Online Supplement:

From Sound to Structure: What Total Roughness Misses in Complex Music

Angela Zhang

This supplement provides additional figures, full numerical results underlying selected main-text figures, classification results, robustness analyses, and secondary tests. All core data-processing and statistical methods are described in the Methods section of the main paper. Tables and figures use the prefix "S" to distinguish them from those in the main text.

Additional Synthetic Benchmark Panels

Figure S1 provides two additional views of the synthetic roughness benchmark that complement Figure 1 in the main text. Panel A illustrates the time-domain interaction between two tones. When two tones are sounded together, their waveforms interfere and produce amplitude fluctuations known as *beating*. The beating rate equals the frequency difference between the two tones. For a minor third (6:5 ratio, shown in red), the frequency difference is smaller than for a major third (5:4, shown in blue), so the beating is slower. In the Sethares model, however, roughness does not depend simply on beating rate. It depends on where the frequency separation falls on the critical-band-scaled roughness kernel. At typical musical registers, the minor third lies closer to the peak of this kernel and therefore receives a higher modeled roughness value.

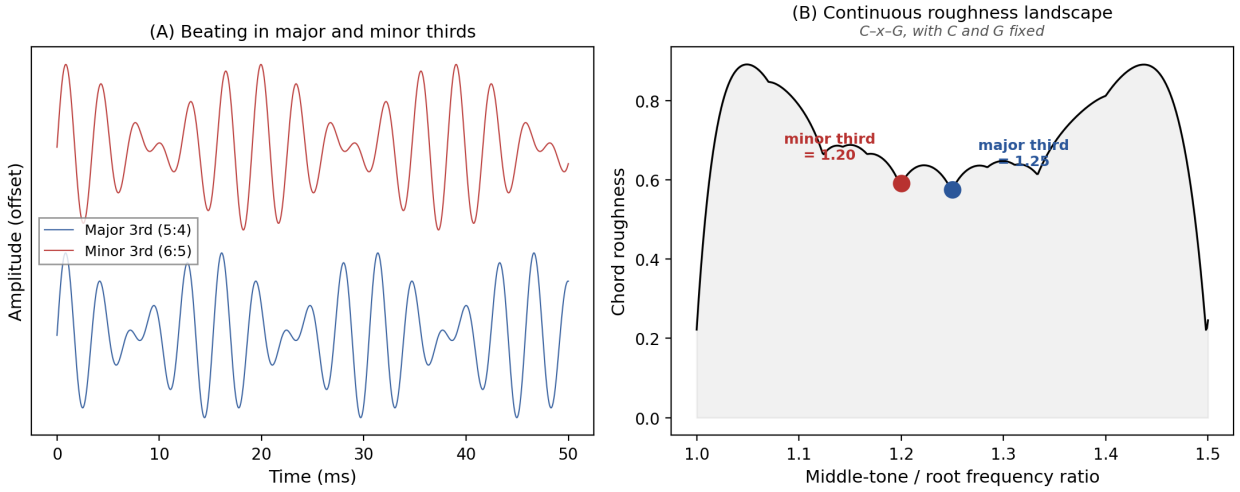


Figure S1: Additional synthetic benchmark panels. **(A)** Time-domain waveforms for the major third (5:4, blue) and minor third (6:5, red). The minor third has a smaller frequency difference and therefore slower beating, but lies closer to the peak of the critical-band-scaled roughness kernel. **(B)** Continuous roughness landscape for a C–*x*–G triad with C and G fixed and the middle tone swept across frequency ratios (eight harmonics per note). The minor third at ratio 1.20 lies above the major third at 1.25.

Panel B places this comparison in a broader context by sweeping the middle tone of a C–*x*–G triad continuously across frequency ratios from 1.0 to 1.5. The root C and fifth G are held fixed while the middle tone varies, and chord roughness is computed at each position using eight harmonics per note. The resulting roughness landscape shows that the minor third (ratio 1.20, red dot) lies slightly above the major third (ratio 1.25, blue dot), consistent with the synthetic benchmark in the main text. Both intervals lie in a local valley of the roughness landscape, although the minor third remains slightly rougher than the major third.

S1. Progressive Spectral Complexity

Table S1: Progressive spectral-complexity analysis in the random sample of 3,000 tracks using the primary minimum-amplitude weighting. Signed d is calculated as minor-like minus major-like, so positive values indicate the minor-rougher direction. "Max. pairs" gives the maximum number of unique pairs possible at each value of N. Correlations are calculated across tracks at each value of N.

N	Max. pairs	d (mode)	r (RMS)	r (entropy)	r (harmonic ratio)
2	1	+0.21	0.73	−0.02	−0.18
3	3	+0.18	0.76	−0.01	−0.20
5	10	+0.12	0.79	+0.03	−0.22
7	21	+0.08	0.81	+0.08	−0.23
10	45	+0.06	0.82	+0.17	−0.26
15	105	+0.04	0.83	+0.32	−0.31

The table documents the transition summarized in main text Figure 2. The mode-related roughness effect falls from $d = +0.21$ with two peaks to $+0.04$ with 15, while the correlation with spectral entropy rises from essentially zero to $+0.32$. The correlation with harmonic ratio also becomes more negative (-0.179 to -0.311), showing that roughness becomes more strongly associated with harmonic versus percussive content as additional spectral components enter the calculation.

As an additional robustness check, I apply an alternative temporal-envelope roughness estimator using 24 ERB-spaced bandpass filters and a Gaussian roughness-sensitivity function peaking at 70 Hz.[2] This estimator also produces only a small major–minor contrast ($|d| = 0.06$). The near-null result is therefore not unique to the Sethares implementation, although this comparison does not validate either estimator against perceived roughness in the corpus.

S2. Mode Information in the Interaction-Separation Profile

Table S2 reports the classification analysis corresponding to the cancellation analysis in the main text. The question is whether retaining information about *where* pairwise roughness occurs improves mode discrimination beyond total roughness and simpler spectral summaries.

Retaining the frequency-separation structure of pairwise roughness improves mode classification from AUC 0.582 to 0.637 after amplitude-distribution features are already included ($\Delta\text{AUC} = +0.055$, $p < 0.001$). The amplitude-distribution features alone and the harmonicity-plus-total-roughness baseline each reach AUC

Table S2: Mode classification using the interaction-separation profile (artist-grouped 5-fold cross-validation, 9,244 tracks). The baseline contains peak harmonicity and total roughness. "Amplitude distribution" denotes the three spectral amplitude-distribution features (top-3 amplitude share, N_{eff}, and spectral entropy), and "Profile" denotes the eight roughness contributions from the frequency-separation bins. ΔAUC is measured relative to the harmonicity + total-roughness baseline unless otherwise indicated.

Model	AUC	ΔAUC
Harmonicity + total	0.555	—
+ Amplitude distribution	0.582	+0.027
+ Profile	0.615	+0.060
+ Amplitude distribution + Profile	0.637	+0.082
Profile alone	0.608	—
Amplitude distribution alone	0.555	—
Incremental artist-level bootstrap comparison:		
Harmonicity + total + amplitude distribution → + profile		+0.055 ($p < 0.001$)

0.555, whereas the separation profile alone reaches 0.608. When amplitude-distribution features are added to the baseline, however, AUC increases to 0.582. The separation profile therefore contains substantial tonal information beyond that captured by total roughness and amplitude-distribution features.

S3. Which Part of the Profile Carries Mode Information?

Table S3 separates the mode information carried by frequency arrangement, amplitude weighting, and the complete roughness calculation.

The count profile, which uses only the number of frequency pairs per separation bin and ignores amplitudes, produces the largest improvement over the baseline (+0.076, compared with +0.067 for the amplitude profile and +0.060 for the roughness profile). This indicates that much of the mode information in the separation structure is already present in frequency arrangement. However, count and roughness profiles are not redundant. In profile-only models, combining them improves AUC by +0.029 relative to roughness alone and by +0.023 relative to count alone (both $p < 0.001$), reaching AUC 0.637. Frequency arrangement therefore carries substantial tonal information on its own, while amplitude and roughness weighting retain additional information beyond pair counts.

S4. Alternative Separation Representations

Table S4 tests whether the interaction-separation result depends on the particular eight-bin specification used in the main analysis.

Every alternative representation improves classification over the baseline, with ΔAUC ranging from +0.052 to +0.072. For non-overlapping bin schemes, the cancellation coefficient C ranges from 0.54 (8-quantile) to 0.75 (12-quantile), indicating substantial cancellation across all tested partitions. The primary 8-bin scheme yields $C = 0.71$. Gaussian radial basis function (RBF) representations are smooth and do

Table S3: Comparison of three separation-profile representations (artist-grouped 5-fold cross-validation, 9,244 mode-quartile tracks). The count profile records the number of pairs in each bin; the amplitude profile sums $\min(A_i, A_j)$ within each bin; and the roughness profile sums the full Sethares pairwise contribution. ΔAUC in the upper panel is measured relative to the harmonicity-plus-total-roughness baseline. AUC values are rounded; ΔAUC is computed from unrounded values. Bracketed intervals are 95% bootstrap CIs based on 2,000 resamples: track-level resampling for individual AUC values and paired artist-level resampling for ΔAUC, using the same resampled artists for both models in each comparison.

Model	AUC	ΔAUC
Baseline (harmonicity + total)	0.555 [0.539, 0.575]	—
+ Count profile	0.630	+0.076
+ Amplitude profile	0.621	+0.067
+ Roughness profile	0.615	+0.060
+ Count + Roughness	0.637 [0.618, 0.656]	0.082 [0.051, 0.101]
Count alone	0.614	—
Amplitude alone	0.610	—
Roughness alone	0.608	—
Count + Roughness alone	0.637	—
Profile-only paired bootstrap comparisons:		
Count + Roughness vs. Roughness alone		+0.029 ($p < 0.001$)
Count + Roughness vs. Count alone		+0.023 ($p < 0.001$)

Table S4: Robustness of mode classification and cancellation to alternative representations of frequency separation (artist-grouped 5-fold cross-validation; 9,244 tracks). The baseline model includes peak harmonicity and total roughness (AUC = 0.555). For non-overlapping bin schemes, the cancellation coefficient is defined as $C = 1 - |\sum_b \Delta_b| / \sum_b |\Delta_b|$. Quantile-based schemes include an additional overflow bin for pairs beyond the largest quantile boundary, yielding one more feature than the stated number of quantile bins. All reported ΔAUC values are significant at $p < 0.001$.

Representation	# features	C	ΔAUC
Original 8 bins	8	0.71	+0.060
Equal-width, 6 bins	6	0.72	+0.054
Equal-width, 10 bins	10	0.61	+0.060
Equal-width, 12 bins	12	0.72	+0.057
Quantile, 8 bins	9	0.54	+0.067
Quantile, 12 bins	13	0.75	+0.072
Gaussian RBF, 6 centers	6	—	+0.052
Gaussian RBF, 10 centers	10	—	+0.054
Gaussian RBF, 15 centers	15	—	+0.054

not partition pairs into discrete bins, so C is not defined for them. These results indicate that the mode information captured by the frequency-separation structure is robust to changes in bin boundaries and to alternative representations of frequency separation.

S5. Happy–Sad Boundary Analysis

Table S5 reports a secondary test of whether the interaction-separation profile (ISP), defined as the eight-bin vector of level-normalized pairwise roughness contributions across frequency-separation ranges, contributes information about happy–sad expression beyond activity-related acoustic features, tonal organization, and mode.

Table S5: Happy–sad classification using artist-grouped 5-fold cross-validation (301 artists, 1,311 tracks with complete ISP data from the mode-quartile sample; 893 happy, 418 sad). "Activity" includes RMS signal level, spectral flux, onset rate, and tempo; "Harm" includes harmonic ratio and spectral flatness. ISP denotes the interaction-separation profile. ΔAUC values are evaluated using 2,000 paired artist-level bootstrap resamples. AUC values are rounded, while ΔAUC is calculated from the underlying unrounded values. p_{raw} denotes the unadjusted p-value and p_{adj} the Holm-adjusted p-value across the four ΔAUC comparisons in this table.

Model	AUC	ΔAUC of ISP
Without mode:		
Activity + Harm	0.817	—
Activity + Harm + ISP	0.829	$+0.011$ ($p_{\text{raw}} = 0.042$, $p_{\text{adj}} = 0.084$)
With mode:		
Activity + Harm + Mode	0.863	—
Activity + Harm + Mode + ISP	0.865	$+0.002$ ($p_{\text{raw}} = 0.33$, $p_{\text{adj}} = 0.33$)
Mode contribution:		
Mode beyond Activity + Harm		$+0.046$ ($p < 0.001$)
Mode beyond Activity + Harm + ISP		$+0.036$ ($p < 0.001$)

Without controlling for mode, adding the ISP to the Activity + Harm model produces a nominally significant but small improvement in classification ($\Delta\text{AUC} = +0.011$, $p_{\text{raw}} = 0.042$, unadjusted). Once mode is included, however, the incremental contribution of the ISP falls to only $+0.002$ and is not statistically significant ($p_{\text{raw}} = 0.33$, $p_{\text{adj}} = 0.33$). Mode itself contributes substantially more predictive information: it improves AUC by $+0.046$ beyond the Activity + Harm baseline and by $+0.036$ even when the ISP is already included (both $p < 0.001$).

Thus, the ISP provides a small amount of predictive information about happy–sad expression when mode is omitted, but essentially no additional information once mode and the broader acoustic controls are included. This supports interpreting the recovered separation structure primarily in relation to tonal-mode organization rather than as an independent predictor of happy–sad expression.

S6. Amplitude Distribution and Active-Pair Strength

To quantify the relationship between spectral amplitude distribution and pair strength, I regress each track's level-normalized mean active-pair amplitude,

$$\bar{W}_i^{\text{norm}} = \frac{1}{N_{\text{active},i}} \sum_{(j,k):\, z_{jk} < 1.5} \frac{\min(A_j, A_k)}{\text{RMS}_i},$$

on a major-mode indicator and spectral controls. I then add the three amplitude-distribution measures: top-3 amplitude share, N_{eff}, and spectral entropy.

Table S6: OLS regression of level-normalized mean active-pair amplitude on mode and spectral amplitude distribution (9,244 tracks: 4,622 minor-like and 4,622 major-like). Percentage changes are measured relative to the mode-only coefficient.

Model	$\hat{\beta}_{\text{major}}$	Change vs. mode only
Mode only	$+4.56$ ($p < 0.001$)	—
+ Spectral controls	$+4.89$ ($p < 0.001$)	$+7\%$
+ Amplitude distribution	$+1.37$ ($p < 0.001$)	-70%

The conventional spectral controls (centroid, bandwidth, harmonic ratio, and spectral flatness) do not attenuate the major–minor pair-strength difference; the coefficient increases slightly from $+4.56$ to $+4.89$. After the three amplitude-distribution measures are added, the coefficient falls to $+1.37$, a 70% reduction relative to the mode-only estimate. This attenuation indicates that amplitude distribution shares substantial variation with pair strength. Because both are constructed from the same retained-peak amplitudes, however, this relationship should be interpreted as associational rather than as causal mediation.

Table S7. Acoustic Characteristics of Mode-Quartile Groups

S8. Additional Robustness Analyses and Scope of Interpretation

This section examines the sensitivity of the main findings to mode-score construction, sample selection, roughness measurement, genre and artist structure, and analysis windows. It also clarifies the remaining limitations of the audio data, synthetic experiments, and amplitude-equalization counterfactual.

Mode-score robustness across key profiles

The mode score is an acoustic proxy derived from the Krumhansl key profiles rather than an independently verified major/minor label. To assess sensitivity to profile choice, I compare the Krumhansl classification with classifications derived from Temperley's and Aarden's key profiles on the 9,244 mode-quartile tracks. The Krumhansl classification agrees with the three-profile majority-vote consensus on 90% of quartile tracks (Cohen's $\kappa = 0.79$). Restricting the quartile comparison to tracks on which all three methods agree retains 66% of the original quartile sample and yields a small minor-rougher difference ($d = +0.08$, 95% CI $[-0.01,$

Table S7: Acoustic characteristics of mode-quartile groups (4,622 major-like and 4,622 minor-like recordings). Group values are means. Cohen's d is the standardized mean difference (minor-like minus major-like). The 95% confidence intervals for d are obtained by artist-level bootstrap resampling (2,000 iterations). "Full-spectrum spectral entropy" is the Shannon entropy of the normalized STFT magnitudes across all frequency bins (in bits, using base-2 logarithms). It differs from the retained-peak entropy $H = -\sum_k p_k \ln p_k$ (in nats) defined in the main text, which is computed over at most 15 retained peaks.

Feature	Major-like mean	Minor-like mean	d	95% CI for d
RMS signal level	0.076	0.073	-0.07	$[-0.13, -0.01]$
Spectral centroid (Hz)	1732	1698	-0.05	$[-0.11, +0.02]$
Spectral bandwidth (Hz)	1847	1826	-0.04	$[-0.11, +0.02]$
Spectral flatness	0.0083	0.0087	$+0.03$	$[-0.03, +0.10]$
Full-spectrum spectral entropy (bits)	4.83	4.83	$+0.01$	$[-0.05, +0.07]$
Harmonic ratio	0.489	0.476	-0.11	$[-0.17, -0.05]$
Tempo (BPM)	121.2	120.8	-0.01	$[-0.07, +0.05]$
Onset rate (s^{-1})	3.42	3.35	-0.04	$[-0.10, +0.02]$
Spectral flux	0.041	0.040	-0.05	$[-0.11, +0.01]$

$+0.16$]). Although this estimate is larger than the original quartile estimate ($d = +0.03$), its magnitude remains small.

This is an internal agreement check among algorithmic classifications, not a comparison against independently verified labels. The central comparison should therefore be interpreted as distinguishing tracks with more versus less major-like chroma profiles, not necessarily major versus minor mode per se.

Continuous and quartile analyses

The quartile split used for effect-size reporting excludes approximately half the dataset from the group comparison. Complementary evidence comes from continuous mode-score analyses that use all 18,486 tracks: the continuous mode score is almost uncorrelated with level-normalized roughness ($r = -0.01$, $p = 0.052$). The quartile comparisons are retained because they provide interpretable standardized mean differences (Cohen's d), support the group-based decomposition analysis, and facilitate direct comparisons with the amplitude-equalization counterfactual. The continuous and quartile results are consistent: the quartile comparison yields $d = +0.03$ (95% CI $[-0.04, +0.09]$), supporting the same near-zero conclusion.

Peak selection and the ERB-band robustness check

Because the 15 dominant spectral peaks are selected by amplitude, the "frequency arrangement" and "amplitude distribution" components of the decomposition are not fully independent. A peak's inclusion depends on its amplitude, so the frequency locations available for the arrangement analysis are influenced by the amplitude structure of the spectrum. This coupling means that the two components should be interpreted as complementary descriptions of the same spectral data rather than as causally independent mechanisms.

To assess whether the aggregate near-zero result depends on the amplitude-driven peak-selection procedure, I conduct an alternative analysis using 30 auditory filters with center frequencies equally spaced on the equivalent rectangular bandwidth (ERB) scale[3,4] from 50 to 5,000 Hz. For each frame, I compute

the energy in each filter band and represent the band by its root-energy amplitude and energy-weighted centroid frequency. I then use these band-level amplitudes and frequencies to calculate pairwise roughness via the Sethares formula. The filter bank is fixed, but the centroid frequencies vary across frames with the within-band energy distribution.

This approach avoids selecting the 15 strongest spectral peaks, although the representative frequencies and active-band threshold remain dependent on spectral energy. The analysis includes $N = 9{,}244$ quartile tracks (4,622 minor-like and 4,622 major-like). A filter band is considered active if its root-energy amplitude exceeds 10^{-10}; frames with fewer than two active bands are excluded from the per-track average.

The resulting mode-quartile difference is $d = +0.01$ (95% CI $[-0.09, +0.11]$, obtained by artist-level bootstrap resampling with 2,000 resamples), and the correlation between ERB-band roughness and peak-based roughness is $r = 0.93$. The similar near-zero estimate supports the robustness of the aggregate finding to replacing amplitude-selected peaks with filter-band representations, but does not establish independence from spectral amplitude structure or independently validate the cancellation decomposition.

Genre, instrumentation, and artist dependence

The two mode-quartile groups are similar on the acoustic features reported in Table S7 (all $|d| \leq 0.11$) but differ in genre composition. Pop is more prevalent among major-like than minor-like tracks (21% versus 11%), whereas electronic music is more prevalent among minor-like than major-like tracks (29% versus 19%). Because genre is associated with instrumentation and production, these differences raise the possibility that genre composition contributes to the aggregate roughness result or the cancellation pattern.

Several analyses address this concern. In the full corpus of 54,080 tracks with extractable audio (27,040 in the lower and upper mode-score quartiles), a within-genre stratified analysis across 51 genres includes 25,429 mode-quartile tracks in strata with at least 30 tracks per group. The median within-genre effect size is $d = +0.04$, and the mean is $d = +0.05$, indicating small mean and median within-genre differences. Each track is assigned a single primary genre (its first listed genre tag); individual genre strata range from approximately 60 to over 3,000 quartile tracks. An analogous within-instrument stratification across 27 instruments, requiring at least 20 tracks per group, yields similarly small summary effect sizes (median $d = +0.04$, mean $d = +0.03$). These within-genre and within-instrument analyses address the aggregate mode–roughness difference but do not directly test whether the cancellation pattern also holds within each stratum.

Mixed-effects models. A linear mixed-effects model[1] with artist random intercepts estimates the association between continuous mode score and level-normalized roughness in the mood-tagged subset ($N = 18{,}486$; 1,533 artists; Table S8a). The outcome and predictor are standardized. The mode coefficient is $\hat{\beta} = -0.0173$ (SE $= 0.0060$, $p = 0.004$). The artist-level ICC is 0.45, indicating substantial clustering of roughness within artists. Replication in the larger corpus ($N = 54{,}079$; 3,549 artists) yields $\hat{\beta} = -0.0120$ (SE $= 0.0035$, $p < 0.001$), with an artist-level ICC of 0.40. One full-corpus track is excluded for a missing mode score.

Genre-adjusted model. Adding genre covariates to the full-corpus mixed-effects model reduces the magnitude of the mode coefficient by 9.5% ($\hat{\beta} = -0.0109$, $p = 0.002$; Table S8a). The association therefore remains small after adjustment for the included genre covariates.

Thus, the stratified analyses yield small mean and median mode–roughness differences within genres and instruments, while the mixed-effects models yield small standardized mode coefficients after accounting for artist clustering and the included genre covariates. These results support the robustness of the small aggregate mode–roughness association. They do not rule out residual confounding or establish that the cancellation pattern holds within individual genres, instruments, or artists. The corpus is predominantly Western, which limits generalization to other musical traditions.

Table S8a: Mixed-effects models of the mode–roughness association. The outcome is standardized level-normalized roughness, and the predictor is standardized continuous mode score. Artist random intercepts are included in all models. Genre covariates are indicator variables for genres with ≥ 30 tracks, with remaining genres pooled as "other". Confidence intervals are Wald-type 95% intervals ($\hat{\beta} \pm 1.96 \times \mathrm{SE}$).

Sample	Additional covariates	N	Artists	$\hat{\beta}$	SE	95% CI	p
Mood-tagged	None	18,486	1,533	−0.0173	0.0060	[−0.029, −0.005]	0.004
Full corpus	None	54,079	3,549	−0.0120	0.0035	[−0.019, −0.005]	<0.001
Full corpus	Genre	54,079	3,549	−0.0109	0.0035	[−0.018, −0.004]	0.002

ICC (artist): 0.45 (mood-tagged), 0.40 (full corpus), 0.35 (genre-adjusted).

Temporal stability across analysis windows

The main analyses use the first 30 seconds of each recording, which may not capture the full musical form. As a robustness check, I compute raw roughness separately for three 30-second windows within the first minute: 0–30, 15–45, and 30–60 seconds. The mode-quartile differences remain small across all three windows (Table S8b). Cross-window roughness correlations are also high ($r \geq 0.84$; Table S8b), although overlap between adjacent windows may contribute to their similarity. These results support temporal stability of the near-zero group difference in raw roughness within the first minute; they do not directly establish the same stability for level-normalized roughness. Later sections of recordings, particularly those with substantial structural changes, remain unexamined.

Table S8b: Mode-quartile roughness differences across analysis windows ($N = 9{,}244$ quartile tracks for 0–30 s; $N = 9{,}244$ for 15–45 s; $N = 9{,}102$ for 30–60 s due to shorter recordings). Cohen's d is the standardized mean difference (minor-like minus major-like). This check uses raw (unnormalized) roughness; the primary analysis uses level-normalized roughness, which yields $d = +0.03$ for the 0–30 s window. Confidence intervals are 95% artist-level bootstrap intervals (2,000 resamples). Cross-window Pearson correlations of track-level raw roughness, computed using tracks available in both windows, are shown below the diagonal.

Window	Cohen's d	95% CI for d	r(0–30)	r(15–45)	r(30–60)
0–30 s	+0.01	[−0.08, +0.11]	—		
15–45 s	+0.03	[−0.06, +0.13]	0.93	—	
30–60 s	+0.05	[−0.04, +0.14]	0.84	0.95	—

Dynamics sensitivity

RMS normalization adjusts for overall signal level but does not control for within-track dynamic range variation. Among the $N = 9{,}244$ mode-quartile tracks, minor-like tracks have slightly higher frame-level RMS standard deviation than major-like tracks (5.45 versus 5.17 dB; $d = +0.11$). As a sensitivity check, I add RMS dynamics (frame-level RMS standard deviation and 90th–10th percentile range in dB) and spectral covariates (centroid, bandwidth, and flatness) to an OLS model of level-normalized roughness on continuous mode score. The outcome and all predictors are standardized, and confidence intervals use HC1 heteroskedasticity-robust standard errors, without clustering by artist. The mode coefficient changes from $\hat{\beta} = -0.024$ (95% CI $[-0.044, -0.005]$) to $\hat{\beta} = -0.031$ (95% CI $[-0.050, -0.012]$), indicating that the association is not attenuated by adjustment for these measured dynamics and spectral characteristics.

Dividing tracks into terciles of frame-level RMS standard deviation yields small estimated mode-group differences, with 95% artist-level bootstrap CIs based on 2,000 resamples: $d = +0.10$ $[+0.004, +0.184]$ for low dynamics (RMS SD 0.6–3.9 dB), $d = +0.05$ $[-0.04, +0.14]$ for medium dynamics (3.9–6.0 dB), and $d = -0.01$ $[-0.11, +0.10]$ for high dynamics (6.0–18.8 dB). Cancellation coefficients are point estimates of $C = 0.29$, 0.62, and 0.88, respectively. The qualitative bin-level sign pattern is preserved across strata. These descriptive estimates suggest variation in cancellation with within-track level dynamics, but do not establish a causal explanation or a statistically supported trend across strata. RMS dynamics measures temporal level variation rather than directly identifying compression or mastering practices; residual confounding by production practices therefore remains possible.

Mono audio and channel limitations

The MTG-Jamendo preview files used in the main analyses are mono, so the mono conversion step leaves these files unchanged. Channel-specific information, including stereo panning and interchannel phase relationships, is therefore unavailable, although effects of stereo production may remain in the mono mix. Assessing sensitivity to channel handling would require stereo recordings and a separate analysis.

Scope of synthetic experiments

The synthetic experiments use triads in twelve-tone equal temperament (12-TET), with just intonation as a robustness check. Under the tested conditions, harmonic organization preserves the minor-rougher contrast as spectral complexity increases. However, these experiments use idealized conditions, including exactly harmonic partials and controlled amplitude decay, that differ from those of real recordings. The results may not generalize to non-triadic chords, other tuning systems, or the dense, overlapping spectra characteristic of complex musical recordings.

Interpretive scope of the equalization analysis

The amplitude-equalization counterfactual is a within-model manipulation. It shows how the Sethares roughness model responds when the amplitude distribution across retained peaks is altered. It is not an acoustic resynthesis or perceptual experiment. The increase from $d = +0.03$ to $d = +0.24$ after equalization indicates

that, within the model and conditional on the retained peak frequencies, the observed amplitude distribution attenuates the minor-rougher contrast. Whether listeners would perceive corresponding resynthesized sounds as more or less rough is a separate empirical question that requires perceptual testing.

Peak-separation constraint sensitivity

The primary analysis identifies spectral peaks using `scipy.signal.find_peaks` with no explicit minimum-distance constraint between peaks. As a sensitivity check, I repeat peak detection with a minimum separation of three FFT bins (approximately 32.3 Hz at a sampling rate of 22,050 Hz and an FFT length of 2,048 samples; `distance=3`). This analysis uses a random sample of 2,000 mode-quartile tracks (1,001 minor-like and 999 major-like) and level-normalized roughness. Confidence intervals for both Cohen's d and the cancellation coefficient C use 2,000 artist-level bootstrap resamples.

Track-level roughness values remain strongly correlated with those from the unconstrained procedure ($r = 0.98$). Roughness decreases by an average of 13.4% across tracks, calculated as the mean track-level percentage decrease. The estimated mode-quartile difference changes from $d = +0.028$ (95% CI [-0.071, $+0.124$]) to $d = -0.023$ (95% CI [-0.127, $+0.075$]), remaining small under both specifications. Within this 2,000-track sensitivity sample, the eight-bin sign pattern is identical under both conditions ($++-------$), with positive minor-minus-major differences in the two closest-separation bins and negative differences in the remaining six bins. The cancellation coefficient is $C = 0.735$ (95% CI [0.120, 0.983]) without the constraint and $C = 0.779$ (95% CI [0.192, 0.982]) with it.

These results support the robustness of the small aggregate difference and the qualitative cancellation pattern to the tested peak-separation constraint, while indicating sensitivity of roughness magnitudes to this parameter.

References

1. R. H. Baayen, D. J. Davidson, D. M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language. Vol. 59, pg. 390–412, 2008, `https://doi.org/10.1016/j.jml.2007.12.005`.

2. P. Daniel, R. Weber. Psychoacoustical roughness: implementation of an optimized model. Acta Acustica united with Acustica. Vol. 83, pg. 113–123, 1997.

3. B. R. Glasberg, B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. Hearing Research. Vol. 47, pg. 103–138, 1990, `https://doi.org/10.1016/0378-5955(90)90170-T`.

4. B. C. J. Moore, B. R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. Journal of the Acoustical Society of America. Vol. 74, pg. 750–753, 1983, `https://doi.org/10.1121/1.389861`.