Generalising AI-Music Detection: Decoder Artefacts, Self-Supervised Features, and the Limits of Hand-Crafted Cues

A follow-up to "Detecting AI-Generated Piano Music with a Spectrogram CNN: A Proof of Concept, and a Study in Shortcuts."

Daniel Bordovský[1]

Abstract

Our previous study showed that a spectrogram CNN can distinguish one generator's piano output from human recordings with high in-distribution confidence, but fails to generalise: it collapsed on an unseen generator and on an unseen genre, and part of its signal was traceable to data-collection confounds. This follow-up turns from that single working classifier to the broader question it raised — *how can AI-music detection be made to generalise across generators?* — and reports a small empirical study of the most-discussed answer: detecting the artefacts left by the neural decoders that all current generators use to render audio. We first survey the candidate approaches (supervised spectrogram classification, self-supervised foundation features, neural-decoder artefact detection, musical-structure analysis, and watermarking) with their respective strengths and limitations. We then test, on a multi-generator dataset spanning seven systems (AudioLDM, MusicGen, Mustango, Riffusion, Stable Audio, Suno, Udio) plus real music, whether the decoder artefact can be captured by a simple, training-free signal-processing feature. A naive spectral-periodicity ("comb-strength") feature failed (AUC 0.37): it measures musical harmonicity, which real instrumental music has in abundance. A temporal-stationarity feature — exploiting that decoder artefacts are frozen in frequency while music moves — recovered a real but modest signal (AUC 0.68). Crucially, it separated codec-based generators (AudioLDM, MusicGen) well but was blind to the two most prominent commercial systems, Suno and Udio, whose polished output suppresses the artefact. We conclude that the decoder-artefact route is the most principled path to generator-agnostic detection, but that the artefact in today's best generators is too subtle for hand-crafted features and requires a learned model (e.g. the autoencoder round-trip of Afchar et al.). No single approach is a silver bullet; a practical detector is an ensemble, and the field is an arms race.

[1]Independent Researcher. Correspondence to: borrtex@borrtex.com

1. Introduction

In our previous work we built a spectrogram-CNN detector for AI-generated piano and found, after removing confounds, that it worked well *in distribution* but did not generalise: it misclassified an unseen generator and defaulted to "real" on an unseen genre. We argued the core unsolved problem is generalisation, a detector trained on one generator's idiosyncrasies learns *that generator*, and the vendor can change those idiosyncrasies at will.

This follow-up asks how that problem might be escaped, and reports an empirical test of the most promising idea. The structure is: a survey of candidate detection strategies (Section 2), the hypothesis that all current generators share a *neural-decoder* fingerprint (Section 3), our experiments testing whether that fingerprint can be caught by simple features (Section 4), and a comparative discussion of where each approach helps and where it breaks (Section 5). As with the first study, this is exploratory single-operator work on small samples; the contribution is conceptual and the honest mapping of what does and does not work, not a production system.

2. The space of detection strategies

We distinguish five families.

(a) Supervised spectrogram (or waveform) classification. Train a network to separate AI from human examples directly; our previous approach. *Strength:* simple, strong in-distribution, needs no special knowledge of the generators. *Weakness:* learns generator and genre-specific cues; generalises poorly; vulnerable to confounds (loudness, bandwidth, codec) unless carefully controlled.

(b) Self-supervised foundation features. Replace the from-scratch network with a pretrained audio model (MERT for music; wav2vec 2.0 / HuBERT for speech) as a frozen feature extractor, training only a small classifier on top. *Strength:* richer representations, far better generalisation and data efficiency. *Weakness:* it remains a learned discriminator bounded by its training distribution, so while it may generalise better than a from-scratch CNN, it still degrades on a genuinely unseen generator, inheriting the same out-of-distribution wall rather than escaping it; still ultimately a learned discriminator that can latch onto distributional cues.

(c) Neural-decoder artefact detection. Target the computational fingerprint left by the neural decoders all current generators use to turn latent representations into waveforms. This is the focus of this paper and is developed in Section 3.

(d) Musical-structure analysis. Detect AI not by surface texture but by higher-level musical behaviour: timing regularity, harmonic logic, instrument interaction, long-range coherence. *Strength:* most robust in principle, because it does not depend on a rendering artefact a vendor can remove. *Weakness:* the hardest to build; requires music-information-retrieval features.

(e) Watermarking and provenance. Rely on signals the generator embeds (e.g. inaudible watermarks) or content-provenance metadata. *Strength:* near-perfect when present and intact. *Weakness:* only works with cooperating generators, can be stripped or degraded by editing, and does nothing for adversarial or older content.

3. The neural-decoder artefact hypothesis

The appeal of approach (c) is that it attacks generalisation at its root. Whatever their differences in architecture, training data, genre, or version, current generative music systems share a final step: a neural decoder that upsamples a compact representation into a waveform, typically through *transposed-convolution* ("deconvolution") layers. Recent work (Afchar et al., Deezer Research[2]) proves that such layers necessarily produce systematic, periodic spectral peaks — the audio analogue of the "checkerboard" artefact familiar from AI image forensics — and that this artefact is a property of the *architecture*, not of the weights or training data.

This is a stronger invariance than anything available to a supervised classifier. A supervised detector keyed to a generator's mastering or bandwidth fails when the vendor retrains or re-masters, a *weights* change. The deconvolution artefact survives weight changes, fine-tuning, genre shifts and version bumps; only a change to the *upsampling architecture itself* (e.g. anti-aliased upsampling, or waveform-domain diffusion) removes it. The artefact is therefore the most generator-agnostic cue currently known.

Two ways to exploit it have been proposed. The direct route detects the spectral-peak signature itself. The learned round-trip route (the published Deezer method[3]) auto-encodes real music through several neural codecs and trains a classifier to distinguish originals from their reconstructions; because the two differ only in the decoder artefact, the model learns the artefact without any confound, and importantly a model trained only on reconstructions then flags fully prompt-generated music. The round-trip needs no AI training data at all, only real music and a few codecs.

[2]Afchar, D., Meseguer-Brocal, G., Akesbi, K., & Hennequin, R. (2025). A Fourier Explanation of AI-music Artifacts. Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR), Daejeon, South Korea.

[3]Afchar, D., Meseguer-Brocal, G., & Hennequin, R. (2025). AI-Generated Music Detection and its Challenges. IEEE ICASSP 2025.

The limitations of the whole route should be stated up front: a generator that abandons transposed-convolution upsampling escapes it; legitimate audio that has passed through a neural codec (increasingly common in streaming and telephony) can false-positive; and it detects *rendering by a neural decoder*, which is not identical to *authorship by an AI*. Within those bounds, however, it is the most principled basis for a generalisable detector.

4. Can the artefact be caught by a simple feature?

We asked the practical question for a resource-limited builder: can the decoder artefact be detected by a cheap, training-free signal-processing feature, rather than the round-trip's learned model?

4.1 Data

We used the AIME dataset (lossless audio from twelve generators plus real music), drawing a small per-generator sample for fast iteration: seven generators present in our subset: AudioLDM, MusicGen, Mustango, Riffusion, Stable Audio, Suno, Udio, with roughly eight to ten tracks each (64 AI tracks in total), against 60 real tracks. Audio was analysed at 44.1 kHz with a long FFT (8192–16384) to resolve narrow peaks. Sample sizes are small and results are directional, not definitive.

4.2 The artefact is visible, when generators are not pooled

In the long-term average spectrum of a *single* generator versus real music, the AI curve is visibly "hairier" carrying additional narrow spectral peaks, and a baseline-removed view exposes them as spikes that the real curve lacks. Pooling several generators, however, smears the signal: each decoder places its comb at *different* frequencies, so averaging many generators blurs many combs into general hairiness. This already implies that any detector keyed to *specific* peak frequencies is generator-specific; a generalisable feature must detect the *presence* of a comb, not its location.

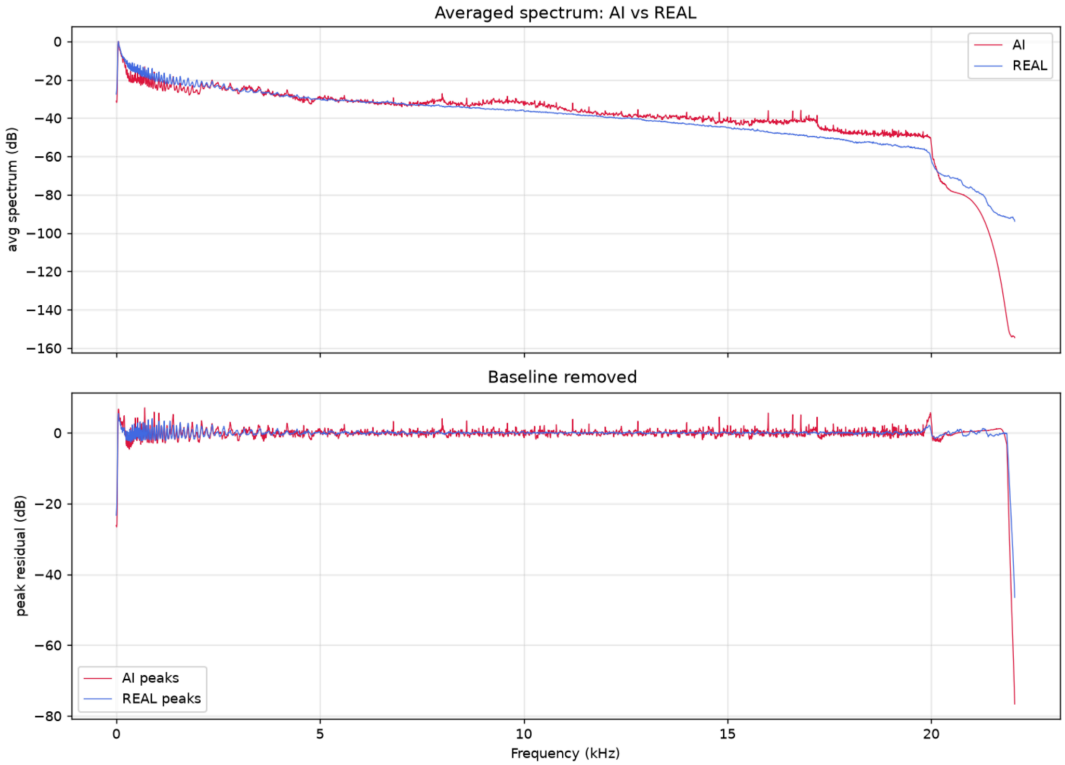

Figure 1. Long-term averaged spectrum of AI versus real music (top), and the same with the smooth baseline removed to expose narrow peaks (bottom).

4.3 Experiment 1 — spectral periodicity

To detect if there is a regular comb at any spacing, we computed the autocorrelation of the baseline-removed average spectrum per track, taking the strongest periodic component as a score. **This failed: AUC 0.37** (below chance; real music scored *higher* than pooled AI, 0.344 vs 0.308). The reason is that musical harmonics *are* a regular comb; every note's overtones are evenly spaced by its fundamental, so tonal instrumental music produces strong spectral periodicity of its own, swamping the faint decoder comb. The feature measured harmonicity, which real music has in abundance.

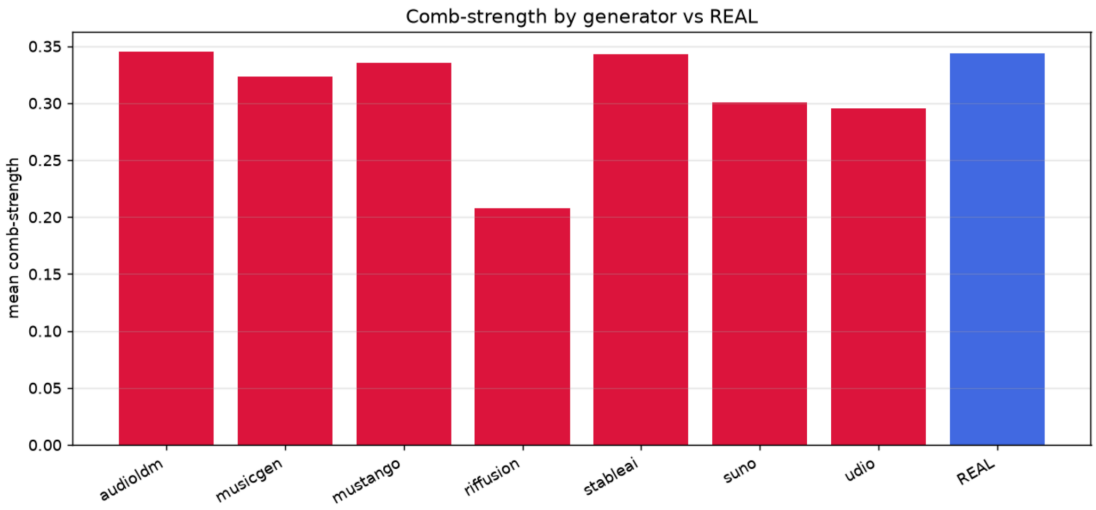

Figure 2. Comb-strength by generator versus real music. Each bar is the mean "comb-strength" score, the strength of the strongest periodic component in a track's baseline-removed average spectrum, found by autocorrelation, averaged over that generator's tracks (red) or the real set (blue). The idea: a decoder comb is a regularly-spaced pattern, so a high score should flag AI regardless of where the comb sits. For the feature to work, every red bar would have to stand clearly above the blue one. It doesn't.

4.4 Experiment 2 — temporal stationarity

The flaw in Experiment 1 was averaging over time, which discards the one property that distinguishes the artefact from music: a decoder comb sits at the *same* frequencies in *every* frame, whereas musical harmonics move as the notes change. We therefore measured **persistent** spectral lines — taking, per frequency bin, the low-percentile energy over time (the energy present in nearly every frame), removing the smooth frequency baseline to isolate narrow lines, and scoring their strength in a mid band (10–18 kHz, chosen below both classes' bandwidth cutoffs so the brightness/bandwidth envelope could not confound the result).

This recovered a real but modest signal: AUC 0.68. A usable generic detector would need to sit far higher, and the pooled number actually understates the problem, because it averages two opposite behaviours rather than a uniform mediocrity.

Per-generator mean scores (real = 1.56) were: AudioLDM 4.20, MusicGen 3.36, Stable Audio 2.18, Riffusion 2.16, Mustango 2.13, Suno 1.61, Udio 0.80.

The pattern is the important result. The feature cleanly separates the codec-based generators (AudioLDM and MusicGen, which use Encodec/DAC-style decoders) from real music, exactly as the artefact theory predicts. But it is blind to the two most prominent commercial systems: Suno scores at the real-music level and Udio *below* it, ranking it inverted as "more real" than real music. These are the highest-quality, most-trained generators, and their decoder artefacts are correspondingly the most suppressed.

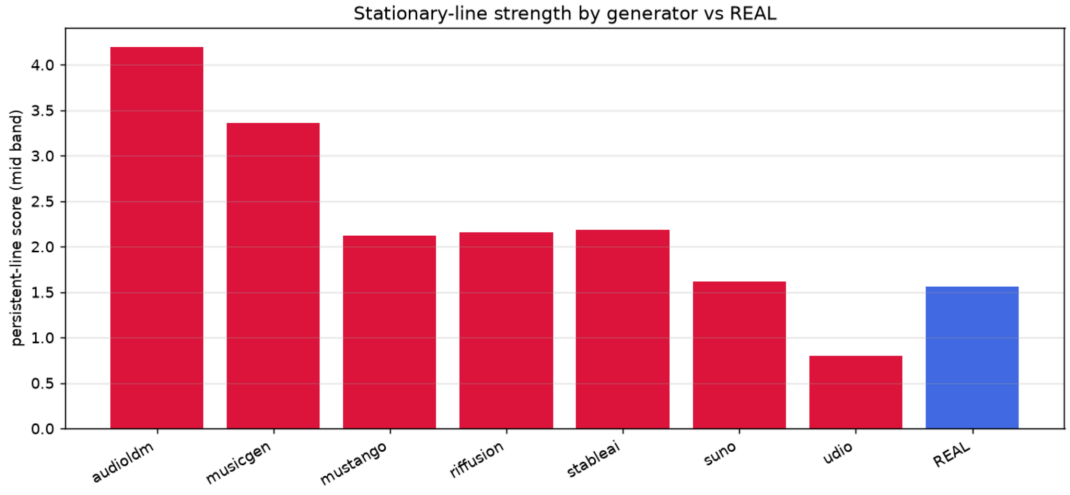

Figure 3. Stationary-line strength by generator versus real (blue). Each bar is the mean strength of persistent, narrow spectral lines in the 10–18 kHz band: the "frozen comb" signature.

4.5 Interpretation

Across two attempts the hand-crafted route topped out near AUC 0.68, strong on codec-based generators and blind on the polished commercial ones. The decoder artefact is genuinely present and detectable in principle — Experiment 2 confirms it is not merely the spectral envelope, but in the systems that matter most commercially it is too faint for a single statistic to isolate from the music. This is precisely the gap that motivates the *learned* round-trip: a trained model can integrate many weak, subtle cues that no closed-form feature captures, which is why the published state of the art trains a network rather than thresholding a spectrum.

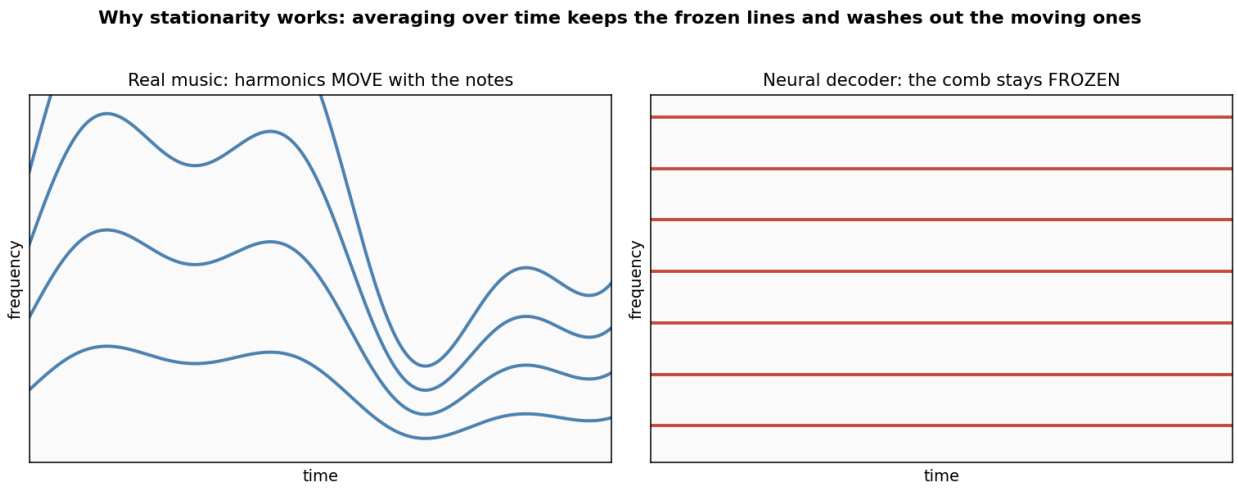

Figure 4. Schematic, not real data; musical harmonics (left) move with the notes, so a given frequency's energy comes and goes, a neural-decoder comb (right) stays at the same frequencies in every frame. Measuring the energy present in nearly every frame therefore keeps the frozen lines and washes out the moving ones, the idea behind Experiment 2, which a plain spectral average (Experiment 1) could not exploit.

5. Discussion

There seems to be a persistent trade-off between in-distribution strength and generalisation: the supervised classifier is the strongest on generators it has seen and the weakest on those it has not, while the decoder-artefact route inverts this — generic by design but weak on any specific high-quality system. Second, every approach has an escape hatch for the generator: re-mastering defeats the spectral classifier, architecture changes defeat the artefact detector, better fidelity

defeats both, and only watermarking (which the generator must volunteer) is exact. AI-music detection is structurally an arms race.

For a single builder, the pragmatic reading is therefore an ensemble rather than a single perfect lens: a supervised model for strong in-distribution accuracy on known generators, combined with a decoder-artefact signal for some generalisation to unseen ones, with honest uncertainty when the lenses disagree or when an input is unlike anything seen in training. Our results temper expectations for the cheap version of the second lens: a hand-crafted artefact feature contributes most for codec-based generators and little for Suno/Udio, so the supervised model remains the workhorse for the systems most people actually want to detect.

6. Limitations

These are small, exploratory experiments: roughly eight to ten tracks per generator and sixty real tracks, so the AUC figures are directional and the per-generator means are noisy. The AI and real corpora differ in content and provenance, so some separation may reflect dataset differences rather than synthesis, as in the previous study. The two hand-crafted features are deliberately simple; a richer feature set or a learned model could exceed AUC 0.68, and we did not implement the round-trip here. The "real" set again includes sample-library productions, so the contrast remains generative-AI synthesis versus real-acoustic source. We did not evaluate against neural-codec-processed real audio, which the artefact route would be expected to mis-flag.

7. Conclusion

The generalisation failure documented in our first study has a principled candidate solution — detecting the neural-decoder artefact shared by all current generators — and we set out to test whether a single enthusiast could exploit it cheaply. The artefact is real and, with the right feature (temporal stationarity rather than naive periodicity), partially detectable without any training. But it is modest, and in the polished commercial generators that dominate real-world use it is suppressed below what a hand-crafted statistic can reliably catch. The honest conclusion is twofold: the decoder-artefact route is the most promising direction for generator-agnostic detection and deserves the learned, round-trip treatment that the published state of the art gives it; and in the meantime, a supervised classifier remains the dependable core, with artefact and foundation-feature lenses as complementary, partial generalisation aids.

Notes. Experiments used librosa for spectral analysis and numpy for the features; audio came from the AIME dataset and real-music collections (FMA, MTG-Jamendo). AUC values and per-generator means come from small single-run samples and are reported as approximate. The neural-decoder artefact theory and round-trip method are due to Afchar, Meseguer-Brocal and Hennequin (Deezer Research); see "Detecting music deepfakes is easy but actually hard" and "A Fourier Explanation of AI-music Artifacts."

References

[1] Afchar, D., Meseguer-Brocal, G., Akesbi, K., & Hennequin, R. (2025). A Fourier Explanation of AI-music Artifacts. Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR), Daejeon, South Korea. https://arxiv.org/abs/2506.19108

[2] Afchar, D., Meseguer-Brocal, G., & Hennequin, R. (2025). AI-Generated Music Detection and its Challenges. IEEE ICASSP 2025. https://arxiv.org/abs/2501.10111