde{\mathbf{X}}_{\text{LTAS}}$ and $\tilde{\mathbf{R}}_{\text{LTAS}}$ represent the LTAS of the restored recordings and the reference test set, respectively.

The results are presented at the top of Table 1. The compared versions of BABE-2 consistently outperform the original recordings, the LTAS-EQ and BEHM-GAN baselines [22], and the predecessor BABE [7]. An ablation study highlights the critical role of noise regularization for better performance. Additionally, the results show that employing LTAS-based initialization strategies improves the FAD scores.

To delve deeper into the impact of the restoration methods on different recordings, Fig. 2 illustrates the frequency-response estimates derived from LTAS analysis. This figure includes three plots corresponding to distinct solo piano recordings, each labeled with its respective year and identifier, depicting the magnitude in dB across a logarithmically spaced frequency axis. The blue lines represent the frequency-response estimates for the original recordings, calculated based on the ratio between the LTAS of each recording and that of the reference test set, following Eq. (9). The remaining lines in each plot denote the LTAS ratio for each restoration method, as detailed in the legend.

Ideally, in Fig. 2, an equalized recording should exhibit a relatively flat frequency response, centered around 0 dB. It can be observed that the original recordings' responses possess an irreg-

Table 1: *Objective evaluation on historical recordings. The best results of each column for each experiment are bolded.*

Experiment, method	VGGish	FAD ↓ CLAP	Encodec	LTAS dist. ↓
Piano: Original	2.37	0.33	8.11	-1.15 dB
LTAS-EQ	3.33	0.19	8.12	-1.77 dB
BEHM-GAN	1.82	0.21	6.65	-1.37 dB
BABE	1.50	0.15	7.72	-2.68 dB
BABE-2	1.45	**0.12**	4.65	-3.02 dB
BABE-2 w/o noise reg.	1.46	0.15	7.25	-2.37 dB
BABE-2 w/ LTAS-EQ init.	1.50	**0.12**	**4.56**	**-3.42 dB**
BABE-2 w/ LTAS-EQ obj.	**1.20**	**0.12**	5.21	-2.57 dB
Vocals (Caruso): Original	19.34	0.47	31.01	-0.29 dB
LTAS-EQ	19.59	0.52	23.41	**-1.96 dB**
BABE	14.48	0.31	26.10	-1.06 dB
BABE-2	11.36	0.28	21.08	-1.65 dB
BABE-2 w/ LTAS-EQ init.	11.32	0.29	20.53	-1.84 dB
BABE-2 w/ LTAS-EQ obj.	**11.11**	**0.27**	**15.04**	-1.95 dB
Vocals (Melba): Original	11.56	0.63	41.34	-0.94 dB
LTAS-EQ	11.99	0.66	37.04	**-2.39 dB**
BABE	5.33	0.36	32.68	-0.88 dB
BABE-2	3.71	0.31	27.78	-0.84 dB
BABE-2 w/ LTAS-EQ init.	**3.44**	**0.30**	**22.12**	-1.56 dB
BABE-2 w/ LTAS-EQ obj.	3.58	0.32	23.64	-2.23 dB

ular shape, characterized by deficiencies in both high and low-frequency energy and marked coloration, i.e. fluctuations, in the mid-frequency range. The "LTAS-EQ" method manages to mitigate the coloration to some extent but, in doing so, it amplifies background noise and distortion artifacts. This harms the FAD scores, as evidenced in Table 1. The "BABE" method, indicated by the cyan dashed line, increases the high-frequency energy, yet not enough to align with the target equalization curve. In contrast, the "BABE-2" method achieves a more balanced equalization, particularly effective when combined with the LTAS-EQ objective. This analysis showcases the superiority of the BABE-2 method.

4.2. Evaluation of Singing Voice Recordings

We evaluate our restoration method to recordings from two iconic singers: Enrico Caruso (tenor) and Nellie Melba (soprano). For this purpose, we pretrained a diffusion model with a collection of studio quality modern singing voice recordings from different sources[6], all of them sampled at 44.1 kHz. This model was later fine-tuned for each target singer in particular, using smaller refer-

---

[6]See Appendix A for more information regarding the training data.
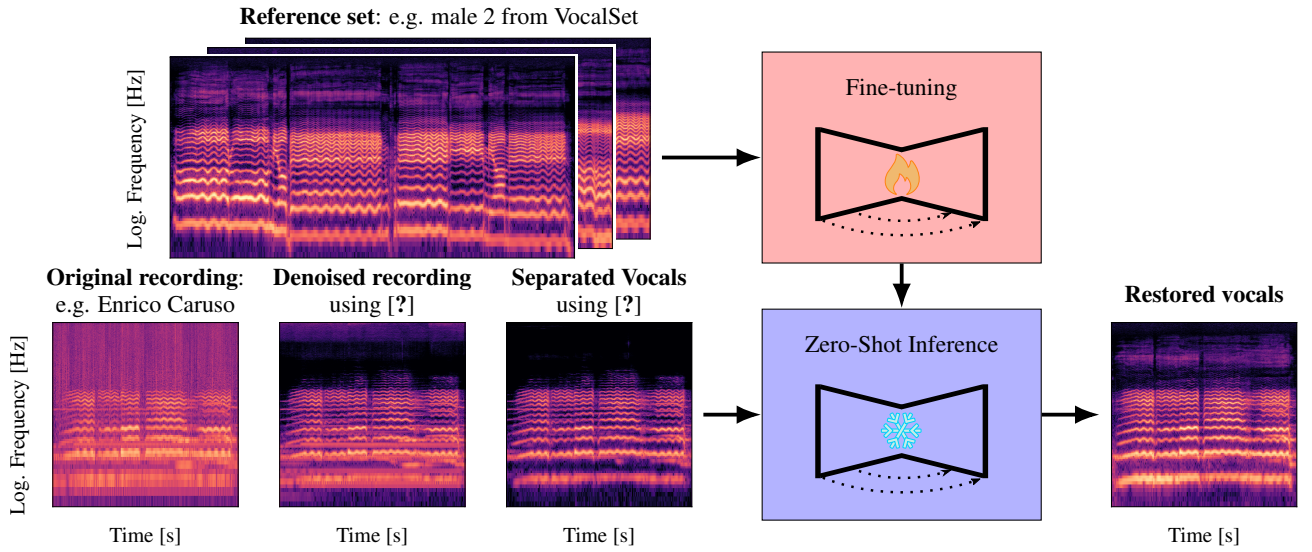
Figure 3: *Singing voice restoration pipeline.*

ence sets extracted from selected singers from Vocalset [23], as we elaborate in Sec. 5. We opted for standard fine-tuning, resuming the training from the pretrained weights.

The process for restoring singing voice recordings is shown in Fig. 3. Similar to the approach detailed in Sec. 4.1, all recordings were trimmed to 1 min and denoised [5] as an initial step. Our evaluation is exclusively aimed at enhancing vocal tracks, and hence, the instrumental parts were separated using HT-Demucs [24]. While this is effective, the source separation method is not fully optimized for historical recordings and tends to introduce noticeable artifacts. These artifacts, however, are outside the scope of this study, as we are not focusing on source separation.

We collected two evaluation sets consisting of 32 recordings from Enrico Caruso and 25 from Nellie Melba. The objective evaluation mirrors the one conducted for piano music, employing the same metrics for assessment. The outcomes for both Caruso and Melba are detailed in Table 1. For the Caruso recordings, the BABE-2 version utilizing the LTAS-EQ objective achieved the most favorable FAD scores. On the other hand, for Melba's recordings, the BABE-2 version with LTAS-EQ initialization showed slightly superior performance. It is noteworthy that, despite recording significantly lower FAD scores, the LTAS-EQ baseline registered the smallest LTAS distance.

5. DISCUSSION: RESTORING HISTORICAL VOICES

One of the key contributions from our research lies in our efforts to restore early-20th-century recordings of classical opera singers. These recordings are often so much degraded that some of the unique vocal characteristics of the singer are no longer preserved. The role of BABE-2 is to reconstruct, or hallucinate, the missing pieces, based on statistical patterns derived from the training data. This underscores the critical importance of selecting appropriate training data, as it significantly impacts the accuracy and authenticity of the audio restoration, ensuring the reconstructed voices closely mirror the original timbres and nuances.

Our methodology involves the strategic fine-tuning of diffusion models, which have been pre-trained on a broad spectrum of singing voice recordings, with a more focused dataset of high-quality recordings from a unique and carefully selected reference singer. The reference singers are chosen for their vocal qualities, which closely resemble those of the historical figures we seek to restore. For this purpose, we employ VocalSet [23], a dataset comprising studio-quality a cappella recordings from 20 professional singers. Each singer contributed with approximately 30 min of data, encompassing a diverse array of vocal techniques.

5.1. Searching for Reference Voices

The task of identifying voices that resemble a particular opera singer is not trivial. Opera singers can be divided into different voice categories, "fach", based on, for example, vocal range and timbre, permeability, or flexibility of their voices [25]. To get the full impression of the singer's voice characteristics on a historical recording, it is necessary not only to listen but also to consult the writings of contemporaries and to study the singer's repertoire. Enrico Caruso (1871-1921) was known for his distinctive voice and his interpretations of tenor roles in the Italian opera repertoire for example Cavaradossi from Puccini's *Tosca* or Canio from *Pagliacci* by Leoncavallo [26]. Both roles require stamina and permeability in low and high register, and Caruso's voice has been described as having a baritonal quality in the low and brilliance of the tenor voice in the higher register. Caruso's recording career spanned 20 years and during that time his technique stabilized, and his voice matured and gained depth [27, 28]. Therefore, we only seek an average estimate of his voice.

We found that the voice from VocalSet's *Male-2* possesses a similar warmth and depth in the low and middle registers as Caruso, together with the power and clarity in the higher register, hence we used it for fine-tuning Caruso's recordings. The famous aria "Vesti la giubba" from Pagliacci, included among the audio examples on our webpage, is a perfect example of Caruso's voice but also of his dramatic abilities. On the webpage, you also find a version of the same aria sung by Beniamino Gigli (1890-1957), who was considered a legitimate heir of Caruso and their voices have similarities, but Gigli's voice sounds more metallic [29]. Gigli's restored example was produced using the dramatic
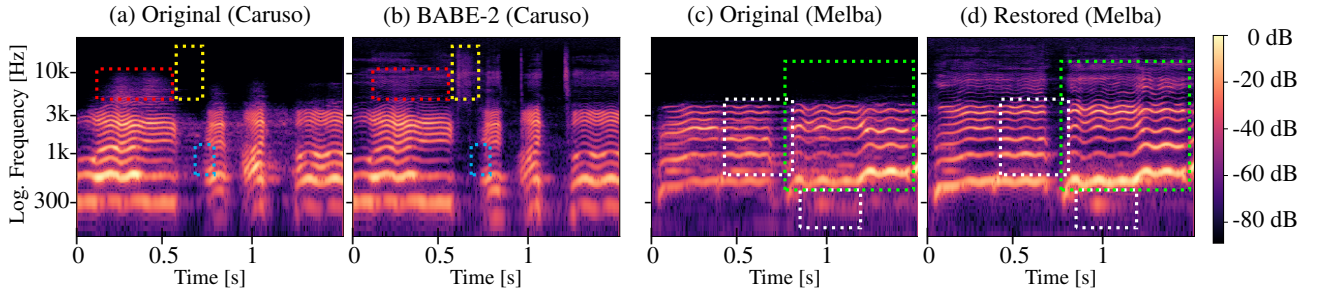
Figure 4: *Spectrogram representations of two vocal restoration examples. The colored boxes highlight key points discussed in Sec. 5.2.*

voice from VocalSets's *Male-11*.

Nellie Melba (1861–1931) was complimented on the beauty of her voice and her flawless technique with equalized scale from top to bottom. Her colleague John McCormack described her voice as a lyric soprano with a beautiful tone and her technique as perfection. She had the ability to sing coloraturas with bird-like quality, and her voice was light and agile [30]. The beauty of voice is a subjective matter, but we found a similar light lyric quality in the voice from VocalSet's *Female-1*, which we used for fine-tuning. Like Caruso, Melba had a long recording career, and she made a total of 150 recordings between 1904 and 1926, including a varied repertoire, but focusing mainly on opera [31]. One of Melba's most celebrated roles was Marguerite from Gounod's *Faust*. As a representative example, we selected Marguerite's aria "Ah! Je ris je me voir si belle", a.k.a. "Jewel aria", from *Faust*.

We additionally restored a version of the same aria from the soprano Adelina Patti (1843–1919), who sang the same repertoire as Melba, and their voices had similarities [32]. Still, we used a different voice, *Female-5*, for fine-tuning because, in this recording, the impression of her voice is darker and lower compared to Melba's. The main reason for this difference might be that, at the time of the recording, Patti was already over 60 years old [32].

Despite our reliance on the written literature, the main strategy for finding pairs between reference and original singers was analytic listening. With the limited bandwidth and the disturbing coloration of the original medium, some imagination was required when figuring out the original color of the voice. Consequently, evaluating the fidelity of our results posed challenges, leading us to focus on a qualitative analysis.

5.2. Qualitative Analysis

Figure 4 shows spectrograms of two restoration examples: a 1.5-second excerpt from Enrico Caruso's "Vesti la Giubba" (Victor 6001-A, 1910) in (a) and (b), and a same-length excerpt from Nellie Melba's "Jewel aria" (Victor 88066, 1910) in (c) and (d). Both recordings show a bandwidth limit of approximately 4 kHz. In both cases, the original recording effective bandwidth barely exceeds 4 kHz. It can be observed that the limited bandwidth affects more critically the soprano voice than the tenor, as the former is at least an octave higher. Therefore, the restored voice of Caruso is recognisable and the hallucinated overtones create a warmer and fuller sound. In Melba's case, the result is more controversial. Even if the overtones are clearly audible, the voice sounds slightly breathy and lacks its ringing. The more overtones are generated, the more of the singer's personal color can be heard from the examples yielding a more natural-sounding result, although, in this case, part of the color has been hallucinated by the generative model. In

Melba's case, the generated timbre is slightly different from what we can hear on the recordings.

Thanks to the extensive range of vocal techniques included by VocalSet, BABE-2 demonstrates a notable capacity for adapting to the distinct formant characteristics of various vowels, as illustrated in the green-colored box in Figs. 4(c, d). However, it is observed that subtle portamentos and singer legato lines are sometimes truncated, as shown in the cyan-colored box in Fig. 4(a, b). This behavior is expected as soft (low-energy) singing has a much lower weight than louder passages in the reconstruction cost. Also, further research is needed on how to better identify the vowel modification of the soprano voice in particular, i.e., the rounding of vowels at high pitches.

One limitation of the proposed method is that it is not designed to tackle nonlinear degradations, such as harmonic distortion. Thus, one would expect that the model would fail when it encounters distortion artifacts. However, this does not seem to always be the case. The red box in Fig. 4(a, b) highlights an example where Caruso's loud singing provoked a significant amount of harmonic distortion in the recording. In this example, BABE-2 suppresses the distortion and regenerates a clean high-frequency spectrum. We attribute this behavior to the effect of the noise regularization implemented in our method and introduced in Sec. 3.3.

Due to scale limitations, BABE-2 does not have language-level control of the singing voice distribution. Since the frequency range of certain consonants, such as fricatives, is outside the range captured in the original recording, BABE-2 hallucinates consonants at statistically plausible locations, but these may not correspond to the lyrics. This is the case of the incorrectly placed fricative in the yellow box in Fig. 4(a, b).

The white boxes in Fig. 4(c, d) show some of the consequences of the source separation preprocessing. The bleeding from the instrumental track is present in the restored recording, especially in the low-frequency region. Nevertheless, we also notice that a side-effect of BABE-2 is that it additionally suppresses some of the bleeding between the harmonics, producing an even cleaner vocal separation. This behavior is expected, as these artifacts are not included in the training data.

6. CONCLUSIONS

This study introduces BABE-2, enhancing the restoration of historical music recordings through a generative equalization strategy. Objective evaluations demonstrate its superiority over the prior version, particularly in enhancing recordings of iconic vocalists Enrico Caruso and Nellie Melba. In addition to music enthusiasts, BABE-2 allows researchers to get a more accurate auditory picture of specific features of a recording or interpretation. Although cer-

tain limitations are identified, indicating areas for future research, this work represents a notable advancement in the domain of historical music restoration. BABE-2 demonstrates unprecedented potential to revitalize aged audio recordings, achieving levels of clarity and fidelity previously unattainable.

7. ACKNOWLEDGMENTS

8. REFERENCES

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

[2] M. Kob and T. A. Weege, "How to interpret early recordings? Artefacts and resonances in recording and reproduction of singing voices," *Computational Phonogram Archiving*, pp. 335–350, 2019.

[3] V. Välimäki, S. González, O. Kimmelma, and J. Parviainen, "Digital audio antiquing—Signal processing methods for imitating the sound quality of historical recordings," *J. Audio Eng. Soc.*, vol. 56, no. 3, pp. 115–139, Mar. 2008.

[4] T. G. Stockham, T. M. Cannon, and R. B. Ingebretsen, "Blind deconvolution through digital signal processing," *Proc. IEEE*, vol. 63, no. 4, pp. 678–692, 1975.

[5] 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. (ICASSP)*, Singapore, May 2022, pp. 841–845.

[6] J.-M. Lemercier, J. Richter, S. Welker, E. Moliner, V. Välimäki, and T. Gerkmann, "Diffusion models for audio restoration," *arXiv preprint arXiv:2402.09821*, 2024.

[7] E. Moliner, F. Elvander, and V. Välimäki, "Blind audio bandwidth extension: A diffusion-based zero-shot approach," *arXiv*, 2024.

[8] 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.

[9] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2022.

[10] E. Moliner, J. Lehtinen, and V. Välimäki, "Solving audio inverse problems with a diffusion model," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Rhodes, Greece, Jun. 2023.

[11] H. Chung, J. Kim, M. T. Mccann, et al., "Diffusion posterior sampling for general noisy inverse problems," in *Proc. Int. Conf. Learning Representations (ICLR)*, Kigali, Rwanda, May 2023.

[12] H. Chung, J. Kim, S. Kim, and J. C. Ye, "Parallel diffusion models of operator and image for blind inverse problems," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 6059–6069.

[13] S. Sadat, J. Buhmann, D. Bradely, O. Hilliges, and R.M Weber, "CADS: Unleashing the diversity of diffusion models through condition-annealed sampling," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, May 2015.

[15] A. Wright and V. Välimäki, "Perceptual loss function for neural modeling of audio systems," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 251–255.

[16] E. Moliner and V. Välimäki, "Diffusion-based audio inpainting," *J. Audio Eng. Soc.*, vol. 72, Mar. 2024.

[17] C. Hawthorne, A. Stasyuk, A. Roberts, et al., "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proc. Int. Conf. Learning Representations (ICLR)*, May 2019.

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

[19] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou, "Adapting Frechet audio distance for generative music evaluation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024.

[20] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pre-training with feature fusion and keyword-to-caption augmentation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023.

[21] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023.

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

[23] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, "Vocalset: A singing voice dataset," in *ISMIR*, 2018, pp. 468–474.

[24] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023.

[25] J.B Steane, "Fach," *Groove Music Online*, 2002.

[26] J. Freestone, *Enrico Caruso: His Recorded Legacy*, TS Denison, Minneapolis, 1961.

[27] B. Gentili, "The birth of 'modern' vocalism: The paradigmatic case of Enrico Caruso," *J. Royal Musical Assoc.*, vol. 146, no. 2, pp. 425–453, 2021.

[28] R. Celletti and A. Blyth, "Caruso, Enrico," *Groove Music Online*, 2013.

[29] D. Shawe-Taylor and A. Blyth, "Gigli, Beniamino," *Groove Music Online*, 2001.

[30] L. A. G. Strong, "John McCormack: The story of a singer," *The Macmillan Company*, 1941.

[31] D. Shawe-Taylor, "Melba, Dame Nellie," *Groove Music Online*, May 2009.

[32] E. Forbes, "Patti, Adelina," *Groove Music Online*, 2001.

[33] R. Huang, F. Chen, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3945–3954.

[34] Y. Wang, X. Wang, P. Zhu, J. Wu, H. Li, H. Xue, Y. Zhang, L. Xie, and M. Bi, "Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis," in *Interspeech*, 2022.

[35] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013, pp. 1–9.

[36] J. Koguchi, S. Takamichi, and M. Morise, "PJS: Phoneme-balanced japanese singing-voice corpus," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2020, pp. 487–491.

[37] L. Zhang, R. Li, S. Wang, L. Deng, J. Liu, Y. Ren, J. He, R. Huang, J. Zhu, X. Chen, et al., "M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus," *Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets and Benchmarks Track*, vol. 35, pp. 6914–6926, 2022.

[38] S. Choi, W. Kim, S. Park, S. Yong, and J. Nam, "Children's song dataset for singing voice research," in *Proc. ISMIR*, 2020, vol. 4.

[39] Lingxiao Yang, Shutong Ding, Yifan Cai, Jingyi Yu, Jingya Wang, and Ye Shi, "Guidance with spherical gaussian constraint for conditional diffusion," *arXiv preprint arXiv:2402.03201*, 2024.

[40] 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.

A. SUPPLEMENTARY ALGORITHM SPECIFICATIONS

This appendix includes additional specifications and implementation insights regarding the BABE-2 algorithm. Many aspects covered here were omitted from the main paper either because they build upon established work or due to their detailed, technical nature.

A.1. Training

The training objective is based on the one expressed in Eq. 3. However, we apply the preconditioning strategy proposed by Karras et al. [9], which aimed to improve the diffusion model training dynamics. The denoiser in Eq. 3 is reparameterized as:

$$D_\theta(\boldsymbol{x}_\tau, \tau) = c_{\text{skip}}(\tau)\boldsymbol{x}_\tau + c_{\text{out}}(\tau)F_\theta(c_{\text{in}}(\tau)\boldsymbol{x}_\tau, \frac{\ln(\sigma(\tau))}{4}), \quad (15)$$

where c_{in}, c_{skip} and c_{out} are time-dependent preconditioning parameters defined in [9], and F_θ corresponds to the optimizable neural

Table 2: Hyperparameters chosen in our experiments

Training params.	Piano	S. Voice (pre)	S. Voice (f-t)
Learning rate	2×10^{-4}	2×10^{-5}	1×10^{-4}
Batch size	4	4	4
Sampling rate	22.05 kHz	44.1 kHz	44.1kHz
Segment length	8.35 s	5.94 s	5.94 s
EMA rate	0.9999	0.9999	0.9999
σ_{data}	6.3×10^{-2}	1	1
Training its.	850k	325k	8k

network architecture, as detailed in Sec. A.2. The loss weighting term $\lambda(\tau)$ in Eq. 3 is defined as: $\lambda(\tau) = 1/c_{\text{out}}(\tau)^2$, also following the guidelines from [9].

The diffusion model employed for piano restoration was trained using the MAESTRO dataset [17]. For the singing voice experiments, we pretrained a new diffusion model with a collection of studio quality modern singing voice recordings from different sources [33, 34, 35, 36, 37, 38]. This pre-training phase spanned 325k iterations over three days on a single NVIDIA A100-80GB GPU. The singing voice models were later fine-tuned with a smaller reference set extracted from selected singers from Vocalset [23], as we elaborate on in Sec. 5. We employed standard fine-tuning, resuming the training from the pre-trained weights. The fine-tuning process involved 8k training iterations, requiring about two hours on the same GPU.

The preconditioning parameters used in Eq. 15 depend on the average power of the training data σ_{data}^2. For the MAESTRO dataset, this value was estimated as $\sigma_{\text{data}} = 6.3 \times 10^{-2}$. For the singing voice experiments in particular, as we were working with heterogeneous datasets, we decided to normalize the training data samples. During the pre-training stage, we apply Root-Mean-Square (RMS) normalization, thus resulting with the parameter $\sigma_{\text{data}} = 1$. For fine-tuning, in order to preserve the volume-dependent characteristics of singing voice, we estimated the average RMS value from a representative set of the fine-tuning dataset, in our case VocalSet, and normalize all the training samples with it. For this purpose, we estimated the value $\sigma_{\text{norm}} = 0.04$.

Additionally, as often employed for training diffusion models, we track an exponential moving average (EMA) of the neural network weights during training. The EMA weights are later used during the inference stage. Some of the most important parameters regarding training are summarized in Table 2.

A.2. Neural Network Architecture

In this study, we employed the CQT-Diff+ architecture [16] as the backbone for our diffusion model, denoted as F_θ. This architecture incorporates optimizable layers with an invertible Constant-Q Transform (CQT), characterized by varying time resolutions across octave bands. The architecture adopts a U-Net-style encoder-decoder structure with skip connections, primarily consisting of 2D-Convolutional layers that process along both time and frequency axes.

For our experiments involving piano music and singing voice, we utilized distinct parameter configurations, yet both setups comprise approximately 40 million trainable parameters. Specifically, the piano music configuration aligns with that employed in BABE [7], utilizing a CQT that spans 7 octaves with 64 frequency bins

per octave. In contrast, the singing voice experiments required adaptation to a higher sampling rate, resulting in a CQT covering 8 octaves but with 32 frequency bins per octave. This adjustment in frequency resolution, particularly for singing voice, proved beneficial for generating a more coherent spectrum in the produced audio, a critical aspect for tasks like equalization. Furthermore, an important consideration in both experimental configurations is the exclusion of frequencies below 86 Hz, achieved through high-pass filtering. While this decision prevents the model from restoring content within this low-frequency range, it is not deemed problematic for our purposes. This is because the musical signals we aim to process typically do not contain significant information in such low frequencies.

A.3. Inference

The inference algorithm was designed to share the same structure as BABE [7], but including BABE-2's new contributions detailed in the main paper. Some of the most relevant hyperparameters employed for inference are summarized in Table 3.

We use the exponentially-warped discretization from [9]:

$$\tau_{i<T} = \left(\sigma_{\text{start}}^{\frac{1}{\rho}} + \frac{i}{T-1} \left(\sigma_{\text{min}}^{\frac{1}{\rho}} - \sigma_{\text{start}}^{\frac{1}{\rho}} \right) \right)^{\rho}, \quad (16)$$

where the hyperparameters σ_{start}, σ_{min}, T and ρ define the time schedule. Intuitively, the starting noise level σ_{start} should be large enough to mask all the imperfections from the audio signal used at initialization, and σ_{min} should be negligible when compared with the audio signal magnitude. The role of the parameter ρ is to warp the time discretization and concentrate more steps towards lower noise levels. The parameter T defines the number of discretized steps. With it, one can effectively trade-off audio quality per inference speed. While using a larger T can potentially lead to better results, we decied to fix $T = 51$ throughout all our experiments. It must be noted that, since we employ a second-order sampler, the number of neural function evaluations required during inference is twice the value of T.

Following [10, 7], we normalize the likelihood score step size $\xi(\tau)$ as follows to enhance robustness:

$$\xi(\tau) = \xi' \sqrt{N} / (\sigma(\tau) \|\nabla_{\boldsymbol{x}_\tau} \| \boldsymbol{y} - \mathbf{m} \odot \hat{\boldsymbol{x}}_0 \|^2 \|^2), \quad (17)$$

a technique also recognized as a spherical Gaussian constraint in recent literature [39]. The parameter ξ' plays a crucial role by balancing audio quality and fidelity to observations. Higher ξ' values aim to minimize the cost function effectively, although excessively large values may lead to artifacts in the audio. We observed that introducing stochasticity in the inference phase allows enables the application of larger ξ' values while maintaining audio quality. This stochasticity is introduced through the S_{churn} parameter of the sampler, and the noise regularization from Sec. 3.3.

The filter parameters are optimized through stochastic gradient descent using the *Adam* optimizer. The learning rate was searched qualitatively by trial and error, and was defined to a value of 10. We adopted the default momentum parameters from the *PyTorch*'s implementation. Optimizing the filter parameters is computationally less demanding than updating audio signals, as it avoids the need for neural functions evaluations. Consequently, we employed a relatively high number of filter optimization iterations for each sampling step, specifically 100.

Table 3: Hyperparameters chosen in our experiments

Inference params.	Piano	Singing Voice
Start noise level σ_{start}	0.5	10
Min. noise level σ_{min}	4×10^{-5}	1×10^{-3}
Time warping ρ	13	13
Discretization steps T	51	51
Sampling stochasticity S_{churn}	10	10
Step size ξ'	1.0	0.5
Noise regularization γ	0.25	1
Filter learning rate	10	10
Num. optimization its. per step	100	100
BCR decay rate β	0.1	0.1
BCR weight γ_{BCR}	10	10

A.4. Filter Parameterization

The proposed filter parameterization is detailed in Sec. 3.1. For completeness, we report its mathematical formulation, which is as follows:

$$H_\phi(f)[\text{dB}] = \begin{cases} A_{\text{lim-}} \log_2 \frac{f}{f_{-S'}} + \sum_{i=-S'}^{-1} A_i & f < f_{-S'} \\ \vdots & \vdots \\ A_{-2} \log_2 \frac{f}{f_{-1}} + A_{-1} + A_{-2} & f_{-2} \le f < f_{-1} \\ A_{-1} \log_2 \frac{f}{f_0} & f_{-1} \le f < f_0 \\ A_1 \log_2 \frac{f}{f_0} & f_0 \le f < f_1 \\ A_2 \log_2 \frac{f}{f_1} + A_1 & f_1 \le f < f_2 \\ \vdots & \vdots \\ A_{\text{lim+}} \log_2 \frac{f}{f_S} + \sum_{i=1}^{S} A_i & f_S \le f, \end{cases}$$

$$(18)$$

where f_i (Hz) represent cutoff frequencies and A_i (dB) are the decay slopes. The set of adjustable parameters is defined as:

$$\phi = \{ f_i, A_i \mid i = 1, \ldots, S \}, \quad (19)$$

where S is the number of breakpoints.

For the reasons specified in Sec. 3.1, we use a reduced number of $S = 5$ stages in our experiments. This comprises a total of 9 optimizable parameters. We initialized the 5 breakpoints to, and the rest to all: $f_{-2} = 50 \, \text{Hz}$, $f_{-1} = 500 \, \text{Hz}$, $f_0 = 1 \, \text{kHz}$, $f_1 = 1.5 \, \text{kHz}$, and $f_2 = 2 \, \text{kHz}$, while all the slopes are initialized to 0 dB. Considering the the $A_{\text{lim+}}$ slope is fixed at -80 dB, this initialization produces a flat magnitude response with a steep band-limit at 2 kHz, a relatively low cut-off frequency. The initialization plays a relevant role in the performance of the method. However we notice that, in this case, the filter initialization influence is not as critical as in the prior version, BABE. We attribute this observed improvement in robustness to some of the incorporated additions, specifically the noise regularization and the Adam optimizer.

A.5. Block-Autoregressive Inference

Due to memory constraints, the diffusion models employed in this work are designed to process short audio segments within the range of a few seconds. However, in the context of audio restoration, it is necessary to be able to restore entire recordings which may last several minutes long.

Adapting diffusion models to function on a frame-by-frame basis while ensuring coherence between subsequent frames is a known trick in the field [40]. One can just enforce consistency with the last fragment of the previous frame by incorporating an "outpainting" objective on top of the reconstruction. Given the last fragment of the previous restored frame repositioned at the start of the current one $\mathbf{x}_{\text{prev}}$, the proposed algorithm can be adapted to a block-autoregressive inference scenario by substituting the reconstruction cost function in Eq. (12) for:

$$
\begin{aligned}
C_{\text{audio}}^{\text{B-AR}}(\mathbf{y}, \mathbf{x}_{\text{prev}}, H_\phi(\hat{\mathbf{x}}_0)) = \| \mathbb{1}_{[0,t_{\text{ov.}}]} \odot (\mathbf{x}_{\text{prev}} - \hat{\mathbf{x}}_0) \\
+ \mathbb{1}_{[t_{\text{ov.}},t_{\text{end}}]} \odot (\mathbf{y} - H_\phi(\hat{\mathbf{x}}_0)) \|_2^2, \quad (20)
\end{aligned}
$$

where $\mathbb{1}_{[t_{\text{start}},t_{\text{end}}]}$ represents a binary mask that is equal to 1 in the time interval between t_{start} and t_{end} and 0 otherwise, $t_{\text{ov.}}$ is the overlap time, t_{end} points to the end of the segment, and $\odot$ denotes element-wise multiplication. In our experiments we use an overlap factor $t_{\text{ov.}}/t_{\text{end}}$ of 10%. This trick was already employed in BABE [7], but, although mentioned, the adapted cost function for bandwidth extension was not properly formalized, and neither is anywhere else in the related literature, as far as we are aware.

This adaptation can be equivalently applied to the filter reconstruction cost function in Eq. (**??**), allowing us to estimate the filter parameters with the updated objective. Note however that, if the degradation is considered to be time-invariant, there is no need to re-estimate the filter parameters ϕ among different frames, as only once is needed. Nevertheless, we use the block-autoregressive inference in our experiments for evaluation purposes.