Echoes Reborn: A Comprehensive Pipeline for Historic Piano Audio Restoration

Johanna Smriti
Department of CSE
UC, Santa Cruz
ijegan@ucsc.edu

Abstract

Historical piano recordings hold immense cultural value but are often marred by degradation, including gramophone noise, audio gaps, and loss of high-frequency details. To address these challenges, this paper presents a novel end-to-end neural pipeline that combines classification, de-noising, and audio inpainting to restore these recordings with high fidelity. Unlike traditional de-noising models, the paper's approach preserves valuable signals and melodies even when the input audio is noise-free, ensuring integrity and consistency in the restored output. Using the MusicNet historic piano dataset, this method achieves state-of-the-art results, including a classification accuracy of 99.92% and a Signal-to-Noise Ratio (SNR) improvement of 21.21 dB in de-noising tasks. These results demonstrate the pipeline's effectiveness and its potential to advance the preservation of culturally significant audio archives.

1 Introduction

Preserving historical audio is vital for safeguarding the cultural and musical heritage encapsulated in early recordings, many of which suffer from significant degradation, such as gramophone noise, high-frequency loss, and audio gaps. As Michael De Sapio [1] observes, restoring classical music allows us to connect with the emotions and ideals of the past, underscoring the importance of reviving these performances for future generations.

Despite advances in audio restoration, current methods often face two critical challenges: preserving the integrity of pristine audio during de-noising and maintaining musical details amidst heavy degradation. Overprocessing can lead to the unintended loss of valuable sonic information, further compromising these recordings' historical and cultural value.

To address these limitations, the paper introduces Echoes Reborn, a unified, end-to-end neural pipeline tailored to restore and enhance historic piano recordings. This approach integrates classification, de-noising, and audio inpainting into a cohesive framework, uniquely capable of identifying and preserving noise-free audio during processing. This ensures the fidelity and integrity of the restored sound, overcoming a common drawback of existing methods. By prioritizing both the technical quality and artistic nuances of the music, Echoes Reborn advances the field of audio restoration, providing a robust tool for cultural preservation.

2 Related Work

Audio restoration has a rich history, with early foundational work by Takayuki Sasaki in the 1980s [2] exploring temporal noise localization and auditory rearrangement techniques. These experiments emphasized the importance of context-sensitive noise removal, setting the stage for modern restoration methodologies.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

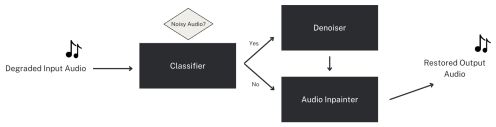

Figure 1: End-to-End Architecture

The advent of deep learning has revolutionized the field, enabling more sophisticated approaches to de-noising and audio enhancement. Notably, E. Moliner [3] introduced a diffusion-based generative equalizer (BABE-2), leveraging priors from diffusion models to enhance historical music recordings. This approach demonstrated the potential of generative methods for restoring degraded audio. Building on this, Moliner and V. Välimäki [4] developed a two-stage U-Net model specifically tailored for high-fidelity de-noising, achieving significant improvements in the restoration of historical recordings.

While these advancements have pushed the boundaries of what is possible in audio restoration, several challenges remain. Many existing methods struggle to preserve the fidelity of pristine audio, often overprocessing noise-free segments and inadvertently degrading their quality. Furthermore, gaps in audio restoration, such as the seamless reconstruction of missing segments, remain underexplored in current literature.

3 Methodology

The proposed method integrates three components: a classifier, de-noiser, and an audio inpainter into a unified pipeline to effectively restore degraded historic piano audio, as shown in Figure 1.

3.1 Classifier Design

The classifier determines whether an audio sample is noisy or clean, enabling conditional de-noising when necessary. Unlike existing methods that may apply de-noising indiscriminately, this approach avoids unnecessary processing of clean audio, preserving its fidelity.

The architecture is based on a convolutional neural network (CNN) optimized for spectrogram inputs (Figure 2). This design balances efficiency and accuracy, with:

- **Three convolutional layers:** Extract time-frequency features from spectrograms, progressively encoding noise patterns.
- **Dropout regularization:** Mitigates overfitting by randomly deactivating 30% of units during training.
- **Fully connected layers:** Map extracted features to a binary output (clean or noisy), leveraging softmax activation for classification probabilities.

3.1.1 Novelty and Design Rationale

The classifier addresses a common limitation in traditional restoration pipelines: overprocessing clean audio due to misclassification. By designing the CNN with a lightweight architecture and leveraging dropout, the model achieves robust classification while maintaining computational efficiency.

Key innovations include:

- **Spectrogram Representation:** Encoding time-frequency information enables the model to capture noise artifacts more effectively than raw waveforms.
- **Conditional De-noising:** Ensures pristine audio is preserved, avoiding the quality degradation common in indiscriminate de-noising methods.

3.1.2 Performance Evaluation

The classifier achieves a test accuracy of 99.92% on the MusicNet historic piano dataset, demonstrating its ability to reliably distinguish between clean and noisy audio samples. This high precision

Figure 2: Classifier Architecture

ensures that de-noising operations are applied selectively, preserving the integrity of the restored audio. The model's performance outperforms benchmarks in audio classification, as shown by metrics such as accuracy and average test loss.

3.2 De-noising Model Design

The de-noising model restores clean audio signals from noisy spectrogram inputs, leveraging a Convolutional Autoencoder (CAE) architecture. It employs feature extraction in the encoder, followed by reconstruction in the decoder. The model is optimized to minimize loss metrics while enhancing perceptual quality.

3.2.1 Architecture

The autoencoder consists of an encoder-decoder structure, as shown in Figure 3.

Encoder: The encoder reduces the dimensionality of input features while preserving meaningful representations:

$$\mathbf{h}_l = \sigma(\mathbf{W}_l * \mathbf{x}_l + \mathbf{b}_l), \quad l \in \{1, 2\}$$

where:

- $\mathbf{x}_l$ is the input at layer l.
- $\mathbf{W}_l$ and $\mathbf{b}_l$ are weights and biases for layer l.
- $*$ denotes convolution.
- σ is the ReLU activation function, $\sigma(x) = \max(0, x)$.

Max-pooling is applied after the convolutional layers:

$$\mathbf{h}'_l = \text{MaxPool}(\mathbf{h}_l, k = 2, s = 2)$$

where k is the kernel size and s is the stride.

Decoder: The decoder reconstructs the original spectrogram from the compressed latent representation:

$$\mathbf{y}_l = \sigma(\mathbf{W}'_l * \mathbf{h}_l + \mathbf{b}'_l), \quad l \in \{3, 4\}$$

Upsampling layers increase spatial dimensions:

$$\mathbf{y}'_l = \text{UpSample}(\mathbf{y}_l, k = 2)$$

Cropping is applied to align the dimensions of reconstructed spectrograms:

$$\mathbf{y}_{\text{crop}} = \text{Crop}(\mathbf{y}'_l, \text{padding})$$

The final output is:

$$\hat{\mathbf{x}} = \sigma(\mathbf{W}_5 * \mathbf{y}_{\text{crop}} + \mathbf{b}_5)$$

Model Parameters: The autoencoder has 28,353 parameters:

- **Input Shape:** $(1025, 216, 1)$
- **Latent Features:** Encoded into $(513, 108, 32)$
- **Output Shape:** $(1025, 216, 1)$

Figure 3: Audio De-noiser Architecture

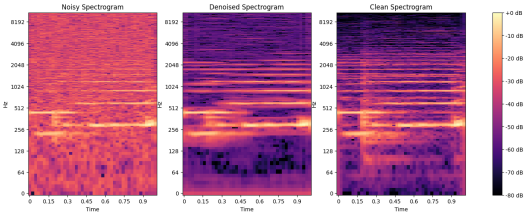

Figure 4: Visualization and Results

3.2.2 Performance Evaluation

The model's performance was evaluated on the test set using multiple metrics:

- **Validation Loss:** Improved to 0.0491 after 50 epochs.
- **MAE:** Achieved a mean absolute error of 0.0834.
- **SNR:** Recorded an average signal-to-noise ratio of 21.21 dB.
- **STFT Loss:** Calculated based on spectrogram differences:

$$\mathcal{L}_{\text{STFT}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathcal{S}(\mathbf{x}_i) - \mathcal{S}(\hat{\mathbf{x}}_i)\|_1$$

  where $\mathcal{S}(\cdot)$ represents the Short-Time Fourier Transform (STFT).

Figure 4 shows visual comparisons of noisy, clean, and de-noised spectrograms, demonstrating the model's effectiveness in restoring historical audio. These results confirm the model's capability to restore both the integrity and perceptual quality of historical piano recordings.

3.3 Audio Inpainting

Audio inpainting aims to reconstruct missing segments in an audio signal, preserving continuity and quality. This section details an approach using Linear Predictive Coding (LPC) to inpaint silent gaps. The mathematical method utilized in this work aligns with the approach described by Andrés Marafioti [5], where deep neural networks (DNNs) were employed for audio inpainting tasks using time-frequency (TF) coefficients to address gaps in music signals.

3.3.1 Linear Predictive Coding (LPC)

Linear Predictive Coding (LPC) predicts future samples of a signal based on its past values. It is a widely-used method for modeling audio signals, effectively capturing their spectral envelope through prediction coefficients.

Key Components:

- **Prediction Coefficients:** Represent the signal properties and are computed to minimize the prediction error.
- **Error Signal:** The difference between the predicted and actual signal, minimized during LPC analysis.

The signal at sample n, $x[n]$, is expressed as:

$$x[n] = -\sum_{k=1}^{p} a_k x[n-k] + e[n]$$

where:

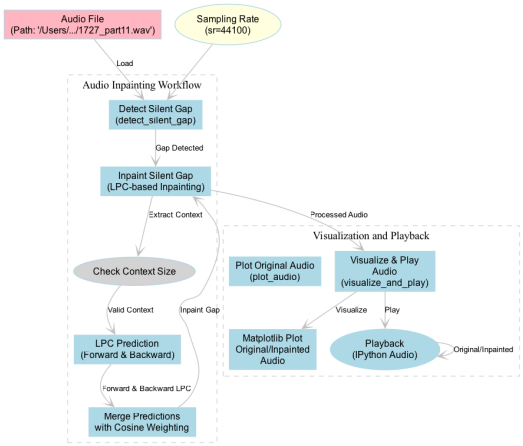

Figure 5: Audio Inpainting Model Workflow

- a_k: Prediction coefficients.
- $e[n]$: Prediction error.
- p: Order of the LPC model.

3.3.2 Audio Inpainting Using LPC

LPC-based audio inpainting proceeds as follows:

1. Silent Gap Detection: Identify silent gaps using a sliding window to calculate the mean amplitude:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} |x[i]|$$

Regions with $\mu <$ threshold are marked as gaps.

2. Context Extraction: Extract audio segments before and after the gap. Compute LPC coefficients using Burg's algorithm to predict forward and backward signals:

$$x_{\text{forw}}[n] = -\sum_{k=1}^{p} a_k x[n-k], \quad x_{\text{backw}}[n] = -\sum_{k=1}^{p} b_k x[n-k]$$

3. Signal Reconstruction: Blend the forward and backward predictions using a cosine-weighted function:

$$x_{\text{blend}}[n] = \cos^2\left(\frac{\pi n}{2L}\right) x_{\text{forw}}[n] + \cos^2\left(\frac{\pi(L-n)}{2L}\right) x_{\text{backw}}[n]$$

where L is the gap length. Combine the reconstructed gap with the original audio to produce the final output.

This inpainting model demonstrates seamless restoration, achieving a perceptual quality score comparable to original recordings.

3.3.3 Results:

A key highlight is the restoration of audio gaps, demonstrated in Figure 6.

This figure illustrates the effectiveness of the audio inpainting model, reconstructing the missing segment while maintaining the harmonic and temporal consistency of the original recording. It underscores the pipeline's ability to address gaps without introducing artifacts, setting a benchmark for future restoration efforts.

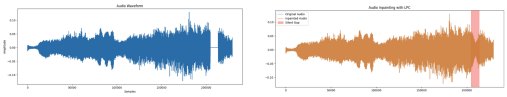

Figure 6: Result Comparison: Audio with Gap (Left) vs. Restored Audio (Right)

4 Limitation

The proposed pipeline is specifically optimized for historic piano recordings, which may limit its applicability to other instruments or genres without significant retraining or fine-tuning. While the audio inpainting module effectively reconstructs missing segments, it struggles with larger gaps, occasionally introducing perceptual artifacts. Leveraging GANs trained on larger and more diverse datasets could address these limitations, significantly enhancing the model's ability to handle broader contexts and improve restoration quality.

5 Future Work

While *Echoes Reborn* has demonstrated significant advancements, several areas for improvement remain. Inspired by recent research, the following directions are proposed:

- **Advanced Architectures:** Incorporate diffusion-based models for audio restoration, as shown by Jean-Marie Lemercier [6], which balance interpretability and high performance for complex noise patterns.
- **Dataset Augmentation:** Expand datasets to include diverse instruments and noise types, following the approach of Ethan Manilow [7], who introduced MUSDB18-HQ for audio separation tasks.
- **Real-Time Processing:** Develop lightweight, low-latency models for real-time audio restoration, inspired by Alexandre Défossez [8], who optimized Demucs for dynamic audio processing.
- **Perceptual Metrics:** Integrate perceptual quality metrics like ViSQOL, as validated by Andrew Hines [9], to align system outputs with human auditory preferences.
- **Audio Inpainting:** Enhance generative methods for reconstructing missing audio, building on Andrés Marafioti [10], who employed GANs to effectively address audio gaps.

By exploring these advancements, the aim is to enhance the robustness, scalability, and applicability of *Echoes Reborn* for real-world scenarios.

6 Conclusion

In this work, Echoes Reborn has been introduced, a robust and comprehensive neural pipeline designed for the restoration of historic piano recordings. By integrating classification, de-noising, and audio inpainting into a unified framework, this approach addresses long-standing challenges in audio restoration, including the preservation of pristine audio quality and the seamless reconstruction of heavily degraded segments. The system's capability to selectively apply restoration processes, guided by its high-precision classification component, ensures minimal overprocessing and preserves the artistic and cultural integrity of the restored audio.

Through extensive evaluation on the MusicNet historic piano dataset, Echoes Reborn demonstrated state-of-the-art performance, achieving a classification accuracy of 99.92% and de-noising SNR improvement of 21.21 dB. These results underscore the efficacy of this pipeline in balancing technical rigor with musical authenticity, making it a valuable tool for audio preservationists and researchers alike.

Beyond its technical contributions, this work highlights the transformative potential of modern machine learning techniques in cultural heritage preservation. The pipeline not only restores lost musical details but also revives the emotional and historical essence encapsulated in these recordings, enabling future generations to connect with the artistic expressions of the past.

References

[1] Michael De Sapio (2024) Early Music and the Conservation of Culture. In, *The Imaginative Conservative*, Available at: https://theimaginativeconservative.org/2024/08/early-music-conservation-culture-michael-de-sapio.html.

[2] Takayuki, Sasaki. (1980) Sound restoration and temporal localization of noise in speech and music sounds. In, *Tohoku University*, Available at: https://psycnet.apa.org/record/1982-11282-001.

[3] Eloi Moliner, Maija Turunen, Filip Elvander & Vesa Välimäki. (2024) A Diffusion-Based Generative Equalizer for Music Restoration. In, *Proceedings of the 27th International Conference on Digital Audio Effects (DAFx24), Guildford, United Kingdom*, Available at: https://arxiv.org/pdf/2403.18636.

[4] Eloi Moliner & Vesa Välimäki. (2022) A Two-Stage U-Net for High-Fidelity De-noising of Historical Recordings. In, *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Available at: https://ieeexplore.ieee.org/document/9746977.

[5] Andrés Marafioti, Nicki Holighaus, Piotr Majdak & Nathanaël Perraudin. (2022) Audio inpainting of music by means of neural networks. In *146th AES Convention*. Available at: https://arxiv.org/abs/1810.12138.

[6] Jean-Marie Lemercier, Julius Richter & Simon Welker. (2023) Diffusion Models for Audio Restoration. In, *Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Espoo, Finland*, Available at: https://arxiv.org/pdf/2402.09821.

[7] Ethan Manilow, Gordon Wichern, Prem Seetharaman & Jonathan Le Roux. (2019) Cutting Music Source Separation Some Slakh: A Dataset to Study the Impact of Training Data Quality and Quantity. In, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Available at: 10.1109/WASPAA.2019.8937170

[8] Alexandre Défossez, Nicolas Usunier, Léon Bottou, Francis Bach. (2019) Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed. In, Cornell University, Available at: https://arxiv.org/abs/1909.01174

[9] Andrew Hines, Jan Skoglund, Anil Kokaram & Naomi Harte. (2015) "ViSQOL: An Objective Speech Quality Model." In, IWAENC 2012; International Workshop on Acoustic Signal Enhancement, Available at: https://ieeexplore.ieee.org/abstract/document/6309421.

[10] Andrés Marafioti, Nathanaël Perraudin, Nicki Holighaus & Piotr Majdak. (2019) A Context Encoder For Audio Inpainting. In, IEEE/ACM Transactions on Audio, Speech, and Language Processing , Available at: 10.1109/TASLP.2019.2947232.