

Supplementary Information

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

Amaç Erdem · 14 October 2026

This companion document preserves the design, implementation detail and complete mechanism-level result coverage behind the ten-page article. It distinguishes the production instruction from perceived emotion, the generation request from its audio alternatives, and a native MI scalar from a supervised decoder fitted to that scalar. The accompanying numerical archive is part of the research record: printed rounding does not replace the original values.

Inventory	Completed scope
Original corpus	342 recordings; 172 generation requests; three intended texts; one displayed voice
Main factorial	90 requests; 30 cells; speech/song × five emotions × three texts
Model outputs	84 mechanisms; 889 dimensions; 131 combinations; 26 regional and four neurochemical proxies
Scalar coverage	556 contrasts × five summaries × 1,050 outputs = 2,919,000 effects
Predictive coverage	1,680 original probes; 308 canonical permutation tests; 300 gain-condition probes
Temporal / geometric coverage	220 full main trajectory contrasts; seven summaries for all contrasts; 1,890 distance matrices
Paired level control	342 gain copies and frozen-engine reruns; exact weighted-score reconstruction

Reading map

S1 describes the corpus and units. S2 prints every distinct S01 prompt prefix, with all request-level strings retained electronically. S3 defines the output registry. S4–S5 report every canonical task and regularization sensitivity. S6 lists all 84 mechanisms with full names. S7–S8 retain all 45 main native contrasts and the level/component analysis. S9–S11 cover gain-wide patterns, acoustic relations, time and geometry. S12 records numerical validation and the archive map for Study 1. S13 preserves the relocated original figures; S14 retains its complete methods. S15–S20 report the executed spectral intervention, all controls and complete mechanism sensitivity rankings.

Reporting conventions

Percentages are balanced accuracies unless stated otherwise. An em dash denotes an unavailable or undefined estimate, not zero. Model units have no externally established perceptual scale. Regional and neurochemical names are computational labels. No human quality-control gate or human emotion-labeling target is used. All main and supplemental analyses remain exploratory with respect to this assembled corpus.

S1 | Corpus, experimental units and intended texts

S01

The woman carried a small wooden basket through the garden and placed it on the table beside the open window.

S02

The driver checked the number on the ticket and put the folded paper into the pocket of his coat.

S03

A small lamp stood beside the books while the clock on the wall marked the end of another afternoon.

Production cohorts (disjoint inventory)

Production cohort	Requests	Recordings	Texts
elevenlabs	21	42	S01
emotion	99	196	S01, S02, S03
factorial	42	84	S01, S02, S03
factorial:bridge	10	20	S01

The main factorial and scientific panels are overlapping selections from these disjoint production cohorts. There are 170 two-alternative requests and two one-alternative requests. One unavailable alternative occurs in the main factorial; the other occurs outside it. Request averaging prevents the doubled alternatives from being counted as independent replication. The complete request-to-recording mapping is in `data/request-registry.tsv` and the original `catalog.json`.

Production settings and information boundaries

Displayed model: Eleven v4. Displayed voice: Anna—Young Female Conversational Voice. Stability: 0.5. Similarity: 0.75. Language: Auto. Output: 44.1-kHz, 128-kbps MP3. Exact provider model ID, voice ID and seed were not recoverable. Six of 342 automatic transcripts disagree with their intended text; all are retained. Automatic word boundaries are available as audit records, but no verified word-aligned analysis is claimed.

The fresh bridge uses S01 with all ten main delivery–emotion conditions, one request per cell and two alternatives per request. Its production order was randomized within the fresh block. It therefore checks transfer across a production block of known text, not across voice, language or previously unseen text. The staged historical corpus does not supply the same complete block randomization.

S2 | Exact generation instructions

Every row below is an exact prefix found in a distinct S01 generation string; append the S01 sentence printed in S1. The bare-sentence condition has no prefix. Full request IDs, exact texts for all three sentences, generation cohorts and alternatives are preserved in the tab-separated registry. Emotion words are production directions, not post hoc listener labels.

Instruction label	Exact prefix before S01
Natural speech	(no prefix)
Gentle cadence	[speaking with a gently regular cadence]
Regular rhythm	[speaking with an evenly spaced, strongly rhythmic cadence]
Melodic speech	[speaking with a gently melodic lilt]
Unaccompanied singing	[singing a simple unaccompanied melody]
Slightly breathy	[slightly breathy voice]
Very breathy, voiced	[very breathy voice, still voiced, not whispering]
mad	[mad]
proud	[proud]
playful	[playful]
peaceful	[peaceful]
content	[content]
bitter	[bitter]
stressed	[stressed]
confused	[confused]
sad	[sad]

The broad emotion panel contains aggressive, amazed, anxious, bitter, bored, confused, content, curious, despair, excited, mad, peaceful, playful, proud, repelled, sad, scared, stressed, thoughtful and tired, plus untagged speech. The main five-way comparison uses untagged, excited, sad, mad and peaceful. “Unspecified” in machine-readable files denotes the untagged condition. These memberships are fixed; recordings are not selected for apparent prompt success.

S2 | Exact generation instructions (continued)

Every row below is an exact prefix found in a distinct S01 generation string; append the S01 sentence printed in S1. The bare-sentence condition has no prefix. Full request IDs, exact texts for all three sentences, generation cohorts and alternatives are preserved in the tab-separated registry. Emotion words are production directions, not post hoc listener labels.

Instruction label	Exact prefix before S01
scared	[scared]
curious	[curious]
repelled	[repelled]
amazed	[amazed]
anxious	[anxious]
despair	[despair]
excited	[excited]
bored	[bored]
tired	[tired]
thoughtful	[thoughtful]
aggressive	[aggressive]
excited	[excited, singing a simple unaccompanied melody]
sad	[sad, singing a simple unaccompanied melody]
mad	[mad, singing a simple unaccompanied melody]
peaceful	[peaceful, singing a simple unaccompanied melody]

The broad emotion panel contains aggressive, amazed, anxious, bitter, bored, confused, content, curious, despair, excited, mad, peaceful, playful, proud, repelled, sad, scared, stressed, thoughtful and tired, plus untagged speech. The main five-way comparison uses untagged, excited, sad, mad and peaceful. “Unspecified” in machine-readable files denotes the untagged condition. These memberships are fixed; recordings are not selected for apparent prompt success.

S3 | The complete MI response registry

Every analysis indexes outputs in the order declared by `experiment/analysis/comprehensive/catalog.json`. The first 889 entries are the declared mechanism dimensions; combinations and proxy projections extend the pool to 1,050. The full row-by-row registry, including undefined estimates, is `data/all-readouts.csv`. The 84 complete mechanism identities appear in S6; the original architecture metadata retain each dimension name, layer, source, dependencies and horizon declarations.

Output family	Declared dimensions	Interpretation
C ³ mechanisms	889	84 implemented mechanisms, including their internal dimension groups
Combination scores	131	Unstretched combinations of model responses
RAM projections	26	Regionally named computational mappings
Neurochemical proxies	4	DA, NE, OPI and 5HT model variables

Function	F1	F2	F3	F4	F5	F6	F7	F8
Dimensions	127	110	113	138	132	70	132	67

Regional projection identifiers

A1_HG, STG, STS, IFG, dlPFC, vmPFC, OFC, ACC, SMA, PMC, AG, TP, VTA, NAcc, caudate, amygdala, hippocampus, putamen, MGB, hypothalamus, insula, IC, AN, CN, SOC, PAG.

Dimension groups and interpretation

E, M, P, F and C identify declared internal groups in the saved architecture. Their exact meanings and availability vary by mechanism. They are not automatically interchangeable with human extraction, memory, present experience, prediction or conscious cognition. A label cannot establish a temporal integration window or neural substrate. MI is run offline with fresh state for each recording, and some mechanisms use centred or future context. No output in this study measures brain activation or transmitter concentration.

Native affect and the fitted native-feature families

The two-feature family is `VMM:C1:valence_state` plus `AAC:P2:perceptual_arousal`. It is evaluated by the same fitted ridge decoder used for every feature family. The six-feature family is explicitly indexed in `catalog.json`. Neither fitted family should be called an existing native categorical emotion-recognition system. Native-score contrasts preserve raw model units and the original mapping; probe training is an additional analysis.

S4 | Prediction tasks and generalization targets

Task	Classes	Test N	Chance %	Generalization
Five emotions · both	5	90	20.0	unseen sentence
Five emotions · speech	5	45	20.0	unseen sentence
Five emotions · song	5	45	20.0	unseen sentence
Three emotions · both	3	54	33.3	unseen sentence
Three emotions · speech	3	27	33.3	unseen sentence
Three emotions · song	3	27	33.3	unseen sentence
Four tagged emotions	4	72	25.0	unseen sentence
Speech / song	2	90	50.0	unseen sentence
Ten joint conditions	10	90	10.0	unseen sentence
Six speech emotions	6	54	16.7	unseen sentence
Speech → song	5	45	20.0	unseen sentence
Song → speech	5	45	20.0	unseen sentence
21 labels · same text	21	63	4.8	request block, same sentence
Seven delivery instructions	7	21	14.3	request block, same sentence
Fresh cohort · same text	5	10	20.0	fresh cohort, same S01 text

The test counts are requests, not recordings. Each held-out sentence occurs in exactly one test fold for the relevant task; both alternatives remain together and are averaged before fitting. Cross-delivery tasks train only on the opposite delivery and on other sentences. Same-text tasks hold out generation blocks but preserve lexical identity across training and test. Class-balanced accuracy is not directly comparable across different class counts, production panels or generalization targets.

Labels, scores and nulls are retained

For every task, the prediction directory contains class order, request indices, fold membership, continuous decision scores, predicted and true labels, and the null distribution where defined. The canonical task tables use ridge $\alpha = 1$. The unrestricted collection of 84 mechanism probes is descriptive. Canonical inferential families contain 22 feature sets and 14 valid tasks, giving 308 tests; binary delivery has no valid within-stratum label shuffle and therefore no permutation P value.

S4 | Canonical decoding results

Task	Acoust.	Native 2	Native 6	C ³	All MI	A+C ³	RAM	Neuro.
Five emotions · both	53.3	41.1	52.2	55.6	57.8	55.6	52.2	41.1
Five emotions · speech	57.8	35.6	46.7	68.9	66.7	68.9	40.0	40.0
Five emotions · song	51.1	44.4	51.1	51.1	48.9	48.9	40.0	28.9
Three emotions · both	70.4	61.1	63.0	68.5	68.5	70.4	57.4	38.9
Three emotions · speech	77.8	44.4	70.4	81.5	77.8	81.5	66.7	40.7
Three emotions · song	70.4	66.7	63.0	74.1	70.4	74.1	51.9	51.9
Four tagged emotions	63.9	51.4	61.1	68.1	68.1	69.4	61.1	48.6
Speech / song	71.1	63.3	66.7	60.0	60.0	60.0	63.3	62.2
Ten joint conditions	34.4	23.3	32.2	47.8	45.6	51.1	26.7	15.6
Six speech emotions	57.4	48.1	48.1	61.1	63.0	63.0	42.6	38.9
Speech → song	53.3	37.8	48.9	51.1	51.1	51.1	48.9	35.6
Song → speech	42.2	33.3	37.8	40.0	40.0	37.8	42.2	28.9
21 labels · same text	66.7	22.2	41.3	85.7	88.9	90.5	61.9	31.7
Seven delivery instructions	85.7	47.6	52.4	90.5	90.5	90.5	66.7	47.6
Fresh cohort · same text	80.0	70.0	80.0	100.0	100.0	100.0	80.0	50.0

All values are balanced accuracy (%). “A+C³” concatenates acoustics and C³. All MI includes the 889 mechanism dimensions, 131 combinations, 26 regional projections and four neurochemical proxies. The same training-only imputation, centring, scaling, active-dimension criterion and fixed ridge regularization apply to each representation. The figure and electronic tables retain the corresponding combination, function and dimension-group results.

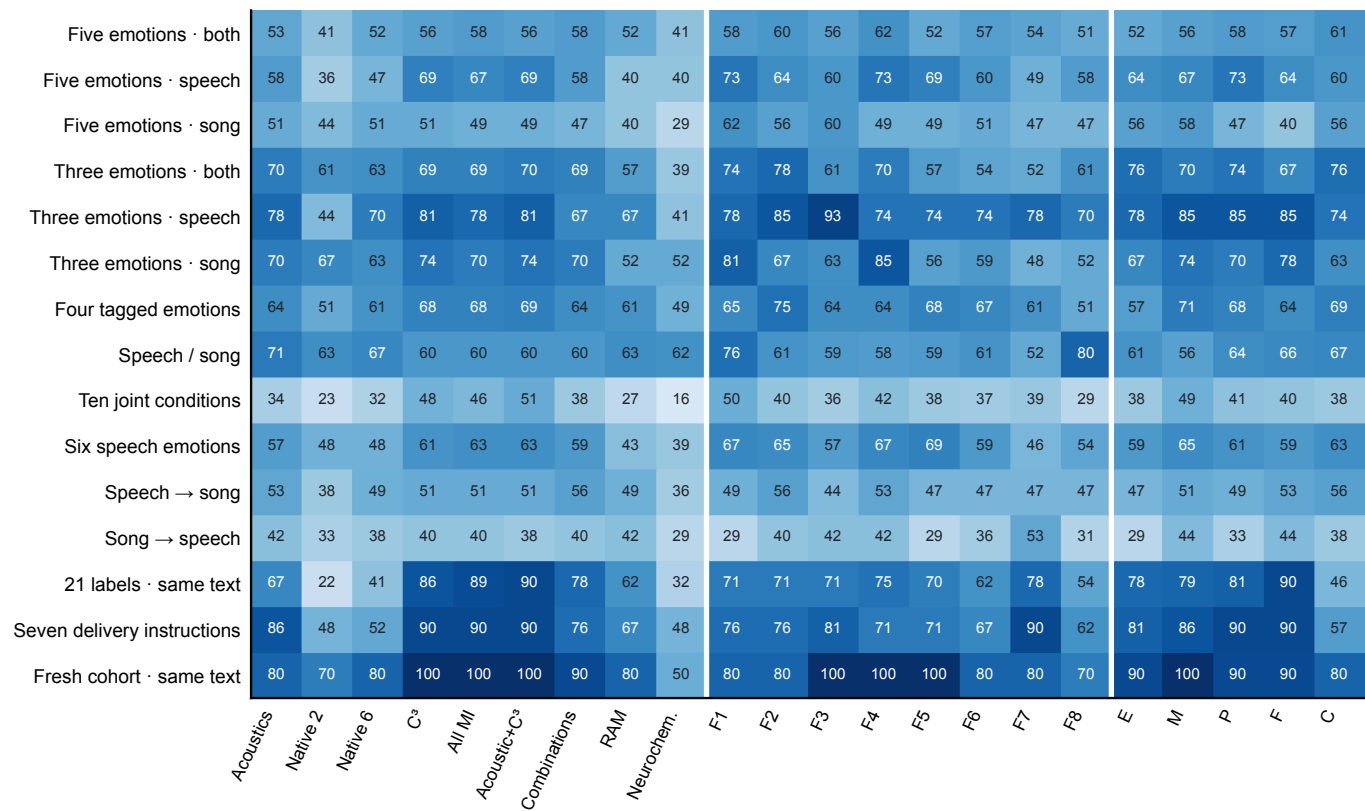
Scalar agreement is not the decoding target

A native-pair decoder can classify an instruction by learning a new boundary even when native valence has the opposite sign from an intended valence ordering. Thus, neither native-pair accuracy nor its change after gain matching repairs the native score automatically. Equally, C³ exceeding a baseline percentage in a particular task does not constitute a significance test of incremental benefit.

Null tests and small held-out sets

The smallest attainable Monte Carlo P value is 0.001 with 999 permutations. Multiple models share requests and output variables, so the valid canonical tests enter a global BY family. Perfect bridge performance means ten correctly classified generation requests, one per condition, on already known text. It does not establish a zero population error rate. Fold-wise performance and all continuous predictions remain available for reanalysis.

S4 | The complete canonical performance landscape



Supplementary Fig. 1 | All canonical feature families and all prediction tasks. Cells show balanced accuracy (%) at fixed ridge $\alpha = 1$. Blue increases with accuracy; each cell is printed, with no significance stars or post hoc selection. Function groups F1–F8 and internal dimension groups E/M/P/F/C retain their architecture identifiers. Values are rounded for print; exact estimates, class recalls, confusion matrices and fold scores are retained electronically. Chance differs between task rows; a bright cell does not by itself indicate above-null performance.

Task coverage is deliberately exhaustive. Adding the untagged condition, pooling or separating delivery, excluding untagged prompts, and requiring transfer between delivery modes each changes the target. A model should not be ranked across such rows as if they were repeated estimates of one common ability. The 21-label, seven-delivery and fresh-bridge tasks specifically reuse known text.

The inferential family differs from the plotted family size: binary delivery has no valid label exchange under the retained cohort restrictions. Its accuracy remains descriptive. No alternative, frame or interpolated point is promoted to an independent sample. Full canonical tables and all individual-mechanism results remain in prediction-summary.json.

S4 | Function-group decoding

Task	F1	F2	F3	F4	F5	F6	F7	F8
Five emotions · both	57.8	60.0	55.6	62.2	52.2	56.7	54.4	51.1
Five emotions · speech	73.3	64.4	60.0	73.3	68.9	60.0	48.9	57.8
Five emotions · song	62.2	55.6	60.0	48.9	48.9	51.1	46.7	46.7
Three emotions · both	74.1	77.8	61.1	70.4	57.4	53.7	51.9	61.1
Three emotions · speech	77.8	85.2	92.6	74.1	74.1	74.1	77.8	70.4
Three emotions · song	81.5	66.7	63.0	85.2	55.6	59.3	48.1	51.9
Four tagged emotions	65.3	75.0	63.9	63.9	68.1	66.7	61.1	51.4
Speech / song	75.6	61.1	58.9	57.8	58.9	61.1	52.2	80.0
Ten joint conditions	50.0	40.0	35.6	42.2	37.8	36.7	38.9	28.9
Six speech emotions	66.7	64.8	57.4	66.7	68.5	59.3	46.3	53.7
Speech → song	48.9	55.6	44.4	53.3	46.7	46.7	46.7	46.7
Song → speech	28.9	40.0	42.2	42.2	28.9	35.6	53.3	31.1
21 labels · same text	71.4	71.4	71.4	74.6	69.8	61.9	77.8	54.0
Seven delivery instructions	76.2	76.2	81.0	71.4	71.4	66.7	90.5	61.9
Fresh cohort · same text	80.0	80.0	100.0	100.0	100.0	80.0	80.0	70.0

Values are balanced accuracy (%). The number of dimensions differs across functions: 127, 110, 113, 138, 132, 70, 132 and 67. Features are standardized within training folds and divided by the square root of the active count before fitting. This makes the numerical regularization convention explicit; it does not equate the intrinsic complexity, redundancy or causal specificity of different groups.

Mechanism-specific findings should be prospectively tested

The strongest observed mechanism-level speech result, IACM at 82.2%, is part of F3; MMP at 80.0% is part of F4. These are observed rankings among 84 available mechanisms. The full C^3 result, 68.9%, is lower in this particular five-way speech task. This pattern motivates dimensionality and regularization tests, but does not prove that the best-ranked mechanism is uniquely responsible for instruction information. All 84 entries, including low-performing ones, are printed in S6.

S4 | Internal dimension groups and permutation results

Task	E	M	P	F	C	Comb.	C ³ P	C ³ BY q
Five emotions · both	52.2	55.6	57.8	56.7	61.1	57.8	0.001	0.0072
Five emotions · speech	64.4	66.7	73.3	64.4	60.0	57.8	0.001	0.0072
Five emotions · song	55.6	57.8	46.7	40.0	55.6	46.7	0.001	0.0072
Three emotions · both	75.9	70.4	74.1	66.7	75.9	68.5	0.001	0.0072
Three emotions · speech	77.8	85.2	85.2	85.2	74.1	66.7	0.001	0.0072
Three emotions · song	66.7	74.1	70.4	77.8	63.0	70.4	0.001	0.0072
Four tagged emotions	56.9	70.8	68.1	63.9	69.4	63.9	0.001	0.0072
Speech / song	61.1	55.6	64.4	65.6	66.7	60.0	—	—
Ten joint conditions	37.8	48.9	41.1	40.0	37.8	37.8	0.001	0.0072
Six speech emotions	59.3	64.8	61.1	59.3	63.0	59.3	0.001	0.0072
Speech → song	46.7	51.1	48.9	53.3	55.6	55.6	0.001	0.0072
Song → speech	28.9	44.4	33.3	44.4	37.8	40.0	0.001	0.0072
21 labels · same text	77.8	79.4	81.0	90.5	46.0	77.8	0.001	0.0072
Seven delivery instructions	81.0	85.7	90.5	90.5	57.1	76.2	0.001	0.0072
Fresh cohort · same text	90.0	100.0	90.0	90.0	80.0	90.0	0.001	0.0072

Accuracy columns are percentages; P and q columns are probabilities. A dash for binary delivery denotes a target fixed within the allowed strata. The C³ P values assess instruction–representation association conditional on exchangeable labels; they do not compare C³ against acoustics. All 308 canonical P values and both BH and BY adjustments are retained, including tests that fail the threshold. There are 291 canonical BY discoveries at $q < 0.05$.

The scalar result differs: among 318,150 exact contrast–output tests, 9,689 meet BH $q < 0.05$ and none meet BY $q < 0.05$. The minimum scalar BY q is 0.1293326. These are different test families, not contradictory summaries of the same null hypothesis. Multivariate prediction and a selected scalar contrast should not be substituted for one another.

S5 | Regularization sensitivity without model selection

Task	Acoustics 0.1 / 1 / 10	C ³ 0.1 / 1 / 10	All MI 0.1 / 1 / 10
Five emotions · both	47.8 / 53.3 / 61.1	48.9 / 55.6 / 52.2	50.0 / 57.8 / 52.2
Five emotions · speech	48.9 / 57.8 / 57.8	60.0 / 68.9 / 48.9	60.0 / 66.7 / 46.7
Five emotions · song	51.1 / 51.1 / 55.6	48.9 / 51.1 / 51.1	48.9 / 48.9 / 48.9
Three emotions · both	72.2 / 70.4 / 79.6	64.8 / 68.5 / 72.2	59.3 / 68.5 / 70.4
Three emotions · speech	85.2 / 77.8 / 85.2	85.2 / 81.5 / 77.8	77.8 / 77.8 / 77.8
Three emotions · song	63.0 / 70.4 / 63.0	74.1 / 74.1 / 74.1	81.5 / 70.4 / 74.1
Four tagged emotions	61.1 / 63.9 / 75.0	69.4 / 68.1 / 62.5	69.4 / 68.1 / 61.1
Speech / song	68.9 / 71.1 / 72.2	64.4 / 60.0 / 63.3	63.3 / 60.0 / 61.1
Ten joint conditions	31.1 / 34.4 / 35.6	42.2 / 47.8 / 41.1	42.2 / 45.6 / 40.0
Six speech emotions	50.0 / 57.4 / 59.3	66.7 / 61.1 / 53.7	66.7 / 63.0 / 51.9
Speech → song	37.8 / 53.3 / 57.8	53.3 / 51.1 / 44.4	53.3 / 51.1 / 44.4
Song → speech	48.9 / 42.2 / 35.6	48.9 / 40.0 / 37.8	46.7 / 40.0 / 40.0
21 labels · same text	66.7 / 66.7 / 69.8	90.5 / 85.7 / 68.3	90.5 / 88.9 / 68.3
Seven delivery instructions	81.0 / 85.7 / 81.0	85.7 / 90.5 / 76.2	85.7 / 90.5 / 71.4
Fresh cohort · same text	80.0 / 80.0 / 90.0	100.0 / 100.0 / 70.0	100.0 / 100.0 / 70.0

Each entry gives balanced accuracy (%) at $\alpha = 0.1, 1$ and 10 , in that order. These 90 additional fits supplement the 1,590 $\alpha = 1$ fits. No test-set accuracy chooses a regularization setting, no “best alpha” performance replaces the canonical value, and no new permutation family is attached to the sensitivity fits.

Training-only preprocessing

For each fold, missing values are imputed with training medians, with a zero fallback only for all-missing training dimensions. Means and standard deviations come from training requests. The active-dimension floor is $16 \text{ float32 epsilons} \times \max(1, \text{maximum absolute training value})$. Features are standardized and divided by $\sqrt{\text{active count}}$. A dual ridge projection with an intercept maps one-hot class labels to test scores. The algebra is independently checked against scikit-learn with maximum discrepancy 7.216×10^{-16} .

S6 | All 84 mechanism results (1–28)

Complete mechanism names and declared dimension counts. Speech and song columns are five-way held-out-sentence balanced accuracies (%), each from 45 test requests. ICC is the median finite request-level ICC(1,1) over that mechanism’s dimensions. These rankings are descriptive; no mechanism is omitted for poor performance.

ID	Full mechanism name	Group	Dims	Speech %	Song %	ICC
BCH	Brainstem Consonance Hierarchy	F1	16	60.0	55.6	0.972
CSG	Consonance-Saliency Gradient	F1	12	62.2	44.4	0.954
MIAA	Musical Imagery Auditory Activation	F1	11	51.1	35.6	0.968
MPG	Melodic Processing Gradient	F1	10	60.0	48.9	0.938
PCCR	Pitch Chroma Cortical Representation	F1	11	66.7	51.1	0.948
PNH	Pythagorean Neural Hierarchy	F1	11	48.9	48.9	0.966
PSCL	Pitch Saliency in Cortical Lateralization	F1	16	62.2	42.2	0.974
SDNPS	Stimulus-Dependent Neural Pitch Saliency	F1	10	57.8	62.2	0.975
SDED	Sensory Dissonance Early Detection	F1	10	57.8	57.8	0.973
TPIO	Timbre Perception-Imagery Overlap	F1	10	44.4	48.9	0.987
TPRD	Tonotopy-Pitch Representation Dissociation	F1	10	55.6	64.4	0.970
HTP	Hierarchical Temporal Prediction	F2	12	71.1	57.8	0.962
SPH	Spatiotemporal Prediction Hierarchy	F2	14	51.1	48.9	0.976
ICEM	Information Content Emotion Model	F2	13	66.7	53.3	0.951
PWUP	Precision-Weighted Uncertainty Processing	F2	10	68.9	66.7	0.974
WMED	Working Memory-Entrainment Dissociation	F2	11	46.7	48.9	0.879
UDP	Uncertainty-Driven Pleasure	F2	10	44.4	51.1	0.964
CHPI	Cross-Modal Harmonic Predictive Integration	F2	11	51.1	55.6	0.971
IGFE	Individual Gamma Frequency Enhancement	F2	9	57.8	31.1	0.970
MAA	Multifactorial Atonal Appreciation	F2	10	55.6	51.1	0.968
PSH	Prediction Silencing Hypothesis	F2	10	60.0	51.1	0.975
SNEM	Sensory Novelty and Expectation Model	F3	12	33.3	44.4	0.823
IACM	Inharmonicity-Attention Capture Model	F3	11	82.2	46.7	0.966
BARM	Brainstem Auditory Response Modulation	F3	10	44.4	42.2	0.912
STANM	Spectrotemporal Attention Network Model	F3	11	64.4	53.3	0.901
AACM	Aesthetic-Attention Coupling Model	F3	10	53.3	40.0	0.896
AMSS	Attention-Modulated Stream Segregation	F3	11	42.2	35.6	0.930
ETAM	Entrainment, Tempo and Attention Modulation	F3	11	48.9	53.3	0.941

S6 | All 84 mechanism results (29–56)

Complete mechanism names and declared dimension counts. Speech and song columns are five-way held-out-sentence balanced accuracies (%), each from 45 test requests. ICC is the median finite request-level ICC(1,1) over that mechanism’s dimensions. These rankings are descriptive; no mechanism is omitted for poor performance.

ID	Full mechanism name	Group	Dims	Speech %	Song %	ICC
DGTP	Domain-General Temporal Processing	F3	9	55.6	51.1	0.921
SDL	Salience-Dependent Lateralization	F3	9	33.3	31.1	0.836
NEWMD	Neural Entrainment-Working Memory Dissociation	F3	10	46.7	51.1	0.952
PWSM	Precision-Weighted Salience Model	F3	9	51.1	57.8	0.926
MEAMN	Music-Evoked Autobiographical Memory Network	F4	12	55.6	42.2	0.973
MMP	Musical Mnemonic Preservation	F4	12	80.0	44.4	0.921
RASN	Rhythmic Auditory Stimulation Neuroplasticity	F4	11	62.2	51.1	0.888
PMIM	Predictive Memory Integration Model	F4	11	66.7	55.6	0.936
OII	Oscillatory Intelligence Integration	F4	10	66.7	66.7	0.975
HCMC	Hippocampal-Cortical Memory Circuit	F4	11	55.6	44.4	0.959
RIRI	RAS-Intelligent Rehabilitation Integration	F4	10	42.2	40.0	0.956
MSPBA	Musical Syntax Processing in Broca's Area	F4	11	60.0	60.0	0.958
VRIAP	VR-Integrated Analgesia Paradigm	F4	10	46.7	35.6	0.883
CMAPEC	Cross-Modal Action-Perception Common Code	F4	10	60.0	42.2	0.974
DMMS	Developmental Music Memory Scaffold	F4	10	40.0	40.0	0.978
CSSL	Cross-Species Song Learning	F4	10	66.7	66.7	0.970
CDEM	Context-Dependent Emotional Memory	F4	10	51.1	46.7	0.976
SRP	Striatal Reward Prediction	F5	19	42.2	40.0	0.931
AAC	Autonomic-Arousal Circuit	F5	14	35.6	44.4	0.880
VMM	Valence-Mode Mapping	F5	12	64.4	33.3	0.959
PUPF	Pleasure-Uncertainty Prediction Function	F5	12	44.4	31.1	0.734
CLAM	Closed-Loop Affective Modulation	F5	11	51.1	44.4	0.969
MAD	Musical Anhedonia Disconnection	F5	11	55.6	57.8	0.948
NEMAC	Nostalgia-Evoked Memory-Affect Circuit	F5	11	40.0	33.3	0.962
STAI	Spectral-Temporal Aesthetic Integration	F5	12	37.8	46.7	0.971
DAP	Developmental Affective Plasticity	F5	10	51.1	46.7	0.970
CMAT	Cross-Modal Affective Transfer	F5	10	60.0	46.7	0.956
TAR	Therapeutic Affective Resonance	F5	10	46.7	37.8	0.922

S6 | All 84 mechanism results (57–84)

Complete mechanism names and declared dimension counts. Speech and song columns are five-way held-out-sentence balanced accuracies (%), each from 45 test requests. ICC is the median finite request-level ICC(1,1) over that mechanism’s dimensions. These rankings are descriptive; no mechanism is omitted for poor performance.

ID	Full mechanism name	Group	Dims	Speech %	Song %	ICC
DAED	Dopamine Anticipation-Experience Dissociation	F6	8	37.8	40.0	0.787
MORMR	mu-Opioid Receptor Music Reward	F6	7	37.8	37.8	0.763
RPEM	Reward Prediction Error in Music	F6	8	35.6	40.0	0.964
IUCP	Inverted-U Complexity Preference	F6	6	46.7	51.1	0.964
MCCN	Musical Chills Cortical Network	F6	7	40.0	42.2	0.918
MEAMR	Music-Evoked Autobiographical Memory Reward	F6	6	24.4	22.2	0.365
SSRI	Social Synchrony Reward Integration	F6	11	48.9	60.0	0.641
LDAC	Liking-Dependent Auditory Cortex	F6	6	51.1	48.9	0.969
IOTMS	Individual Opioid Tone Music Sensitivity	F6	5	24.4	40.0	0.936
SSPS	Saddle-Shaped Preference Surface	F6	6	44.4	40.0	0.928
PEOM	Period Entrainment Optimization Model	F7	11	44.4	55.6	0.912
MSR	Musician Sensorimotor Reorganization	F7	11	55.6	40.0	0.952
GSSM	Gait-Synchronized Stimulation Model	F7	11	46.7	51.1	0.946
HMCE	Hierarchical Musical Context Encoding	F7	11	42.2	42.2	0.835
HGSIC	Hierarchical Groove State Integration Circuit	F7	11	40.0	48.9	0.953
ASAP	Action Simulation for Auditory Prediction	F7	11	44.4	46.7	0.920
DDSMI	Dyadic Dance Social Motor Integration	F7	11	44.4	42.2	0.803
VRMSME	VR Music Stimulation Motor Enhancement	F7	11	48.9	51.1	0.850
SPMC	SMA-Premotor-M1 Motor Circuit	F7	11	35.6	48.9	0.902
NSCP	Neural Synchrony Commercial Prediction	F7	11	53.3	48.9	0.920
CTBB	Cerebellar Theta-Burst Balance	F7	11	44.4	35.6	0.874
STC	Singing Training Connectivity	F7	11	44.4	55.6	0.906
EDNR	Expertise-Dependent Network Reorganization	F8	10	48.9	40.0	0.960
TSCP	Temporal Synaptic Consolidation Processor	F8	10	55.6	37.8	0.974
CDMR	Context-Dependent Mismatch Response	F8	11	44.4	35.6	0.969
SLEE	Synaptic Long-term Encoding Engine	F8	13	42.2	20.0	0.961
ESME	Error-Signal Modulated Encoding	F8	11	42.2	35.6	0.977
ECT	Expertise Compartmentalization Trade-off	F8	12	46.7	40.0	0.926

S7 | All main native-affect contrasts

Signed first-minus-second differences, averaged equally over the three texts. V denotes native VMM valence and A denotes native AAC arousal. Exact P values are available only for eligible, within-cohort comparisons. BY values use the full 318,150-test scalar family. Cross-cohort contrasts remain descriptive; the complete registry retains their confounding flags.

Main contrast	ΔV	ΔA	P(V)	P(A)	BY(V)	BY(A)
Sp:untagged – Sp:excited	0.01717	-0.00014	0.0003	0.8965	0.1293	1.0000
Sp:untagged – Sp:sad	-0.01253	0.00934	0.0003	0.0003	0.1293	0.1293
Sp:untagged – Sp:mad	0.02785	-0.00430	0.0003	0.0005	0.1293	0.2396
Sp:untagged – Sp:peaceful	-0.02055	0.00925	0.0003	0.0003	0.1293	0.1293
Sp:untagged – So:untagged	-0.00603	0.00843	—	—	—	—
Sp:untagged – So:excited	-0.00645	0.00246	—	—	—	—
Sp:untagged – So:sad	-0.00764	0.01611	—	—	—	—
Sp:untagged – So:mad	0.02781	-0.00458	—	—	—	—
Sp:untagged – So:peaceful	-0.01035	0.01291	—	—	—	—
Sp:excited – Sp:sad	-0.02970	0.00948	0.0003	0.0003	0.1293	0.1293
Sp:excited – Sp:mad	0.01067	-0.00416	0.0005	0.0045	0.2396	1.0000
Sp:excited – Sp:peaceful	-0.03772	0.00939	0.0003	0.0003	0.1293	0.1293
Sp:excited – So:untagged	-0.02320	0.00857	—	—	—	—
Sp:excited – So:excited	-0.02362	0.00260	—	—	—	—
Sp:excited – So:sad	-0.02481	0.01625	—	—	—	—
Sp:excited – So:mad	0.01064	-0.00444	—	—	—	—
Sp:excited – So:peaceful	-0.02752	0.01305	—	—	—	—
Sp:sad – Sp:mad	0.04038	-0.01364	0.0003	0.0003	0.1293	0.1293
Sp:sad – Sp:peaceful	-0.00802	-0.00009	0.0020	0.9145	0.8497	1.0000
Sp:sad – So:untagged	0.00650	-0.00091	—	—	—	—
Sp:sad – So:excited	0.00608	-0.00688	—	—	—	—
Sp:sad – So:sad	0.00490	0.00677	—	—	—	—
Sp:sad – So:mad	0.04035	-0.01392	—	—	—	—

The electronic arrays also retain median, within-recording SD, p10 and p90 contrasts; standardized mean differences; request-pair ordering probabilities; and pointwise bootstrap bounds. They are not replaced by these rounded native-mean tables. Full hypotheses span 556 comparisons and all 1,050 outputs.

S7 | All main native-affect contrasts (continued)

Signed first-minus-second differences, averaged equally over the three texts. V denotes native VMM valence and A denotes native AAC arousal. Exact P values are available only for eligible, within-cohort comparisons. BY values use the full 318,150-test scalar family. Cross-cohort contrasts remain descriptive; the complete registry retains their confounding flags.

Main contrast	ΔV	ΔA	P(V)	P(A)	BY(V)	BY(A)
Sp:sad – So:peaceful	0.00219	0.00357	—	—	—	—
Sp:mad – Sp:peaceful	-0.04840	0.01355	0.0003	0.0003	0.1293	0.1293
Sp:mad – So:untagged	-0.03388	0.01273	—	—	—	—
Sp:mad – So:excited	-0.03430	0.00676	—	—	—	—
Sp:mad – So:sad	-0.03548	0.02040	—	—	—	—
Sp:mad – So:mad	-0.00003	-0.00028	—	—	—	—
Sp:mad – So:peaceful	-0.03819	0.01721	—	—	—	—
Sp:peaceful – So:untagged	0.01452	-0.00082	—	—	—	—
Sp:peaceful – So:excited	0.01410	-0.00679	—	—	—	—
Sp:peaceful – So:sad	0.01292	0.00685	—	—	—	—
Sp:peaceful – So:mad	0.04837	-0.01383	—	—	—	—
Sp:peaceful – So:peaceful	0.01021	0.00366	—	—	—	—
So:untagged – So:excited	-0.00042	-0.00597	—	—	—	—
So:untagged – So:sad	-0.00160	0.00767	—	—	—	—
So:untagged – So:mad	0.03385	-0.01301	—	—	—	—
So:untagged – So:peaceful	-0.00431	0.00448	—	—	—	—
So:excited – So:sad	-0.00119	0.01365	0.7732	0.0003	1.0000	0.1293
So:excited – So:mad	0.03426	-0.00704	0.0003	0.0003	0.1293	0.1293
So:excited – So:peaceful	-0.00389	0.01045	0.4263	0.0003	1.0000	0.1293
So:sad – So:mad	0.03545	-0.02068	0.0003	0.0003	0.1293	0.1293
So:sad – So:peaceful	-0.00271	-0.00319	0.1288	0.0013	1.0000	0.5511
So:mad – So:peaceful	-0.03816	0.01749	0.0003	0.0003	0.1293	0.1293

The electronic arrays also retain median, within-recording SD, p10 and p90 contrasts; standardized mean differences; request-pair ordering probabilities; and pointwise bootstrap bounds. They are not replaced by these rounded native-mean tables. Full hypotheses span 556 comparisons and all 1,050 outputs.

S8 | Native directions under the paired gain intervention

Contrast / output / mode	Original mean	Common-RMS mean	Aligned texts original → gain
excited – sad / valence / speech	-0.029705	-0.026638	0/3 → 0/3
excited – sad / valence / song	-0.001187	0.003316	1/3 → 1/3
excited – sad / arousal / speech	0.009479	0.000302	3/3 → 2/3
excited – sad / arousal / song	0.013646	0.006727	3/3 → 3/3
peaceful – mad / valence / speech	0.048399	0.044167	3/3 → 3/3
peaceful – mad / valence / song	0.038158	0.035537	3/3 → 3/3
peaceful – mad / arousal / speech	-0.013550	0.001718	3/3 → 0/3
peaceful – mad / arousal / song	-0.017489	0.000541	3/3 → 2/3

“Aligned” refers to the predefined instruction-level sign: positive excited-minus-sad valence and arousal; positive peaceful-minus-mad valence; negative peaceful-minus-mad arousal. A pooled sign can differ from the majority of text-specific signs. For example, song peaceful-minus-mad arousal becomes slightly positive on average while two of three text differences remain negative.

The common target is -29.517516 dBFS, the minimum original RMS. Each copied waveform receives a constant gain no greater than one. The operation preserves sample rate and sample count; no pitch shifting, time stretching, dynamic compression, denoising or new synthesis is performed. It does not equate perceived loudness or remove relative temporal-envelope differences. All 342 original audio inputs and all 342 gain outputs are accounted for.

Gain sensitivity and native interpretation

Excited-minus-sad speech arousal decreases by 96.8% in its equal-text mean, whereas excited-minus-sad valence remains negative in every text. This makes a single explanation for both native scores untenable within the tested intervention. It does not prove that all arousal information is waveform level, nor that the residual valence mismatch is independent of every acoustic variable. The exact algebraic decomposition below identifies what changed in this implementation.

S8 | Exact weighted-score components: speech

Values are contributions to a contrast in model units, after the implemented weights. E–S is excited minus sad; P–M is peaceful minus mad. O is original and G common RMS. Sum the five VMM terms plus its clipping adjustment for native valence, or the four AAC terms plus its clipping adjustment for native arousal. The reconstruction error is zero. These are algebraic contributions, not independent interventions on neural mechanisms.

Weighted component	E–S O	E–S G	P–M O	P–M G
VMM · valence V0	-0.008929	-0.007424	0.018328	0.015586
VMM · pleasantness H3	-0.007752	-0.003740	0.015453	0.008143
VMM · reward evaluation	-0.003443	-0.006271	0.003601	0.009094
VMM · happy minus sad	-0.004033	-0.003654	-0.000127	0.000199
VMM · inverse roughness	-0.005548	-0.005548	0.011144	0.011144
VMM · clipping adjustment	0.000000	0.000000	-0.000000	-0.000000
AAC · emotion times amplitude	0.005359	0.001324	-0.007569	-0.000734
AAC · ANS times SCR	-0.001238	-0.001863	0.001730	0.002811
AAC · chills proxy	0.000836	-0.000522	-0.001372	0.000867
AAC · HR times emotion	0.004522	0.001363	-0.006339	-0.001226
AAC · clipping adjustment	0.000000	0.000000	-0.000000	-0.000000

Model equations

VMM:C1 = clip[0.25 V0 + 0.20 H3 pleasantness + 0.20 reward evaluation + 0.20 (happy – sad) + 0.15 (1 – roughness)].
 AAC:P2 = clip[0.25 emotional arousal × amplitude + 0.25 ANS × SCR proxy + 0.25 chills proxy + 0.25 heart-rate proxy × emotional arousal]. clip restricts the sum to the implemented 0–1 interval. Clipping adjustments are retained explicitly rather than assumed absent.

A constant gain can change both an explicit amplitude factor and upstream nonlinear representations. Therefore the amount attributed to a product term is a decomposition of this score, not a unique causal allocation of the gain effect. The full per-frame, per-recording original and gain component arrays are retained in `experiment/analysis/comprehensive/components/`.

S8 | Exact weighted-score components: song

Values are contributions to a contrast in model units, after the implemented weights. E–S is excited minus sad; P–M is peaceful minus mad. O is original and G common RMS. Sum the five VMM terms plus its clipping adjustment for native valence, or the four AAC terms plus its clipping adjustment for native arousal. The reconstruction error is zero. These are algebraic contributions, not independent interventions on neural mechanisms.

Weighted component	E–S O	E–S G	P–M O	P–M G
VMM · valence V0	0.000751	0.002307	0.016139	0.013539
VMM · pleasantness H3	-0.002248	0.001901	0.014510	0.007577
VMM · reward evaluation	-0.001399	-0.003596	0.002948	0.009091
VMM · happy minus sad	-0.003774	-0.002779	-0.002062	-0.001294
VMM · inverse roughness	0.005483	0.005484	0.006623	0.006623
VMM · clipping adjustment	0.000000	0.000000	-0.000000	-0.000000
AAC · emotion times amplitude	0.006017	0.003017	-0.008687	-0.000954
AAC · ANS times SCR	-0.000063	-0.000581	0.000938	0.002365
AAC · chills proxy	0.002864	0.001864	-0.002016	0.000643
AAC · HR times emotion	0.004829	0.002427	-0.007723	-0.001514
AAC · clipping adjustment	0.000000	0.000000	-0.000000	-0.000000

Model equations

VMM:C1 = clip[0.25 V0 + 0.20 H3 pleasantness + 0.20 reward evaluation + 0.20 (happy – sad) + 0.15 (1 – roughness)].
 AAC:P2 = clip[0.25 emotional arousal × amplitude + 0.25 ANS × SCR proxy + 0.25 chills proxy + 0.25 heart-rate proxy × emotional arousal]. clip restricts the sum to the implemented 0–1 interval. Clipping adjustments are retained explicitly rather than assumed absent.

A constant gain can change both an explicit amplitude factor and upstream nonlinear representations. Therefore the amount attributed to a product term is a decomposition of this score, not a unique causal allocation of the gain effect. The full per-frame, per-recording original and gain component arrays are retained in [experiment/analysis/comprehensive/components/](#).

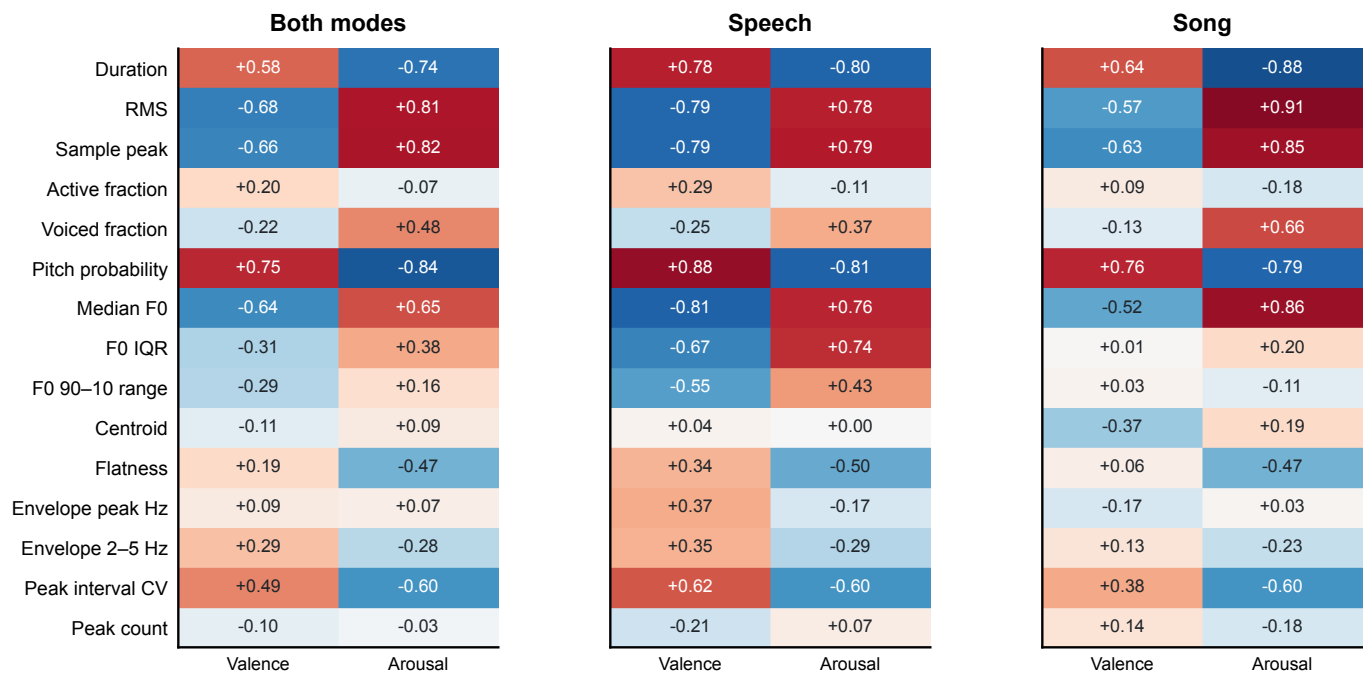
S9 | Gain sensitivity across every canonical model family

Five emotions · both	+7	-1	+0	-1	+6	-7	-22	+9	-8	-2	-4	+6	-7	-3	-1	-1	-2	-1	-2	+6
Five emotions · speech	+20	+22	+4	+7	+4	+13	-9	+4	-7	+4	+2	-2	-2	+20	-13	+9	+0	-2	+9	+7
Five emotions · song	+4	-4	-2	+0	+0	-2	-9	+0	-4	-7	-2	-4	-11	-9	-9	-2	-7	+0	+0	-16
Three emotions · both	-15	-13	-2	-2	-7	-6	-9	-6	-19	-9	-2	-4	-2	-2	-7	-15	-19	-13	+0	-13
Three emotions · speech	+11	+0	-11	-11	+0	-4	-15	+11	-15	-19	-4	-0	-7	-15	+0	+0	-11	-19	-11	-7
Three emotions · song	-7	-26	+4	+4	-4	-4	+4	+7	+4	+4	-30	+0	-15	-15	-19	+7	+0	+7	-19	+4
Four tagged emotions	+4	+0	+7	+6	+11	-8	-21	+8	+3	-4	+3	+1	-4	+6	+3	+10	-3	+0	+11	+1
Speech / song	+3	-12	+0	+0	-4	+1	-8	-16	-3	-6	+2	-3	-6	-2	-11	-3	+2	-9	-8	-2
Ten joint conditions	+6	-2	-3	+0	+8	+10	-3	-1	-10	+2	-7	+6	-8	-4	+3	+4	-7	-0	-2	-2
Six speech emotions	-6	-2	+2	+0	+0	+11	-2	+2	-11	+4	+4	-6	-4	+7	-13	-4	+0	+0	+0	+2
Speech → song	-2	-7	-13	-9	-9	-11	-11	-9	-16	+2	-9	-16	-7	-4	-9	+0	-13	-13	-11	-13
Song → speech	+7	+7	+2	+2	+4	-4	-7	+11	+4	+7	+4	+0	+4	+0	+13	+13	-2	+2	+9	-13
21 labels · same text	+0	+6	+3	-2	+11	-0	-5	+2	-3	+0	+2	-2	-2	+2	-3	+5	+6	+5	-2	-3
Seven delivery instructions	-5	+0	+0	+0	+5	+0	-5	-10	-5	+0	-5	+5	-5	+0	-14	+0	-5	-5	-5	-10
Fresh cohort · same text	-10	-10	+0	+0	+10	+10	+10	+10	+10	-10	+0	+0	+0	+10	+0	+0	+0	+10	+10	+10
	Native 2	Native 6	C ³	All MI	Combinations	RAM	Neurochem.	F1	F2	F3	F4	F5	F6	F7	F8	E	M	Q	L	U

Supplementary Fig. 2 | Change in balanced accuracy after common-RMS reanalysis. Each cell is gain-condition accuracy minus original accuracy, in percentage points. Blue denotes improvement and red decline; the scale is symmetric around zero. All 15 tasks and 20 MI feature families are included. The same task definitions, request folds and fixed ridge setting are used. Processed copies do not add independent replication. No processed-acoustic baseline, new null test or best-preprocessing selection is implied. Exact values, gain predictions and request-level correct-to-wrong/wrong-to-correct changes are preserved.

Five-way speech C³ improves from 68.9% to 73.3%, whereas song changes from 51.1% to 48.9%. The native two-feature speech probe improves from 35.6% to 55.6%, despite the persistent excited-minus-sad native-valence reversal. These observations reinforce the distinction between a fitted decision rule and the direction of a native scalar. A positive cell is a descriptive paired sensitivity result, not a claim that gain matching universally improves MI.

S10 | Acoustic associations and predictability



Supplementary Fig. 3 | Native outputs and sentence-centred acoustic variation. Spearman correlations between the 15 acoustic descriptors and native valence/arousal, separately for both modes (90 requests), speech (45) and song (45). Sentence means are removed before ranking. Red is positive and blue negative on a fixed -1 to $+1$ scale. Each descriptor uses complete pairs; the underlying count is retained. This figure is descriptive and supplies neither causal attribution nor an independent test of every cell.

The archive contains six scopes: main all, main speech, main song, broad emotion, delivery and all requests. Each has raw and sentence-centred Pearson and Spearman coefficients against all 1,050 outputs. No acoustic descriptor is used to exclude a recording. Undefined pitch or other measurements remain missing and are handled only at the relevant estimator or within-fold imputation stage.

Acoustic-to-MI prediction

Ridge regression predicts every MI readout from the acoustic baseline under full sentence exclusion. Aggregate R^2 , mean absolute error, correlation, all held-out predictions and fold R^2 are retained. Median R^2 is 0.487 for C^3 , 0.463 for combinations, 0.109 for RAM and 0.473 for neurochemical proxies. Native valence and arousal yield 0.703 and 0.319. Negative predictions are retained; R^2 is undefined for three C^3 dimensions. These values do not identify “non-acoustic cognition” in the unexplained residual.

S11 | Temporal measurements and geometric scope

Temporal quantity	Definition / unit
Normalized area	Trapezoidal integral divided by final model-frame time; native units
Range	Maximum minus minimum model value; native units
Total variation	Sum of absolute frame differences divided by duration; units/s
Linear slope	Least-squares slope against model time; units/s
Peak time	Time of maximal response; seconds; undefined for constant trajectories
Late – early	Mean final 20% minus mean initial 20%; native units
Lag correlation	Correlation between response and its one-second shifted version; undefined when denominator vanishes

Frame interpolation creates 101 normalized-phase samples and 91 absolute samples on 0–4.5 s, a support interval available in every recording. The 220 main trajectory contrasts cover main pairs and delivery-by-emotion interactions. Metric contrasts cover all 556 comparisons. Metrics are computed per recording; finite alternative metrics are averaged per request with valid-alternative counts retained. Undefined contributing request metrics keep their contrast undefined; zero-weight unrelated requests cannot propagate NaN into a valid contrast.

Geometry

Condition centroids are standardized using the SD across 90 main request means. Near-constant dimensions are removed. Euclidean distance is divided by $\sqrt{\text{active count}}$; correlation distance is also stored. There are 105 feature families, nine panels and two metrics, or 1,890 matrices. Emotion transfer includes ten pairwise contrasts per feature family, with cross-delivery vector cosine, song/speech norm ratio and within-mode distance-rank agreement. The C^3 speech–song emotion-geometry correlation is 0.721212. No P value treats the ten shared edges as independent observations.

What is not identified

A peak in a model trajectory is not neural latency. Phase-normalized time is not word alignment. “Memory” or “forecast” dimensions in an offline architecture are not evidence of a measured cognitive timescale. Study 1 contains generated anchors rather than intermediate DSP morphs. Study 2, reported in S15–S20, separately evaluates a controlled spectral intervention. Neither experiment estimates a continuous human perceptual boundary.

S12 | Independent numerical verification

Verification check	Result
all 342 original recordings included	PASS
172 unique requests	PASS
alternatives partition recordings once	PASS
equal weight request pooling	PASS
frozen plan hash	PASS
original catalog inputs unchanged	PASS
contrast coefficients sum to zero	PASS
all scalar differences reconstruct	PASS
scipy exact partitions match	PASS
BY correction reproduces	PASS
request disjoint all splits	PASS
sentence disjoint all new text splits	PASS
independent sklearn ridge match	PASS
orthogonal variance closure	PASS
undefined temporal values scoped to contributors	PASS
342 gain runs complete	PASS
native weighted component closure	PASS
attenuation only	PASS

All 18 checks passed. The independent ridge comparison differs by at most $7.216449660063518 \times 10^{-16}$. Native weighted components close with zero maximum discrepancy. Exact partition enumeration and BY adjustment are independently reproduced. These checks validate implementation consistency, data inclusion and numerical bookkeeping; they do not validate the biological meaning of MI’s labels.

S12 | Reproducibility and preservation map

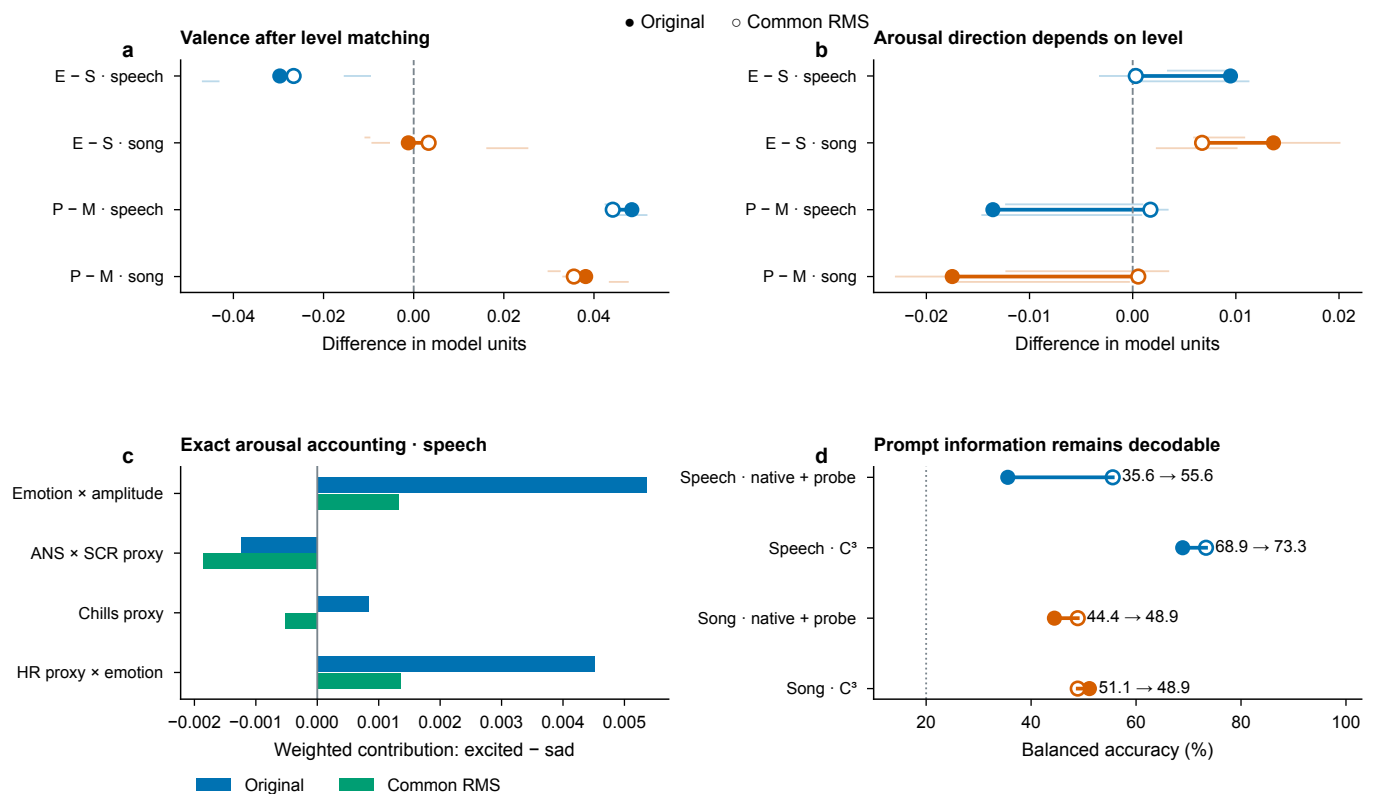
Archive location (under experiment/)	Contents and unit
manifest.json; protocol/; receipts/	Exact generation conditions, intended text, requests and production receipts
audio/; raw/; alignment/	All original recordings, provider downloads and transcript/time estimates
analysis/mi/; provenance/mi-engine/	Original model outputs and pinned engine source
controls/comprehensive-gain/	342 constant-gain waveform copies and generation metadata
analysis/comprehensive/catalog.json	Master record, request, readout and feature-family ordering
scalar-inputs.npz; contrasts.npz	Recording/request summaries; 556 contrasts; exact P/q and bootstrap arrays
temporal-*.npz	Two time grids, full curves, seven metrics and metric contrasts
prediction/; prediction-summary.json	Original held-out scores, labels, nulls, folds and 1,680 probe results
gain-mi/; gain-prediction/; gain-comparison.npz	342 reruns, 300 gain probes, paired differences and native component contrasts
variance-repeatability.npz	Eight variance terms, ICCs and every alternative-pair difference
acoustic-associations.npz; acoustics-predict-MI.npz	All correlations and held-out acoustic-to-model predictions
representational-distances.npz; geometry.json	1,890 matrices and all emotion-transfer vector summaries
validation.json; analysis-identity.json	Independent checks, pinned identities and execution provenance

Unqualified result filenames in the table reside in analysis/comprehensive/. Manuscript-specific additions are outside experiment/: data/ includes all-readouts.csv, all-mechanisms.csv, full request/recording registries, exact supplement tables and source JSON for each figure. code/ contains the manuscript, numerical figure scripts and PDF builders. qa/ contains archive hashes, layout records and page-render verification. The original SHA inventory is preserved separately; experiment/ links to the unchanged source archive. Keep that source location with any redistribution.

Reproduction order

Use the original pinned analysis environment and run comprehensive_prepare, statistics, prediction, acoustics_geometry and temporal; use the frozen MI environment for comprehensive_gain, then gain_analysis, audit and report. The manuscript figures read existing arrays and do not regenerate stimuli or refit MI. Replacement of virtual environments or downloadable model caches is allowed only with the recorded dependencies; original audio, prompts, scientific arrays, code and identities remain preserved. The package README records the exact commands used for this edition.

S13 | Study 1: paired level intervention



Supplementary Fig. 4 | Matching waveform RMS alters native arousal without erasing C³ 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.

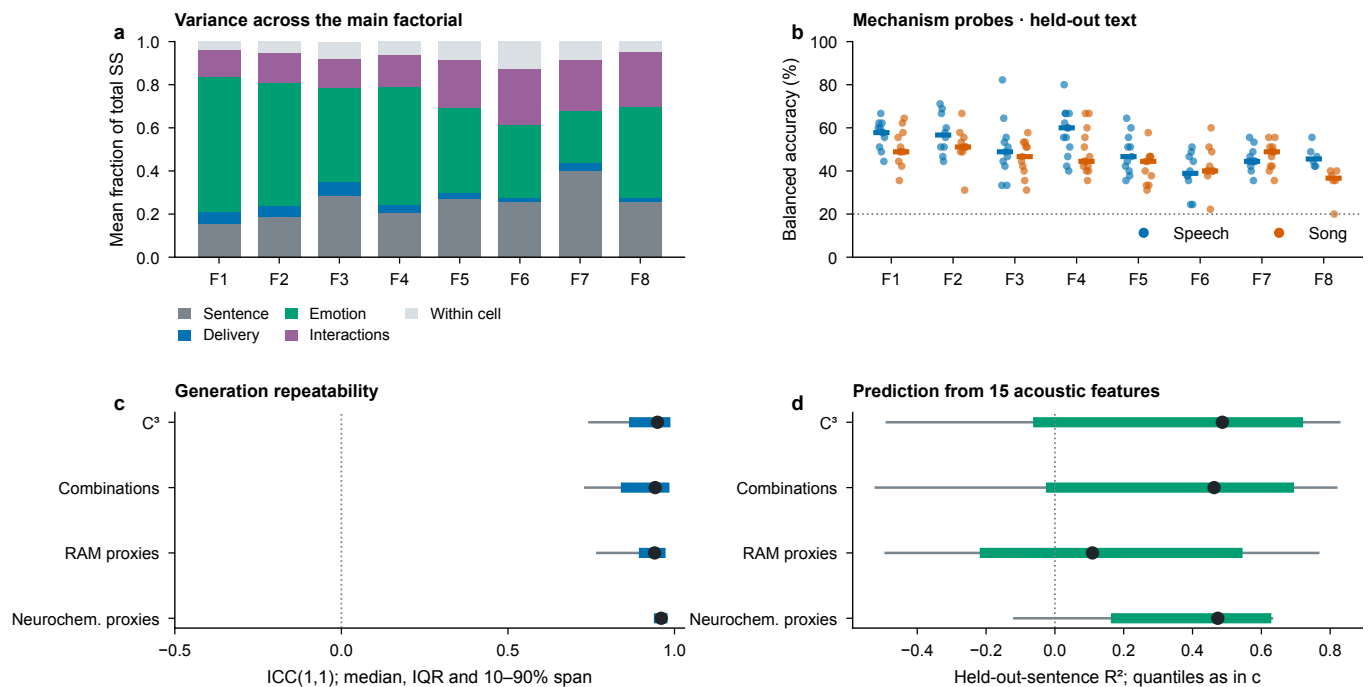
A constant gain intervention provides a direct test of whether observed scores depend on waveform level. Each recording was attenuated to the lowest RMS in the original corpus, preserving duration, sample count, pitch and the relative temporal envelope. Required attenuation ranged from 0 to -10.8538 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 (Supplementary 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 C³ 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.

S13 | Study 1: the complete response repertoire



Supplementary 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 C³ 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 R³/T³ variables may contribute to C³ 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.

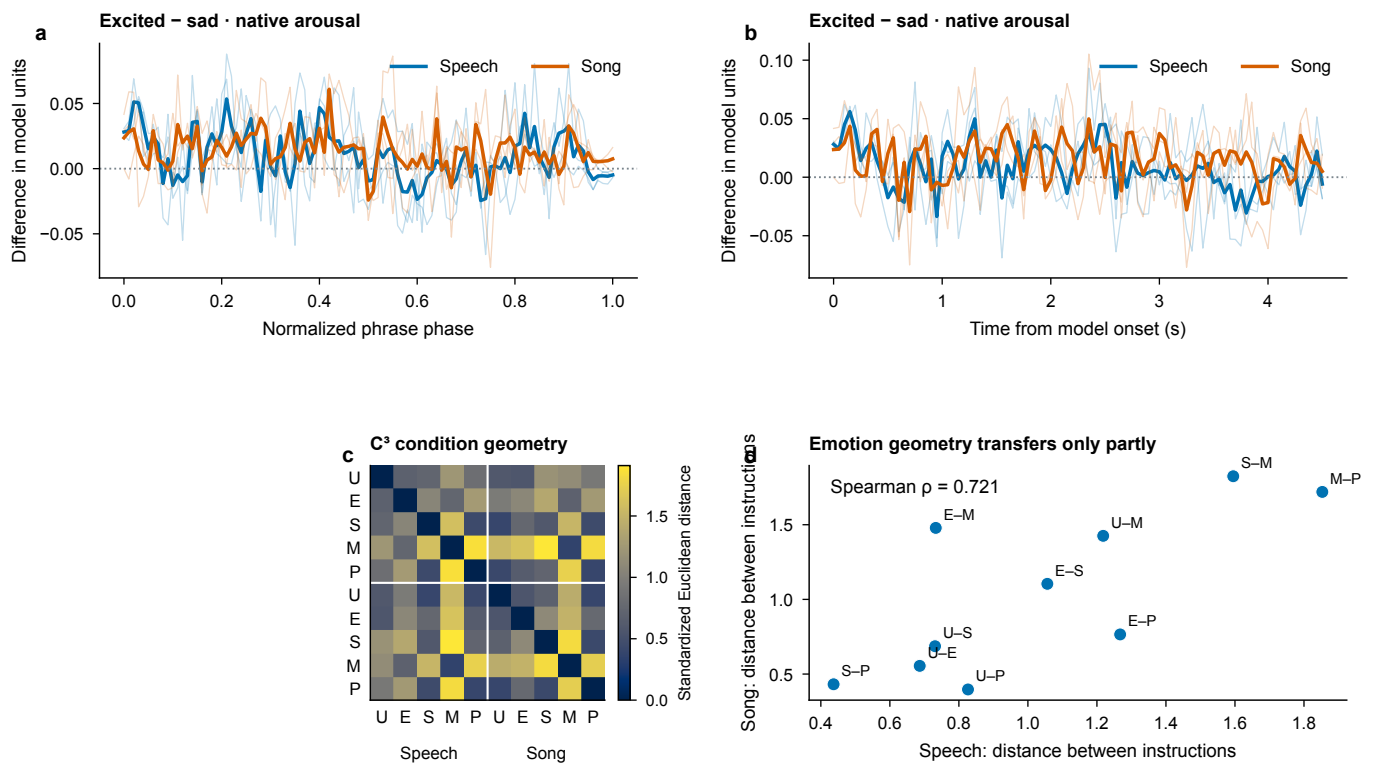
A balanced variance decomposition separated text, delivery, emotion, their interactions and within-cell request variation. Native valence allocated 55.3% of total sum of squares to emotion, 24.6% to sentence and 1.9% to within-cell variation; native arousal allocated 71.6%, 4.2% and 4.7%, respectively. 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 (Supplementary 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 C³ 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 R² = 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 C³, 0.942 for combinations, 0.940 for regional projections and 0.960 for neurochemical proxies. Median held-out acoustic-prediction R² values were 0.487, 0.463, 0.109 and 0.473, respectively. Three C³ dimensions had undefined R²; 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.

S13 | Study 1: time and geometry



Supplementary 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^3 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 (Supplementary 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^3 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.²⁷

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.

S14 | Complete Study 1 methods from the preceding edition

These methods preserve the original study in full. Their corpus and gain-control values refer to Study 1. Study 2 uses a separate six-source intervention design and a different frozen common-RMS target. Numbered references refer to the integrated main article. No new model inference or new hypothesis test was performed for the editorial revision.

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;²¹ 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.

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.⁵⁶ 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 $16 \text{ float32 epsilons} \times \max(1, \text{maximum absolute training value})$ 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.²² 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.^{23,24} 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 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 $\sqrt{\text{active dimensions}}$. 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 7.2×10^{-16} ; native component closure error was zero. NumPy/SciPy and scikit-learn versions are recorded with source hashes.²⁶ 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.

S15 | Study 2: source lock, sampling and independence

The intervention reuses six archived recordings from Study 1: one selected take/alternative per excited or sad instruction and per text. Selection used the first available take and lexicographic alternative, without transformed MI scores. The text is matched within each pair, but the two expressive recordings do not have identical prosody or segment timing. No new speakers, requests or independently labelled emotions were added.

Text / instruction	Source ID	Duration (s)	F0 median (Hz)
S01 / excited	emotion-E01-01-a	5.956	232.98
S01 / sad	emotion-E13-01-a	6.426	197.04
S02 / excited	emotion-S02-E01-01-a	4.989	248.26
S02 / sad	emotion-S02-E13-01-a	5.486	191.43
S03 / excited	emotion-S03-E01-01-a	5.956	207.56
S03 / sad	emotion-S03-E13-01-a	6.034	175.55

Evaluation family	Count	Purpose
Primary full files	1,596	49 joint anchors + 5x5 controls; shared conditions deduplicated
Unprotected renderer	54	Pitch-guard sensitivity; original pilot inputs retained
Isolated peak correction	72	One-operation numerical control on existing inputs
Near-endpoint full files	192	8 additional q x 6 sources x 2 envelopes x 2 levels
All analytic full-file evaluations	1,914	1,836 unique decoded waveforms; 1,842 file hashes
Feature clamps	72	3 feature holds x 24 endpoints; not additional audio
Exact no-change feature controls	24	Reproduce endpoint arrays with no feature replacement

Evaluation count is not independent sample size. Conditions share source clips; corrected-engine evaluations reuse audio, and some WAV files differ only in metadata. The raw source archive retains original, derived and repeated files separately. Benchmark and engineering repeats are excluded from the analytic count. The actual fixed-source comparisons are descriptive, without population P values or anchor-based confidence intervals.

S16 | Study 2: renderer, registration and deviations

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 $q = 0, 1/48, \dots, 1$, complemented by a $5 \times 5 \beta/\rho$ 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.

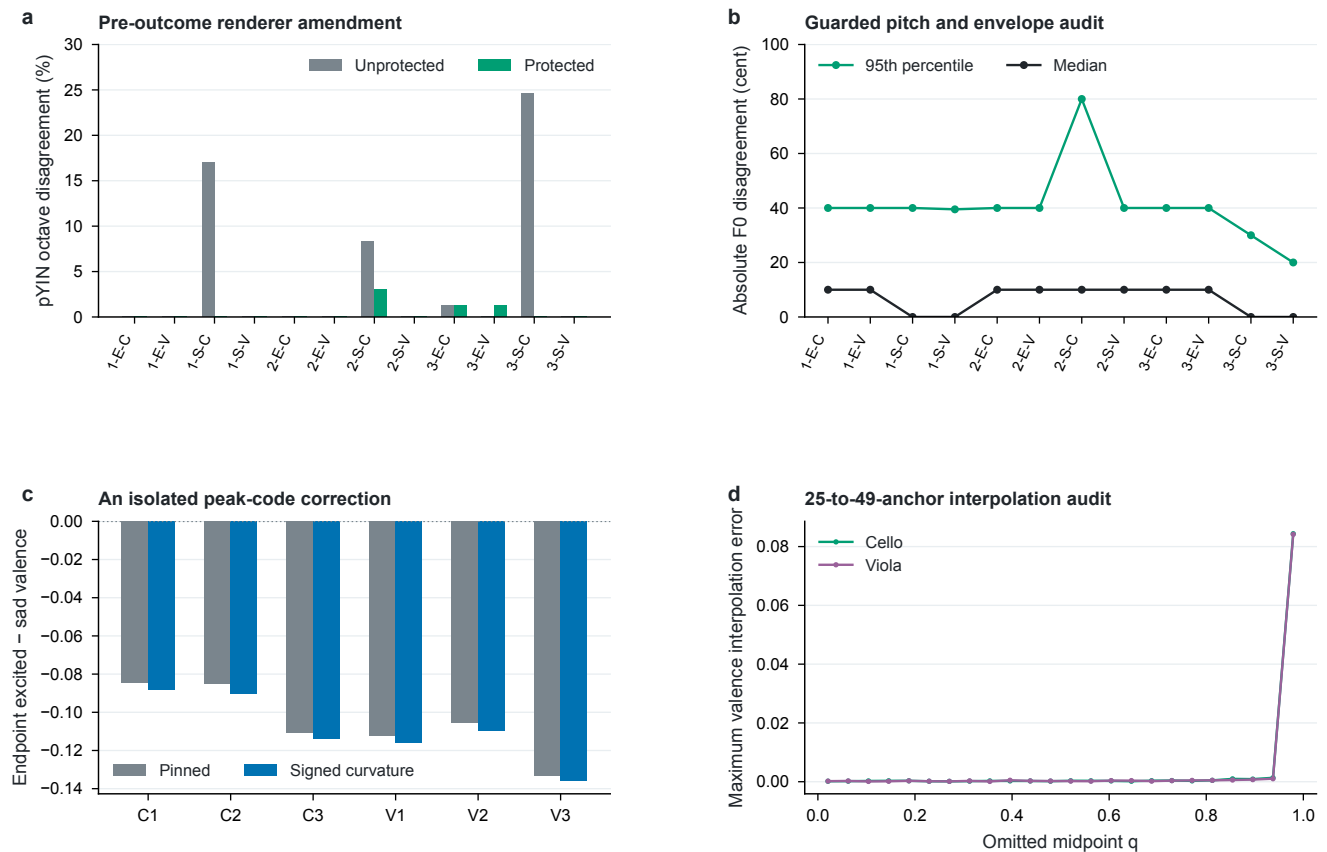
What changed before the primary transformed outcomes

The source-phase renderer was amended after pYIN exposed octave ambiguity in the unprotected pilot. The amendment was recorded before transformed MI outputs were inspected. All initial pilot audio and diagnostics remain in archive/unprotected-envelope/. The main grid uses the guarded renderer; 54 original pilot conditions were separately passed through MI. This sequence is not an independent preregistration, and the later near-endpoint refinement is explicitly post-primary. Neither change was selected to obtain a favourable excited–sad ordering.

What the instrument envelopes do not establish

The template bank contains measured cello and viola spectral envelopes from non-vibrato solo samples. Twenty-nine of 32 candidates passed the label-constrained screen. The cello bank brackets source F0; viola uses its lowest available envelope outside the sampled lower register. No transposition compensates for this gap. The first harmonics remain protected, consonantal/phonetic information can remain in voiced spectra, and removing residual energy changes amplitude around unvoiced portions. The result is a controlled partial envelope intervention, not a natural bowing performance or a complete phonetic-to-instrument conversion. Instrument similarity was not independently optimized or perceptually measured.

S16 | Acoustic and implementation controls



Supplementary Fig. 9 | Renderer, pitch, implementation and sampling controls. **a**, Endpoint octave-tracker disagreement before and after protecting lower harmonics. **b**, Median and 95th-percentile absolute F0 disagreement for guarded endpoints. Labels identify text, instruction and envelope. These are estimator diagnostics, not perceptual equivalence. **c**, Native endpoint excited-minus-sad valence with the pinned and isolated signed-curvature implementation; all remain negative. **d**, Error at omitted midpoints in a 25-anchor approximation to 49 evaluated points. All deviations remain in the result record.

The guarded 54-condition pilot yielded median absolute F0 disagreements of 0–10 cents, maximum 95th percentile 80 cents and maximum octave disagreement 3.0806%. The minimum 100-ms envelope correlation was approximately 0.849. Pitch is not assumed exactly preserved by a nominal source-phase constraint. Sample count and rate match the source, and identity conditions use the exact decoded samples.

At a spectral maximum, quadratic curvature is ordinarily negative. The pinned code applies a positive minimum clamp to that denominator. The isolated sensitivity copy preserves its sign outside a 10^{-8} magnitude floor. On six pure tones (83.3, 131.7, 197.3, 220, 333.3 and 440 Hz), absolute frequency errors fell from 5.85–10.05 Hz to 0.038–0.170 Hz. Across 72 full-file controls, valence RMS difference from the primary output was 0.00590; native endpoint contrasts remained negative. This identifies one numerical defect, not the entire cause of the response boundary. The primary engine was not modified.

S17 | All paired native-affect endpoints

Level	Envelope	Text	Source E – S	Endpoint E – S
native	cello	S01	-0.02668	-0.08436
native	cello	S02	-0.01842	-0.08523
native	cello	S03	-0.04720	-0.11077
native	viola	S01	-0.02668	-0.11220
native	viola	S02	-0.01842	-0.10546
native	viola	S03	-0.04720	-0.13308
matched	cello	S01	-0.02674	-0.07406
matched	cello	S02	-0.01315	-0.07176
matched	cello	S03	-0.04819	-0.10024
matched	viola	S01	-0.02674	-0.10017
matched	viola	S02	-0.01315	-0.09140
matched	viola	S03	-0.04819	-0.12088

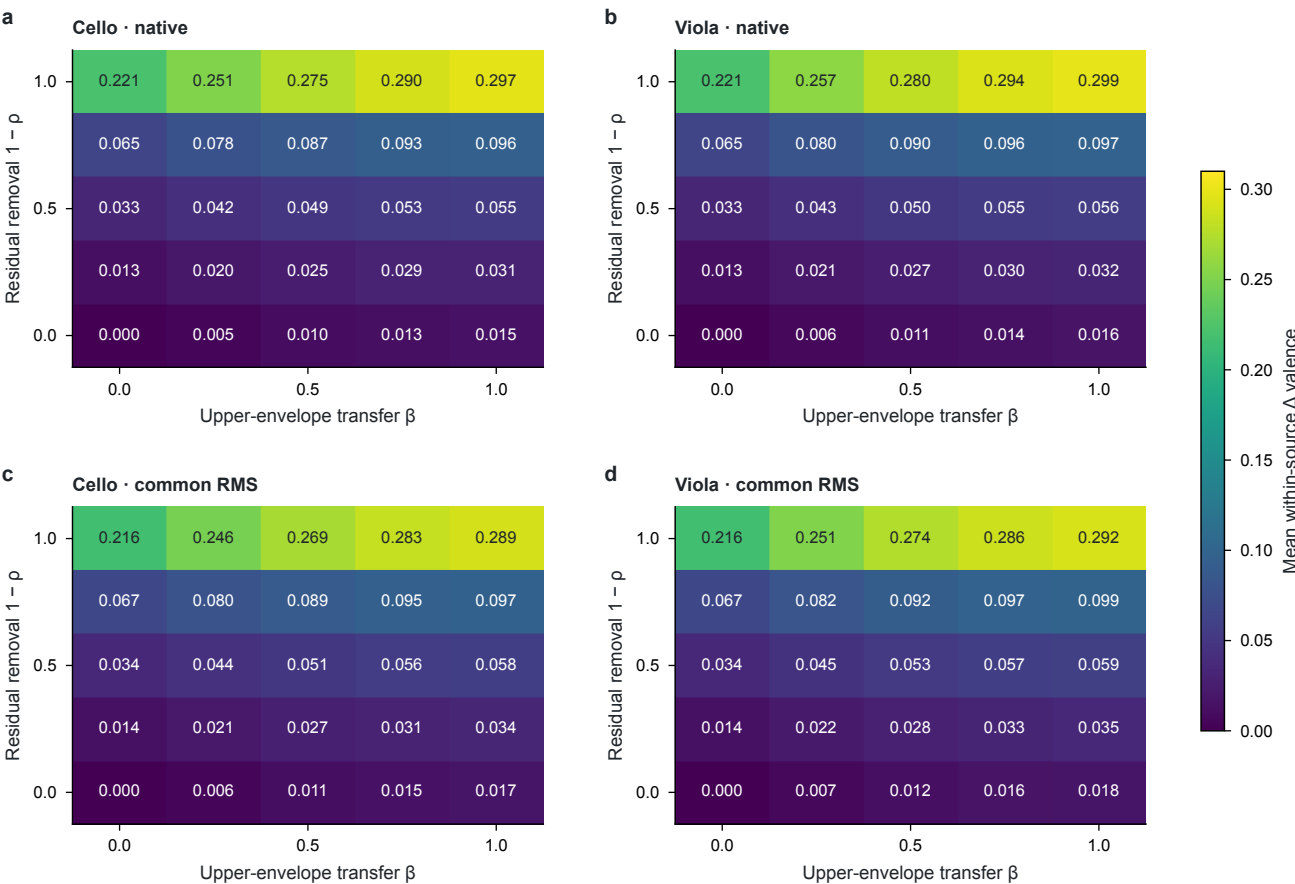
All values are full-recording means of VMM:C1:valence_state. Every primary paired-anchor valence contrast is negative (588/588). In the separate boundary refinement, 95/96 paired contrasts are negative; the sole positive is printed in S19. Native arousal contrasts are positive in 294/294 conditions; common-RMS arousal contrasts are positive in 94/294, with all six common-RMS endpoints negative. The RMS target is -37.38699170118202 dBFS, not Study 1's -29.517516 dBFS.

Factorial endpoint decomposition

Level	Envelope	Envelope only	Residual only	Joint	Interaction
native	cello	+0.01495	+0.22083	+0.29668	+0.06090
native	viola	+0.01591	+0.22083	+0.29879	+0.06205
matched	cello	+0.01699	+0.21629	+0.28950	+0.05622
matched	viola	+0.01781	+0.21629	+0.29159	+0.05749

Each mean gives equal weight to six source clips. Interaction = $V(1,0) - V(1,1) - V(0,0) + V(0,1)$. All source-specific values remain in data/morph/factorial-interactions.csv. These are finite differences of a deterministic response, not an ANOVA over independent speakers.

S17 | Complete factorial surfaces



Supplementary Fig. 7 | All four factorial response surfaces. Cells show the mean within-source native-valence change over six fixed source clips, separately for native/common-RMS level and cello/viola envelope. The shared scale permits comparison. Horizontal position varies upper-envelope transfer; vertical position varies residual removal. Every cell is evaluated with fresh state on the full file. Printed rounding does not replace the saved source values.

The largest isolated change comes from residual removal, while the joint intervention includes a positive interaction. This does not identify the residual with a phonetic class. The full-grid condition identifiers, gain, waveform hashes and MI receipts are retained; shared zero-transfer inputs are not duplicated as independent observations.

S18 | Feature interventions and weighted-term attribution

Level	Envelope	Source-held features	Mean valence	Δ from endpoint
native	cello	roughness	+0.29939	-0.13596
native	cello	pleasantness	+0.27890	-0.15645
native	cello	consonance block	+0.15421	-0.28114
native	viola	roughness	+0.29999	-0.13747
native	viola	pleasantness	+0.28196	-0.15550
native	viola	consonance block	+0.15585	-0.28161
matched	cello	roughness	+0.31347	-0.13790
matched	cello	pleasantness	+0.31409	-0.13728
matched	cello	consonance block	+0.18595	-0.26541
matched	viola	roughness	+0.31409	-0.13937
matched	viola	pleasantness	+0.31681	-0.13665
matched	viola	consonance block	+0.18730	-0.26616

The original source’s full feature trajectory replaces endpoint roughness (R^3 index 0), pleasantness (index 4) or the seven-feature consonance block (indices 0–6). All temporal requests and C^3 outputs are then recomputed, retaining fresh state. These are hybrid vectors and need not correspond to a realizable acoustic signal. No-change checks reproduce endpoint arrays exactly. Single-feature effects overlap and are not summed as independent fractions. Under the full block hold, inverse endpoint instruction ordering remains.

At native level, the full-block hold suppresses 94.8% of the mean cello endpoint rise and 94.2% of the viola rise. This percentage uses $\text{mean}(\text{endpoint} - \text{counterfactual}) / \text{mean}(\text{endpoint} - \text{source})$; it is not variance explained, a classification accuracy or a human mediation fraction. The original source’s own affect score is not substituted for the endpoint output.

Arithmetic closure is separately verified

All five valence weighted terms, all four arousal weighted terms and both clipping adjustments are saved frame by frame. Reconstruction error is zero. Roughness also enters upstream temporal and mechanism terms, so the direct $0.15 \times (1 - \text{roughness})$ contribution does not exhaust its computational influence. Paired term differences, every counterfactual full array and output hashes remain in the source archive; see `paired-component-decomposition.csv` and `counterfactual-summary.json`.

S19 | All near-endpoint probes and the positive exception

q	ρ	Min E – S	Max E – S	Max endpoint gap	Positive / 12
0.99	1.00e-02	-0.07618	-0.00014	0.174747	0
0.995	5.00e-03	-0.07686	+0.00063	0.164067	1
0.999	1.00e-03	-0.07629	-0.00151	0.138994	0
0.9999	1.00e-04	-0.08040	-0.01561	0.107309	0
0.99999	1.00e-05	-0.08892	-0.03392	0.075670	0
0.999999	1.00e-06	-0.10218	-0.06319	0.049631	0
0.9999999	1.00e-08	-0.13246	-0.07491	0.003208	0
0.999999999	1.00e-10	-0.13302	-0.07416	0.002869	0

Each row contains 24 full-file evaluations and 12 paired contrasts. The sole positive comparison is S02, cello envelope, common RMS, $q = 0.995$: +0.0006344318389892578. It becomes negative again at the zero-residual endpoint; there is no consistent recovery across texts or instruments. This exception is retained and the primary statement “588/588 negative” is restricted to the predeclared 49-anchor grid.

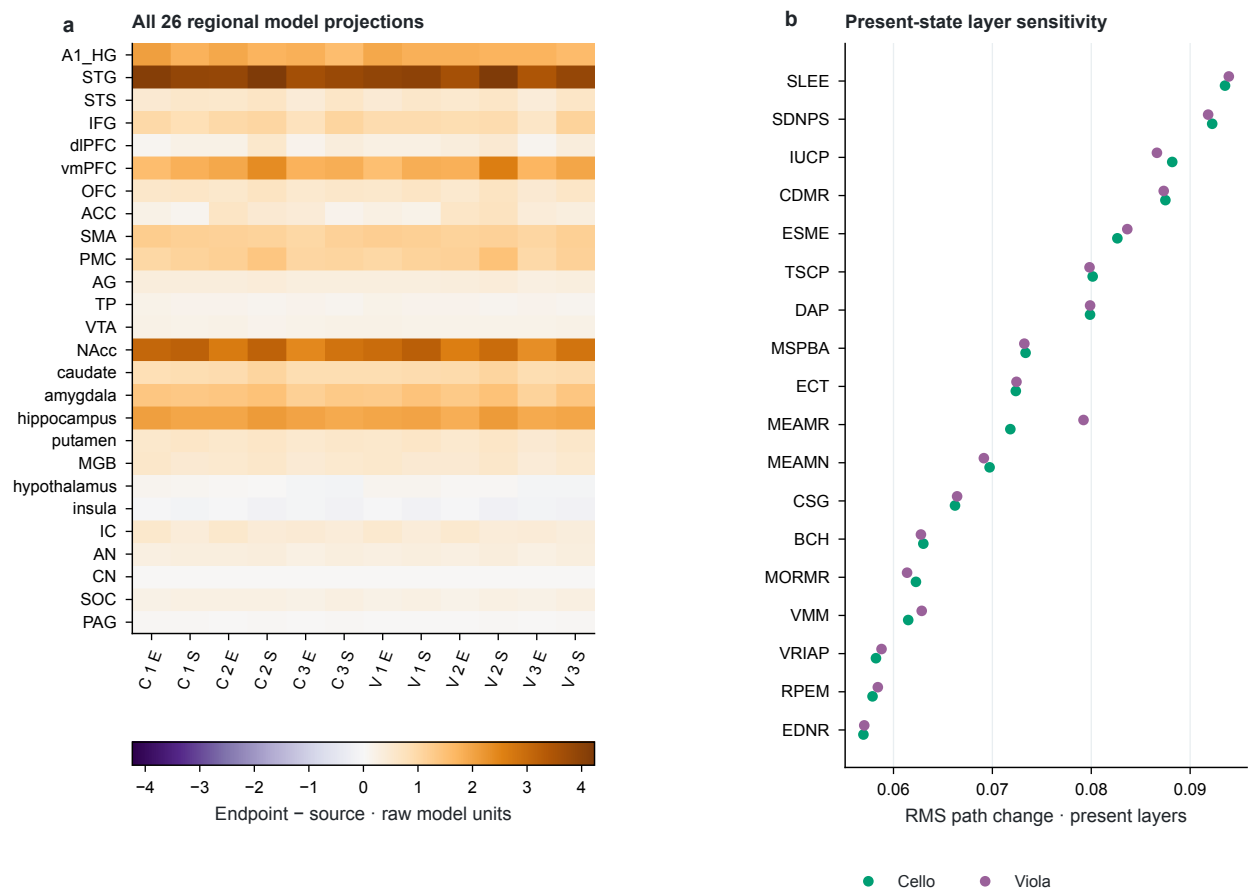
The endpoint gap is $|V(q) - V(1)|$ within source, envelope and level. At $q = 0.9999$, the waveform error $20 \log_{10}[\text{RMS}(y_q - y_1)/\text{RMS}(y_1)]$ ranges from -86.70 to -81.84 dB; the largest score gap is 0.1073093. At the smallest tested positive residual the maximum is 0.0028686. These are numerical sensitivity diagnostics and make no claim of inaudibility or listener thresholds.

Fixed-source strata

Source	Voiced	Unvoiced active	Quiet
emotion-E01-01-a	680	284	63
emotion-E13-01-a	611	413	84
emotion-S02-E01-01-a	496	314	50
emotion-S02-E13-01-a	428	454	64
emotion-S03-E01-01-a	688	284	55
emotion-S03-E13-01-a	628	367	45

All 1,050 outputs are summarized over these three source-defined masks at every primary condition. “Unvoiced active” exceeds -40 dB of the source peak frame RMS, and the remaining frames define quiet. The masks are unchanged under transformation. Inverse valence endpoint contrasts persist in source-voiced frames; outputs still reflect full-file temporal context. These masks are not word or phoneme alignments.

S20 | Regional projections and present-state layers



Supplementary Fig. 8 | All regions and the present-state subset. **a**, All 26 regional endpoint-minus-source projections for every native source/envelope condition. C/V denote cello/viola; E/S denote instructions. The colour scale has raw projection units and is unbounded by 0–1. **b**, Eighteen most sensitive native cello mechanisms in metadata-defined Present or Cognitive Present layers, with viola shown in identical order. The full subset contains 195 dimensions; all 889 remain in the primary analysis. Rankings are descriptive and selected for display, not independently confirmed.

Largest native cello regional path RMS values are STG 1.5701, NAcc 1.0261, A1/HG 0.9290, hippocampus 0.8687 and vmPFC 0.6698. These are projections of model activity, not measured brain activation. Native cello endpoint proxy means change by DA +0.123712, NE -0.049066, OPI -0.259996 and 5HT +0.092091. Higher valence need not increase every reward-labelled proxy.

S20 | All 84 mechanism sensitivities (1–28)

Rank / ID	Mechanism name	Cello RMS	Viola RMS	Present RMS C
1 / IUCP	Inverted-U Complexity Preference	0.0984	0.0965	0.0882
2 / SDNPS	Stimulus-Dependent Neural Pitch Saliency	0.0969	0.0964	0.0922
3 / ESME	Error-Signal Modulated Encoding	0.0897	0.0905	0.0826
4 / SLEE	Synaptic Long-term Encoding Engine	0.0889	0.0891	0.0935
5 / MORMR	mu-Opioid Receptor Music Reward	0.0793	0.0785	0.0623
6 / CDMR	Context-Dependent Mismatch Response	0.0773	0.0774	0.0875
7 / TSCP	Temporal Synaptic Consolidation Processor	0.0763	0.0763	0.0801
8 / DAP	Developmental Affective Plasticity	0.0756	0.0746	0.0799
9 / DMMS	Developmental Music Memory Scaffold	0.0749	0.0755	0.0357
10 / ECT	Expertise Compartmentalization Trade-off	0.0744	0.0745	0.0724
11 / MEAMR	Music-Evoked Autobiographical Memory Reward	0.0743	0.0814	0.0718
12 / SDED	Sensory Dissonance Early Detection	0.0741	0.0738	0.0548
13 / CSG	Consonance-Saliency Gradient	0.0722	0.0723	0.0662
14 / EDNR	Expertise-Dependent Network Reorganization	0.0679	0.0675	0.0570
15 / PNH	Pythagorean Neural Hierarchy	0.0678	0.0682	0.0518
16 / DAED	Dopamine Anticipation-Experience Dissociation	0.0666	0.0656	0.0552
17 / PMIM	Predictive Memory Integration Model	0.0652	0.0649	—
18 / CDEM	Context-Dependent Emotional Memory	0.0650	0.0674	0.0429
19 / MEAMN	Music-Evoked Autobiographical Memory Network	0.0645	0.0641	0.0697
20 / VMM	Valence-Mode Mapping	0.0643	0.0647	0.0615
21 / RPEM	Reward Prediction Error in Music	0.0640	0.0641	0.0579
22 / VRIAP	VR-Integrated Analgesia Paradigm	0.0625	0.0676	0.0582
23 / BCH	Brainstem Consonance Hierarchy	0.0604	0.0599	0.0630
24 / IACM	Inharmonicity-Attention Capture Model	0.0587	0.0585	0.0451
25 / LDAC	Liking-Dependent Auditory Cortex	0.0581	0.0572	0.0426
26 / SSPS	Saddle-Shaped Preference Surface	0.0550	0.0550	0.0331
27 / IOTMS	Individual Opioid Tone Music Sensitivity	0.0529	0.0558	0.0444
28 / TPIO	Timbre Perception-Imagery Overlap	0.0525	0.0519	0.0284

RMS = $\sqrt{\text{mean over sources, anchors and mechanism dimensions of squared recording-mean change from the source}}$. Cello determines rank; viola is shown without reordering. “Present” uses declared layer names. All four instrument/level ranks, frame-resolved RMS, all dimensions, signed endpoints and quantiles remain in the numerical archive. Raw output scales are not a common biological effect size.

S20 | All 84 mechanism sensitivities (29–56)

Rank / ID	Mechanism name	Cello RMS	Viola RMS	Present RMS C
29 / MIAA	Musical Imagery Auditory Activation	0.0501	0.0489	0.0392
30 / MSPBA	Musical Syntax Processing in Broca's Area	0.0501	0.0499	0.0734
31 / SPH	Spatiotemporal Prediction Hierarchy	0.0498	0.0488	0.0385
32 / RIRI	RAS-Intelligent Rehabilitation Integration	0.0472	0.0468	0.0520
33 / PSCL	Pitch Salience in Cortical Lateralization	0.0472	0.0467	0.0462
34 / CMAT	Cross-Modal Affective Transfer	0.0470	0.0472	0.0226
35 / STAI	Spectral-Temporal Aesthetic Integration	0.0462	0.0461	0.0313
36 / NSCP	Neural Synchrony Commercial Prediction	0.0429	0.0418	0.0265
37 / CHPI	Cross-Modal Harmonic Predictive Integration	0.0418	0.0417	0.0210
38 / PEOM	Period Entrainment Optimization Model	0.0417	0.0415	0.0417
39 / PWUP	Precision-Weighted Uncertainty Processing	0.0415	0.0408	0.0271
40 / ICEM	Information Content Emotion Model	0.0414	0.0404	0.0416
41 / HTP	Hierarchical Temporal Prediction	0.0406	0.0437	0.0274
42 / SSRI	Social Synchrony Reward Integration	0.0406	0.0420	0.0194
43 / MAA	Multifactorial Atonal Appreciation	0.0401	0.0402	0.0392
44 / IGFE	Individual Gamma Frequency Enhancement	0.0394	0.0391	0.0169
45 / SRP	Striatal Reward Prediction	0.0394	0.0395	0.0448
46 / PSH	Prediction Silencing Hypothesis	0.0389	0.0380	0.0238
47 / AACM	Aesthetic-Attention Coupling Model	0.0385	0.0379	0.0247
48 / UDP	Uncertainty-Driven Pleasure	0.0382	0.0377	0.0157
49 / NEMAC	Nostalgia-Evoked Memory-Affect Circuit	0.0353	0.0348	0.0208
50 / CMAPCC	Cross-Modal Action-Perception Common Code	0.0334	0.0329	0.0380
51 / TAR	Therapeutic Affective Resonance	0.0328	0.0333	0.0335
52 / HMCE	Hierarchical Musical Context Encoding	0.0326	0.0319	0.0271
53 / SPMC	SMA-Premotor-M1 Motor Circuit	0.0324	0.0320	0.0358
54 / MAD	Musical Anhedonia Disconnection	0.0323	0.0319	0.0124
55 / HCMC	Hippocampal-Cortical Memory Circuit	0.0316	0.0312	0.0345
56 / RASN	Rhythmic Auditory Stimulation Neuroplasticity	0.0309	0.0312	0.0054

RMS = $\sqrt{\text{mean over sources, anchors and mechanism dimensions of squared recording-mean change from the source}}$). Cello determines rank; viola is shown without reordering. “Present” uses declared layer names. All four instrument/level ranks, frame-resolved RMS, all dimensions, signed endpoints and quantiles remain in the numerical archive. Raw output scales are not a common biological effect size.

S20 | All 84 mechanism sensitivities (57–84)

Rank / ID	Mechanism name	Cello RMS	Viola RMS	Present RMS C
57 / TPRD	Tonotopy-Pitch Representation Dissociation	0.0300	0.0300	0.0241
58 / MCCN	Musical Chills Cortical Network	0.0289	0.0292	0.0195
59 / STC	Singing Training Connectivity	0.0286	0.0289	0.0225
60 / CLAM	Closed-Loop Affective Modulation	0.0277	0.0277	0.0407
61 / VRMSME	VR Music Stimulation Motor Enhancement	0.0275	0.0269	0.0257
62 / CSSL	Cross-Species Song Learning	0.0272	0.0270	0.0217
63 / MMP	Musical Mnemonic Preservation	0.0260	0.0262	—
64 / HGSIC	Hierarchical Groove State Integration Circuit	0.0256	0.0252	0.0292
65 / PCCR	Pitch Chroma Cortical Representation	0.0240	0.0237	0.0302
66 / DDSMI	Dyadic Dance Social Motor Integration	0.0238	0.0232	0.0191
67 / SNEM	Sensory Novelty and Expectation Model	0.0226	0.0223	0.0267
68 / GSSM	Gait-Synchronized Stimulation Model	0.0220	0.0217	0.0215
69 / PWSM	Precision-Weighted Salience Model	0.0218	0.0213	0.0214
70 / CTBB	Cerebellar Theta-Burst Balance	0.0212	0.0212	0.0123
71 / ASAP	Action Simulation for Auditory Prediction	0.0211	0.0205	0.0197
72 / AMSS	Attention-Modulated Stream Segregation	0.0208	0.0208	0.0109
73 / ETAM	Entrainment, Tempo and Attention Modulation	0.0203	0.0201	0.0150
74 / MSR	Musician Sensorimotor Reorganization	0.0202	0.0199	0.0256
75 / WMED	Working Memory-Entrainment Dissociation	0.0185	0.0180	0.0170
76 / MPG	Melodic Processing Gradient	0.0184	0.0188	0.0151
77 / DGTP	Domain-General Temporal Processing	0.0181	0.0182	0.0123
78 / PUPF	Pleasure-Uncertainty Prediction Function	0.0174	0.0173	0.0086
79 / BARM	Brainstem Auditory Response Modulation	0.0155	0.0151	0.0066
80 / AAC	Autonomic-Arousal Circuit	0.0154	0.0155	0.0051
81 / SDL	Salience-Dependent Lateralization	0.0146	0.0146	0.0095
82 / NEWMD	Neural Entrainment-Working Memory Dissociation	0.0121	0.0121	0.0195
83 / STANM	Spectrotemporal Attention Network Model	0.0084	0.0083	0.0081
84 / OII	Oscillatory Intelligence Integration	0.0067	0.0067	0.0092

RMS = sqrt(mean over sources, anchors and mechanism dimensions of squared recording-mean change from the source)). Cello determines rank; viola is shown without reordering. “Present” uses declared layer names. All four instrument/level ranks, frame-resolved RMS, all dimensions, signed endpoints and quantiles remain in the numerical archive. Raw output scales are not a common biological effect size.

S20 | Archive map, validation and preserved evidence

Path under morph-experiment/	Retained content
sources.json; manifest.json	All six sources, 1,596 primary coordinates, waveform IDs and hashes
audio/; prepared/; mi/	Full WAVs, source F0/masks/phase/envelopes, every MI frame and receipt
analysis/	All 1,050-readout means, five summaries, paths, ranks, strata, contrasts and controls
counterfactuals/	72 full feature-intervention arrays and 24 exact no-change checks
endpoint-refinement/	192 full-file boundary evaluations with exact q and waveform hashes
secondary-unprotected/	54 sensitivity outputs; initial pilot audio/diagnostics retained in archive/
sensitivity-peak-curvature/	72 isolated one-operation controls; separate engine pin and diff
provenance/mi-engine/	855 pinned primary sources and original adapter
protocol/; qc/	Source lock, pre-outcome amendment, gain target, pitch checks and verification
data/morph/ in manuscript package	Figure-specific values, complete exported rankings/summaries and source hashes

All source waveforms and MI output hashes were verified. Original-source arrays reproduce Study 1 exactly across 89 arrays for each of six files; repeated inference and weighted-term closure have zero discrepancy. The primary adapter SHA-256 is 67ba775b5d889ea6bb3b599e78c0c5c1936e621e15061e189da78b23a4badecb. The execution runtime uses Python 3.11.8, torch/torchaudio 2.2.0, librosa 0.10.1, NumPy 1.26.4 and SciPy 1.11.4. The manuscript renders existing arrays and does not rerun or alter either experiment.

The source archive is on SRC-9, accessed through an explicit symbolic link. The portable aggregate-analysis bundle omits full raw MI/audio arrays and the original Philharmonia sample bank; these remain locally archived. The instrument samples must not be redistributed as-is. All six new figures have editable vector PDF/SVG, 450-dpi PNG, numerical source JSON and a SHA-256 input map. The earlier six main figures and three supplementary figures remain byte-identical assets.