Instruction information and spectral sensitivity in a musical model of speech and song

Amaç Erdem
Musical Intelligence research project · Independent research

Abstract

A model can retain information about vocal expression without assigning its native affect scores the expected direction. We examined this distinction in Musical Intelligence (MI), a system designed for musical listening, using an instruction factorial and a subsequent acoustic intervention. The first study analysed 342 synthetic recordings from 172 requests. Across held-out sentences, a fixed probe decoded five speech instructions from 889 C^3 dimensions with 68.9% balanced accuracy, compared with 57.8% from 15 acoustic descriptors and 35.6% from two native affect outputs. Yet native valence was lower for excited than sad speech in every text. A second study selected six recordings and independently varied an instrument-derived upper-harmonic envelope and residual retention, yielding 1,596 primary full-recording evaluations. Mean native valence rose from 0.139 to 0.435 for cello-derived and 0.437 for viola-derived envelopes, while all 588 primary paired-anchor contrasts retained the inverse instruction ordering. Residual removal produced a substantially larger shift than envelope transfer alone. Restoring the source consonance-feature trajectories suppressed 94.8% and 94.2% of the mean endpoint rises. Arousal ordering depended on waveform level, and 192 near-endpoint probes exposed substantial score differences between very close waveforms. These results distinguish decodable instruction information, native affect mapping and numerical sensitivity. They identify a computational route through which speech-derived texture changes affect scores, without establishing consonant-specific causality, natural instrument equivalence or human perceptual thresholds.

Introduction

How does a natural sound become a perceptual object, and when does its organization become musical? These questions require more than detecting that two waveforms differ. A useful computational account must specify which acoustic distinctions survive into its internal representations, which distinctions are collapsed, and how the resulting variables acquire an interpretation. Speech provides a demanding case: words, voice quality, timing, pitch and expressive intent arrive together. Holding the intended words constant permits a comparison of vocal form, but does not remove meaning from a listener's experience.

Research on vocal expression and musical performance identifies partially shared acoustic routes for communicating emotion.[1] Speech and melody also depend differently on spectral and temporal information: controlled degradation of sung material can dissociate their perceptual intelligibility.[2] These observations motivate a common representational analysis without assuming that speech and song are interchangeable. An established human corpus, RAVDESS, uses matched statements in emotional speech and song;[3] the present study instead uses generative instructions to vary delivery while preserving an explicit, inspectable production reference.

A music-oriented model offers a specific transfer test. Fixed words can retain different realizations of frication, bursts, phonation and vowel transitions. These distinctions may alter features used by a musical affect computation without changing the intended emotional instruction. A speech-versus-song comparison alone cannot locate that dependence: singing still contains words, while pitch, duration and energy also change. The relevant question is therefore whether a controlled intervention in speech-derived acoustic texture changes native affect direction, and which implemented pathway carries that change.

Study 1 audits instruction information, native affect scores and level dependence in a frozen corpus. Study 2 follows the resulting mismatch with a source-phase-preserving spectral intervention, separating upper-harmonic envelope transfer from attenuation of the residual outside a harmonic mask. This second study extends the transformation logic of the birdsong experiment but does not assume that the endpoint is a natural instrument performance or perceived music. Abstraction from acoustic detail and induction from natural regularities remain motivating questions; the present tests concern the model's explicit dependence on controlled inputs.

We distinguish three outcomes throughout: information recovered by a supervised probe, the direction of the model's own scores, and stability under acoustic or numerical intervention. Anchor density is a property of stimulus sampling, not replication. The six intervention sources belong to the original corpus, and neither study measures human emotion judgments. The combined evidence explains a large shift within the computation while leaving the inverse excited–sad valence ordering unresolved.

Study 1 | Instruction factorial and experimental units

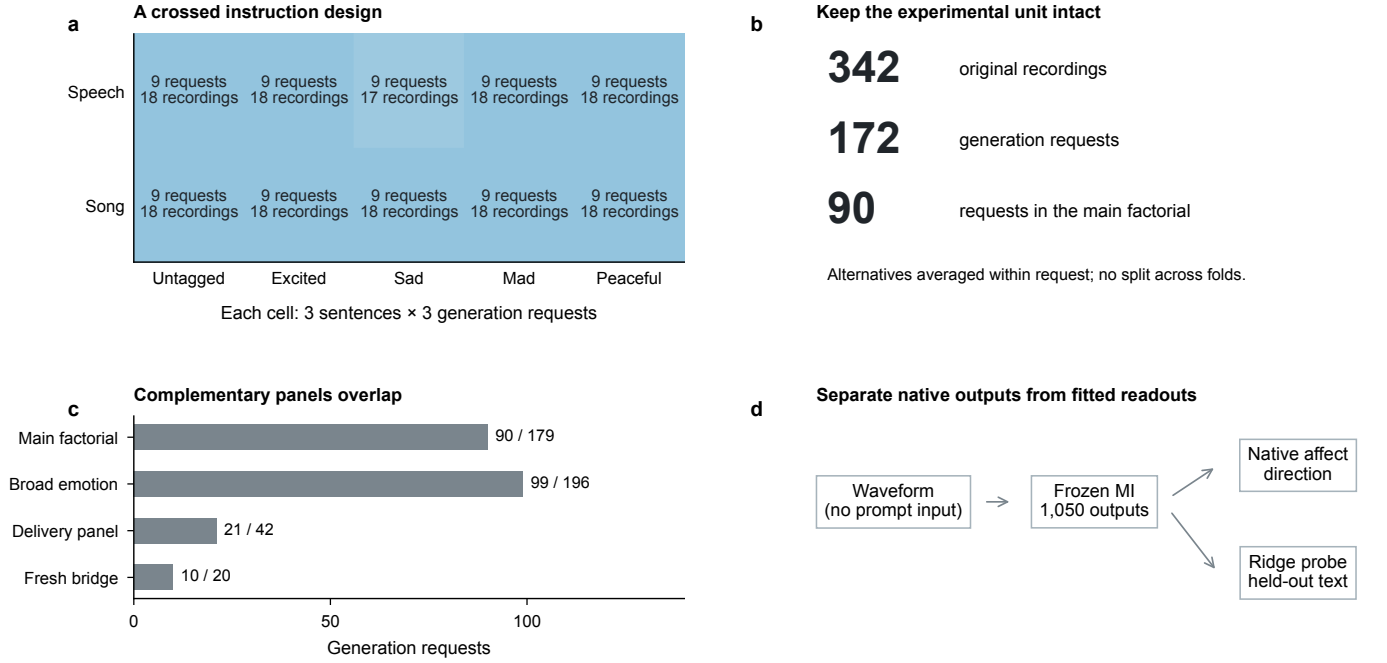

Fig. 1 | Same intended words, crossed expressive instructions and preserved generation units. a, Main factorial: three texts × two delivery modes × five emotional conditions × three requests, yielding 90 requests and 179 recordings. Counts are pooled across texts; one alternative is unavailable. Untagged means no explicit emotional instruction, not verified emotional neutrality. **b,** The complete corpus contains 342 recordings from 172 requests; available alternatives are averaged before inference. **c,** Panel sizes are requests / recordings. Panels overlap and must not be added. The fresh bridge contains one request per speech/song × emotion cell for the already known S01 text. **d,** MI receives waveforms only. Native-score interpretation and supervised instruction decoding are evaluated separately; all alternatives of a request remain together.

The main design crossed three English sentences with requested speech or unaccompanied song and five emotional conditions: untagged, excited, sad, mad and peaceful. Each of the 30 cells contained three generation requests. The corpus was assembled in stages: the initial three-emotion factorial contributed 107 recordings; a two-emotion extension contributed 72. A broader emotional panel, a seven-delivery panel and a fresh, randomized production bridge expanded the available comparisons without increasing the number of distinct texts. These panels share recordings, so the corpus-wide total is 342, not the sum of panel totals (Fig. 1).

Generation used the displayed Eleven v4 model and a single displayed voice, Anna—Young Female Conversational Voice, with stability 0.5 and similarity 0.75. Instructions were entered as natural-language audio tags, consistent with the provider's documented performance-direction interface.[4] Speech prompts combined an emotional tag with the intended sentence; song prompts additionally requested "singing a simple unaccompanied melody". The actual prompt strings, production order, receipts, available alternatives and file hashes are retained. The interface did not expose a recoverable model checkpoint, voice identifier or seed. Consequently, the existing recordings, rather than a claim of exactly regenerable synthesis, define the experiment.

The statistical unit is the generation request. Alternative renderings from the same request are not treated as independent participants or independent replications. For each recording, model trajectories were summarized separately; available alternatives were then averaged within request. Main contrasts weight the three fixed sentences equally. Prompt labels are the reference throughout. No human rating, automatic-transcription disagreement or agreement with MI determined inclusion. A transcript audit recorded exact matches for 336 of 342 files; all six disagreements remain in every eligible analysis. This audit documents lexical uncertainty rather than establishing perceptual or semantic equivalence.

The evidence hierarchy follows the question being asked. Native contrasts ask whether an implemented scalar follows a specified instruction ordering. Held-out-sentence probes ask whether information generalizes to another one of the three texts. Same-text generation tests ask a weaker, different question about production variability. The bridge asks whether an existing mapping transfers to a fresh production block of known text. None of these tests, alone or together, establishes how a human listener would label the audio. This separation is essential because the comprehensive analysis exposes 1,050 correlated outputs and many opportunities to select a persuasive result.

The response archive is deliberately broader than the headline affect comparison. Five scalar summaries retain changes in central tendency, spread and extremes; the temporal analysis retains within-phrase organization; grouped prediction asks what survives a text change; and the paired level intervention probes a specific dependence of the computation. This design allows a negative native-score result to coexist with a positive distributed-information result without treating either as an exception to be hidden. It also prevents the number of frames, derived dimensions or processed copies from being mistaken for independent replication.

Native affect separates some instructions and reverses others

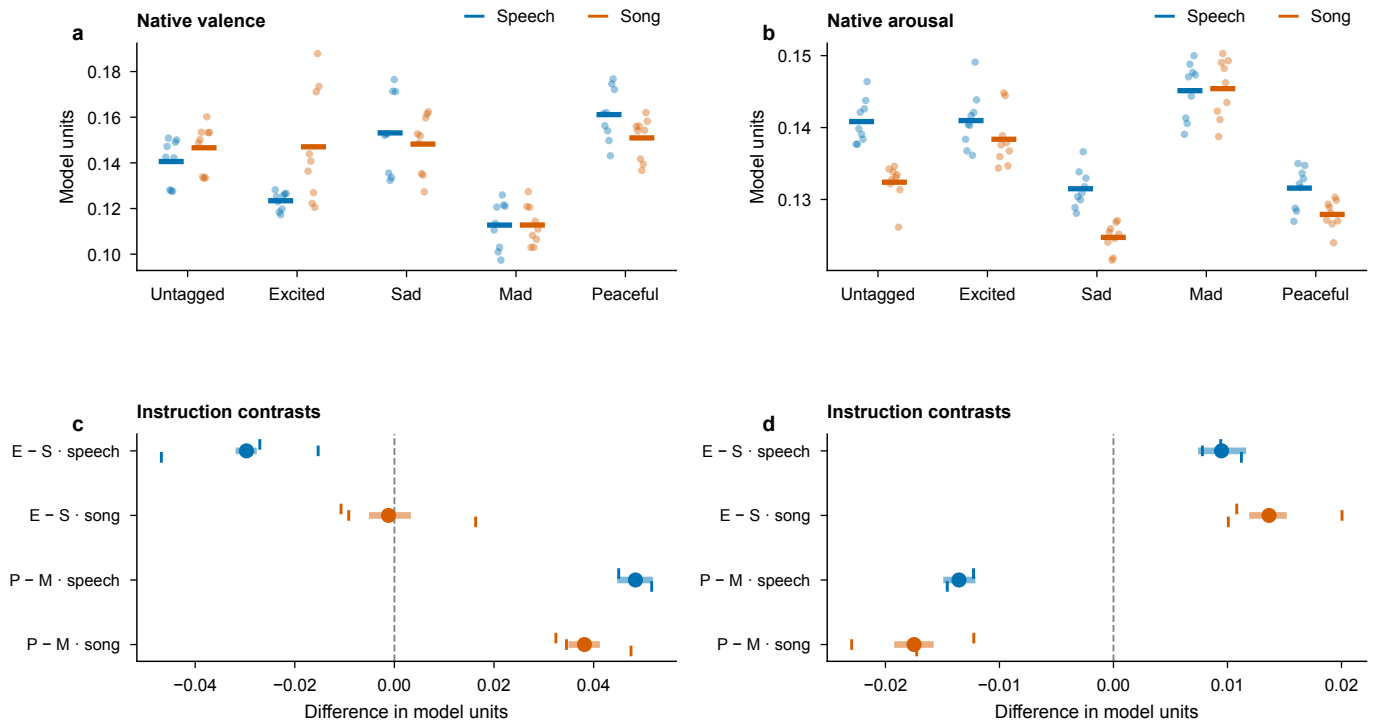

Fig. 2 | Instruction-dependent native affect is not uniformly directionally calibrated. a,b, Request means for valence (VMM:C1) and arousal (AAC:P2); nine requests per emotion and delivery mode, with three requests from each text. Points show requests; short horizontal marks show pooled means. **c,d,** Excited minus sad and peaceful minus mad contrasts. Dots are equal-text means, vertical ticks are the three text-specific contrasts, and thick lines are pointwise 95% percentile intervals from 1,999 within-cell request bootstraps. Intervals are conditional on the fixed texts and unadjusted for selection or multiplicity. Native units are retained. Speech is blue and song orange. The intended excited-minus-sad valence direction is positive, yet all three speech contrasts are negative.

The native valence result was contrast-specific. Excited-minus-sad speech produced a mean difference of −0.029705 model units, with text-specific differences of −0.027019, −0.015310 and −0.046785. Thus, the direction opposed the positive-valence ordering specified for this comparison in every text. Song did not reproduce a uniform reversal: its mean difference was −0.001187, with one positive and two negative text-specific differences. Peaceful-minus-mad valence was positive in all three texts for both speech and song, averaging 0.048399 and 0.038158, respectively. MI therefore showed a selective mismatch rather than an absence of instruction sensitivity (Fig. 2).

Arousal initially appeared more consistent. Excited-minus-sad differences were positive in all three texts for both speech and song, averaging 0.009479 and 0.013646. Peaceful-minus-mad differences were negative in all three texts, averaging −0.013550 and −0.017489. These are results about the direction of model outputs relative to intended production contrasts. Excited versus sad and peaceful versus mad provide useful operational hypotheses; they do not place all possible emotions on a universal two-dimensional scale, and the untagged condition is not an independently established neutral anchor.

The complete scalar analysis evaluates 556 contrasts using five within-recording summaries across all 1,050 outputs, yielding 2,919,000 effect values. Exact tests are available for 303 unique eligible contrasts, comprising 318,150 contrast–output hypotheses. Global Benjamini–Hochberg adjustment gives 9,689 findings at $q < 0.05$; the dependence-robust Benjamini–Yekutieli adjustment gives none, with minimum $q = 0.1293$.[5,6] Both results are reported because dependence and family definition materially change the inferential conclusion. The main native contrasts are consequently interpreted through their direction, magnitude and controlled sensitivity, not promoted as a set of multiplicity-robust scalar discoveries.

The inferential resolution is constrained by replication. A within-text comparison has three requests against three, only 20 possible label partitions, and a minimum two-sided exact P value of 0.1. Pooling contrasts over three texts increases permutation resolution but does not turn three sentences into a representative sample of language. Bootstrap intervals describe generation variation conditional on those texts. The apparently consistent arousal ordering is therefore a candidate property of this frozen corpus and model, whose dependence on waveform level is tested directly below.

Speech–song differences are also evaluated directly rather than inferred by comparing significance labels. The archive includes 30 sentence-specific and ten equal-text delivery-by-emotion interactions. These differences of differences ask whether an emotional contrast changes with requested musical delivery. Their availability matters because a near-zero song average can conceal opposing text-specific effects, as in excited-minus-sad valence. Neither a smaller mean nor an absent discovery alone establishes equivalence between delivery modes.

Distributed C³ retains information that two affect scores miss

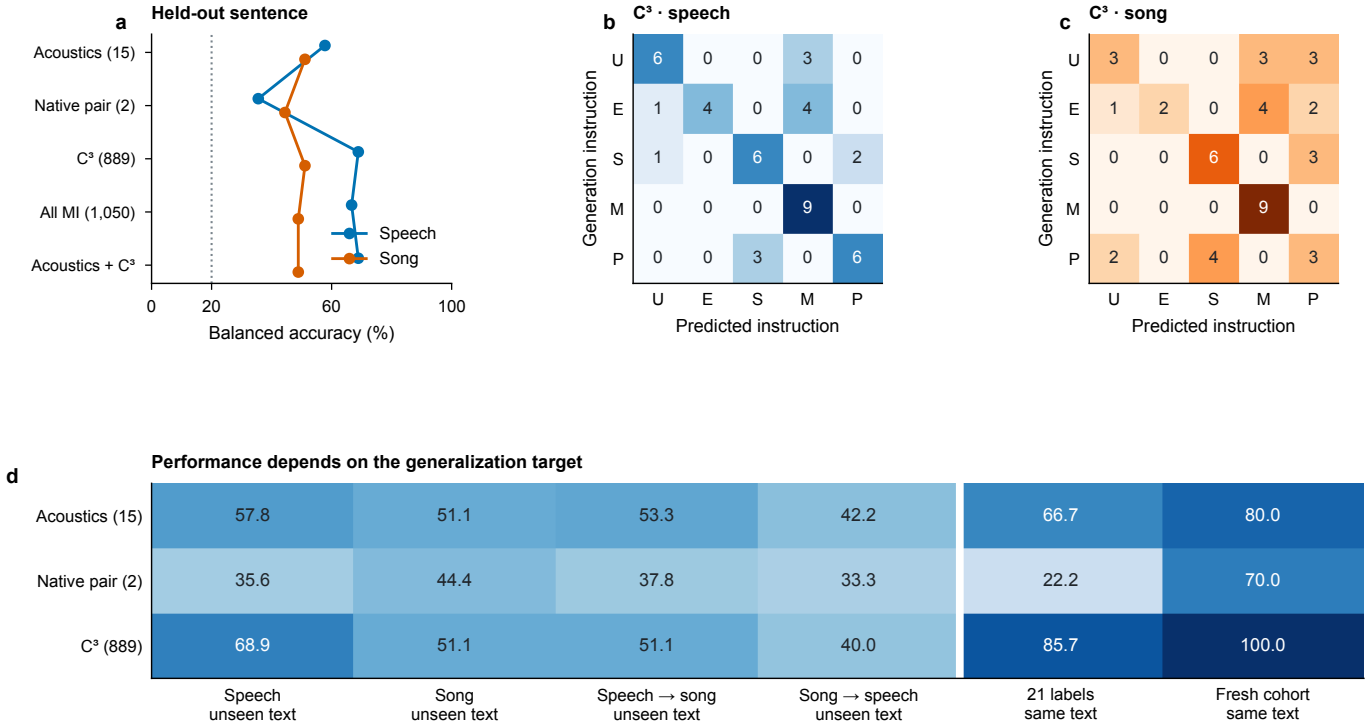

Fig. 3 | Native-score interpretation and fitted decoding answer different questions. a, Five-way emotion decoding with an identical fixed ridge probe and sentence-held-out folds. Each mode contributes 45 test requests, nine per class; dotted line, nominal 20% chance. "Native affect + probe" fits the same decoder to two outputs and is not a native emotion classifier. **b,c,** C³ confusion matrices, pooled over held-out texts; entries are request counts (U, untagged; E, excited; S, sad; M, mad; P, peaceful). **d,** Balanced accuracy (%) across distinct generalization targets. The 21-label task tests 63 requests on one text; the fresh bridge tests ten requests on known text. These columns are not equivalent replications of unseen-text performance. The speech C³–acoustic difference is descriptive, not a tested superiority claim.

Five-way speech decoding reached 68.9% balanced accuracy from C³, corresponding to 31 correct predictions among 45 equally weighted test requests. The 15-feature acoustic baseline reached 57.8% (26/45), and a fitted probe using native valence and arousal reached 35.6% (16/45). Adding acoustics to C³ did not improve the speech result, while concatenating all 1,050 MI outputs yielded 66.7%. For song, C³ and acoustics both reached 51.1% (23/45); the native pair reached 44.4%, and the combined acoustic–C³ representation reached 48.9%. More outputs therefore did not monotonically improve generalization (Fig. 3).

Errors were not distributed uniformly across labels. The C³ probe correctly classified all nine mad requests in each delivery mode. Excited requests were harder: four of nine were correct in speech and two of nine in song. Peaceful speech yielded six correct predictions, whereas peaceful song yielded three; confusions with sad accounted for three and four cases, respectively. Thus, the same overall accuracy can conceal qualitatively different class-specific information. The confusion matrices preserve those differences instead of collapsing them into a single performance claim.

The C³ speech result demonstrates accessible instruction information under the specified folds. Its restricted label-permutation P value was 0.001 and its globally adjusted BY q value was 0.00722; C³ song and the speech acoustic baseline also passed this family-level criterion. Across 308 canonical decoder tests, 291 survived global BY adjustment. These nulls assess correspondence between representations and labels conditional on the permitted production strata. They do not test the 11.1-percentage-point difference between C³ and acoustics, prove a cognitive explanation, or convert a fitted classifier into an already calibrated component of MI. The distinction between represented information and what a probe learns is central to interpreting diagnostic classifiers.[7]

Transfer also depended on the test regime. Training C³ on speech and testing song on an excluded text gave 51.1%; the reverse direction gave 40.0%. Same-text tasks were easier: C³ reached 85.7% for 20 emotional instructions plus an untagged condition and 90.5% for seven delivery instructions. The joint acoustic–C³ representation reached 90.5% in the 21-label task. These tasks retain the same lexical material across training and test and should not be described as stronger evidence of text generalization. Their role is to show that the response repertoire distinguishes a wide instruction set when text is controlled.

The fresh bridge was predicted correctly for all ten requests by C³, compared with eight by acoustics and seven by the native pair. This is a useful production-block check, but its single request per cell, familiar S01 text and shared voice make 10/10 a narrow finding. The manuscript therefore retains all 15 tasks and three regularization settings rather than selecting the most favourable percentage. Cross-validation on small samples is intrinsically uncertain;[8] the archive preserves fold identities and every held-out prediction so that errors, dependence and generalization scope can be inspected directly.

Study 2 | Spectral intervention shifts valence without repairing its ordering

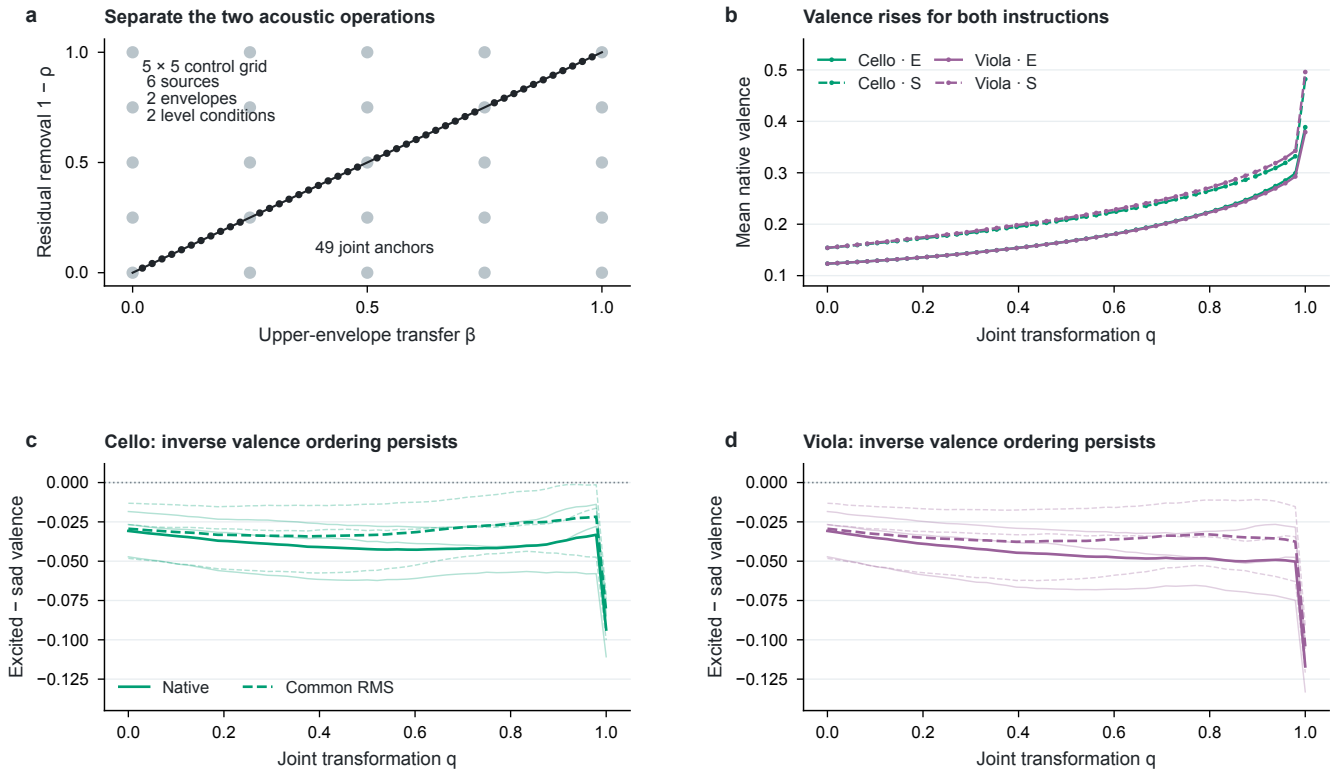

Fig. 4 | Acoustic intervention and native valence trajectories. a, Upper-harmonic envelope transfer β and residual multiplier ρ are independently controlled; the diagonal follows β = q, ρ = 1 − q at 49 evaluated anchors. A 5 × 5 grid supplies factorial controls. **b,** Mean native valence over three fixed texts, separately for excited (E, solid) and sad (S, dashed) sources; colours denote cello- or viola-derived envelopes. Dots mark evaluated anchors. **c,d,** Excited-minus-sad valence for cello and viola: thin lines are individual texts, thick lines their mean; solid denotes native component energy and dashed common RMS. No confidence band treats anchors as independent observations. All 588 primary contrasts are negative. Connecting lines are visual guides; Fig. 6 examines the steep final interval.

Six complete speech recordings were selected without reference to their MI scores: the first available take and lexicographically first alternative for excited and sad instructions in each text. A deterministic renderer retained source phase and sample count while transferring an instrument-derived upper-harmonic envelope and attenuating the complementary residual. Unlike learned timbre-transfer systems,[9] this renderer was not trained to imitate an instrument. The hypothesis concerns an acoustic intervention, not a validated speech-to-cello conversion.

An independent pitch audit prompted a documented amendment before transformed MI outcomes were inspected. Unconstrained transfer produced octave-tracker disagreement in 8.3–24.6% of co-voiced frames at sad cello endpoints. Protecting the lower harmonics reduced the maximum across the guarded pilot to 3.08%; median absolute F0 disagreement was 0–10 cents and the maximum 95th percentile was 80 cents. Remaining deviations are reported rather than treated as exact prosody preservation (Supplementary Fig. 9).

Across six sources, mean native valence increased from 0.13867 to 0.43535 for cello and 0.43746 for viola envelopes. Individual native endpoint increases ranged from 0.2380 to 0.3607. However, every primary excited-minus-sad contrast remained negative: three texts × two instruments × 49 anchors × two level conditions = 588 descriptive comparisons. Native endpoint differences ranged from −0.0844 to −0.1330 and were more negative than their source differences (Fig. 4).

Thus the intervention altered absolute score level without recovering the proposed instruction ordering. The two envelopes produced similar mean endpoint shifts, but the experiment does not establish which instrument is spectrally closest to this voice. Cello samples bracketed the tracked range; some viola comparisons used a lower-edge envelope outside the sampled instrument register, without octave transposition. These conditions do not constitute six independent speakers or independent natural instrument performances.

Residual attenuation and consonance features account for most of the shift

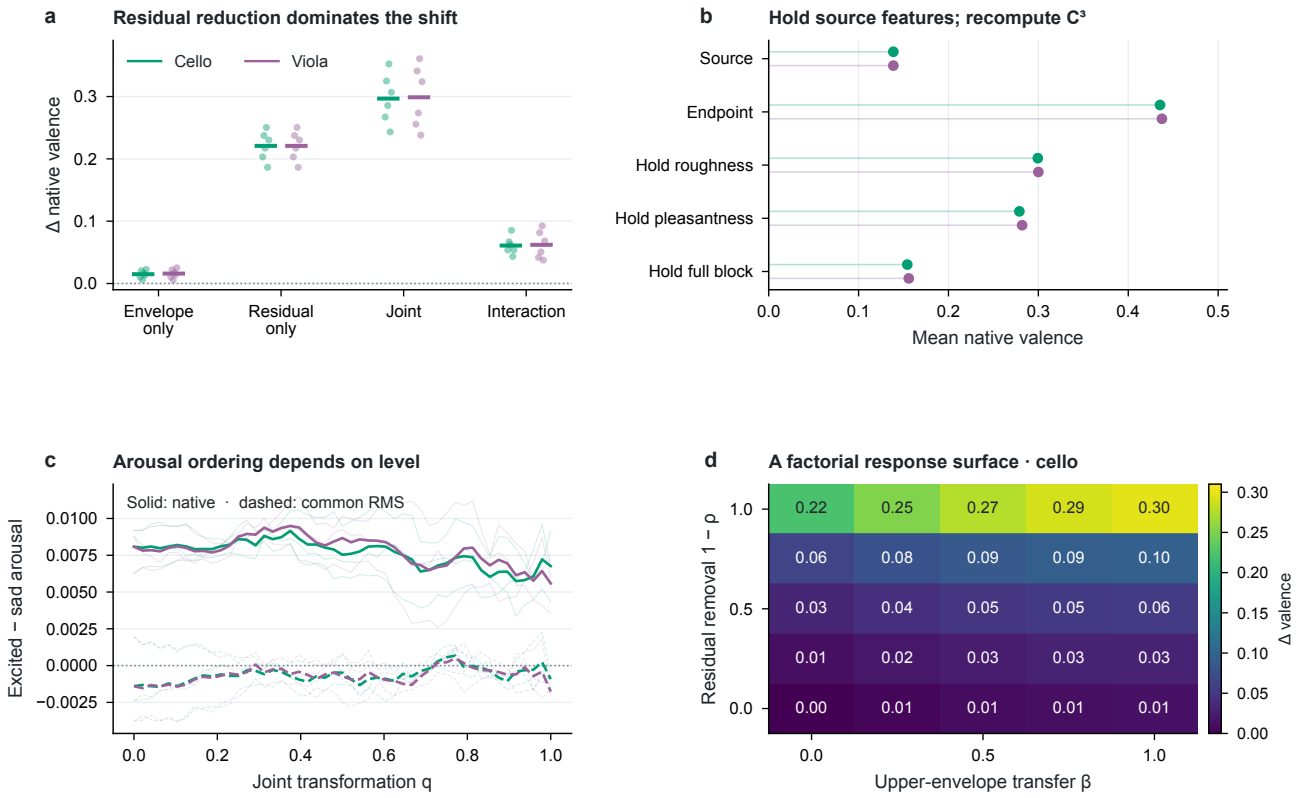

Fig. 5 | Separate signal interventions from model-feature interventions. a, Native valence changes at factorial corners: envelope-only, residual-only and joint endpoint shifts relative to source; interaction = joint − envelope-only − residual-only. Small dots are the six fixed source clips; horizontal marks their mean. **b,** Mean native endpoint valence after restoring source roughness, pleasantness or the seven-feature consonance block and recomputing all downstream responses. These hybrid feature vectors need not represent realizable sounds. **c,** Excited-minus-sad arousal; thin lines show texts, thick lines means, with native (solid) and common-RMS (dashed) conditions. **d,** Cello/native factorial surface, mean within-source valence change. All four surfaces appear in Supplementary Fig. 7. No panel measures listener emotion.

The factorial controls separate two contributors that a single joint path would confound. Mean native valence changed by +0.01495 (cello) or +0.01591 (viola) under upper-envelope transfer alone, compared with +0.22083 under residual removal alone. Joint changes were +0.29668 and +0.29879; finite-difference interactions were +0.06090 and +0.06205 (Fig. 5a,d). This interaction is a deterministic comparison of selected signal conditions, not a population interaction test.

Feature interventions then tested the implemented route. At each of 24 source-by-instrument-by-level endpoints, the roughness feature, pleasantness feature or full seven-feature consonance block was replaced by the corresponding original-source trajectory, and all temporal requests and C^3 outputs were recomputed with fresh state. Seventy-two interventions were accompanied by 24 no-change controls, each reproducing its endpoint exactly. Holding the full block suppressed 94.8% of the mean native cello rise and 94.2% of the viola rise (Fig. 5b). Individual roughness and pleasantness holds overlap computationally and must not be added as independent mediated fractions.

The inverse excited–sad endpoint ordering survived those feature holds. The experiment therefore explains much of the absolute valence rise without showing that the same route is sufficient to repair instruction ordering. Residual attenuation is not equivalent to removing consonants: the residual includes signal components left by a particular pitch-dependent mask, while phonetic identity also resides in voiced spectra and transitions. The intervention changes unvoiced energy and cannot isolate letters, lexical meaning or a listener's response to phonemes.

Arousal supplied a distinct dependence. Excited exceeded sad in all 294 native primary comparisons but only 94 of 294 common-RMS comparisons; all common-RMS endpoint contrasts were negative (Fig. 5c). The matched target was −37.38699 dBFS, fixed from the technical pilot. This sensitivity demonstrates that native arousal ordering is not invariant to level under the tested intervention; it does not establish equal perceived loudness.

The response extends across C³, with a numerically sensitive boundary

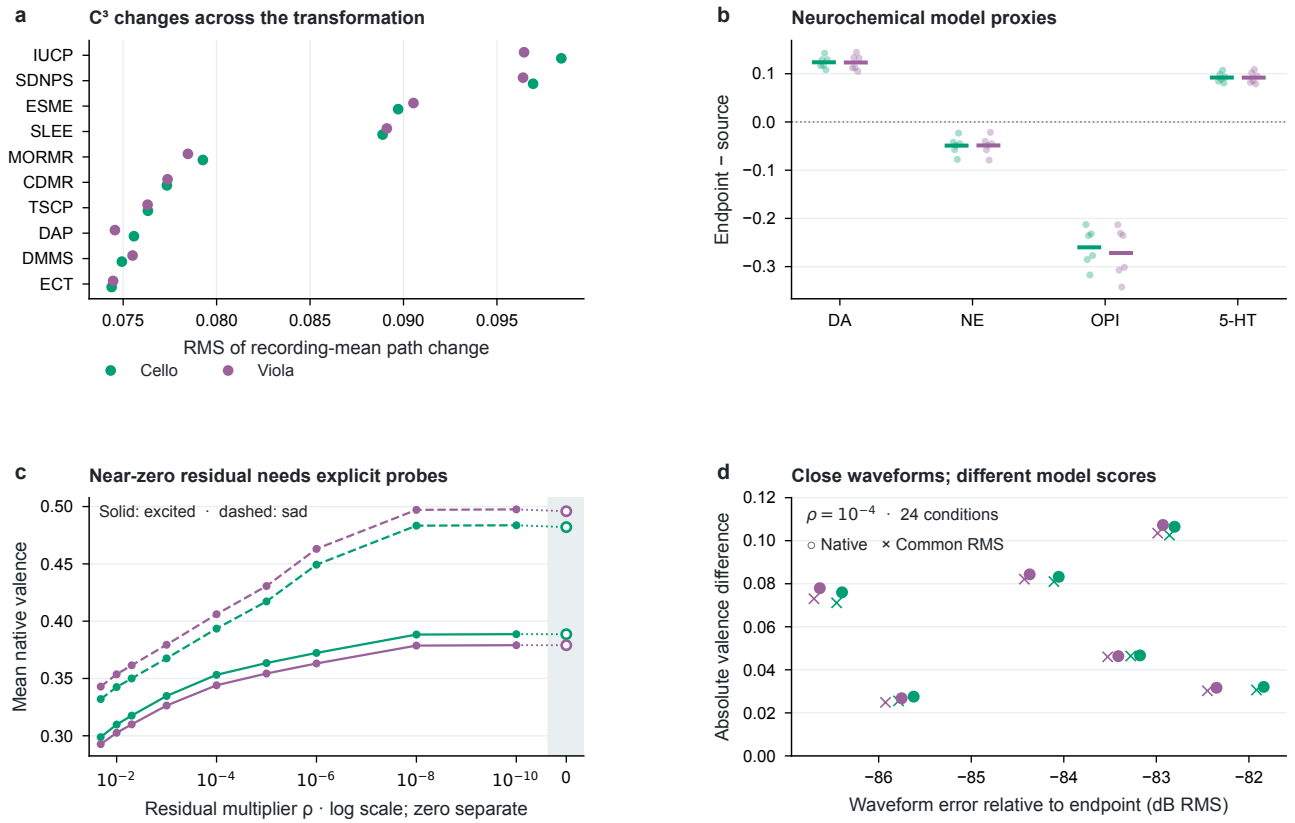

Fig. 6 | Distributed model response and the near-zero-residual boundary. a, Ten mechanisms with greatest native cello path RMS, compared with viola in the same order; RMS averages squared changes in recording means across six sources, 49 anchors and each mechanism's dimensions. Full rankings are retained. **b,** Endpoint changes in four neurochemical proxies; dots are sources and marks are means. **c,** Native valence near the endpoint; colours denote instruments and line styles instructions. Residual retention uses logarithmic spacing; zero is drawn separately. **d,** At $\rho = 10^{-4}$, waveform error relative to the zero-residual endpoint versus absolute valence difference across 24 source/instrument/level conditions. These are numerical sensitivity measurements, not auditory thresholds or biological responses.

The complete saved response includes 889 C³ dimensions, 131 unstretched combinations, 26 raw regional projections and four neurochemical proxies. IUCP, SDNPS, ESME and SLEE ranked highest by native cello recording-mean path RMS, at 0.0984, 0.0969, 0.0897 and 0.0889. Full names, all 84 ranks and 195 dimensions assigned to present-state layers are retained. Neither a high-ranked mechanism nor its memory-related name establishes a measured cognitive process.

Native cello endpoints changed the DA, NE, opioid and serotonin proxies by +0.1237, −0.0491, −0.2600 and +0.0921 on average. Higher valence therefore did not correspond to a uniform increase across reward-labelled outputs. Regional projections retain unbounded native units and are not compared numerically with bounded affect scores (Supplementary Fig. 8). Most mechanisms use centred or future context; their trajectories are not estimates of real-time human response latency.

Sampling diagnostics revealed a sensitive final interval. Withholding alternate anchors produced maximum native valence interpolation errors of 0.0844 and 0.0842 for cello and viola. We therefore evaluated eight additional near-endpoint levels, yielding 192 full-file analyses. At $\rho = 10^{-4}$, waveform error relative to the endpoint was only −86.7 to −81.8 dB RMS, while the largest valence gap was 0.1073 (Fig. 6c,d). The smallest tested residual reduced that gap to 0.00287. Of 96 additional paired contrasts, one was weakly positive (+0.000634; S02, common-RMS cello, $q = 0.995$), then negative again at the endpoint.

A separate numerical audit identified a sign-destroying clamp in quadratic peak interpolation.[10] Correcting only that operation in an isolated engine copy reduced pure-tone frequency errors and changed valence, but all native endpoint instruction contrasts remained negative in 72 control evaluations. Fifty-four unprotected-renderer evaluations likewise retained that ordering. These controls constrain simple explanations without validating either implementation biologically. The primary pinned model and the original study remain unchanged.

Discussion

Information, mapping and intervention are different outcomes

Across two connected studies, MI retains information about expressive instructions while its native valence score follows a different ordering. The spectral intervention substantially changes valence and identifies a strong computational dependence on the consonance feature block, yet does not consistently restore excited-above-sad ordering. This combination is more informative than either a positive decoder score or a large morph response alone: representation sensitivity, score interpretation and numerical robustness must be evaluated separately.

What the intervention explains

The large residual-only shift and the feature-clamp results show that the tested acoustic operation reaches valence chiefly through the implemented consonance block. They do not show that consonants cause the original reversal. The intervention removes a decomposition residual, including unvoiced energy, and transfers only upper-harmonic envelopes. The lower harmonics are protected, other phonetic cues remain, and the hybrid feature holds may not correspond to physical sounds. The stronger inference is model-specific: restoring the source values of this block removes most of the absolute endpoint rise under these conditions, while the inverse instruction contrast persists.

Music-to-speech transfer remains a conditional hypothesis

Direct experiments show that acoustic-affect mappings need not be identical across domains. Manipulations of pitch, rate and intensity can have conditional effects in speech and music,[11] and vocal valence relations depend on emotion family.[12] Shared psychoacoustic cues nevertheless predict affect in both domains,[13] while cross-domain learning has yielded asymmetric transfer.[14] More recent representation comparisons find layer- and emotion-dependent similarity,[15] and speaking-to-singing transfer has been more reliable for activation than valence.[16] These results motivate explicit domain tests; they do not license assigning every MI mismatch to a musical mode of listening.

At least three explanations remain: a miscalibrated native mapping, domain-dependent cue weighting, and a mismatch among intended expression, perceived expression, sensory pleasantness and felt emotion. Sad music can dissociate perceived from felt affect,[17] but the present data do not measure either. Study 1 also found negative excited-minus-sad valence in two song texts, preventing a simple speech-failure/music-success interpretation. Emotion-dependent articulation changes segmental spectra,[18] yet phoneme identity and causal contribution require segment-aligned interventions. The present residual operation is an intermediate computational test, not that phonetic experiment.

Numerical stability precedes perceptual interpretation

Large score differences between nearly identical endpoint waveforms qualify the meaning of a smooth-looking morph plot. A line between sampled anchors cannot establish a continuous response, and a computational transition cannot be named a perceptual threshold without an appropriate measurement. The isolated peak correction improves an elementary numerical diagnostic but does not repair the emotion contrast, suggesting that this single operation is not a sufficient explanation. Source-fixed voiced-stratum summaries also retain the inverse endpoint contrast, so averaging newly quiet frames alone is insufficient; full-file context still affects those estimates.

Limits and the next discriminating test

One synthetic voice and three fixed texts limit both studies. The intervention uses a selected take per instruction/text rather than new independent generations; anchor counts are not sample sizes for population inference. Prompt labels describe production conditions, and no rating enters the analysis as a quality gate. Naturalness, intelligibility and musicality of the transformed sounds were not established. Anatomical and neurochemical labels are model projections. A successful decoder or a named mechanism does not validate human computation,[19] and unchanged-waveform, exposure-dependent phenomena such as the speech-to-song illusion require additional state or experimental context.[20]

A decisive follow-up should cross additional voices and held-out texts with phoneme-aligned vowel, consonant and transition interventions, separating residual suppression from formant and envelope changes while controlling pitch, duration and level. Worded and wordless renditions should test phonetic dependence directly. Sensory pleasantness and expressed emotion require separately specified targets. Numerical perturbation tests and adaptive anchor refinement should precede perceptual claims. The current findings supply falsifiable model dependencies and expose an unresolved ordering, rather than settling whether natural sound becomes music through abstraction or induction.

Methods | Study 1: expression and level

Study 1: corpus and experimental units

The frozen corpus contains 342 recordings from 172 generation requests; the main 3-text × 2-delivery × 5-emotion × 3-request factorial contains 90 requests and 179 recordings. Broad-emotion, delivery and fresh-bridge panels overlap and are not added to these totals. The generator displayed Eleven v4, Anna—Young Female Conversational Voice, stability 0.5, similarity 0.75 and 44.1-kHz/128-kbps MP3. Exact checkpoint, voice ID and seed were unavailable. Exact prompts and alternatives define the archived material. Six automatic-transcript disagreements were retained. No human review score or MI agreement determined inclusion. Available alternatives are averaged within request, and texts are equally weighted in pooled contrasts.

Frozen engine and native scores

Both studies use the identical float32 decoding adapter and 855-file engine pin, with fresh state per full recording. The input is 44.1-kHz mono; the mel representation has 128 bands, FFT 2,048 and hop 256, followed by log(1 + power) and maximum normalization with a 10^{-8} floor. The waveform also reaches the acoustic extractor, so the complete pipeline is not assumed gain-invariant. C^3 dimensions, unstretched combinations, raw RAM and neurochemical proxies are retained at every frame. VMM:C1 is the clipped sum of 0.25V0 + 0.20 sensory pleasantness + 0.20 reward evaluation + 0.20(happy − sad) + 0.15(1 − roughness). AAC:P2 combines four equally weighted arousal-labelled products; all terms and clipping adjustments are archived. Exact closure is arithmetic consistency, not biological validation.

Acoustic baseline and level control

Fifteen descriptors cover duration, waveform RMS/peak, active and voiced fractions, pitch probability, F0 median/ranges, centroid, flatness and envelope modulation/peak timing. A separate 16-kHz analysis copy uses FFT 1,024, hop 160 and pYIN over 65–1,000 Hz.[21] Missing descriptors are retained and handled within estimators. Study 1 applies attenuation-only gain $g = r^*/\mathrm{RMS}(x)$, with r^* the smallest source RMS (−29.517516 dBFS), and reruns all 342 files. This is distinct from Study 2's lower pilot-frozen target. Neither manipulation equates perceived loudness.

Contrasts, uncertainty and multiplicity

Five scalar summaries for 1,050 outputs and 556 declared contrasts yield 2,919,000 effects. Eligible tests enumerate 20 or 8,000 within-cohort label partitions; 303 unique tests per output enter a common 318,150-test BH/BY family.[5,6] No scalar survives global BY correction. Pointwise intervals use 1,999 request resamples for 55 equal-text mean/interaction contrasts; they neither generalize over texts nor correct selection. Negative and undefined estimates remain recorded. Detailed contrast membership, standardization, variance decompositions, repeatability, all temporal metrics and 1,890 geometry matrices are preserved in Supplementary S1–S14.

Held-out prediction

Fifteen tasks and 106 feature families give 1,590 fixed-ridge fits ($\alpha = 1$), with 90 $\alpha = 0.1/10$ sensitivities. All alternatives of a request stay together. Unseen-text tasks exclude every request of the test sentence; cross-delivery tasks use the opposite delivery from other texts. Same-text and bridge tasks use their declared generation/cohort splits. Each training fold supplies imputation medians, means, SDs and near-constant-feature filtering; features are divided by sqrt(retained dimension count). Balanced accuracy is mean class recall. A dual ridge implementation agrees with scikit-learn to 7.2×10^{-16}.[22] Canonical tests use 999 exchangeability-restricted permutations, including refitting, with P = (1 + exceedances)/1,000; 308 valid tests form a global BY family.[23,24] Individual mechanism ranks are exploratory. All 300 gain-condition probes retain the same folds without a new inferential family.

Variance, repeatability and acoustic relations

The balanced 90-request main design is decomposed into orthogonal sums of squares for sentence, delivery, emotion, their three two-way interactions, the three-way interaction and within-cell request variation (degrees of freedom 2, 1, 4, 2, 8, 4, 8 and 60). Fractions divide by total sum of squares. ICC(1,1) and ICC(1,3) use 30 conditions with three requests each; negative and undefined estimates are retained. All 170 two-alternative request differences are stored. Acoustic associations are Pearson and Spearman correlations in six declared scopes, with and without within-sentence centring and complete pairs per acoustic descriptor. Acoustic-to-MI regression uses the same training-only ridge preprocessing and sentence exclusion; aggregate R^2 uses pooled held-out residuals against the observed grand-mean denominator, with fold-specific R^2 also retained.

Temporal and geometric analyses

Frame times define normalized phase t/t_{last}. Linear interpolation gives 101 phase samples and 91 common-support samples from 0 to 4.5 s; neither creates new trials. Duration-normalized trapezoidal area, range, total variation per second, least-squares slope per second, peak time, final-minus-initial 20% means and one-second lag correlation are computed before alternative averaging. Only contributing requests determine whether a metric contrast is undefined. Geometry uses condition centroids after scaling by main-request feature SD, removing near-constant dimensions and normalizing Euclidean distance by sqrt(active dimensions). Correlation distances, inter-family rank similarity and cross-delivery emotion-vector cosines/norm ratios accompany all nine panels. Shared matrix edges are not treated as independent observations.

Interpretation of inferential families

The scalar and decoding families address different hypotheses and have different test counts. Zero BY discoveries among 318,150 scalar tests is compatible with above-null multivariate decoding among 308 canonical tests: a distributed pattern can support prediction without any selected scalar satisfying a large-family threshold. Conversely, decoder significance does not locate a unique mechanism. The 84 individual-mechanism families are included for coverage and hypothesis generation, not selected for confirmatory testing after ranking. Regularization sensitivities and gain-condition results are preserved without choosing the condition that maximizes accuracy. No confidence interval treats folds, distance edges, interpolated samples or derived outputs as independent biological observations.

Methods | Study 2: spectral intervention

Study 2: locked sources and template bank

For each text's excited/sad condition, the first take and lexicographically first available alternative were selected before transformed MI analysis. Sources last 4.99–6.43 s. Twenty-four initial solo non-vibrato string samples were extended by eight cello notes; three label-screen failures were excluded, leaving 13 cello and 16 viola templates.[25] Central 20–80% magnitude spectra define harmonic envelopes using measured F0. Log-harmonic envelopes are smoothed and interpolated in log F0; the cello bank spans 64.8–527.6 Hz, viola 130.4–443.6 Hz. Depending on source, 0.3–25.3% of voiced frames lie below C3. Viola edge envelopes are used without octave transposition, not claimed to be physically playable notes.

Signal construction and pre-outcome amendment

Source F0 is estimated by pYIN on a 22.05-kHz copy (frame 1,024, hop 128; 65–700 Hz). At 44.1 kHz, a 2,048-point STFT and 256-sample hop define a pitch-dependent harmonic mask. Let H0 be its reconstructed harmonic component and $R = x - H0$. The output is $y(\beta,\rho) = g[H\beta + \rho R]$. A raised-cosine transition over harmonic orders 2.5–3.5 protects lower harmonics; only the upper envelope is transferred with source phase and separate upper-band energy normalization. This guard followed an independent pitch audit before transformed scores were inspected. Identity conditions use exact decoded source samples. No explicit transpose, time warp or note/rhythm quantization is applied. Residual attenuation intentionally changes unvoiced amplitude.

Harmonic mask and envelope construction

Invalid F0 frames are interpolated only for defining harmonic locations; the voiced gate is Gaussian-smoothed with SD one frame. At each bin, harmonic order is the nearest integer (minimum one). With distance d to that harmonic and width $w = \min(0.22F0, 1.6 \times 44{,}100/2{,}048)$, mask magnitude is $\exp[-0.5(d/w)^4]$ times the voiced gate; DC is zeroed. Template amplitudes at 80 harmonics are floored at 10^{-4} of their maximum, log-smoothed with SD 0.6 harmonics and normalized by harmonic energy. Source amplitudes use a three-bin local maximum before harmonic interpolation. The centred log target/source ratio is clipped at ±36 dB and smoothed by one frame. For each β, the transferred upper band is energy-normalized, recombined with protected lower harmonics, then normalized to the source harmonic-frame energy. Inverse STFT returns the original sample count. Source phase and timing are retained by construction, but independent pitch tracking and envelope correlations quantify imperfect acoustic equivalence.

Grid, full-file evaluation and level

The joint path uses $\beta = q$ and $\rho = 1 - q$ at $q = 0, 1/48, \ldots, 1$, complemented by a 5×5 β/ρ grid at 0, 0.25, 0.5, 0.75 and 1, two envelopes and two level conditions. Shared conditions are deduplicated, giving 1,596 primary evaluations. Native component energy is compared with attenuation-only common RMS at −37.386992 dBFS, frozen as the smallest pilot RMS minus 6.0206 dB. Every full recording uses fresh MI state. Recording means define path contrasts; all frame arrays are retained. Mechanism RMS averages squared mean changes across six sources, 49 anchors and its dimensions. Fixed voiced/unvoiced/quiet source masks define a separate stratum audit, not phoneme labels.

Feature, renderer and numerical controls

Three source-trajectory clamps (R^3 index 0, index 4, or indices 0–6) are applied to 24 endpoints, recomputing temporal requests and C^3: 72 interventions plus 24 exact no-op checks. The archive also contains 54 unprotected-renderer evaluations and 72 isolated signed-curvature peak-code controls. Eight post-primary q values, 0.99 through 0.9999999999, supply 192 boundary evaluations. Total analytic full-file evaluations are 1,914, representing 1,836 unique decoded waveforms; 1,842 file hashes differ partly through WAV metadata. Additional benchmarks/repeats are excluded. Anchor and feature-clamp comparisons are descriptive, with no population P values or anchor-derived confidence intervals.

Boundary metrics and fixed source strata

The additional q values are 0.99, 0.995, 0.999, 0.9999, 0.99999, 0.999999, 0.99999999 and 0.9999999999. Relative waveform error is $20 \log10[\mathrm{RMS}(yq - y1)/\mathrm{RMS}(y1)]$ within the same source, envelope and level condition. It is not a calibrated listening-level comparison. The 25-to-49 diagnostic compares each omitted midpoint with the arithmetic mean of its two neighbours. Source strata are pYIN-voiced frames; unvoiced frames whose source frame RMS exceeds 1% of that source's maximum; and the remaining quiet frames. Masks are frozen across all anchors. Full-file responses are summarized within these masks; no cropped-clip rerun or phoneme segmentation is implied. The isolated peak correction preserves the sign of quadratic curvature when its magnitude exceeds 10^{-8}, replacing the original positive lower clamp. Its separate code pin and complete diff are retained.

Reproducibility and scope

All six original sources reproduce Study 1 arrays exactly; repeat inference and weighted-score closure errors are zero. The primary model pin is unchanged. NumPy/SciPy-based analyses retain all finite outputs, source coordinates, differences, fixed masks, code hashes and limitations.[26] Seventy-eight of 84 mechanisms request centred/future context, with maximum declared horizon 36 s; VMM uses 15 s. The 5.8-ms hop is an offline sampling interval, not biological latency. Full original methods, all endpoints, all mechanism ranks, renderer diagnostics and the boundary exception are printed or indexed in Supplementary S13–S20. All numerical arrays remain in the two linked source archives.

References

1. Juslin, P. N. & Laukka, P. Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol. Bull.* ***129****, 770–814 (2003).* 10.1037/0033-2909.129.5.770.

2. Albouy, P., Benjamin, L., Morillon, B. & Zatorre, R. J. Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science* ***367****, 1043–1047 (2020).* 10.1126/science.aaz3468.

3. Livingstone, S. R. & Russo, F. A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* ***13****, e0196391 (2018).* 10.1371/journal.pone.0196391.

4. ElevenLabs. ElevenLabs Audio Tags list for Eleven v4, with 50+ examples. *Provider documentation; accessed 14 October 2026.* ElevenLabs documentation.

5. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* ***57****, 289–300 (1995).* 10.1111/j.2517-6161.1995.tb02031.x.

6. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* ***29****, 1165–1188 (2001).* 10.1214/aos/1013699998.

7. Hewitt, J. & Liang, P. Designing and interpreting probes with control tasks. *Proc. EMNLP-IJCNLP, 2733–2743 (2019).* 10.18653/v1/D19-1275.

8. Varoquaux, G. Cross-validation failure: small sample sizes lead to large error bars. *NeuroImage* ***180****, 68–77 (2018).* 10.1016/j.neuroimage.2017.06.061.

9. Engel, J., Hantrakul, L., Gu, C. & Roberts, A. DDSP: Differentiable Digital Signal Processing. *ICLR (2020).* Source.

10. Smith, J. O. Spectral Audio Signal Processing: quadratic interpolation of spectral peaks. *Online text, equations 6.30–6.31.* Source.

11. Ilie, G. & Thompson, W. F. A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Percept. 23, 319–330 (2006).* 10.1525/mp.2006.23.4.319.

12. Belyk, M. & Brown, S. The acoustic correlates of valence depend on emotion family. *J. Voice 28, 523.e9–523.e18 (2014).* 10.1016/j.jvoice.2013.12.007.

13. Coutinho, E. & Dibben, N. Psychoacoustic cues to emotion in speech prosody and music. *Cogn. Emot. 27, 658–684 (2013).* 10.1080/02699931.2012.732559.

14. Coutinho, E. & Schuller, B. Shared acoustic codes underlie emotional communication in music and speech—Evidence from deep transfer learning. *PLoS ONE 12, e0179289 (2017).* 10.1371/journal.pone.0179289.

15. Sun, Y., Zhao, Z., Richmond, K. & Li, Y. Exploring acoustic similarity in emotional speech and music via self-supervised representations. *Proc. IEEE ICASSP, 1–5 (2025).* 10.1109/ICASSP49660.2025.10890304.

16. Zhang, B., Mower Provost, E., Swedberg, R. & Essl, G. Predicting emotion perception across domains: a study of singing and speaking. *Proc. AAAI 29, 1328–1334 (2015).* 10.1609/aaai.v29i1.9334.

17. Kawakami, A., Furukawa, K., Katahira, K. & Okanoya, K. Sad music induces pleasant emotion. *Front. Psychol. 4, 311 (2013).* 10.3389/fpsyg.2013.00311.

18. Kienast, M. & Sendlmeier, W. F. Acoustical analysis of spectral and temporal changes in emotional speech. *Proc. ISCA Workshop Speech and Emotion, 92–97 (2000).* Source.

19. Feather, J., Leclerc, G., Mądry, A. & McDermott, J. H. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nat. Neurosci.* ***26****, 2017–2034 (2023).* 10.1038/s41593-023-01442-0.

20. Deutsch, D., Henthorn, T. & Lapidis, R. Illusory transformation from speech to song. *J. Acoust. Soc. Am.* ***129****, 2245–2252 (2011).* 10.1121/1.3562174.

21. Mauch, M. & Dixon, S. pYIN: a fundamental frequency estimator using probabilistic threshold distributions. *Proc. IEEE ICASSP, 659–663 (2014).* 10.1109/ICASSP.2014.6853678.

22. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* ***12****, 2825–2830 (2011).* Source.

23. Ojala, M. & Garriga, G. C. Permutation tests for studying classifier performance. *J. Mach. Learn. Res.* ***11****, 1833–1863 (2010).* Source.

24. Winkler, A. M., Webster, M. A., Vidaurre, D., Nichols, T. E. & Smith, S. M. Multi-level block permutation. *NeuroImage* ***123****, 253–268 (2015).* 10.1016/j.neuroimage.2015.05.092.

25. Philharmonia Orchestra. Sound samples: library and licensing. *Accessed 14 October 2026.* Source.

26. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* ***17****, 261–272 (2020).* 10.1038/s41592-019-0686-2.

27. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* ***2****, 4 (2008).* 10.3389/neuro.06.004.2008.

Data and code availability

The local delivery links to the complete original factorial archive (experiment/) and spectral-intervention archive (morph-experiment/). Both retain audio, framewise outputs, exact conditions, code pins and unrounded arrays. Figure-specific source values accompany editable PDF/SVG artwork; the supplement maps all result families. No public repository accession or DOI has been assigned. Original instrument samples are not redistributed as-is.

Authorship and model-development interest

Amaç Erdem developed the research question and MI framework and assembled the computational study. The author's role in developing the evaluated model is an intellectual interest relevant to interpretation. This manuscript and its analysis presentation were prepared with AI-assisted coding, synthesis and editorial support; responsibility for scientific claims and any submission remains with the author.

Research scope

No participant, brain or neurotransmitter measurements are reported. Both studies concern one archived synthetic voice and three intended texts; derived anchors are not independent replications. This research manuscript does not imply journal acceptance or peer review.