Restauro zero-shot con demucs degli audio di Musicgen

August 29, 2025

Aurora Di Giovanna

Abstract

I modelli audio generativi producono spesso output con una qualità sonora inferiore agli standard professionali. L'obiettivo di questo progetto è l'implementazione di una pipeline di restauro che usi un approccio zero-shot basata su demucs. Il progetto si concentra sull'analisi spettrale e percettiva del miglioramento ottenuto.

1. Restauro audio

1.1. Generazione audio utilizzando Musicgen

Gli audio vengono generati utilizzando il modello Musicgen-small (2), dando dei prompt di generi diversi. Ogni audio ha una durata di 5s che può essere aumentata o diminuita.

1.2. Restauro utilizzando demucs

Per restaurare gli audio generati si è scelto di riconvertire un modello pre-addestrato per la separazione: **Demucs**, un modello di deep learning sviluppato da Meta AI introdotto in (4) e migliorato con un approccio ibrido in (3). Nonostante demucs sia stato creato per la separazione è stato studiato ed usato anche per il restauro audio, come nel denoising (10) e nel declipping (6). Questo modello ha una struttura **Encoder-Decoder a U-Net** .

- **L'encoder**: l'encoder di Demucs processa l'audio in due modi paralleli:
  1. **Waveform**: il segnale di audio grezzo (forma d'onda) viene processato in diversi strati convoluzionali 1D;
  2. **Spectogram**: lo spettrogramma viene poi processato da strati convoluzionali 2D.

  L'encoder risponde alla domanda "Cosa c'è in questo audio?" prendendo lo spettrogramma degradato e processandolo in livelli, per ogni livello esegue due azioni: applica dei filtri per estrarre le caratteristiche (texture, forma...) e il downsampling finché non arriva al bottleneck, dove abbiamo una rappresentazione densa perché le dimensioni dei dati vengono ristrette.
- **Decoder**: il decoder di Demucs andrà a ricostruire l'intero segnale audio partendo dalla rappresentazione compressa.
- **U-Net**: L'architettura U-Net, introdotta originariamente per la segmentazione di immagini biomediche (8), è stata successivamente adattata con successo anche all'audio, ad esempio con Wave-U-Net (9). Questra struttura ha la forma di una U, con il processo di encoding andiamo ad eseguire la discesa e lentamente viene ridotta la risoluzione temporale e viene aumentata la profondità, ovvero il numero dei canali, costringendo il modello ad imparare suoni sempre più compressi. Il punto più profondo della "U" è il **bottleneck**. In questo livello Demucs inserisce degli strati a breve memoria che permettono ad esso di ricordare i pattern musicali. Con il processo di encoding andiamo ad eseguire la risalita aumentando la risoluzione temporale e in questa fase gli strati del percorso di discesa vengono connessi ai loro corrispondenti del percorso di risalita grazie alle skip connections, che permettono di combinare dati a bassa risoluzione con dettagli ad alta risoluzione.

2. Analisi Visiva e quantitativa

2.1. Analisi Spettrale

L'analisi spettrale fornisce una rappresentazione visiva del processo di restauro nel dominio della frequenza. Per ogni audio campione del dataset sono stati generati due spettrogrammi in scala Mel:

- Nella *Figure 1.* troviamo l'audio generato da Musicgen che presenta un taglio nello nelle alte frequenze ovvero una mancanza di nitidezza. Inoltre si nota anche la presenza di rumore maggiormente nelle basse frequenze.
- Nella *Figure 2.* troviamo l'audio restaurato che presenta delle alte frequenze ricostruite e meno rumore nelle basse frequenze, riducendo visivamente le macchie.

Email: Aurora Di Giovanna <digiovanna.2127738@studenti.uniroma1.it>.

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

Figure 1. Input di Musicgen

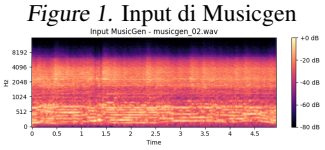


Figure 2. Audio restaurato

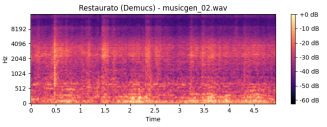


Figure 3. Demucs

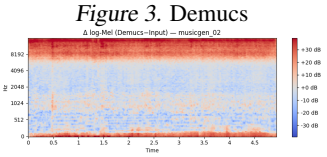
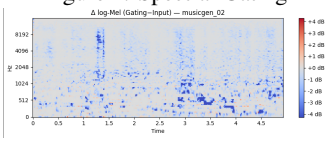


Figure 4. Spectral Gating



2.2. Baseline: Spectral Gating e metriche proxy di valutazione

Per valutare l'utilità di questo modello eseguiamo un metodo di paragone: una baseline. In questo progetto la baseline è rappresentata dallo *Spectral Gating* basato sul dominio della frequenza. Questo metodo consiste nell'analizzare lo spettrogramma STFT (a breve termine) del segnale audio e per ciascuna frequenza stimare una soglia di rumore. Lo Spectral Gating non ricostruisce ma elimina le basse frequenze ritenute bande rumorose. Viene quindi rimosso il rumore senza aggiungere altri dati. Per una valutazione quantitativa del restauro usiamo tre metriche proxy, confrontando input-Demucs e input-Spectral Gating :

1. **HBER (Highband Energy Gain)**: la funzione *highband_energy_gain* misura la variazione di energia nel segnale nelle alte frequenze. Quindi un aumento di energia in questa funzione è positivo poiché indica nitidezza e riduzione del rumore.
2. **Spectral Flatness:** la funzione *spectral_flatness_mean* calcola la piattezza spettrale del segnale di output. Il valore della spectral flatness più è vicino a uno = rumore bianco, più è vicino a zero = segnale tonal (5; 7)
3. **Silence Noise Floor**: la funzione *silence_noise_floor* calcola il livello di rumore di fondo durante gli intervalli di silenzio. Si basa sul calcolo dell'RMS (Roat Mean Square), più è basso il suo valore e meno rumore sarà presente nell'audio e ha un valore soglia di -45 dB (1; 11).

Possiamo notare che lo Spectral-Gating agisce come filtro passa-basso, come in *Figure 3.*, e che il rumore a bassa energia viene rimosso o limitato (zone scure dello spettrogramma).

Con le misure proxy andiamo a quantificare le conseguenze dell'applicazione dello Spectral-Gating:

Table 1. Risultati delle metriche proxy per il file musicgen_02.wav

file	Flat_in	Flat_out	ΔHBER_dB
musicgen_02.wav	0.001	0.051	-21.732

Noise Floor (dB)	Input	Output
musicgen_02.wav	NaN	NaN

3. Conclusioni

In conclusione, questo progetto ha dimostrato l'utilità di una pipeline di restauro zero-shot basata su Demucs, rivelando un miglioramento del segnale audio nelle alte frequenze e al tempo stesso delle difficoltà nel restauro di audio complessi, soprattutto nelle bassse frequenze. La baseline basata sul Spectral Gating è risultata soddisfacente tuttavia può portare all'eliminazione di segnali utili e all'inserimento di nuovi rumori. Demucs ha quindi mostrato una migliore efficienza nel riconoscere rumore e segnale utile (segnale musicale), usando la separazione delle sorgenti.

Ulteriori esperimenti sono riportati nel notebook allegato.

References

[1] Objective measurement of active speech level. Technical report, ITU-T Recommendation P.56, 1993.

[2] Jade Copet et al. Simple and controllable music generation. In *NeurIPS*, 2023.

[3] Alexandre Défossez. Hybrid spectrogram and waveform source separation. In *arXiv:2111.03600*, 2021.

[4] Alexandre Défossez et al. Music source separation in the waveform domain. In *arXiv:1911.13254*, 2019.

[5] James D Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on Selected Areas in Communications*, 6(2):314–323, 1988.

[6] A. Marafioti et al. Audio declipping with deep neural networks. In *ICMLA*, 2019.

[7] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification). Technical report, IRCAM, 2004.

[8] Olaf Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[9] Dominic Stölter et al. Wave-u-net: A multi-scale neural network for audio source separation. In *ISMIR*, 2018.

[10] Cassia Valentini-Botinhao et al. Speech enhancement using deep recurrent neural networks. In *Interspeech*, 2016.

[11] A. Varga and H. Steeneken. Assessment for automatic speech recognition: Noisex-92. *Speech Communication*, 12(3):247–251, 1993.