

# Instruction information and level-sensitive affect in a musical model of speech and song

Amaç Erdem

Musical Intelligence research project · Independent research

## Abstract

Speech and song organize the same words into different acoustic forms, but a computational system can distinguish those forms without assigning them an appropriate affective meaning. We investigated this distinction in Musical Intelligence (MI), an executable model designed for musical listening, with a distributed  $C^3$  response and native affect readouts. A corpus of 342 synthetic recordings from 172 generation requests varied emotional and musical-delivery instructions while retaining three intended texts. The main factorial contained 90 requests crossing text, speech or song, and five emotional conditions. Across held-out sentences, a fixed ridge 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. This descriptive advantage did not extend to song, for which  $C^3$  and acoustics both achieved 51.1%. Native valence was lower for excited than sad speech in all three texts, including after common-RMS reanalysis. In contrast, the excited-minus-sad speech-arousal difference declined from 0.009479 to 0.000302 model units after level matching, while  $C^3$  decoding remained above its original accuracy. Exact reconstruction traced these score changes to the implemented weighted terms. Comprehensive scalar, temporal, geometric and prediction analyses therefore distinguish accessible instruction information from native affect mapping and amplitude sensitivity. The results characterize one frozen model and one synthetic voice; they establish neither perceived emotion nor a biological mechanism, but identify precise targets for model revision and independent testing.

## 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.<sup>1</sup> Speech and melody also depend differently on spectral and temporal information: controlled degradation of sung material can dissociate their perceptual intelligibility.<sup>2</sup> 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;<sup>3</sup> the present study instead uses generative instructions to vary delivery while preserving an explicit, inspectable production reference.

The distinction between abstraction and induction offers two possible accounts of this transformation. An abstraction account emphasizes the information discarded when a rich acoustic event is represented through a simpler structure. An induction account emphasizes regularities from which a listener derives structure. Both require a way to identify which details carry weight, yet

neither is established by observing sensitivity to arbitrary signal changes. For a computational model, the first requirement is an auditable relationship between controlled inputs and the full pattern of outputs, followed by a clear separation between sensitivity, prediction and interpretation.

The immediate problem is narrower than a theory of musical origins. Does a model designed around musical sound retain information about emotional instructions in speech? If it does, do its own affect scores express the expected ordering, and do those scores remain stable under a controlled change in waveform level? These are distinct tests. A supervised probe may recover an instruction label from a distributed representation even when the model's native valence or arousal scalar points in the opposite direction. Conversely, a plausible scalar ordering may depend heavily on a nuisance variable that changes with production style.

We address these questions using MI's frozen  $C^3$  implementation. The study extends the experimental logic of our birdsong-to-instrument work from a transformation path to a factorial set of generated vocal anchors. The speech corpus does not yet instantiate a continuous morph, nor does it measure where a listener changes perceptual category. Its contribution is an exhaustive, reproducible audit of the available model response: native scores, all mechanisms, regional and neurochemical projections, temporal trajectories, geometric relations and held-out prediction. The central result is a dissociation between information accessible in  $C^3$  and the interpretation of its music-oriented affect readouts in a speech domain.

## Experimental design and a hierarchy of evidence

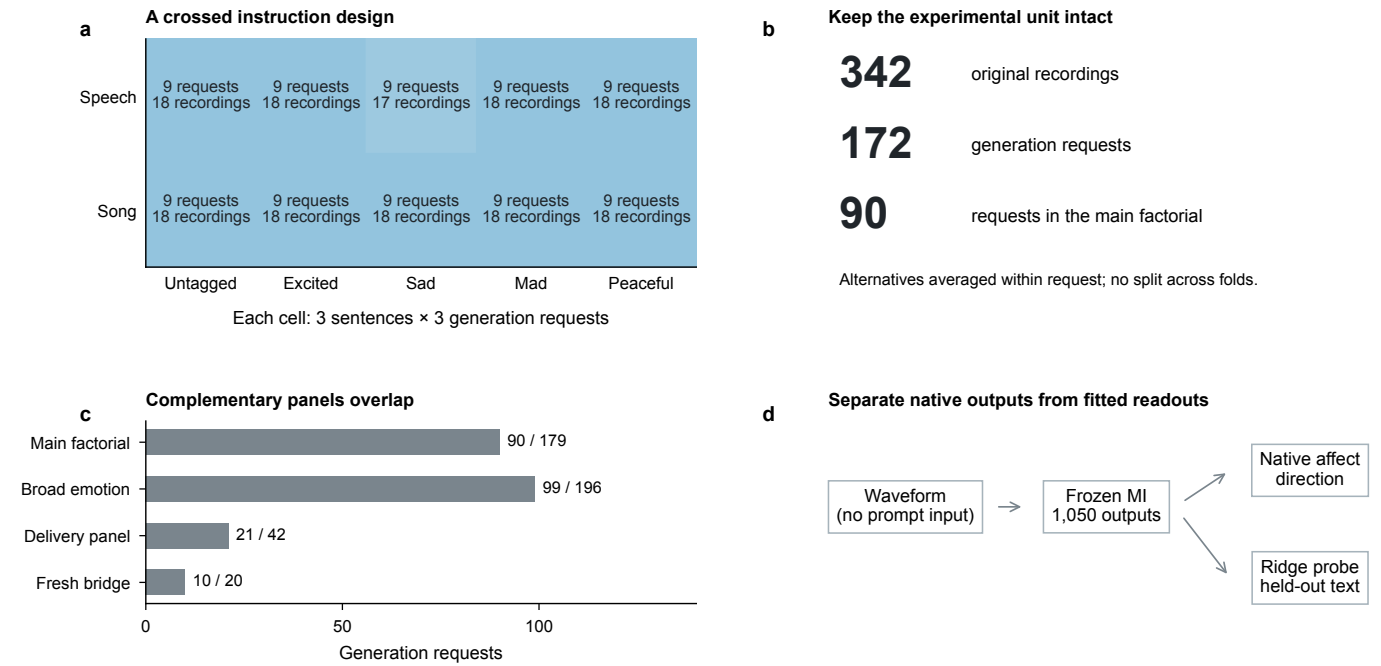


**Fig. 1 | Same intended words, crossed expressive instructions and preserved generation units.** **a**, Main factorial: three texts × two delivery modes × five emotional conditions × three requests, yielding 90 requests and 179 recordings. Counts are pooled across texts; one alternative is unavailable. 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, consistent with the provider’s documented performance-direction interface.<sup>4</sup> 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

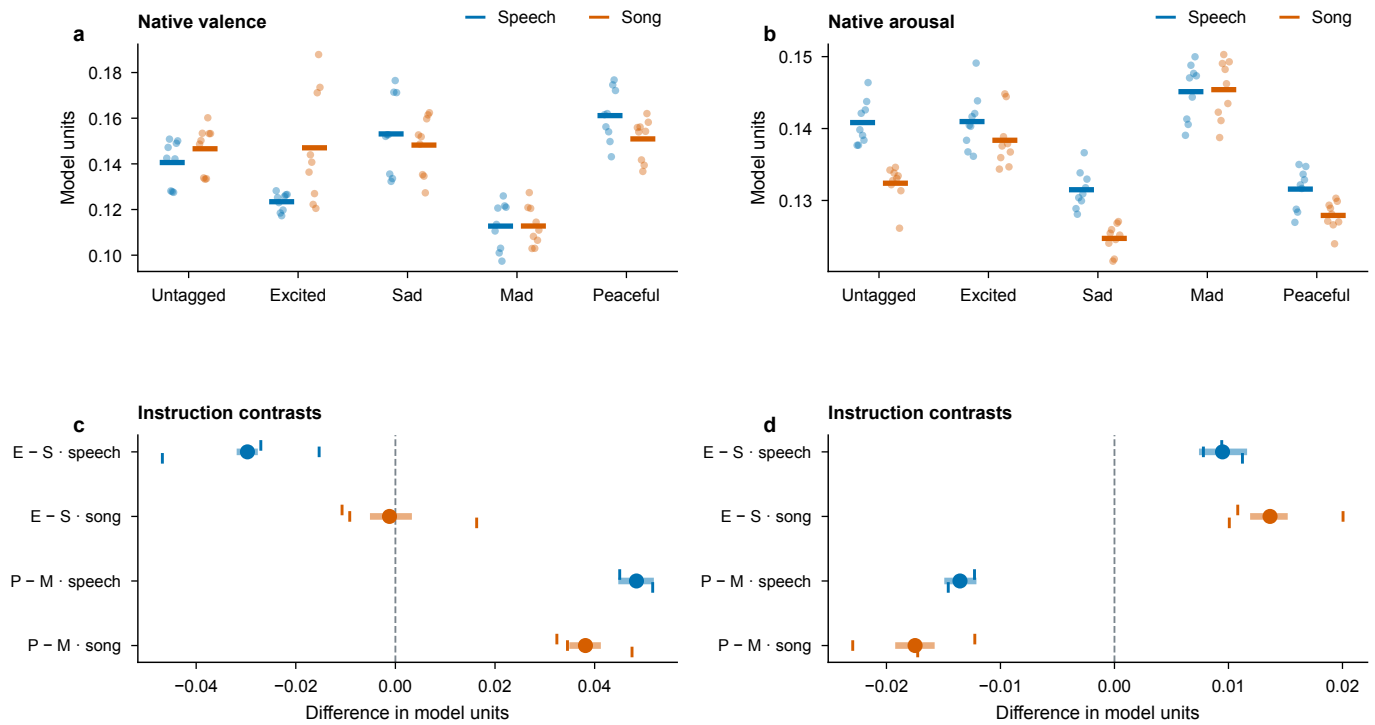


**Fig. 2 | Instruction-dependent native affect is not uniformly directionally calibrated.** **a,b**, Request means for valence (VMM:C1) and arousal (AAC:P2); nine requests per emotion and delivery mode, with three requests from each text. Points show requests; short horizontal marks show pooled means. **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  $q = 0.1293$ .<sup>5,6</sup> 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 C<sup>3</sup> retains information that two affect scores miss

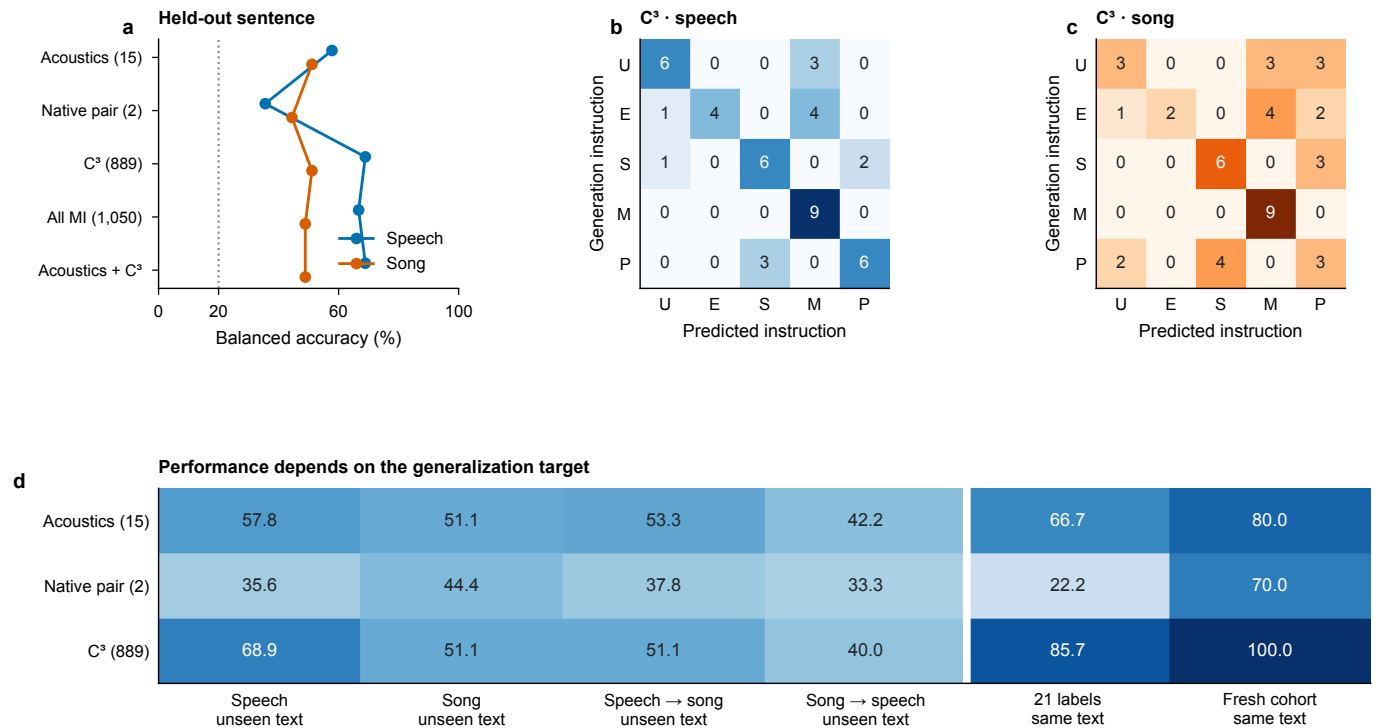


**Fig. 3 | Native-score interpretation and fitted decoding answer different questions.** **a**, Five-way emotion decoding with an identical fixed ridge probe and sentence-held-out folds. Each mode contributes 45 test requests, nine per class; dotted line, nominal 20% chance. “Native affect + probe” fits the same decoder to two outputs and is not a native emotion classifier. **b,c**, C<sup>3</sup> 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 C<sup>3</sup>–acoustic difference is descriptive, not a tested superiority claim.

Five-way speech decoding reached 68.9% balanced accuracy from C<sup>3</sup>, 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 C<sup>3</sup> did not improve the speech result, while concatenating all 1,050 MI outputs yielded 66.7%. For song, C<sup>3</sup> and acoustics both reached 51.1% (23/45); the native pair reached 44.4%, and the combined acoustic–C<sup>3</sup> representation reached 48.9%. More outputs therefore did not monotonically improve generalization (Fig. 3).

Errors were not distributed uniformly across labels. The C<sup>3</sup> 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 C<sup>3</sup> 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; C<sup>3</sup> 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 C<sup>3</sup> 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.<sup>7</sup>

Transfer also depended on the test regime. Training C<sup>3</sup> on speech and testing song on an excluded text gave 51.1%; the reverse direction gave 40.0%. Same-text tasks were easier: C<sup>3</sup> reached 85.7% for 20 emotional instructions plus an untagged condition and 90.5% for seven delivery instructions. The joint acoustic–C<sup>3</sup> 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 C<sup>3</sup>, 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;<sup>8</sup> the archive preserves fold identities and every held-out prediction so that errors, dependence and generalization scope can be inspected directly.

## A paired level intervention dissociates arousal from instruction information

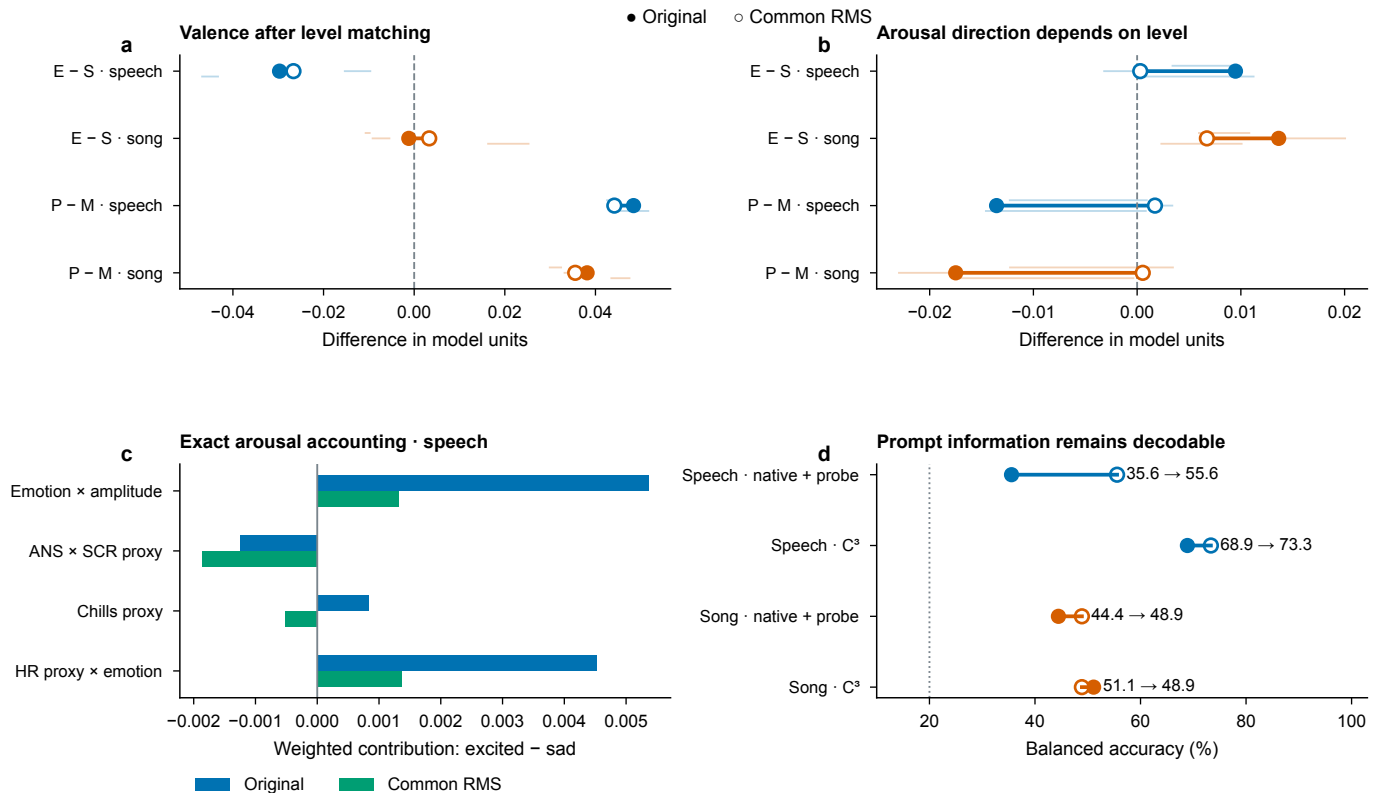


**Fig. 4 | Matching waveform RMS alters native arousal without erasing C<sup>3</sup> instruction decoding.** All 342 recordings were reanalysed after constant attenuation to  $-29.5175$  dBFS; the original and processed waveform identities were verified. **a,b**, Original (filled) and common-RMS (open) native contrasts; large symbols are equal-text means, thin segments connect each text's pair. **c**, Exact weighted contributions to the excited-minus-sad speech-arousal contrast; terms sum to the implemented score difference, with zero reconstruction error. Names denote model variables, not measured physiology. **d**, Original and common-RMS held-out-sentence decoding, using the same tasks and fixed probe. Values are percentages. The intervention adds processed copies, not new independent requests; no acoustic baseline or new permutation inference is imputed for processed audio.

A constant gain intervention provides a direct test of whether observed scores depend on waveform level. Each recording was attenuated to the lowest RMS in the original corpus, preserving duration, sample count, pitch and the relative temporal envelope. Required attenuation ranged from 0 to  $-10.8538$  dB. The complete frozen engine was rerun on all 342 floating-point WAV copies. This operation changes a physically defined signal statistic; it does not equate perceptual loudness, remove differences in dynamics, or emulate a listener adjusting playback volume after MI has analysed the original file.

The excited-minus-sad speech-arousal difference declined from 0.009479 to 0.000302, a 96.8% reduction in this mean contrast. Its three text-specific signs changed from three positive to two positive and one negative. Peaceful-minus-mad speech arousal reversed from  $-0.013550$  to  $+0.001718$ , with all three texts becoming positive. Song showed a smaller excited-minus-sad reduction, from 0.013646 to 0.006727; all three text-specific contrasts remained positive. Its peaceful-minus-mad mean changed from  $-0.017489$  to  $+0.000541$  although two of the three text-specific contrasts remained negative. Aggregate reversal and unanimous reversal are therefore distinguished (Fig. 4).

Valence behaved differently. Excited-minus-sad speech remained negative in every text after matching, averaging

$-0.026638$ . Peaceful-minus-mad valence remained positive in every text and both delivery modes. Equalizing RMS therefore did not repair the excited-sad native-valence mismatch. Nor did it remove the information available to a fitted C<sup>3</sup> decoder: five-way speech accuracy increased descriptively to 73.3%, whereas song accuracy was 48.9%. The corresponding native-pair probes reached 55.6% and 48.9%. These paired decoding changes were retained as sensitivity results, without a new significance test or post hoc choice of the better preprocessing condition.

The implementation makes score attribution exact at the algebraic level. AAC arousal combines four weighted terms, including emotional arousal multiplied by acoustic amplitude, ANS and SCR proxies, a chills proxy, and heart-rate proxy multiplied by emotional arousal. For excited-minus-sad speech, the direct amplitude-product contribution fell from approximately  $+0.00536$  to  $+0.00132$ ; other terms changed as well. VMM valence combines five terms, all of which contributed negatively to the original excited-minus-sad speech contrast. Weighted reconstructions matched the saved scores exactly. This accounting locates the numerical change inside the implementation, but is not a causal neural decomposition: gain also changes upstream computations, and 96.8% is a contrast reduction rather than a mediation fraction.

## The full response repertoire is structured, repeatable and acoustically accessible

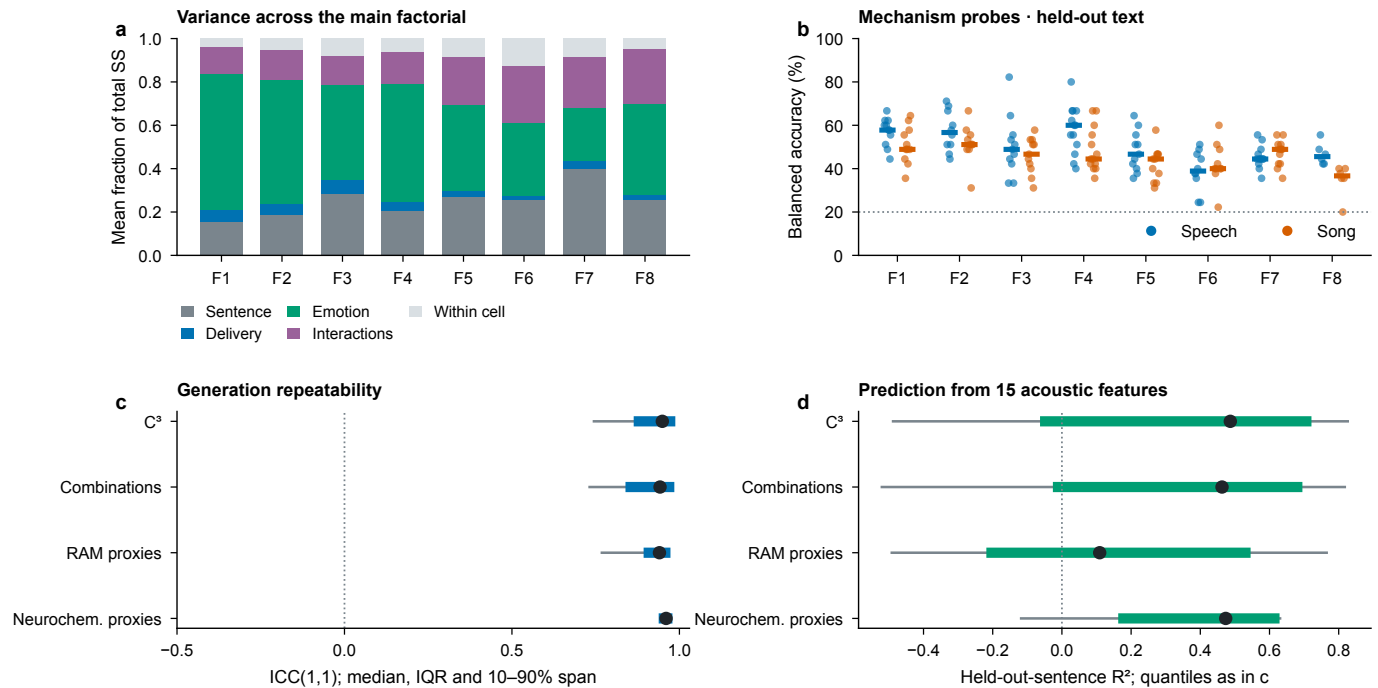


**Fig. 5 | Architecture-wide summaries retain the entire measured repertoire.** **a**, Orthogonal variance decomposition of 90 main request means, summarized as the mean fraction across dimensions in each C<sup>3</sup> function group. Interactions combine all two- and three-factor terms. Dimensions with undefined fractions are not assigned zero. **b**, Five-way speech and song probe accuracy for every mechanism, grouped by function; points are mechanisms and short bars are medians. These 84-mechanism rankings are descriptive, not multiplicity-tested selections. **c**, Request-level ICC(1,1) across 30 fixed conditions. **d**, Held-out-text prediction of MI outputs from 15 acoustic descriptors. Points mark medians, thick lines interquartile ranges, and thin lines 10th–90th percentiles across finite output estimates. Negative values and full tails remain in source data. Output labels do not establish neural validity.

The audit covers all 84 mechanisms, their 889 dimensions, 131 combination scores, 26 regional projections and four neurochemical proxies. Upstream R<sup>3</sup>/T<sup>3</sup> variables may contribute to C<sup>3</sup> computations but are not cognitive endpoints here. The eight declared functions and internal dimension prefixes retain their saved identities (Supplementary S3). Those labels do not themselves establish a psychological process, a temporal memory span or a neural substrate.

A balanced variance decomposition separated text, delivery, emotion, their interactions and within-cell request variation. Native valence allocated 55.3% of total sum of squares to emotion, 24.6% to sentence and 1.9% to within-cell variation; native arousal allocated 71.6%, 4.2% and 4.7%, respectively. These descriptive proportions show that repeatable instruction differences can coexist with an incorrect expected sign. They are not causal effect shares because historical generation cohorts partially track experimental conditions. Interaction terms and group-level summaries expose the structure that a single global emotion score conceals (Fig. 5).

Between-condition repeatability was high for the native outputs: ICC(1,1) was 0.972 for valence and 0.932 for arousal. The unit here is a request-level alternative average, observed three times within each of 30 fixed cells. The ICC therefore describes this controlled stimulus repertoire, not test–retest reliability in humans, and depends on the range of conditions included. The archive additionally preserves all 170 available within-request alternative differences, three-request ICCs, negative estimates

and undefined estimates. A reproducible mismatch is scientifically different from an unstable one, but reliability does not establish the meaning of the measured variable.

The strongest descriptive mechanism-level speech decoders included IACM (82.2%) and MMP (80.0%), exceeding the full C<sup>3</sup> probe in this sample. Selecting a mechanism after inspecting 84 such results is not an independently validated feature-selection procedure, and these rankings should guide prospective tests rather than establish mechanism superiority. Acoustic-to-output ridge prediction, also holding out complete sentences, gave R<sup>2</sup> = 0.703 for native valence and 0.319 for arousal. These values quantify recoverability from a limited acoustic baseline; they do not partition a model into acoustic and cognitive fractions. Regionally named outputs and neurochemical proxies are included in the same audit, but all remain deterministic model projections without brain recordings or biochemical measurements.

Across the entire repertoire, median ICCs were 0.949 for C<sup>3</sup>, 0.942 for combinations, 0.940 for regional projections and 0.960 for neurochemical proxies. Median held-out acoustic-prediction R<sup>2</sup> values were 0.487, 0.463, 0.109 and 0.473, respectively. Three C<sup>3</sup> dimensions had undefined R<sup>2</sup>; all others contributed to the displayed finite-value distributions. These summaries reveal substantial variation in acoustic predictability even among highly repeatable outputs. The weaker median for regional projections is not evidence of greater biological realism: a poor fit from 15 descriptors can also reflect nonlinear transforms, omitted acoustics or extrapolation across texts.

## Time-resolved responses and geometry expose additional structure

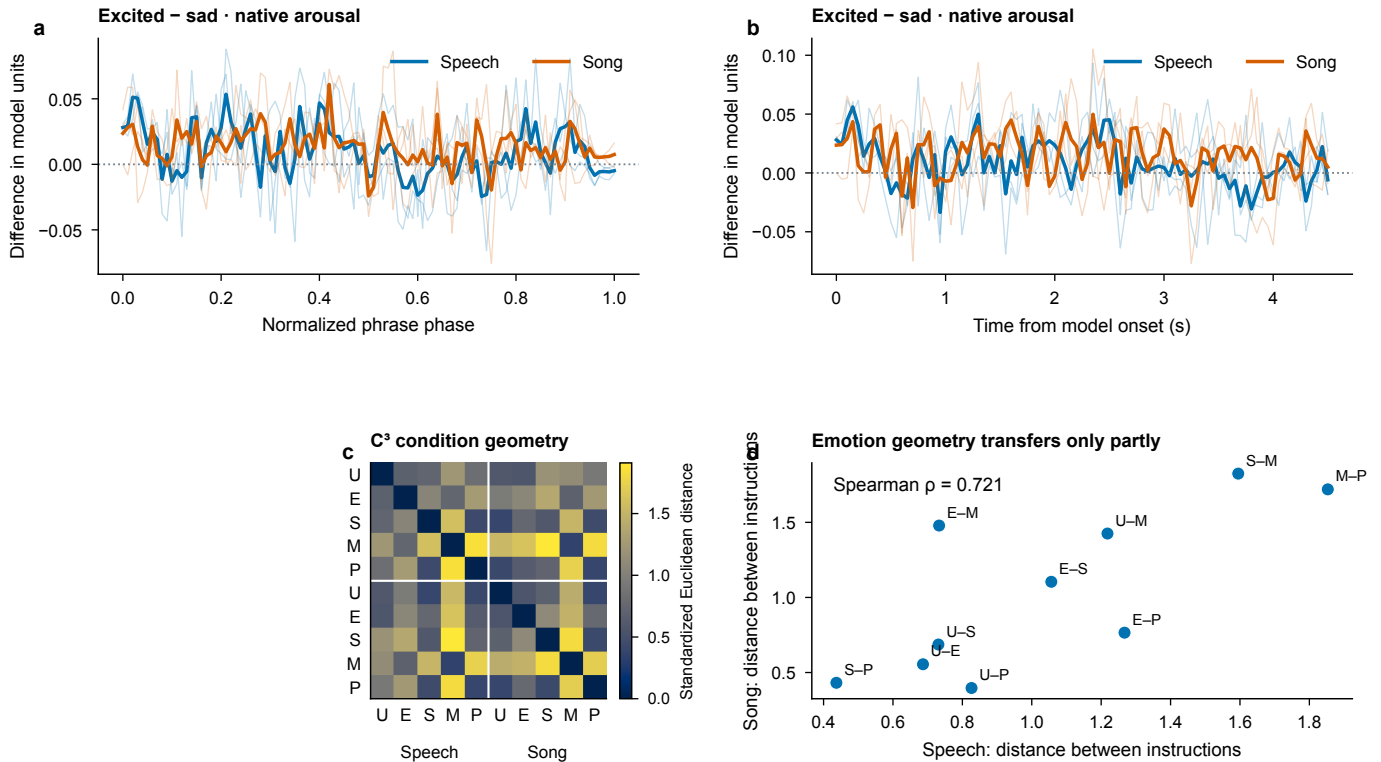


**Fig. 6 | The same instruction contrast has a trajectory and a representational geometry.** **a,b**, Original excited-minus-sad native-arousal trajectories on normalized phrase phase and common absolute-time support (0–4.5 s). Thick curves average texts equally; thin curves show the three text contrasts. Each condition mean averages three requests after within-request alternative averaging. The 101 phase points and 91 time points are interpolation grids, not independent observations or word-aligned latencies. **c**, C<sup>3</sup> distances between ten delivery–emotion centroids, standardized using main-request feature variation and normalized by active dimension count. **d**, The ten within-speech emotion distances against corresponding within-song distances; labels identify emotion pairs. Their rank correlation is descriptive because distance-matrix edges are dependent. Full matrices and all seven trajectory summaries are archived.

Whole-recording averages cannot specify when a difference occurs. We therefore retained 220 main trajectory contrasts on two complementary clocks: 101 equally spaced points from the first to the final model frame, and 91 points over the absolute 0–4.5 s interval supported by every recording. Phase alignment permits comparison of relative phrase progression; absolute alignment retains elapsed time but compares potentially different lexical positions. Neither alignment is a substitute for verified word boundaries. The excited-minus-sad arousal contrast varies through a phrase, and the visibility of that variation depends on which clock is used (Fig. 6).

Seven additional summaries characterize each trajectory: duration-normalized area, range, total variation per second, linear slope per second, peak time, late-minus-early level, and one-second lag correlation. Each of the 556 contrasts is available for these summaries across all outputs. Peak time and lag correlation are explicitly undefined for relevant constant trajectories; undefined values in an unrelated request are prevented from contaminating a contrast through zero-weight arithmetic. A large trajectory peak identifies where an implemented response differs, not the latency of a neuronal or perceptual process. Offline future or centred context in MI further prevents a causal online interpretation of these time courses.

Representational distances reveal a complementary form of consistency. The ten emotion-pair distances within speech and their counterparts within song had a C<sup>3</sup> Spearman correlation of 0.721. This suggests partial preservation of relative instruction geometry despite the difference in classifier performance and the asymmetry of cross-delivery transfer. A similar arrangement of condition distances does not require identical response vectors or a perfectly transferable decision boundary. The archive retains both Euclidean and correlation distances, emotion-contrast vector cosines and relative vector norms, allowing these possibilities to be separated. Representational similarity is used here as a descriptive organizational tool, not a test based on independent distance edges.<sup>9</sup>

The complete geometric inventory contains 1,890 matrices: 105 feature families, nine panels and two distance definitions. In parallel, acoustic associations cover 15 descriptors against all outputs in six scopes, both raw and sentence-centred, with Pearson and rank correlations. These complementary views matter because a strong pooled association may be driven by text or delivery, whereas an instruction contrast or held-out-text regression asks a more specific question. None of these analyses constructs an unobserved continuous speech-to-music path. The existing anchors establish endpoints and directions of model change; any interpolated acoustic morph or perceptual transition remains a separate experiment.

## Discussion

### A music-oriented model confronted with speech

MI separates instruction information, native affect direction and level robustness. C<sup>3</sup> supports above-null held-text speech decoding, yet excited-minus-sad valence is negative in every text, including at common RMS. A model designed for musical listening is being transferred to speech; it is not a speech-trained emotion estimator. The reversal cannot be assigned a cause from its sign alone.

### Shared cues do not require identical mappings

Direct experiments support conditional rather than universal cue mappings. Ilie and Thompson manipulated intensity, rate and pitch in music and speech: higher pitch increased speech pleasantness, whereas lower pitch favoured music pleasantness, with a tempo interaction.<sup>10</sup> Belyk and Brown found that vocal valence–pitch and valence–amplitude relations depended on emotion family.<sup>11</sup> Conversely, shared psychoacoustic features predicted affect in both domains,<sup>12</sup> and cross-domain recurrent models transferred useful information; speech-trained models predicted musical valence well, while music-to-speech transfer remained weaker than speech-domain training.<sup>13</sup> Dataset and target-distribution differences can contribute to such asymmetries. This literature motivates testing MI's transfer, not excusing any mismatch as intrinsically musical.

Sun and colleagues compared speech- and music-pretrained representations on matched-text spoken and sung RAVDESS recordings; transfer varied by layer and emotion.<sup>14</sup> Their random recording splits and target-domain adaptation preclude direct comparison with our held-text frozen-model assay. Zhang and colleagues found activation more transferable than valence in matched speech and song.<sup>15</sup> Spectrotemporal dissociations in speech and melody perception<sup>2</sup> motivate stimulus controls but do not explain valence reversal.

### Three explanations that the present data cannot separate

First, MI's native affect mapping may be miscalibrated. Second, cue weighting appropriate to musical listening may transfer unevenly to emotional speech. Third, the comparison may join different targets: intended expression, perceived expression, sensory pleasantness and felt emotion. These are not mutually exclusive. VMM mixes sensory-pleasantness and reward-labelled terms with other components; their names do not establish which target the composite measures. Human studies of sad music show that perceived and felt emotion can diverge.<sup>16</sup> Thus, a higher score for sad speech is neither automatically a valid musical judgment nor, without a specified target, sufficient evidence of a general emotion-prediction failure.

A specific transfer hypothesis concerns phonetic structure. Frication, bursts and vowel–consonant transitions alter acoustic organization; emotion-dependent articulation changes segmental spectra even when words are fixed.<sup>17</sup> A music-oriented readout may interpret these changes as affect without conditioning on phonetic function. VMM defines consonance as  $1 - \text{roughness}$  and feeds it into valence directly and indirectly. The archived direct inverse-roughness contribution to excited-minus-sad valence is

negative in all three speech texts but positive in all three song texts. These whole-recording contributions do not locate consonants or establish their causal role. Sung words also retain phonetic structure.

The observed pattern narrows but does not settle these alternatives. Excited-minus-sad valence is negative in all three speech texts, but also in two song texts; the third song text is positive. Lower five-way decoding in song prevents a general claim of music superiority. Delivery changes pitch, duration, phonation and energy together. A difference between two model outputs does not establish that humans weight those cues differently. Direct delivery-by-emotion contrasts are archived, but no scalar finding survives the global BY correction. Domain dependence therefore remains a literature-motivated hypothesis, not an established perceptual explanation.

### Limits of the inference

Prompts remain production conditions, not verified perceived-emotion labels; no human rating or model agreement determined inclusion. The fixed voice, three texts and staged generation limit generalization. Cohort restrictions cannot remove confounding built into production history, and the fresh bridge uses one familiar text. The small acoustic baseline is not a comprehensive specialist benchmark. Even successful decoding can exploit acoustic transformations without validating named mechanisms,<sup>18</sup> and a post hoc probe cannot retroactively calibrate the frozen native output.

Meaning is not removed by withholding text from MI: waveforms retain linguistic cues. Repetition can also turn an unchanged phrase into perceived song,<sup>19</sup> whereas fresh-state analysis of identical audio cannot reproduce an exposure-dependent change. Neither musical abstraction nor induction from natural regularities is established here. Temporal and anatomical labels cannot bridge these gaps without independent tests.

### A falsifiable test of domain-dependent interpretation

Cross more voices and held-out texts with matched speech and song; randomize generation blocks and control waveform level. Specify separate targets for expressed emotion and sensory pleasantness before fitting shared or domain-specific mappings. Evaluate both transfer directions and direct domain-by-emotion interactions under acoustic controls. A shared calibration that repairs both domains would support a general mapping problem; reproducible gains from domain-specific mappings would support conditional cue weighting. Neither result alone establishes a human mechanism. Independent speech and song corpora with target-specific listener measurements can test which interpretation generalizes; this is a future experiment, not an inclusion gate for these recordings.

Phoneme-aligned trajectories and matched worded versus wordless renditions could test whether phonetic segments drive the valence discrepancy, with controls for level, pitch and duration. Component interventions can locate computational causes, not listener mediation. Continuous morphing requires controlled endpoints and spacing. The archive preserves failed predictions and uninformative dimensions for these tests.

## Methods

### Study status, units and corpus assembly

This is an exploratory computational study of a fixed generated-audio corpus, not a participant experiment. The comprehensive analysis plan was frozen before its execution but after earlier native-affect results had been inspected; it is not represented as an independent preregistration. The 172 generation requests produced 342 available alternatives. The main  $3 \times 2 \times 5 \times 3$  design contains 90 requests and 179 files. Broad emotion, delivery and bridge panels overlap this factorial. All analyses retain their declared memberships. No human review score enters a feature, outcome, target or exclusion rule. The nine historical review records are preserved only as provenance. Supplementary Information gives all texts, prompt categories, panel memberships and analysis scopes.

### Generation reference and lexical information

The reference label is the exact generation instruction. Untagged speech uses the bare sentence; tagged speech prefixes an emotional direction; song combines the emotional direction, when present, with a request for a simple unaccompanied melody. The broader panel includes 20 tagged emotions and an untagged condition on S01; six emotional conditions span all three sentences. The delivery panel has seven instructions on S01. The bridge comprises ten newly requested S01 conditions. The generation UI reported Eleven v4, one Anna voice, stability 0.5, similarity 0.75, automatic language and MP3 at 44.1 kHz/128 kbps. Exact provider model ID, voice ID and seed are unavailable. Archived automatic transcripts and word-time estimates are audit data; six lexical disagreements are retained, and no verified word-aligned response latency is estimated.

### Frozen model and feature inventory

Each recording is decoded through the pinned soundfile float32 adapter at native 44.1 kHz. Fresh model state is used for each file. The front end uses 128 mel bands, an FFT length of 2,048 and hop length of 256, with log compression and the saved normalization implementation. The offline engine may use centred or future context. Native trajectories and unstretched combination scores are retained rather than rescaled per output for presentation. The pinned adapter and 855 engine-source files are identity-verified; complete fingerprints are retained in the provenance archive. A prior, differing audio-loading adapter was rejected rather than pooled. The output registry contains 889 declared mechanism dimensions, 131 combinations, 26 RAM projections and four neurochemical proxies.  $R^3/T^3$  are not analysis endpoints; implementation dependencies remain documented.

### Input normalization and the interpretation of level control

The spectrogram is a power mel representation followed by  $\log(1 + \text{value})$ , divided by its maximum across frequency and time with a floor of  $10^{-8}$ . This normalization does not make the entire model gain-invariant. The acoustic extractor also receives the original waveform, and subsequent operations can depend on amplitude-related variables. Moreover, a maximum-normalized logarithmic representation need not be identical after constant waveform scaling. The gain intervention therefore reruns the complete path rather than multiplying a saved scalar or assuming which

terms must remain unchanged. Component arrays record the resulting upstream and downstream changes together.

### Aggregation and native-score definitions

Frame mean, median, standard deviation and the 10th and 90th percentiles are computed within each recording, then averaged over available alternatives within each request. Means of requests define condition estimates; sentences contribute equally to averaged contrasts. VMM:C1 is the clipped sum of  $0.25 \times V0$  valence,  $0.20 \times H3$  sensory pleasantness,  $0.20 \times$  reward evaluation,  $0.20 \times (\text{happy} - \text{sad})$ , and  $0.15 \times (1 - \text{roughness})$ . AAC:P2 is the clipped sum of four equally weighted terms: emotional arousal  $\times$  amplitude, ANS  $\times$  SCR proxy, chills proxy, and heart-rate proxy  $\times$  emotional arousal. All components are reconstructed at their implemented frame resolution before summary; clipping adjustments are retained. A zero reconstruction residual establishes arithmetic closure, not causal mediation.

### Acoustic baseline

Fifteen interpretable descriptors are extracted from a mono, 16-kHz analysis copy, separately from MI input preprocessing. They comprise duration, waveform RMS and sample peak in dBFS, active fraction, voiced fraction, median pitch probability, median F0, F0 interquartile and 90th–10th percentile ranges in semitones, mean spectral centroid, mean spectral flatness, envelope-modulation peak (1–12 Hz), its 2–5-Hz power fraction, envelope-peak interval coefficient of variation, and envelope-peak count. The acoustic STFT uses a 1,024-sample window and 160-sample hop. Active frames exceed the larger of –45 dBFS and 5% of maximal frame RMS. pYIN estimates F0 over 65–1,000 Hz;<sup>20</sup> pitch summaries exclude inactive/invalid frames. Descriptor failures remain missing values, not recording exclusions. Full definitions and code are preserved.

### Paired common-RMS intervention

For original waveform  $x$ , RMS is  $\sqrt{\text{mean}(x^2)}$ ; gain is  $g = r^*/\text{RMS}(x)$ , where  $r^*$  is the smallest original RMS. Thus  $0 < g \leq 1$  and  $x' = gx$ . The target is –29.517516 dBFS and the largest attenuation is –10.853763 dB. Copies are saved as floating-point WAV without quantization clipping. Sample count, rate and file identities are checked, and all 342 are rerun using the identical frozen adapter. These are paired transformations of existing samples, not additional generation units. Three hundred canonical gain-condition probes are reported descriptively. There is no processed-audio acoustic baseline, perceived-loudness normalization, automatic gain-selection step or new permutation family for these sensitivity results.

### Preserving production and analysis identity

The source manifest retains request-level provenance separately from recording-level file identity. Two requests have one available alternative; 170 have two. Within-request summaries use available alternatives with equal weight, and requests remain equally weighted thereafter. Original MP3 files, decoded-input checks, processed WAV files and model-output receipts are all retained. The original research archive and its SHA-256 inventory remain intact; the manuscript package links to that archive and adds source tables without overwriting the experimental result files. Missing provider identifiers remain explicit null values rather than being reconstructed from a display name.

## Methods | Statistical analysis and reproducibility

### Contrasts, uncertainty and multiplicity

The 556 contrasts cover main sentence-specific pairs (135), main equal-text averages (45), delivery-by-emotion differences of differences by sentence (30) and averaged (10), sentence differences (30), broad-emotion pairs (240), delivery pairs (21), and bridge pairs (45). Eligible mean tests exchange labels only among comparable within-cohort requests. Absolute two-sided statistics enumerate all 20 partitions for one 3-versus-3 comparison or 8,000 joint partitions for three texts; duplicated tests across panels are counted once. There are 303 unique tests per output. All 318,150 P values enter one BH and one BY family.<sup>5,6</sup> The latter is emphasized for arbitrary dependence. Pointwise 95% percentile intervals use 1,999 request resamples within each cell for 55 equal-text main and interaction contrasts. They do not generalize over texts or correct selection. Standardized effects divide by pooled within-cell request variation; near-constant denominators remain undefined.

### Predictive tasks and leakage control

Fifteen tasks are evaluated using 106 feature families: acoustics, two and six native-affect variables, all  $C^3$  dimensions, all MI outputs, combinations, RAM, neurochemical proxies, eight functions, five dimension prefixes, all 84 mechanisms and acoustics plus  $C^3$ . The core analysis yields 1,590 fits at fixed ridge  $\alpha = 1$ ; 90 further fits retain  $\alpha = 0.1$  and 10 for selected canonical families without choosing a winner. For unseen-text tasks, every request containing the held-out sentence is excluded from training. Cross-delivery tasks train on the opposite delivery from other sentences. Same-text tasks hold out generation blocks; the bridge trains on historical main requests and tests fresh S01 requests. Splits preserve every request's alternatives. Balanced accuracy is the mean class recall; macro-F1, confusion matrices, class recalls, fold identities and continuous decision scores are archived.

### Training, nulls and sensitivity

Each training fold supplies feature medians for missing-value imputation, means and standard deviations. Dimensions below a numerical floor of 16 float32 epsilons  $\times \max(1, \text{maximum absolute training value})$  are excluded. Retained standardized features are divided by the square root of their count. An intercept-bearing ridge model predicts one-hot labels; a dual-form implementation is independently checked against scikit-learn.<sup>21</sup> Canonical tests use 999 restricted label permutations, repeating the label-dependent fit under the same folds, with  $P = (1 + \text{number of null scores at least observed}) / (1 + 999)$ . Permutation inference follows the requirement to preserve exchangeability structure.<sup>22,23</sup> Emotion shuffles retain sentence, delivery and cohort strata; other tasks use their saved valid strata. A target fixed within strata, such as binary delivery, has no reported permutation P. The 308 valid canonical tests form one global BY family. Individual mechanism probes have no confirmatory P value. Seed: 202610145.

### 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 pre-processing 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  $t/t_{\text{last}}$ . 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  $\sqrt{\text{active dimensions}}$ . 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.

### Validation and research record

Eighteen independent checks cover partition enumeration, scalar reconstruction, ridge agreement, split disjointness, held-out-text exclusion, variance closure, undefined-value handling and original/processed waveform identity. The largest independently checked ridge discrepancy was  $7.2 \times 10^{-16}$ ; native component closure error was zero. NumPy/SciPy and scikit-learn versions are recorded with source hashes.<sup>24</sup> The delivery includes immutable experiment files, all unrounded arrays and predictions, exact prompts, audio and gain copies, the full output registry, figure-specific source data, figure code, editable manuscript and supplementary tables. Findings are not withheld when weak, negative or undefined. The supplementary inventory identifies each result's file, unit and scope; reproduction requires the recorded environment and frozen model source rather than access to a current provider model.

### 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.

## References

- 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](https://doi.org/10.1037/0033-2909.129.5.770).
- 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](https://doi.org/10.1126/science.aaz3468).
- 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](https://doi.org/10.1371/journal.pone.0196391).
- ElevenLabs. ElevenLabs Audio Tags list for Eleven v4, with 50+ examples. *Provider documentation*; accessed 14 October 2026. [ElevenLabs documentation](https://elevenlabs.com/docs/audio-tags).
- 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](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x).
- 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](https://doi.org/10.1214/aos/1013699998).
- Hewitt, J. & Liang, P. Designing and interpreting probes with control tasks. *Proc. EMNLP-IJCNLP*, 2733–2743 (2019). [10.18653/v1/D19-1275](https://doi.org/10.18653/v1/D19-1275).
- 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](https://doi.org/10.1016/j.neuroimage.2017.06.061).
- 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](https://doi.org/10.3389/neuro.06.004.2008).
- 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](https://doi.org/10.1525/mp.2006.23.4.319).
- 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](https://doi.org/10.1016/j.jvoice.2013.12.007).
- Coutinho, E. & Dikken, N. Psychoacoustic cues to emotion in speech prosody and music. *Cogn. Emot.* **27**, 658–684 (2013). [10.1080/02699931.2012.732559](https://doi.org/10.1080/02699931.2012.732559).
- 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](https://doi.org/10.1371/journal.pone.0179289).
- 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](https://doi.org/10.1109/ICASSP49660.2025.10890304).
- 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](https://doi.org/10.1609/aaai.v29i1.9334).
- Kawakami, A., Furukawa, K., Katahira, K. & Okanoya, K. Sad music induces pleasant emotion. *Front. Psychol.* **4**, 311 (2013). [10.3389/fpsyg.2013.00311](https://doi.org/10.3389/fpsyg.2013.00311).
- Kienast, M. & Sendlmeier, W. F. Acoustical analysis of spectral and temporal changes in emotional speech. *Proc. ISCA Workshop Speech and Emotion*, 92–97 (2000). [Journal full text](#).
- 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](https://doi.org/10.1038/s41593-023-01442-0).
- Deutsch, D., Henthorn, T. & Lapidis, R. Illusory transformation from speech to song. *J. Acoust. Soc. Am.* **129**, 2245–2252 (2011). [10.1121/1.3562174](https://doi.org/10.1121/1.3562174).
- Mauch, M. & Dixon, S. pYIN: a fundamental frequency estimator using probabilistic threshold distributions. *Proc. IEEE ICASSP*, 659–663 (2014). [10.1109/ICASSP.2014.6853678](https://doi.org/10.1109/ICASSP.2014.6853678).
- Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011). [Journal full text](#).
- Ojala, M. & Garriga, G. C. Permutation tests for studying classifier performance. *J. Mach. Learn. Res.* **11**, 1833–1863 (2010). [Journal full text](#).
- 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](https://doi.org/10.1016/j.neuroimage.2015.05.092).
- 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](https://doi.org/10.1038/s41592-019-0686-2).

## Data and code availability

The delivery package links to the complete original experimental archive, including both audio conditions, model outputs, analysis code, prompts, provenance and unrounded arrays; manuscript text, figure sources and result tables are included separately. `experiment/README.txt` and `experiment/analysis/comprehensive/DATA_GUIDE.txt` describe the original analysis. `data/` contains manuscript derivations and complete readout/mechanism tables. This is an accompanying local research archive; no public repository accession or permanent DOI has been assigned.

## 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 data, brain measurements or neurotransmitter measurements are reported. Historical informal review records are retained for provenance and excluded from analyses. The computational findings concern a single archived voice, three intended texts and a frozen model implementation. This document is a research manuscript, not a claim of peer review or journal acceptance.