Project ID 2 (Restauro audio tramite reti spettrali: confronto tra rappresentazioni lineari e Mel)

October 28, 2025

Marika Maria Rago

Abstract

I recenti modelli generativi producono musica a partire da prompt testuali, ma le tracce generate presentano spesso lievi degradazioni percettive. In questo lavoro propongo una pipeline di restauro audio basata su **reti neurali spettrali**, finalizzata a migliorare la qualità sonora dei brani generati da modelli di sintesi. Il modello *Spectral Enhancement* è stato addestrato per operare su due differenti rappresentazioni del dominio tempo-frequenza: lo spettrogramma lineare e quello Mel. I risultati indicano che ciascun modello apprende pattern specifici del proprio dominio, pur mostrando una limitata capacità di generalizzare a degradazioni non viste.

1. Introduzione e Motivazione

La generazione automatica di musica mediante reti neurali ha rappresentato un significativo progresso nella creatività computazionale, consentendo la sintesi di brani coerenti a partire da semplici prompt testuali. Tra i modelli più recenti, MusicGen (Copet et al., 2023) ha mostrato come l'integrazione di rappresentazioni linguistiche e acustiche consenta di generare contenuti musicali plausibili in modo controllabile. Tuttavia, nonostante l'elevata coerenza strutturale, le tracce prodotte evidenziano ancora limitazioni percettive, tra cui perdita di dettaglio armonico, attenuazione delle alte frequenze e presenza di artefatti di compressione o quantizzazione. Tali imperfezioni, pur mantenendo intatta la componente semantica e melodica del brano, ne compromettono la fedeltà timbrica e la sensazione di realismo.

In questo contesto, il **restauro audio generativo** mira a colmare il divario qualitativo tra segnali sintetizzati e produzioni ad alta fedeltà. Come nell'elaborazione di immagini e video generati da modelli neurali, una fase di *post-processing* dedicata consente di migliorare le caratteristiche percettive del contenuto senza alterarne la struttura semantica. L'impiego di **reti neurali spettrali** per l'enhancement consente di modellare la relazione tra rappresentazioni degradate e segnali puliti, fornendo un approccio flessibile e adattabile a differenti tipi di distorsione.

Email: Marika Maria Rago <rago.2090371@studenti.uniroma1.it>.

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

2. Lavori correlati

MusicGen (Copet et al., 2023) rappresenta uno dei più avanzati modelli di generazione musicale sviluppati da Meta AI. La sua architettura si fonda su un *Language Model* autoregressivo di tipo Transformer, addestrato a prevedere sequenze di token audio discreti derivanti da EnCodec, un autoencoder convoluzionale che impiega la *Residual Vector Quantization* (RVQ) per comprimere il segnale audio in rappresentazioni latenti a basso frame rate. Questa strategia consente di modellare l'audio come una sequenza discreta, permettendo la generazione controllata di brani coerenti a partire da prompt testuali. Nonostante l'elevata coerenza strutturale, le produzioni di MusicGen presentano ancora limitazioni percettive, come già detto.

BABE-2 (Moliner et al., 2024) (Blind Audio Bandwidth Extension 2) costituisce un recente avanzamento nello stato dell'arte del *music enhancement* neurale basato su modelli di diffusione. Il metodo affronta il problema del restauro di registrazioni musicali degradate attraverso il framework del *Diffusion Posterior Sampling* (DPS), stimando simultaneamente un filtro di degradazione parametrico e la ricostruzione audio. Grazie a questa formulazione, BABE-2 è in grado di ricostruire le componenti spettrali ad alta frequenza e di riequilibrare la distribuzione energetica del segnale, ottenendo miglioramenti significativi in termini di timbrica e chiarezza percettiva. Tuttavia, le prestazioni elevate del modello sono bilanciate da un notevole costo computazionale, poiché l'inferenza richiede molteplici iterazioni di campionamento e ottimizzazione dei parametri.

Il presente lavoro propone un approccio più leggero e mirato basato su reti convoluzionali multi-scala in stile U-Net, testato su due gruppi di degradazioni.

3. Il mio approccio

L'implementazione del progetto è disponibile su GitHub: https://github.com/marika-rago/projectID2ML.

Dataset e degradazioni Per l'addestramento e la valutazione del modello è stato utilizzato il dataset **FMA Small**, che contiene 8k clip musicali. Da questo sono stati selezionati casualmente circa il 16% dei file, ogni clip viene ridotta a segmenti di 3 secondi, campionati a 32 kHz e normalizzati in ampiezza. Le degradazioni applicate simulano difetti reali di registrazioni o di codifica digitale, producendo coppie $(x_{\text{degraded}}, x_{\text{clean}})$ per l'apprendimento supervisionato. Sono stati definiti due gruppi distinti di degradazioni, il **Gruppo A** (rumore additivo e fenomeni periodici, *quantizzazione*, *tonal stripes*, *noise*), e il **Gruppo B** (distorsioni non lineari e convolutive, *clipping*, *reverbero*, *low-pass*, *distorsione armonica*). Questa distinzione consente di valutare la capacità di generalizzazione tra distorsioni di diversa natura.

Architettura Il modello **Spectral Enhancement Net** è composto da moduli gerarchici ispirati alla struttura della *U-Net*. Alla base si trovano i blocchi *MultiScaleCNN*, che eseguono convoluzioni parallele con kernel 3×3, 5×5 e 7×7 per catturare pattern locali e globali a diverse scale. Le feature risultanti vengono poi elaborate da *ResidualBlock* con normalizzazione e *dropout*, aumentando la stabilità dell'addestramento e riducendo l'overfitting. L'*encoder* comprime progressivamente lo spettrogramma tramite convoluzioni con *stride*, concentrando l'informazione sulle componenti spettrali più rilevanti. Il *decoder* ricostruisce la struttura originaria mediante upsampling bilineare e *skip connection* simmetriche, preservando i dettagli armonici e di banda fine. Il *bottleneck* centrale integra tre blocchi residuali che fungono da livello di trasformazione non lineare profonda, mentre lo strato finale applica una convoluzione 1×1 per generare lo spettrogramma restaurato. A seconda del dominio di rappresentazione, la normalizzazione dell'uscita varia, nei **Mel-spettrogrammi** è utilizzata una funzione sigmoide, mentre nei **spettrogrammi lineari** il range è limitato tramite *clamping* in $[0, 1]$ durante l'addestramento.

Funzione di perdita È stata progettata una loss composita, **HybridLoss**, che combina sei termini per modellare diversi aspetti della qualità percettiva: *Charbonnier Loss* (fedeltà puntuale), *Do-No-Harm Loss* (evita modifiche non necessarie), *High-Frequency Loss* (recupero delle alte frequenze), *Spectral Flatness Loss* (coerenza tonale), *Band-Weighted L1 Loss* (coerenza per banda) e *Energy Consistency Loss* (bilanciamento energetico).

Addestramento e metriche L'addestramento utilizza *AdamW* (lr 10^{-4}) e *ReduceLROnPlateau scheduler*, per un massimo di 25 epoche con *early stopping*. È stato impiegato il *mixed precision training* per ridurre i tempi di calcolo e la memoria GPU. Durante la validazione, sono state monitorate **MSE**, **MAE**, **Cosine Similarity** e **LSD** (solo per spettrogrammi Mel).

Risultati Sono state addestrate due versioni del modello: una sullo **spettrogramma lineare** e una sullo **spettrogramma Mel**. Il primo mantiene una rappresentazione diretta delle frequenze, mentre il secondo enfatizza la percezione umana nelle bande logaritmiche, consentendo di valutare l'influenza della rappresentazione sul restauro spettrale. I risultati quantitativi (Table1) evidenziano un comportamento fortemente *domain-dependent*, entrambi i modelli ottengono le migliori prestazioni nei test **in-domain**, mentre perdono efficacia nei casi **out-of-domain**, mostrando limitata generalizzazione verso degradazioni non viste. Il modello **Mel** risulta più efficace sulle distorsioni additive, migliorando MSE e LSD, mentre quello **Lineare** mostra maggiore stabilità complessiva e coerenza nelle metriche di similarità. Una valutazione qualitativa è stata condotta applicando il modello lineare a un brano generato da **MusicGen**. Lo spettrogramma restaurato è stato ricostruito in dominio audio con *Griffin-Lim*, e l'ascolto conferma che il modello preserva la qualità della traccia originale. Entrambi i modelli mantengono un comportamento conservativo, riconoscendo la buona qualità dell'input e limitando l'intervento a quanto necessario.

Table 1. (Δ) tra audio degradato e restaurato.

Train/Test	ΔMSE↑	ΔL1↑	ΔCOS↓	ΔLSD↑
(Mel) A → A	+0.0036	+0.0033	-0.0106	+0.6132
(Mel) A → B	-0.0006	-0.0052	+0.0026	-0.4602
(Mel) B → A	-0.0010	-0.0137	+0.0076	-1.2787
(Mel) B → B	+0.0165	+0.0319	-0.0180	+2.7853
(Lin) A → A	+0.0074	+0.0143	-0.0323	—
(Lin) A → B	-0.0005	-0.0037	+0.0028	—
(Lin) B → A	-0.0005	-0.0102	+0.0066	—
(Lin) B → B	+0.0087	+0.0130	-0.0205	—

4. Sviluppi futuri

Un possibile sviluppo futuro consiste nell'applicazione di tecniche di **continual learning**, per estendere la robustezza del modello a diverse classi di degradazioni senza necessità di riaddestramento completo. Un'altra opzione potrebbe essere l'adozione di modelli **diffusion-based** per il restauro, che, pur essendo computazionalmente più onerosi, offrono prestazioni di qualità superiore in compiti di *music enhancement* e ricostruzione percettiva.

References

Copet, J., Kreuk, F., Défossez, A., Synnaeve, G., Adi, Y., Gat, I., Remez, T., and Kant, D. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023. URL https://arxiv.org/pdf/2306.05284.

Moliner, E., Turunen, M., Elvander, F., and Välimäki, V. Babe-2: A diffusion-based generative equalizer for music restoration. *arXiv preprint arXiv:2403.18636*, 2024. URL https://arxiv.org/html/2403.18636v2.