Reduced Speech–Tools Separation in Late TRIBE Representations Across New Sound Sets

New items from recurring sources, with a separate test on new source collections

Brain Researcher

Date: 2026-08-13

Contents

Abstract 1

1. Historical context for the later sound analysis 3

2. A complete sound-category screen motivated the speech–tools test 8

3. Three pre-specified sound sets showed the pattern in new items 10

4. How the speech–tools representation changed 12

5. The pattern did not generalize to new source collections 14

6. How Brain Researcher made the investigation auditable 15

7. Limits and the next test 17

8. Conclusion 18

Methods and provenance 19

References 21

Abstract

TRIBE, the TRImodal Brain Encoder, maps audio, text, and video stimuli into internal model representations and predicted cortical-response features. A historical Brain Researcher (BR) campaign explored seven stimulus branches with different outcomes. Its historical spoken story–math signal did not sustain the original late-layer interpretation under larger-set and acoustic checks; a later analysis localized a bounded brain-alignment signal to the audio projector instead. Together with a noisier auditory candidate, that historical trajectory made audio a productive domain for focused search, but it did not prospectively select the eventual contrast. The current question was broader: which sound-category distinctions change most between early and late TRIBE layers, and what kind of representational change do they show? The analyses in this report concern model representations, not observed fMRI.

An open discovery panel evaluated all 15 pairs among six sound categories using 48 items and four source-held-out folds. Speech versus tools had the third largest absolute early-to-late change in classification AUC ($\Delta\mathrm{AUC} = -.333$). It was not the winner under the pre-specified ranking rule. A post-hoc geometry analysis instead showed a clean reduction in speech–tools separation while mostly retaining the learned direction, and new speech and tool sounds could be assembled with tighter acoustic matching. A researcher therefore authorised this contrast for a pre-specified follow-up study.

Three mutually non-overlapping 48-item panels from the same four recurring source collections then showed late-minus-early changes in normalized separation of $-.694$, $-.447$, and $-.547$. Negative separation change appeared in 11 of 12 source-by-panel cells. A decomposition of the final panel showed that three sources retained the direction while becoming less separated; AudioSet Strong was the explicit exception. A final endpoint used four entirely new source collections. Its aggregate direction remained negative ($\Delta S = -.198$; 3/4 sources), but the pre-specified permutation test did not support the primary hypothesis (raw $p = .132$; Holm-adjusted $p = .396$).

The resulting claim is deliberately bounded. Speech–tools compression repeated across new items from recurring source identities, but it did not pass the new-source endpoint. BR preserved the complete search landscape, recorded the researcher-owned choice, fixed each follow-up panel before evaluation, and retained both the AudioSet exception and the terminal non-support. It did not establish a source-general law, a semantic or neural mechanism, observed-fMRI effects, or scientific confirmation.

What this report establishes

Across three pre-specified panels of new items drawn from recurring source collections, late TRIBE layers showed a repeatable reduction in normalized speech–tools separation. The same direction was weaker and statistically unsupported in a four-new-source endpoint. The evidence therefore supports a recurring-source, new-item representation pattern, not a source-general, semantic, neural, or mechanistic claim.

1. Historical context for the later sound analysis

TRIBE is a multimodal brain-encoding model that accepts video, audio, and text [1]. Its internal layers provide a way to ask which distinctions are preserved, strengthened, or weakened as stimuli move through the model. The historical BR campaign scouted seven stimulus branches spanning HCP language and social contrasts and IBC math, theory-of-mind, auditory, RSVP, and biological-motion contrasts [2, 3]. The branches did not share commensurate scores or packaging, so they cannot be ranked as one formal modality-selection experiment. They produced a mixture of a model-only positive, noisy candidates, representation-sensitive diagnostics, and packaging failures. This entire section supplies historical context only; it is not evidence for the current speech–tools trajectory.

1.1 Historical campaign structure

The scientific objective was not simply to maximize one contrast score. Each branch followed the same operational chain:

$$\text{stimulus question} \rightarrow \text{stimulus set} \rightarrow \text{model representation} \rightarrow \text{contrast score} \rightarrow \text{follow-up} \rightarrow \text{decision}.$$

A branch became interpretable only when a concrete contrast was constructed, scored, changed by a real follow-up, and assigned a bounded status such as continue, stop, or redesign. This distinction mattered because a weak score could mean that the proposed biological axis was absent, but it could also mean that temporal information had been lost during stimulus packaging, that the chosen representation was mismatched to the domain, or that a supposed follow-up had repeated the same stimulus set. BR therefore checked whether each follow-up actually changed the stimulus set and preserved branch trajectories instead of treating every terminal number as a scientific hypothesis.

The historical campaign used an archived TRIBE v2 model release and its released checkpoint [1]. Depending on the branch, the retained artifacts were model representations or predicted cortical-response vectors on `fsaverage5` with 20,484 vertices. These predicted responses are model outputs, not observed participant fMRI. Later layer-level analyses captured the audio and text projectors and selected early and late encoder modules. Those analyses could localize a model representation, but they did not by themselves establish a neural effect.

1.2 Historical context: the seven stimulus branches

Table 1 restores the branch-level context that motivated the later audio search. The labels summarize the latest audited state of each trajectory rather than its most flattering intermediate score.

Scores across predicted-response, encoder-feature, and explicit motion-feature spaces are not commensurate, so they are deliberately not ranked together.

Table 1: Historical stimulus branches and the design information each trajectory contributed. This table provides search context; it is not evidence for the later speech–tools claim.

Branch	Question	Latest class	Consequence for the search
HCP language	Spoken story versus spoken arithmetic	Revised model-only finding	The initial item signal motivated deeper audio work, but the original late-encoder interpretation failed larger-set and acoustic tests; the bounded rescue localized brain alignment to the audio projector.
HCP social	Social versus mechanical motion	Packaging-sensitive failure	A balanced follow-up became weaker, supporting stop or redesign rather than a biological null.
IBC math	Arithmetic principle versus lexical control	Noisy candidate	The follow-up retained the lexical control and showed why condition construction had to be audited.
IBC theory of mind	Belief versus physical questions	Noisy candidate	The branch moved from broad story labels to the more specific question contrast.
IBC auditory	Speech versus five other sound families	Noisy candidate	A nonzero but item-sensitive signal motivated larger stimulus sets and explicit acoustic controls.
IBC RSVP	Temporally presented target versus distractor language	Packaging failure	Static labels discarded the task-defining timing and probe structure, so the current pipeline could not test the intended question.
IBC biological motion	Intact versus scrambled point-light motion	Representation-sensitive candidate	Static and TRIBE-video variants were weak; explicit motion features recovered only a block-sensitive candidate, pointing to representation redesign rather than a neural claim.

The branch landscape and trajectories in Figures 1 and 2 make two parts of the exploration visible that a summary paragraph cannot. First, a positive-looking branch did not automatically advance. Second, a failed branch could still specify which part of the experiment needed redesign.

1.3 Historical context: why the campaign narrowed to audio

Within that contextual survey, the clearest initial item-level result separated spoken stories from spoken arithmetic. Crucially, the search did not stop at that attractive result. Extended-acoustic adjustment collapsed the predicted-response contrast ($p = .190$), the original late-minus-early statistic did not reproduce on the larger 81-item set ($T = -.187$ unadjusted and $T = -3.81$ with extended acoustic adjustment), and late encoder features anti-aligned with a published HCP language-task group map (mean $r = -.28$) [2]. A limited follow-up analysis instead identified the audio projector as the strongest aligned feature (group-map $r = .64$). This corrected trajectory, together with the noisy IBC auditory branch, made audio a productive domain for broader search without specifying which category contrast should win.

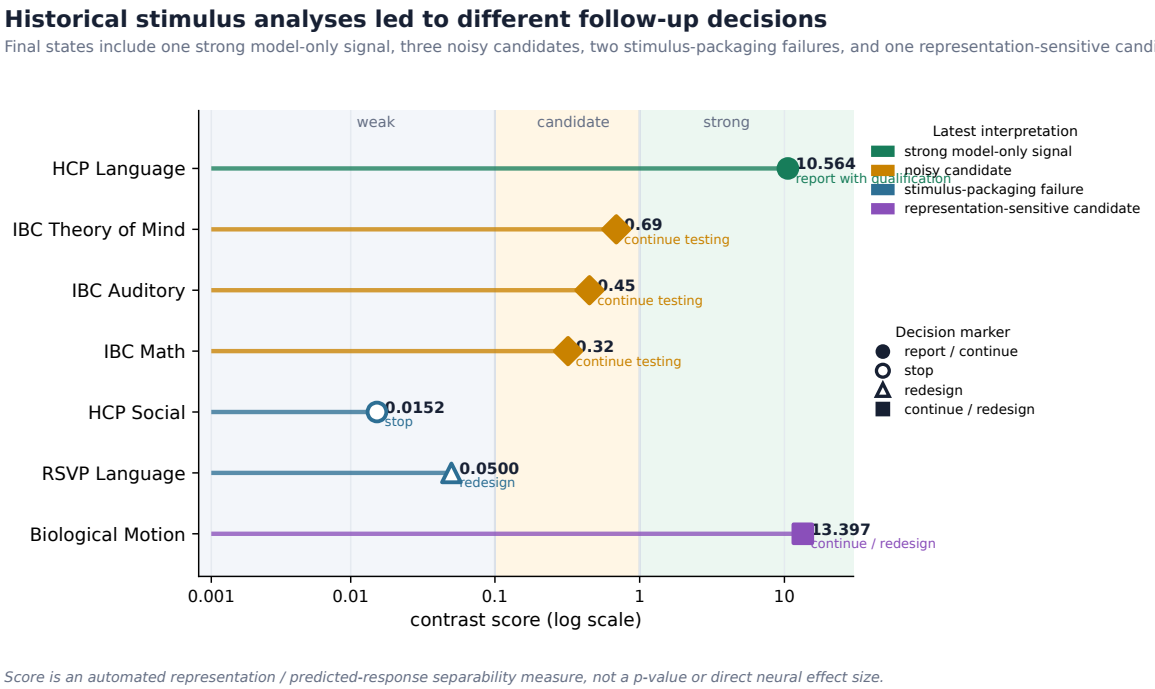
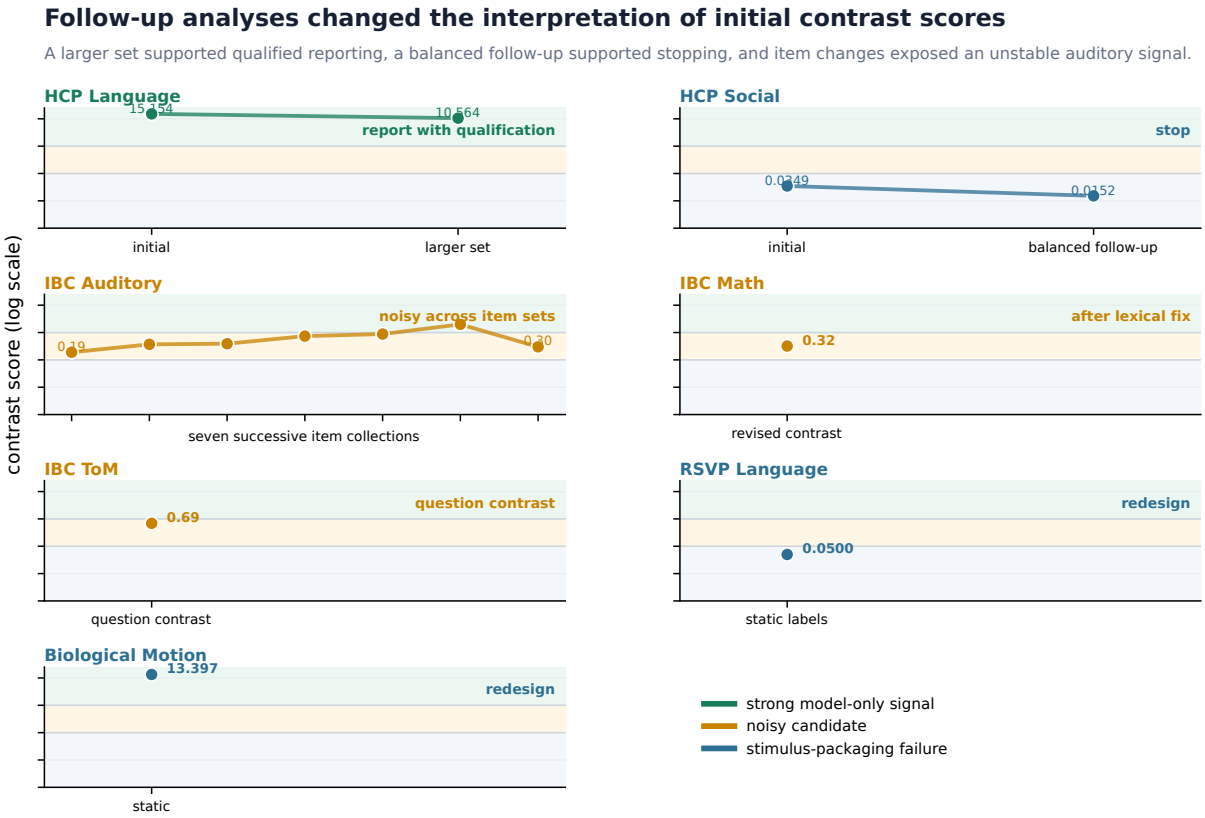


Figure 1: Latest audited state of the seven historical branches. Horizontal positions are branch-specific prioritization scores, colors encode audited hypothesis classes, and markers encode recorded decisions. Because the underlying feature spaces differ, the plot is a branch management landscape rather than a common effect-size comparison.

Thus, audio was selected because two independent branch trajectories indicated that auditory representations were scientifically active and experimentally redesignable, not because the campaign had already found speech–tools. This was motivation, not a pre-specified selector: the historical campaign did not prospectively name speech–tools, and none of its model-side results establish the later geometry or an observed neural effect.

The follow-up study did not assume that story versus math was the right answer. It stepped back and asked a broader question: among ordinary sound categories, which pairwise distinctions change most from early to late TRIBE layers? This turned the earlier result into a search direction rather than a conclusion.

The current report starts at that point. It does not reanalyse the old story–math hypothesis, and it does not merge the two evidence streams. The historical campaign explains why the search focused on sound; all quantitative claims below come from the later speech–tools line.



Single-point branches show the latest evaluated analysis; workflow changes without an audited score are not assigned values.

Figure 2: Historical branch scores over successive materializations and follow-ups. HCP social weakened after a valid balanced follow-up; auditory remained nonzero but noisy; and biological motion changed class only after the representation, rather than merely the videos, was redesigned. Values are model-only or feature-space diagnostics, not observed-fMRI statistics.

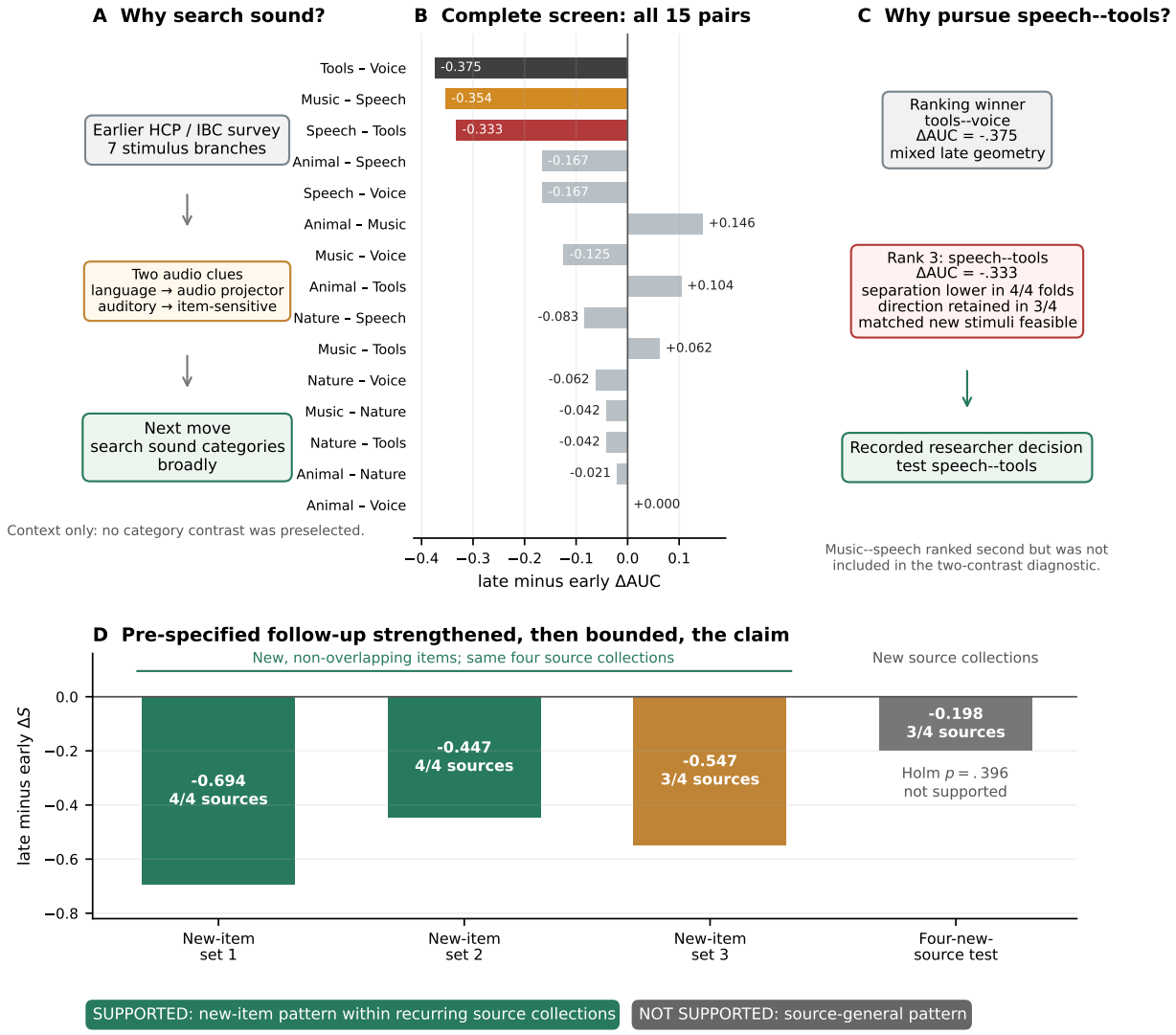


Figure 3: From a complete sound screen to a bounded speech–tools result. (A) Earlier exploratory work motivated a search in sound but did not select a category contrast. (B) The complete screen evaluated all 15 category pairs; speech–tools ranked third by absolute early-to-late AUC change. (C) Its follow-up was selected after a post-hoc geometry diagnostic and a recorded researcher decision, rather than by ranking alone. Music–speech ranked second but was not included in that two-contrast diagnostic. (D) Three pre-specified sets of new items from recurring source collections met the directional criteria, whereas the stronger four-new-source test did not support the source-general claim. The recurring-source sets are directional evidence, not independent source-population replications.

2. A complete sound-category screen motivated the speech–tools test

The discovery panel contained 48 sounds from six categories: animal, music, nature, speech, tools, and voice. The analysis evaluated all 15 category pairs using four held-out source folds. For each pair, it compared how well early and late layer families separated the two categories. The screen was deliberately complete: speech–tools was not inspected in isolation after the fact.

The three largest absolute early-to-late AUC changes were tools–voice ($-.375$), music–speech ($-.354$), and speech–tools ($-.333$). The pre-specified ranking rule selected tools–voice. Speech–tools therefore ranked third and cannot be described as the prospectively selected discovery winner.

Figure 4 shows the complete candidate denominator rather than only the three contrasts later discussed. The panel was a technical re-analysis of the original ordered 48-item discovery set under the pre-specified four-fold implementation, not an independent dataset, and the $\Delta\mathrm{AUC}$ values are descriptive. No inferential p-value was specified for choosing among the 15 pairs.

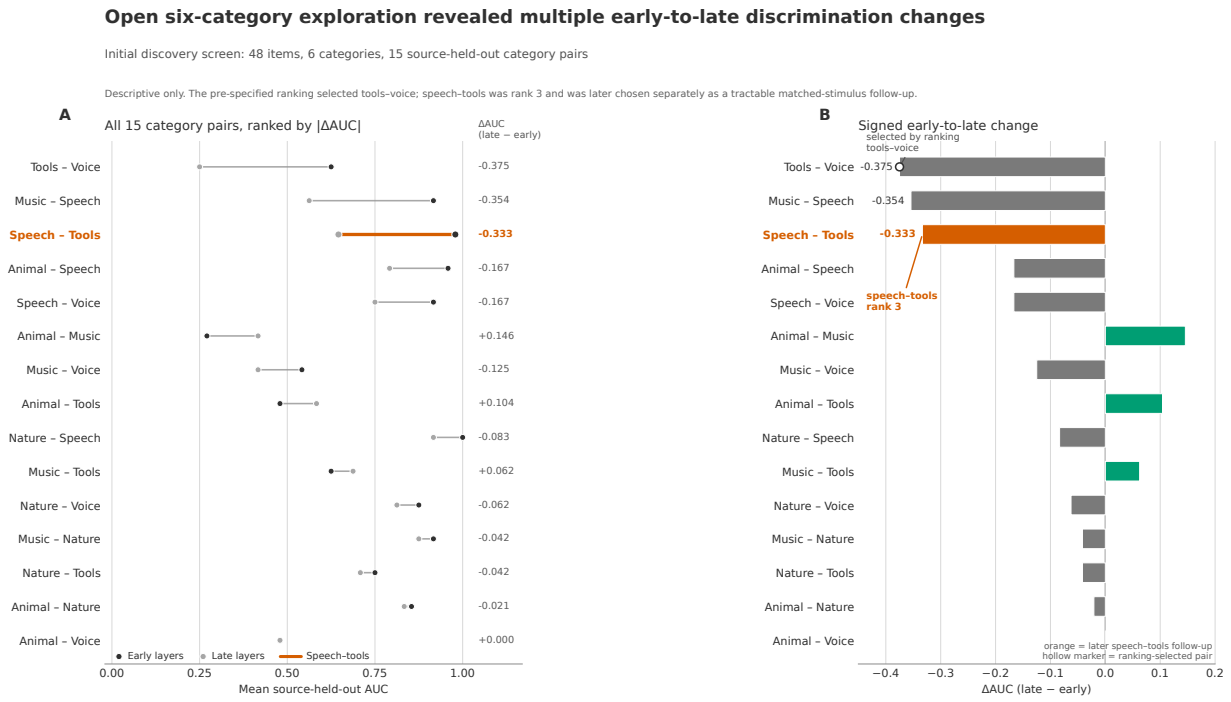


Figure 4: All 15 pairwise contrasts among animal, music, nature, speech, tools, and voice. Dumbbells connect early and late source-held-out AUC; annotations show late-minus-early change. The pre-specified absolute-change winner was tools–voice. Speech–tools ranked third and was adopted only after the separate geometry diagnosis and researcher decision.

The numerical winner did not become the scientific hypothesis by default. A post-hoc analysis decomposed category separation into two questions: did early and late layers preserve the same direction, and did the magnitude of separation change? This follows the broader logic of analysing representation geometry [4], but the fixed-axis metrics used here are specific to this report. Tools–voice produced a mixed pattern with late-axis misalignment and no consistent magnitude reduction. For speech–tools, the early direction was positive in all four source folds; late layers retained that direction in three of four folds, while normalized separation was lower in all four. Music–speech ranked second but was not included in this two-contrast geometry diagnostic, so speech–tools cannot be called the objectively best of the top three.

Speech–tools nevertheless offered an interpretable falsifier and a practical route to prospective testing: new speech and tool sounds could be assembled in acoustically matched sets. Those properties made it a tractable question, not an algorithmically optimal one.

A researcher, not BR, authorised the choice. BR then specified the representation geometry, the decision rule, and the next stimulus panel before evaluation. The ranking defined what had been discovered; the geometry diagnostic and the researcher decision defined what was worth trying to falsify. This selection history is part of the result: the follow-up was reasoned and explicit, but it was still post-hoc relative to the original 15-pair screen.

3. Three pre-specified sound sets showed the pattern in new items

The transition from a selected question to an evaluable sound set was itself part of the record. Two initial setup runs produced no reviewable scientific estimate: one failed because required local text and audio model assets were unavailable, and the other failed model validation because nonexistent image and video backbones were configured. These are infrastructure failures, not null or negative results. The first scientific evaluation followed only after those setup failures had been resolved.

The three scientific evaluations each used 48 newly assembled sounds. We call them new-item sets 1–3. Their item sets did not overlap, but all three drew from the same four source collections. Each set was specified before scoring.

3.1 Building acoustically matched sound sets

Each new-item set contained 24 speech and 24 tool sounds, with six items per condition from each of four WavCaps source subsets: AudioSet Strong, BBC Sound Effects, FreeSound, and SoundBible [5]. The panels therefore balanced condition and source identity rather than allowing one large collection to dominate the aggregate. Previously exposed item identities were excluded before model scoring.

The most fully audited construction record is new-item set 3. Its candidate item list contained 240 rows; exclusion of 48 earlier items left 192 eligible candidates. A score-blind minimax selector then chose 48 items under exact source-by-condition cell counts, source-specific speech/tools event-window histograms, and a maximum absolute pool-standardized mean difference of .5 for seven acoustic descriptors: log duration, RMS, spectral bandwidth, spectral centroid, spectral flatness, 85% spectral rolloff, and zero-crossing rate. The solver returned `optimal`; the largest observed standardized difference was .463 and the pooled value was .374.

WhisperX-derived event windows were also matched within source and condition [6]. For AudioSet Strong, BBC Sound Effects, and SoundBible, all six items in each condition had one event window; in FreeSound, each condition had five one-window and one two-window item. These controls constrain the measured duration, spectrum, amplitude, and segmentation proxies. They do not equate lexical content, semantics, transcript accuracy, or every feature available to TRIBE. The selector was blind to the model scores later used for evaluation.

The primary statistic was $\Delta S = S_{\text{late}} - S_{\text{early}}$, where S is normalized speech–tools separation. A negative value means that normalized speech–tools separation was lower in the late layer

family. Whether the learned direction was retained was assessed separately using the normalized cross-inner-product.

The pre-specified directional rule was conjunctive: aggregate $\Delta S < 0$; negative ΔS in at least three of four sources; early reference AUC above $.5$ in at least three sources; and positive early–late direction concordance in at least three sources. All three new-item sets met that rule. Aggregate ΔS was $-.694$ in set 1, $-.447$ in set 2, and $-.547$ in set 3. Negative ΔS appeared in 4/4 sources, 4/4 sources, and 3/4 sources, respectively: 11 of 12 source-by-panel cells overall.

Table 2: Pre-specified new-item sets using recurring source collections.

Set	Aggregate ΔS	Negative sources	Outcome
New-item set 1	$-.693741$	4/4	Met all directional criteria
New-item set 2	$-.446720$	4/4	Met all directional criteria
New-item set 3	$-.547118$	3/4	Met all directional criteria

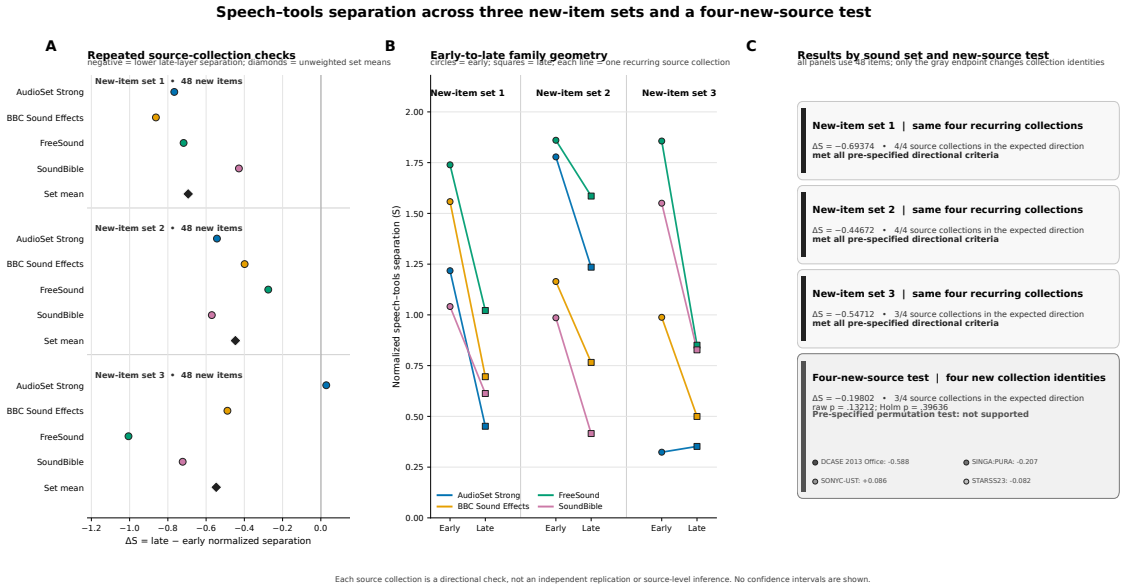


Figure 5: New-item repetition and the new-source endpoint. Circles are source-by-panel stability units and diamonds are unweighted panel means. The three plotted new-item sets met their pre-specified directional criteria using new items from recurring sources. The four-new-source endpoint used four new source collections; its mean remained negative, but its primary hypothesis was not supported after permutation and Holm correction.

These are meaningful repetitions because the items were new and the analysis was specified before evaluation. They are not four-source replications in the population sense. The same source identities recur across sound sets, and the source-by-set cells are stability units rather than independent draws from a defined source population [7]. The pre-specified analyses also did not specify a conventional confidence interval or p-value. The correct evidence state is therefore *pre-specified new-item repetition within recurring sources*.

4. How the speech–tools representation changed

New-item set 3 was used to ask what the negative ΔS meant. BBC Sound Effects, FreeSound, and SoundBible all showed weaker normalized separation in late layers while retaining the learned speech–tools direction. For each of those sources, the negative sign survived every single-item deletion. This argues against a simple disappearance of the distinction: the late representation usually kept the axis but expressed it with less normalized separation.

AudioSet Strong was different. Its full-source ΔS was slightly positive at $+.0284$, and 6 of 12 leave-one-item-out deletions changed the sign relative to that small positive estimate. The exception is not an inconvenient outlier to remove. It is evidence that the pattern is not universal and may depend on how a source collection samples speech and tool sounds.

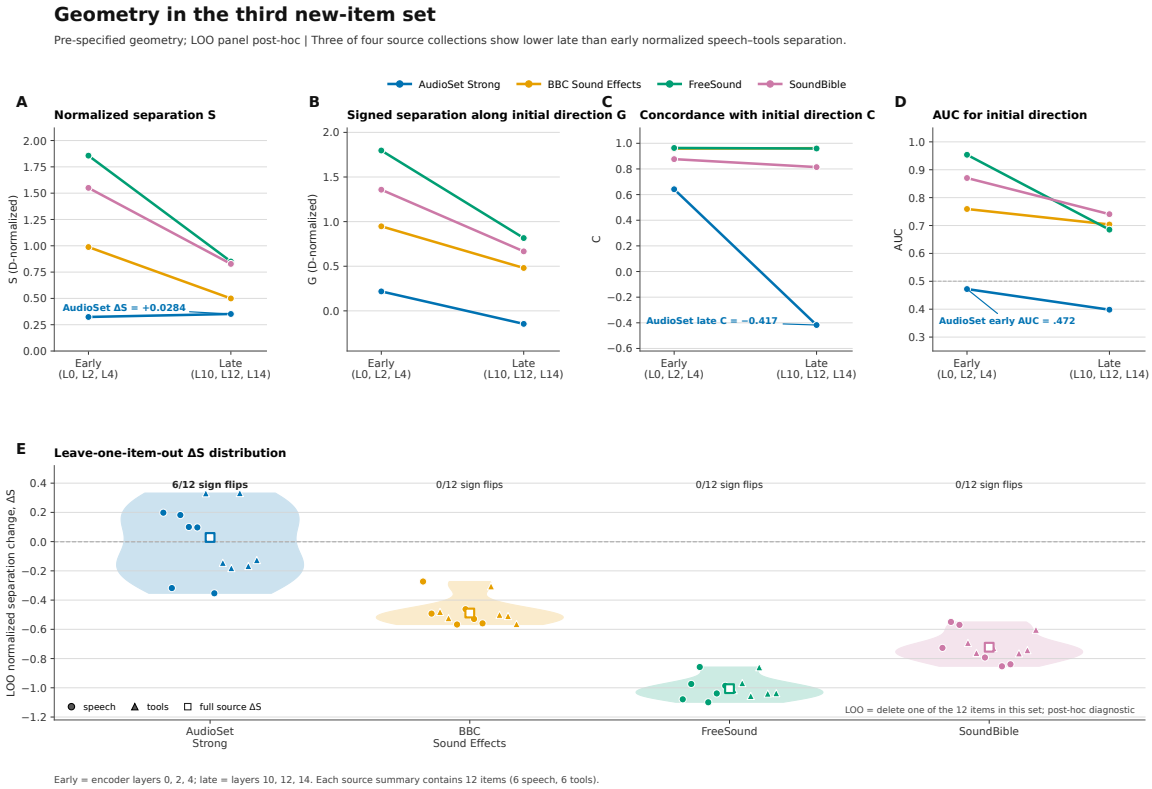


Figure 6: Geometry of new-item set 3. Three source collections retain the learned direction while showing lower late separation. AudioSet Strong is displayed as the explicit exception. The leave-one-item-out analysis is a post-hoc stability diagnostic; it explains the pre-specified result but does not change its evidence tier.

Table 3: Source-level geometry in new-item set 3. S is normalized separation and C is alignment with the reference direction defined in the initial screen.

Source	Early S	Late S	ΔS	Early C	Late C
AudioSet Strong	.3236	.3520	+.0284	.6417	$-.4172$
BBC Sound Effects	.9881	.4996	$-.4885$	.9587	.9618
FreeSound	1.8565	.8506	-1.0059	.9635	.9594
SoundBible	1.5503	.8278	$-.7225$	.8765	.8145

The most specific supported interpretation is therefore *mostly direction-preserving reduction in speech–tools separation*. The result does not show that speech and tool information disappears, that late layers are universally compressed, or that the geometry is semantic rather than acoustic or dataset-specific.

5. The pattern did not generalize to new source collections

The recurring-source panels left one central uncertainty: would the pattern hold when the source collections themselves changed? The final endpoint used 48 sounds selected without inspecting their TRIBE scores from DCASE Office Live [8], SINGA:PURA [9], SONYC-UST [10], and STARSS23 [11]. None of these collections appeared in the earlier panels. Its inferential scope is these four fixed collections, not a population of all possible sound sources.

The aggregate direction remained negative, but the effect was smaller: $\Delta S = -.1980$, with the predicted sign in 3/4 sources. Under the pre-specified balanced-label permutation, the primary hypothesis had raw $p = .1321$ and Holm-adjusted $p = .3964$ across four hypothesis families [12]. Thus the pre-specified test did not support the primary source-general claim.

This result changes the conclusion. The three recurring-source panels support a repeatable new-item pattern, but the line stops short of source-general confirmation. The new-source endpoint is not a weak positive to be added to the earlier panels. It is the prespecified stress test that did not support the stronger claim.

6. How Brain Researcher made the investigation auditable

BR did not supply the scientific answer. It made the trajectory inspectable:

- It retained all 15 discovery pairs rather than presenting only the speech–tools contrast.
- It distinguished the pre-specified ranking winner from the later researcher-authorised choice and preserved that authority boundary.
- It kept two infrastructure failures separate from the three later scientific evaluations.
- It recorded the statistic, item panel, and directional rule before each follow-up evaluation.
- It preserved the AudioSet Strong exception rather than averaging it away.
- It routed the new-source endpoint through independent scientific review and retained the terminal non-support.

This is claim governance rather than autonomous scientific authority. BR helped turn an open screen into a sequence of explicit questions whose results could strengthen, narrow, or stop the claim.

6.1 Researcher decisions and independent review

The trajectory was not a sequence of unowned agent decisions. Table 4 identifies which transition was mechanical, which required researcher authority, and which was assigned by independent review. In particular, the pre-specified ranking rule did not authorise the speech–tools follow-up study; the researcher did. Likewise, a successful run did not authorise confirmation. Neither the recurring-source study nor the new-source endpoint was designated as confirmed.

6.2 Why the without-BR comparison did not support a scientific conclusion

In a separately initiated goal-only baseline without direct BR use, a computational agent did not stop after a shallow start. It extracted representations for 192 audio clips, created a fixed 96/96 computational split, selected an acoustic-geometry contrast, and reported reproducible positive numbers on both subsets. A later audit nevertheless concluded that the analysis did not preserve a valid prospective-validation design. The second subset had not remained access-sealed before candidate selection; the later lock narrowed the eligible analysis; the reported positive effect used a summary/PCA acoustic representation rather than the pre-specified primary analysis; and the pre-specified primary representation was null when implemented.

Table 4: Researcher decisions, review, and scientific interpretation.

Transition	Recorded authority or rule	Resulting state
Historical branch search	BR operated inside a researcher-scoped set of seven stimulus questions; stimulus-set changes and branch decisions were logged	Mixed model-only, candidate, packaging, and redesign outcomes
15-pair discovery screen	Pre-specified absolute-ΔAUC ranking rule	Tools–voice ranked first; speech–tools ranked third
Adoption of speech–tools	Researcher decision, informed by geometry and stimulus constructability	Follow-up question authorised, not confirmed
Follow-up evaluations	Pre-evaluation authority records specified panels, features, statistic, and decision rule	Two setup failures; then three scientific evaluations
Recurring-source outcomes	Deterministic evaluator followed by independent review	Three sets met all directional criteria; confirmation not authorised
Four-new-source endpoint	Pre-specified permutation/Holm evaluator followed by independent review and closeout	Primary claim not supported

The comparison is useful because its scientific failure was not a computational one. The pipeline ran and the positive numbers reproduced, but the final claim was not bound to one preserved protocol. The BR line kept the selection history, pre-specified definitions, and terminal review state attached to each result. The two runs were not matched in prompt, budget, interaction, or stopping rules, so this remains a qualitative contrast rather than a causal estimate of BR's effect.

7. Limits and the next test

Why the conclusion remains bounded

- Speech–tools ranked third in the open screen and was selected after a post-hoc geometry diagnostic and researcher judgement.
- The three supporting panels used new items but recurring source identities; their 12 source-by-panel cells are not independent replications.
- No conventional prospective p-value or confidence interval was specified for the recurring-source panels.
- AudioSet Strong was an observed exception and was sensitive to item deletion.
- The four-new-source endpoint did not pass its pre-specified permutation and multiplicity criteria.
- All results concern TRIBE representations, not observed participant fMRI, neural activation, semantic invariance, or causal mechanism.

The next informative experiment is a new, predeclared source panel with the speech–tools axis, acoustic-matching rules, layer families, statistic, and inference fixed before any representation is inspected. Source collections should be sampled independently of both the recurring panels and the earlier four-new-source test. That experiment should estimate effect size and uncertainty directly rather than rely only on a directional support rule. Observed-fMRI validation would be a separate later tier and should not be implied by another model-only panel.

8. Conclusion

BR used an earlier multimodal stimulus survey to motivate a focused audio question, then exposed the full landscape of 15 pairwise contrasts among six sound categories. Speech–tools was not the ranking winner. It was chosen by a researcher because its geometry suggested a clear, testable reduction in separation and because new matched stimuli could be built.

That choice produced a coherent result. Three pre-specified, non-overlapping sets showed negative late-minus-early separation changes, with the predicted direction in 11 of 12 source-by-panel cells. Geometry analysis showed that the late layers usually retained the speech–tools axis while expressing it less strongly, and it preserved AudioSet Strong as a counterexample. The stronger source-general claim was then not supported by its designated test: a four-new-source endpoint had the same direction but did not pass permutation or multiplicity criteria.

The final result is therefore both positive and bounded. TRIBE shows a repeatable speech–tools geometry pattern across new items from recurring sources. The pattern is not yet supported as source-general, semantic, neural, or mechanistic. BR's scientific value in this case is visible in the complete arc: discovery, explicit human choice, pre-specified repetition, diagnostic decomposition, and a terminal test that was allowed to narrow the claim.

Methods and provenance

Design summary

The discovery analysis used 48 items from six categories and evaluated all 15 pairs over four source-held-out folds. The three follow-up sets each used 48 new, mutually non-overlapping items from four recurring source collections. The pre-specified statistic was $\Delta S = S_{\text{late}} - S_{\text{early}}$, where S is normalized speech–tools separation. New-item set 3 added a post-hoc geometry and leave-one-item-out decomposition. The terminal endpoint used 48 items from four previously unseen source collections and a pre-specified balanced-label permutation with Holm correction across four hypothesis families.

Historical branch representation and score

Historical branches constructed text, audio, video, or task-condition events and retained either TRIBE representations or predicted `fsaverage5` surface responses. For a positive-condition centroid c_P and a negative-condition centroid c_N, the branch-prioritization score was

$$\|c_P - c_N\|_2 \max\{1 - \cos(c_P, c_N), 10^{-6}\}.$$

This score was used to manage a branch, not as a p-value or as a common effect size across representations. A proposed follow-up counted only when its condition signature changed. This stimulus-set-change rule is why the RSVP and early biological-motion failures were classified as packaging or representation problems rather than biological nulls.

Model execution and pre-specified layer families

The speech–tools follow-up sets used pre-provisioned offline assets compatible with the published TRIBE v2 release [1]. The checkpoint and audio and text encoders were fixed before evaluation. Forward hooks captured module 1 of encoder layers 0, 2, and 4 as the early family and layers 10, 12, and 14 as the late family. For each item and layer, the raw hook output had shape $1 \times 200 \times 1152$; averaging over the 200 observations produced one 1,152-dimensional item vector. Each new-item set therefore produced six 48×1152 layer matrices. Family metrics were arithmetic means over their three pre-specified layers.

Fixed reference geometry

The reference axis came from the initial discovery feature set and was kept separate from each follow-up evaluation set. This is a custom fixed-axis representational-geometry analysis, not a direct implementation of standard RSA [4]. At each layer, let δ_{ref} be the reference speech centroid

minus the reference tools centroid, and let D be the root-mean-square norm of the within-condition reference residuals. For an evaluation source, with centroid difference δ_{eval}, the pre-specified metrics were

$$S = \frac{\|\delta_{\text{eval}}\|}{D}, \qquad C = \cos(\delta_{\text{ref}}, \delta_{\text{eval}}), \qquad G = \frac{\langle \delta_{\text{ref}}, \delta_{\text{eval}} \rangle}{\|\delta_{\text{ref}}\| D}.$$

S measures normalized separation magnitude, C measures directional alignment, and G is the signed normalized projection. Reference-axis AUC ranked items by $x \cdot \delta_{\text{ref}}$. The primary source-level estimand was $\Delta S = \text{mean}_{\text{late}}(S) - \text{mean}_{\text{early}}(S)$, and the panel aggregate was the unweighted mean of the four source-level values.

Decision and inferential boundaries

The recurring-source rule required all four conjuncts stated in Section 3 and was a directional support rule, not a conventional significance test. The new-source endpoint instead used its pre-specified balanced-label permutation and Holm adjustment. The leave-one-item-out results and the geometry diagnosis of the original screen were post-hoc explanatory analyses. They did not modify a pre-specified evaluation outcome or raise its evidence tier.

Public evidence and reproducibility

This public report preserves the logical evidence path rather than redistributing private execution records: historical context motivated an open 15-pair sound screen; a post-hoc geometry diagnostic and researcher decision adopted speech–tools; three pre-specified new-item panels tested the selected question; and a four-new-source endpoint narrowed the claim when it did not pass its permutation and Holm criteria. The report source bundle contains the LaTeX text, bibliography, and figures required to regenerate this PDF. The accompanying paper is available at arXiv:2608.19902; public software and portable reproducibility-pack documentation are available at the Brain Researcher public repository and its reproducibility directory. Those public materials document the software and portable verification format, but they are not an immutable replay package for this model-asset-dependent episode. The claim boundaries in this report rest on the stated design and evidence tiers, not on access to private receipts. The historical branch survey is context, not part of the current outcome denominator. The 15-pair screen is an open descriptive screen, the geometry diagnosis and researcher adoption are post-hoc relative to that screen, the three recurring-source panels are pre-specified new-item repetitions, and the new-source endpoint is the only designated test of the stronger source-general claim. Its non-support is therefore retained as a conclusion-changing result rather than averaged into the earlier panels.

The self-contained source directory supports transparent document regeneration: it carries the report text, citation database, report style, and only the figure PDFs cited here. It deliberately does not carry the offline model assets, raw audio selections, intermediate representations, or private execution receipts. A reader can inspect the scientific narrative, plots, equations, captions, and references without mistaking document reproducibility for a replay of the model-dependent evaluation. A future source-general test would require newly assembled source collections and a new predeclared protocol before any relevant representations are inspected.

References

[1] Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Teon Brooks, Katelyn Begany, Joséphine Raugel, Hubert Banville, and Jean-Rémi King. A foundation model of vision, audition, and language for in-silico neuroscience, 2026.

[2] Deanna M. Barch, Gregory C. Burgess, Michael P. Harms, et al. Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage*, 80:169–189, 2013. doi: 10.1016/j.neuroimage.2013.05.033.

[3] Ana Luísa Pinho, Alexis Amadon, Torsten Ruest, et al. Individual brain charting, a high-resolution fMRI dataset for cognitive mapping. *Scientific Data*, 5:180105, 2018. doi: 10.1038/sdata.2018.105.

[4] Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis: Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

[5] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024. doi: 10.1109/TASLP.2024.3419446.

[6] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-accurate speech transcription of long-form audio. In *INTERSPEECH 2023*, pages 4489–4493, 2023. doi: 10.21437/Interspeech.2023-78.

[7] Jacob Westfall, David A. Kenny, and Charles M. Judd. Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143(5):2020–2045, 2014. doi: 10.1037/xge0000014.

[8] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, 2015. doi: 10.1109/TMM.2015.2428998.

[9] Kenneth Ooi, Karn N. Watcharasupat, Santi Peksi, Furi Andi Karnapi, Zhen-Ting Ong, Danny Chua, Hui-Wen Leow, Li-Long Kwok, Xin-Lei Ng, Zhen-Ann Loh, and Woon-Seng Gan. A strongly-labelled polyphonic dataset of urban sounds with spatiotemporal context. In *Proceedings of the 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 982–988, 2021. Dataset release: https://doi.org/10.5281/zenodo.5645825.

[10] Mark Cartwright, Jason Cramer, Ana Elisa Mendez Mendez, Yu Wang, Ho-Hsiang Wu, Vincent Lostanlen, Magdalena Fuentes, Graham Dove, Charlie Mydlarz, Justin Salamon, Oded Nov, and Juan Pablo Bello. SONYC-UST-V2: An urban sound tagging dataset with spatiotemporal context. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, 2020. Dataset release: https://doi.org/10.5281/zenodo.3966543.

[11] Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, et al. STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. In *Advances in Neural Information Processing Systems 36*, pages 72931–72957, 2023. doi: $10.52202/075280\text{-}3189$. Dataset release: https://doi.org/10.5281/zenodo.7880637.

[12] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. doi: $10.2307/4615733$.