

MRS: A Multi-Dimensional Reference-Distance Metric for Evaluating the Technical Realism of AI-Generated Music

Xuanzhen Zhu

Moodify Sound Laboratory, Rongjing Wenchuan (Shenzhen) Technology Co., Ltd., Shenzhen, China

Intended venue: Journal of the Audio Engineering Society; manuscript type: Research Paper draft for internal review and experimental completion

Abstract—Objective evaluation of generated audio commonly relies on paired-reference perceptual models or distribution-level embedding distances. Music post-processing, however, requires a track-level measure that is interpretable enough to identify whether a candidate output deviates in spectrum, dynamics, transients, stereo space, texture, temporal stability, or processing artifacts. This paper presents the Moodify Reality Score (MRS), a reference-conditioned metric implemented as 29 handcrafted features organized into seven acoustic groups. Each feature is standardized against a curated reference distribution; group distances are aggregated into a nonnegative primary distance, D_R , and optionally transformed into a bounded presentation score, $100e^{-D_R}$. The paper defines the metric semantics, mathematical properties, implementation details, and a controlled-degradation validation protocol covering spectral, dynamic, spatial, and temporal defects. A code audit identifies current limitations, including small reference sets in the smoke-test script, provisional feature weights, incomplete perceptual calibration, and a short-clip schema inconsistency. MRS is therefore proposed as an explainable engineering metric and research hypothesis, not as a validated universal measure of musical quality. A formal validation plan combining controlled distortions, MUSHRA-style listening tests, ablation analysis, and comparisons with FAD, KAD, PEAQ, and ViSQOLAudio is specified.

Keywords—AI-generated music; audio quality assessment; explainable metrics; music information retrieval; objective evaluation

1 Introduction

Generative music systems can produce complete songs at low marginal cost, but the technical condition of a rendered full mix remains variable. A recording may contain excessive high-frequency hardness, weak low-frequency support, flattened dynamics, unstable stereo width, inconsistent later sections, clipping, or narrow spectral spikes. These defects are not equivalent to compositional failure: a musically meaningful song may still require post-processing before release.

Evaluation is difficult because “quality” contains several non-identical concepts. Signal fidelity describes difference from a paired clean reference. Perceptual quality predicts listener judgments under a specific test domain. Distributional realism compares sets of generated and reference audio. Production readiness concerns whether an individual track has controllable, interpretable technical defects. Artistic quality concerns intention, originality, emotion, and cultural meaning. A metric that conflates these concepts may become easy to report but difficult to trust.

This paper introduces the Moodify Reality Score (MRS), a track-level, non-paired, reference-conditioned metric for technical realism. Here, *technical realism* means proximity to a curated distribution of production-ready audio in a selected acoustic feature space. It does not mean that a recording was made by a human,

nor does it measure composition, originality, semantic plausibility, cultural value, or emotional truth.

The design goals are:

1. provide a decomposable distance rather than an unexplained scalar;
2. operate without a paired clean version of the same song;
3. expose which acoustic group contributes to a poor score;
4. support before/after comparison in post-processing experiments;
5. permit reference sets and weights to be recalibrated; and
6. remain explicit about the difference between implementation evidence and perceptual validation.

The contribution is a metric formulation, an open implementation in the Moodify repository, a set of mathematical and engineering properties, and a validation protocol. The paper does not claim that MRS presently predicts human preference better than established metrics.

2 Positioning Relative to Existing Metrics

2.1 Paired-Reference Perceptual Models

ITU-R BS.1387 specifies objective measurement of perceived audio quality and is commonly associated with PEAQ [1]. ViSQOLAudio similarly

compares a degraded signal with a reference and has been validated primarily for codec and transmission conditions [2, 3]. These methods are appropriate when a clean or original reference for the same content exists. AI-generated music post-processing often lacks such a reference: the generated full mix is the only available source.

2.2 Distribution-Level Metrics

Fréchet Audio Distance (FAD) compares distributions in an embedding space and was introduced for evaluating music enhancement algorithms [4]. Kernel Audio Distance (KAD) replaces the Gaussian assumption with a kernel-based discrepancy and reports improved sample efficiency and perceptual alignment in its validation setting [5]. These methods are valuable for comparing systems or datasets, but their aggregate nature and learned representations do not directly explain why one individual track is problematic.

2.3 Subjective Standards

Subjective assessment remains necessary when a metric claims perceptual validity. ITU-R BS.1534-3 defines MUSHRA for intermediate-quality audio systems, including assessor selection, training, stimulus presentation, and grading procedures [6]. Loudness matching and true-peak control should follow established measurement practice such as ITU-R BS.1770-5 and EBU R 128 [7, 8].

3 Metric Semantics and Design Requirements

3.1 Distance First, Score Second

The scientifically primary output is the nonnegative distance D_R . Lower values indicate proximity to the selected reference distribution. A bounded score may be generated for user interfaces, but both quantities should be reported to avoid semantic confusion:

$$\text{lower } D_R \text{ is better,} \quad \text{higher MRS is better.} \quad (1)$$

The distinction also preserves an open-ended research variable: D_R is not capped, whereas the presentation score lies in $(0, 100]$.

3.2 Reference-Conditioned, Not Universal

MRS is conditional on the reference corpus. A reference distribution built from acoustic, jazz, and orchestral masters will encode a different statistical center from one built from hypercompressed electronic music. The metric should

therefore report the reference-set identity, version, sample count, genre composition, and pre-processing policy. Style-conditioned references may be preferable to a single universal pool.

3.3 Track-Level Explanation

A global score is useful only if it can be decomposed. MRS reports seven component distances and their changes before and after processing. The largest reduction identifies the main measured gain; the largest increase identifies the principal penalty. These labels are diagnostic hypotheses, not causal proof.

4 Feature Architecture

The implementation uses 29 scored features grouped into seven dimensions. Additional raw values may be extracted for reporting, but only the features in Table 2 enter the current distance calculation.

4.1 Spectrum

For mono signal $x[n]$, the implementation applies a Hann window and computes the magnitude spectrum $X[k]$. Relative energy for band b is

$$E_b = \frac{\sum_{k \in \mathcal{B}_b} |X[k]|^2}{\sum_k |X[k]|^2 + \epsilon}. \quad (2)$$

The six bands are 20–60, 60–250, 250–500, 500–2000, 2000–5000, and 8000–16000 Hz. Spectral centroid and rolloff are normalized by Nyquist frequency. This normalization reduces sensitivity to global gain, although it does not remove reference-corpus bias or frequency-band design choices.

4.2 Dynamics and Transients

Crest factor is peak amplitude divided by RMS amplitude. Dynamic range is estimated from the 95th minus 5th percentile of RMS values in 100 ms windows. The transient group uses 10 ms frames, normalized spectral change between adjacent frames, and variation of short-time energy.

4.3 Space

Stereo descriptors are computed from left and right channels. Let

$$m[n] = \frac{l[n] + r[n]}{2}, \quad s[n] = \frac{l[n] - r[n]}{2}. \quad (3)$$

The implementation includes Pearson correlation, side-to-mid energy ratio, standard-deviation width ratio, and a penalty for negative correlation. Mono input is duplicated to two

Table 1: Conceptual positioning of MRS. “Reference” denotes the information required at evaluation time.

Method	Evaluation unit	Reference	Interpretability	Primary intended use
PEAQ	signal pair	paired clean signal	medium; psychoacoustic model variables	codec and impairment assessment
ViSQOLAudio	signal pair	paired clean signal	medium; patch similarity	codec and transmission quality
FAD	set versus set	unpaired reference distribution	low at track level	generative or enhancement-system comparison
KAD	set versus set	unpaired reference distribution	low at track level	sample-efficient generative-system comparison
MRS	single track versus statistics	unpaired, curated reference distribution	high; seven component distances	post-processing diagnosis and candidate comparison

Table 2: MRS feature groups in the current implementation.

Group	Count	Features	Default weight
Spectrum	9	normalized centroid, normalized 95% rolloff, flatness, and relative energy in sub, bass, low-mid, mid, presence, and air bands	0.20
Dynamics	3	crest factor, short-window dynamic range, short-time RMS standard deviation	0.15
Transients	3	mean spectral flux, spectral-flux deviation, normalized short-time energy change	0.10
Space	4	left-right correlation, mid-side energy ratio, stereo width, anti-correlation anomaly	0.15
Texture	3	local spectral roughness proxy, high-frequency smoothness, narrow spectral-spike density	0.20
Temporal	4	segment RMS variation, late-versus-early feature change, RMS drift, centroid drift	0.15
Artifacts	3	clipping ratio, abnormal-peak ratio, spectral-spike score	0.05
Total	29		1.00

channels, producing correlation 1 and width 0; this behavior must be considered during reference construction.

4.4 Texture, Temporal Stability, and Artifacts

Texture is represented by local spectral variance, high-frequency smoothness, and narrow spectral spikes. Temporal stability divides the track into early, middle, and late thirds and measures drift in RMS and spectral centroid, plus a late-versus-early feature difference. Artifact descriptors count samples near digital full scale, peaks beyond six standard deviations, and narrow spectral spikes.

These features are deliberately interpretable but incomplete. They do not yet model auditory masking, harmonic plausibility, musical structure, source separation quality, or semantic consistency.

5 Reference-Distance Formulation

5.1 Reference Statistics

For a reference corpus $\mathcal{R}$, each feature is summarized by mean μ_{gj} and standard deviation σ_{gj} . The current implementation floors the standard deviation at 0.01 to avoid unstable division. For track feature x_{gj} in group g , the standardized squared deviation is

$$z_{gj}^2 = \left(\frac{x_{gj} - \mu_{gj}}{\sigma_{gj} + \epsilon} \right)^2. \quad (4)$$

The component distance is the root mean square of standardized deviations:

$$d_g = \sqrt{\frac{1}{n_g} \sum_{j=1}^{n_g} z_{gj}^2}. \quad (5)$$

The total technical-reality distance is

$$D_R = \sum_{g=1}^G \alpha_g d_g, \quad \alpha_g \geq 0, \quad \sum_g \alpha_g = 1. \quad (6)$$

The optional bounded display score is

$$\text{MRS} = 100 \exp(-D_R). \quad (7)$$

5.2 Mathematical Properties

Proposition 1 (nonnegativity). Every $d_g \geq 0$ and, for nonnegative weights, $D_R \geq 0$.

Proposition 2 (bounded presentation score). Equation 7 implies $0 < \text{MRS} \leq 100$, with equality at 100 when $D_R = 0$.

Proposition 3 (monotonicity). The derivative is

$$\frac{\partial \text{MRS}}{\partial D_R} = -100e^{-D_R} = -\text{MRS} < 0, \quad (8)$$

so any increase in total distance strictly decreases the presentation score.

Proposition 4 (component transparency). If only group q changes, then

$$\Delta D_R = \alpha_q \Delta d_q, \quad (9)$$

which exposes the exact weighted contribution of that group. This transparency is lost if only the bounded score is retained.

Table 3 illustrates the nonlinear transform. Each additional unit of D_R multiplies the score by $e^{-1} \approx 0.368$, making weight and variance calibration consequential.

Table 3: Illustrative distance-to-score mapping.

D_R	0	0.25	0.5	1	2
MRS	100.0	77.9	60.7	36.8	13.5

6 Controlled-Degradation Validation Protocol

The repository contains an executable five-experiment script. Four experiments test directional correctness under synthetic degradations; a fifth produces before/after comparisons for selected Moodify treatments. The protocol is summarized in Table 4.

6.1 Directional Hypotheses

For controlled item x and degraded item $T(x)$, the basic hypothesis is

$$D_R(T(x)) > D_R(x), \quad \text{MRS}(T(x)) < \text{MRS}(x). \quad (10)$$

A stronger component hypothesis requires the intended group distance to increase. The metric should also be tested for specificity: a spectral degradation should not consistently appear only as a spatial defect.

6.2 Severity and Monotonicity

Binary pass/fail testing is insufficient. Each transform should be applied at multiple severities $s_1 < \dots < s_K$ and evaluated for rank monotonicity:

$$\rho_s = \text{Spearman}(s, D_R). \quad (11)$$

A robust metric should exhibit positive severity-distance correlation in the intended group while avoiding unstable reversals.

6.3 Loudness Control

Before/after candidates must be loudness matched; otherwise a louder version may be preferred or may alter level-dependent features. Programme loudness and true peak should be measured according to ITU-R BS.1770-5, with normalization policy documented [7].

7 Implementation Audit and Current Evidence

7.1 What Is Implemented

The public module computes the seven feature groups, constructs reference statistics, calculates group distances, returns D_R and MRS, and compares before/after files. The validation script implements four controlled-degradation families and an optional Moodify case-comparison stage. The code therefore supports reproducible experimentation at the implementation level.

7.2 What Is Not Yet Proven

The repository does not currently include a complete benchmark package containing audio stimuli, frozen reference statistics, execution logs, listening-test responses, and statistical analysis. Consequently, the present manuscript does not report correlation coefficients, confidence intervals, or superiority over FAD, KAD, PEAQ, or ViSQOLAudio.

The current smoke-test script builds reference statistics from up to three available WAV files, with a fallback to two files. Such a reference set is far too small for scientific calibration and can produce unstable means and variances. It is suitable only for software-directionality checks.

A repository summary contains 30 treatment records, of which six have completed internal feedback and 24 remain pending. All six completed records marked the processed output as better than before, but the coverage is only 20%; the feedback is not documented as blinded, randomized, multi-listener, or statistically independent. It is therefore excluded from MRS perceptual-validation claims.

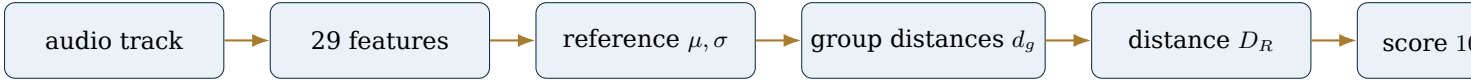


Figure 1: MRS computation. The primary scientific variable is the open-ended distance D_R ; the 0-100 score is a monotonic presentation transform.

Table 4: Controlled-degradation families implemented in the current validation script.

Experiment	Transform	Intended defect	Directional hypothesis
Spectrum	+8 dB above 6 kHz; -6 dB above 8 kHz; eight-region magnitude flattening	excessive brightness, dullness, reduced spectral shape	degraded MRS below baseline; spectrum distance increases
Dynamics	4:1 hard compression above about -12 dBFS; clipping at 0.6; +6 dB gain applied to second half	overcompression, limiting, level instability	degraded MRS below baseline; dynamics or artifact distance increases
Space	side channel multiplied by 3; right channel delayed 2 ms; stereo collapsed to mono	excessive width, phase inconsistency, loss of width	overwide and delayed versions score below baseline; mono is treated cautiously
Temporal	second half receives +9.5 dB high-frequency boost, gain drift up to +50%, and amplified spikes	late-section deterioration	degraded MRS below baseline; temporal distance increases
Treatment comparison	three track types crossed with three Moodify presets, using loudness-matched after versions when available	inspect whether component gains and penalties are interpretable	report ΔD_R , ΔMRS , main gain, and main penalty; not a superiority test

7.3 Code-Level Caveats

The audit identified several issues that should be resolved before benchmark release:

1. default weights are expert priors and have not been learned or perceptually calibrated;
2. the temporal function uses a different key name in the short-clip return path than the feature list expects, causing the affected value to fall back to zero;
3. the fifth validation experiment records comparisons as interpretable cases rather than applying a directional pass criterion;
4. artifact detection is intentionally simple and does not yet include a full perceptual distortion model;
5. several features are correlated, so equal within-group averaging does not imply equal independent information.

These limitations are part of the research problem rather than details to hide.

8 Formal Validation Plan

A submission-ready study should use a preregistered protocol with frozen code, reference sets,

and analysis scripts.

8.1 Corpus Design

The corpus should include AI-generated music from multiple generation systems and genres, plus legally usable production-ready references. Training, calibration, and held-out evaluation partitions must be separated. Reference statistics should be released by version and, where appropriate, stratified by genre or production style.

8.2 Distortion Benchmark

Each source should be transformed by multiple distortion families and at least four severity levels. Candidate families include spectral tilt, narrow resonances, low-pass loss, hard compression, clipping, transient smearing, channel delay, excessive width, late-section drift, and spectral spikes. The evaluation should report directional accuracy, severity rank correlation, component specificity, and cross-genre stability.

8.3 Subjective Listening

A MUSHRA-style experiment should compare hidden reference or best-available anchors, degraded variants, and post-processed candidates. Listeners should rate overall technical quality and separate attributes such as naturalness, harshness, spatial coherence, transient integrity, and artifact severity. Expert and non-expert listeners should be analyzed separately. Stimuli must be loudness matched and presented under documented playback conditions.

8.4 Metric Comparison

MRS should be compared with PEAQ, ViSQOLAudio, FAD, and KAD where their input assumptions are satisfied. Track-level FAD or KAD use requires careful interpretation because these methods are naturally distributional. Report Spearman and Kendall correlations, bootstrap confidence intervals, and mixed-effects models with listener and track as random effects.

8.5 Ablation and Calibration

Ablation should remove each group or feature family to estimate marginal contribution. Weight calibration must be performed only on a training partition and tested on unseen data. The bounded score transform should also be compared with linear and logistic alternatives; the primary D_R should remain available regardless of interface mapping.

9 Discussion

MRS occupies a specific position between perceptual models and distributional generative metrics. Its advantage is not a claim of universality but track-level explanation. A producer can see that a candidate moved closer to the reference distribution in texture while moving farther away in dynamics. This supports diagnosis, regression testing, and controlled iteration.

The same interpretability creates risk. Hand-crafted features may encode the developers' assumptions more strongly than human perception. A metric can also encourage optimization toward its own coordinate system, producing technically conforming but artistically homogeneous audio. MRS should therefore be used as a gate and diagnostic aid, not as the sole objective for automatic processing.

The reference corpus is an explicit value judgment. "Real" cannot be treated as a single timeless acoustic state. The most defensible interpretation is operational: closer to a named, versioned, curated reference distribution under a stated feature definition. Publishing this context is necessary for reproducibility and fairness.

10 Conclusion

This paper defined MRS as a multi-dimensional, reference-conditioned metric for the technical realism of individual AI-generated music tracks. The metric uses 29 interpretable features in seven groups, reports an open-ended distance D_R , and optionally maps that distance to a bounded score. Its mathematical semantics are clear, and controlled-degradation experiments are implemented. However, the current evidence is engineering evidence rather than full perceptual validation.

The next milestone is not to add more formulas. It is to freeze a benchmark, enlarge and version the reference corpus, correct implementation inconsistencies, run multi-severity distortions, conduct a MUSHRA-style listening study, compare against established metrics, and publish both successful and failed cases. Only after this process should MRS be presented as a validated audio-quality metric.

Code and Data Availability

The implementation is available in the public Moodify repository under the Apache License 2.0. The current repository includes metric code, validation scripts, and treatment records, but not a complete publication benchmark with redistributable audio stimuli. A frozen release, reference-statistics file, and experiment manifest should accompany formal submission.

Conflict of Interest

The author is the founder of Rongjing Wenchuan (Shenzhen) Technology Co., Ltd. and the creator of Moodify. This commercial relationship may influence system design and interpretation. The manuscript therefore separates implemented facts, preliminary internal records, and uncompleted validation requirements.

Generative-AI Assistance Disclosure

Generative-AI tools were used to assist with manuscript organization, language drafting, and formatting. The human author defined the research direction and technical concepts and is responsible for verifying all equations, code descriptions, references, claims, and interpretations. No synthetic experimental result is reported as observed data.

References

- [1] International Telecommunication Union, "Method for Objective Measurements of Per-

- ceived Audio Quality,” Recommendation ITU-R BS.1387-2 (2023 May).
- [2] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, “ViSQOLAudio: An Objective Audio Quality Metric for Low Bitrate Codecs,” *J. Acoust. Soc. Am.*, vol. 137, no. 6, pp. EL449–EL455 (2015), doi: 10.1121/1.4921674.
- [3] C. Sloan, N. Harte, D. Kelly, A. C. Kokaram, and A. Hines, “Objective Assessment of Perceptual Audio Quality Using ViSQOLAudio,” *IEEE Trans. Broadcast.*, vol. 63, no. 4, pp. 693–705 (2017), doi: 10.1109/TBC.2017.2704421.
- [4] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, “Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms,” in *Proc. Interspeech*, pp. 2350–2354 (2019), doi: 10.21437/Interspeech.2019-2219.
- [5] Y. Chung, P. Eu, J. Lee, K. Choi, J. Nam, and B. S. Chon, “KAD: No More FAD! An Effective and Efficient Evaluation Metric for Audio Generation,” arXiv:2502.15602 (2025).
- [6] International Telecommunication Union, “Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems,” Recommendation ITU-R BS.1534-3 (2015 Oct.).
- [7] International Telecommunication Union, “Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level,” Recommendation ITU-R BS.1770-5 (2023 Nov.).
- [8] European Broadcasting Union, “Loudness Normalisation and Permitted Maximum Level of Audio Signals,” EBU R 128, version 5.0 (2023 Nov.).
- [9] M. Torcoli, T. Kastner, and J. Herre, “Objective Measures of Perceptual Audio Quality Reviewed: An Evaluation of Their Application Domain Dependence,” arXiv:2110.11438 (2021).
- [10] X. Zhu, “Moodify: AI Music Post-Processing and Listening Engineering System,” software repository, version 0.1.0 and experimental research modules (2026), <https://github.com/huliye24/moodify-o3is>.