Relazione: “ID 2: Audio Restoration for Generative Models-Improving MusicGen Outputs” Giorgia Cecchi 2069895

Negli ultimi anni i modelli generativi di musica basati su reti neurali hanno raggiunto traguardi significativi. Tra i più noti c’è MusicGen, sviluppato da Gabriel Défossez e dal team di Meta AI: un modello **text-to-music** che permette di generare brani musicali a partire da semplici prompt testuali. Usandolo, però, ho notato che le tracce prodotte presentano spesso **fruscio ad alta frequenza (hiss)** e piccoli artefatti di sintesi, che riducono la qualità percepita dell’audio.

L’obiettivo del mio progetto è stato applicare tecniche di **audio restoration** alle tracce generate da MusicGen. La domanda che mi sono posta è stata: ***“Posso ridurre il rumore tipico di MusicGen senza snaturarne la musicalità?”*.**

Per poter rispondere ho sperimentato diverse soluzioni. In un primo momento ho provato approcci neurali già pronti, come **Demucs e Facebook Denoiser**, eliminavano gran parte del rumore, ma allo stesso tempo snaturavano il suono, rendendolo poco naturale e lontano dall’originale. Di conseguenza ho scelto di concentrarmi su **tre modelli Digital Signal Processing (DSP)**, più leggeri e trasparenti, con lo stesso fine: ridurre il fruscio senza alterare l’identità musicale delle tracce. In particolare, letteratura si trovano molte soluzioni basate su metodi spettrali, statistici o neurali, ma nel mio caso i DSP si sono dimostrati l’opzione più adatta.

**Metodologia:**

Per mettere in pratica queste idee ho seguito uno schema di lavoro ben preciso:

- Ho configurato l’ambiente di lavoro installando le librerie necessarie.
- Ho caricato **MusicGen (facebook/musicgen-small)** da Hugging Face e generato cinque clip musicali a partire da prompt testuali.
- Ho salvato e organizzato i file audio in formato .wav (16 kHz) per prepararli alla fase di restauro.
- Ho applicato i tre modelli DSP e li ho confrontati con lo stesso **protocollo di valutazione**: stesse metriche per tutti, più un ascolto soggettivo.
- Infine, ho analizzato i risultati con grafici e tabelle riassuntive per rendere il confronto chiaro.

**Metriche di valutazione:**

Per valutare i modelli ho usato tre metriche che mi permettono di misurare sia la riduzione del rumore che la fedeltà musicale:

- **HF-band energy (dB):** l’ho usata per misurare l’energia sopra i 6 kHz, cioè nella zona dove si concentra il fruscio. Se il valore diminuisce significa che il rumore è stato ridotto.
- **HPSS ratio (Harmonic Percussive Source Separation)**: mi è servita per valutare il rapporto tra componente armonica e percussiva. In questo modo ho potuto capire se il filtro andava a intaccare la musicalità o se la preservava.
- **Log-Mel distance**: l’ho utilizzata per confrontare lo spettrogramma log-mel dell’audio originale con quello elaborato. Se la distanza rimane bassa, vuol dire che il timbro è rimasto fedele.

**Modello 1: NoiseReduce (DSP)**

Per il primo modello ho scelto di utilizzare **NoiseReduce,** un algoritmo di tipo DSP basato su **gating spettrale.** Ho analizzato l’audio a piccole finestre: quando una frequenza era vicina al livello del rumore la attenuavo, lasciando intatte le altre.

**Strategia applicata**: Il mio obiettivo era ridurre il fruscio caratteristico di molte tracce MusicGen concentrato soprattutto nelle bande sopra i 6 kHz, senza alterare le frequenze più basse. Per farlo ho seguito tre passaggi:

- Ho isolato le frequenze ≥ 6 kHz con un filtro passa-alto.
- Ho applicato NoiseReduce con parametri leggeri **solo** su quella banda.
- Ho ricombinato la banda pulita con il resto del brano originale.

In questo modo ho ridotto il rumore in maniera mirata, senza modificare il timbro complessivo del brano.

**Valutazione dei risultati:** Dall’analisi delle metriche ho visto che tutte le tracce hanno superato i tre criteri (100% OK) dimostrando la validità di questo modello.

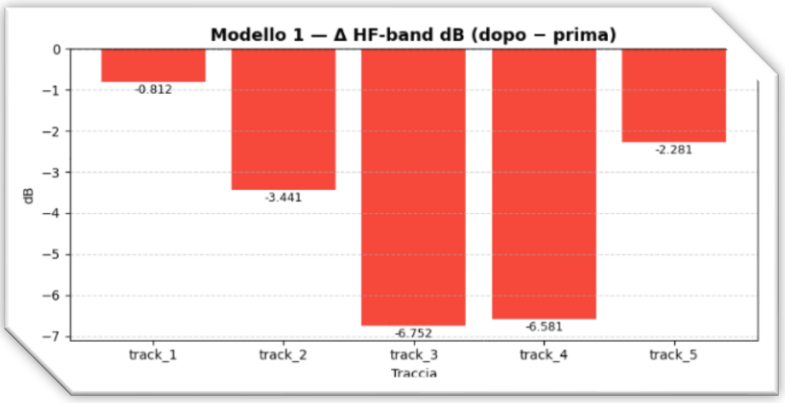
In particolare:

- L’energia nelle alte frequenze è diminuita in modo costante (**Δ HF-band dB < 0**), indicando che il fruscio è stato ridotto.
- Il **Δ HPSS ratio** è rimasto sempre vicino allo zero o leggermente positivo: mostrandomi che le componenti armoniche sono state preservate e che non si sono introdotti squilibri musicali.

- Infine, la **Log-Mel distance** si è mantenuta sotto la soglia di riferimento (0.20), con valori medi intorno a 0.17. Questo risultato mi conferma che la timbrica del brano è rimasta coerente con l'originale, senza introdurre cambiamenti percettibili.

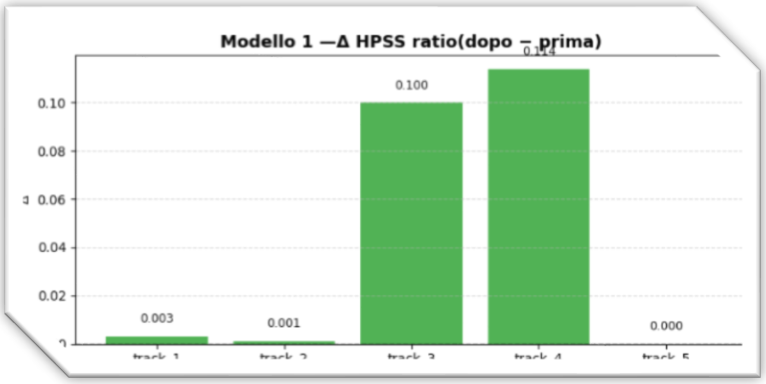
**Tabella 1:**

| | track | before_hf_band_db | after_hf_band_db | delta_hf_band_db | before_hpss_ratio | after_hpss_ratio | delta_hpss_ratio | log_mel_distance | OK_HFband | OK_HPSS | OK_LogMel | OK_count | criteri_ok | Verdetto |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | track_1 | 35.113 | 34.301 | -0.812 | 5.322 | 5.326 | 0.003 | 0.168 | True | True | True | 3 | 3/3 | OK |
| 1 | track_2 | -1.968 | -5.409 | -3.441 | 105.813 | 105.814 | 0.001 | 0.169 | True | True | True | 3 | 3/3 | OK |
| 2 | track_3 | 30.753 | 24.000 | -6.752 | 12.364 | 12.464 | 0.100 | 0.177 | True | True | True | 3 | 3/3 | OK |
| 3 | track_4 | 25.169 | 18.588 | -6.581 | 143.980 | 144.094 | 0.114 | 0.179 | True | True | True | 3 | 3/3 | OK |
| 4 | track_5 | -7.487 | -9.768 | -2.281 | 92.600 | 92.600 | 0.000 | 0.174 | True | True | True | 3 | 3/3 | OK |



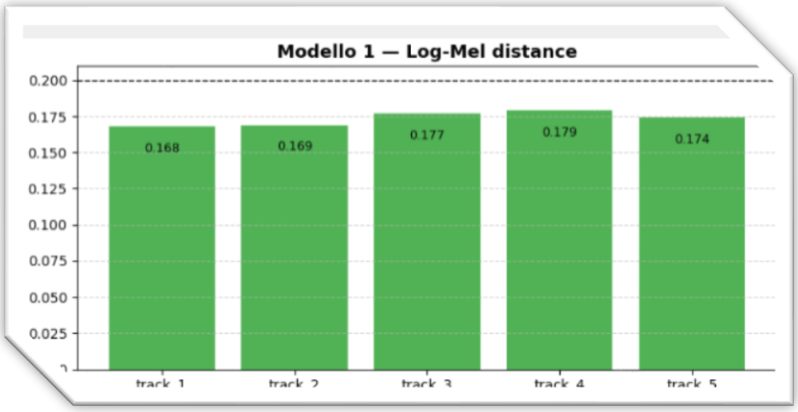

**Δ HF-band dB**
**- Rosso** = Δ < 0 (attenuazione HF → fruscio ridotto)
**- Verde** = Δ ≥ 0 (nessuna riduzione / aumento HF)

**Figura 1:** vedo che le tracce 3 e 4 ottengono i miglioramenti più significativi (oltre −6 dB), mentre le tracce 1 e 5 evidenziano differenze più piccole.




**Δ HPSS ratio**
**- Verde** = Δ ≥ 0 (armoniche stabili/valorizzate)
**- Rosso** = Δ < 0 (armoniche penalizzate)

**Figura 2:** vedo che viene confermata la stabilità delle armoniche, con piccoli incrementi positivi nelle tracce 3 e 4, quindi la musicalità viene preservata.




**Log-Mel distance**
**- Verde** = < 0.20 (variazione timbrica contenuta)
**- Rosso** = ≥ 0.20 (oltre baseline, variazione più marcata)

**Figura 3:** tutti i valori si mantengono sotto la soglia di riferimento (TH_LOGMEL=0.20), segno che non si verificano alterazioni rilevanti nelle tracce.

**Commento personale:** All'ascolto ho percepito miglioramenti soprattutto nelle tracce più rumorose, in particolare la 3 e la 4, dove il fruscio si riduce sensibilmente, pur non scomparendo del tutto, rendendo l'audio più pulito e gradevole. Nelle tracce che erano già abbastanza pulite, invece, le differenze sono minime o quasi impercettibili: considero questo un aspetto positivo, perché significa che il modello interviene solo dove serve, evitando modifiche inutili. Nel complesso considero

questo approccio efficace, capace di migliorare la qualità percepita senza compromettere la naturalezza e l'equilibrio musicale.

**Modello 2**: **High-Band Spectral Subtraction (HB-SS)**

In questo secondo modello ho **usato la sottrazione spettrale** mirata alle alte frequenze (≥6 kHz). La mia idea era che il fruscio prodotto da MusicGen fosse soprattutto stazionario e concentrato in quella banda; quindi, non aveva senso intervenire sulle frequenze medio-basse dove si trovano le armoniche principali.

**Strategia applicata:** Ho calcolato lo spettrogramma a breve termine (STFT), stimato il profilo medio del rumore *N(k)* dai frame più silenziosi e per ogni frequenza ≥ 6 kHz ho applicato la formula di sottrazione spettrale:

$$|\mathbf{Y}(\mathbf{k},\mathbf{t})| = \mathbf{max}(|\mathbf{X}(\mathbf{k},\mathbf{t})| - \boldsymbol{\alpha}\cdot\mathbf{N}(\mathbf{k}), \boldsymbol{\beta}\cdot|\mathbf{X}(\mathbf{k},\mathbf{t})|)$$

con parametri: $\alpha = 1.2$ (oversubtraction, rimozione più aggressiva del rumore) e $\beta = 0.04$ (spectral floor, evita la comparsa di musical noise). Infine, ho ricostruito l'audio usando la fase originale.

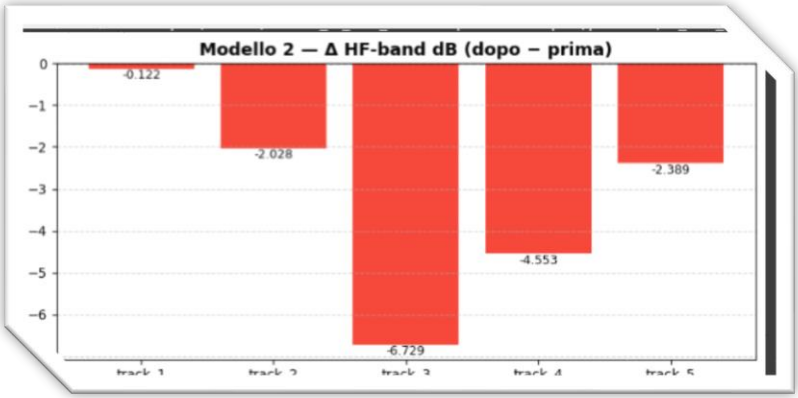
**Valutazione dei risultati:** Dalla tabella noto che:

- il **Δ HF-band dB** risulta sempre negativo, segno che il fruscio è stato attenuato in modo sistematico;
- Il **Δ HPSS ratio** rimane sostanzialmente invariato o leggermente positivo, a conferma che le componenti armoniche sono state preservate;
- La **Log-Mel distance** si colloca tra 0.15 e 0.22 (media ≈ 0.17), mantenendo quindi una buona fedeltà timbrica; questo fa sì che non tutte le clip soddisfano i 3 criteri contemporaneamente (≈ 40% OK).

In generale posso dire che il modello funziona bene, ma con qualche compromesso sulla coerenza timbrica rispetto al primo modello.

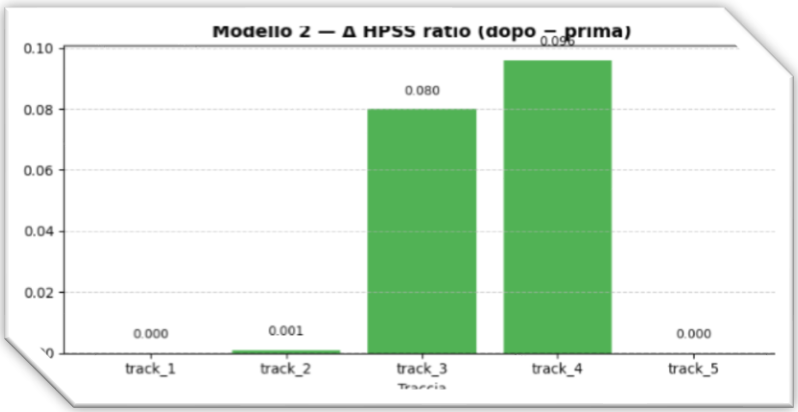
**Tabella 2:**

| | track | before_hf_band_db | after_hf_band_db | delta_hf_band_db | before_hpss_ratio | after_hpss_ratio | delta_hpss_ratio | log_mel_distance | OK_HFband | OK_HPSS | OK_LogMel | OK_count | criteri_ok | Verdetto |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | track_1 | 35.113 | 34.991 | -0.122 | 5.322 | 5.323 | 0.000 | 0.229 | True | True | False | 2 | 2/3 | OK |
| 1 | track_2 | -1.968 | -3.997 | -2.028 | 105.813 | 105.814 | 0.001 | 0.159 | True | True | True | 3 | 3/3 | OK |
| 2 | track_3 | 30.753 | 24.024 | -6.729 | 12.364 | 12.445 | 0.080 | 0.228 | True | True | False | 2 | 2/3 | OK |
| 3 | track_4 | 25.169 | 20.616 | -4.553 | 143.980 | 144.076 | 0.096 | 0.215 | True | True | False | 2 | 2/3 | OK |
| 4 | track_5 | -7.487 | -9.876 | -2.389 | 92.600 | 92.600 | 0.000 | 0.182 | True | True | True | 3 | 3/3 | OK |

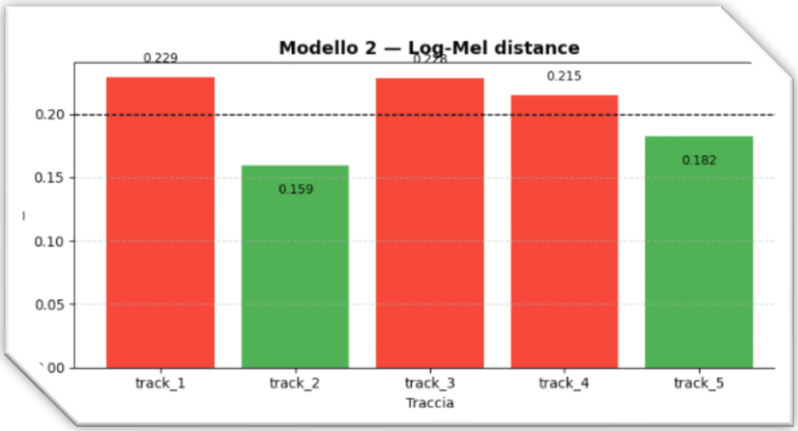



**Δ HF-band dB**
**- Rosso** = Δ < 0 (attenuazione HF → fruscio ridotto)
**- Verde** = Δ ≥ 0 (nessuna riduzione / aumento HF)

**Figura 1:** Osservo che l'andamento del **Δ HF-band dB** è sempre negativo e più marcato nelle tracce 3 e 4: confermando che il modello interviene in modo mirato dove il fruscio è più presente.




**Δ HPSS ratio**
**- Verde** = Δ ≥ 0 (armoniche stabili/valorizzate)
**- Rosso** = Δ < 0 (armoniche penalizzate)
**Figura 2:** Noto che il Δ HPSS ratio resta vicino allo zero. Questo andamento mi conferma che l'equilibrio tra parte armonica e percussiva è stato mantenuto stabile e in alcuni casi con un leggero miglioramento.




**Log-Mel distance**
**- Verde** = < 0.20 (variazione timbrica contenuta)
**- Rosso** = ≥ 0.20 (oltre baseline, variazione più marcata)

**Figura 3:** Vedo che la Log-Mel distance rimane bassa e abbastanza uniforme: i cambiamenti timbrici sono minimi e, anche se in alcuni casi supera leggermente la soglia (che uso solo come riferimento comparativo), considero comunque il risultato soddisfacente.

**Commento personale:** Nelle tracce più rumorose, come la 3 e in parte la 4 ho percepito un controllo più evidente del fruscio: l'intervento è percepibile, anche se non elimina completamente il rumore. Nelle tracce già abbastanza pulite, come la 1 e la 5, l'azione del modello è quasi impercettibile. Questo mi conferma che il metodo è trasparente: riduce il fruscio senza introdurre artefatti e lascia intatta la brillantezza del timbro. Rispetto al Modello 1, sento che questo metodo è meno aggressivo: non interviene in modo netto, ma lavora con più delicatezza, riuscendo comunque a ridurre il fruscio e a preservare la naturalezza e la timbrica originale.

**Modello 3: MCRA (Minimum Controlled Recursive Averaging) + filtro HF "safe"**

Per il terzo modello ho scelto di usare l'algoritmo **MCRA**.Questo modello mi permette di stimare in tempo reale lo spettro del rumore e di applicare un filtro di Wiener, il quale attenua in modo dinamico le bande più rumorose lasciando intatte le componenti musicali. A differenza del Modello 1, che agiva solo sulle alte frequenze, qui l'intervento è **full band,** cioè su tutto lo spettro, con alcune attenzioni per mantenere la naturalezza del suono.

**Strategia applicata:**

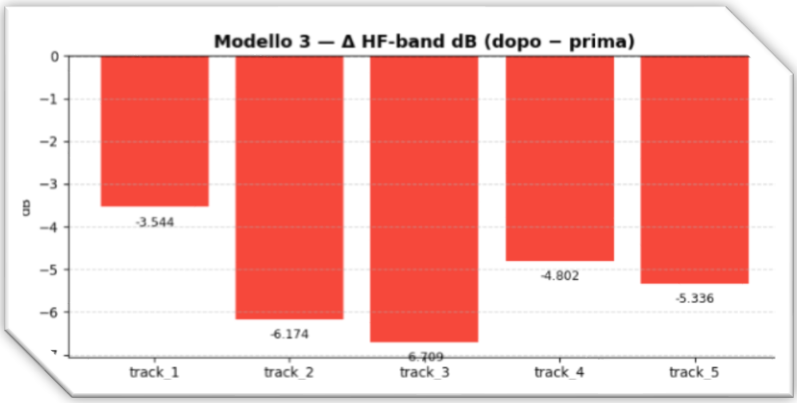
- Ho stimato lo spettro di potenza con STFT e ricavato il profilo del rumore con MCRA;
- ho applicato un filtro di Wiener per attenuare le zone più rumorose;
- ho impostato un guadagno minimo (*g_floor*) per evitare che alcune bande venissero annullate.

**Valutazione dei risultati:** Dalla tabella noto che:

- Il **Δ HF-band dB** è sempre negativo, segno di una riduzione costante del fruscio.
- Il **Δ HPSS ratio** è sempre positivo: le armoniche sono preservate e talvolta persino valorizzate.
- **La Log-Mel distance** rappresenta il punto critico: 4 tracce restano tra 0.43–0.49, mentre la traccia 1 sale a 0.79, superando la soglia di 0.20.

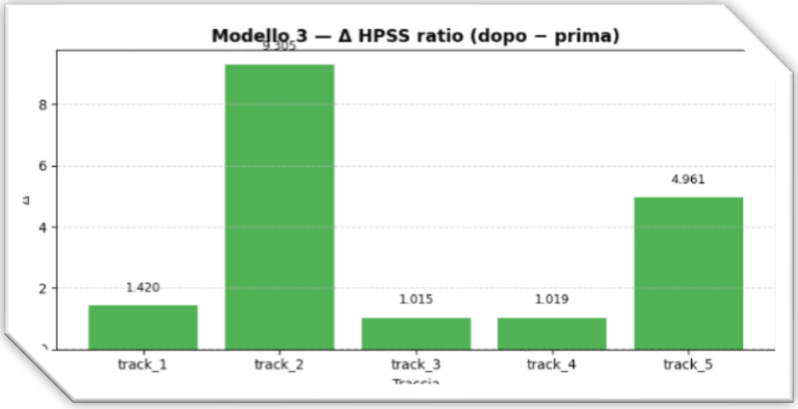
**Tabella 3:**

| | track | before_hf_band_db | after_hf_band_db | delta_hf_band_db | before_hpss_ratio | after_hpss_ratio | delta_hpss_ratio | log_mel_distance | OK_HFband | OK_HPSS | OK_LogMel | OK_count | criteri_ok | Verdetto |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | track_1 | 35.113 | 31.570 | -3.544 | 5.322 | 6.742 | 1.420 | 0.796 | True | True | False | 2 | 2/3 | OK |
| 1 | track_2 | -1.968 | -8.142 | -6.174 | 105.813 | 115.118 | 9.305 | 0.438 | True | True | False | 2 | 2/3 | OK |
| 2 | track_3 | 30.753 | 24.044 | -6.709 | 12.364 | 13.379 | 1.015 | 0.449 | True | True | False | 2 | 2/3 | OK |
| 3 | track_4 | 25.169 | 20.368 | -4.802 | 143.980 | 144.998 | 1.019 | 0.428 | True | True | False | 2 | 2/3 | OK |
| 4 | track_5 | -7.487 | -12.823 | -5.336 | 92.600 | 97.561 | 4.961 | 0.489 | True | True | False | 2 | 2/3 | OK |



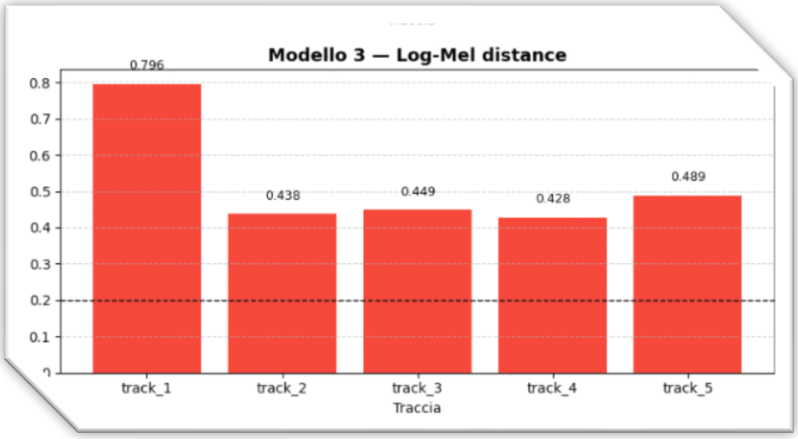

**Δ HF-band dB**
**- Rosso** = Δ < 0 (attenuazione HF → fruscio ridotto)
**- Verde** = Δ ≥ 0 (nessuna riduzione / aumento HF)

**Figura 1**: vedo che i valori sono sempre negativi, con attenuazioni comprese tra −3.5 e −6.7 dB. Questo conferma che il modello ha agito in modo deciso e uniforme, riducendo il fruscio su tutte le tracce.




**Δ HPSS ratio**
**- Verde** = Δ ≥ 0 (armoniche stabili/valorizzate)
**- Rosso** = Δ < 0 (armoniche penalizzate)

**Figura 2**: Osservo valori sempre positivi, con un picco nella traccia 2 (+9.3). Questo mi conferma che le armoniche sono state preservate e in certi casi risultano persino più evidenti rispetto al rumore.




**Log-Mel distance**
**- Verde** = < 0.20 (variazione timbrica contenuta)
**- Rosso** = ≥ 0.20 (oltre baseline, variazione più marcata)

**Figura 3:** vedo che quasi tutte le tracce restano tra 0.43 e 0.49, mentre la traccia 1 (≈0.80) indica una variazione timbrica più marcata.

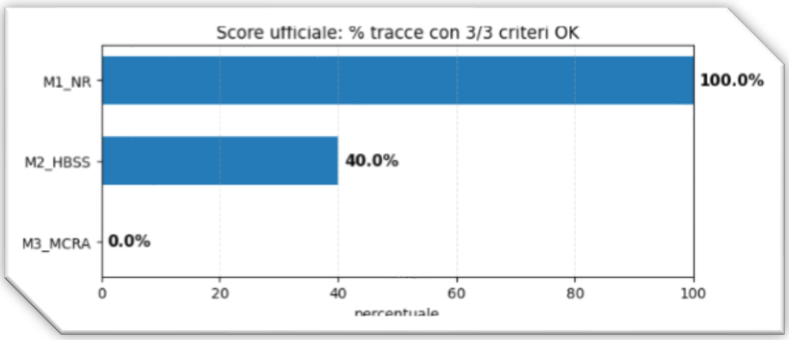
**Commento personale:** All'ascolto, anche nella traccia con Log-Mel distance elevata non percepisco peggioramenti: il cambiamento è poco evidente e non introduce anomalie fastidiose. Nelle tracce più rumorose il fruscio risulta meglio controllato, mentre nei brani già puliti la differenza è quasi nulla. Il modello riduce il rumore in modo costante e deciso (Δ HF-band sempre negativo), con armoniche preservate o persino valorizzate (Δ HPSS positivo). Il limite principale è la Log-Mel distance, che in alcune tracce cresce sensibilmente, indicando variazioni timbriche più marcate, anche se non penalizzanti all'ascolto. Nel complesso lo considero efficace e robusto, adatto quando serve un intervento full band deciso, accettando però un minore livello di trasparenza rispetto agli altri due modelli.

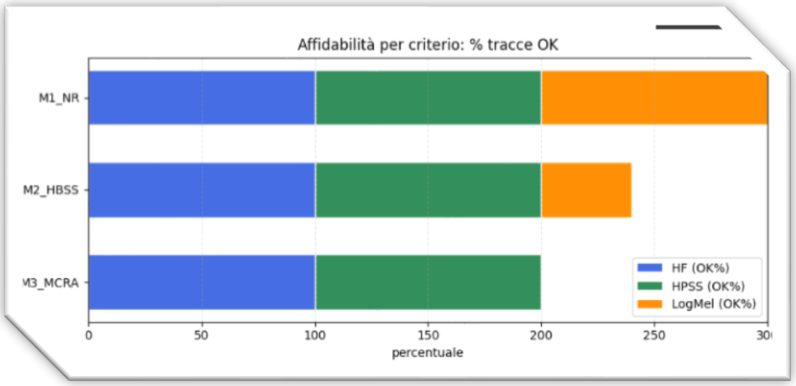
**Conclusione:**

Il mio obiettivo era ridurre il fruscio tipico delle tracce generate da MusicGen mantenendo naturalezza e qualità del suono. Dai risultati posso dire di esserci riuscita: tutti e tre i modelli hanno ridotto l'energia nelle alte frequenze senza compromettere le componenti armoniche, anche se con differenze nel livello di trasparenza e fedeltà timbrica.

Dai grafici emergono differenze chiare tra i tre modelli:

- Il **Modello 1 (NR)** il più equilibrato, con il **100%** delle tracce che rispettano tutti i criteri.
- Il **Modello 2 (HBSS)** è efficace quando il rumore è concentrato sulle alte frequenze, ma meno costante (≈40% OK).
- Il **Modello 3 (MCRA** riduce il rumore in modo deciso su tutto lo spettro, ma sacrifica la trasparenza timbrica, fallendo sempre sulla Log-Mel.




Nel complesso considero i tre modelli strumenti complementari, ciascuno con pregi e limiti. Dall'analisi dei risultati, però, il **Modello 1** per me è stato il più efficiente: ha ridotto in modo costante il fruscio senza introdurre alterazioni timbriche, risultando la soluzione più bilanciata. Il **Modello 2** e il **Modello 3** li vedo invece come opzioni utili in casi specifici, anche se meno trasparenti.

Con questo progetto ho potuto rispondere alla domanda che mi ero posta all'inizio. La mia risposta è **sì**: è possibile ridurre il fruscio tipico di MusicGen senza perdere musicalità. Ci sono riuscita soprattutto grazie ad approcci DSP leggeri, che si sono rivelati semplici da applicare ed efficaci, molto più dei modelli neurali che avevo testato in partenza.