

Expressive information across complementary processing layers in speech and song

Amaç Erdem

Musical Intelligence Research

Abstract

Musical Intelligence (MI) represents expressive speech and song through differentiated mechanisms whose processing layers make measurable, complementary contributions. We mapped this organization in a fixed auditory architecture using 342 original ElevenLabs recordings from 172 generation requests. A matched core crossed three texts, speech or song, and five production tags. The analysis covered all 84 mechanisms, 326 declared layers and 889 channels; MI received waveforms and was not fitted to the tags. A separate fixed ridge readout recovered five speech tags from IACM at 82.2% agreement, compared with 68.9% from the complete C^3 vector. Three-tag agreement reached 27/27 in speech with RASN and 26/27 in song with CSSL, across held-out texts. Layer combinations explained these results. In CSSL, removing E, M, P or F reduced 26/27 to 19/27, 22/27, 24/27 or 20/27; the complete representation exceeded every proper subset. In IACM, adding P increased agreement from 29/45 to 37/45 although P alone reached only 18/45. A nearest-centroid control recovered five song tags from a single SDED channel at 41/45, exposing information missed by the original readout. Candidate-wide permutation tests supported tag-related structure across all four core tasks. These findings establish a detailed computational account of where expressive production information is accessible and how processing-layer outputs complement one another. The resulting mechanism and layer predictions provide explicit targets for independent acoustic, behavioral and neuroscientific tests.

Introduction

Expressive speech and song combine pitch, timing, spectral structure and amplitude into recognizable vocal patterns. A computational account should explain how these patterns are represented across processing stages, and which stages contribute information that other stages do not supply. Matched speech and song corpora make this question experimentally tractable by varying vocal expression while preserving intended verbal material. RAVDESS provides one established example with human performances and perceptual ratings.¹ Generative audio offers a complementary route: explicit production instructions define an inspectable reference for examining the responses of a computational architecture.

MI is an executable auditory architecture with separately accessible mechanism outputs. Its C^3 response comprises 84 mechanisms and 889 channels, organized into 326 declared processing layers. This structure makes it possible to examine the model at several resolutions: the complete response, individual mechanisms, their layers and individual channels. The scientific object is this organization. A single score obtained by reading the complete vector cannot reveal which computations preserve

a distinction or whether several output layers contribute complementary information.

We therefore conducted a comprehensive representational study of the fixed engine. Every declared mechanism entered the original analysis; the present decomposition examines every layer and channel. Candidate membership follows the engine registry and metadata, rather than an expectation that a mechanism's name should match speech, song or emotion. Detailed figures identify structures discovered in that complete inventory. Simple external readouts quantify how consistently production tags can be recovered when the sentence being evaluated is absent from readout estimation. They measure information accessible in the stored response and do not alter MI.

The principal finding is selective complementarity. Individual mechanisms preserve strong expressive distinctions in both speech and song, and particular layer combinations increase their accessibility. CSSL requires its complete set of output layers for its highest observed three-tag song result, whereas IACM's five-tag speech result is retained by E+M+P. These differences specify where the architecture adds useful representational structure and where a smaller output set is sufficient.

A fixed engine, an exhaustive response inventory



Full archive: 342 original recordings / 172 requests. Alternatives remain within their generation request.

Fig. 1 | Production conditions and comprehensive measurement of a fixed auditory architecture. **a**, The matched core contains 90 generation requests and 179 recordings: three texts × two delivery modes × five tags × three requests. The three-tag subset contains excited, sad and untagged; the five-tag task adds mad and peaceful. **b**, MI processes waveforms without production labels. All 84 mechanisms, 326 layers and 889 channels enter the inventory. **c**, External readouts are estimated from two texts and evaluated on the third. Their parameters are separate from the fixed MI engine. Alternatives are combined within their request before evaluation.

The full corpus contains 342 recordings from 172 requests to ElevenLabs, using the displayed Eleven v4 model and Anna—Young Female Conversational Voice. The matched core crosses three English texts, requested speech or unaccompanied song, and five production tags: excited, sad, mad, peaceful and untagged. Each cell contains three requests. One request has only one returned alternative, giving 179 core recordings. The main three-tag comparison uses excited, sad and untagged, with 27 requests per mode; the five-tag comparison uses 45 per mode (Fig. 1).

Tags are archived generation instructions. They specify intended performance through the provider's audio-tag interface,² and are not an independent classification of the returned waveform. Untagged denotes the absence of an explicit emotion instruction. All original recordings remain eligible, including unsuccessful or ambiguous realizations of a requested expression or singing style. This preserves the production reference without making MI agreement a criterion for accepting a stimulus.

The architecture was fixed throughout this study. The extraction adapter instantiates the registered mechanisms, processes each waveform under *torch.no_grad()*, and saves all outputs. It accepts neither tag labels nor a tag-dependent optimization objective. The canonical registry resolves

duplicate mechanism names by a documented first-function-wins rule. Its resulting 84 mechanisms and 889 channels were already included in the original whole-mechanism analysis. The new layer and channel analyses expand resolution within that same fixed inventory.

Analysis retains the generation request as its unit. Each channel is summarized within a recording, and available alternatives are averaged within request. The complete-C³ comparator concatenates 889 channels; it is not an arithmetic average over mechanisms. No signed mean across unlike mechanisms or layers serves as a primary outcome. This distinction matters because the present question concerns information available at identifiable locations in the engine, rather than the behavior of an aggregate response score.

Input verification reproduced the hashes of all 342 recordings, reconstructed 1,590 archived prediction matrices and checked 43 task splits. Recomputing the four core tasks for all 84 mechanisms reproduced archived readout scores to a maximum absolute difference of 1.14×10^{-7} . These checks establish continuity between the original engine response and the mechanism-level findings reported here.

Strong expressive distinctions persist across texts

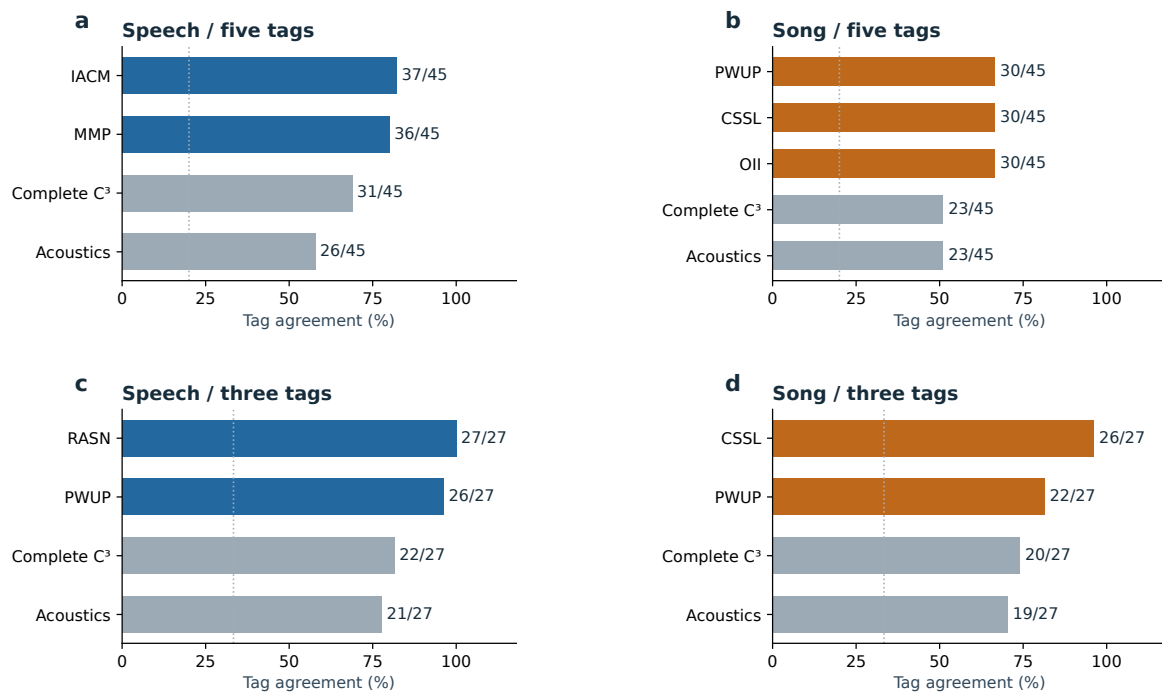


Fig. 2 | Mechanism-specific information in speech and song. All panels use the original fixed ridge readout, the same request units and sentence-held-out evaluation. Bars show agreement with production tags; labels give correct requests / total requests. Three-tag and five-tag tasks are distinct comparisons. Mechanisms shown are leading results from the complete 84-mechanism inventory, supplied in Supplementary Fig. S1. Gray bars give the complete 889-channel C³ response and the 15-feature acoustic comparison. Dotted lines indicate nominal class-balanced chance.

Individual mechanisms recover production-tag distinctions that are less accessible to a readout of the complete C³ vector. In five-tag speech, IACM reaches 37/45 (82.2%) and MMP reaches 36/45 (80.0%), compared with 31/45 (68.9%) for complete C³ and 26/45 (57.8%) for the acoustic comparison. IACM is consistent across the three evaluated texts, with 13/15, 12/15 and 12/15 correct requests. These results identify a concentrated representation of speech-expression differences within a broader auditory response (Fig. 2).

The three-tag task exposes particularly clear distinctions. RASN separates all 27 speech requests, with 9/9 for each held-out text; PWUP reaches 26/27. CSSL reaches 26/27 in song, with 9/9, 9/9 and 8/9 across texts. The corresponding complete-C³ results are 22/27 for speech and 20/27 for song. The 100% and 96.3% values describe these three-tag comparisons. They retain their own class sets and denominators and do not replace the more demanding five-tag results.

Five-tag song has a structured error pattern. PWUP, CSSL and OII each reach 30/45 (66.7%), compared with 23/45

(51.1%) for complete C³ and the acoustic comparison. PWUP separates excited, mad and sad at 9/9 each, while peaceful reaches 2/9 and untagged 1/9. The complete-C³ readout recovers only 2/9 excited requests. Thus, difficulty reading excited song from the complete vector does not imply that the architecture's strongest mechanism separates only two excited examples. Class-specific matrices preserve this distinction (Supplementary Fig. S2).

The broad inventory also identifies tasks for which combining mechanisms is advantageous. For 21 tags within the same S01 text, complete C³ reaches 54/63 (85.7%), while the strongest individual mechanisms reach 36/63. Seven delivery instructions are separated at 19/21 by complete C³. These same-text panels address different questions from the main cross-text comparisons, but they show that the benefit of a compact or distributed representation depends on the distinction being measured. The architecture cannot be characterized adequately by either its best single mechanism or its complete-vector score alone.

E, M, P and F provide complementary output information

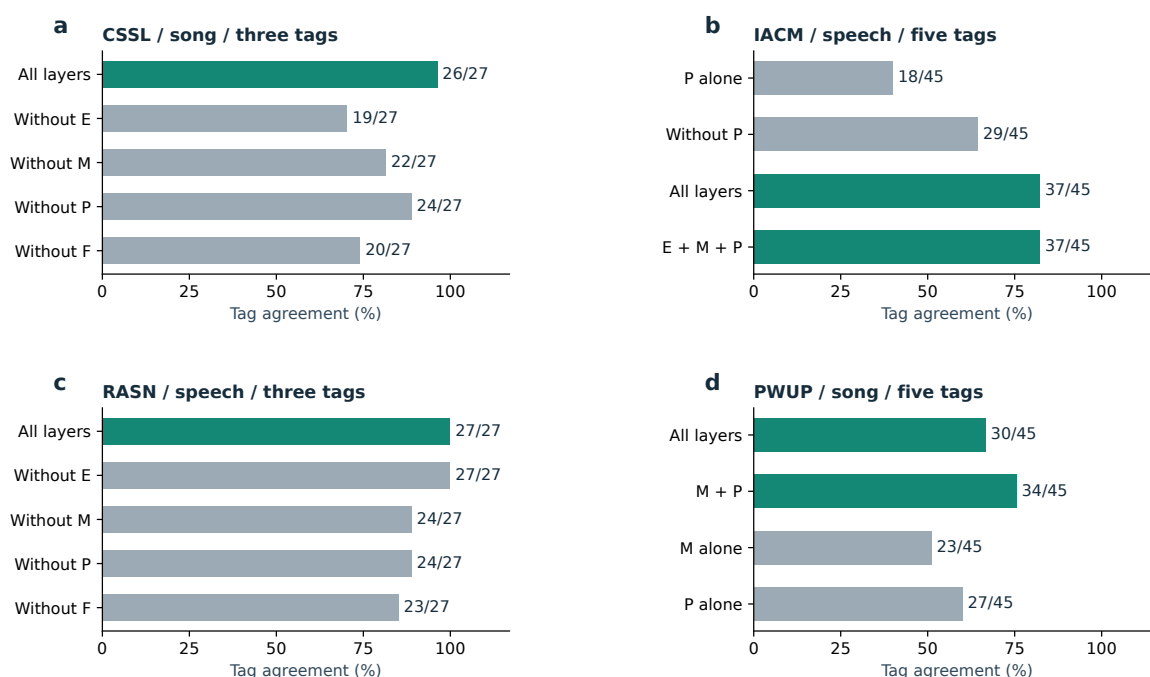


Fig. 3 | Layer combinations add measurable discrimination. **a**, CSSL three-tag song: removing any one output layer lowers the complete 26/27 result. **b**, IACM five-tag speech: P alone reaches 18/45, but adding it to E+M+F increases agreement from 29/45 to 37/45; E+M+P also reaches 37/45. **c**, RASN three-tag speech: M+P+F retains 27/27. **d**, PWUP five-tag song: M+P reaches 34/45, compared with 30/45 for all layers. Counts use the same fixed ridge procedure. These are exclusions of stored outputs followed by readout refitting; they do not disable upstream computations inside MI. All subsets and scaling controls are supplied in the supplement.

Layer-level decomposition explains why a mechanism's joint response can be more informative than its components considered separately. In CSSL's three-tag song task, E, M, P and F individually reach 20/27, 15/27, 16/27 and 14/27. Their joint response reaches 26/27. Removing E reduces this to 19/27; removing M, P or F yields 22/27, 24/27 or 20/27. Across all 16 possible subsets, the complete response uniquely gives the highest observed agreement under the original readout. Each output layer therefore contributes to the best joint CSSL result in this task (Fig. 3a).

IACM demonstrates that standalone accuracy can substantially underestimate a layer's contribution. Its P layer alone reaches 18/45 in five-tag speech. Adding P to the remaining layers raises agreement from 29/45 to 37/45, an increase of eight correct requests or 17.8 percentage points. E+M+P also reaches 37/45, indicating that F supplies no additional correct predictions for this particular readout and task. The contribution is selective: P carries information that becomes useful in combination even though it is a weak standalone discriminator (Fig. 3b).

RASN shows another pattern. Its complete response separates three speech tags at 27/27, and M+P+F retains that result. Removing M or P from the complete response reduces agreement to 24/27; removing F reduces it to 23/27. In five-tag song, PWUP's M+P subset reaches 34/45 (75.6%), above the complete mechanism's 30/45. These contrasts make layer complementarity testable at a specific mechanism and task, rather than treating additional layers as uniformly beneficial (Fig. 3c,d).

The analysis covers 1,304 single-layer exclusions across the complete inventory, plus 1,152 subset/readout combinations for six detailed mechanisms. A control that preserves the complete mechanism's scaling denominator retains the key losses: IACM without P reaches 26/45 instead of 37/45, and CSSL without E reaches 20/27 instead of 26/27. The contribution of these outputs therefore survives this normalization control. Some effects change with the external reader, as expected when usefulness depends on the geometry available to that reader. Supplementary Fig. S4 reports both scaling and reader controls without collapsing them into a common effect.

Individual channels expose fine expressive structure

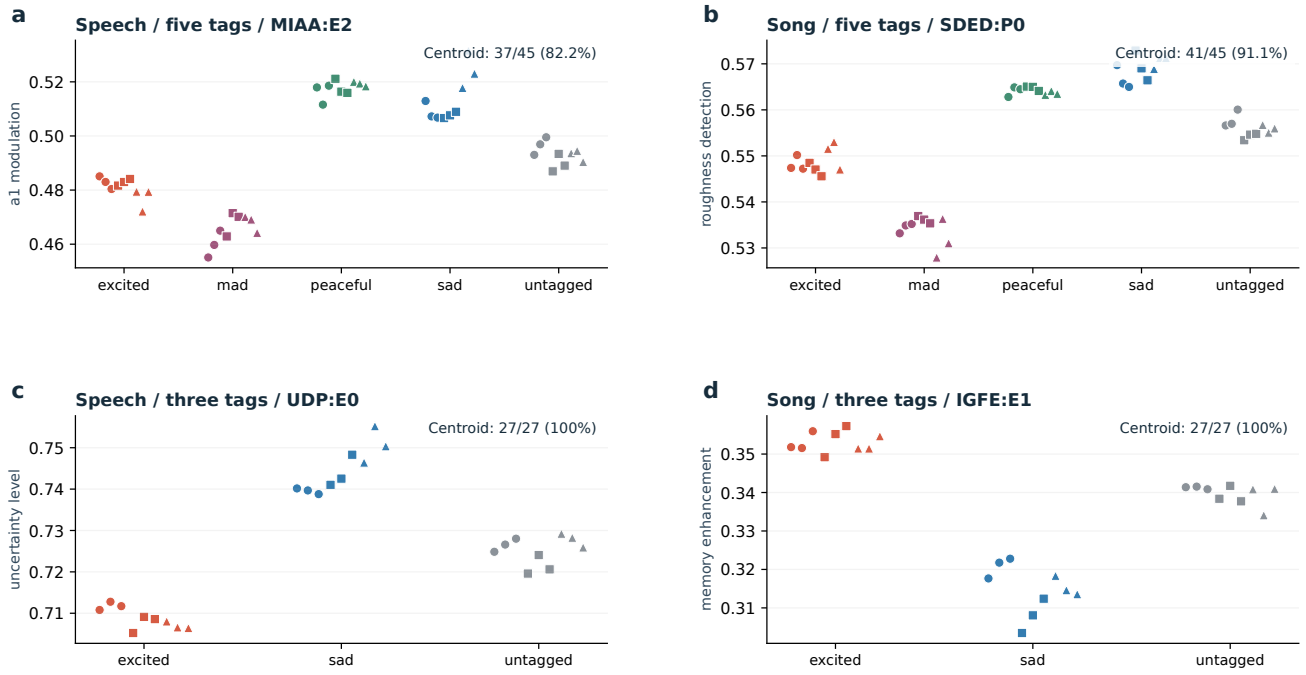


Fig. 4 | Request-level channel values and an alternative measurement of separability. Points show original channel values for each request; circles, squares and triangles identify S01, S02 and S03. Insets report sentence-held-out nearest-centroid agreement using only the displayed channel. **a**, MIAA:E2, five-tag speech, 37/45. **b**, SDED:P0, five-tag song, 41/45. **c**, UDP:E0, three-tag speech, 27/27. **d**, IGFE:E1, three-tag song, 27/27. These examples follow the complete channel audit. Native units are retained, and each channel has its own scale.

The channel inventory reveals that some expressive distinctions are accessible within a single continuous output. This result required a readout suited to the geometry of scalar data. For balanced multiclass tasks, the original centered linear-ridge rule can restrict a single channel to the classes with extreme fitted slopes. Intermediate classes may therefore be inaccessible to that rule even when the observed values form distinguishable ranges. We added nearest-centroid measurement to all 84 mechanisms, 326 layers and 889 channels, alongside the same comparators and held-out texts.

This alternative exposes strong five-tag song structure. SDED:P0:roughness_detection reaches 41/45 (91.1%), and IGFE:E1:memory_enhancement reaches 40/45 (88.9%). SDED:P0 recovers all nine mad and all nine peaceful requests, eight excited, seven sad and eight untagged requests. In five-tag speech, MIAA:E2:a1_modulation reaches 37/45 (82.2%). For three tags, UDP:E0:uncertainty_level and IGFE:E0:igf_match each reach 27/27 in speech, while IGFE:E1 reaches 27/27 in song. Figure 4 displays the underlying request values so that these results can be read as structure in model outputs as well as classification counts.

The second reader changes the interpretation of low scalar ridge accuracy: a failure of one measurement rule does not establish the absence of information in the channel. The analysis retains both readers and all their outputs. The 91.1% song result is an additional, explicitly identified measurement of five-tag separability; it is not substituted for the original whole-mechanism ridge result. Together, the two analyses locate both distributed complementarity and compact expressive signatures in the same fixed engine.

The full search was included in restricted permutation controls. In each of 999 permutations, the label mapping was recomputed across all candidates. The maximum observed agreement in each core task exceeded the corresponding permuted maxima for the 84-mechanism ridge family and the 1,299-candidate nearest-centroid family ($P = 0.001$ in each task; within-reader four-task BY $q = 0.00208$). These tests support tag-related structure after accounting for the stated candidate searches. They test information relative to a restricted label null, rather than the superiority of one channel over another or the significance of every individual layer-removal difference.

Temporal profiles and task structure locate the contribution

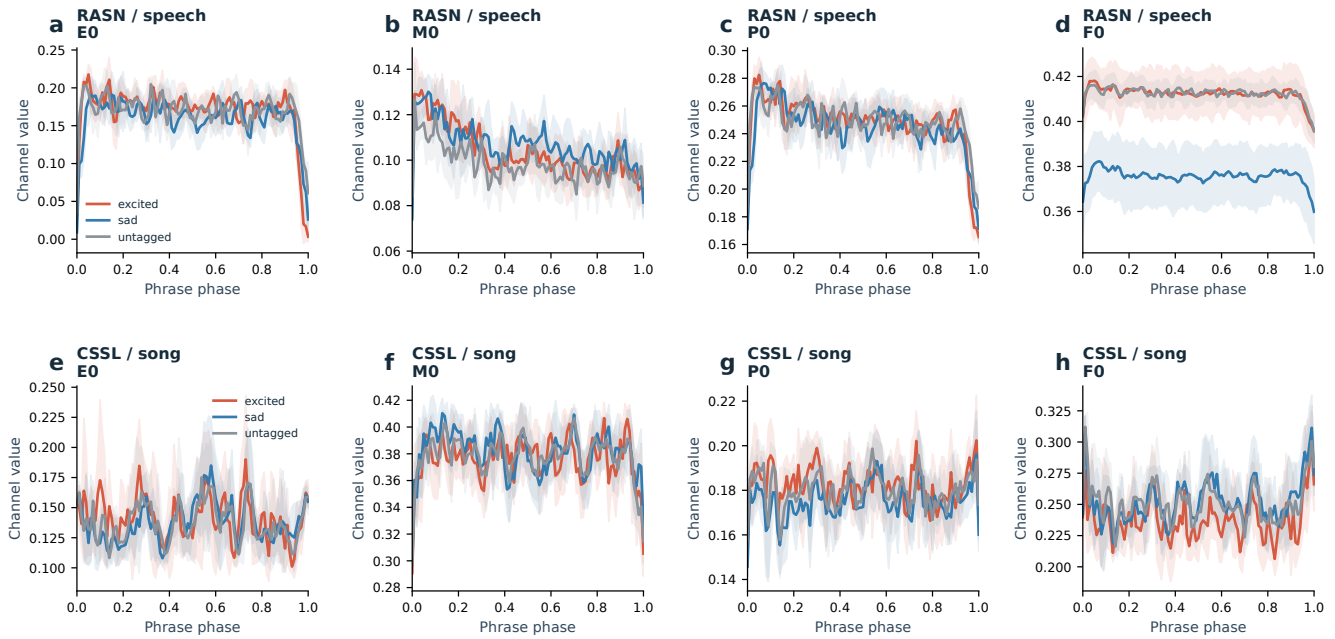


Fig. 5 | Processing-layer outputs retain distinct temporal profiles. Representative E0, M0, P0 and F0 channels are shown for RASN in speech and CSSL in song. Lines show the mean of nine requests per tag and shading one request standard deviation; the three tags are excited, sad and untagged. Time is normalized within each recording before alternatives are averaged. Separate channel axes preserve their native units. Normalized phase does not align words, and these offline model traces are not measurements of neural latency.

Channel means provide a compact description, but the underlying responses evolve throughout each recording. Separate E, M, P and F trajectories expose how tag-dependent differences are expressed within a mechanism over time. Figure 5 shows the E0/M0/P0/F0 channels of RASN and CSSL using every eligible request in the three-tag comparisons. The resulting profiles make the processing layers inspectable in their own units. Differences in raw amplitude, baseline and trajectory should therefore be examined within a channel rather than averaged across unrelated outputs.

Temporal summaries also change which distinctions are accessible to a fixed reader. RASN's five-tag speech agreement rises from 28/45 with channel means to 36/45 with ten consecutive normalized-time blocks. CSSL's five-tag speech agreement rises from 30/45 with means to 37/45 with seven temporal descriptors. The effect is not uniformly positive: CSSL's three-tag song result decreases from 26/27 with means to 18/27 with ten time blocks. Temporal detail can expose useful response structure while also increasing the dimensional demands on a readout estimated from a small number of requests.

Transfer between requested speech and song is more limited than separation within a mode. With the evaluation text excluded from readout estimation, CSG and STANM reach 26/45 for speech-to-song transfer, and HTP reaches 26/45 in the reverse direction. In the broader six-tag speech panel, IACM reaches 40/54 and MMP 39/54. Separate binary comparisons identify excited versus sad at 18/18 in 17 speech mechanisms and 18 song mechanisms. Speech peaceful versus sad is harder, with the strongest mechanism reaching 14/18. These observations locate the remaining difficulty in particular mode and tag relations rather than assigning the architecture one overall performance level.

The analysis consequently yields a concrete set of representational findings: cross-text expressive distinctions in individual mechanisms; complementary contributions from particular processing layers; compact tag-related signatures in individual channels; and sensitivity to temporal summary and delivery mode. Each finding refers to a named output, an explicit comparison and an archived numerical result. That specificity turns a broad architecture into a set of testable computational predictions.

Discussion

Complementarity is a measurable property of the representation

The central result is that the fixed MI architecture organizes expressive production information across outputs whose contributions can be measured separately and in combination. CSSL's complete E+M+P+F response exceeds every proper subset in the three-tag song comparison. IACM's P outputs add eight correct five-tag speech identifications despite weak standalone agreement. These results demonstrate useful complementarity at the level of accessible model output. They explain why reading each layer separately, or reading the entire architecture through a single classifier, can miss the contribution of a specific combination.

A comprehensive examination of an existing engine

The study examines an already implemented architecture. All registered mechanisms, all declared layers and all channels were eligible because they exist in the engine. Their inclusion did not depend on a favorable name or a preliminary score. Identifying informative channels after this complete examination is the discovery process itself. The resulting inventory can be evaluated for completeness, reproducibility, label-null behavior and stability across the available texts. A different objective, choosing one channel automatically before seeing the evaluation text, requires a selection algorithm. We report that additional experiment in Supplementary Note S3; its performance is not substituted for the response map being investigated here.

The measurement rule determines what can be read

A readout is a measurement of accessibility, not an exhaustive definition of the representation. This distinction is well established in work on probing internal representations.³ Here it has a concrete mathematical consequence: a centered linear rule can fail to recover several ordered classes from one scalar even when those classes occupy separable regions. The nearest-centroid control reveals such information without modifying MI. Conversely, strong agreement in a single channel shows that the full hierarchy is not required for every measured distinction. The informative result is the observed allocation of compact and complementary information, together with the conditions under which each becomes accessible.

Speech and song expose different uses of the same architecture

The strongest speech mechanisms and the strongest song mechanisms do not coincide across all tasks. RASN and CSSL are especially effective in the three-tag speech and song comparisons, respectively; IACM leads five-tag speech, while a single SDED channel gives the strongest five-tag song separation under the centroid reader. These names identify computational implementations. They do not, by themselves, demonstrate neuroplasticity, cross-species learning, attention capture or a neural dissonance response. Their value here is to locate a reproducible pattern in a specified computation that can subsequently be tested against controlled acoustics and human observations.

Scope and inferential limits

The findings concern one synthetic voice, three fixed texts and a staged generation history. Production tag, realized vocal expression and perceived emotion remain distinct variables. Cohort differences constrain causal attribution, and the study does not estimate performance over a population of speakers or sentences. Candidate-wide permutation tests account for the reported search families under their restricted exchangeability assumptions. They do not retroactively preregister the study, adjust every historical analytic choice, or establish a population-level advantage for each highlighted mechanism. The layer comparisons are observed changes in held-out agreement; no unreported significance test is implied for their differences.

Mechanism-level predictions for the next experiments

The response map supplies specific follow-up predictions. Additional independently generated or human-produced texts can test whether the CSSL layer-removal pattern persists in three-tag song and whether IACM's P contribution persists in five-tag speech. Controlled spectral and temporal manipulations can isolate the acoustic conditions associated with SDED:P0, RASN and CSSL. Listener ratings can then test whether the same outputs track perceived expression beyond the production instructions. Neurophysiological tests would require separately specified observables and timing assumptions, particularly because the present engine uses offline context. The contribution of this study is the precise computational target for each test: a mechanism, a layer combination, an observable output and a direction of change.

Methods | Corpus and fixed architecture

Corpus and production reference

We analyzed the archived original-waveform corpus generated through the displayed Eleven v4 interface with the displayed Anna—Young Female Conversational Voice, stability 0.5 and similarity 0.75. The interface did not provide a recoverable model checkpoint, voice identifier or synthesis seed. The corpus is therefore defined by its archived files, receipts, prompt strings and hashes. The matched design includes three intended English texts, two requested delivery modes, five tags and three requests per cell. Song prompts additionally request singing a simple unaccompanied melody. The initial excited/sad/untagged panel and the mad/peaceful extension were generated in different stages. Broader speech-tag, delivery-instruction and fresh-generation panels overlap this core and are identified in the request catalog.

Inclusion and analysis unit

All 342 originals are retained. No recording was excluded using listening quality, agreement with MI, perceived expression or transcription accuracy. The archived transcript audit reports exact matches for 336 of 342 files; its disagreements do not change inclusion. Returned alternatives share a generation request. Within-recording summaries are calculated before averaging available alternatives with equal weight within request. Request means, rather than frames or alternative recordings, enter all reported decoding comparisons. All reported analyses use the original speech and requested-song waveforms.

Fixed extraction and declared output membership

The pinned adapter decodes audio as float32, averages stereo channels to mono and requires a 44.1-kHz sample rate. It computes a 128-bin mel spectrogram with a 2,048-sample Fourier window and a 256-sample hop, applies log1p and the adapter's per-clip mel scaling, and runs R^3 extraction, H^3 context construction and the C^3

executor. Mechanisms are newly instantiated for each clip. The computation is performed without gradient recording or tag-dependent parameter optimization. Some context demands use centered or future samples; the analysis is offline. Metadata supplies each mechanism's channel names, layer membership, processing role and context demands.

Registry policy and exhaustive enumeration

The adapter traverses the exports of functions F1–F8 and retains each mechanism name once. The canonical first-function-wins policy retains F2 IGFE, F1 PNH, F1 TPRD and F2 MAA where duplicate names occur. This produces 84 mechanisms, 326 metadata-defined layers and 889 mechanism channels. The broader archive contains 1,050 outputs, including combinations and regional/neurochemical proxies. The primary response map here concerns the 889 mechanism channels. E/M/P/F labels are interpreted through each mechanism's schema: M can denote Memory or Temporal Integration, and MMP uses R/P/F/C. Three reserved channels are constant across requests and remain documented.

Channel summaries and external comparators

The primary feature for each channel is its full-recording time mean, followed by averaging within request. Temporal sensitivities use five distributional summaries, seven archived temporal descriptors or ten consecutive normalized-time blocks. The acoustic comparator retains all 15 archived descriptors: duration, RMS level, peak level, active fraction, voiced fraction, pitch probability, median F0, F0 interquartile and 90th–10th percentile ranges, spectral centroid, spectral flatness, envelope modulation peak and 2–5-Hz fraction, envelope peak-interval variability and envelope peak count. The native two-output comparator uses VMM:C1:valence_state and AAC:P2:perceptual_arousal. All comparator results remain separate from the mechanism inventory.

Methods | Measuring accessible information

Readout estimation and held-out texts

MI was not trained on this corpus. Only the external mapping from a stored response vector to a production tag was estimated. In each main fold, this mapping uses requests from two texts and predicts the third; repeating the procedure yields one held-out prediction per request. Imputation, centering, scaling and constant-feature removal use only the two estimation texts. Five-tag folds contain 30 estimation and 15 evaluation requests; three-tag folds contain 18 and nine. All alternatives remain inside their request. Since class counts are equal, balanced accuracy equals the reported correct-request fraction.

Original ridge measurement

The original reader uses centered one-hot targets, an intercept and ridge regularization $\alpha = 1$. After standardization, features are divided by the square root of the number of active features. The class with the largest fitted score is selected. No parameter search selects a new headline ridge result. Values $\alpha = 0.1$ and 10, and alternative temporal summaries, are retained as sensitivities for all 84 mechanisms. The complete-C³ reader receives all 889 channels as separate features. Its score measures that vector under this reader, rather than a cross-mechanism arithmetic average.

Nearest-centroid measurement

The channel audit motivated an additional reader because the balanced scalar ridge rule can restrict decisions to extreme class slopes. Following the same fold-specific preprocessing, each class is represented by the mean of its estimation-request vectors. The evaluation vector receives the label of its nearest class mean in squared Euclidean distance. This reader was applied to all 1,299 mechanism/layer/channel candidates and three comparators across four tasks, giving 5,208 evaluations. Eighty-four independent comparisons against scikit-learn reproduced the predictions exactly.⁴ The reader is an explicitly documented adaptive extension, and its results are labeled separately throughout.

Layer exclusions and combinations

Each layer exclusion removes that layer's stored output channels and re-estimates the external reader. The engine and the remaining stored responses are unchanged. All 326 layers are excluded once in each of four tasks, giving 1,304 comparisons. All 16 subsets are evaluated for six mechanisms, four tasks and three readout settings, giving 1,152 comparisons. Settings are original ridge, ridge retaining the complete-mechanism scaling denominator, and nearest centroid. These analyses quantify usefulness of accessible output channels. They do not rerun the engine with an internal stage disabled and therefore do not identify causal necessity of upstream computations.

Restricted permutation tests and discovery scope

We use 999 permutations restricted within text $\times$ delivery mode $\times$ generation cohort, recomputing the external label mapping after each shuffle.⁵ For each task, the null statistic is the maximum agreement across all 84 mechanisms for ridge or all 1,299 mechanism/layer/channel candidates for nearest centroid. Monte Carlo P is $(1 + \text{number of null maxima at least as large as the observed maximum})/1,000$. BY adjustment is applied across the four task-maximum tests within each reader.⁶ This controls the stated search family under the conditional exchangeability assumption. The analysis remains exploratory and conditional on these recorded texts, voice and cohorts.

Additional panels and reproducibility

Same-text panels use generation-request-block holdout; the fresh bridge evaluates ten new requests for a known text. Speech $\leftrightarrow$ song transfer excludes the evaluated text from the opposite-mode estimation set. Binary tasks use separate two-class mappings. Supplementary Note S3 describes automatic feature selection as a distinct secondary question. The delivery includes source tables, archived predictions, editable figure code, numerical assertions and a file-hash manifest. Full extraction depends on the pinned engine and original audio archive identified in provenance; the delivered manuscript package does not claim an independently regenerable synthesis service.

References and research record

1. Livingstone, S. R. & Russo, F. A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **13**, e0196391 (2018). doi:10.1371/journal.pone.0196391.
2. ElevenLabs. ElevenLabs Audio Tags list for Eleven v4, with 50+ examples. *Provider documentation*; accessed 16 October 2026. <https://elevenlabs.io/blog/elevenlabs-audio-tags-list>.
3. Hewitt, J. & Liang, P. Designing and interpreting probes with control tasks. *Proc. EMNLP-IJCNLP*, 2733–2743 (2019). doi:10.18653/v1/D19-1275.
4. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011). <https://jmlr.org/papers/v12/pedregosa11a.html>.
5. Ojala, M. & Garriga, G. C. Permutation tests for studying classifier performance. *J. Mach. Learn. Res.* **11**, 1833–1863 (2010). <https://jmlr.org/papers/v11/ojala10a.html>.
6. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001). doi:10.1214/aos/1013699998.

Data and code availability

The accompanying local package contains the manuscript source, vector figure sources, complete mechanism/layer/channel result tables, prediction records and verification manifests. The original audio and pinned engine remain in the separately identified research archive. The package README specifies these dependencies and their hashes. No public repository accession or permanent DOI is asserted.

Authorship and disclosure

Amaç Erdem developed MI and the research question and assembled the computational study. The author's role as developer of the evaluated architecture is relevant to interpretation. Analysis implementation, manuscript preparation and editorial revision used AI-assisted coding and writing; responsibility for the scientific claims remains with the author.

Research status

This is a computational research manuscript. No human participant responses or neural measurements are reported. Behavioral and neuroscientific validation remain prospective. The local revision is dated 16 October 2026; publication or peer-review status is not implied by this document.