COMPUTATIONAL DISSECTION OF SOUND-TO-MEANING HIERARCHIES IN SPEECH FOUNDATION MODELS AND THE HUMAN BRAIN

Anonymous authors
Paper under double-blind review

ABSTRACT

Whether speech foundation models and the human brain abstract sound into meaning along a convergent route remains unresolved: prior comparisons rely on brain scores, which are hampered by low-level confounds and do not address compositional meaning. We introduce a spoken multi-definition-to-target paradigm in which acoustic, phonetic, and lexical form vary while meaning converges: differently worded definitions, read by multiple speakers, converge on a targeted semantic concept. We dissect eleven speech foundation models spanning masked prediction (HuBERT, WavLM), contrastive prediction (wav2vec 2.0), weak supervision (Whisper), and audio-LLM encoders (Qwen2-Audio, Qwen3-Omni), as well as intracranial recordings from 29 patients performing the same task. Layer-wise decoding and representational geometry analyses in the models reveal cascaded dynamics: speaker identity separates first, followed by surface form and phonetics, each persisting across most layers, whereas lexico-syntactic category and target semantics emerge last, in narrow bands whose depth is set by the training objective, and the semantic target is decodable from held-out definitions. Temporal response function encoding with interpretable speech and linguistic features shows that continuous acoustic and prosodic features trade off against categorical linguistic features across depth, whereas cortex abstracts along a graded, nested hierarchy from core auditory regions to distributed frontal, parietal, and medial hubs, in which part of speech (POS) behaves as a lexical property and constituency structure engages posterior temporal and parietal cortex. Temporal generalization decoding shows that the two systems share dynamics for phonetic and statistical word-level information but diverge for rule-based structure and sentence-level semantic categories. Low-level information is thus spread across depth in the models but localized in cortex, and higher-order structure shows the reverse pattern. By testing speech models for brain-like abstraction rather than brain-like prediction, this work establishes a framework that locates where and how artificial and biological speech networks converge and diverge, and points to the model layers that could serve neuroprostheses decoding meaning as well as speech.

1 INTRODUCTION

During speech comprehension, the brain transforms a continuous acoustic waveform into discrete linguistic units and, ultimately, into meaning. Intracranial recordings with electrocorticography (ECoG) and stereo-electroencephalography (sEEG) have mapped much of this transformation at millimeter and millisecond resolution (Gwilliams et al., 2025a): populations tuned to phonetic features (Mesgarani et al., 2014), to speaker-normalized intonation (Tang et al., 2017), and to sustained versus transient responses to acoustic onsets (Hamilton et al., 2018; Forseth et al., 2020). Primary and nonprimary auditory cortex operate in parallel (Hamilton et al., 2021), preserving lower-level information as higher-order structures accrue (Keshishian et al., 2023). Meaning is assembled by a far more distributed network: picture naming recruits a lexico-semantic hub in middle fusiform gyrus (Forseth et al., 2018), modality-invariant lexical access engages inferior frontal, orbitofrontal, medial parietal, and medial temporal cortex (Murphy et al., 2023), and phrase composition draws on posterior superior temporal sulcus (Murphy et al., 2022; 2024).

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.

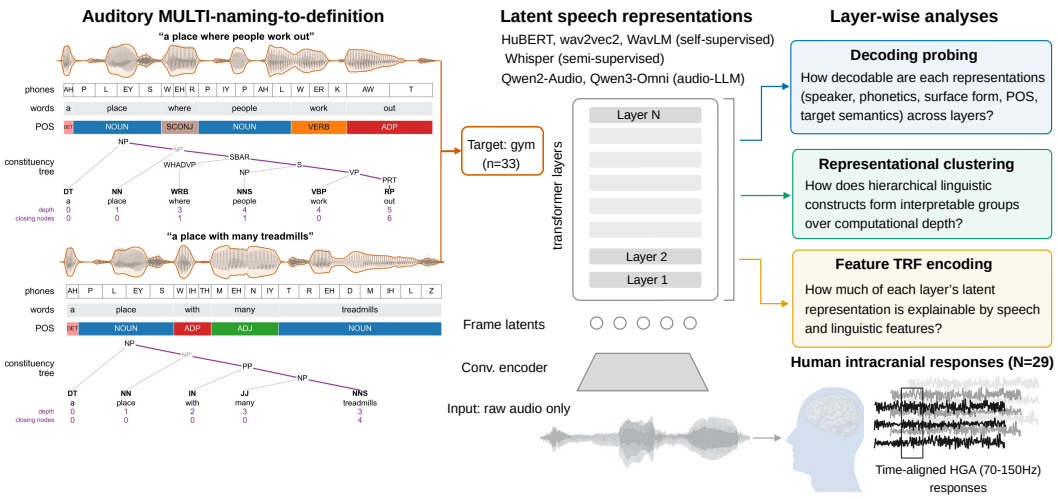

Figure 1: Analytical paradigm. Left: in the spoken multi-definition-to-target task, target concepts are each described by several definitions and speakers; the target word is never spoken. Two example definitions are annotated with linguistic features. Middle: speech foundation models receive raw audio only; time-aligned embeddings are extracted for analysis. Right: decoding probes, representational clustering, and feature-based TRF encoding, applied to pooled states and to intracranial recordings from 29 patients performing the same task.

Speech foundation models ascribe a similar sense of "meaning" to acoustic representations but differ chiefly in supervision: HuBERT and WavLM predict masked pseudo-labels (Hsu et al., 2021; Chen et al., 2022), wav2vec 2.0 solves a contrastive task (Baevski et al., 2020), Whisper is weakly supervised on transcripts (Radford et al., 2023), and audio-LLMs (Qwen2-Audio, Qwen3-Omni) couple an audio encoder to a language model (Chu et al., 2024; Xu et al., 2025). Self-supervised speech models (S3Ms) never observe text, and existing studies ask how linguistic structure emerges in them without explicit supervision (de Heer Kloots et al., 2026b; Choi et al., 2024; He et al., 2025a).

Efforts to relate the two systems score how well model embeddings predict neural activity (Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022) or fine-tune models on brain responses (Moussa et al., 2025; Vattikonda et al., 2025; Xiao et al., 2026). Such scores are hard to interpret: embeddings entangle levels (He et al., 2025b), low-level speech features explain much of the alignment (Oota et al., 2024; Shimizu et al., 2025), and whether alignment reflects prediction, linguistic abstraction, or the intrinsic dimensionality of neural activity remains contested (Antonello & Huth, 2024; Cheng et al., 2026). In naturalistic paradigms (Hamilton & Huth, 2020), such aggregate scores cannot reveal which level drives a brain-model correspondence, and composed meaning (Toneva et al., 2022) has not been examined in speech models at the temporal resolution of comprehension. An encoding score also underdetermines the relation it reports: a linear map from model states to neural responses shows that the two spaces are predictively related, not that they share a geometry, and because both are driven by the same stimulus, alignment can be inherited from shared input rather than shared computation (Guest & Martin, 2023; Schönmann et al., 2026). Comparisons computed within each system, on targets defined by linguistic properties rather than model features, involve no such map.

We dissect both hierarchies with a spoken multi-definition-to-target stimulus design (Figure 1) that extends definitional naming (Forseth et al., 2018), wherein differently worded definitions read by five speakers converge on 33 target concepts. We apply layer-wise decoding, representational geometry, temporal generalization, and interpretable encoding models to six model families and to intracranial recordings from 29 patients, contributing (i) **an interpretable framework that holds meaning fixed while form varies**, with matched human intracranial recordings; (ii) **a dissectible linguistic hierarchy in computational substrates**, in which lower-order codes persist across depth while lexico-syntactic and semantic codes emerge last in narrow, objective-dependent bands; (iii) **shared and divergent dynamics between cortical and artificial networks**: cortex and models share dynamics for phonetic and word-statistical information, not for rule-based structure or target semantic

categories; and (iv) **an inverted organization of feature representations**: low-level and categorical linguistic features segregate across model depth, whereas cortex abstracts along a gradient from auditory cortex to distributed hubs, so low-level information is distributed across model depth but localized in cortex, and the reverse holds for higher-order structure.

2 Related Work

Model-brain alignment without compositional meaning. Encoding models map model representations onto neural activity (Huth et al., 2016): for text language models, brain predictivity tracks next-word prediction (Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022) and model scale (Antonello et al., 2023); for speech, self-supervised representations predict cortical responses (Millet et al., 2022; Vaidya et al., 2022), successive stages map onto the ascending auditory pathway (Li et al., 2023; Tuckute et al., 2023), and model and cortical feature hierarchies converge (Mischler et al., 2024; Keshishian et al., 2026). Previous work using Whisper embeddings mapped a coarse sensorimotor-to-language network in the human brain with ECoG (Goldstein et al., 2025). These analyses regress unconstrained embeddings onto neural data, so a correspondence can be carried by features shared across levels (Oota et al., 2024), which disentanglement (He et al., 2025b) and interpretable feature spaces (Benara et al., 2024; Shimizu et al., 2025) address only post hoc. None of this work isolates compositional meaning, studied in the brain with controlled contrasts from minimal phrases (Murphy et al., 2022) to definitions converging on a referent (Murphy et al., 2023) and in text by isolating supra-word meaning (Toneva et al., 2022), but never dissociated from form in a model-brain comparison. Inferring shared mechanism from predictive alignment *affirms the consequent* (Guest & Martin, 2023): a model implementing the cortical operation would predict cortex, but so would any system driven by the same input. *Divergence* is therefore the discriminating outcome: if a model implemented the same operation at a level, the dynamics of decodability at that level should match, and where they do not, the model is excluded as an implementation of that operation, whereas a match remains compatible with common input alone.

Probing and representational dynamics within S3Ms. Probing reads a property from a representation with a simple classifier (Hupkes et al., 2018; Hewitt & Liang, 2019). In speech models, it has mapped acoustic, phonetic, speaker, and word-level information (Pasad et al., 2021; 2023; 2024; Pasad, 2025) as well as its emergence during training (de Heer Kloots et al., 2026a). Previous work, however, targeted properties transparent in the input, which a probe can read from form, rather than composed meaning (Choi et al., 2024). Closest to our aim, minimal-pair probing shows that speech models organize grammatical and conceptual contrasts hierarchically across layers (He et al., 2025a), and residual disentanglement isolates higher-order components of language-model embeddings that better predict cortex (He et al., 2025b); neither, however, dissociates meaning from the surface form that carries it, or compares the resulting codes with neural dynamics. Temporal generalization (King & Dehaene, 2014) compares how representations evolve: in MEG during story listening, each linguistic level is maintained by a dynamic code evolving at a level-specific speed (Gwilliams et al., 2025b).

3 Methods

3.1 The multi-definition-to-target paradigm

The stimuli are spoken definitions of 33 target concepts, four per concept, each read by different male and female speakers; the target word never occurs in any definition (Figure 1). Each claim is tested under the split that isolates it: (i) GroupKFold over definitions (five folds) for speaker, phonetic, and word-level properties and for all encoding models; (ii) LeaveOneSpeakerOut for 132-way recognition of surface form; (iii) LeaveOneDefinitionOut, which holds out the f-th definition of every concept for all speakers in fold f, for 33-way decoding of the target.

3.2 Neural data and speech foundation models

We recorded from 29 patients undergoing intracranial monitoring for epilepsy (Appendix C) as they listened to the definitions and were prompted to name the target of each definition. Broadband high-

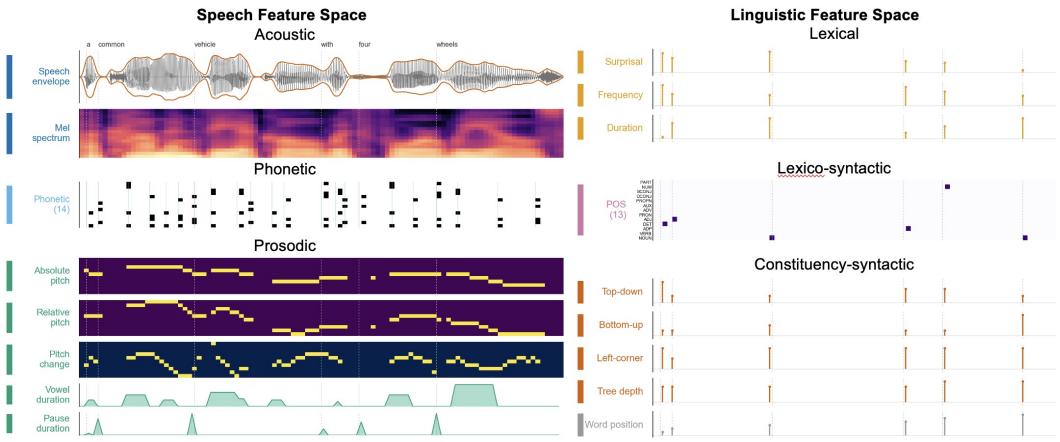

Figure 2: The stimulus feature bank for one utterance ("a common vehicle with four wheels"). Speech features: envelope and 16-band mel spectrogram (acoustic); 14 binary phonological features at phone onsets (phonetic); absolute pitch, relative pitch, and pitch change in 10 bins, with vowel and pause durations (prosodic). Linguistic features: surprisal, frequency, and duration at word onsets (lexical); a 13-way part-of-speech indicator (lexico-syntactic); top-down, bottom-up, and left-corner node counts, tree depth, and word position (constituency-syntactic). Colored bars mark the variance-partition bands; a sentence-onset term enters every model.

gamma activity (HGA; 70 to 150 Hz) was downsampled to 100 Hz and z-scored against a -300 to -50 ms pre-stimulus baseline. Electrodes were localized by co-registering the pre-operative T1-weighted MRI with the post-implantation CT in AFNI and projected onto each patient's FreeSurfer cortical surface; group maps are displayed on the standard TT_N27 surface (Appendix C).

We analyze eleven frozen checkpoints from six model families (Table 1; Appendix Table A3); layer $\ell \geq 1$ is the output of block ℓ. A phone, word, or definition covering frames $\mathcal{T}$ is represented at layer ℓ by its mean state $\bar{\boldsymbol{h}}^{(\ell)} = |\mathcal{T}|^{-1} \sum_{t \in \mathcal{T}} \boldsymbol{h}_t^{(\ell)}$.

3.3 Analyses

Layer-wise decoding and representational geometry. At each layer, a linear probe (Alain & Bengio, 2016), an L2-regularized multinomial logistic regression $\hat{y} = \arg\max_k [\boldsymbol{W}\bar{\boldsymbol{h}}^{(\ell)} + \boldsymbol{b}]_k$ on standardized features ($C = 1$), reads speaker, phone, surface form, part of speech, and target concept under their splits (class counts in Table 1). We report out-of-fold accuracy against chance $1/k$ and each property's peak on relative depth ℓ/L. For representational geometry, we cluster each layer's pooled states with label-free k-means and report the normalized mutual information (NMI) between clusters and labels (Figure 4; Appendix Figures A2 and A3).

Temporal generalization across systems. A decoder trained at each time t is tested at every time t' (King & Dehaene, 2014) on phoneme-, word-, and definition-locked targets from the alignment, the parse (Qi et al., 2020), and the design (Appendix C). With training items $\boldsymbol{X}_t \in \mathbb{R}^{n \times p}$ and one-hot labels $\boldsymbol{Y} \in \{0,1\}^{n \times k}$,

$$\hat{\boldsymbol{W}}_t = \left(\boldsymbol{X}_t^\top \boldsymbol{X}_t + \alpha \boldsymbol{I}\right)^{-1} \boldsymbol{X}_t^\top \boldsymbol{Y}, \qquad \mathrm{TG}(t, t') = \mathrm{BA}\Big(\boldsymbol{y}', \arg\max_k \left[\boldsymbol{X}'_{t'} \hat{\boldsymbol{W}}_t\right]_{:,k}\Big), \tag{1}$$

with $\alpha = 10^3$, held-out items $\boldsymbol{X}'_{t'}$ under each target's split, row-wise $\arg\max$, balanced accuracy BA, and per-cell empirical chance. Neural features are a participant's channels and model features one layer's frames, both in 20 ms bins; model matrices use HuBERT Large, wav2vec 2.0 Large, and Whisper Medium at layers 19, 12, and 24, near their probing peaks (Table 1).

Encoding and variance partitioning. Both systems are encoded from one bank of nine feature families (Figure 2, Table A1). A multivariate temporal response function (mTRF; Crosse et al., 2016) predicts a response $y_c(t)$, the HGA of channel c or one of 30 principal components of a

Table 1: Peak layer-wise probe accuracy, with the best layer in parentheses (0: input embedding). Surface form is tested leave-one-speaker-out and semantics leave-one-definition-out; bold: column maximum; †: Qwen2-Audio's final block, available only after temporal pooling, is excluded.

Model	Objective	L	Speaker	Phonetic	Surface	Lexico-synt.	Semantics
Classes			5	38	132	13	33
Chance			0.20	0.03	0.01	0.08	0.03
HuBERT Base	Masked	12	1.00 (0)	0.97 (10)	1.00 (2)	0.95 (9)	0.58 (8)
HuBERT Large	Masked	24	1.00 (0)	0.97 (21)	1.00 (5)	**0.97** (18)	**0.64** (18)
WavLM Base	Masked	12	1.00 (0)	0.97 (10)	1.00 (3)	0.96 (9)	0.57 (9)
WavLM Large	Masked	24	1.00 (0)	**0.98** (21)	1.00 (7)	**0.97** (19)	0.63 (17)
wav2vec 2.0 Base	Contrastive	12	1.00 (0)	0.97 (8)	1.00 (2)	0.95 (7)	0.53 (6)
wav2vec 2.0 Large	Contrastive	24	1.00 (0)	0.97 (14)	1.00 (3)	0.96 (12)	0.59 (10)
Whisper Base	Weak ASR	6	1.00 (0)	0.95 (3)	1.00 (1)	0.92 (5)	0.52 (5)
Whisper Small	Weak ASR	12	1.00 (0)	0.96 (8)	1.00 (4)	0.96 (11)	0.58 (11)
Whisper Medium	Weak ASR	24	1.00 (0)	0.96 (13)	1.00 (8)	**0.97** (23)	**0.64** (22)
Qwen2-Audio-7B	Audio-LLM	31†	1.00 (0)	0.97 (23)	1.00 (16)	**0.97** (29)	0.60 (30)
Qwen3-Omni AuT	Audio-LLM	32	1.00 (0)	0.72 (22)	1.00 (17)	0.95 (30)	0.63 (31)

layer's states, from time-lagged features $\boldsymbol{x}_b$ of each band b,

$$\hat{y}_c(t) = \sum_b \sum_\tau \boldsymbol{w}_{c,b,\tau}^\top \boldsymbol{x}_b(t-\tau), \qquad \min_{\boldsymbol{w}_c} \left\| \boldsymbol{y}_c - \sum_b \boldsymbol{X}_b \boldsymbol{w}_{c,b} \right\|_2^2 + \sum_b \frac{\alpha}{\gamma_b} \left\| \boldsymbol{w}_{c,b} \right\|_2^2, \tag{2}$$

where $\boldsymbol{X}_b$ stacks band b's lagged features. For HGA, lags span -100 to 500 ms at 100 Hz and banded ridge (Nunez-Elizalde et al., 2019; Dupré la Tour et al., 2022) tunes one penalty per band by nested, definition-disjoint cross-validation; for models, lags span -100 to 400 ms at 50 Hz under a single penalty ($\gamma_b \equiv 1$). Variance is partitioned over five cumulative tiers (Table A2): a family's unique contribution is the held-out R^2 lost when it is removed from the tier model $\mathcal{A}_k$, $\Delta R_f^2 = R^2(\mathcal{A}_k) - R^2(\mathcal{A}_k \setminus f)$. For models, each family is instead removed from the full model. Statistical inference is in Appendix C.

4 Results

4.1 Hierarchical codes unfold uniquely in speech models

Table 1 gives peak probe accuracies. Speaker identity is at ceiling from the input embedding and surface form saturates within a few layers; phonetic and lexico-syntactic labels saturate in the middle-to-late part of each model. The semantic target reaches roughly twenty times chance in every model under leave-one-definition-out, so the target code generalizes across definitions. Peak semantic accuracy nevertheless remains below its phonetic and lexical counterparts.

Figure 3 characterizes how each probe saturates across model depth. Speaker identity, surface form, and phone identity rise early and stay near their peak thereafter (Appendix Figure A1); part of speech and target semantics rise slowly and peak in narrow bands placed by the training objective: late but not last in the masked-prediction models, at mid-depth in contrastive wav2vec2-large, and in the final layers of Whisper and the audio-LLM encoders. The same ordering holds in every model, across objectives and architectures, and semantic confusions stay within semantic neighborhoods.

Label-free representational geometry (Figure 4) mirrors the probes, and target semantics reaches the highest normalized mutual information of any level, its dissimilarity structure recovering semantic categories absent at the input, in contrast with phonetics (Appendix Figures A2 and A3). Both analyses show that lower-order representations persist across depth, whereas higher-order lexico-syntactic and semantic representations occupy narrow, objective-dependent bands.

4.2 Encoding models dissect the anatomy of speech models

Treating each layer as a multichannel response to the feature bank, Figure 5 shows a ladder: acoustic families are best expressed at the input and decay, phonetic features rise to a broad plateau, and

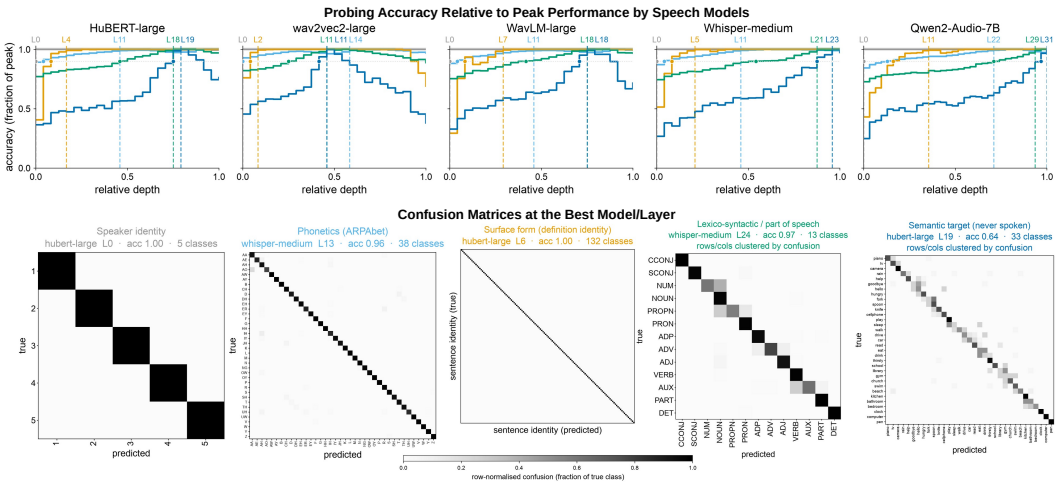

Figure 3: Depth of the decoding hierarchy. Top: probe accuracy relative to each task's peak for HuBERT-large, wav2vec2-large, WavLM-large, Whisper-medium, and Qwen2-Audio-7B; dashed lines mark peak layers (speaker, grey; phone, light blue; surface form, orange; POS, green; target semantics, dark blue). Bottom: row-normalized confusion matrices at each task's best model and layer.

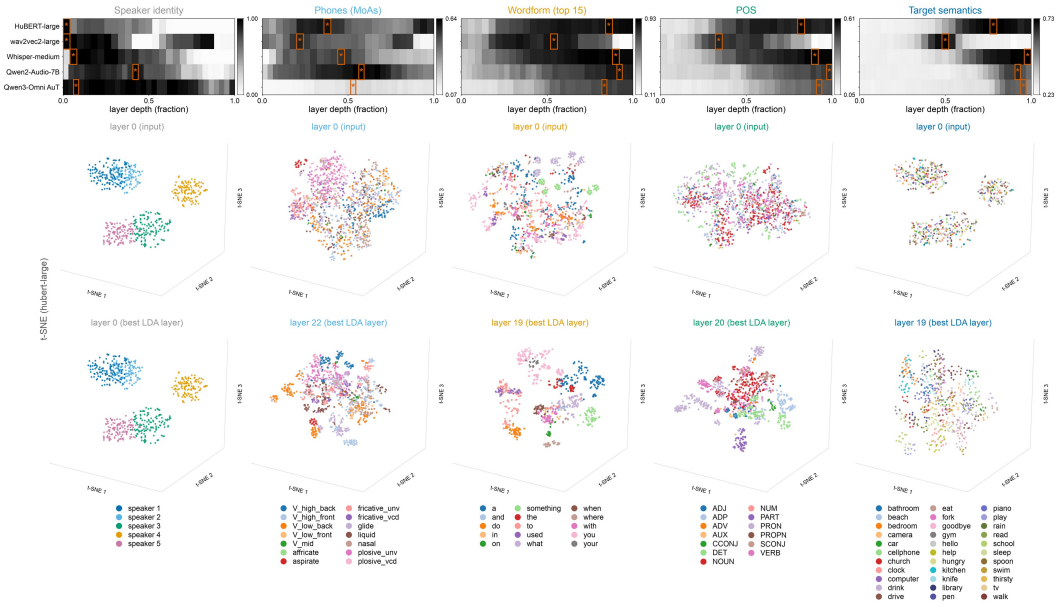

Figure 4: Representational geometry across depth. Top: normalized mutual information (NMI) between the label-free k-means partition of each layer's pooled states and the labels, across relative depth, for speaker identity, phones, the 15 most frequent word forms, POS, and target semantics; orange boxes mark the best layer. Bottom: t-SNE of HuBERT-large states at layer 0 and at each level's best linear-discriminant layer.

word-level, part-of-speech, and constituency families appear last in peaked bands placed by the objective (Appendix Figures A4 to A6). Nonetheless, higher-order families gain share as low-level families give it up but never dominate (Figure 5, middle), and S3Ms carry multiple linguistic families jointly in their mid-to-late layers, whereas Whisper carries joint representations only in its last layer (Appendix Figure A7). Figure 6 shows how the families co-vary across depth: acoustic-prosodic and symbolic (phonetics, word-level, part of speech, constituency) families segregate into distinct

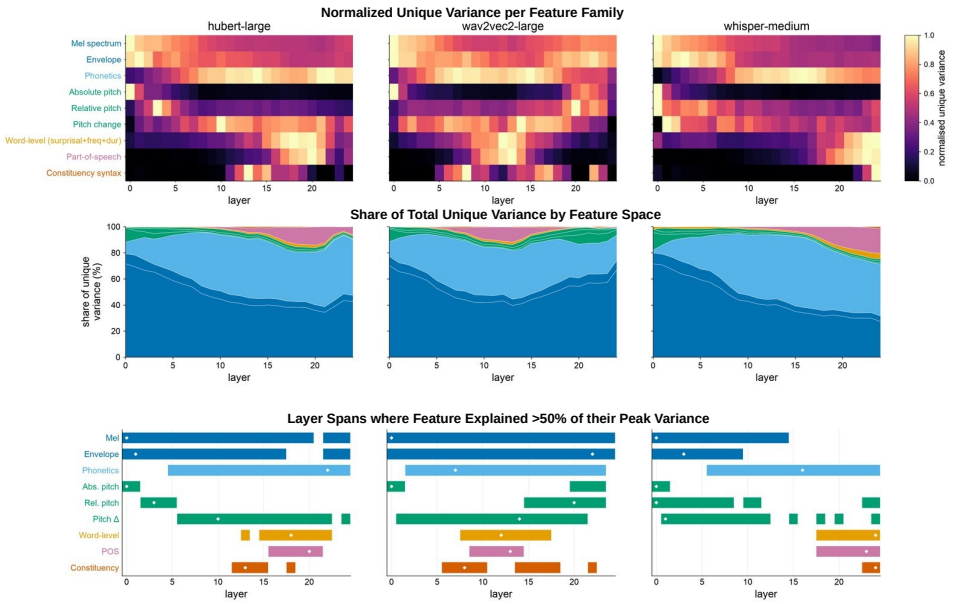

Figure 5: Layer-wise variance partitioning of model representations. Each layer's states (30 principal components) of HuBERT-large, wav2vec2-large, and Whisper-medium are predicted by a forward mTRF from the feature bank. Top: unique R^2 per family and layer, normalized within family. Middle: each family's share of the layer's unique variance. Bottom: layers where a family exceeds 50% of its peak (bars) and its peak layer (diamond).

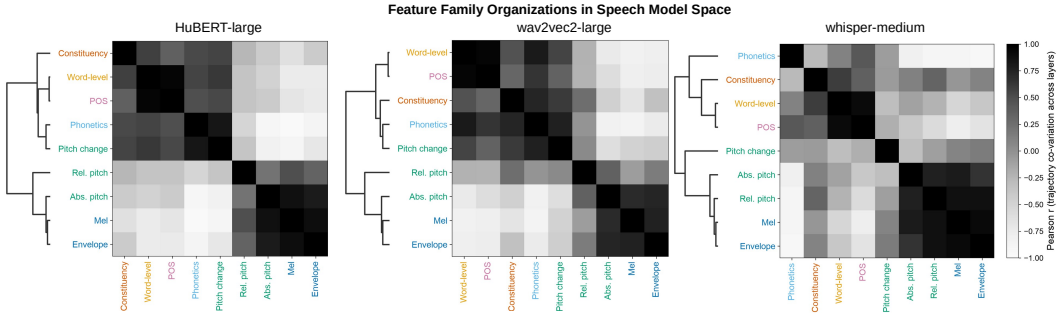

Figure 6: Co-emergence of feature families across model depth. Pearson correlation, across layers, between two families' unique-R^2 trajectories for HuBERT-large, wav2vec2-large, and Whisper-medium; rows and columns ordered by hierarchical clustering.

blocks, positively correlated within and strongly negatively correlated between, so across depth the models trade surface form off against structure.

4.3 Convergence and Divergence via Temporal Generalization

To test whether cortex and the models share the dynamics of each level's code, temporal generalization decoding is applied to phoneme-, word-, and sentence-aligned model states and HGA (Figure 7). For lower-level speech representations, cortex resembles the models: every phoneme-level target and the surface-level lexical features are diagonal-dominant in both systems, and the speaker code matches when aligned to sentence offset. They diverge on compositional structure: node closure and tree depth form broad, sustained late blocks in cortex that the models carry as thin diagonals or not at all. The similarity panel quantifies the split: phoneme-level and statistical word-level targets are shared, node closure, tree depth, and the cross-concept semantic category are not, and the target concept is moderately shared, so the divergence concerns rule-based structure and sentence-level semantics, not meaning altogether (per-target decoder maps: Appendix Figures A8 to A10).

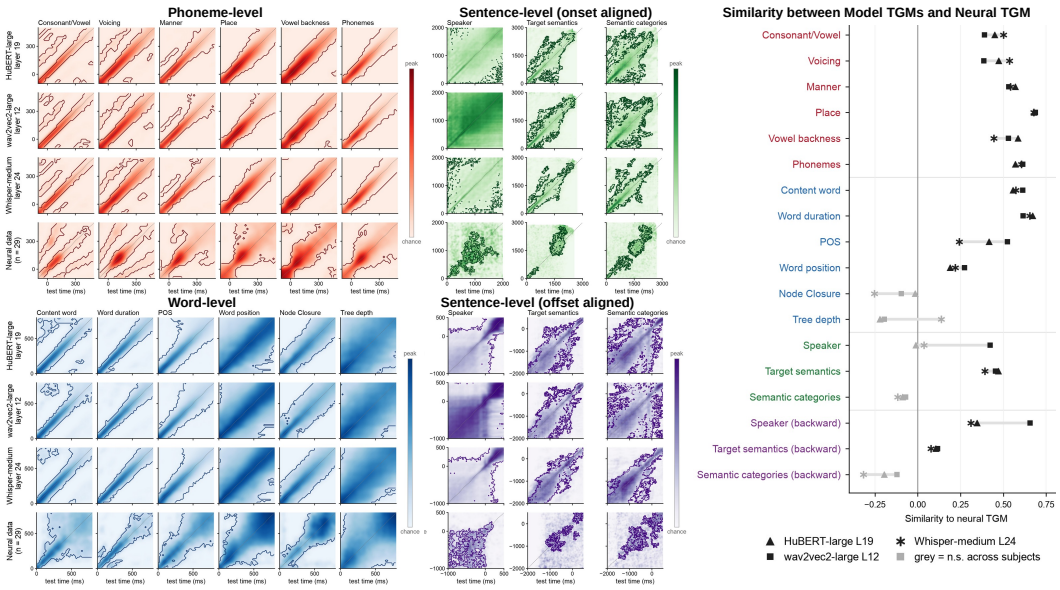

Figure 7: Temporal generalization matrices (TGMs) in speech models and cortex. Decoders trained at each time (vertical) are tested at every time (horizontal); color is balanced accuracy from per-cell chance to each panel's peak. Rows: HuBERT-large (layer 19), wav2vec2-large (12), Whisper-medium (24), and the cross-participant median. Left: phoneme-level (orange) and word-level (blue) targets; middle: sentence-level targets aligned to onset (green) and offset (purple). Contours: above chance in every fold (models) or in a significant majority of participants. Right: similarity of each model's TGM to the neural TGM (grey: not significant).

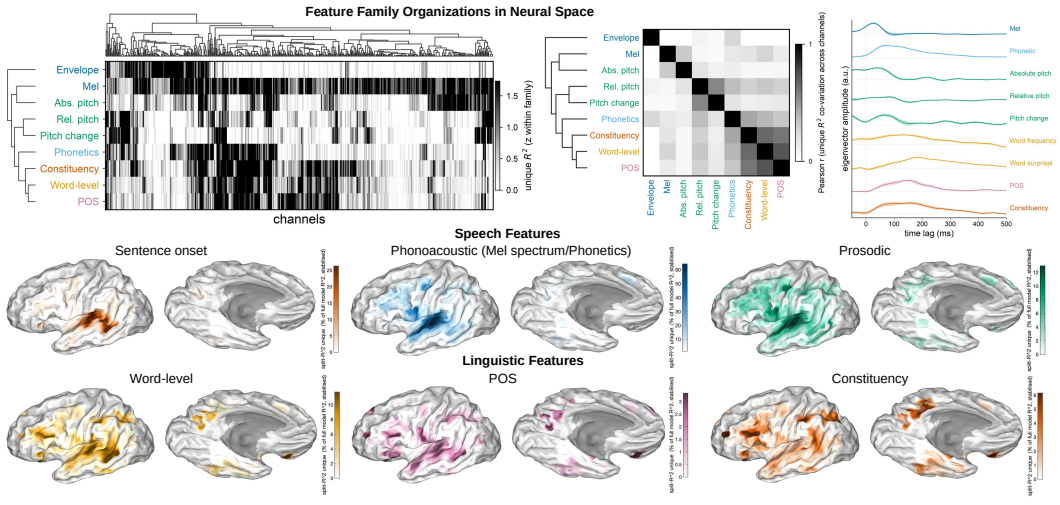

Figure 8: Feature-family organization in cortex. Top left: unique R^2 per family and channel from banded-ridge encoding models (held-out split-R^2), z-scored within family and biclustered, for the 482 of 1,604 left-hemisphere channels passing the held-out R^2 floor. Top middle: Pearson correlation of unique R^2 across channels. Top right: each family's dominant temporal mode (bootstrap SD; only peak latency is comparable across rows). Bottom: unique R^2 as a percentage of full-model R^2 per family on the standard TT_N27 surface.

4.4 A NESTED CORTICAL HIERARCHY INVERTS THE MODELS' ORGANIZATION

By contrast, the representational hierarchy in cortex (Figure 8) is nested, with categorical linguistic constructs extracted on top of one another; the dominant temporal modes follow the same order in

time. Part of speech clusters with the word-level family rather than with constituency, behaving as a lexical property rather than as compositional syntax. Sentence-onset and phonoacoustic variance concentrate in posterior superior temporal gyrus, planum temporale, and the supratemporal plane (Hamilton et al., 2018; Forseth et al., 2020); lexical, lexico-syntactic, and compositional syntactic families engage a fronto-parietal-temporal network. Superior temporal gyrus, which encodes lexical alongside phonetic features (Zhang et al., 2026), carries word-level but little constituency variance, and posterior temporal cortex separates anterior lexical from posterior compositional encoding. Along their own axes, the two systems thus diverge: the models spread low-level speech features across depth and confine lexical and compositional information to a narrow band of layers, whereas cortex localizes the former and distributes the latter.

5 Discussion

Joint interpretability of artificial and biological speech networks. To our knowledge, this is the first evidence that self-supervised speech models, trained without text, carry representations of compositional meaning that generalize across wordings, a capacity previously characterized as absent or weak (Choi et al., 2024; Pasad, 2025), and the first comparison of artificial and biological speech networks by the temporal dynamics of decodability and by an interpretable feature space rather than by prediction of neural activity, with targets defined by linguistic rules rather than model features. Reading both systems with the same probes, feature bank, and temporal generalization keeps the comparison symmetric and separates two questions: a probe shows that information is available to an external decoder, not that the system uses it (Ritchie et al., 2019), whereas temporal generalization asks whether a code, once formed, is maintained over the interval in which it must be acted on. On that criterion the two systems part ways at the level the paradigm isolates: the target concept is linearly available in every model, but the constituency targets, whose values depend on bracketing rather than on the words themselves, appear in the models as thin diagonals or not at all while cortex sustains them for hundreds of milliseconds, as representations that preserve lexical content while collapsing constituent structure would. Neither convergence nor divergence identifies an operation, since shared dynamics constrain the space of admissible implementations rather than specify one (Ross & Bassett, 2024), but the divergences are the discriminating results (Section 2). The feature bank likewise identifies what each system represents, not the transformations defined over it; recovering those transformations is the natural next step, and the layer bands and cortical hubs reported here mark where to look.

New constraints on the neurobiology of language. With sEEG of high spatiotemporal resolution, this study adds a spatial dimension to the dynamic coding account from magnetoencephalography (Gwilliams et al., 2025b): distributed hubs discretize compositional and relational information on top of the spectrotemporal code of auditory cortex (Forseth et al., 2018; Murphy et al., 2023), and sites beyond canonical speech-processing regions (Hamilton et al., 2021; Norman-Haignere et al., 2022; Keshishian et al., 2023; Zhang et al., 2026) support the extraction of meaning through lexical and constituency-based syntactic operations.

Implications for communication brain-computer interfaces. Speech neuroprostheses already decode model representations, from intermediate HuBERT units (Metzger et al., 2023; Littlejohn et al., 2025; Silva et al., 2024) to the GPT layer encoding language statistics (Tang et al., 2023). Mid-to-late layers carrying every level suggest a target space for decoding meaning as well as speech, and because the semantic band sits at a depth set by the training objective, that layer should be chosen per model rather than at a fixed relative depth; because only lower-level dynamics are shared, such decoders should learn compositional targets from pooled neural data (Singh et al., 2025), toward aphasia neuroprostheses (Li et al., 2026).

Limitations and outlook. Intracranial coverage was determined by clinical need, not experimental design: electrode placement differs across participants, samples cortex sparsely and unevenly, and is restricted here to the left hemisphere. Denser and bilateral coverage will be needed to confirm the proposed hubs. Within these limits, the results show where artificial and biological speech networks converge and diverge.

ETHICS STATEMENT

All participants were semi-chronically implanted with intracranial electrodes for the clinical localization of pharmaco-resistant epilepsy and provided written informed consent before taking part in research. All experimental procedures were reviewed and approved by the Committee for the Protection of Human Subjects (CPHS). Participants were identified only by coded study IDs, and because the recordings contain HIPAA-protected information, raw neural recordings cannot be shared publicly.

REPRODUCIBILITY STATEMENT

The full codebase and the experimental analysis results that underlie every figure and table are provided in the supplementary material.

AI USE STATEMENT

In this work, we used generative AI tools only for tasks with recommended disclosure under the ICLR 2027 AI Policy for Authors: (1) revising the manuscript to improve readability and correct grammar; (2) formatting the manuscript, tables, and references to the official ICLR 2027 template; and (3) editing the code that renders figures and tables to improve their display. AI tools were never used in the conceptualization of the study, the design of the stimuli, experimental data collection, data processing, proving mathematical claims, or generating synthetic datasets. We have reviewed all AI-assisted work: all authors read the revised text to confirm that it preserves the intended claims, and the formatted manuscript was checked against the ICLR 2027 author guidelines. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Richard Antonello and Alexander Huth. Predictive coding or just feature discovery? An alternative account of why language models fit brain data. *Neurobiology of Language*, 5(1):64–79, 2024. doi: 10.1162/nol_a_00087.

Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fMRI. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21895–21907, 2023.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020.

Vinamra Benara, Chandan Singh, John X. Morris, Richard J. Antonello, Ion Stoica, Alexander G. Huth, and Jianfeng Gao. Crafting interpretable embeddings for language neuroscience by asking LLMs questions. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3944.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Marc Brysbaert and Boris New. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990, 2009. doi: 10.3758/BRM.41.4.977.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, 2022. doi: 10.1038/s42003-022-03036-1.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, 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, 2022. doi: 10.1109/JSTSP.2022.3188113.

Emily Cheng, Aditya R. Vaidya, and Richard Antonello. Abstraction induces the brain alignment of language and speech models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. Self-supervised speech representations are more phonetic than semantic. In *Proceedings of Interspeech 2024*, pp. 4578–4582, 2024. doi: 10.21437/Interspeech.2024-1157.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Michael J. Crosse, Giovanni M. Di Liberto, Adam Bednar, and Edmund C. Lalor. The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, 10:604, 2016. doi: 10.3389/fnhum.2016.00604.

Marianne de Heer Kloots, Martijn Bentum, Hosein Mohebbi, Charlotte Pouw, Gaofei Shen, and Willem Zuidema. Tracking the emergence of linguistic structure in self-supervised models learning from speech. *arXiv preprint arXiv:2604.02043*, 2026a.

Marianne de Heer Kloots, Paul Boersma, and Willem Zuidema. Linguists should learn to love speech-based deep learning models. *Behavioral and Brain Sciences*, 49, 2026b. doi: 10.1017/S0140525X26104713.

Tom Dupré la Tour, Michael Eickenberg, Anwar O. Nunez-Elizalde, and Jack L. Gallant. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728, 2022. doi: 10.1016/j.neuroimage.2022.119728.

K. J. Forseth, G. Hickok, P. S. Rollo, and N. Tandon. Language prediction mechanisms in human auditory cortex. *Nature Communications*, 11(1):5240, 2020. doi: 10.1038/s41467-020-19010-6.

Kiefer James Forseth, Cihan Mehmet Kadipasaoglu, Christopher Richard Conner, Gregory Hickok, Robert Thomas Knight, and Nitin Tandon. A lexical semantic hub for heteromodal naming in middle fusiform gyrus. *Brain*, 141(7):2112–2126, 2018. doi: 10.1093/brain/awy120.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022. doi: 10.1038/s41593-022-01026-4.

Ariel Goldstein, Haocheng Wang, Leonard Niekerken, Mariano Schain, Zaid Zada, Bobbi Aubrey, Tom Sheffer, Samuel A. Nastase, Harshvardhan Gazula, Aditi Singh, Aditi Rao, Gina Choe, Catherine Kim, Werner Doyle, Daniel Friedman, Sasha Devore, Patricia Dugan, Avinatan Hassidim, Michael Brenner, Yossi Matias, Orrin Devinsky, Adeen Flinker, and Uri Hasson. A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature Human Behaviour*, 9(5):1041–1055, 2025. doi: 10.1038/s41562-025-02105-9.

Olivia Guest and Andrea E. Martin. On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, 6(2):213–227, 2023. doi: 10.1007/s42113-022-00166-x.

Laura Gwilliams, Ilina Bhaya-Grossman, Yizhen Zhang, Terri Scott, Sarah Harper, and Deborah Levy. Computational architecture of speech comprehension in the human brain. *Annual Review of Linguistics*, 11:209–226, 2025a. doi: 10.1146/annurev-linguistics-031120-111245.

Laura Gwilliams, Alec Marantz, David Poeppel, and Jean-Rémi King. Hierarchical dynamic coding coordinates speech comprehension in the human brain. *Proceedings of the National Academy of Sciences*, 122(42):e2422097122, 2025b. doi: 10.1073/pnas.2422097122.

Liberty S. Hamilton and Alexander G. Huth. The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 35(5):573–582, 2020. doi: 10.1080/23273798.2018.1499946.

Liberty S. Hamilton, Erik Edwards, and Edward F. Chang. A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Current Biology*, 28(12):1860–1871.e4, 2018. doi: 10.1016/j.cub.2018.04.033.

Liberty S. Hamilton, Yulia Oganian, Jeffery Hall, and Edward F. Chang. Parallel and distributed encoding of speech across human auditory cortex. *Cell*, 184(18):4626–4639.e13, 2021. doi: 10.1016/j.cell.2021.07.019.

Linyang He, Qiaolin Wang, Xilin Jiang, and Nima Mesgarani. Layer-wise minimal pair probing reveals contextual grammatical-conceptual hierarchy in speech representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35338–35353, Suzhou, China, 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1790.

Linyang He, Tianjun Zhong, Richard Antonello, Gavin Mischler, Micah Goldblum, and Nima Mesgarani. Far from the shallow: Brain-predictive reasoning embedding through residual disentanglement. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926, 2018.

Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 (7600):453–458, 2016. doi: 10.1038/nature17637.

Menoua Keshishian, Serdar Akkol, Jose Herrero, Stephan Bickel, Ashesh D. Mehta, and Nima Mesgarani. Joint, distributed and hierarchically organized encoding of linguistic features in the human auditory cortex. *Nature Human Behaviour*, 7(5):740–753, 2023. doi: 10.1038/s41562-023-01520-0.

Menoua Keshishian, Gavin Mischler, Samuel Thomas, Brian Kingsbury, Stephan Bickel, Ashesh D. Mehta, and Nima Mesgarani. Parallel hierarchical encoding of linguistic representations in the human auditory cortex and recurrent automatic speech recognition systems. *Nature Machine Intelligence*, 8(2):257–269, 2026. doi: 10.1038/s42256-026-01185-0.

J.-R. King and S. Dehaene. Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203–210, 2014. doi: 10.1016/j.tics.2014.01.002.

J. Li, A. Singh, and N. Tandon. Transitioning from speech to language brain–computer interfaces. *Nature Reviews Bioengineering*, 4:702–704, 2026. doi: 10.1038/s44222-026-00460-4.

Yuanning Li, Gopala K. Anumanchipalli, Abdelrahman Mohamed, Peili Chen, Laurel H. Carney, Junfeng Lu, Jinsong Wu, and Edward F. Chang. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, 26(12):2213–2225, 2023. doi: 10.1038/s41593-023-01468-4.

Kaylo T. Littlejohn, Cheol Jun Cho, Jessie R. Liu, Alexander B. Silva, Bohan Yu, Vanessa R. Anderson, Cady M. Kurtz-Miott, Samantha Brosler, Anshul P. Kashyap, Irina P. Hallinan, Adit Shah, Adelyn Tu-Chan, Karunesh Ganguly, David A. Moses, Edward F. Chang, and Gopala K. Anumanchipalli. A streaming brain-to-voice neuroprosthesis to restore naturalistic communication. *Nature Neuroscience*, 28(4):902–912, 2025. doi: 10.1038/s41593-025-01905-6.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of Interspeech 2017*, pp. 498–502, 2017. doi: 10.21437/Interspeech.2017-1386.

Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F. Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014. doi: 10.1126/science.1245994.

Sean L. Metzger, Kaylo T. Littlejohn, Alexander B. Silva, David A. Moses, Margaret P. Seaton, Ran Wang, Maximilian E. Dougherty, Jessie R. Liu, Peter Wu, Michael A. Berger, Inga Zhuravleva, Adelyn Tu-Chan, Karunesh Ganguly, Gopala K. Anumanchipalli, and Edward F. Chang. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620:1037–1046, 2023. doi: 10.1038/s41586-023-06443-4.

Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Rémi King. Toward a realistic model of speech processing in the brain with self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33428–33443, 2022.

Gavin Mischler, Yinghao Aaron Li, Stephan Bickel, Ashesh D. Mehta, and Nima Mesgarani. Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence*, 6(12):1467–1477, 2024. doi: 10.1038/s42256-024-00925-4.

Omer Moussa, Dietrich Klakow, and Mariya Toneva. Improving semantic understanding in speech language models via brain-tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KL8Sm4xRn7.

Elliot Murphy, Oscar Woolnough, Patrick S. Rollo, Zachary J. Roccaforte, Katrien Segaert, Peter Hagoort, and Nitin Tandon. Minimal phrase composition revealed by intracranial recordings. *Journal of Neuroscience*, 42(15):3216–3227, 2022. doi: 10.1523/JNEUROSCI.1575-21.2022.

Elliot Murphy, Kiefer J. Forseth, Cristian Donos, Kathryn M. Snyder, Patrick S. Rollo, and Nitin Tandon. The spatiotemporal dynamics of semantic integration in the human brain. *Nature Communications*, 14(1):6336, 2023. doi: 10.1038/s41467-023-42087-8.

Elliot Murphy, Patrick S. Rollo, Katrien Segaert, Peter Hagoort, and Nitin Tandon. Multiple dimensions of syntactic structure are resolved earliest in posterior temporal cortex. *Progress in Neurobiology*, 241:102669, 2024. doi: 10.1016/j.pneurobio.2024.102669.

Sam V. Norman-Haignere, Laura K. Long, Orrin Devinsky, Werner Doyle, Ifeoma Irobunda, Edward M. Merricks, Neil A. Feldstein, Guy M. McKhann, Catherine A. Schevon, Adeen Flinker, and Nima Mesgarani. Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nature Human Behaviour*, 6(3):455–469, 2022. doi: 10.1038/s41562-021-01261-y.

Anwar O. Nunez-Elizalde, Alexander G. Huth, and Jack L. Gallant. Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage*, 197:482–492, 2019. doi: 10.1016/j.neuroimage.2019.04.012.

Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. Speech language models lack important brain-relevant semantics. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8503–8528, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.462.

Ankita Pasad. *What do Speech Foundation Models Learn? Analysis and Applications*. PhD thesis, Toyota Technological Institute at Chicago, 2025. arXiv:2508.12255.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 914–921, 2021. doi: 10.1109/ASRU51503.2021.9688093.

Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. doi: 10.1109/ICASSP49357.2023.10096149.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics*, 12: 372–391, 2024. doi: 10.1162/tacl_a_00656.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 101–108. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-demos.14.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023.

J. Brendan Ritchie, David Michael Kaplan, and Colin Klein. Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science*, 70(2):581–607, 2019. doi: 10.1093/bjps/axx023.

Lauren N. Ross and Dani S. Bassett. Causation in neuroscience: Keeping mechanism meaningful. *Nature Reviews Neuroscience*, 25(2):81–90, 2024. doi: 10.1038/s41583-023-00778-7.

Inés Schönmann, Jakub Szewczyk, Floris P. de Lange, and Micha Heilbron. Stimulus dependencies, rather than next-word prediction, can explain pre-onset brain encoding in naturalistic listening designs. *eLife*, 14:RP106543, 2026. doi: 10.7554/eLife.106543.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118.

Riki Shimizu, Richard J. Antonello, Chandan Singh, and Nima Mesgarani. Interpretable embeddings of speech explain and enhance the brain encoding performance of audio models. *arXiv preprint arXiv:2507.16080*, 2025.

Alexander B. Silva, Kaylo T. Littlejohn, Jessie R. Liu, David A. Moses, and Edward F. Chang. The speech neuroprosthesis. *Nature Reviews Neuroscience*, 25:473–492, 2024. doi: 10.1038/s41583-024-00819-9.

Aditya Singh, Tessy Thomas, Jinlong Li, Greg Hickok, Xaq Pitkow, and Nitin Tandon. Transfer learning via distributed brain recordings enables reliable speech decoding. *Nature Communications*, 16(1):8749, 2025. doi: 10.1038/s41467-025-63825-0.

C. Tang, L. S. Hamilton, and E. F. Chang. Intonational speech prosody encoding in the human auditory cortex. *Science*, 357(6353):797–801, 2017. doi: 10.1126/science.aam8577.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023. doi: 10.1038/s41593-023-01304-9.

Mariya Toneva, Tom M. Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, 2(11):745–757, 2022. doi: 10.1038/s43588-022-00354-6.

Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H. McDermott. 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, 2023. doi: 10.1371/journal.pbio.3002366.

Aditya R. Vaidya, Shailee Jain, and Alexander G. Huth. Self-supervised models of audio effectively explain human cortical responses to speech. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21927–21944. PMLR, 2022.

N. Vattikonda, Aditya R. Vaidya, Richard J. Antonello, and Alexander G. Huth. BrainWavLM: Fine-tuning speech representations with brain responses to language. *arXiv preprint arXiv:2502.08866*, 2025.

Mingqing Xiao, Kai Du, and Zhouchen Lin. Beyond representational alignment with brain-guided language models for robust reasoning. *Nature Machine Intelligence*, 2026. doi: 10.1038/s42256-026-01278-w.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Yizhen Zhang, Matthew K. Leonard, Ilina Bhaya-Grossman, Laura Gwilliams, and Edward F. Chang. Human cortical dynamics of auditory word form encoding. *Neuron*, 114(1):167–180.e6, 2026. doi: 10.1016/j.neuron.2025.10.011.

A VARIANCE PARTITIONING SCHEME

The cortical encoding models are fit in five cumulative tiers (Table A1): each tier model contains its own bands and those of every tier below it, and every model includes a sentence-onset nuisance term. Phonological features follow Mesgarani et al. (2014), surprisal is computed with GPT-2 XL (Radford et al., 2019), and word frequency comes from SUBTLEX-US (Brysbaert & New, 2009). A family's unique contribution is the held-out R^2 of its tier model minus that of the reduced model in Table A2 (Section 3.3). Vowel and pause durations and word position enter as covariates and receive no unique term.

Table A1: Feature tiers of the cortical encoding models (dimensions in parentheses).

Tier	Bands added
Phonoacoustic	mel spectrogram (16), envelope (1), phonological features (14)
Prosodic	absolute pitch (10), relative pitch (10), pitch change (10), vowel duration (1), pause duration (1)
Lexical	word duration (1), word frequency (1), word surprisal (1)
Lexico-syntactic	part-of-speech tags (13)
Constituency	top-down, bottom-up, and left-corner node counts and tree depth (4); word position (1)

B SUPPLEMENTARY FIGURES

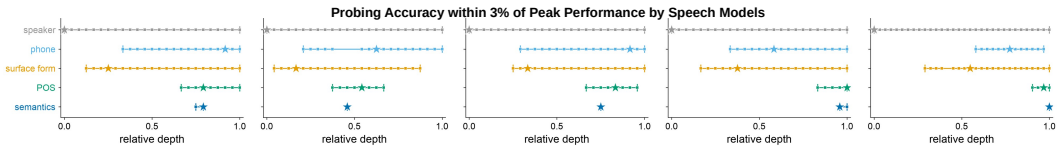

Figure A1: Layers within 3% of peak probe accuracy. For each task and model (HuBERT-large, wav2vec2-large, WavLM-large, Whisper-medium, and Qwen2-Audio-7B), the span of layers whose accuracy lies within 3% of the task's peak (bar) and the peak layer (star), against relative depth; colors as in Figure 3.

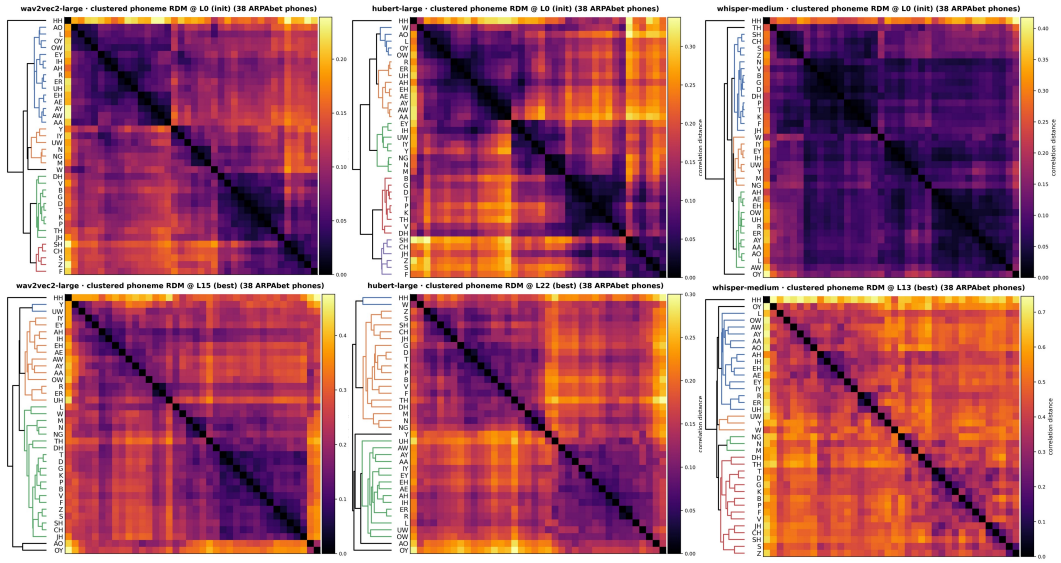

Figure A2: Phoneme representational dissimilarity at the input and at the best phone layer. Correlation distance between the span-pooled representations of the 38 ARPAbet phones that occur in the corpus, at layer 0 (top) and at each model's best phone layer (bottom; wav2vec2-large layer 15, HuBERT-large layer 22, Whisper-medium layer 13), with rows and columns ordered by hierarchical clustering (dendrograms). Color scales differ between panels.

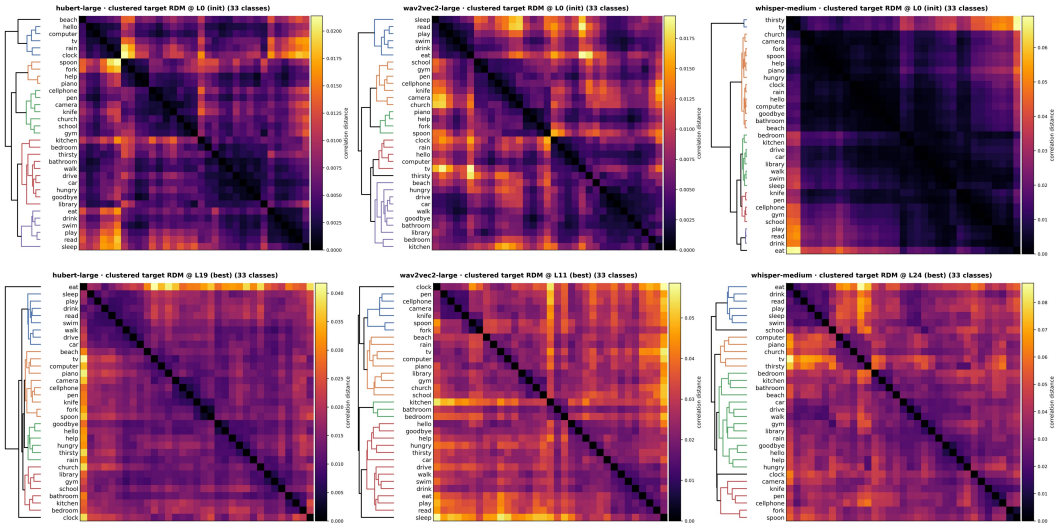

Figure A3: Target-concept representational dissimilarity at the input and at the best target layer. Correlation distance between the span-pooled representations of the 33 never-spoken targets, averaged over their definitions, at layer 0 (top) and at each model's best target layer (bottom; HuBERT-large layer 19, wav2vec2-large layer 11, Whisper-medium layer 24), with rows and columns ordered by hierarchical clustering (dendrograms). Color scales differ between panels.

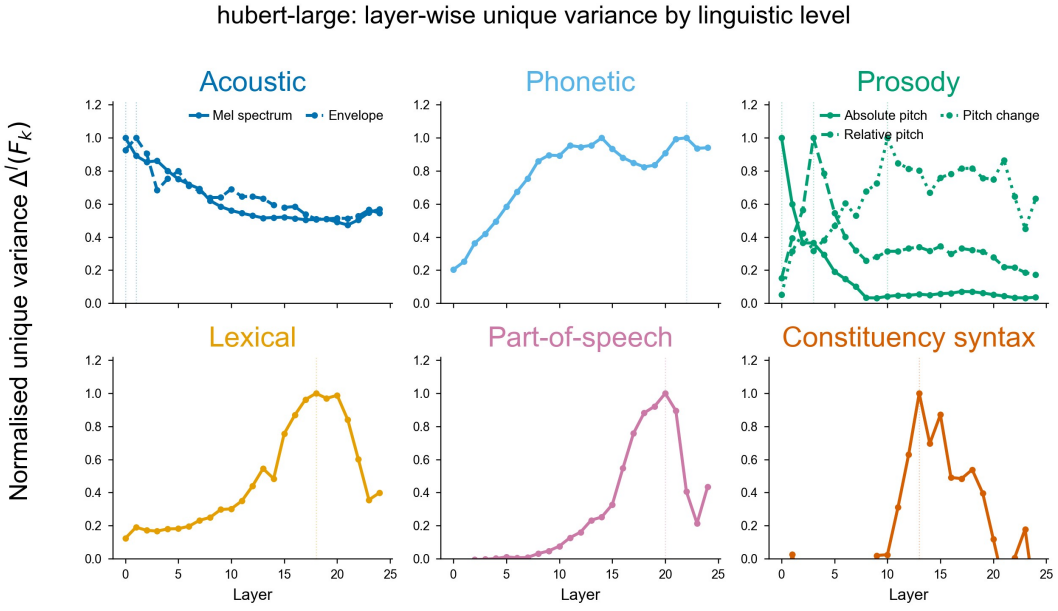

Figure A4: Layer-wise unique variance by linguistic level, HuBERT-large. Each family's unique R^2 from the model-side TRF partition of Figure 5 is plotted against layer, normalized to its own peak (dotted vertical lines mark the peak layer). Panels: acoustic (mel spectrum, envelope), phonetic, prosody (absolute pitch, relative pitch, pitch change), lexical (surprisal, frequency, duration), part of speech, and constituency syntax.

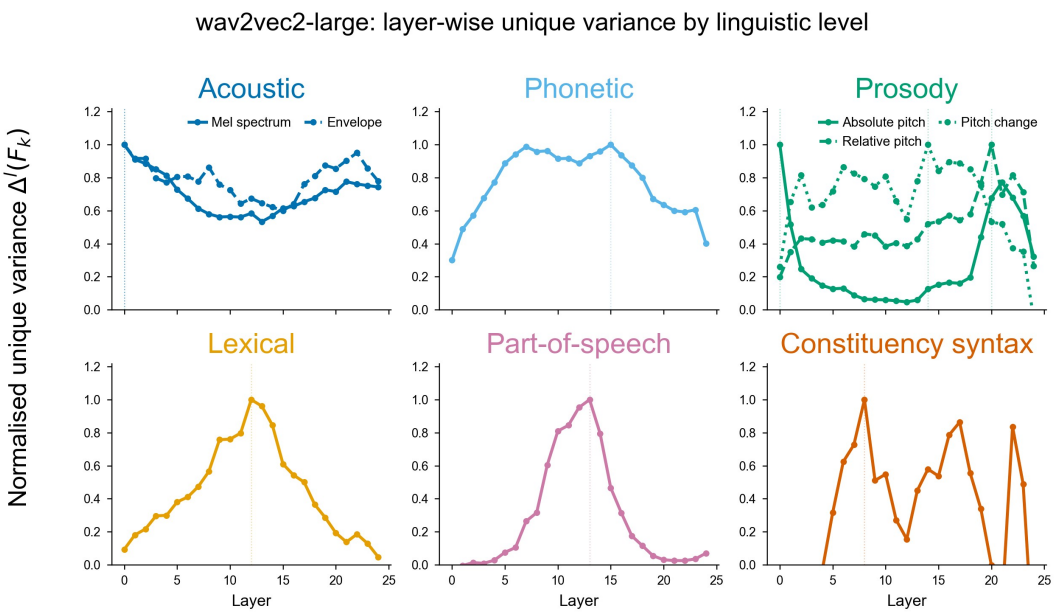

Figure A5: Layer-wise unique variance by linguistic level, wav2vec2-large. Each family's unique R^2 from the model-side TRF partition of Figure 5 is plotted against layer, normalized to its own peak (dotted vertical lines mark the peak layer). Panels: acoustic (mel spectrum, envelope), phonetic, prosody (absolute pitch, relative pitch, pitch change), lexical (surprisal, frequency, duration), part of speech, and constituency syntax.

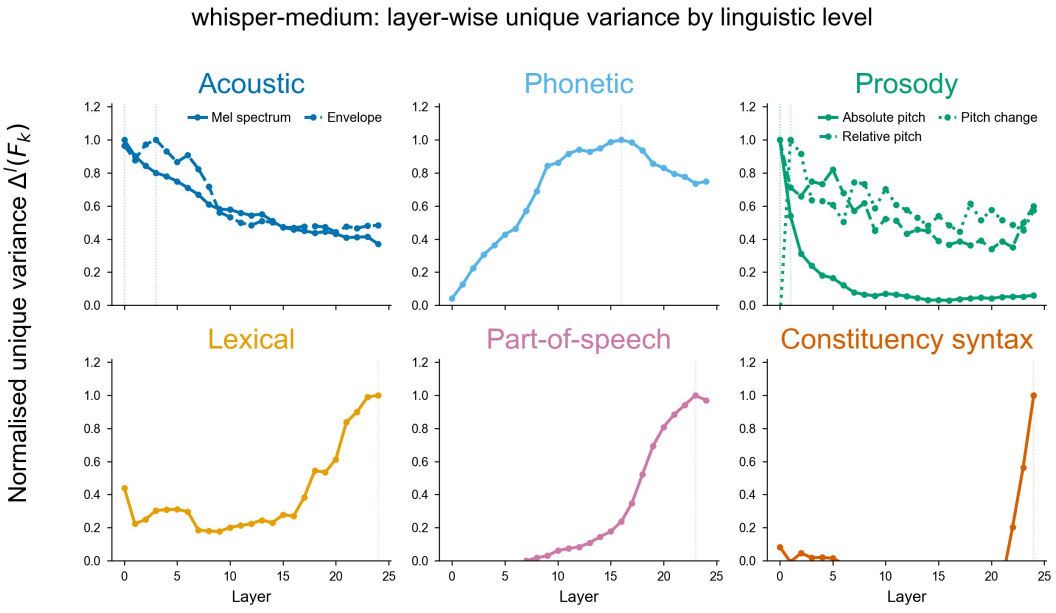

Figure A6: Layer-wise unique variance by linguistic level, Whisper-medium. Each family's unique R^2 from the model-side TRF partition of Figure 5 is plotted against layer, normalized to its own peak (dotted vertical lines mark the peak layer). Panels: acoustic (mel spectrum, envelope), phonetic, prosody (absolute pitch, relative pitch, pitch change), lexical (surprisal, frequency, duration), part of speech, and constituency syntax.

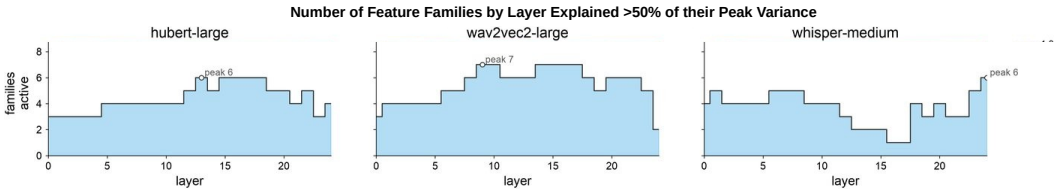

Figure A7: Coexistence of feature families across model depth. Number of feature families whose unique variance exceeds 50% of their own peak at each layer, for HuBERT-large, wav2vec2-large, and Whisper-medium, from the partition of Figure 5.

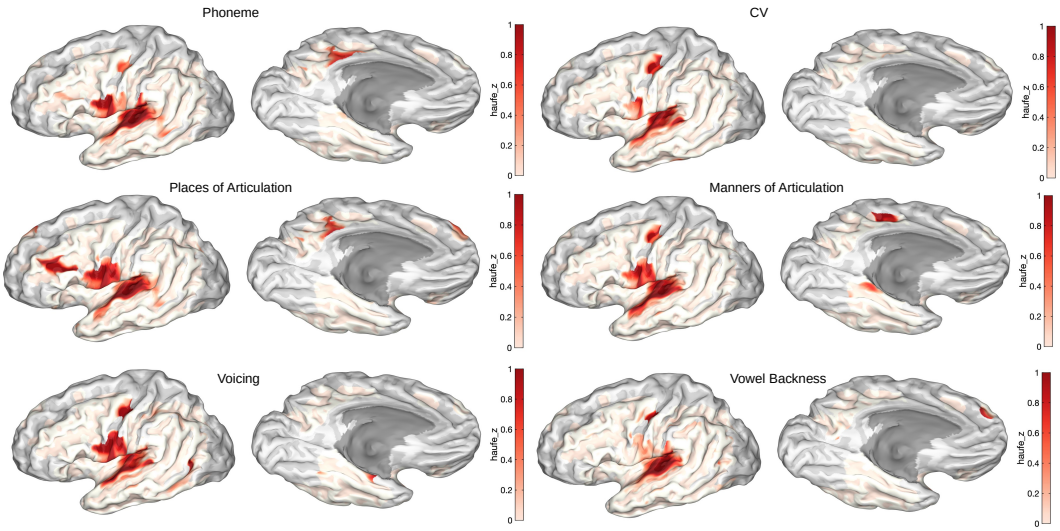

Figure A8: Cortical contribution maps of the phoneme-level temporal generalization decoders (orange panels of Figure 7). Decoder weights were converted to activation patterns with the Haufe transform, z-scored, and projected onto the standard TT_N27 left-hemisphere surface (lateral and medial views; color bar: z-scored Haufe pattern). Panels, left to right and top to bottom, correspond to the Figure 7 targets phoneme identity (Phonemes), consonant versus vowel (Consonant/Vowel), place of articulation (Place), manner of articulation (Manner), voicing (Voicing), and vowel backness (Vowel backness).

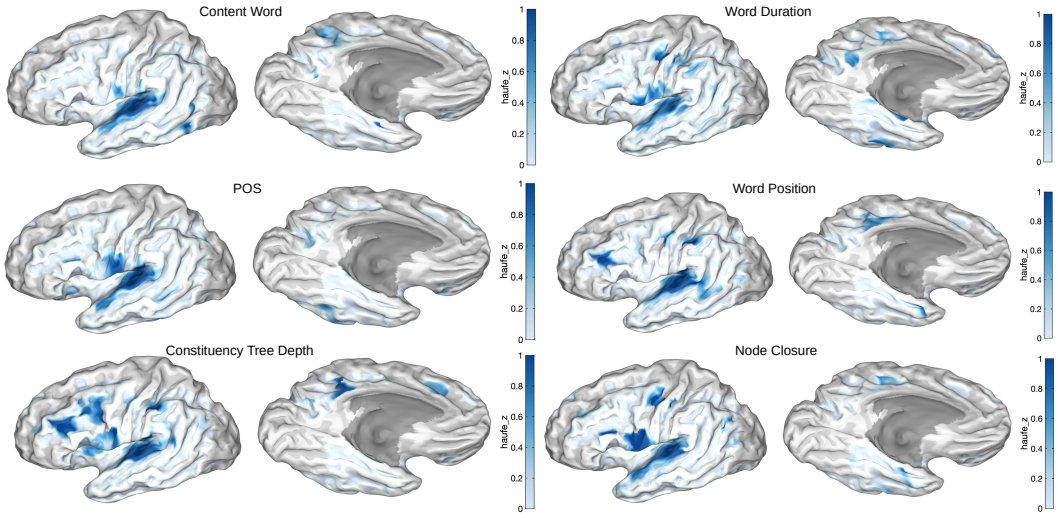

Figure A9: Cortical contribution maps of the word-level temporal generalization decoders (blue panels of Figure 7), computed as in Figure A8. Panels, left to right and top to bottom, correspond to the Figure 7 targets content word, word duration, part of speech (POS), word position, tree depth, and node closure.

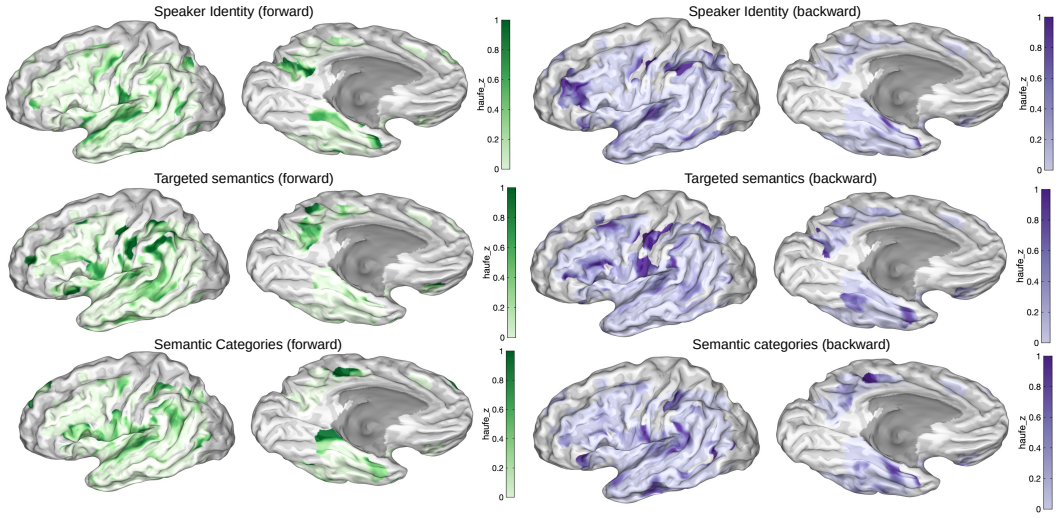

Figure A10: Cortical contribution maps of the sentence-level temporal generalization decoders (green and purple panels of Figure 7), computed as in Figure A8. Left: decoders aligned to sentence onset (forward; green); right: decoders aligned to sentence offset (backward; purple). Rows correspond to the Figure 7 targets speaker identity (top), target concept (middle, labeled targeted semantics), and semantic category (bottom).

Table A2: Contrasts defining each family's unique contribution, $\Delta R^2 = R^2(\text{tier model}) - R^2(\text{reduced model})$.

Unique contribution	Tier model	Reduced model
Phonetics	Phonoacoustic	Mel + envelope
Mel	Phonoacoustic	Phonetics + envelope
Envelope	Phonoacoustic	Mel + phonetics
Absolute pitch	Phonoacoustic + prosodic	Phonoacoustic + prosodic, without absolute pitch
Relative pitch	Phonoacoustic + prosodic	Phonoacoustic + prosodic, without relative pitch
Pitch change	Phonoacoustic + prosodic	Phonoacoustic + prosodic, without pitch change
Lexical	Phonoacoustic + prosodic + lexical	Phonoacoustic + prosodic
Lexico-syntactic	Phonoacoustic + prosodic + lexical + lexico-syntactic	Phonoacoustic + prosodic + lexical
Constituency	Phonoacoustic + prosodic + lexical + lexico-syntactic + constituency + word position	Phonoacoustic + prosodic + lexical + lexico-syntactic + word position

C ADDITIONAL METHODS

Neural recording and preprocessing. Intracranial EEG was recorded with a NeuroPort system (Blackrock Neurotech) at 2 kHz from sEEG probes implanted for clinical seizure localization. We excluded channels with line noise, epileptiform activity or other artifacts, as well as channels within the seizure-onset zone. Clean channels were re-referenced to their common average. Line noise and its harmonics were removed with zero-phase, second-order Butterworth band-stop filters. Broadband high-gamma activity (HGA; 70–150 Hz) was extracted with a frequency-domain bandpass Hilbert transform. Each channel's HGA was then z-scored against trial-level pre-stimulus baseline from -300 to -50 ms. Audio was recorded synchronously at 30 kHz through a lapel microphone on the same system. Word and phoneme boundaries were obtained with the Montreal Forced Aligner (McAuliffe et al., 2017) and checked manually in Praat. HGA was smoothed with a 100 ms Gaussian kernel, downsampled to 100 Hz, and epoched from 100 ms before the first word onset to 500 ms after the last word offset.

Temporal generalization targets. Targets come from the alignment, parse, and design, with continuous properties split at corpus terciles. *Phoneme-locked* targets (-100 to 500 ms from phone onset) map each aligned phone to consonant versus vowel, voicing, manner, place (consonants only), backness (vowels only), and identity. *Word-locked* targets (0 to 800 ms from word onset) are each word's content or function status, part of speech, length in phones, duration, position (first, medial, final), and constituent closure, i.e., whether it closes at least one phrase in a Stanza parse (Qi et al., 2020). *Definition-locked* targets, read forward from onset and backward from offset, are speaker, target concept, and its pre-registered eight-way semantic category. Each target is decoded under its own split (Section 3.1).

Statistical inference. Trajectory correlations are tested with an effective sample size $n_{\text{eff}} = L(1-\rho_1\rho_2)/(1+\rho_1\rho_2)$ given lag-1 autocorrelations ρ_1, ρ_2, with false discovery rate control (Benjamini & Hochberg, 1995) and a detrended replication. Group temporal generalization matrices are medians across participants, contoured where a significant majority exceed their own chance (sign test, $p < 0.05$, uncorrected); model contours mark cells above chance in every fold, a descriptive criterion. Encoding analyses use the channels with held-out $R^2 \geq 0.05$.

Model checkpoints. Table A3 lists the frozen encoders.

Table A3: Frozen encoders given audio only (SSL: self-supervised). Layer 0 is the input to block 1. Qwen2-Audio's last block exists only after pooling to 40 ms and is excluded (layers 0 to 31).

Family	Objective	Checkpoints	Blocks L	Frame
HuBERT	SSL: masked prediction	Base, Large	12, 24	20 ms
WavLM	SSL: masked, denoising	Base, Large	12, 24	20 ms
wav2vec 2.0	SSL: contrastive	Base, Large	12, 24	20 ms
Whisper	Weakly sup. ASR	Base, Small, Medium	6, 12, 24	20 ms
Qwen2-Audio	Audio-LLM	7B	32	20 ms
Qwen3-Omni (AuT)	Audio-LLM	30B-A3B	32	80 ms