PRESERVING WHAT MATTERS: HUMAN-RELEVANT MANIFOLD GEOMETRY OF SPEECH REPRESENTATIONS UNDER ACOUSTIC INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust representations should preserve what matters across perturbations. Yet invariance alone cannot determine whether a representation has isolated task-relevant identity or simply discarded information required downstream. Here, we introduce functional geometry under intervention, a framework for testing when representational geometry is functionally meaningful. We hold sentence identity fixed while manipulating acoustic realization through clean speech, noise, and two speech-enhancement interventions that produce opposite effects on human comprehension. We characterize sentence manifolds across 429 frozen model-layer representations and test whether their geometry preserves sentence identity under an acoustic condition excluded from representation construction. We find that late Whisper encoders develop a coordinated geometric regime with greater inter-sentence separability and reduced within-sentence intervention radius and effective dimension. Across seven direct-audio encoders, these geometric properties reliably track leave-one-condition-out identity retrieval, and the same pattern is reproduced by mean reciprocal rank. Pairwise transfer further reveals an important distinction hidden by aggregate robustness measures: some representations preserve identity only between acoustically neighboring conditions, whereas others support identity across the full intervention space. Human fMRI provides an independent constraint. Beneficial enhancement selectively shifts Inferior Frontal Gyrus (IFG) representations toward the geometry evoked by clean speech, while cross-intervention identity preservation and cortical predictivity remain related but non-equivalent. Together, these findings show that robust representations cannot be identified from invariance alone. A representation is functionally robust when the object survives an unseen intervention and remains available in a form useful to a downstream system.

1 INTRODUCTION

A fundamental challenge in representation learning is to preserve task-relevant identity across nuisance variation while retaining the structure required for downstream computation. In vision, object identity should remain stable across changes in pose, illumination, and corruption; in speech, linguistic content should survive variation in speaker, channel, noise, and signal-processing transformations. Object-manifold theory provides a geometric framework for understanding this problem by treating different realizations of the same object as a manifold in representation space and relating its organization to the separability of task-relevant identities (Chung et al., 2018; Cohen et al., 2020). Across deep vision, speech, and language systems, learning can progressively reorganize these manifolds, separating meaningful structure from nuisance variation and producing increasingly disentangled representations (Cohen et al., 2020; Stephenson et al., 2019; Mamou et al., 2020).

However, invariance alone does not establish whether a representation preserves useful structure. A representation may become insensitive to perturbations because it has successfully isolated task-relevant information, or because it has discarded information required for later computation. The same ambiguity applies to representational geometry. Although manifold capacity, radius, and dimension provide important descriptions of representational organization, they do not imply a universal computational advantage. Depending on the task, useful representations may rely on com-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

pactness, high dimensionality, factorized structure, or orthogonal organization (Rigotti et al., 2013; Fusi et al., 2016; Bernardi et al., 2020; Mante et al., 2013; Flesch et al., 2022). Therefore, geometric structure cannot be interpreted from appearance alone; it must be evaluated through the function it supports.

Speech provides a natural setting for studying this problem because linguistic identity must remain stable despite substantial acoustic variation. Speech enhancement creates a controlled intervention in which sentence identity is preserved while acoustic realization changes. In our paradigm, deep-neural enhancement (DNN-SE) and classical enhancement (Classic-SE) address the same degraded-speech problem but produce opposite effects on human comprehension: relative to unprocessed Noisy speech, DNN-SE improves comprehension (+0.065), whereas Classic-SE reduces comprehension (-0.101). This behavioral dissociation transforms acoustic manipulation from a simple signal-quality comparison into an interpretable intervention space for examining how representations preserve linguistic identity under beneficial and harmful transformations.

Here, we introduce *functional geometry under intervention*, a framework for evaluating when representational geometry supports meaningful robustness. The framework links controlled perturbations, manifold organization, held-out identity preservation, and human cortical relevance, asking whether representations maintain task-relevant identity under unseen transformations rather than merely exhibiting invariance under observed conditions. To quantify representational organization, we use *Mean-Field Theoretic Manifold Analysis* (MFTMA), which characterizes object-manifold geometry through measures of capacity, radius, and effective dimension. These measures describe how representations separate task-relevant identities while accommodating within-object variability, providing a principled connection between geometry and linear separability (Chung et al., 2018; Cohen et al., 2020).

Geometric structure alone, however, does not establish functional robustness. We therefore evaluate whether manifold organization supports identity preservation under held-out acoustic conditions using leave-one-condition-out (LOCO) evaluation and examine human cortical relevance through model-to-brain encoding. Applying this framework reveals distinct representational strategies: late Whisper encoder representations exhibit coordinated geometric untangling across acoustic interventions, whereas self-supervised speech representations such as wav2vec 2.0 and WavLM show more localized adaptation. Model-to-brain encoding further identifies bilateral Inferior Frontal Gyrus (IFG; including IFG pars triangularis and pars opercularis) as an independent cortical constraint on representational robustness. Together, these analyses establish functional geometry under intervention as a framework for determining when representational structure remains meaningful under changing conditions. Our study makes three contributions:

- **Controlled intervention framework.** We establish a comprehension-centered intervention paradigm in which sentence identity is preserved while acoustic conditions are systematically manipulated. The contrasting effects of DNN-SE and Classic-SE provide a functional basis for evaluating representation robustness beyond signal alone.
- **Functional validation of representational geometry.** We demonstrate that manifold organization becomes meaningful when it supports held-out identity preservation. Across 96 sentence identities and four acoustic realizations, MFTMA combined with strict leave-one-condition-out evaluation provides a principled test of whether geometric structure generalizes beyond observed perturbations.
- **Beyond aggregate invariance.** We show that robustness is not captured by a single invariance measure. Intervention transfer reveals distinct forms of identity preservation, while cortical encoding provides an independent constraint linking artificial representations with human neural responses.

2 Related Work

Representational geometry provides a natural language for asking what robust representations preserve. Object-manifold theory relates effective radius, dimension, center structure, and classification capacity to separability (Kriegeskorte & Kievit, 2013; Chung et al., 2018; Chung & Abbott, 2021; Cohen et al., 2020; Perich et al., 2025). Across deep networks, learning can progressively reorganize object manifolds to expose task-relevant variables (Cohen et al., 2020; Stephenson et al., 2019;

Mamou et al., 2020). We adopt the same formalism but use sentence identity as the object and controlled acoustic intervention as within-object variation. Geometry is nevertheless task dependent. Intrinsic neural manifolds constrain motor adaptation, preserved low-dimensional modes can support multiple behaviors, and attention or context can selectively warp or orthogonalize representational spaces (Sadtler et al., 2014; Gallego et al., 2017; 2018; Çukur et al., 2013; Nastase et al., 2017; Mante et al., 2013; Flesch et al., 2022). Recent work likewise documents learning-dependent geometric reorganization in sensory and prefrontal systems (Cheng et al., 2025; Hu et al., 2026; Wójcik et al., 2026). These results motivate a functional criterion: compression or separability matters only insofar as it supports the computation of interest.

Speech is a stringent test because linguistic identity must survive large acoustic variation. Auditory cortex becomes increasingly noise robust beyond primary auditory regions, while degraded-speech comprehension engages temporal, parietal, sensorimotor, and inferior-frontal systems (Kell & McDermott, 2019; Suarez Omedas & Williamson, 2026; Davis & Johnsrude, 2003; Hickok & Poeppel, 2007; Peelle, 2018). Modern speech models impose different computational pressures on this problem: wav2vec 2.0 and WavLM learn self-supervised acoustic representations, whereas Whisper is optimized for large-scale weakly supervised ASR under diverse acoustic conditions (Baevski et al., 2020; Chen et al., 2022; Radford et al., 2023). Prior model-to-brain work has largely asked which model families or processing stages best predict cortical responses (e.g., Tuckute et al. (2023)). Our question is different: whether representations that preserve sentence identity across a controlled acoustic intervention are also those that remain cortically predictive. By separating held-out identity preservation from brain predictivity, we avoid treating geometric compactness as a self-validating robustness score.

3 METHODS

3.1 EXPERIMENTAL DESIGN, BEHAVIOR, AND FMRI

Twenty-five participants heard 96 approximately 10-word sentences across four acoustic conditions (Clean, DNN-SE, Classic-SE, and unprocessed Noisy), with 24 trials per condition assigned using Latin-square counterbalancing across four runs. Noisy speech used stationary speech-shaped noise at -3 dB SNR; DNN-SE used SEMamba (Chao et al., 2024), and Classic-SE used the classical MMSE estimator of Ephraim & Malah (1985). After each sentence, participants completed a two-alternative comprehension question and rated perceived quality on a five-point scale. Computational analyses used all 96 sentence identities in all four acoustic forms, yielding 384 matched waveforms.

MRI data were acquired using a Siemens MAGNETOM Skyra 3T scanner (EPI: $\mathrm{TR} = 2$ s, $\mathrm{TE} = 24$ ms, 38 slices, voxel size $3.4 \times 3.4 \times 4.0\ \mathrm{mm}^3$; T1-weighted MP-RAGE: 1-mm isotropic resolution). Functional data were preprocessed using SPM12, including slice-timing correction, motion correction, coregistration, MNI normalization, and 4-mm FWHM spatial smoothing. Single-trial beta patterns were estimated using a general linear model with one auditory-event regressor per trial and six motion regressors. Voxel patterns were analyzed within bilateral auditory, temporal, parietal, frontal, and sensorimotor regions of interest (ROIs), including Heschl's gyrus (HG), superior and middle temporal gyri (STG, MTG), angular gyrus (AG), inferior frontal pars opercularis and pars triangularis (IFGoper, IFGtri), and precentral and postcentral gyri (PreCG, PostCG). Acquisition details, preprocessing procedures, and complete ROI definitions are provided in Appendices B and C.

3.2 MODEL PANEL AND FROZEN REPRESENTATIONS

The primary direct-audio panel comprises wav2vec2-base/large (Baevski et al., 2020), WavLM-base-plus/large (Chen et al., 2022), and Whisper-tiny/medium/large encoders (Radford et al., 2023). Whisper `decoder_mean` and `decoder_last` are secondary streams; GPT-2 XL (Radford et al., 2019) and Phi-4-mini-instruct (Abouelenin et al., 2025) serve as token-dependent linguistic controls and are not treated as equivalent direct-audio competitors. All models were frozen. Audio was converted to mono 16 kHz and truncated or zero-padded to 4 s; valid hidden states were mean-pooled to one sentence vector per layer. Layers were mapped to normalized depth and summarized into pre-specified Early, Middle, and Late thirds for primary cross-model inference; full trajectories were retained for descriptive analyses.

3.3 CORTICAL INTERVENTION GEOMETRY AND CANONICAL SENTENCE MANIFOLDS

For each participant and cortical system, we represented each acoustic condition in standardized log-radius/log-dimension space, $\mathbf{g} = [z(\log R_M), z(\log D_M)]^T$. The participant-specific Noisy $\rightarrow$ Clean vector $\mathbf{v} = \mathbf{g}_{\text{Clean}} - \mathbf{g}_{\text{Noisy}}$ defines the target direction; each enhancement displacement is $\Delta\mathbf{g}_e = \mathbf{g}_e - \mathbf{g}_{\text{Noisy}}$. We decompose it into (i) Clean-directed alignment $A = \cos(\Delta\mathbf{g}_e, \mathbf{v})$, (ii) signed projected displacement $P = \Delta\mathbf{g}_e^T(\frac{\mathbf{v}}{\|\mathbf{v}\|_2})$, and (iii) orthogonal reorganization $O = \|\Delta\mathbf{g}_e - \frac{P\mathbf{v}}{\|\mathbf{v}\|_2}\|$. These measures respectively capture direction, distance along the Clean-defined axis, and residual off-axis change. DNN-SE and Classic-SE were compared within participant with BH-FDR correction across four pre-specified cortical systems; these are associative state-space measures, not causal localization claims.

For each model layer, the four acoustic realizations of sentence $i = 1, \ldots, 96$ define one canonical object manifold $M_i = \{h_{i,\text{Clean}}, h_{i,\text{DNN-SE}}, h_{i,\text{Classic-SE}}, h_{i,\text{Noisy}}\}$, with 96 manifolds analyzed jointly using MFTMA (Chung et al., 2018; Cohen et al., 2020). Capacity α_M indexes inter-sentence separability; radius R_M quantifies intervention-induced within-sentence spread; and dimension D_M estimates the effective complexity of the sampled intervention subspace. Because each manifold contains only four acoustic samples, D_M is interpreted specifically as sampled intervention-subspace complexity. Exact anchor-point definitions are given in Appendix C.

3.4 FUNCTIONAL VALIDATION I: LOCO RETRIEVAL AND PAIRWISE TRANSFER

We tested whether representational geometry preserves sentence identity under unseen acoustic conditions using leave-one-condition-out (LOCO) retrieval. For each model $\times$ stream $\times$ layer, representations from three acoustic conditions were averaged to form one prototype per sentence. Representations from the held-out condition were then queried against all 96 sentence prototypes using cosine similarity. Top-1 accuracy was the primary metric (chance $= \frac{1}{96} = 1.04\%$), with mean reciprocal rank (MRR) as a secondary metric. Clean, DNN-SE, Classic-SE, and Noisy were held out in turn, ensuring that no condition served as a privileged canonical target. To distinguish local from broadly generalizable identity preservation, we additionally computed directional 4×4 pairwise transfer matrices. For each ordered condition pair, one condition provided the queries and the other the sentence prototypes. This analysis tests whether identity transfers primarily between acoustically neighboring conditions or remains stable across the broader intervention space. Full pairwise transfer results are reported in Appendix A.2.

3.5 FUNCTIONAL VALIDATION II: LOCO MODEL-TO-BRAIN ENCODING

For each participant, model representation, layer, cortical system, and held-out condition, ridge encoding was trained on the 72 trials from the other three conditions and evaluated on the 24 held-out trials. Centering, scaling, and optional dimensionality reduction were performed strictly within training conditions. Held-out voxelwise correlations were Fisher-z transformed and aggregated within ROI and system. IFG is the primary synthesis target because its selection preceded model-to-brain fitting; full four-system predictivity is reported in Appendix A.

3.6 STATISTICAL EVALUATION

Behavior was evaluated with participant-level mixed-effects models and planned condition contrasts. Cortical intervention geometry used paired participant-level tests with BH-FDR correction across four pre-specified systems and nonparametric sensitivity analysis. Geometry-to-function inference was restricted to seven direct-audio encoders summarized at pre-specified Early, Middle, and Late stages (21 model-stage observations). Adjacent layers were not treated as independent; inference was clustered by model. Associations of α_M, R_M, and D_M with LOCO Top-1 retrieval were assessed by Spearman rank regression with model-clustered standard errors and BH-FDR correction across the three descriptors; MRR was a pre-specified secondary outcome. A 10,000-resample model-cluster bootstrap retained all stages within each sampled model. For visualization only, coefficients were sign-aligned as $\{\rho_{\text{capacity}}, -\rho_{\text{radius}}, -\rho_{\text{dimension}}\}$, so larger values consistently indicate stronger association with better retrieval without altering the underlying correlations or uncertainty. Because the descriptors are strongly collinear, joint coefficients are interpreted de-

scriptively rather than as unique causal contributions. Neural encoding treated participants as the independent units of inference, with all preprocessing fit strictly on training data before evaluation on held-out data.

4 RESULTS

4.1 BEHAVIOR DEFINES BENEFICIAL AND HARMFUL INTERVENTION

The two enhancement methods had opposite effects on comprehension. Relative to Noisy speech, DNN-SE increased comprehension by 0.065 (Bonferroni $p = .025$), whereas Classic-SE reduced it by 0.101 ($p < .001$). Perceived quality showed a related but distinct pattern: DNN-SE increased quality by 0.782 ($p < .001$), while Classic-SE did not reliably differ from Noisy (-0.224; $p = .248$). We therefore define DNN-SE as behaviorally beneficial and Classic-SE as harmful based on recovered sentence meaning rather than signal quality alone. This dissociation turns acoustic processing into a functional intervention: sentence identity is fixed while the transformation either supports or impairs recovery of meaning.

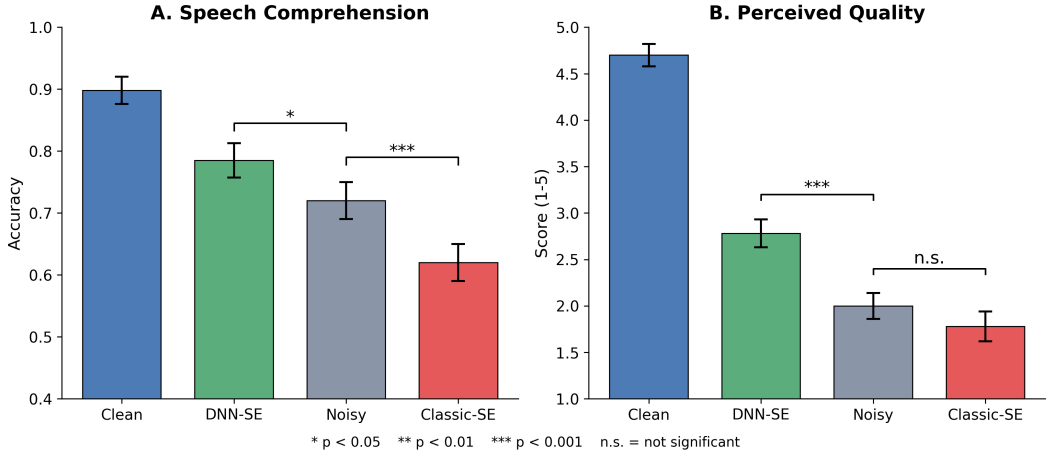


Figure 1: Behavioral effects distinguish beneficial and harmful acoustic interventions. **(A)** DNN-SE improves speech comprehension relative to Noisy speech, whereas Classic-SE further reduces comprehension. **(B)** Perceived quality shows a related but non-identical pattern: DNN-SE improves quality relative to Noisy speech, whereas Classic-SE does not differ significantly from Noisy speech. Error bars indicate SEM. $*$ $p < .05$, $**$ $p < .01$, $***$ $p < .001$; $n.s.$, not significant.

4.2 BENEFICIAL ENHANCEMENT FOLLOWS THE CLEAN-DEFINED IFG DIRECTION

The corrected DNN-SE versus Classic-SE difference was selective to bilateral IFG (Figure 2). DNN-SE produced stronger Clean-directed alignment than Classic-SE (0.578 vs. 0.239; $\Delta = 0.339$; 95% CI $[0.093, 0.586]$; $p = .009$; $q = .036$) and greater projected displacement (1.177 vs. 0.474; $\Delta = 0.703$; 95% CI $[0.220, 1.185]$; $p = .006$; $q = .025$), which also survived paired Wilcoxon analysis ($q = .022$). Orthogonal reorganization did not differ ($q = .186$), and no other cortical system showed a corrected enhancement effect. Thus, successful enhancement shifted IFG representations farther along the Clean-defined direction without increasing off-axis reorganization.

4.3 LATE ASR ENCODERS EXHIBIT COORDINATED UNTANGLING

Across 429 model-layer representations and $41{,}184$ sentence manifolds, model families followed distinct geometric trajectories (Figure 3). Whisper encoders showed the clearest coordinated untangling from Early to Late stages: capacity increased from 0.406 to 0.536, radius contracted from 2.277 to 1.294, and effective dimension decreased from 2.463 to 2.196. wav2vec 2.0 showed a

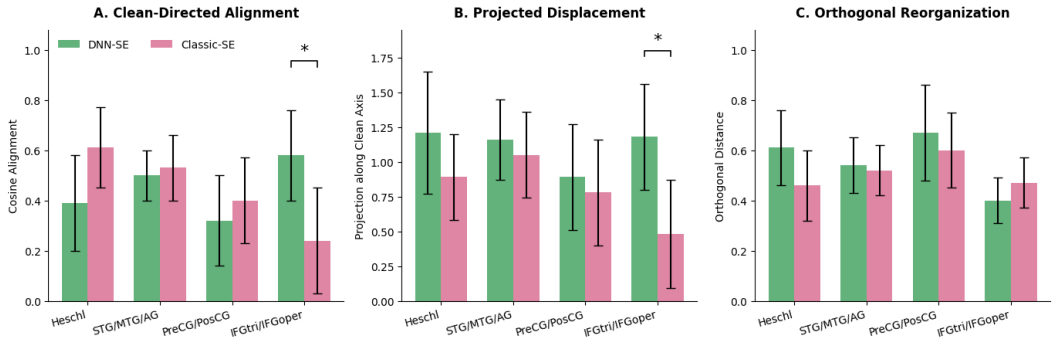


Figure 2: Cortical geometric decomposition under speech enhancement ($N = 25$). Relative to Classic-SE, DNN-SE shows stronger Clean-directed alignment (A, $q = .036$) and greater projected displacement along the Clean-defined axis (B, $q = .025$) in bilateral IFG, whereas orthogonal reorganization (C) does not differ reliably ($q = .186$). Asterisks indicate $q < .05$, BH-FDR corrected across the four pre-specified cortical systems. Error bars indicate SD.

weaker transition, whereas WavLM primarily showed partial compression. Whisper-medium encoder layer 24 exhibited the strongest single direct-audio regime ($\alpha_M = 0.596$, $R_M = 1.130$, $D_M = 2.094$), with late Whisper-large layers following a similar trajectory. Thus, in the strongest ASR encoders, sentence identities become more separable while acoustic interventions occupy a smaller and simpler within-identity space.

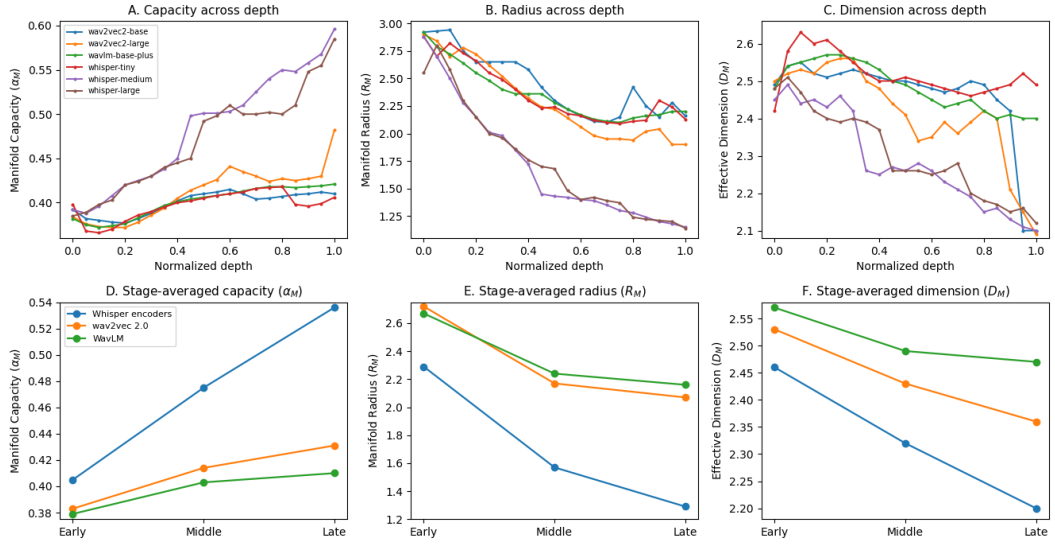


Figure 3: Coordinated sentence-manifold untangling across model depth. (A-C) Layerwise trajectories of manifold capacity (α_M), radius (R_M), and effective intervention-subspace dimension (D_M) across the direct-audio encoders. (D–F) Early/Middle/Late family averages. Whisper encoders show the clearest coordinated increase in separability with contraction of intervention-induced radius and dimension.

4.4 Coordinated geometry predicts held-out identity preservation

We next asked whether this geometric regime preserves sentence identity under an unseen intervention. Across the seven direct-audio encoders summarized at pre-specified Early, Middle, and Late stages ($N = 21$), LOCO Top-1 retrieval increased with capacity ($\rho = .643$, $q = .012$) and decreased with radius ($\rho = -.604$, $q = .012$) and intervention-subspace dimension ($\rho = -.573$, $q = .021$).

MRR reproduced the same pattern (capacity $\rho = .586$, $q = .028$; radius $\rho = -.542$, $q = .028$; dimension $\rho = -.534$, $q = .033$).

The descriptors were strongly coupled ($|r| = .939$–$.975$), and model-cluster bootstrap comparisons did not identify a uniquely stronger association: radius versus capacity, 95% CI $[-.176, .140]$; radius versus dimension, 95% CI $[-.224, .201]$. Bootstrap uncertainty was broad because only seven model clusters were independent (Appendix D), and per-layer analyses are reported in Appendix D.1. We therefore interpret the result as coordinated untangling rather than a uniquely dominant geometric descriptor: representations that better separate sentence identities while contracting the intervention space also better preserve identity under a held-out condition.

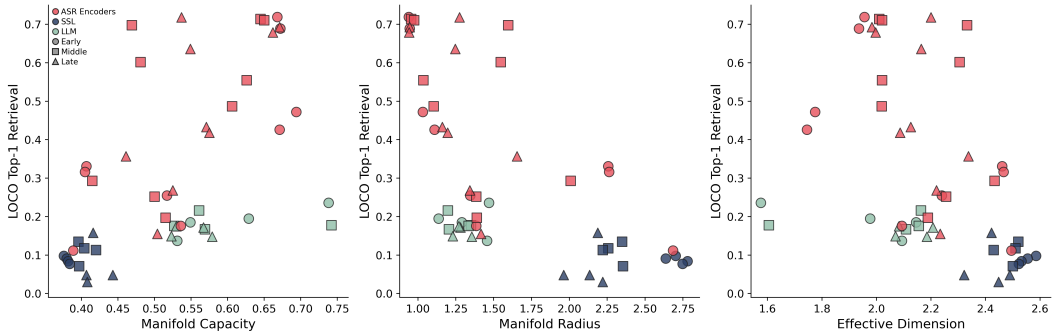


Figure 4: Sentence-manifold geometry and held-out identity preservation. Stage-level capacity, radius, and effective dimension are related to four-fold LOCO Top-1 retrieval across the seven direct-audio encoders ($N = 21$ model $\times$ stage summaries). Primary inference uses model-clustered rank regression; MRR gives the same directional pattern. Secondary representation streams are shown for context but are excluded from the primary inferential population.

4.5 Pairwise transfer separates local from intervention-general identity

Aggregate LOCO retrieval can obscure how identity generalizes across acoustic conditions. As shown in Figure 5, wav2vec2-large encoder layer 13 exhibits local acoustic-neighborhood transfer, with strong Clean-to-DNN-SE and DNN-SE-to-Clean transfer (0.531 and 0.542) but weak transfer across the remaining condition pairs (0.000–0.094). In contrast, Whisper-large decoder layer 13 exhibits broad intervention-general identity transfer across all conditions, including Clean-to-Noisy and Noisy-to-Clean transfer (0.760 and 0.885). These panels are descriptive exemplars rather than a matched causal comparison. The full nine-panel analysis in Appendix A.2 shows that this distinction extends beyond the two examples. wav2vec2 and WavLM representations predominantly support local transfer, whereas Whisper-medium and Whisper-large approach intervention-general transfer. Although both exemplars are large-family variants at the same nominal layer index, they differ in architecture, depth, parameterization, and training objective and therefore do not isolate a causal source of the difference. Together, these results show that local acoustic stability does not necessarily imply preservation of a common sentence identity across the full intervention space.

4.6 Functional preservation and cortical relevance impose different constraints

LOCO cortical encoding provides a complementary human-relevance constraint. IFG predictivity varied with architecture, representation stream, and processing stage rather than tracking transfer monotonically. Whisper encoders increased from approximately $r = .0076$ at Early stages to $r = .0175$ at Late stages, and late Whisper `decoder_mean` states reached approximately $r = .0157$; WavLM remained near zero, whereas token-derived controls retained non-trivial predictivity despite weaker acoustic transfer. Stage profiles are reported in Appendix A.1.

Figure 6 makes this dissociation explicit. Late Whisper-medium and Whisper-large combine strong transfer with comparatively high IFG predictivity, whereas wav2vec2-base retains appreciable IFG prediction despite weaker transfer. GPT-2 XL and Phi-4-mini-instruct are comparatively IFG-

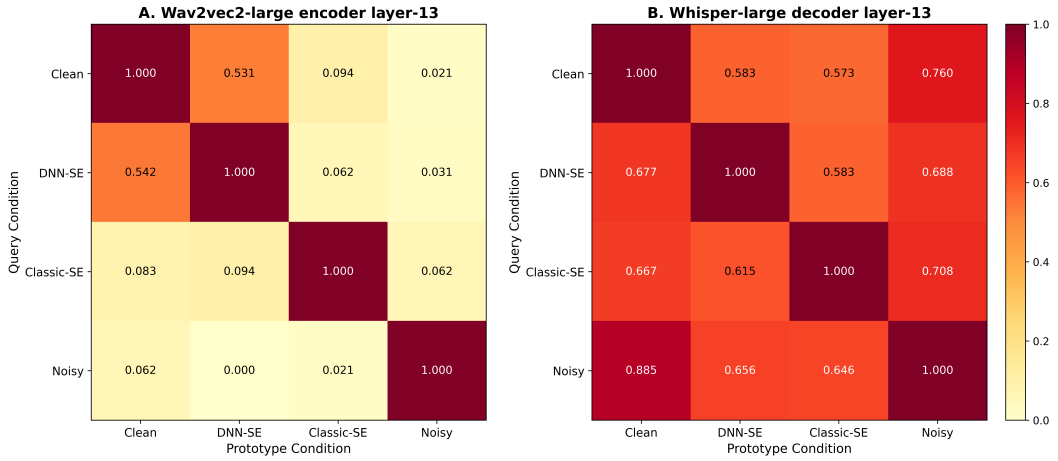


Figure 5: Pairwise transfer exposes local versus intervention-general identity. (A) wav2vec2-large encoder layer 13 primarily preserves identity between the acoustically neighboring Clean and DNN-SE conditions. (B) Whisper-large decoder layer 13 transfers sentence identity broadly across all four acoustic regimes, including Clean-to-Noisy and Noisy-to-Clean. The panels are descriptive exemplars; all nine representative states are shown in Appendix A.2.

predictive despite modest acoustic transfer, while high-transfer Whisper decoder states do not maximize IFG predictivity. Because these streams differ in information pathway, they are not directly matched competitors. Instead, the comparison shows that preserving sentence identity across acoustic intervention and predicting higher-order cortical responses are related but distinct representational properties.

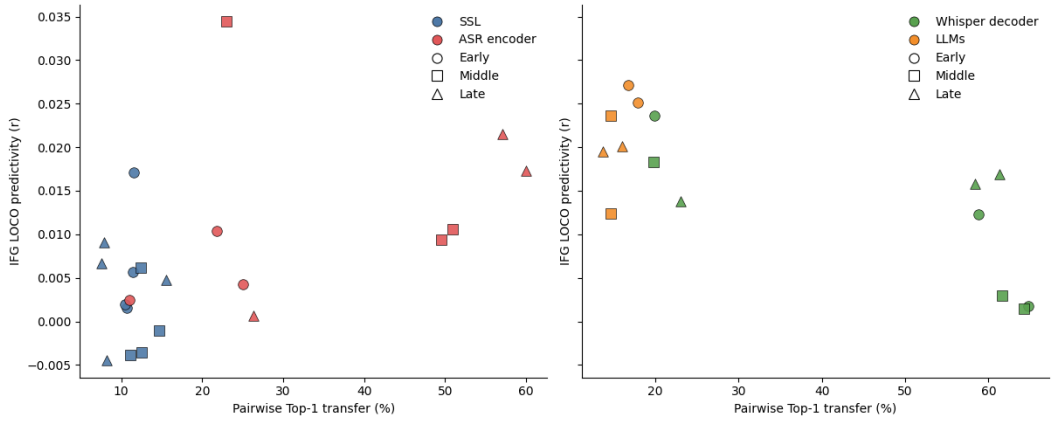


Figure 6: Beyond invariance: intervention-general content preservation versus IFG relevance. (A) Direct-audio encoders show a broadly positive but architecture-dependent relation between pairwise cross-condition transfer and IFG LOCO predictivity. (B) Decoder-conditioned and token-derived streams reveal a complementary dissociation. Points denote model × stream × stage summaries. The figure is a two-axis descriptive synthesis, not a one-dimensional model ranking.

5 Discussion

5.1 Functional geometry requires a held-out consequence

The central contribution of this work is a criterion for deciding when representational geometry matters. Geometry describes how a representation is organized, but not what that organization pre-

serves. Functional geometry under intervention resolves this ambiguity by linking geometry to a held-out consequence. MFTMA characterizes how acoustic intervention reorganizes sentence manifolds. LOCO retrieval tests whether that organization preserves identity under a condition excluded from prototype construction, while cortical encoding asks whether the surviving representation remains relevant to a human downstream system. These levels are complementary rather than interchangeable. A compact manifold that fails held-out retrieval is not useful invariance, and successful retrieval alone does not establish human-like computation.

Across direct-audio encoders, better held-out identity preservation was associated with greater manifold capacity, smaller intervention radius, and lower intervention-subspace dimension. Because these descriptors are strongly coupled, the evidence supports a coordinated compact-and-separable regime rather than a single dominant factor. This pattern should not be interpreted as a universal preference for compact or low-dimensional representations, since expanded, high-dimensional structure can support mixed selectivity, context separation, and flexible computation (Rigotti et al., 2013; Fusi et al., 2016; Bernardi et al., 2020; Mante et al., 2013; Flesch et al., 2022). Here, geometric contraction is meaningful because it accompanies successful identity preservation under an unseen acoustic intervention.

Model trajectories reinforce this interpretation. Whisper encoders show the clearest coordinated untangling and strongest held-out identity preservation, whereas wav2vec 2.0 and WavLM show weaker or more localized changes. These differences are descriptive rather than causal because the model families differ in architecture, scale, and training. Pairwise transfer further distinguishes local acoustic smoothness from intervention-general identity. Some representations preserve identity between acoustically similar conditions but fail across larger intervention boundaries, whereas others maintain a common sentence identity across Clean, enhanced, and noisy speech. Thus, local stability alone does not establish robustness across the broader intervention space.

Human cortical relevance provides an additional constraint. Beneficial enhancement selectively reorganized IFG toward the Clean-defined direction, while cortical predictivity did not track cross-condition transfer monotonically. Some representations predicted IFG despite weaker acoustic transfer, whereas highly transferable representations did not necessarily maximize cortical predictivity. Functional identity preservation and cortical predictivity therefore provide complementary rather than interchangeable evidence. Together, these results support a conjunctive criterion for robust representation learning. A robust representation should preserve object identity across perturbations while retaining information that supports downstream function.

5.2 Limitations

This study examines four controlled acoustic realizations of each sentence, comprising Clean speech, stationary noise, one DNN enhancement method, and one classical enhancement method. Generalization across speakers, channels, nonstationary noise, languages, and enhancement families remains to be established. With only four realizations per sentence, D_M reflects the complexity of the sampled intervention subspace rather than unrestricted nuisance dimensionality. Geometry-to-function inference is limited to seven independent direct-audio model clusters. Although clustered analyses and MRR support a coherent geometric regime, bootstrap uncertainty limits fine-grained comparisons among descriptors. IFG encoding likewise provides a complementary measure of human cortical relevance rather than a global ranking of speech models.

6 Conclusion

Functional geometry under intervention provides a stricter test of robustness than invariance alone. Across speech models, coordinated manifold organization predicts whether sentence identity survives an unseen acoustic intervention, while pairwise transfer distinguishes local acoustic smoothness from intervention-general identity. Human fMRI provides a complementary cortical constraint, showing that successful enhancement reorganizes IFG toward the Clean-defined direction without making cortical predictivity equivalent to functional preservation. Robust representations should preserve object identity across perturbations while retaining information that supports downstream function.

AI USE STATEMENT

Generative AI tools were used to support interpretation of results, refinement of research hypotheses, and feedback on research methodology and experimental analyses. They were also used for language editing, manuscript restructuring, presentation refinement, literature and reference checking, and formatting assistance. Generative AI was not used to generate experimental data. All AI-assisted suggestions were critically reviewed by the authors against the underlying analyses and relevant source literature. The authors made all final methodological, analytical, and interpretive decisions and take full responsibility for the contents, results, citations, and scientific claims of this work.

ETHICS STATEMENT

All experimental procedures involving human participants adhered to the Declaration of Helsinki and were formally approved by the Institutional Review Board (Research Ethics Committee). Participants were free to withdraw from the study at any time without penalty and received monetary compensation commensurate with local institutional standards for their participation time. All behavioral scores and functional neuroimaging datasets were fully de-identified, assigned randomized subject codes to protect participant privacy. The acoustic interventions and computational speech models investigated in this study pose no foreseeable dual-use or societal harms, aiming instead to illuminate the principles of robust speech representation and inform the design of restorative auditory and speech technologies.

REPRODUCIBILITY STATEMENT

To ensure complete reproducibility of the empirical findings and computational models presented in this work, we document all experimental, algorithmic, and statistical specifications across the main text and appendices. Human fMRI and behavior experiment procedures are detailed in Appendix B and C. Implementation detail and model specification can be found in Appendix C. Analysis metrics and statistical analyses are elaborated mostly in Appendix C and D.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025. doi: 10.48550/arXiv.2503.01743.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

Silvia Bernardi, Marcus K. Benna, Mattia Rigotti, et al. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4):954–967.e21, 2020. ISSN 0092-8674. doi: 10.1016/j.cell.2020.09.031. URL https://www.sciencedirect.com/science/article/pii/S0092867420312289.

Rong Chao, Wen-Huang Cheng, Moreno La Quatra, et al. An investigation of incorporating mamba for speech enhancement. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 302–308, 2024. doi: 10.1109/SLT61566.2024.10832332.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, October 2022. ISSN 1941-0484. doi: 10.1109/jstsp.2022.3188113.

Yu-Ang Cheng, Mehdi Sanayei, Xing Chen, et al. A neural geometry approach comprehensively explains apparently conflicting models of visual perceptual learning. *Nature Human Behaviour*, 9(5):1023–1040, 2025. doi: 10.1038/s41562-025-02149-x.

SueYeon Chung and L.F. Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology*, 70:137–144, October 2021. ISSN 0959-4388. doi: 10.1016/j.conb.2021.10.010.

SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3), 2018. ISSN 2160-3308. doi: 10.1103/physrevx.8.031003.

Uri Cohen, SueYeon Chung, Daniel D Lee, et al. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 11(1):746, 2020. doi: 10.1038/s41467-020-14578-5.

Tolga Çukur, Shinji Nishimoto, Alexander G Huth, et al. Attention during natural vision warps semantic representation across the human brain. *Nature neuroscience*, 16(6):763–770, 2013. doi: 10.1038/nn.3381.

Matthew H. Davis and Ingrid S. Johnsrude. Hierarchical processing in spoken language comprehension. *Journal of Neuroscience*, 23(8):3423–3431, 2003. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.23-08-03423.2003. URL https://www.jneurosci.org/content/23/8/3423.

Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):443–445, 1985. doi: 10.1109/TASSP.1985.1164550.

Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, et al. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, 110(7):1258–1270.e11, 2022. ISSN 0896-6273. doi: 10.1016/j.neuron.2022.01.005. URL https://www.sciencedirect.com/science/article/pii/S0896627322000058.

Stefano Fusi, Earl K Miller, and Mattia Rigotti. Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37:66–74, 2016. ISSN 0959-4388. doi: 10.1016/j.conb.2016.01.010. URL https://www.sciencedirect.com/science/article/pii/S0959438816000118. Neurobiology of cognitive behavior.

Juan A. Gallego, Matthew G. Perich, Lee E. Miller, et al. Neural manifolds for the control of movement. *Neuron*, 94(5):978–984, 2017. ISSN 0896-6273. doi: 10.1016/j.neuron.2017.05.025. URL https://www.sciencedirect.com/science/article/pii/S0896627317304634.

Juan A Gallego, Matthew G Perich, Stephanie N Naufel, et al. Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nature communications*, 9(1):4233, 2018. doi: 10.1038/s41467-018-06560-z.

Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402, 2007. doi: 10.1038/nrn2113.

Bo Hu, Nesibe Z Temiz, Chi-Ning Chou, et al. Representational learning by optimization of neural manifolds in an olfactory memory network. *Nature Neuroscience*, pp. 1–12, 2026. doi: 10.1038/s41593-026-02429-3.

Alexander JE Kell and Josh H McDermott. Invariance to background noise as a signature of non-primary auditory cortex. *Nature communications*, 10(1):3958, 2019. doi: 10.1038/s41467-019-11710-y.

Nikolaus Kriegeskorte and Rogier A Kievit. Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*, 17(8):401–412, 2013. doi: 10.1016/j.tics.2013.06.007.

Jonathan Mamou, Hang Le, Miguel Del Rio, et al. Emergence of separable manifolds in deep language representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6713–6723. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/mamou20a.html.

Valerio Mante, David Sussillo, Krishna V Shenoy, et al. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013. doi: 10.1038/nature12742.

Samuel A. Nastase, Andrew C. Connolly, Nikolaas N. Oosterhof, et al. Attention selectively reshapes the geometry of distributed semantic representation. *Cerebral Cortex*, 27(8):4277–4291, 08 2017. ISSN 1047-3211. doi: 10.1093/cercor/bhx138.

Jonathan E Peelle. Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, 39:204–214, 2018. doi: 10.1097/AUD.0000000000000494.

Matthew G Perich, Devika Narain, and Juan A Gallego. A neural manifold view of the brain. *Nature Neuroscience*, 28(8):1582–1597, 2025. doi: 10.1038/s41593-025-02031-z.

Alec Radford, Jeffrey Wu, Rewon Child, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Tao Xu, et al. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Mattia Rigotti, Omri Barak, Melissa R Warden, et al. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013. doi: 10.1038/nature12160.

Patrick T Sadtler, Kristin M Quick, Matthew D Golub, et al. Neural constraints on learning. *Nature*, 512(7515):423–426, 2014. doi: 10.1038/nature13665.

Cory Stephenson, Jenelle Feather, Suchismita Padhy, et al. *Untangling in invariant speech recognition*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Tomas Suarez Omedas and Ross S. Williamson. Noise-invariant representations of sound emerge along the canonical cortical hierarchy. *PLOS Biology*, 24(7):e3003915, 07 2026. doi: 10.1371/journal.pbio.3003915.

Greta Tuckute, Jenelle Feather, Dana Boebinger, et al. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Biology*, 21(12):e3002366, 12 2023. doi: 10.1371/journal.pbio.3002366.

Michał J. Wójcik, Jake P. Stroud, Dante Wasmuht, et al. Learning shapes neural geometry in the primate prefrontal cortex. *Nature Neuroscience*, 29:1966–1975, 2026. doi: 10.1038/s41593-026-02333-w.

A FULL CORTICAL LOCO ENCODING RESULTS

The main text focuses on IFG because it was independently identified by the human intervention contrast. For completeness, this appendix reports the remaining pre-specified cortical systems and the IFG stage profiles omitted from Figure 6 to save main-text space. The clearest non-IFG pattern is that Heschl's gyrus is the most predictable cortical system across most direct-audio families, providing an auditory-fidelity reference. IFG is emphasized not because it is globally easiest to predict, but because it is selectively implicated by the beneficial-versus-harmful intervention contrast.

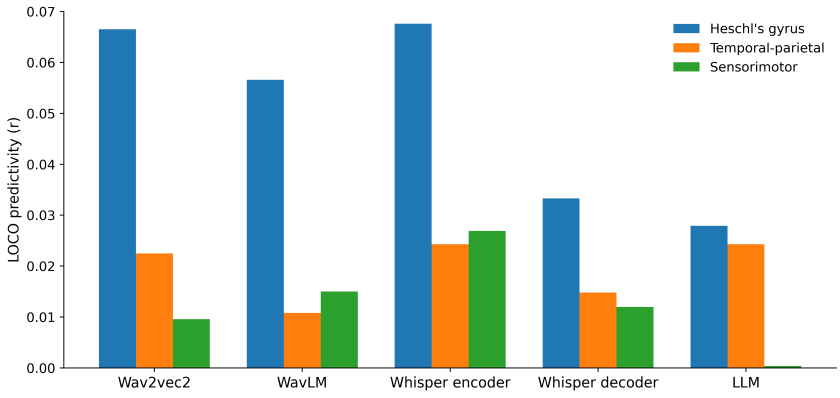


Figure A1: Late-stage cortical LOCO encoding across non-IFG systems. Heschl's gyrus is the strongest cortical target for most direct-audio model families, consistent with its role in early auditory analysis. Temporal-parietal and sensorimotor systems show lower and more architecture-dependent predictivity. IFG is omitted here because it is analyzed as the primary functional cortical target in the main text.

A.1 IFG STAGE PROFILES

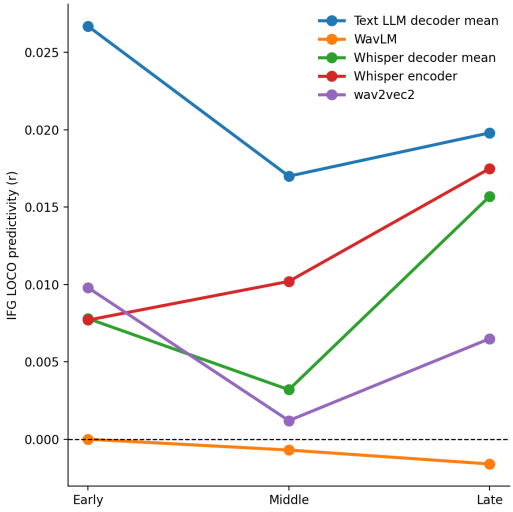


Figure A2: Condition-general IFG LOCO predictivity across Early, Middle, and Late stages. Predictivity varies by architecture and representation stream rather than tracking cross-intervention transfer monotonically. This panel was moved from the main text so that Figure 6 focuses only on local versus global transfer; the values are interpreted as a descriptive stage profile unless supported by the pre-specified participant-level inference.

A.2 EXTENDED PAIRWISE TRANSFER EXAMPLES

The full nine-panel transfer set provides the context for the two main-text exemplars. wav2vec2-base/large and WavLM-base-plus/large are largely confined to local acoustic neighborhoods; Whisper-medium joins Whisper-large in broad intervention-general transfer; Whisper-tiny and the token-derived controls occupy intermediate regimes. Figure 5 therefore uses wav2vec2-large encoder layer 13 and Whisper-large decoder_mean layer 13 because they make the local-versus-global distinction maximally legible, not because the effect exists only in those two states. Both are large-family variants at the same nominal layer index, but the comparison remains descriptive rather than a causal test of architecture or objective.

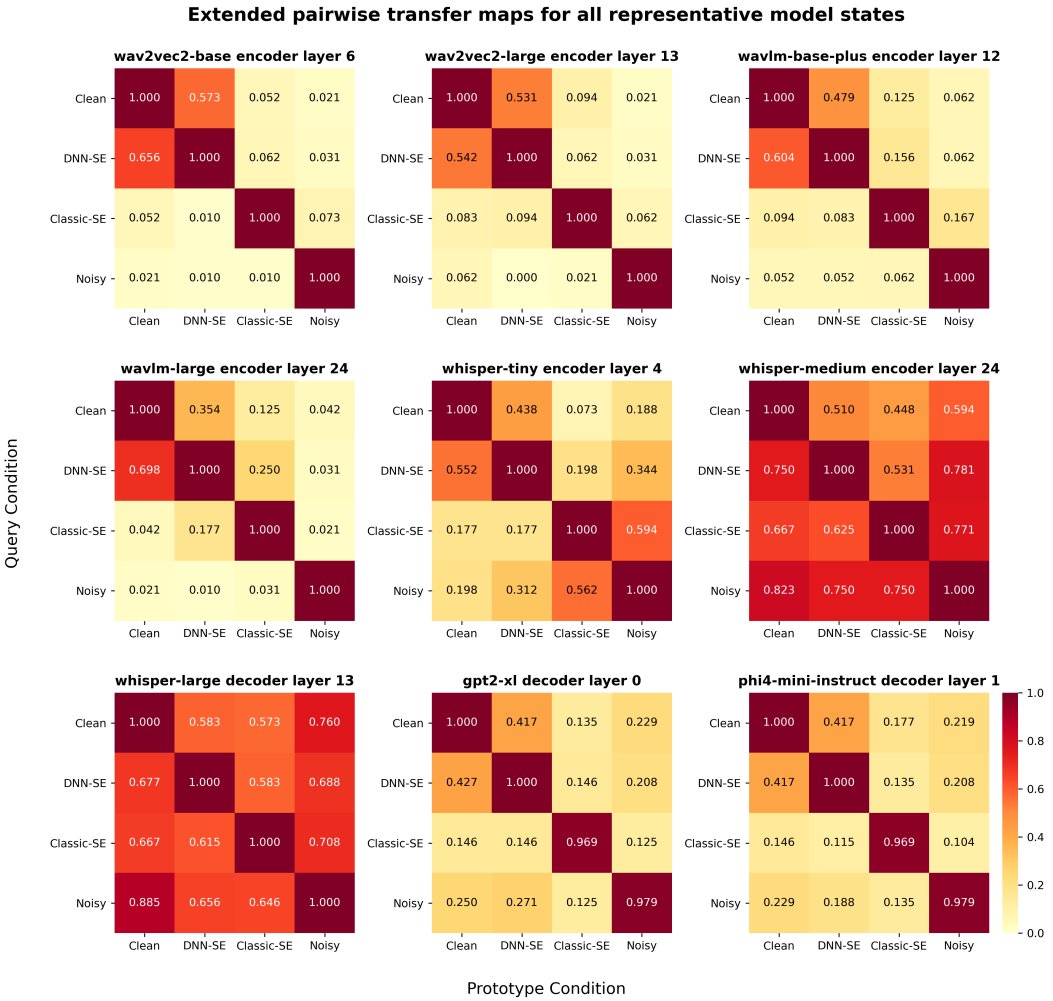


Figure A3: Pairwise transfer (Top-1 accuracy) between all condition pairs for the representative layer/stream of each model. Rows denote query condition; columns denote prototype condition; the diagonal equals 1 by construction. Panels, left-to-right and top-to-bottom: wav2vec2-base, wav2vec2-large, WavLM-base-plus, WavLM-large, Whisper-tiny, Whisper-medium, Whisper-large, GPT-2 XL, and Phi-4-mini-instruct. The broader panel shows that local trapping is not unique to wav2vec2-large and broad transfer is not unique to Whisper-large.

B Brain Regions

B.1 Bilateral anatomical regions

The broader dataset contains bilateral anatomical ROIs spanning auditory, temporal, parietal, frontal, and sensorimotor cortex. The primary four-system analysis aggregates these ROIs into a pre-specified speech-processing hierarchy; the full anatomical map is shown below for transparency.

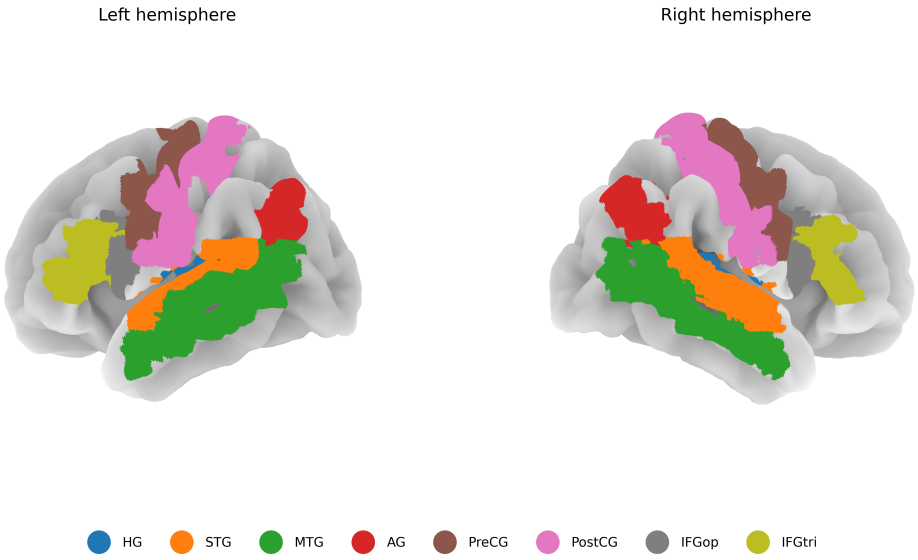


Figure B1: Bilateral anatomical ROIs in the broader fMRI dataset. IFG comprises bilateral IFGtri and IFGoper; complete system definitions are described in the main Methods and extended Methods below.

C Extended Methods

C.1 Experimental Setup, fMRI Acquisition

Twenty-five healthy subjects participated in the fMRI experiment. Each participant heard 96 spoken sentences, with 24 trials assigned to each of four acoustic conditions: Clean, DNN-SE, Classic-SE, and Noisy speech. Sentences contained approximately 10 words and lasted approximately 3 s. Noisy stimuli were generated using stationary speech-shaped noise at -3 dB SNR. DNN-SE stimuli were generated using SEMamba, whereas Classic-SE stimuli used the minimum mean-square-error log-spectral amplitude estimator. Each trial comprised a 1-s fixation period, a 3-s auditory-presentation interval, a second 1-s fixation period, and a 1.5-s behavioral-response period. Participants rated perceived quality on five-point scales and completed a two-alternative forced-choice comprehension question.

Structural and functional images were acquired using a Siemens MAGNETOM Skyra 3 T MRI scanner. T1-weighted MP-RAGE anatomical images were acquired at 1-mm isotropic resolution. Functional images were acquired using gradient-echo EPI with TR $= 2000\,\mathrm{ms}$, TE $= 24\,\mathrm{ms}$, flip angle=90 degrees, field of view $220\times220\,\mathrm{mm}^2$, 38 slices, and voxel dimensions $3.4\times3.4\times4.0\,\mathrm{mm}^3$. SPM12 preprocessing included slice-timing correction, rigid-body motion correction, structural-functional coregistration, normalization to MNI space, and 4-mm FWHM Gaussian smoothing. Single-trial beta estimates were obtained with a general linear model containing one 3-s auditory-event regressor per trial, convolved with the canonical hemodynamic response function, plus six rigid-body motion regressors.

For participant s, condition c, sentence i, and cortical ROI r, the single-trial multivoxel beta pattern was written as:

$$\mathbf{X}^{\text{fMRI}}_{s,c,i,r} \in \mathbb{R}^{V_r} \tag{C1}$$

where V_r is the number of voxels in ROI r. The 24 trial-level patterns within condition c were concatenated as:

$$\mathbf{X}^{\text{fMRI}}_{s,c,r} = [\mathbf{X}^{\text{fMRI}}_{s,c,1,r}, \ldots, \mathbf{X}^{\text{fMRI}}_{s,c,24,r}] \in \mathbb{R}^{V_r \times 24} \tag{C2}$$

Thus, each participant contributed a condition-specific response matrix for every ROI and acoustic condition. The primary cortical systems were bilateral Heschl's gyrus; temporal-parietal cortex (STG, MTG, AG); sensorimotor cortex (PreCG, PostCG); and IFG (IFGtri, IFGoper).

C.2 Frozen model feature extraction

All auditory stimuli were converted to mono, resampled to $16\,\text{kHz}$, and truncated or zero-padded to a maximum duration of $4\,\text{s}$ ($64,000$ waveform samples). Pretrained parameters were frozen, models were run in evaluation mode, and no model was fine-tuned on the experimental stimuli, behavioral measurements, or fMRI data. The direct-audio panel comprised wav2vec2-base, wav2vec2-large, WavLM-base-plus, WavLM-large, and Whisper-tiny/medium/large encoders. Whisper decoder states and GPT-2 XL/Phi-4-mini-instruct states were treated as secondary representation streams.

For a speech encoder, the hidden activation at layer l for sentence i in condition c was:

$$\mathbf{H}^{\text{enc}}_{c,i,l} \in \mathbb{R}^{T_{c,i,l} \times D_l} \tag{C3}$$

where $T_{c,i,l}$ is the number of valid temporal frames and D_l is the layer width. Padding frames were excluded. Valid hidden states were mean-pooled to obtain one sentence representation:

$$\mathbf{X}^{\text{enc}}_{c,i,l} = \frac{1}{T_{c,i,l}} \sum_{t=1}^{T_{c,i,l}} \mathbf{H}^{\text{enc}}_{c,i,l}[t,:] \in \mathbb{R}^{D_l} \tag{C4}$$

For Whisper decoder states, let $K_{c,i}$ denote the number of valid generated tokens. We retained two distinct sentence-level streams. The `decoder_mean` representation averaged valid token states:

$$\mathbf{X}^{\text{dec_mean}}_{c,i,l} = \frac{1}{K_{c,i}} \sum_{k=1}^{K_{c,i}} \mathbf{H}^{\text{dec}}_{c,i,l}[k,:] \tag{C5}$$

The `decoder_last` stream used the final valid generated-token state:

$$\mathbf{X}^{\text{dec_last}}_{c,i,l} = \mathbf{H}^{\text{dec}}_{c,i,l}[K_{c,i},:] \in \mathbb{R}^{D_l} \tag{C6}$$

Text-model controls received token sequences derived from the speech-recognition pathway rather than the identical raw waveform. Their states are therefore interpreted as secondary linguistic controls. For `decoder_mean` analyses, valid token states were averaged using the same form as Eq. C5. This input-pathway difference is why text-model controls are not treated as direct competitors for acoustic robustness.

C.3 Mathematical definition of canonical MFTMA

We used the anchor-point mean-field theoretic manifold analysis (MFTMA) formulation of Chung et al. (2018) and Stephenson et al. (2019). Unlike the condition-level formulation in the accompanying Clean-versus-Noisy analysis, the present model benchmark is canonical: each of $P = 96$ sentence identities is one object manifold, and the $K = 4$ acoustic conditions are samples of within-object nuisance variation. For sentence i at model layer l:

$$\mathcal{M}_i = \{\mathbf{X}_{i,\text{Clean}}, \mathbf{X}_{i,\text{DNN-SE}}, \mathbf{X}_{i,\text{Classic-SE}}, \mathbf{X}_{i,\text{Noisy}}\}, i = 1, \ldots, 96 \tag{C7}$$

Let $\mathbf{X}_i = [x_{i,1}, \ldots, x_{i,K}] \in \mathbb{R}^{D_l \times K}$, with $K = 4$. The manifold center is:

$$\mu_i = \frac{1}{K} \sum_{k=1}^{K} X_{i,k} \tag{C8}$$

Representations were centered and normalized relative to the Euclidean norm of the manifold center:

$$\bar{X}_i = \frac{\mathbf{X}_i - \mu_i \mathbf{1}^T}{\|\mu_i\|_2} \tag{C9}$$

A QR decomposition provides an orthonormal basis for the affine subspace spanned by each centered sentence manifold. An affine center coordinate equal to 1 is appended before applying the replica mean-field solver. At zero classification margin ($\kappa = 0$), for standard Gaussian field $\mathbf{T} = (\mathbf{t}, \mathbf{t}_0)$, the solver returns the centered anchor coordinate $\tilde{\mathbf{s}}_i(\mathbf{T})$, the point on manifold i that constrains the linear-separation solution.

The squared effective manifold radius is:

$$R_M^2 = \left\langle \|\tilde{\mathbf{s}}_i(\mathbf{T})\|_2^2 \right\rangle_{i,\mathbf{T}} \tag{C10}$$

Define $\hat{\mathbf{s}}_i(\mathbf{T}) = \tilde{\mathbf{s}}_i(\mathbf{T})/||\tilde{\mathbf{s}}_i(\mathbf{T})||_2$. The effective manifold dimension is:

$$D_M = \langle (\mathbf{t}^T \hat{\mathbf{s}}_i(\mathbf{T}))^2 \rangle_{i,\mathbf{T}}, \quad \hat{s}_i(T) = \frac{\tilde{\mathbf{s}}_i(\mathbf{T})}{\|\tilde{\mathbf{s}}_i(\mathbf{T})\|_2} \tag{C11}$$

R_M characterizes the effective within-sentence extent induced by acoustic intervention, whereas D_M characterizes the number of anchor-point directions contributing to the separating solution. Because each sentence manifold contains four acoustic samples, D_M is interpreted as sampled intervention-subspace complexity rather than the unrestricted dimensionality of all possible acoustic transformations. Functionally, R_M therefore captures perturbation spread within identity, while D_M captures the complexity of the perturbation subspace.

At $\kappa = 0$, inverse classification capacity is estimated from the same anchor-point solutions:

$$\alpha_M^{-1}(0) = \langle \frac{[-t_0 - \mathbf{t}^T \tilde{\mathbf{s}}_i(\mathbf{T})]_+^2}{1 + \|\tilde{\mathbf{s}}_i(\mathbf{T})\|_2^2} \rangle_{i,\mathbf{T}}, \quad [x]_+ = \max(x, 0) \tag{C12}$$

In canonical MFTMA, α_M is the maximal load of randomly labeled object manifolds that can be linearly separated per feature dimension. Because this study contains 96 independently labeled sentence manifolds, α_M has the canonical interpretation of inter-sentence manifold separability. We therefore interpret α_M, R_M, and D_M jointly: separability, within-identity spread, and intervention-subspace complexity are complementary descriptors of untangling, and none is assumed a priori to be functionally sufficient.

C.4 Four-fold LOCO retrieval

The functional retrieval analysis asks whether sentence identity remains recoverable when one acoustic condition is unseen during prototype construction. First, every representation was L2-normalized:

$$\tilde{\mathbf{h}}_{i,c} = \frac{\mathbf{h}_{i,c}}{\|\mathbf{h}_{i,c}\|_2} \tag{C13}$$

For held-out condition c^*, the prototype for sentence i was the mean of the other three normalized acoustic versions, followed by a second L2 normalization:

$$\mathbf{p}_{i,-c^*} = \text{normalize}[\frac{1}{3}\sum_{c\neq c^*}\tilde{\mathbf{h}}_{i,c}] \tag{C14}$$

The held-out query $\mathbf{q}_{i,c^*} = \tilde{\mathbf{h}}_{i,c^*}$ was compared with all 96 sentence prototypes using cosine similarity:

$$\mathbf{S}^{c^*}_{ij} = \mathbf{q}^T_{i,c^*}\mathbf{p}_{j,-c^*} \tag{C15}$$

Top-1 retrieval was correct when the most similar prototype had the same sentence identity:

$$\text{Acc}_{c^*} = \frac{1}{96}\sum_{i=1}^{96}\mathbf{1}_{\arg\max_j \mathbf{S}^{c^*}_{ij}=i} \tag{C16}$$

The four folds held out Clean, DNN-SE, Classic-SE, and Noisy in turn. The primary LOCO score was:

$$\text{Acc}_{\text{LOCO}} = \frac{1}{4}\sum_{c^*\in C}\text{Acc}_{c^*}, \quad C = \{\text{Clean}, \text{DNN-SE}, \text{Classic-SE}, \text{Noisy}\} \tag{C17}$$

Chance Top-1 accuracy is $\frac{1}{96} = 1.04\%$. Mean reciprocal rank is a secondary metric. Sentence-identity bootstrap resampling keeps all four acoustic realizations of a sampled sentence together. Pairwise transfer diagnostics use one condition as the query and another as the prototype condition; these directional 4×4 matrices reveal whether a model is locally invariant only between acoustically neighboring conditions or broadly invariant across intervention boundaries.

C.5 LOCO CORTICAL ENCODING AND INFERENCE

The cortical-generalization analysis uses the same train-three/test-one condition structure as retrieval. For participant s, model representation m, layer l, cortical system r, and held-out condition c^*, $X_{\text{train}} \in \mathbb{R}^{72\times D_l}$ contains model features from the three training conditions and $Y_{\text{train}} \in \mathbb{R}^{72\times V_r}$ contains the corresponding single-trial fMRI beta patterns. Ridge regression estimates the linear encoding matrix:

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \|\mathbf{Y}_{\text{train}} - \mathbf{X}_{\text{train}}\mathbf{W}\|^2_F + \lambda\|\mathbf{W}\|^2_F \tag{C18}$$

All centering, scaling, and optional dimensionality reduction are performed strictly within the training conditions. For the 24 trials in the held-out condition:

$$\hat{\mathbf{Y}}_{\text{test}} = \mathbf{X}_{\text{test}}\hat{\mathbf{W}} \tag{C19}$$

For voxel v, held-out predictivity is the Pearson correlation between observed and predicted trial responses:

$$r_v = \text{corr}(\hat{\mathbf{Y}}_{\text{test}}[:,v], \mathbf{Y}_{\text{test}}[:,v]), \quad z_v = \text{atanh}(r_v) \tag{C20}$$

Voxelwise Fisher-z values are averaged within anatomical ROIs. Constituent ROIs are then combined with equal weighting to form the pre-specified cortical systems. IFG comprises bilateral IFGtri and IFGoper; temporal-parietal cortex comprises bilateral STG, MTG, and AG; sensorimotor cortex comprises bilateral PreCG and PostCG; Heschl's gyrus is analyzed bilaterally as the early auditory system. Fold-level system scores are averaged in Fisher-z space within participant:

$$\bar{z}_{s,m,l,r} = \frac{1}{4} \sum_{c^* \in C} z_{s,m,l,r,c^*} \tag{C21}$$

Participant is the independent group-level unit. Above-zero predictivity is evaluated using participant-level sign-flip inference, and pre-specified paired participant-level contrasts are used for representation-family, stream, or stage comparisons. Multiple-comparison correction is applied within the pre-specified contrast family. IFG is the primary synthesis target because it was independently identified by the human intervention analysis before model-to-brain encoding was performed; Heschl's gyrus and the remaining systems are reported as complementary cortical controls.

D Geometry-to-Function Robustness and Sensitivity Analyses

D.1 Stage-level clustered inference

The primary geometry-to-function analysis uses the seven direct-audio encoders only: Whisper-tiny/medium/large encoders, wav2vec2-base/large, and WavLM-base-plus/large. Early/Middle/Late summaries yield 21 observations from seven independent model clusters. Spearman associations with four-fold LOCO Top-1 are evaluated using rank regression with model-clustered standard errors and BH-FDR correction across the three descriptors. Capacity is positively associated with retrieval ($\rho = .643, p = .0080, q = .0121$), radius is negatively associated ($\rho = -.604, p = .0064, q = .0121$), and dimension is negatively associated ($\rho = -.573, p = .0214, q = .0214$). In the complementary $10,000$-resample model-cluster bootstrap, 95% intervals for the raw coefficients are $[-.026, .912]$, $[-.884, -.034]$, and $[-.913, .048]$, respectively. The broad capacity and dimension intervals reflect the small number of independent clusters and motivate interpreting the three descriptors as a coordinated regime rather than finely ranking their strengths.

D.2 Robustness to retrieval metric

Mean reciprocal rank provides a secondary retrieval criterion that uses the full ranking rather than Top-1 correctness alone. The stage-level pattern is preserved: capacity is positively associated with MRR ($\rho = .586$, model-clustered $p = .0188$, $q = .0282$), while radius ($\rho = -.542, p = .0180, q = .0282$) and dimension ($\rho = -.534, p = .0335, q = .0335$) are negatively associated. Agreement across Top-1 accuracy and MRR indicates that the geometry-function result is not specific to a single retrieval summary.

D.3 Descriptor coupling and comparison of association strength

Capacity (α_M), radius (R_M), and effective dimension (D_M) were strongly coupled across the 21 direct-audio model $\times$ stage summaries (7 models $\times$ 3 stages): $r(\alpha_M, R_M) = -.948$, $r(\alpha_M, D_M) = -.975$, and $r(R_M, D_M) = .939$, with variance-inflation factors 24.63, 10.30, and 21.00, respectively, so a joint regression cannot stably estimate unique descriptor contributions. For the association-strength plot only, we orient all coefficients toward the same functional direction: capacity is displayed as ρ_{capacity}, whereas radius and dimension are displayed as $-\rho_{\text{radius}}$ and $-\rho_{\text{dimension}}$. This sign reversal is purely a plotting convention—because better retrieval is associated with larger capacity but smaller radius/dimension—and does not change the raw correlations, p values, or confidence intervals. Direct model-cluster bootstrap comparisons did not establish a stronger association for radius than capacity (95% CI $[-.176, .140]$) or dimension (95% CI $[-.224, .201]$). The per-layer coefficients are descriptive because neighboring layers are correlated. Together, the results support a coordinated geometric regime rather than a uniquely dominant descriptor.

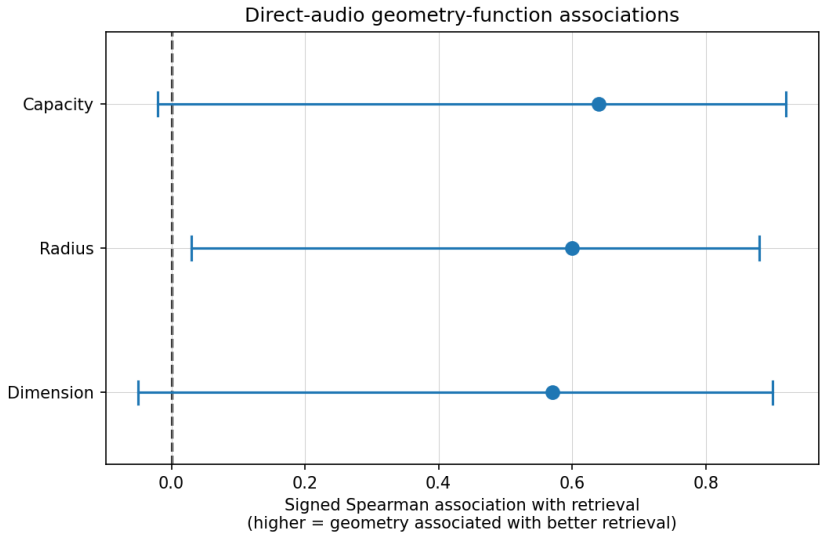


Figure D1: Direction-aligned association strength between sentence-manifold geometry and held-out retrieval. Spearman associations are computed across the 21 direct-audio model $\times$ stage summaries. Capacity is plotted as its raw positive association with retrieval; radius and effective dimension are multiplied by -1 for display because their raw associations are negative. Thus, moving right uniformly means a stronger association with better retrieval; the transformation does not change statistical significance or whether a confidence interval crosses zero. Horizontal bars show 95% model-cluster bootstrap intervals. The interval excludes zero for radius but overlaps zero for capacity and dimension, reflecting uncertainty from seven independent model clusters. Direct bootstrap contrasts do not establish a uniquely stronger descriptor.

D.4 Architecture-specific layerwise trajectories

Within-model layerwise associations reveal that the pooled radius-retrieval relation is not a universal consequence of depth. Radius is strongly negatively associated with Top-1 retrieval within Whisper-tiny ($\rho = -.900$), Whisper-medium ($\rho = -.866$), Whisper-large ($\rho = -.888$), and WavLM-base-plus ($\rho = -.887$), but is positive within wav2vec2-base ($\rho = .637$), wav2vec2-large ($\rho = .196$), and WavLM-large ($\rho = .319$). These coefficients are descriptive because layers are correlated, but their sign heterogeneity is scientifically informative: different architectures and training objectives exhibit different depth-wise coordination between geometry and retrieval. The main-text claim is therefore about representational regimes, strongest in Whisper, rather than a universal rule that reducing radius at every layer must improve retrieval.

Table D1: Within-model layerwise radius-retrieval associations for direct-audio encoders (descriptive only).

Model	**Layers**	**Spearman rho (R_M vs Top-1)**
Whisper-tiny	5	$-.900$
Whisper-medium	25	$-.866$
Whisper-large	33	$-.888$
WavLM-base-plus	13	$-.887$
WavLM-large	25	$+.319$
wav2vec2-base	13	$+.637$
wav2vec2-large	25	$+.196$

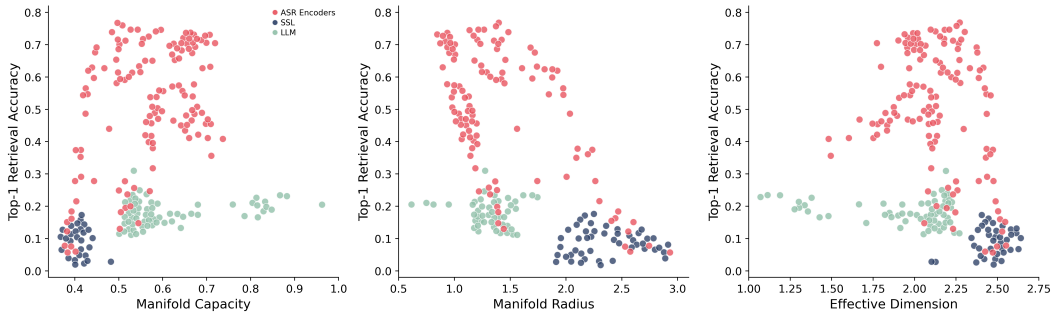


Figure D2: Layerwise geometry–function relationships across evaluated model representations. Across individual layers, higher retrieval is associated descriptively with greater capacity and with smaller radius and effective dimension (direct-audio layers: $\rho = .643$, $-.673$, and $-.583$, respectively). Adjacent layers are strongly correlated, and within-model trajectories differ across architectures; these points are therefore descriptive and are not treated as independent inferential samples.

D.5 Broader representational-class sensitivity

As a sensitivity analysis, we also pooled all 429 model-layer representations, including Whisper decoder streams and token-derived GPT-2 XL/Phi-4 controls. This broader descriptive analysis recovers the point-estimate ordering that motivated the original qualitative description: Top-1 retrieval correlates most strongly with radius ($\rho = -.684$), followed by capacity ($\rho = .589$) and dimension ($\rho = -.496$); all nominal p values are $< 10^{-27}$. A three-predictor LMG relative-importance decomposition of the pooled joint $R^2 = .416$ assigns 40.8% of explained variance to radius, 34.8% to capacity, and 24.4% to dimension. We do not use this ranking as primary evidence because the pooled sample mixes representational classes with different information pathways and treats correlated layers as observations. Its appropriate interpretation is cross-class sensitivity: radius has the largest pooled point estimate, while the primary direct-audio, cluster-aware analysis supports a coordinated regime without a decisive winner.

Table D2: Geometry-to-retrieval associations across primary and sensitivity analyses.

Analysis	**Capacity** α_M	**Radius** R_M	**Dimension** D_M
Primary: direct-audio stage Top-1	$\rho = .643, q = .012$	$\rho = -.604, q = .012$	$\rho = -.573, q = .021$
Robustness: direct-audio stage MRR	$\rho = .586, q = .028$	$\rho = -.542, q = .028$	$\rho = -.534, q = .033$
Sensitivity: all 429 layers	$\rho = .589$	$\rho = -.684$	$\rho = -.496$

D.6 Capacity nonlinearity is not a primary conclusion

In the pooled 429-representation analysis, adding a quadratic capacity term improves fit over a linear-only model ($\Delta R^2 = .063, p < 10^{-8}$) with an estimated vertex near $\alpha_M = .711$. This effect is substantially weaker in the primary direct-audio encoder population ($\Delta R^2 = .016, p = .0097$), and its estimated vertex ($\alpha_M = .638$) lies above the maximum observed direct-audio capacity ($\alpha_M = .596$). The apparent downturn therefore occurs outside the observed primary range. We consequently do not describe capacity as non-monotonic in the main text. The supported direct-audio result is a positive capacity-retrieval association, while secondary representation classes show that high geometric separability alone is not sufficient for acoustic intervention-general identity.