Long ringing tails and dark spectra make music sound calmer, and sadder

A preregistered test on 502 held-out clips from the DEAM music emotion dataset, with shuffle controls and an instrument checked on synthetic music

Zain Dana Harper

Version 1.0, 9 November 2026. Finding F39 in the findings ledger. Strong signal (tier B*).

Abstract

People often reach for music with long, ringing sustain and a warm, dark tone when they want to wind down. We asked whether those two properties of the sound go with how calm and how pleasant listeners judge a clip to be, beyond loudness, genre and year. On 1,744 rated 45-second clips from the DEAM dataset, we measured how long each note rings after its onset and how far the spectrum leans toward low frequencies, and tested the link on 502 clips by 270 artists kept out of all fitting. A calm index built from a longer tail and a darker spectrum correlates -0.394 with rated arousal (95% interval -0.476 to -0.304), with loudness, genre and year held fixed. Adding the two measures raises held-out explained variance in arousal by 0.110 (0.034 to 0.182). We forecast that the same sounds would be rated more pleasant; they were rated less pleasant, -0.375 (-0.455 to -0.280), and -0.227 with arousal also held. In these ratings, calm sits closer to subdued than to content.

1 What this means for a reader

Music that rings on and sounds dark is heard as calmer, across every genre in the set. It is also heard as sadder. If you build a playlist to relax, these two properties will move the music toward calm, and they will also move it toward melancholy. The ratings describe what listeners judged the music to express, and say nothing about what it made them feel.

2 Introduction and motivation

The acoustic side of musical emotion has mostly been studied on single notes. Eerola, Ferrer and Alluri [2] had listeners rate isolated instrument sounds for valence, tension and energy and found that features such as attack time and spectral brightness tracked those ratings. Korsmit and colleagues [3] pooled several such datasets and found that some acoustic links held across them and others did not. These studies hold the musical context fixed by removing it.

Reverberation has its own small literature. Mo, Wu and Horner [4] added reverberation to eight wind and bowed-string sounds and found that its length shifted their emotional character, most strongly toward romantic and mysterious and less toward sad. Wilkie and Harrison [5] varied reverberation time in recorded music heard over headphones and found that longer reverberation raised ratings of sublimity, nostalgia and wonder while leaving unease and vitality about the same.

We asked a question one step further from the laboratory: in full mixed music by many artists, does a longer ringing tail after each note and a darker spectrum go with lower rated arousal, beyond loudness and genre, on artists the analysis never saw? We also asked, more tentatively, whether the same sounds would be rated as more pleasant. The decision this informs is which acoustic controls a music tool can trust to move perceived calm, and what else they move with it.

Searches run on 9 November 2026 found the DEAM benchmark paper [1], the studies above and related work

on isolated tones. They returned nothing that measures an onset-locked tail in full mixed music. That is what the searches returned; it does not show that no such study exists.

3 Preregistration and design

The preregistration, a dated plan hash-locked before any result is seen, was written on 29 October 2026 and locked at 16:24 UTC. Acoustic features were extracted from the audio before the lock. No rating was joined to any feature before it; while checking copies of the annotation file, the first two rows (two clips) were seen and nothing else of the ratings.

The plan fixed the data, the measures that had passed an instrument check, the artist split, the controls, the predictions and the rule for each verdict, which is the pass or fail of one preregistered rule set read off by code. Every rule is fail-closed: a verdict passes only when all of its rules hold. The predictions were:

- P1, primary: the calm index correlates negatively with arousal on the test clips, its interval excludes zero, and it lies below the 5th percentile of a shuffle control.
- P2a and P2b: the tail alone correlates negatively with arousal, and the spectral slope alone positively, each with an interval excluding zero.
- P3, stated as possible: the calm index correlates positively with valence, its interval excludes zero, and it lies above the 95th percentile of a shuffle control.
- P4: adding the tail and the slope to the controls raises explained variance in arousal on the test clips, with an interval above zero and above the 95th percentile of a shuffle control.
- P5: P1 still holds with tempo and spectral entropy added to the controls.

The forecasts written before the run were a P1 correlation near -0.15 (range -0.05 to -0.30) and a P3 correlation near +0.05, likely to fail. The same plan also held two questionnaire tests on music absorption and personality. Their values were seen before the lock through a test of the analysis code, so they are not blind tests, and they are left out of this paper.

One change came after the lock and before feature extraction: the audio loader switched from an FFT resampler to a polyphase one. The instrument check's synthetic clips are made at the analysis rate and never pass through the loader, so the check is unaffected.

4 Data and provenance

The data are DEAM, the MediaEval Database for Emotional Analysis in Music [1]. It holds Creative Commons music from the Free Music Archive, Jamendo and MedleyDB, with ratings from crowd workers. The release is licensed CC BY-NC, non-commercial use with attribution; this paper reports statistics only and redistributes no audio or ratings.

The ratings are arousal and valence: arousal runs from calm to excited and valence from unpleasant to pleasant, each on a nine-point scale averaged over raters for each clip. They measure perceived emotion, what a listener judges the music to express, and not what the listener feels.

We used the 1,744 clips of 45 seconds from the 2013 and 2014 rating rounds, by 899 artists. The 58 full songs from 2015 were left out because their length, rater pool and rating method differ. Genre comes from the dataset's metadata: 11 sampling genres, with one clip under another label pooled as Other.

The official host refused connections on the day of download. The audio came from a public mirror; the archive holds 1,802 files, the count the dataset paper gives, but there is no official checksum to compare it with. The per-clip annotations came from three independent public copies with identical bytes. The metadata is the official file, retrieved through the Internet Archive. The mirror's provenance is a limit, stated below.

5 Methods

Measures

Each clip was converted to mono at 22,050 Hz and cut to its first 45 seconds. The energy envelope is the root mean square over 256-sample windows at a 64-sample hop.

The onset-locked tail is a decay time measured after each note onset. Onsets are envelope peaks at least 0.1 s apart and above the envelope's 60th percentile. For each onset, the log envelope from the peak to the next minimum is fitted by least squares against time, giving a slope b_i. Each onset with a negative slope gives a decay time constant in seconds,

$$\tau_i = -\frac{1}{b_i} \tag{1}$$

and the tail T_{tail} is the median of τ_i over those onsets. This follows whichever of the note's own decay and the room's decay is slower. In dense music it cannot separate the two.

The spectral slope s is the least-squares slope of $\log_{10}$ power against $\log_{10}$ frequency from a Welch spectrum, between 110 Hz and 5.5 kHz. A more negative slope means a darker sound.

The calm index combines the two, with each measure standardised by its mean μ and standard deviation σ on the training clips only:

$$C = \frac{\ln T_{tail} - \mu_T}{\sigma_T} - \frac{s - \mu_s}{\sigma_s} \tag{2}$$

A higher C means a longer tail and a darker spectrum.

Split

Artists were put in a random order with a fixed seed, and 30 percent of them, with all their clips, form the held-out test set: 502 clips by 270 artists. The other 1,242 clips are the training set. Every verdict reads the test set.

Statistics

The controls in every confirmatory model are loudness in decibels relative to full scale, 11 genre indicators and rating year, with an intercept. The main statistic is the partial correlation: each of the two quantities is regressed on the controls by least squares, and the residuals e_x and e_y are correlated, with sums over the clips:

$$r_{xy} = \frac{\sum e_x e_y}{\sqrt{\sum e_x^2 \sum e_y^2}} \tag{3}$$

Residuals from a fit with an intercept have mean zero, so this is the ordinary correlation of the residuals.

The held-out R2 gain fits arousal on the training clips twice, once on the controls and once on the controls plus $\ln T_{tail}$ and s, and scores both fits on the test clips:

$$R^2 = 1 - \frac{\sum_i (y_i - y_{pred,i})^2}{\sum_i (y_i - y_{mean})^2} \tag{4}$$

$$G = R_1^2 - R_0^2 \tag{5}$$

where y_{pred} is a fit's prediction for a test clip, y_{mean} the mean test rating, R_0^2 the test score of the controls-only fit, R_1^2 that of the fit with the two measures, and G the gain.

Intervals are 95% intervals, the 2.5th to 97.5th percentiles of 2,000 cluster bootstrap resamples of the test set, where whole artists are drawn with replacement.

6 Controls and planted checks

Instrument check on synthetic music

Before the lock, every measure was run as a planted check: synthetic music built so that the right answer is known. Each of 40 clips held 30 seconds of harmonic notes on a beat, with room decay time, note decay, attack, tempo, spectral tilt and wet mix set at random and varying at once. Each measure's median error was compared with the error of the same measure against shuffled targets. A measure counted as strong when it beat the shuffled targets' 5th percentile and its median error was at most half the shuffled median. The ratio of the two medians is in Figure 1.

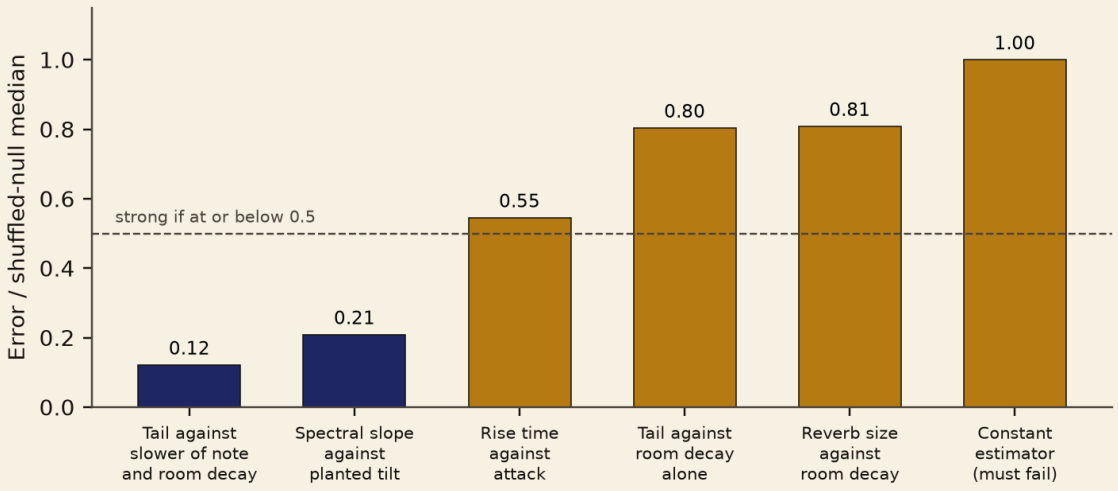


Figure 1. Instrument check on 40 synthetic music clips with planted values. Each bar is the measure's median error divided by its error against shuffled targets; strong measures sit at or below 0.5. Drawn from the instrument file.

The tail recovered the slower of note and room decay with an error ratio of 0.12 and a rank correlation of 0.95; the spectral slope recovered the planted tilt with 0.21 and 0.99. The library's reverb-size measure failed in dense music (0.81), and the tail against room decay alone failed too (0.80), so room reverberation cannot be separated from note decay with these tools and enters no verdict. The best attack measure reached 0.55, short of strong, so attack enters no verdict either. A constant estimator, included as a false-success check, scored 1.00 and failed, as it must.

Shuffle controls

Each control removes the structure under test and is scored the same way as the real data. For P1 and P3, the calm index was shuffled across test clips 1,000 times and the partial correlation recomputed each time; the spread of those draws is the null distribution. For P4, the tail and slope rows were shuffled within the training set and within the test set, 1,000 times, and the whole fit repeated. Table 1 gives the bars.

Table 1. Shuffle controls, 1,000 draws each, and the real value each is compared with.

Control	Mean	5th pct	95th pct	Real value
Calm index shuffled, arousal	0.000	-0.076	0.074	-0.394
Calm index shuffled, valence	-0.002	-0.078	0.077	-0.375
Tail and slope shuffled, R2 gain	-0.001	-0.008	0.003	0.110

7 Results

Figure 2 and Table 2 give the confirmatory results on the 502 test clips. A calm index built from a longer tail and a darker spectrum correlates -0.394 with rated arousal (95% interval -0.476 to -0.304), with loudness, genre and year held fixed. That is far outside the shuffle range and about 2.6 times the size forecast. Each measure carries the link on its own: the tail at -0.301 and the slope at +0.324, where a higher slope means a brighter sound.

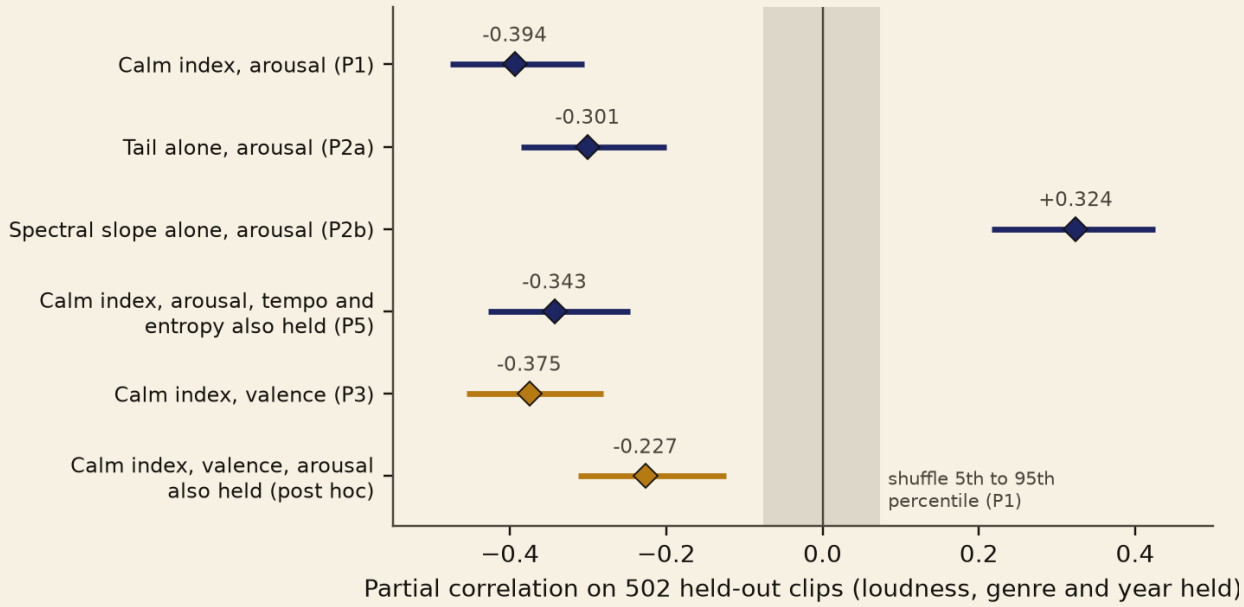


Figure 2. Partial correlations on 502 held-out clips with 95% artist-bootstrap intervals. Navy: arousal. Ochre: valence, forecast to be positive. The grey band is the 5th to 95th percentile of the shuffled calm index for P1. Drawn from the result and post hoc files.

Table 2. Confirmatory results on 502 held-out clips by 270 artists. Controls: loudness, genre, year.

Test	Value	95% interval	Verdict
P1 calm index, arousal	-0.394	-0.476 to -0.304	pass
P2a tail, arousal	-0.301	-0.386 to -0.200	pass
P2b spectral slope, arousal	0.324	0.217 to 0.426	pass
P3 calm index, valence	-0.375	-0.455 to -0.280	fail
P4 R2 gain, arousal	0.110	0.034 to 0.182	pass
P5 P1 with tempo and entropy held	-0.343	-0.428 to -0.246	pass

Adding the two measures raises held-out explained variance in arousal by 0.110 (0.034 to 0.182). The controls alone explain 0.160 of the test variance. With tempo and spectral entropy added to the controls, the calm index still correlates -0.343 with arousal.

We forecast that the same sounds would be rated more pleasant; they were rated less pleasant, -0.375 (-0.455 to -0.280), and -0.227 with arousal also held. P3 failed on all three of its rules. This is an honest null of the forecast, reported with the same weight as the passes: arousal and valence correlate 0.54 on the test clips, so part of the valence drop rides on arousal, and the rest (-0.227, interval -0.312 to -0.123) does not.

Post hoc checks

The following were run after the verdicts and carry no verdict. In busy music the next onset cuts each tail short,

so the tail could stand in for note rate. Note rate has a median of 4.9 events per second and correlates -0.17 with the log tail. With note rate added to the controls, P1 is -0.370 (-0.444 to -0.284) and the tail alone -0.285; with note rate, tempo and entropy all added, P1 is -0.324.

Inside each of the 11 genres, on all 1,744 clips, the calm index correlates negatively with arousal, from -0.25 in hip-hop to -0.62 in international music. Figure 3 shows each genre.

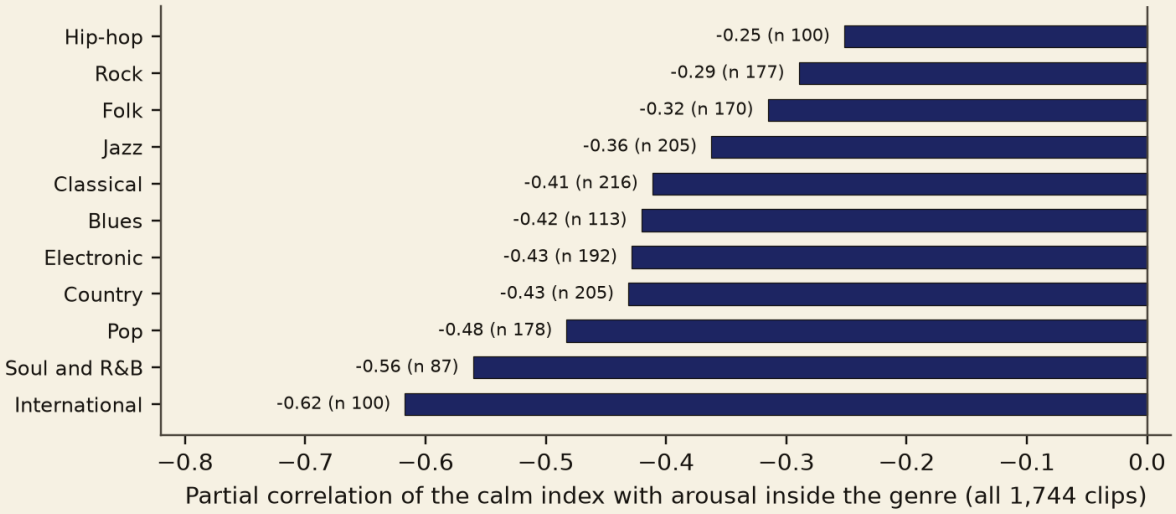


Figure 3. Partial correlation of the calm index with arousal inside each genre, all 1,744 clips, loudness and year held. Post hoc. Drawn from the post hoc file.

On the whole sample, P1 is -0.414 and the valence correlation -0.395. Among exploratory measures on the test clips, the failed reverb-size measure reads in the same direction as the tail (partial correlation with arousal -0.237), and so does the envelope's autocorrelation decay (-0.198); neither can be read as room reverberation. Rise time shows no relation (0.024), which is a null of a weak instrument and says nothing about attack itself.

8 Discussion

The arousal result agrees with what single-note studies found for brightness [2] [3] and extends it to full mixes by many artists, on artists held out from fitting, with a tail measure checked on synthetic music before use. The size, about -0.39, is larger than we forecast. It survives tempo, spectral entropy and note rate, and holds inside every genre.

The valence result is the more interesting one. A reader might expect calm music to be rated pleasant. In DEAM it is rated less pleasant, and the drop is not only the shared variance with arousal. Wilkie and Harrison [5] found that longer reverberation raised sublimity and nostalgia and left unease and vitality about the same, and Mo, Wu and Horner [4] found reverberation moved instrument sounds toward romantic and mysterious and, less strongly, sad. Both fit a reading in which long dark tails carry a wistful or melancholy calm.

For a music tool, the practical reading is that a sustain or darkness control moves perceived calm and perceived sadness together. A tool that offers calm through these controls should say so.

The tier B* places this as a strong signal worth expanding. Three things would raise it: a second rated corpus, a listening test that changes the tail with instrumentation held fixed, and ratings of felt calm.

9 Limitations

- Tail length is partly a proxy for instrumentation and production style. Timbre was not held fixed, so the sound may not cause the rating.
- DEAM ratings are crowd judgments of perceived emotion in 45-second clips. They are not felt calm, relief or any change in mood.
- Room reverberation could not be measured apart from note decay with these tools. The finding is about the combined tail.
- The audio came from a public mirror whose file count matches the published set but which has no official checksum.
- The instrument check used one family of synthetic music. Real mixes can behave differently.
- The genre and note-rate checks were run after the verdicts.
- A later review of dose-response shapes across our findings fitted two lines to this relation and read it as a fall that levels off, with no U shape (quadratic term p 0.012). That review is cited from our ledger and was not re-run for this paper.
- Scoring was done by our own code and was not blind. The ratings come from DEAM's raters and stand apart from the code.

10 Verified

Verified: on 9 November 2026 the counts (1,744 clips, 502 test clips, 270 test artists, 1,242 training clips), the partial correlations -0.394, -0.301, +0.324, -0.375 and -0.343 with their intervals, the R2 gain 0.110 and its interval, the controls-only R2 0.160, all three shuffle controls, the whole-sample values and every verdict were re-read from the run's result file. The post hoc values (-0.227, -0.370, -0.285, -0.324, 0.54, note rate and the 11 genre values) were re-read from the post hoc file, and the instrument ratios and rank correlations from the instrument file. These match the findings ledger. The preregistration's SHA-256 matches the hash recorded in the result file. All three figures were drawn by a script that reads only those three files. The study was not re-run for this paper. The literature was found by searches run on 9 November 2026 and read at the level of abstracts, repository records and publisher pages; the full texts were not read.

11 Corrections

- 29 October 2026: the audio loader's resampler changed after the lock and before feature extraction, recorded at the time; the instrument check does not pass through the loader.
- 29 October 2026: a later review of dose-response shapes in our ledger read the arousal relation as a fall that levels off. This paper states it in the limitations.
- 9 November 2026: first published version. The two questionnaire tests in the same preregistration are left out because they were not blind.

12 Reproduction

- Data: DEAM [1], the 45-second clips with static per-clip arousal and valence means, and the dataset's metadata for artist and genre. Licence CC BY-NC.
- Convert each clip to mono at 22,050 Hz and keep the first 45 seconds. Measure the tail, eq. 1, and the spectral slope; check both on synthetic music with planted decay and tilt first.
- Hold out 30 percent of artists. Build the calm index, eq. 2, with training-set scaling. Compute partial correlations, eq. 3, with loudness, genre and year as controls, and the held-out R2 gain, eqs. 4 and 5. Use 2,000 artist-bootstrap resamples for intervals and 1,000 shuffles for each control.

- Preregistration SHA-256: ae20eac8980611bc07f8e082e3f07c636229179a7ee051c7305b7423e32b89c7. Plan SHA-256: b051a324f52e1dda7802c57aa3a0d6e33ca3f4db0b51b750e810360df73fefc6. Run 20261029T162517Z.
- Cost: the scored analysis and its controls took under a minute on one desktop computer. Feature extraction time was not recorded. The data used about 2.6 GB of disk with the audio unpacked.

13 Appendix A. The preregistered verdicts

Each verdict is a set of rules written before the run. A verdict passes only when every rule holds. Values are read from the run's result file.

Table 3. Preregistered verdicts and the rules that decided each one. The two questionnaire verdicts are omitted (see the preregistration section).

Verdict	Deciding rules	Value	Holds
Instrument	tail strong; slope strong; constant estimator fails	yes; yes; yes	yes
P1	r below 0; interval upper end below 0; r below shuffle 5th percentile (-0.076)	-0.394; -0.304	yes
P2a	r below 0; interval upper end below 0	-0.301; -0.200	yes
P2b	r above 0; interval lower end above 0	0.324; 0.217	yes
P3	r above 0; interval lower end above 0; r above shuffle 95th percentile (0.077)	-0.375; -0.455	no
P4	gain above 0; interval lower end above 0; gain above shuffle 95th percentile (0.003)	0.110; 0.034	yes
P5	r below 0; interval upper end below 0	-0.343; -0.246	yes

14 Glossary

Each term is defined where it is first used in the text. The same definitions, with the field each comes from and what each excludes, are collected here.

- **preregistration**. A dated written plan for a study, fixed and hash-locked before any result is seen: the question, the data, the statistic, the controls, the forecasts and the kill conditions. Field: research method (metascience). Excludes a plan revised after looking at the outcome, and a protocol written only for the record after the run.
- **verdict**. The pass or fail of one preregistered rule set against the run's numbers, read off by code. Field: research method (our usage). Excludes an interpretation written after reading the numbers.
- **arousal and valence**. Two rating scales for emotion: arousal runs from calm to excited and valence from unpleasant to pleasant; in DEAM each is a nine-point scale averaged over crowd raters. Field: psychology of emotion (the circumplex model). Excludes named emotions such as sadness or nostalgia, which the ratings place only as points on the two scales.
- **perceived emotion**. The emotion a listener judges the music to express, as opposed to the emotion the listener feels. Field: music psychology. Excludes felt or induced emotion, physiological response and any change in mood or symptoms.
- **onset-locked tail**. For each note onset found in the energy envelope, the time constant of the envelope's decay from its peak to the next minimum, with the median taken over onsets; it follows whichever of the note's own decay and the room's decay is slower. Field: audio signal processing (our measure). Excludes the reverberation time of the room on its own, which this measure cannot separate in dense music.
- **spectral slope**. The slope of log power against log frequency between 110 Hz and 5.5 kHz; a more negative slope means energy leans toward low frequencies and the sound is darker. Field: audio signal processing. Excludes pitch, loudness and the shape of the spectrum beyond its overall tilt.

- **held-out test set**. Data kept apart from everything that was fitted or standardised, here every clip by 30 percent of the artists, so that a result on it is a check on new material. Field: statistics and machine learning. Excludes data used to fit the model, choose a threshold or compute a scaling.
- **partial correlation**. The correlation between two quantities after each has had the linear effect of a set of control variables removed by least squares. Field: statistics. Excludes a causal effect, and any non-linear dependence on the controls.
- **held-out R2 gain**. The share of variance in the test outcome explained by a model fitted on training data with the added measures, minus the same share for the model without them. Field: statistics. Excludes the in-sample R2 of a fit, which can only grow as variables are added.
- **95% interval**. The range that holds the middle 95 percent of the bootstrap or control draws for a number; we report it beside the number. Field: statistics. Excludes a statement that the true value lies inside with 95 percent probability.
- **cluster bootstrap**. A bootstrap that resamples whole groups, here artists with all their clips, so that clips by one artist count as one draw. Field: statistics. Excludes a bootstrap over single clips, which treats one artist's clips as independent.
- **planted check**. Data built so that the right answer is known, run through the same analysis before the real data, to show that the analysis can see the effect when it is there. Field: research method; called a positive control in experimental science. Excludes any check made on the real data, which cannot show what the analysis misses.
- **control**. A version of the data with the structure under test removed (for example shuffled or pooled), scored the same way as the real data, so the real score has something to beat. Field: statistics and experimental design. Excludes a comparison with a different dataset or with a published number.
- **null distribution**. The spread of scores a control gives over many random draws; its 95th percentile is the bar a real score must clear. Field: statistics. Excludes a theoretical distribution assumed rather than drawn from the data.
- **honest null**. A working test that found no effect beyond a stated bound, a forecast that failed, or an error of our own found and fixed, reported with the same weight as a positive result. Field: research method (our usage). Excludes a void run, and a result left unpublished because it was negative.
- **tier**. How ready a finding is to publish: A candidate, B signal worth expanding (B* strong), C internal, D honest null worth reporting, E not reliable as it stands. Field: our publication register. Excludes a measure of how large or how true an effect is.

15 References

[1] Aljanaki, A., Yang, Y.-H. and Soleymani, M. (2017). Developing a benchmark for emotional analysis of music. PLoS ONE 12(3), e0173392. doi:10.1371/journal.pone.0173392

[2] Eerola, T., Ferrer Flores, R. and Alluri, V. (2012). Timbre and affect dimensions: evidence from affect and similarity ratings and acoustic correlates of isolated instrument sounds. Music Perception 30(1), 49-70. doi:10.1525/mp.2012.30.1.49

[3] Korsmit, I. R., Montrey, M., Wong-Min, A. Y. T. and McAdams, S. (2024). The acoustic properties of affective timbres: consistencies and discrepancies in a synthesis of multiple datasets. Music and Science. doi:10.1177/20592043241256012

[4] Mo, R., Wu, B. and Horner, A. (2015). The effects of reverberation on the emotional characteristics of musical instruments. Journal of the Audio Engineering Society 63(12), 966-979. doi:10.17743/jaes.2015.0082

[5] Wilkie, H. and Harrison, P. M. C. (2025). Reverberation time and musical emotion in recorded music listening. Music Perception 42(4), 329-355.