A Hybrid Pipeline for Restoration of AI–Generated Instrumental Music: Demucs Separation, S–NMF, 2D UNet STFT Mask, and WaveNet Denoising

August 26, 2025

Edoardo Ensoli Andrea Massignan

GitHub: https://github.com/duchannes19/audio-restoration-for-generative-models

1. Introduction

Our goal is a modular, post–generation restoration stack for MusicGen–like outputs(Copet et al., 2023) that (1) is robust to mixed artifacts, (2) preserves transients and timbral brightness, (3) is interactive on commodity hardware, and (4) remains interpretable. We propose a carefully engineered hybrid that couples separation and structured noise modeling with two learned denoisers. Our contribution is an *operationally conservative* integration: pre–clean to avoid biasing masks; S–NMF *per stem* for stationary noise; and a compact *2D* UNet in STFT space with hop–aligned overlap–add. We additionally benchmark a *WaveNet* denoiser (mono, checkpoint–driven) to probe time–domain modeling tradeoffs.

2. Related Work

Demucs(Defossez et al., 2019) and successors are standard for music source separation and often form the front end of restoration pipelines. Nonnegative matrix factorization (NMF) and semi–supervised variants effectively model stationary noise. UNet–style spectrogram masking has been widely used in speech/music enhancement(Jansson et al., 2017). Autoregressive/residual architectures such as WaveNet variants have also been explored for audio denoising/restoration in the waveform domain. Our emphasis is on an *integration strategy* rather than novel architectures.

3. Method

Clean (Mixture)

We first repair pathologies that degrade later stages:

- **Declipping:** dual–direction LPC gap filling on runs detected by percentile thresholding.
- **De–quantization:** multi–cycle dither with residual tracking where a quantization step is detected.
- **Minimum–statistics FD denoise:** conservative floor estimation to reduce steady hiss without damaging transients.

Baseline: Demucs + S–NMF (Per Stem)

The pre–cleaned mixture is upmixed to stereo if needed and separated using `htdemucs`. For each stem magnitude we fit a semi–supervised NMF model with a noise dictionary learned from minimum–statistics frames. A soft Wiener–style mask emphasizes the music reconstruction over estimated noise; a tiny high–band tilt (roughly 0.5–1.5 dB) restores brilliance. Stems are recombined to the final stereo track.

Contribution: 2D UNet STFT Mask (Mono Input)

Input and target. The UNet path operates on *mono* audio (the app downmixes if the degraded input is stereo). We compute an STFT (n_{fft}=1024, hop 256), feed the magnitude $|Y| \in \mathbb{R}^{F\times T}$ to a *2D* UNet (convolutions over frequency and time), and predict a mask $M \in [0,1]^{F\times T}$. The output is $\hat{y} = \text{ISTFT}(M \odot |Y|, \angle Y)$. If the original was stereo, we duplicate the mono UNet output for comparison.

Architecture. Four encoder levels with base channels 32 doubling by depth, GELU activations, GroupNorm; symmetric decoder with transposed–conv upsampling and skip concatenations; sigmoid mask head (UNet style).

Training. We train on 3–4 s mono crops at 32 kHz using degradations that mimic MusicGen: down–up resampling (12–22.05 kHz), bit crush (8–12 bit) or μ-law, soft clipping (1–6 dB), light multi–delay reverb, additive environmental noise at SNR in $[-5, 20]$ dB, and tiny Gaussian dither. Loss combines time–domain L_1 and log–magnitude L_1. Optimization uses AdamW, cosine schedule with warmup, gradient accumulation (effective batch 8), AMP, and optional EMA.

Email: Andrea Massignan, Edoardo Ensoli <massignan.1796802@studenti.uniroma1.it; ensoli.1918623@studenti.uniroma1.it>.

Deep Learning and Applied AI 2025, Sapienza University of Rome, 2nd semester a.y. 2024/2025.

Seamless inference. Hop–aligned OLA with overlap and extra left/right context; each chunk is denoised, core–cropped, and overlap–added with Hann fades and energy normalization.

Additional Baseline: WaveNet Denoiser (Mono)

We add a time–domain WaveNet denoiser as a strong baseline and complementary viewpoint.

I/O and resampling. Inference runs on mono at the model's training rate f_m (from the checkpoint/config). Inputs are auto–resampled from the session rate to f_m and the output is resampled back.

Overlap inference. We process the waveform with chunked inference whose minimum chunk $\geq$ four receptive fields (or $\geq$2 s), with fractional overlap (default 0.1). Fades are applied on overlaps; chunks are averaged to avoid seams.

Checkpoints and size. The app supports `.pt/.pth` checkpoints and model sizes (`small/medium/medium{lite/large`); if the checkpoint carries the full config, it overrides the size selection.

Implementation Notes Aligned with the App

- **Tabs:** Setup, Generate, Degrade, Restore (Demucs), Restore (UNet), Restore (WaveNet), Compare, Reset.
- **UNet checkpoint:** defaults to `./checkpoints/best.pt`; configurable in *Restore (UNet)*.
- **WaveNet checkpoint:** user–provided; model size selectable; device `auto/cuda/cpu`; mono with auto resampling to/from f_m.

4. Experiments and Results

Visual analysis

Given the heterogeneous artifact mix (hiss, hum, quantization harshness, codec haze), scalar metrics (SNR, segSNR, LSD, STOI) only partially reflect perceived quality. We therefore prioritise a *diagnostic, visual* comparison: waveform overlays, difference signals, and log-magnitude spectrograms. These highlight (i) noise-floor suppression, (ii) transient integrity, and (iii) high-frequency behaviour.

Qualitative observations

- **Demucs + S–NMF.** In Fig. 1 the background haze between onsets is strongly reduced vs. the degraded input, with a darker floor across 2–14 kHz. Transient ridges remain well defined, while the difference traces in Fig. 1 cluster tightly around zero except at sharp hits, indicating restrained modification. Residual fine-grain fizz can persist above ∼10 kHz on sustained content.
- **UNet (STFT mask).** Fig. 2 (top) shows effective smoothing of the high-band speckle (10–16 kHz) and removal of quasi-stationary hiss. The overlay/difference panel (Fig. 3) reveals slightly stronger attenuation around peaks (mild transient softening), consistent with the conservative log-spectral masking.
- **WaveNet (time domain).** In Fig. 2 (bottom) transients retain sharp vertical structure with low inter-onset haze; the difference signal remains low except at attacks, pointing to good temporal fidelity. HF energy is reintroduced more coherently than the degraded case, with less speckled noise; occasional brightening is visible on cymbal-like events.

Considerations. Demucs+S–NMF yields the *cleanest spectral floor* with minimal structural change; UNet is the most conservative high-band smoother; WaveNet best preserves *temporal edges* and perceived clarity while keeping residual noise low. The three methods are complementary, supporting the hybrid design.

5. Limitations and Future Work

Limitations: reliance on synthetic degradations, no MOS; UNet path is mono (no inter–channel coherence); WaveNet can drift spectrally (higher LSD) despite strong temporal metrics.

Possible Imporovements: stereo–aware UNet/WaveNet with inter–channel features, multi–objective training to jointly optimize SNR/LSD/STOI, joint separa-

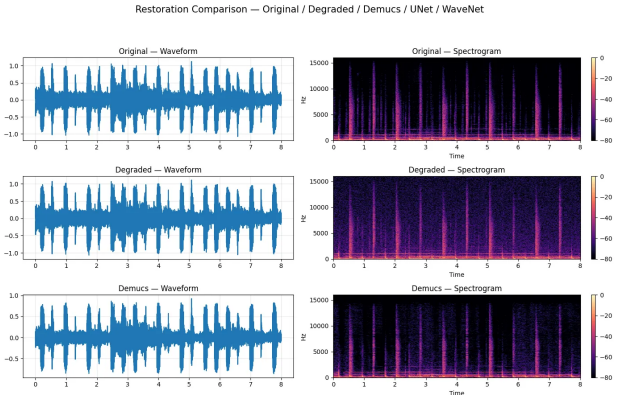

Figure 1.

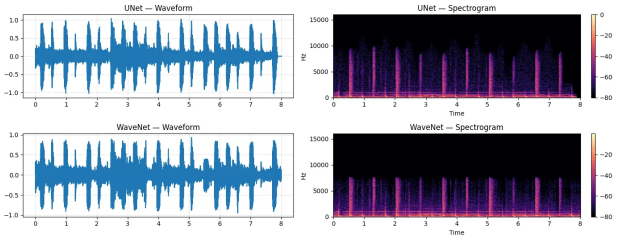

Figure 2.

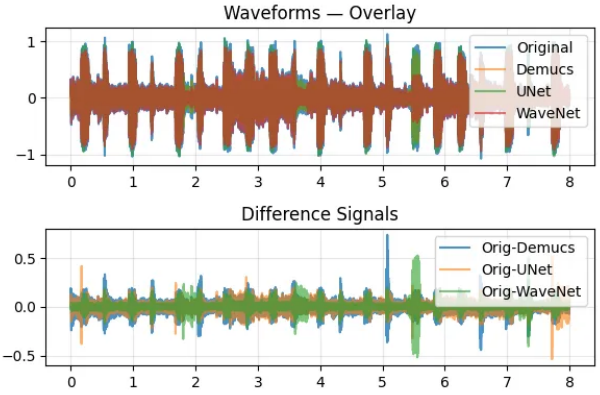

Figure 3.

tion + denoising, streaming/causal inference, and a curated paired dataset of real MusicGen outputs with human–cleaned references.

6. Conclusion

We presented a modular restoration pipeline and app that combine Demucs+S–NMF per stem (stereo–aware), a compact *2D* UNet STFT mask (mono, hop–aligned OLA), and a time–domain WaveNet denoiser (mono, overlap inference). The three paths expose complementary strengths: Demucs+S–NMF minimizes spectral artifacts (LSD), WaveNet maximizes temporal SNR/segSNR and STOI, and UNet provides a conservative mask that is robust and fast.

7. Appendix

We had initially set our sights on a different project: the DjNet one. However, we found it a bit too difficult, especially with regard to how to effectively get MusicGen to cooperate with the task. We had initially planned to use MusicGen's strengths in generating high-quality audio samples, but we encountered challenges in aligning its output with the DjNet framework's requirements. The idea was simple: take the beginning and end of a song and generate a bridge between them. Unfortunately, it was not meant to be. Instead, we pivoted to this project, which we found to be more manageable and interesting. Initially, we explored all possible approaches to denoising, which first led us to Demucs. Eventually, we built a prototype that leveraged Demucs and many other denoising techniques. Once we discovered that the most stable version of this approach used S-NMF per stem, we decided to adopt it. We then thought that a learned denoiser could be a valuable addition to the pipeline. To this end, we built a UNet in the spectrogram domain and trained it on different types of datasets that mimicked the type of noise found in MusicGen samples. Finally, we added a WaveNet denoiser to provide a strong baseline and a complementary viewpoint. Ultimately, we decided to build a simple Gradio app around it so that we could easily test and compare the different approaches. The app also allowed us to visualise the results more intuitively, which helped us to understand the strengths and weaknesses of each method. Overall, this project has been a valuable learning experience.

Roles:

- **Andrea:** Demucs and Application.
- **Edoardo:** WaveNet and Datasets.
- **Both:** UNet.

References

Copet, J., Kreuk, F., Zaveri, J., Defossez, A., Adi, Y., Synnaeve, G., Copet, J., and et al. Simple and controllable music generation, 2023. URL https://arxiv.org/abs/2306.05284.

Defossez, A., Usunier, N., Bottou, L., and Bach, F. Demucs: Deep extractor for music sources with extra unlabeled data remixed, 2019. URL https://arxiv.org/abs/1909.01174.

Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., and Weyde, T. Singing voice separation with deep u-net convolutional networks. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.