Restauro Spettrale e Analisi degli Artefatti per Modelli di Generazione Musicale MusicGen

January 16, 2026

Alessandro Orsini

Abstract

Questo lavoro affronta i limiti qualitativi della generazione musicale basata su Transformer, focalizzandosi sul "vuoto spettrale" introdotto dal codec neurale EnCodec a 32kHz nei modelli MusicGen. Tale limite impone un taglio netto a 16kHz, privando l'audio della brillantezza necessaria per gli standard professionali. Proponiamo un sistema di restauro basato su un'architettura U-Net convoluzionale ottimizzata tramite Multi-Resolution STFT Loss. Per l'addestramento, è stata sviluppata una pipeline di degradazione calibrata statisticamente sui parametri di MusicGen. Sebbene i risultati indichino un successo quantitativo nel recupero della banda passante, l'analisi qualitativa evidenzia un paradosso tra miglioramento numerico e percezione uditiva dovuto all'incoerenza di fase. https://github.com/orsini2155841/Audio_restoration_from_MusicGen

1. Introduzione

I modelli generativi stato-dell'arte come MusicGen-Large, pur eccellendo nella composizione, presentano deficit strutturali nel dominio della frequenza. L'uso del codec EnCodec limita il campionamento a 32kHz, imponendo un filtro "brick-wall" a 16kHz (limite di Nyquist). Questo studio mira ad analizzare matematicamente il divario tra audio sintetico e reale e a implementare un modello di *Audio Super-Resolution* per sintetizzare le armoniche mancanti, riportando il segnale allo standard di 44.1kHz.

Email: Alessandro <orsini.2155841@studenti.uniroma1.it>.

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

2. Analisi degli Artefatti e Related Work

L'analisi esplorativa condotta su 200 campioni evidenzia che MusicGen concentra l'energia sotto i 10.5kHz (Spectral Rolloff medio: 10.4kHz). Un parametro critico rilevato è la *Spectral Flatness* (0.0001 per MusicGen vs 0.26 per MUSDB18-HQ), che indica una "sterilità" timbrica dovuta alla quantizzazione neurale.

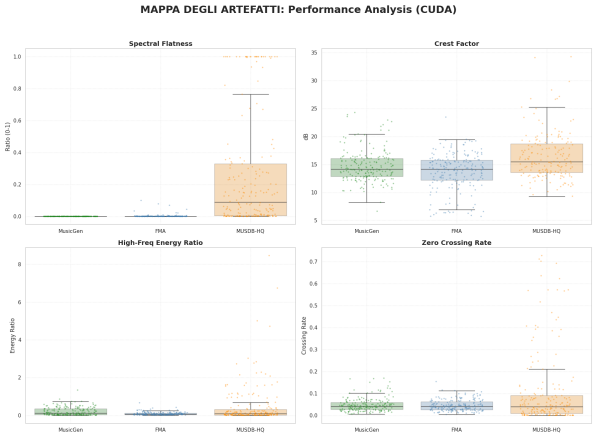

Figure 1. Comparazione delle metriche fisiche tra dataset reali e generati (MusicGen).

In Tabella 1 viene sintetizzato il gap statistico rilevato tra i segnali sintetici e i riferimenti reali, evidenziando la monotonia spettrale del modello generativo.

La letteratura corrente affronta il problema tramite modelli U-Net o approcci basati su spettrogrammi. Il lavoro si appoggia alla *Multi-Resolution STFT Loss* introdotta per i modelli vocoder per garantire stabilità nel dominio tempo-frequenza.

3. Metodologia

Il sistema di restauro si basa su una pipeline di degradazione realistica e un'architettura neurale

Table 1. Sintesi del Gap Statistico rilevato.

Metrica	MGen	Reali	Interpretazione
Rolloff	10.4k	12-13k	Perdita di 2-3 kHz medi.
Varianza	Bassa	Alta	Taglio fisso e sistematico.
Outliers	Assenti	¿20k	Mancanza di "aria" sonora.

U-Net.

Degradazione Controllata. Per addestrare la rete su dati accoppiati, abbiamo sviluppato un simulatore stocastico che trasforma audio HQ in campioni statisticamente identici a MusicGen. Il processo include resampling a 32kHz, filtraggio passabasso (Rolloff target: 10.4kHz) e quantizzazione neurale a 64 livelli per replicare la bassa *flatness* del segnale originale.

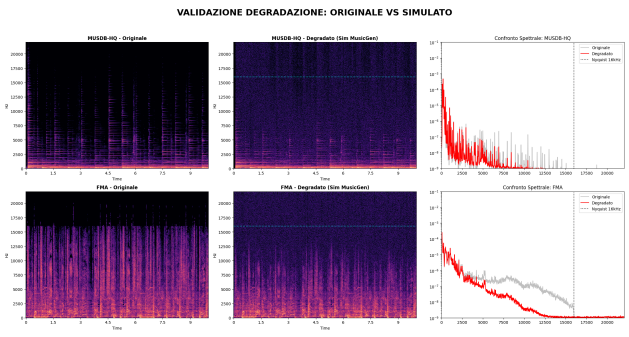

Figure 2. Comparazione tra file originale e degradato

Architettura e Ottimizzazione. Abbiamo implementato una U-Net spettrale che opera sulla magnitudo STFT. La rete utilizza *skip connections* per preservare la fase originale nelle medie frequenze. L'ottimizzazione avviene minimizzando il rapporto tra magnitudo predetta e target:

$$L_{sc}(\text{mag}, \widehat{\text{mag}}) = \frac{\|\text{mag} - \widehat{\text{mag}}\|_F}{\|\text{mag}\|_F}. \quad (1)$$

4. Risultati Sperimentali

Il modello è stato addestrato per 20 epoche sfruttando l'accelerazione GPU Dual T4 (impiegando comunque piu di 8ore). I risultati quantitativi (Tabella 2) mostrano un netto recupero dell'estensione spettrale media.

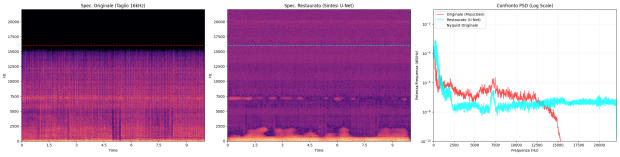

Figure 3. Confronto PSD: MusicGen originale (rosso) vs restaurato (ciano). Si nota il riempimento della banda oltre i 16kHz.

Table 2. Confronto delle prestazioni medie su 200 campioni.

Metrica	MusicGen	Restaurato	Target (HQ)
Rolloff (99%)	10.4 kHz	**16.8 kHz**	13.5 kHz
HF Energy Ratio	0.20	**0.42**	0.47
Centroid Hz	2256	**3120**	3844

5. Discussione e Conclusioni

Nonostante il successo numerico, l'ascolto rivela un limite fondamentale: la rete agisce come un generatore di *shaped noise*. Poiché la fase originale sopra i 16kHz è assente, la ricostruzione introduce artefatti stocastici. Il modello "riempie" il vuoto con energia, ma fallisce nel ricostruire la coerenza armonica fine, portando a un paradosso dove le metriche fisiche migliorano ma la qualità percepita resta degradata.

In conclusione, il solo approccio basato su regressione della magnitudo non è sufficiente. Futuri sviluppi dovranno integrare discriminatori avversariali (GAN) per forzare il realismo timbrico e vocoder neurali per il recupero coerente della fase.

References

[1] Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07, 2008.

[2] Jade Copet, Felix Kreuk, Itai Gat, Tal Talitman, Jade Vyas, Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023.

[3] Alexandre Défossez, Jade Copet, Baptiste Rozière, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[4] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech

synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

[5] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for audio source separation. *arXiv preprint arXiv:1806.03185*, 2018.

[6] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203. IEEE, 2020.