Audio Restoration for Generative Models — Improving MusicGen Outputs

January 4, 2026

Giada Manfredi

Abstract

Recent text-to-music generative models such as MusicGen can produce coherent musical samples directly from textual prompts. Despite their strong semantic consistency, the generated audio often exhibits limited perceptual quality due to noise, quantization artifacts, and reduced spectral detail. In this work, we propose a fully zero-shot audio restoration pipeline that combines machine learning–based source separation with traditional digital signal processing (DSP) techniques. Specifically, we first decompose MusicGen outputs into individual audio stems using Demucs and then apply a DSP-based enhancement pipeline independently to each stem before remixing. Through spectral analysis and quantitative evaluation of low-energy frequency components, we show that the proposed stem-wise restoration consistently achieves superior audio enhancement compared to applying the same processing directly to the mixed audio, with particularly significant improvements in noise reduction.

1. Introduction

Text-to-music generative models such as MusicGen (Jade Copet, 2023) (Alexandre Défossez, 2022) (Ashish Vaswani, 2017) are capable of producing musically coherent audio, but the resulting signals often exhibit perceptual degradation in the form of background noise, codec artifacts, and reduced spectral clarity. These issues limit the usability of generated music in practical scenarios and motivate the need for effective post-processing techniques. Most existing audio restoration approaches operate directly on the mixed audio signal, applying denoising or enhancement without considering the internal structure of musical content. This monolithic treatment can limit the effectiveness of post-processing, as different musical components exhibit distinct spectral and dynamic characteristics.

The main scientific contribution of this work is a simple, fully zero-shot audio restoration pipeline that leverages source separation to improve post-processing effectiveness. Instead of applying enhancement directly to the mixed signal, we first decompose MusicGen-generated audio into individual stems and apply classical DSP-based restoration independently to each source before remixing. We demonstrate that this hybrid processing enables more effective suppression of low-level noise and artifacts compared to direct restoration, without modifying or retraining the generative model.

All materials (code, appendix, audio, figures) are available at `https://github.com/Giada-04/ARGM`.

2. Related works

Traditional audio denoising methods, such as spectral gating (Boll, 1979) and the Wiener filter (Wiener, 1949), estimate noise from the signal and attenuate it in the frequency domain. Spectral gating identifies low-energy regions to approximate the noise spectrum and suppresses those frequencies, while the Wiener filter weights each frequency component according to its estimated signal-to-noise ratio to minimize the mean squared error. Both methods are effective for stationary noise but tend to struggle with the non-stationary. Recent advances in deep learning have addressed these limitations. The advent of architectures such as U-Net (Olaf Ronneberger, 2015), and its audio-specialized variant Wave-U-Net (Daniel Stoller, 2018), paved the way for the development of models like Demucs (Alexandre Défossez, 2019). Demucs not only provides more effective denoising but also represents a state of the art in music source separation model.

However, these learning-based approaches are typically developed and evaluated in supervised settings, relying on paired clean and degraded audio. Consequently, objective evaluation in audio enhancement commonly adopts intrusive perceptual metrics such as PESQ (A.W. Rix, 2001) or STOI (Cees H. Taal, 2011), which require direct comparison with clean reference signals. While effective in

Email: Giada Manfredi <manfredi.2139294@studenti.uniroma1.it>.

Machine Learning 2025, Sapienza University of Rome, 2nd semester a.y. 2024/2025.

supervised speech enhancement benchmarks, such metrics are not applicable to generative music scenarios, where no ground-truth clean audio is available. For this reason, in this work we rely on spectral analysis as a non-intrusive evaluation strategy.

3. Methods

3.1. Baseline: Direct DSP-based audio restoration

As a baseline, we consider a DSP-based audio restoration strategy that operates directly on the mixed audio signal produced by MusicGen.
The DSP pipeline is composed of two stages.
First, spectral denoising is performed using the noisereduce library, which implements an automatic spectral gating algorithm inspired by Audacity. Noise statistics are estimated directly from the signal by identifying low-energy regions and persistent frequency components in the Fourier domain.
Second, dynamic and tonal enhancement is applied using the pedalboard library. The processing chain consists of a noise gate, a dynamic range compressor, a low-shelf equalization filter, and a final gain stage.

3.2. Contribution: DSP-based stem-wise audio restoration

The main contribution of this work is a fully zero-shot audio restoration pipeline that enhances MusicGen outputs by exploiting music source separation prior to DSP-based processing. Instead of directly restoring the mixed audio, we decompose it into individual musical components and process each one independently.

Source separation We use Demucs to decompose MusicGen-generated audio into four stems: vocals, drums, bass, and others. Demucs is based on a U-Net convolutional architecture inspired by Wave-U-Net. The U-Net architecture is an encoder–decoder neural network characterized by a symmetric U-shaped structure. In the encoding path, the audio signal undergoes a progressive reduction in temporal resolution while the number of feature channels increases, allowing the model to extract increasingly abstract and informative representations. The central part of the network, known as the bottleneck, provides a compact representation of the input signal. In the decoding path, the network reconstructs the signal by gradually restoring the original temporal resolution. Skip connections link corresponding layers of the encoder and decoder, enabling the transfer of high-resolution information that would otherwise be lost during compression.

After separation, each stem is processed independently using the same DSP pipeline described in the baseline.

Finally, the restored stems are remixed to produce the final enhanced audio signal.

4. Experimental results

This experiment was conducted on the audio sample generated by MusicGen using the prompt: "*Earthy tones, environmentally conscious, ukulele-infused, harmonic, breezy, easygoing, organic instrumentation, gentle grooves*."
For the analysis of the restored audio, we rely on spectral analysis. Spectral analysis is based on the Fourier Transform, which maps the audio signal from the time domain, describing amplitude variations over time, to the frequency domain, representing the distribution of energy across frequency components.

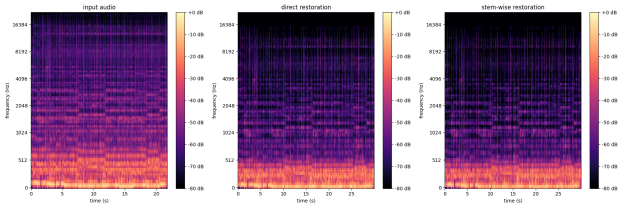

Figure 1.

Qualitative inspection of spectrograms (Figure 1) shows that low-intensity frequency components are noticeably attenuated after restoration, while high-energy musical components remain largely unaffected. Differences between direct restoration and stem-wise restoration are subtle in spectrograms but become evident under quantitative analysis.
To quantify noise reduction, we analyze the 25th percentile of spectral amplitude across frequency bins, which emphasizes low-energy components typically associated with noise and artifacts.

	Input audio	Direct restoration	Stem-wise restoration
25th percentile value	-64.66 dB	-75.77 dB	**-80.00 dB**

These results indicate that applying DSP independently to separated sources enables more effective suppression of low-level noise without degrading musical content.

The four stems before and after restoration and additional results obtained with a different prompt are reported in the appendix.

5. Conclusions

Our results indicate that applying audio restoration after source separation is generally more effective than process-

ing the mixed signal directly, particularly in reducing low-level noise.

The effectiveness of this approach depends on the track's structure. When a track lacks components corresponding to Demucs' predefined stems (i.e., vocals, drums, or bass), separation yields only an informative "others" stem, while the remaining stems are effectively silent. In such cases, source separation not only provides minimal benefit but can actively degrade audio quality by introducing artifacts or additional noise, often resulting in performance comparable to, or worse than, direct restoration.

Future work should explore adaptive methods that apply separation only to tracks with multiple meaningful stems, maximizing restoration quality while minimizing unnecessary processing.

References

Alexandre Défossez, Jade Copet, G. S. Y. A. High fidelity neural audio compression. 2022.

Alexandre Défossez, Nicolas Usunier, L. B. F. B. Music source separation in the waveform domain. 2019.

Ashish Vaswani, Noam Shazeer, N. P. J. U. L. J. A. N. G. L. K. I. P. Attention is all you need. 2017.

A.W. Rix, J.G. Beerends, M. H. A. H. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. 2001.

Boll, S. F. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Audio, Speech and Language Processing*, 1979.

Cees H. Taal, Richard C. Hendriks, R. H. J. J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. 2011.

Daniel Stoller, Sebastian Ewert, S. D. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. 2018.

Jade Copet, Felix Kreuk, I. G. T. R. D. K. G. S. Y. A. A. D. Simple and controllable music generation. 2023.

Olaf Ronneberger, Philipp Fischer, T. B. U-net: Convolutional networks for biomedical image segmentation. 2015.

Wiener, N. Extrapolation, interpolation, and smoothing of stationary time series. 1949.