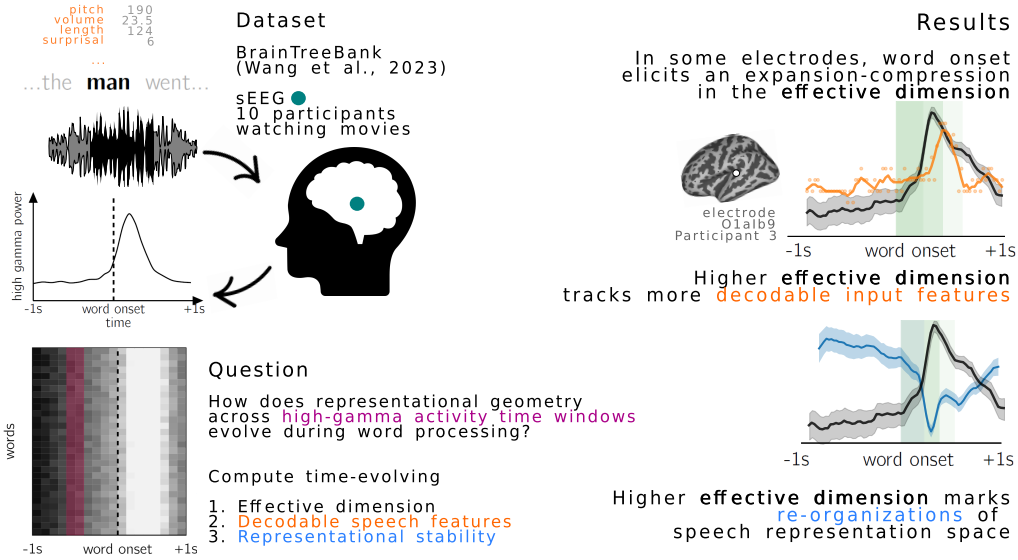
Graphical Abstract

Effective dimensionality tracks the time-varying accessibility of input features during speech comprehension

Emily Cheng*, Christopher Wang, Andrei Barbu, Marco Baroni, Greta Tuckute



*Corresponding author: Emily Cheng emilyshana.cheng@upf.edu

Highlights

Effective dimensionality tracks the time-varying accessibility of input features during speech comprehension

Emily Cheng*, Christopher Wang, Andrei Barbu, Marco Baroni, Greta Tuckute

- During speech processing, dimensionality of neural representations varies over a word.
- Higher dimensionality marks the processing of input-relevant speech features.
- In some electrodes, word onset elicits a peak in representational dimensionality.

*Corresponding author: Emily Cheng emilyshana.cheng@upf.edu

Effective dimensionality tracks the time-varying accessibility of input features during speech comprehension

Emily Cheng*[a], Christopher Wang[c], Andrei Barbu[c], Marco Baroni[a,b], Greta Tuckute[d]

[a]*Universitat Pompeu Fabra, Carrer de la Mercè 12, Barcelona, 08002, Spain*
[b]*ICREA, Passeig de Lluís Companys 23, Barcelona, 08010, Spain*
[c]*MIT CSAIL, 32 Vassar Street, Cambridge, 02139, MA, USA*
[d]*Kempner Institute at Harvard University, 150 Western Avenue, Boston, 02134, MA, USA*

Abstract

How does the brain dynamically build rich meanings from low-level acoustic information during spoken language understanding? Using an intracranial electroencephalography (iEEG) dataset from ten participants watching movies with naturalistic speech (1,688 electrodes), we asked how the evolving geometry of the high-gamma power relates to speech processing at sub-word temporal resolution. We found that the *effective dimension* of the high-gamma response at individual electrodes expands and compresses rapidly over the time course of a single word, and, critically, that a transient expansion in dimensionality tracks the availability of linearly decodable information about the input—a signature predicted by theoretical neuroscience and experimentally confirmed for the first time in the domain of speech comprehension. The link between dimensionality and linear decodability of input-relevant information is further supported by the fact that dimensionality was generally higher during speech than non-speech, changes in dimensionality marked re-organizations of neural representation space, and a subset of electrodes in the temporal and frontal cortex showed a pronounced expansion in dimensionality after word onset. Taken together, our results show that high-gamma dimensionality provides a temporally precise and interpretable marker for when input-relevant features become linearly accessible during natural speech comprehension.

*Corresponding author: Emily Cheng emilyshana.cheng@upf.edu

Keywords: speech processing, representational geometry

1. Introduction

During speech comprehension, the brain extracts and integrates complex meaning from low-level auditory information, all in the span of a couple hundred milliseconds (Brodbeck et al., 2018; Friederici, 2011; Gwilliams, Bhaya-Grossman, et al., 2025). In order to understand how the brain transforms sensory information into higher-order meanings, one dominant paradigm in cognitive neuroscience investigates how information is *represented* over the course of processing (Kriegeskorte & Kievit, 2013; Kriegeskorte et al., 2008). Concretely, patterns of neural activity can be characterized in terms of their similarity structure across stimuli, which reveals how the brain organizes information at different processing stages. Analyzing *representational geometry* as such has been useful in understanding visual processing (Z. Chen et al., 2026; Cohen et al., 2020; DiCarlo & Cox, 2007; Gauthaman et al., 2025; Stringer et al., 2019) and cognitive control and decision-making (Badre et al., 2021; Bernardi et al., 2020; Parthasarathy et al., 2019; Rigotti et al., 2013). In particular, these studies highlight how changes in the representations' *dimensionality* over processing map onto function, notably the building of task-relevant semantics (Fusi et al., 2016). Few analogous studies have been conducted in the domain of speech and language: for functional magnetic resonance imaging (fMRI), neural representations of words were found to be *higher*-dimensional in participants' native language than in a less proficient language (Zhang et al., 2024), and for intracranial electroencephalography (iEEG) and magnetoencephalography (MEG), semantic integration during comprehension has been linked to dimensional *increase* over the course of a sentence (Desbordes et al., 2023). In this work, inspired by recent studies using effective dimensionality as a probe of linguistic processing in large language models (e.g., Cai et al., 2021; Cheng et al., 2025; Valeriani et al., 2023), we present new evidence from human iEEG that, during speech comprehension, the dimensionality of neural responses to speech indeed cues linguistic feature-building in the short time after word onset. In particular, we addressed the following question:

Key research question: Does the effective dimension of the neural (high-gamma) response dynamically reflect the availability of input-relevant features during speech comprehension?

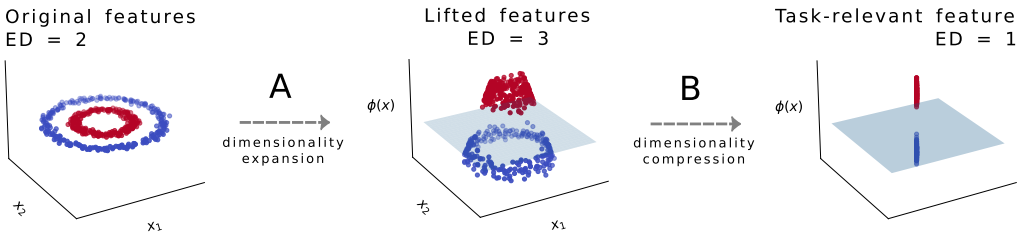


Figure 1: **Dimensionality, expressivity, and extraction of task-relevant information.** Processing may expand (**A**) or compress (**B**) the effective dimension of the feature space. In this illustration—which concatenates **A** and **B** for convenience (dimensionality expansion and compression can happen in isolation or in any order)—the task is binary classification of red and blue points that live in $\mathbb{R}^3$ (ambient dimension = 3). In the original feature space (left), points lie on a two-dimensional plane (ED=2) and are not linearly separable. **(A)** Lifting the points to a space with three effective dimensions via a nonlinear function ϕ (here, a radial basis function) results in a more *expressive* feature space (middle) where the red vs. blue points are linearly separable. **(B)** Compressing the points to a single effective dimension ϕ (vertical axis) gets rid of "nuisance features", which do not contribute to the ultimate classification decision, i.e., are not *task-relevant*.

To answer this question, we quantified the time-varying linear effective dimension of the high-gamma signal at each electrode ($N = 1{,}688$) in 10 participants as they watched movies containing naturalistic speech. In what follows, we show that the effective dimension of the high-gamma power tracks the number of decodable input-relevant features across time. This finding suggests that higher representational dimensionality indexes the (linear) accessibility of input-relevant information, some of which is speech-related (acoustic or linguistic). The claim that dimensionality tracks input-relevant information is supported by three further results. *First*, dimensionality *increases* reliably when speech is present vs. absent. *Second*, changes in dimensionality over the course of a word coincide with re-configurations of the neural representation space, as measured with representational similarity. *Third*, dimensionality is temporally aligned to the speech input, where word onset first elicits a transient expansion and then compression in the dimensionality for a set of electrodes in the temporal lobe. Taken together, our results suggest that the dimensionality of the high-gamma power can signify when, where, and how input-relevant information is made accessible in the brain for downstream processing.

2. Background and Related Work

Speech and language processing in the brain. Speech perception recruits a set of regions in human temporal cortex—in particular, the superior temporal gyrus

(STG) (Norman-Haignere et al., 2015; Overath et al., 2015; Yi et al., 2019). A network of interconnected regions in the temporal and frontal lobes (typically left-lateralized) selectively support linguistic processing across input modalities and languages (Fedorenko et al., 2024; Malik-Moraleda et al., 2022). Of interest to the current work is *representational dimensionality*: although fMRI lacks the temporal resolution to track rapid dimensionality changes within words or sentences, it has revealed low-dimensional structure across voxels, corresponding to interpretable axes such as early-to-late processing stage (Antonello et al., 2021), or easy-to-hard and concrete-to-abstract sentences (Botch & Finn, 2024; Tuckute et al., 2025). Using time-resolved methods, several studies have reported a temporal profile where neural activity gradually increases across the sentence (Fedorenko et al., 2016; Nelson et al., 2017; Regev et al., 2024; Woolnough et al., 2023). Importantly, these studies did not quantify changes in representational geometry; an exception is Desbordes et al. (2023), who linked sentence-level comprehension to increases in dimensionality. We build on this foundation to ask whether the time-varying dimensionality of the neural high-gamma signal tracks changes in the representational geometry of speech inputs at the word-level, and specifically, when input-relevant information becomes linearly accessible during speech comprehension.

Dimensionality of neural representational manifolds and function. Neural response measurements can appear high-dimensional, that is, data may live in a space that has a high *ambient dimension*. Fortunately, these measurements can often be reduced to a lower-dimensional subspace that is both more amenable to analysis and more informative about task semantics (Gao & Ganguli, 2015; Recanatesi et al., 2019). Several works agree that, in theory, the linear dimension of this subspace, or the *effective dimension*, bottlenecks the availability of information to immediate downstream areas (Canatar et al., 2023; Fusi et al., 2016; Ganguli & Sompolinsky, 2012; Harvey et al., 2024; Jazayeri & Ostojic, 2021). Underlying this agreement is the assumption that downstream populations process the information via a *linear* readout, where linearity is imposed as a conservative model of functional complexity (Badre et al., 2021; Jazayeri & Ostojic, 2021). Then, measuring the effective dimension (ED) of the firing activity of a neural population, or in our case, the time-varying high-gamma power at each electrode—a proxy of firing activity (Ray & Maunsell, 2011)—tells us something about the number of linearly decodable pieces of information accessible for that neural population. These studies suggest that the *local* representational complexity across time and space can indicate that information is being processed at a

specific time and location in the brain.

Both *high* and *low* dimensionality may signify task-related semantic abstraction. On one hand, **high-dimensional** geometry indicates the untangling of a rich and expressive feature space from raw inputs, which can then be flexibly used for downstream processing (Fusi et al., 2016; Ganguli & Sompolinsky, 2012). Figure 1A illustrates how dimensionality expansion can make (originally) nonlinear features linearly available. In this example, the data consists of inner and outer ring manifolds which are not linearly separable, and the task is to classify them. Lifting the data to a three-dimensional feature space via a nonlinear function (Radial Basis Function ϕ) allows the two classes to be separated via a linear hyperplane, i.e., making the nonlinear color feature *linearly available* for downstream processing. On the other hand, **low-dimensional** geometry can signal the extraction of abstract, task-relevant information from an array of "nuisance" features. This concept is illustrated by the transformation in Figure 1B, which starts with the three-dimensional feature space (middle plot). If the task is to classify red from blue, then only one task-relevant dimension ϕ (vertical axis) is needed. A key question is when during processing—and under what task demands—these strategies might manifest (Fusi et al., 2016). The neuroscientific literature has found evidence for both low- and high-dimensional signatures. First, the effective dimension of neural recordings is often orders of magnitude lower than their ambient dimension (Cunningham & Yu, 2014). In macaques performing an object categorization task, representational dimensionality *decreases* as one moves further away from visual perceptual areas (Brincat et al., 2018). Along the human visual ventral pathway, low-dimensional sensory manifolds are disentangled, however, via *increases* in representational dimensionality (DiCarlo & Cox, 2007; DiCarlo et al., 2012; Galella et al., 2025; Posani et al., 2025), and recent evidence (Gauthaman et al., 2025; Posani et al., 2025; Stringer et al., 2019) suggests that visual processing requires high-dimensional representation. Concurrent with and most similar to our work, Z. Chen et al. (2026) explore the temporal dynamics of visual representational geometry in an electroencephalography (EEG) and MEG study, showing that the effective dimension of neural representations expands and contracts shortly after stimulus onset, which marks a re-organization of visual representation space. Moving away from vision, Rigotti et al. (2013) have found high-dimensional representations in the macaque prefrontal cortex (PFC) to index behaviorally relevant task abstraction; Bernardi et al. (2020), moreover, found the representational dimensionality in macaque PFC to temporarily expand after stimulus presentation. All of these studies implicate an increase in representational dimensionality as a signature of *rich feature building*, where these features

can be flexibly processed downstream (Fusi et al., 2016; Rigotti et al., 2013).

In language processing, voxelwise variance in fMRI responses to language can be explained by a few principal directions, largely tracking interpretable axes such as *early-to-late* processing stages (as defined in large language models' representational space) (Antonello et al., 2021) or *low-to-high surprisal* and *concrete-to-abstract* (Botch & Finn, 2024; Tuckute et al., 2025). In addition, exposure to spoken words in one's native language vs. a less familiar language elicits *higher*-dimensional voxelwise responses, which Zhang et al. (2024) suggest to signify richer representation of the inputs. Beyond fMRI, how neural representational dimensionality dynamically maps onto language processing at high temporal resolution remains underexplored. Closest to our work, Desbordes et al. (2023) showed that the linear dimension of the broadband voltage signal increases over the course of a sentence, in support of the hypothesis that increased dimensionality signifies semantic integration. Our study also supports this view at a more granular unit of analysis (within each word), where, within one second of word onset, the absolute linear dimension at each electrode is much lower than the ambient dimension, but relative increases in dimensionality signal re-organizations of representation space, and, critically, input-relevant feature building.

3. Methods

3.1. Intracranial Electroencephalography (iEEG) Data

We use the BrainTreeBank (Wang et al., 2024), a publicly available iEEG dataset of 10 participants watching audiovisual movies. Participants ranged from 4 to 19 years old, 5 female, and each watched between 1-6 movies on a laptop screen for an average of total 4.37 hours (min: 0.95hrs, max: 3.93hrs per movie). Of these stimuli, on average, 33% of the movie time contains dialogue (min: 0.23%, max: 0.49%; see Table C.1). The dataset includes movie transcripts, where each word is time-locked to the voltage signal. Word-level annotations range from acoustic (e.g., pitch) to higher-order features (e.g., part-of-speech, surprisal). We use a subset of these features in decoding analyses, explained below in the paragraph *Decoding features*. The neural response data were recorded by stereo- electroencephalographic (sEEG) depth probes, each containing 6-16 0.8 mm diameter contact electrodes. The data then consists of one voltage time series per movie, participant, and electrode. The sampling frequency at each electrode is 2048Hz. Participants have on average 169 electrodes each (min 106, max 246), for a total of 1,688 electrodes, see Figure 2A. Further details can be found in Wang et al. (2024).

3.2. Preprocessing

In line with previous work (Regev et al., 2024), we consider the high-gamma power (frequency band 70-150Hz), which we extract from the raw voltage time-series of each electrode using the procedure in Appendix D: **Preprocessing**. Then, we segment the same high-gamma signal into *two sets of representations* corresponding to *words and non-speech sounds*. The word representations are formed by taking intervals aligned to word onset $\pm$ 1 second. These intervals are then split into 125ms chunks with a stepsize of 25ms. The non-speech representations are formed by taking all sections for which no speech is present, removing all one-second chunks that border speech intervals as a buffer, and then splitting the remaining signal in 125ms chunks with a 25ms stepsize. The resulting high-gamma representation at each timestep has an ambient dimension of $d = 2048\text{Hz} \times 0.125\text{s} = 256$.

3.3. Speech-responsive electrodes

Because we are particularly interested in *speech* processing, we additionally attempted to isolate electrodes that are more responsive to speech than non-speech. We emphasize that the public dataset does not include controlled experimental conditions or independent functional localizer runs (Saxe et al., 2006); accordingly, this "speech subset" is intended only as an auxiliary label for secondary analyses and visualization (e.g., Figure 3).

We define the "speech subset" as electrodes whose high-gamma power is significantly higher for speech (movie segments with dialogue) compared to non-speech sounds (movie segments without dialogue). For each movie, we randomly sample $N = 1000$ 500ms word segments timelocked to word onset ($t = 0$) and 500ms non-speech segments. For each participant and movie, we identify the electrodes whose average high-gamma power over the 500ms is higher for the word segments than for the non-speech segments, as determined by a one-sided Mann-Whitney U test ($\alpha = 0.05$, FDR-corrected over electrodes). Then, for each participant, we keep the electrodes that were significant for every movie. This method identified 165 out of 1688 electrodes (9.7%), located in largely temporal (but also frontal) brain regions typically associated with speech and language processing (Fedorenko et al., 2024; Norman-Haignere et al., 2015), see Figure E.1 for the per-participant electrode locations. See Appendix E for details.

Note that the speech-nonspeech contrast fundamentally differs from *language localizers* used in previous work (Fedorenko et al., 2024; Regev et al., 2024), which compare linguistic inputs with a perceptually matched control condition to isolate brain regions supporting "high-level" language processing (lexico-semantic

and combinatorial processing). In contrast, our method differentiates *speech* from *non-speech*. Moreover, unlike work that often restricts downstream analyses to localized electrodes (Regev et al., 2024), we retain the full electrode set for all primary analyses and use the speech subset only for comparison. Finally, the subset is defined using response magnitude (high-gamma power) and is independent of our primary metric of interest (effective dimension).

3.4. Quantifying processing complexity using the effective dimension

We ask how the *complexity* of speech processing is distributed across time (around word-onset) and across brain regions. Addressing this question first requires a measure of complexity, for which we use the time-varying dimensionality of the neural high-gamma power. For both word and non-speech conditions, we compute the linear effective dimension of each electrode's high-gamma representation for every timestep (estimated, recall, on input vectors given by 256 high-gamma values centered around the timestep).

In particular, we use the Participation Ratio (PR) (Abbott et al., 2011; Chung et al., 2018; Gao et al., 2017; Litwin-Kumar et al., 2017), an estimator that computes a continuous dimensionality measure as a function of the data covariance eigenspectrum $[\lambda_i]_{i=1}^{d=256}$:

$$\mathrm{PR} := \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}. \tag{1}$$

At an extreme where all eigenvalues are equal, the PR is equal to the ambient dimension d; at the other extreme where there is one nonzero eigenvalue, the PR returns 1 effective dimension, as desired. In between, the PR smoothly interpolates between extremes (Del Giudice, 2021).

3.5. Decoding input-relevant information from electrodes

If a representation manifold has higher effective dimension, then, theoretically, it contains more linearly decodable directions (Canatar et al., 2023; Harvey et al., 2024; Schaeffer et al., 2024). We asked whether these directions are related to the speech input by linearly decoding a set of speech input features from the neural high-gamma power of each electrode.

Input features. For each participant and movie, we performed decoding analyses using the word-segmented high-gamma activity from each electrode. We considered a total of *twelve speech features*. The speech features were annotated per word and ranged from low-level acoustic properties (pitch and volume) to lexical

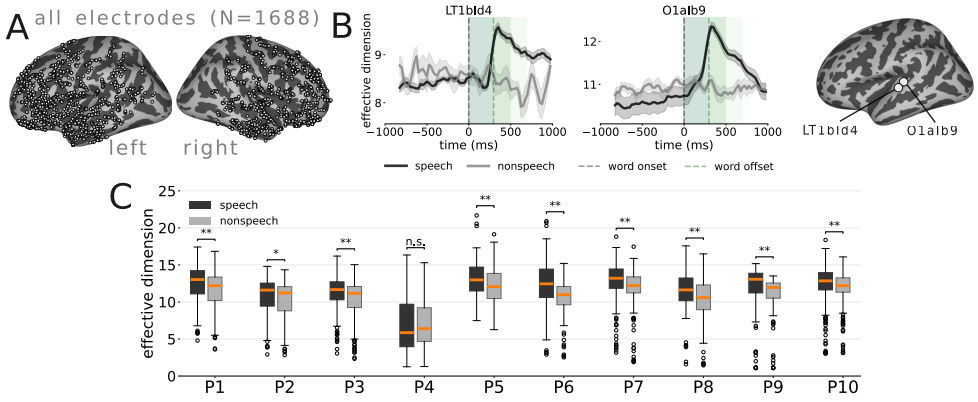


Figure 2: **Dimensionality when speech is present vs. absent**. **(A)** Electrode coverage ($N = 1,688$) electrodes across all participants. For visualization purposes, depth electrode contacts are projected onto the closest point on the brain's surface. **(B)** Two example electrodes (LT1bId4 and O1aIb9; both Participant 3): the effective dimension trajectory over the average pre- and post-word-onset (black) along with the average of non-speech segments of the same duration (gray), $\pm$1SE. Word onset is shown as a dashed gray vertical line, and the average word offset is shown as a dashed green line, with lighter green shadings indicating 1 and 2 SD across words, respectively; note that word onset and offset only apply to speech, as there are no words present for the non-speech trajectories. **(C)** For each participant P1…P10, the distribution of effective dimension over all electrodes is shown for the speech (black) and non-speech (gray) conditions. Values shown are the average effective dimension over the first 500ms after word onset (computed from 125ms, or ambient dimension $d = 256$, chunks with a step size of 25ms) for speech and non-speech conditions—again, word onset and offset only apply to the speech condition, as the non-speech trajectory is computed on random segments of non-speech sounds. For all participants except Participant 4, the average speech effective dimension is *higher* than for non-speech, as determined by a one-sided Mann-Whitney U test ($\alpha = 0.01^{**}, 0.05^{*}$).

surface (word length, number of distinct phonemes) and lexical semantic properties (content vs. function part-of-speech, concreteness, iconicity, valence, and arousal) to syntactic and information-theoretic features (absolute word position in sentence, distance to head position, and GPT2 contextual surprisal (Radford et al., 2019) using the previous 20 seconds). Most features were given by Wang et al. (2024); we additionally constructed the lexical semantic features using the public datasets described in Appendix G: **Decoding features**.

Decoding features. For each electrode, we wanted to characterize how input-relevant information is represented over the course of word processing. We took the same time series chunks used to compute the effective dimension, that is, 125ms chunks within 1 second of word onset, with a step size of 25ms. For each decoding task, the independent

variable was given by binary labels corresponding to the bottom (0) and top (1) quartiles of a given feature—labels are balanced by construction—and the model was a regularized logistic regression (see **Appendix G: Decoding features** for details). For each participant and movie, we conducted decoding experiments over three contiguous blocks, and performed 3-fold cross-validation. Finally, we scored the performance by averaging test accuracies over all splits and movies.

Linking effective dimension and decodability. We wanted to quantify how decodability relates to effective dimension. To the best of our knowledge, this is the first attempt in the literature to quantify the similarity between dimensionality and decodability in speech processing (in other domains, neural response dimensionality has been *qualitatively* related to time-varying task feature decodability (Bernardi et al., 2020; Z. Chen et al., 2026) and to coarse-grained anatomical or functional areas (DiCarlo & Cox, 2007; Galella et al., 2025)). Therefore, there is no established similarity metric. We therefore developed an approach, aiming to be simple and aligned with visual intuitions.

First, for each electrode, we aggregated the decoding performance across word-level features. To do so, we computed the number of features decodable above chance (50%) for each time step, where significance is determined by a one-sided t-test ($\alpha = 0.05$). This measure notably gives equal weight to each feature, reflecting our quantity of interest: how many input-related dimensions are linearly accessible.[1] The result is a "decodability" time series along the same interval (word onset $\pm$1s) and with the same sliding window as the dimensionality trajectories.

Next, we quantified the similarity between decodability and dimensionality. We did not have *a priori* knowledge on when, and for how long, decodability should track effective dimension: (1) features may emerge at different timescales (Hasson et al., 2008), (2) our feature set, though broad, does not capture all potential features being processed, and (3) the interval we consider may include processing of preceding words. These three factors may all influence the effective dimension timecourse, motivating a *search* for time intervals where dimensionality and decodability are similar, rather than imposing a global similarity measure over the whole epoch. We report a statistic $\beta \in [0, 1]$ (higher is more similar), which describes the fraction of the total time course for which dimensionality and decodability trajectories have a significant positive Spearman correlation (with p-value cutoff $\alpha = 0.05$). Examples for β between decodability and dimensionality span-

[1] We also considered mean decoding accuracy over features, but some features were systematically easier to decode, which biased the averages.

ning $[0, 1]$, along with further implementational details, are given in **Appendix G: Decoding features**.

3.6. Linking effective dimension and changes in representational geometry

During speech processing, representation spaces *restructure* over time (Gadonneix et al., 2026; Gwilliams, Bhaya-Grossman, et al., 2025). For each electrode, we quantified the extent of this restructuring via a dynamic Representational Similarity Analysis (RSA) (de Vries & Wurm, 2023; Kriegeskorte & Kievit, 2013; Kriegeskorte et al., 2008) on the high-gamma response to speech (see **Appendix I: Quantifying representational stability across time** for details). That is, for each electrode, using the same representations of speech as the previous analyses—125ms chunks of high-gamma activity with a 25ms step size—we computed an RSA score $\in [-1, 1]$ for each time step. The score at time step t quantifies the extent to which the representational geometry at time $t-1$, given by the pairwise Euclidean distance between responses,[2] predicts the representational geometry at time t: a score of 1 indicates high representational stability, 0 means no correlation, and -1 means that pairwise similarities are inverted. We tested the hypothesis that, because changes in effective dimension signify a re-organization of representation space, they should inversely correlate to RSA. Similarly to the number of decodable features, we made use of the β similarity metric to quantify how well local changes in representational geometry (given by $1 - \text{RSA}$) tracked changes in effective dimension.

4. Results

We first summarize the key results, which are described in detail in the following sections.

Across participants, fluctuations in the dimensionality of the neural high-gamma power at individual electrodes track input processing. We support this claim with four complementary analyses. *First*, the effective dimension was *higher* in the vast majority of electrodes in the presence vs. absence of speech, signaling that higher input complexity is mirrored in higher dimensional representational manifolds (Section 4.1, Figure 2). *Second*, and critically, across participants and in a large set of electrodes, the effective dimension of the high-gamma power was correlated with the number of decodable input features during online processing

[2] Using cosine similarity as a metric yielded extremely similar results.

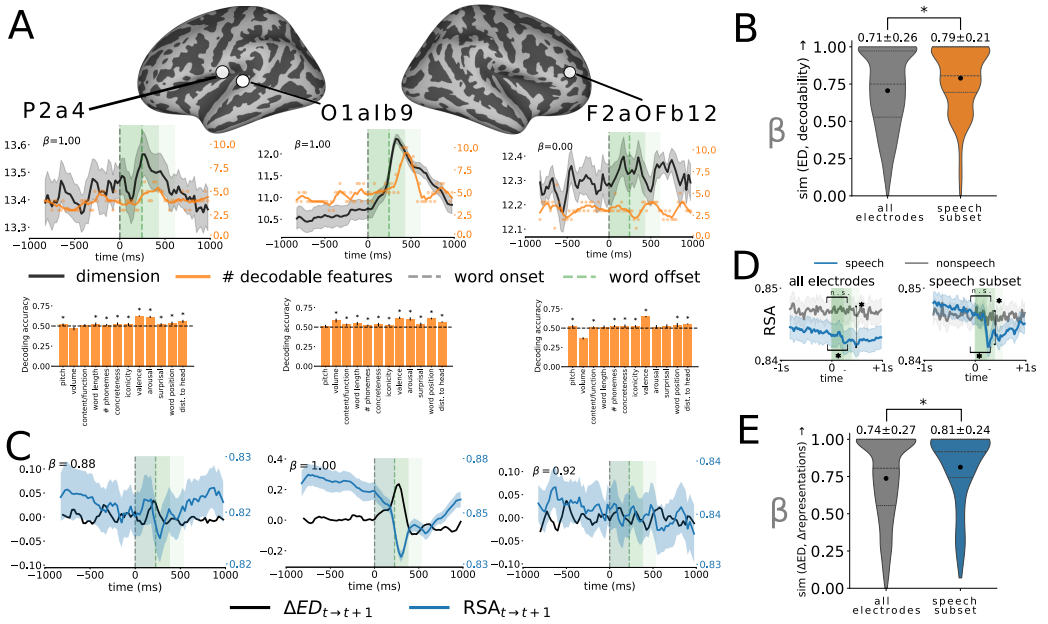


Figure 3: **Effective dimension tracks processing of speech-related information. (A) Example: Dimensionality tracks the number of decodable features. Top:** Example electrodes P2a4, O1aIb9 (Participant 3) and F2aOFb12 (Participant 10)'s locations on the brain. **Middle:** Trajectories for effective dimension (black) and number of decodable features (orange) in electrodes P2a4 and O1aIb9 are highly similar ($\beta = 1$). In contrast, electrode F2aOFb12 (right) with low dimensionality-decoding similarity, $\beta = 0$. **Bottom:** The decodability breakdown per-feature from left to right for P2a4, O1aIb9, and F2aOFb12, where * means significantly higher than chance (dashed line) as determined by a one-sided t-test with $\alpha = 0.05$. **(B) Dimensionality tracks decodability broadly across electrodes.** The distribution over *all electrodes* (left) and *speech-responsive electrodes* of dimensionality-decoding similarity (β). Means are shown at the top, and quartiles marked with dashed lines. A majority of electrodes' dimensionality time courses correlate with feature decodability (mean $\beta = 0.71$). The speech subset shows a significantly *higher* β (mean $\beta = 0.79$) than the set of all electrodes ($p < 5e$-4, Mann-Whitney U test). **(C) Changes in dimensionality track re-organizations of speech representation space.** Change in effective dimension (black) and RSA scores between subsequent time steps (blue, lower is less stable). In electrodes P2a4, O1aIb9, and F2aOFb12 (left to right), RSA scores *negatively correlate* to Δ ED: β between $1 - \text{RSA}$ and Δ ED is near 1. **(D) Representational stability dips shortly after word onset.** For the set of all electrodes (left) and the speech subset (right), there is a statistically significant dip in RSA $\sim$250ms after word onset (blue curves); significance was determined by a one-sided t-test comparing RSA scores in $[-200, 0]$ms and $[300, 500]$ms. The dip does not occur for random non-speech representations (gray curves, n.s.), suggesting a causal link between word onset and representational re-organization. **(E) Across electrodes, increases in dimensionality mark re-organizations of speech representation space.** The distribution of $\beta(1 - \text{RSA}, \Delta\text{ED})$ is shown for the set of all electrodes (left) and the speech subset (right). Over all electrodes, increases in the effective dimension tend to correlate with re-organizations of the space (mean $\beta = 0.74$); the trend is stronger for the speech subset (mean $\beta = 0.81$, significantly higher by a Mann-Whitney U test with $p < 5e-4$).

(Section 4.2, Figure 3). *Third*, changes in effective dimension cued changes in neural representation geometry across speech processing time (Section 4.3, Figure 3). *Fourth*, the temporal dynamics of dimensionality differed across brain areas, where, in particular, electrodes near the auditory cortex and superior temporal cortex displayed a pronounced dimensionality peak after word onset (Section 4.4, Figure 4). *Finally*, we observe that dimensionality and the raw high-gamma power provide overlapping but ultimately different views of speech processing, that is, dimensionality results cannot be explained by the magnitude of the high-gamma power (Section 4.5, Figure 5). Overall, evidence points to the time-varying dimensionality of the high-gamma power as a reliable signature of input feature building.

4.1. Higher processing dimensionality during speech vs. non-speech

We first asked whether the presence of speech is associated with higher-dimensional neural activity than segments without speech. Figure 2B shows two example dimensionality trajectories (channels LT1bId4 and O1aIb9 from Participants 3 and 4, respectively) averaged ± 1 s around word onset and compared to 2 s non-speech segments. In these two electrodes, the dimensionality during speech is higher than that of non-speech for the entire duration. This pattern generalized broadly over participants and electrodes: for all but one participant (Participant 4), the average dimensionality over electrodes was higher in the presence of speech than in its absence, see Figure 2C. Specifically, across all participants, the majority of electrodes (81.8%) showed a higher effective dimension during speech than non-speech, as determined by a one-sided Mann-Whitney U test that compares the first 500ms after word onset to random 500ms non-speech segments (FDR-corrected, $\alpha = 0.05$).[3] This pattern was relatively consistent across participants, see Table F.1. Interestingly, the overall trend of dimensionality increase during speech was true regardless of whether electrodes were speech-responsive. Of the 165 speech-responsive channels, a 67% majority ($N = 110$) displayed a significantly higher effective dimension during speech than non-speech sounds; 24% ($N = 40$) did not have a significant difference, and a minority 15% ($N = 25$) showed a reduced effective dimension ($\alpha = 0.05$). See Table F.1 for per-participant fractions.

In conclusion, we observed that adding an information-rich modality such as speech *increases* dimensionality. This finding provides preliminary support for

[3]We used a nonparametric Mann-Whitney U test as values were not normally distributed, according to a Shapiro-Wilk test ($\alpha = 0.05$).

the hypothesis that dimensionality tracks rich feature-building during comprehension. Crucially, this comparison serves as a prerequisite for the analyses that follow: if dimensionality is to provide a useful marker of speech feature processing, it should first distinguish periods with speech (locked to word onset) from periods without speech (not locked to any analogous event). In the next experiment, we probe the relationship between dimensionality and decodable information explicitly.

4.2. Transient effective dimension tracks decodable information about inputs.

Having established that dimensionality distinguishes speech from non-speech, we next asked whether it tracks specific features related to the speech input. For the vast majority of electrodes, the number of decodable input features tracks the effective dimension of the high-gamma power over the time course of a word. Certain features were more decodable than others in different brain areas; for instance, lexical-semantic properties like arousal were most decodable in temporal and frontal areas typical for language processing (Fedorenko et al., 2024), see Figure G.2 for the breakdown for each of the twelve features. Nonetheless, the *total* number of decodable features correlated positively across time with the effective dimension in the vast majority of electrodes. Figure 3A shows three examples. For electrodes P2a4 and O1aIb9 in Participant 3's left parietal lobe and auditory cortex, trajectories (middle row) for dimensionality (black) and the total number of decodable features (orange) positively correlated for the entire time course ($\beta = 1.0$); in contrast, electrode F2aOFb12 (Participant 10's right frontal lobe) did not positively correlate for any stretch of the time course ($\beta = 0.0$). See Figure G.3 for more electrode examples ranging from $\beta = 0$ to $\beta = 1$.

The high degree of similarity β illustrated by the first two examples was *pervasive across electrodes*. Figure 3B (left) shows the distribution over all electrodes of the similarity β between dimensionality and decodability of all features, where the mean is $\beta = 0.71$ ($\bullet$ in Figure 3B), SD$= 0.26$ and 25th, 50th, and 75th quartiles are marked as dashed lines. In line with expectations, the set of speech-responsive electrodes (orange, right) showed a *higher* mean similarity between dimensionality and number of decodable speech features ($\beta = 0.79$, SD=0.21) than the set of all electrodes. This difference was significant as determined by a one-sided Mann-Whitney U test ($p <$ 5e-4). A similar trend weakly holds for each participant, see Figure G.4. In conclusion, effective dimension tracks the number of decodable features, implying that time-varying dimensionality indicates the linear availability of input information at electrode sites.

4.3. Increases in dimensionality over processing mark re-organizations of representation space

While the time-varying decodability provides one perspective on speech processing dynamics (whether information is linearly accessible), another view is directly given by the temporal evolution of *representation geometry* (the extent to which the representational space of the neural response itself is changing). Here, we show that broadly across electrodes, increases in effective dimension over the time course of a word coincide with reconfigurations of the neural representation space of speech, as measured using Representational Similarity Analysis (RSA).

We first note that, in absolute terms, representation geometry was locally *stable* within the time interval considered: over all electrodes and time steps, the average RSA score between subsequent time steps (25ms) in the two seconds around word onset was $\rho = 0.85$ (SD $= 0.04$) for neural speech representations and $\rho = 0.84$ (SD $= 0.04$) for non-speech representations. This high degree of stability is somewhat expected, as the chosen step size of 25ms is small enough to reflect autocorrelations in both the stimulus (Schönmann et al., 2026) and our choice of representation (125ms). In contrast, pairs of representations at random time points within the same movie are *not* similar (see Appendix I.1 for details): averaged across all representation pairs and electrodes, $\rho = 0$ (SD$= 0.05$) between random speech representations and random lagged representations (mean 39.9 minutes apart, SD $= 34.8$).

Despite the high representational stability over the course of a word, there was a small but significant dip in the average RSA profile over electrodes roughly 250ms after word onset (blue curve in Figure 3D, left). Significance was determined by a one-sided t-test comparing the average RSA scores $[-200, 0]$ms before word onset and $[300, 500]$ms after word onset ($\alpha = 0.05$). This dip was more pronounced for the set of speech-responsive electrodes, seen by the steeper drop in the blue curve in Figure 3D (right) around 250ms post-onset, and does not occur in the processing of non-speech sounds (n.s. at $\alpha = 0.05$, gray curves in Figure 3D), suggesting a causal link between word onset and the re-organization of representation space that follows.[4] Interestingly, representational dynamics were typically dominated by a single slow mode corresponding to the representations' top principal component (PC), discussed in **Appendix J: A closer look at top-PC**

[4]Note that speech-responsive electrodes were selected based on a speech–non-speech contrast rather than a word-onset–based contrast; therefore, the post-word–onset dip does not trivially stem from the selection criterion.

subspaces. This top PC (and other PCs) individually encoded multiple features throughout the speech hierarchy at each time step, aligning with existing work that has found local electrode sites to simultaneously represent low-level and abstract linguistic information (Défossez et al., 2023; Gwilliams, Marantz, et al., 2025; Keshishian et al., 2023).

We tested the hypothesis that lower RSA scores correlated to changes in the effective dimension, as both signatures should appear when speech representations re-organize along different axes. Figure 3C shows for electrodes P2a4 and O1aIb9 from Participant 4 and F2aOFb12 from Participant 10 that, over speech processing, re-organizations of neural representation space coincide with *increases* in the effective dimension. In Figure 3C, each point in the blue curve denotes the RSA score from the previous to the current time step (lower means less stable). Plotted in the same figure is the change in the effective dimension (black curve) computed on the same time steps: the curves are negatively correlated for all example electrodes, where $\beta \approx 1$ between the change in effective dimension and the degree of representational instability given by $1 - \text{RSA}$. Figure 3D shows how this relationship manifests over electrodes: broadly high β scores imply that increases in the effective dimension tend to co-occur with phases of greater representational re-organization for the set of all electrodes (Figure 3D left, $\beta = 0.74$, SD $= 0.27$) and even more so for the speech subset (Figure 3D right, $\beta = 0.81$, SD $= 0.24$; significantly higher than for all electrodes as determined by a Mann-Whitney U test, $\alpha = 0.05$). In contrast, *absolute* changes in effective dimension, i.e., considering both increases and decreases in dimensionality, correlated for a smaller fraction of time with representational instability during speech processing in the set of all electrodes ($\beta = 0.44$, SD $= 0.31$) and the speech subset ($\beta = 0.62$, SD $= 0.34$), see Figure I.1. We hypothesize that *decreases* in the dimensionality may primarily compress features less related to speech processing; as we illustrated in Figure 1B, ridding the space of non-speech–relevant features could lower its dimensionality without severely changing the representational geometry over speech inputs.

Taken together, the dimensionality, decodability, and RSA results appear to support the narrative presented in Figure 1A, suggesting that at local electrode sites, the axes organizing speech-related information rapidly change over processing, where *expanding* the number of effective axes over time corresponds to greater speech representational change and richer encoding of speech features.

4.4. Heterogeneous dimensionality trajectories throughout the brain

Although dimensionality was correlated with decodability and re-organization of the representation space across many, but not all, electrodes (Figure 3), the

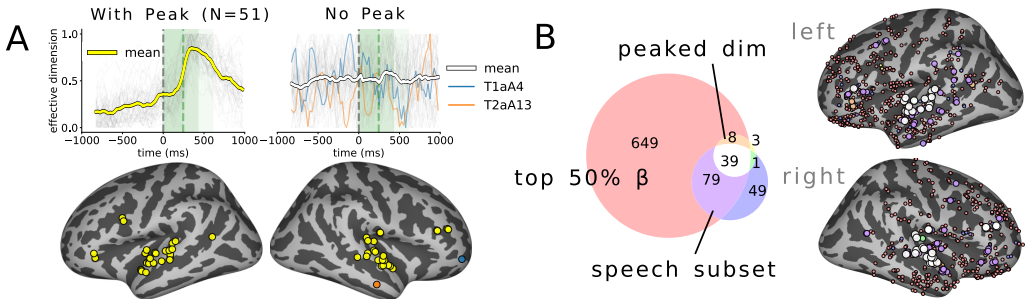


Figure 4: **Electrode subsets. (A) Heterogeneous temporal trajectories in dimensionality for different electrodes. Top left and right:** The mean time-varying effective dimension for the electrodes whose dimensionality displayed a post-word onset peak are shown on the left, and the mean effective dimension for an equally-sized sample of electrodes that do not show a peak is shown on the right. The "no peak" electrodes' dimensionality temporal trajectories were very heterogeneous, as seen by the blue and orange examples in the plot (electrode T1aA4; Participant 5, and T2aA13; Participant 10). The effective dimension is normalized to between 0 and 1 for comparison between participants and electrodes. **Bottom:** The locations of the "peaked" electrodes (yellow) are largely in the auditory cortex and STG, with 7 in frontal regions. The two example electrodes for the non-peaked set are also shown (orange and blue). **(B) Electrode subset distributions across the brain surface.** The distribution of the top 50% electrodes by β for speech feature decodability, corresponding to $\beta \geq 0.75$ (pink), plotted with electrodes that display a post-word onset peak during speech ("peaked dim", yellow), and the speech subset (blue).

shape of their dimensionality trajectories (i.e., time courses) was heterogeneous. Qualitatively, two broad patterns stood out: a word-onset peak and a more variable, non-peaked trajectory. First, related to the "peaked" group: over participants, we manually identified a subset of 51 electrodes out of 1688 that exhibited a pronounced dimensionality peak between 250 and 500ms after word onset. Each of these 51 electrodes' dimensionality trajectories were significantly peaked as proxied by their Pearson correlation to a template Gaussian curve and determined by a permutation test ($\alpha = 0.05$, FDR-corrected over electrodes)—see **Appendix H: Identifying peaked electrodes** for details. Figure 4A (yellow) shows the average dimensionality trajectory across peaked electrodes. Crucially, in these electrodes, the peak is not present for non-speech segments, see Figure H.3. This indicates that the increase in dimensionality after word onset is indeed related to speech processing. These "peaked" electrodes were consistently located near the primary auditory cortex (PAC) and the superior temporal gyri (STG), with only a few in the frontal lobe. One participant out of ten (Participant 6) did not exhibit peaked electrodes, likely due to poor electrode coverage in STG; see Figure E.1 for per-participant electrode placements.

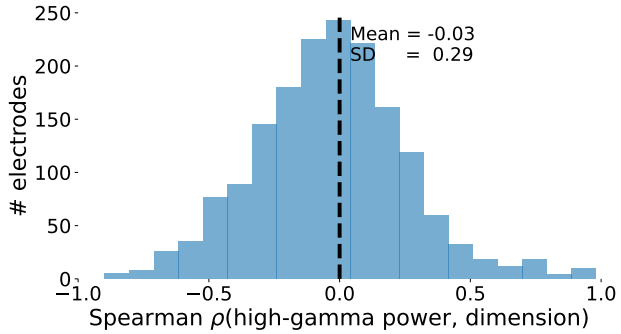


Figure 5: **High-gamma power does not correlate with the effective dimension.** The distribution of the Spearman correlation ρ between the high-gamma power and effective dimension over word onset ±1s is shown over all electrodes. On average, high-gamma power does not correlate highly to effective dimension, seen by the low average $\rho = -0.03$, close to 0 (SD= 0.29).

In contrast, the remaining electrodes showed heterogeneous dimensionality trajectories without a pronounced post-word–onset peak. Figure 4A (top right) shows the mean dimensionality trajectory (white) for a random subsample of $N = 51$ "non-peaked" electrodes, along with two highlighted example trajectories (electrodes T1aA4 and T2aA13, Participants 5 and 10, respectively). Our finding that "peaked" electrodes are generally near the auditory cortex and STG and non-peaked electrodes scattered throughout the brain (Figure 4A) is broadly consistent with the idea that dimensionality dynamics may unfold over different temporal scales across the speech processing hierarchy (C. Chen et al., 2024; Hasson et al., 2008). We return to this idea in the Discussion. Thus, although dimensionality tracked feature decodability across many electrodes, its temporal profile varied substantially across sites. In particular, non-peaked electrodes may still show dimensionality trajectories that correlate with feature decodability without tightly aligning to word onset.

4.5. Effective dimension provides information distinct from high-gamma power

A natural question is whether measuring the effective dimension provides a fundamentally different view from the high-gamma power. We first note that, over all electrodes, the average Spearman correlation between high-gamma dimensionality and magnitude across time was -0.03, SD$= 0.29$, which rules out a consistent relationship between the two (see Figure 5 for the distribution).

Second, we already saw that electrodes deemed speech-responsive through a high-gamma-based selection procedure showed higher similarity β between dimensionality and speech decodability than other electrodes (Figure 3B). But are

high β electrodes for speech decodability guaranteed to be speech-responsive? The evidence suggests not. Figure 4B (left) shows the top 50% of electrodes (pink) ranked by the fraction of the time course during which dimensionality and *speech* decodability are positively correlated ($\beta > 0.77, N = 775$) across all participants. These top-β electrodes are distributed across the brain, within and outside of typical temporal and frontal speech language areas. When comparing the top-β electrodes (pink) with the speech-responsive subset (blue) and the "peaked" electrodes from Figure 4A (yellow), only partial overlap emerged. Notably, a bilateral set of 39 electrodes (white) near the auditory cortex and STG belongs to all three groups: they are speech-responsive, exhibit a post–word–onset dimensionality peak, and their dimensionality trajectories correlate with speech decodability (Figure 4B, right). However, 649 out of 775 top-β electrodes are *not* speech-responsive, and 51 out of 165 speech-responsive electrodes did not exhibit top 50% β values. Overall, these results show that dimensionality and high-gamma-based measures of speech processing provide overlapping but ultimately *different* views of speech processing. Namely, raw high-gamma signal magnitude captures response strength, whereas dimensionality captures the structure of the response space.

5. Discussion

We found the time-varying dimensionality of the high-gamma power to index information about the perceived inputs (acoustic and linguistic): during speech comprehension in movie watching, high-gamma activity sustains time-varying representational subspaces from which input features are decodable with a linear readout. While the effective high-gamma dimension (ED≤ 23 for all participants) remained much lower than the ambient dimension ($d = 256$) throughout processing, relative *expansions* in the dimensionality marked phases of speech feature building and reconfigurations of representation space. This signature confirms theoretical predictions (Fusi et al., 2016; Ganguli & Sompolinsky, 2012; Jazayeri & Ostojic, 2021) and aligns with other experimental work that has found higher dimensionality to track semantic integration (Bernardi et al., 2020; Desbordes et al., 2023). Especially salient are the recent convergent results from Z. Chen et al. (2026), who found that higher effective dimension marks richer stimulus-dependent processing in human visual cortex.

We also determined that effective dimension trajectories varied across electrodes. Some sites, particularly in auditory cortex (AC) and superior temporal gyrus (STG), showed a dimensionality expansion closely tied to word onset,

whereas others exhibited more heterogeneous trajectories, suggesting that dimensionality dynamics need not be uniformly locked to the onset of individual words. The finding that the heterogeneous dimensionality trajectories were distributed over a larger anatomical area compared to the AC/STG areas aligns with Wang et al. (2024), who found speech vs. non-speech decodability, a signature of speech processing, to fluctuate more gradually in the frontal lobe and peak more sharply in the temporal lobe. One possible interpretation is that these distinct dimensionality profiles reflect differences in processing timescale. Neural responses to speech tend to organize along a timescale hierarchy, where sensory inputs are processed in shorter temporal windows than higher-order information (C. Chen et al., 2024; Hasson et al., 2008), and this account corroborates our observation that electrodes with a post-word-onset peak were concentrated near AC/STG, closer to the locus of auditory perception. We note, however, that Regev et al. (2024) found that neural populations sensitive to different temporal scales distributed throughout the fronto-temporal language network (Fedorenko et al., 2024), suggesting that anatomical location alone is unlikely to explain the observed dimensionality profiles. More broadly, non-peaked electrodes may reflect neural populations receiving more diverse inputs, such that their dimensionality trajectories correlate with feature decodability and representational re-organization without being tightly aligned to word onset.

In our study, dimensionality provided additional information about processing *distinct* from the high-gamma power, a central focus in the neurolinguistics literature. We believe that, more broadly, neural representational geometry can serve as a promising tool alongside the high-gamma power to better understand speech and language processing in the brain.

Limitations. Since the data were collected without repeated trials (each participant watched different movies), it was not possible to denoise across the same condition. In addition, electrode placements may reflect sampling bias—a tradeoff inherent to iEEG, as implantation is typically done as part of an invasive medical procedure (Wang et al., 2024). In addition, because the sample size was relatively small ($N = 10$), it is difficult to determine how participant age (4-19 years) impacts our results (no clear trends emerged from analyzing the speech vs. non-speech ED contrasts, for instance).

Future work. There are (at least) two directions for future work. First, one could examine how the presence of a certain feature in speech affects effective dimension trajectories. Especially of interest from a psycholinguistic perspective is

contextual surprisal (Hale, 2001; Levy, 2008), which has well-documented effects on behavioral and neural markers of processing effort like reading times (Shain et al., 2024; Wilcox et al., 2023) and event-related potentials (Frank et al., 2013, 2015). On the same dataset we use, Wang et al. (2024) found higher and delayed voltage responses to higher surprisal words in speech-sensitive electrodes. We hypothesize that words with higher surprisal may show delayed dimensionality responses as a signature of increased processing effort, requiring more time for task-relevant information to emerge.

Second, contemporary computational speech and language models, such as LLMs, are able to process language remarkably well, prompting comparisons to the brain (Caucheteux & King, 2022; Goldstein et al., 2022; Schrimpf et al., 2021; Tuckute et al., 2023, *inter alia*). Relevant to our focus, i.e., linking representational dimensionality and feature abstraction, Cheng et al. (2025, 2026) found that a mid-layer dimensionality peak in LLMs marks a phase of higher-order linguistic processing (as opposed to surface-level processing). Our results qualitatively align with these studies, where we similarly found that *higher* dimensionality tracks richer processing of speech; we note, however, that unlike Cheng et al., low-level features remained decodable for a broad range of electrode sites, and our analysis concerns a different level of granularity, namely single electrode timecourses instead of holistic processing of multi-word inputs by an LLM. Finally, it is known in LLMs that weight matrix rank and representational dimensionality increase over training (Abbe et al., 2023; Cheng et al., 2025). A similar signature has been found for the representation of lexical semantic meaning in a human fMRI study (Zhang et al., 2024), where words in participants' native language (Mandarin Chinese) elicited higher dimensional representations than the same words in a less proficient language (English). How learning language over time dynamically affects its representational dimensionality in the brain, however, remains an open question.

Appendix A. Declaration of generative AI use

During the preparation of this work the author(s) used ChatGPT in order to polish the plotting code. AI was not used for data processing nor writing. After using these tools, the author) reviewed and edited the content as needed and take full responsibility for the content of the published article.

Appendix B. Data and code availability

We will publicly release the code on GitHub at publication. Code built upon the following public data and software:

BrainTreeBank https://braintreebank.dev/; license: cc 4.0

NoRaRe Database https://norare.clld.org/; license: cc 4.0

Compute. Experiments were entirely CPU-based and parallelized on a cluster equipped with 80 CPU cores. Extracting the high-gamma power representations took roughly 5 wall-clock days. Computing the effective dimension and the rolling RSA each took roughly 3 days. Decoding features were by far the most time-consuming, taking roughly two weeks. All other computation was negligible.

Appendix C. Fraction of stimuli time devoted to language

	Stimuli	Time language fraction
0	ant-man	0.31
1	aquaman	0.23
2	avengers-infinity-war	0.24
3	black-panther	0.25
4	cars-2	0.49
5	coraline	0.32
6	fantastic-mr-fox	0.38
7	guardians-of-the-galaxy	0.29
8	guardians-of-the-galaxy-2	0.30
9	incredibles	0.40
10	lotr-1	0.25
11	lotr-2	0.26
12	megamind	0.46
13	sesame-street-episode-3990	0.40
14	shrek-the-third	0.41
15	spider-man-3-homecoming	0.34
16	spider-man-far-from-home	0.34
17	the-martian	0.32
18	thor-ragnarok	0.30
19	toy-story	0.41
20	venom	0.27

Table C.1: **Fraction of stimuli time devoted to language**. This is the fraction of the total movie running time spent in language segments, i.e., while dialogue is being spoken.

Appendix D. Preprocessing

We preprocess the data in line with previous work (Desbordes et al., 2023; Regev et al., 2024). We first remove channels corresponding to computer-induced direct current (DC) triggers. Then, we apply a high-pass filter at 0.5Hz to counter drift (Regev et al., 2024), and remove line noise using notch filters at 60Hz and harmonics up to and including 360Hz. For each channel, we clip voltage signals to ± 5 standard deviations (Desbordes et al., 2023), where the standard deviation is computed with respect to the overall time course within each movie. To mitigate effects of local spatial correlation between channels, we apply Laplacian

referencing as in Wang et al. (2023), where each channel's signal is demeaned with respect to its neighbors on the same wire. Finally, like in prior work (Regev et al., 2024), we consider the high-gamma power ($70-150$Hz) by averaging the absolute Hilbert transform over several frequency bands ($[70,80]\cdots[140,150]$).

Appendix E. Speech-responsive electrodes subset

Language localizer paradigms isolate language-selective channels (or voxels) of interest by comparing their neural activity under contrast conditions. In the literature, this may entail measuring neural activity as participants listen to well-formed sentences vs., for instance, Jabberwocky (syntactically plausible sequences of non-words), and then statistically testing whether the two average neural responses differ (Fedorenko et al., 2024; Regev et al., 2024). Then, if for a given channel the two responses differ with high statistical significance, that voxel or channel is considered *language-selective*.

BrainTreeBank is a naturalistic dataset gathered while participants watched movies (Wang et al., 2024). As such, a language localizer experiment using contrast sets was not performed during data collection. Instead, we detect speech-responsive electrodes by segmenting out speech vs. non-speech contrasts ourselves (movie portions without dialogue). We then compare our speech-localized subset to the publicly available language-selective set from (Wang et al., 2024), finding modest overlap, potentially due to the different signal used (broadband voltage in the case of (Wang et al., 2024)) and different selection criteria.

Methods. For each participant and trial, we randomly sampled 1000 timestamps of the word-segmented and non-speech-segmented representations. Then, for each electrode, we averaged the high-γ power over these 1000 datapoints and conducted a single-tailed Mann-Whitney U test with the null hypothesis that the two conditions' mean high-γ are equal. We then perform False Discovery Rate correction for multiple testing to obtain the final p-values for each electrode's speech selectivity. The p-value cutoff for inclusion in the speech-responsive subset was $\alpha = 0.05$. Finally, for each participant, we selected the electrodes that were in the speech-responsive subset for *all* trials. The final FDR-corrected p-values for each participant are shown in Figure E.1.

Agreement with Wang et al. (2024). There was modest agreement with the *word-responsive* subset identified in Wang et al. (2024). Wang et al. (2024) identify this subset using the broadband voltage signal by considering 100ms windows before and after word onset, spanning windows that start at $\{-500,-400,\cdots,-100\}$ to

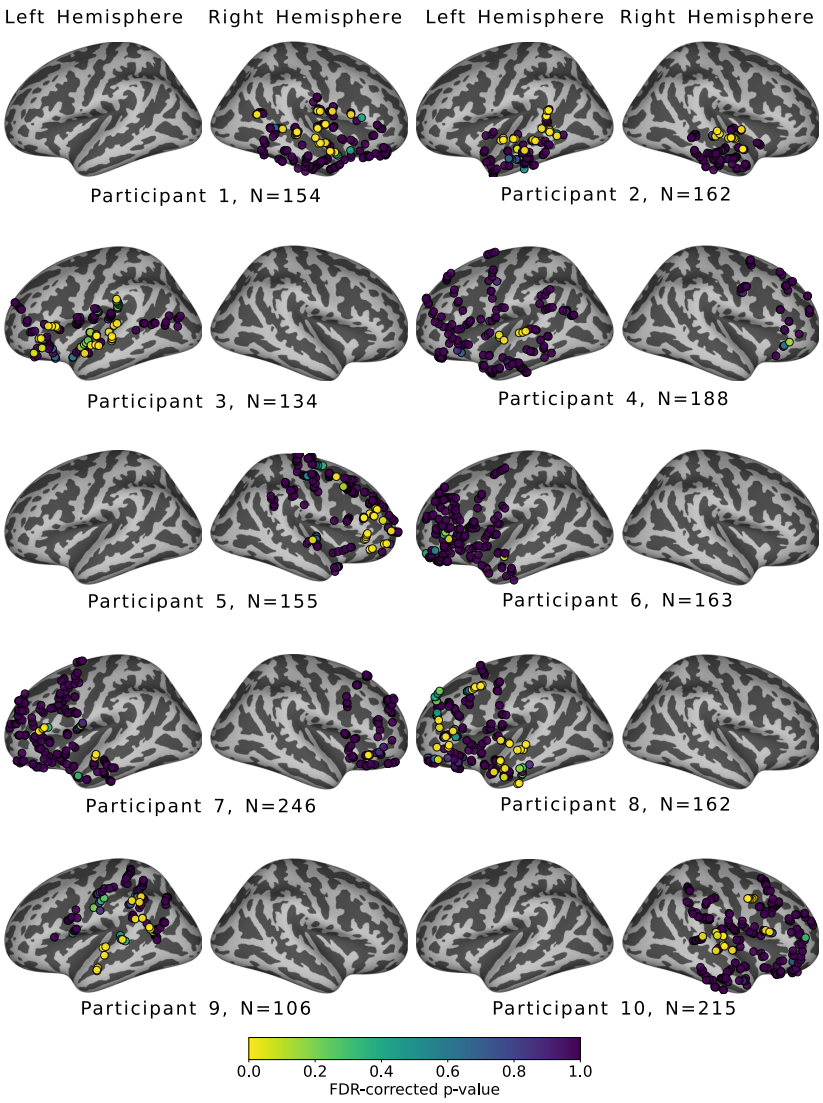


Figure E.1: **Speech subset per-participant breakdown.** For each participant, we plot the speech-responsive subset as found by our localizer. Channels are colored by their p-value (lower is more speech-responsive) as determined by a one-sided Mann-Whitney U test that compares the raw high-gamma power of $N = 1000$ random speech timestamps and $N = 1000$ random non-speech timestamps. Speech-responsive electrodes (yellow) are found in a temporal and frontal brain areas typically associated with speech and language processing.

$\{500, \cdots 1000\}$ms. If mean activity in any pre-onset window was significantly lower than any post-onset window with a two-tailed paired t-test, then the channel was considered word-responsive.

Of the 251 electrodes identified by Wang et al. (2024)'s method, the overlap with our speech subset was 68 electrodes. The speech subset additionally identified 107 electrodes that were selective for speech vs. non-speech sounds. Due to potential differences in preprocessing, especially the use of the broadband voltage as compared to our high-gamma power, we focused on the subset found by our method.

Appendix F. Per-participant distribution of speech vs. non-speech effective dimension

Subject	Speech > Nonspeech (%)	Not significant (%)	Speech < Nonspeech (%)	N_{elec}
1	**94.2**	3.9	1.9	154
2	**72.8**	21.6	5.6	162
3	**75.4**	23.9	0.7	134
4	36.2	16.0	**47.9**	188
5	**95.5**	0.6	3.8	156
6	**100.0**	0.0	0.0	164
7	**93.9**	2.0	4.1	246
8	**85.2**	10.5	4.3	162
9	**92.5**	4.7	2.8	106
10	**78.2**	12.0	9.7	216
Speech subset	**66.7**	24.2	15.2	165
Total	**81.8**	9.3	8.9	1688

Table F.1: **Per-participant percent of electrodes with higher or lower effective dimension during speech.** From left to right, we report the percentage of electrodes whose mean effective dimension over the first 500ms after word onset was *higher*, *not significant*, and *lower* during speech than non-speech, as determined by a nonparametric Mann-Whitney U test with $\alpha = 0.05$. Within each participant, all p-values were Benjamini-Hochberg FDR corrected. For all participants except Participant 4, the overwhelming majority of electrodes had higher effective dimension during speech than without speech.

Appendix G. Decoding features

We considered twelve speech features summarized in Table G.1. These features were selected from a broader initial set of 17 features compiled from the BrainTreeBank annotations (Wang et al., 2024) and from Tjuka et al. (2022), a public database that aggregates datasets of psycholinguistic norms listed on the right hand side of the Table. We dropped five features due to low sample size or conceptual closeness to existing features (e.g., we originally considered imageability, but this was too close to concreteness). We considered two acoustic (pitch and volume), surface-lexical (word length, phoneme count), lexical-semantic (content/function part of speech, concreteness, iconicity, valence, and arousal), syntactic (word position in sentence, distance to syntactic head) and information-theoretic (GPT2 surprisal).

Many of these features, especially the lexico-semantic ones, were interrelated. Figure G.1 plots their fraction overlap between labels (top or bottom quartile),

Feature	Type	Source
Pitch	acoustic	Wang et al. (2024)
Volume	acoustic	Wang et al. (2024)
Word length	lexical-surface	Wang et al. (2024)
# phonemes	lexical-surface	Wang et al. (2024)
Concreteness	lexical-semantic	Brysbaert et al. (2014)
Iconicity	lexical-semantic	Winter et al. (2024)
Valence	lexical-semantic	Mohammad (2025)
Arousal	lexical-semantic	Mohammad (2025)
Dist. to head	syntactic	Wang et al. (2024)
Word position	syntactic	Wang et al. (2024)
GPT2 Surprisal	statistical	Wang et al. (2024)

Table G.1: **Decoding features.**

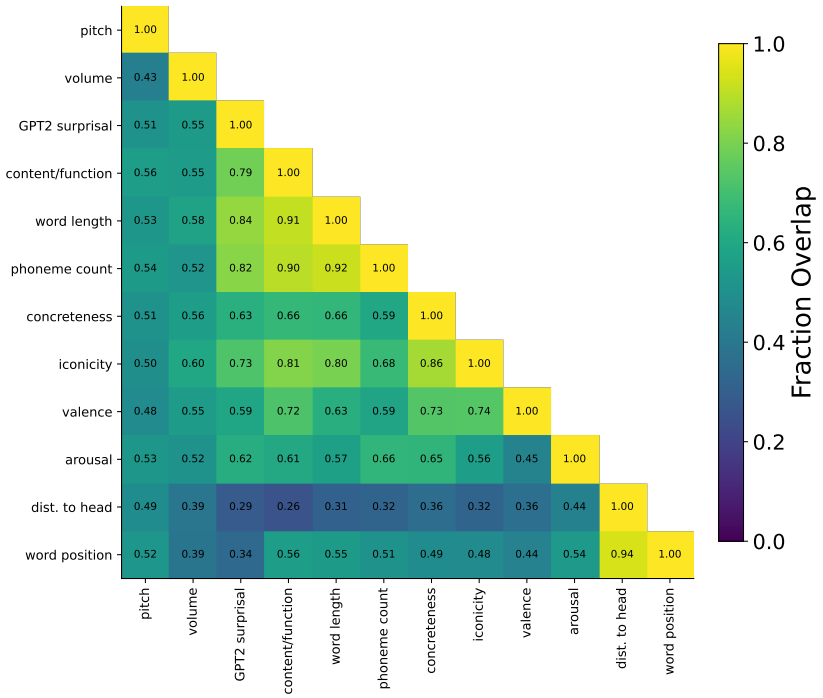


Figure G.1: **Overlap between features.** For each pair of features, the fraction of label overlap between the top and bottom quartiles is shown, i.e., the fraction of entries for which the two features have identical label assignments (chance overlap = 25%).

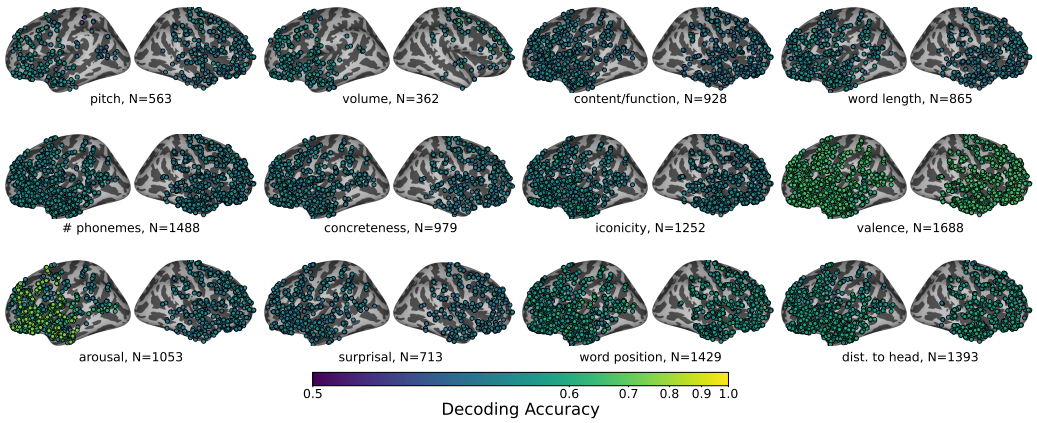


Figure G.2: **The participant-aggregated decoding performance for electrodes performing above chance.** For each feature, we plot the locations of electrodes for which some timepoint in word onset $\pm$ 1 second was decodable above chance, as determined by a one-sided t-test ($\alpha = 0.05$). Most features were decodable only modestly above chance (0.5), exceptions being valence, arousal, and word position, whose test accuracy was closer to 0.6.

computed as the size of the labels' intersection divided by the size of 2 quartiles. Though we note that each feature correlates to multiple other features, we retained all twelve of them, noting that, in practice, they are difficult to disentangle.

Decoding performance over features. For each feature, Figure G.2 shows the maximum test decoding performance over the time course, averaged over all $3 \times T$ cross-validation folds, where T is the number of movies a participant watched. The model was a binary logistic regression with L2 regularization ($\lambda = 1.0$), implemented with `sklearn`'s lbfgs solver for maximum 100000 iterations.

Similar to Wang et al. (2024) and Zahorodnii et al. (2025), who also perform linear decoding experiments on the same dataset, average decoding performance for these features was not high, hovering around 0.5-0.6 (cf. *raw voltage* decoding experiments in (Zahorodnii et al., 2025)). Certain features such as volume, moreover, tended to overfit, resulting in final test accuracies that may be less than 50% (by construction, our chance level).

Computation of β *similarity score.* For each timestep of the feature decoding performance trajectories, we performed a one-sided t-test comparing the test accuracy to 50% ($\alpha = 0.05$) to determine whether the feature was significantly decodable above chance. We then aggregated the number of total features that were

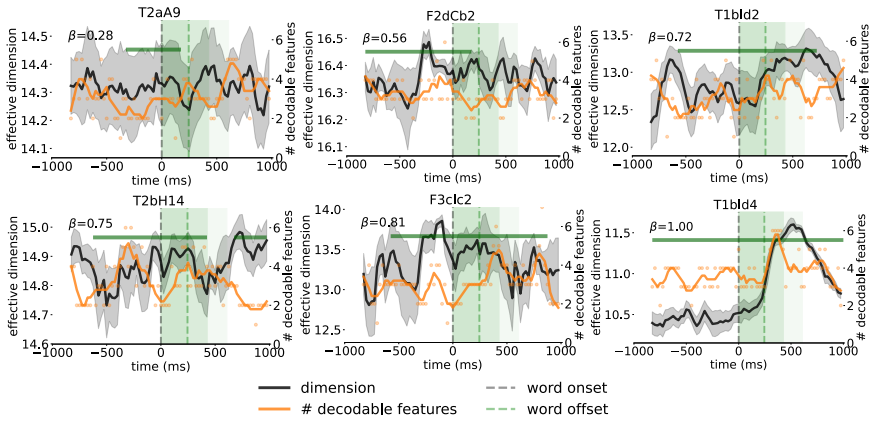


Figure G.3: **Example electrode dimensionality and number of decodable features trajectories, ranging from $\beta = 0$ to $\beta = 1$.** The interval used to calculate β, i.e., the interval where there was a positive correlation between dimensionality and number of decodable features, is marked as a dark green horizontal line.

significantly decodable above chance for each timestep, yielding a "decodability" time series. Because the number of decodable features is noisy and integer-valued, we first smoothed the decodability time series with a mean filter over each point $\pm$ two adjacent points (each point was 25ms apart). We applied the same filter to the dimensionality time series (also noisy) before correlating them.

Recall that, in computing β, we search for time intervals along which decodability and dimensionality positively correlate. To conduct this search, we computed sliding *local* Spearman correlations between the two time series. In particular, we considered the entire time course (word onset $\pm$ 1 second) and computed correlations on time windows of duration 150, 250, 500, 750, and 1000ms, starting at word onset -1 second and stepping along 50ms at a time; we also computed the *global* Spearman correlation over the entire time course.[5] P-values were Benjamini-Hochberg FDR-corrected over all comparisons of the same time window length for each electrode. Finally, to report the final β, we take the union of all such intervals and report their length divided by the entire time course. Several example electrode trajectories from Participant 10 with their corresponding β marked as a horizontal green line, are given in Figure G.3.

[5]We also tried Pearson correlation, Dynamic Time Warping, and Earth Mover's Distance. Ultimately, we opted for Spearman correlation because it is simple, amenable to significance testing, and does not incorporate extra assumptions such as linearity (cf. Pearson).

Appendix G.1. Per-participant breakdown of β

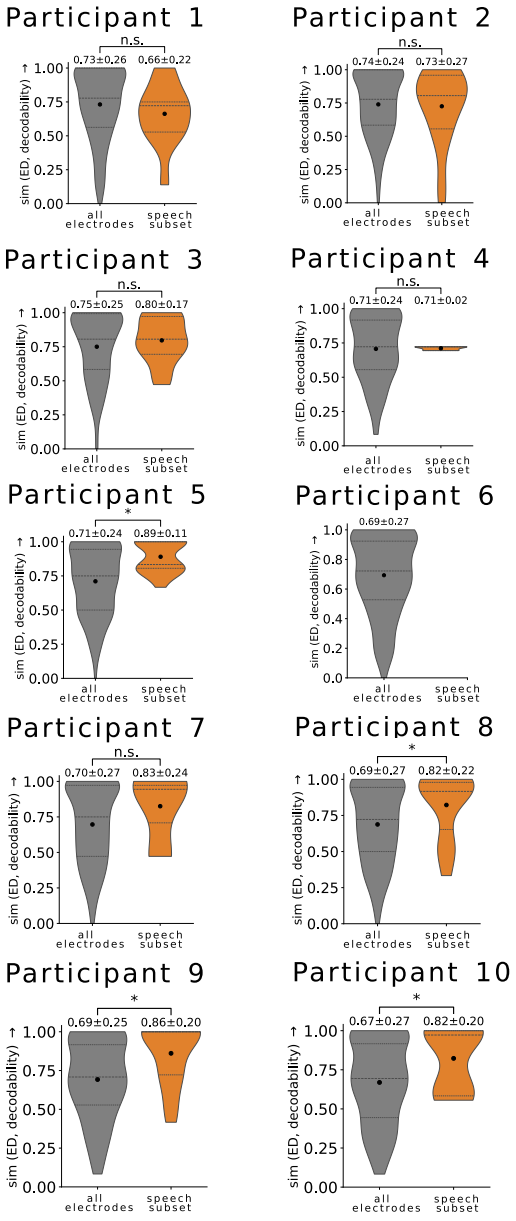


Figure G.4: **Per-participant breakdown of the similarity between dimensionality and decodability (β).** The distribution of β is shown for all electrodes (left) and the speech subset (right). Participant 6 did not have any speech-responsive electrodes at a significance of $\alpha = 0.05^*$. Results per-participant mirror the global distribution in Figure 3B: for all electrodes, β (y-axis) is high, and for the speech subset, β is typically similar or higher. The difference was significant for $4/10$ participants, seen by a Mann-Whitney U test ($\alpha = 0.05$. Some of the non-significant tests were due to low sample size, e.g., Participant 7 had seven and Participant 4 had five speech-responsive electrodes.

Appendix H. Identifying peaked electrodes

When analyzing dimensionality trajectories computed on the speech subset, we noted two broad categories: (1) a dimensionality trajectory that resembles a post-word-onset peak, and (2) heterogeneous trajectories. We first manually filtered each electrode ($N = 1688$), annotating whether they fell into the peaked or non-peaked category. This process selected $N = 51$ peaked electrodes. We note that, in general, it was easy to determine whether an electrode would fall into the peaked or non-peaked category, as the post-word-onset peaks were quite visually distinctive. Then, on each electrode in the peaked set, we confirmed that "peakedness" was statistically significant via a permutation test. We measured peakedness by the following criterion:

1. First, because we are concerned only with shape and not amplitude, we normalize a given peaked dimensionality trajectory to $[0,1]$ to produce a time series x_t.
2. We define a template "peaked" timeseries against which to compare the trajectory. This is given by a right-skewed *exponentially modified Gaussian* (EMG) normalized to $[0,1]$, mean-centered at the electrode trajectory's maximum, $\sigma = \tau = \texttt{len}(x_t)/5$, where σ is the standard deviation and τ the skew parameter. σ and τ, selected heuristically, respectively control the width and the skew of the peak. Changing them slightly did not change the overall results of the permutation test, which determines the present trajectory's *relative* similarity to the EMG with respect to permutations.
3. We measured peakedness with the Pearson correlation R between the normalized dimensionality x_t and the EMG.
4. The p-value of the peakedness score is computed relative to $N = 2000$ permutations of the electrode's time series. Electrodes are considered significant if their p-value is less than $\alpha = 0.05$ after FDR-correction across all 1688 electrodes.

Due to autocorrelation in the time series, we additionally tried "blocked" permutations, a more restrictive method that does not fully destroys temporal structure, but rather shuffles contiguous blocks. Shuffling block sizes of 2 through 10 time steps (corresponding to 50-250ms) did not affect the results.

Directly applying the significance test as a criterion for selecting electrodes yielded 77 electrodes (p-value < 0.05, FDR-corrected across electrodes) that were a superset of the 51 manually annotated ones. The locations of the 77 electrodes are shown in Figure H.1, where they are mostly distributed in frontal and temporal

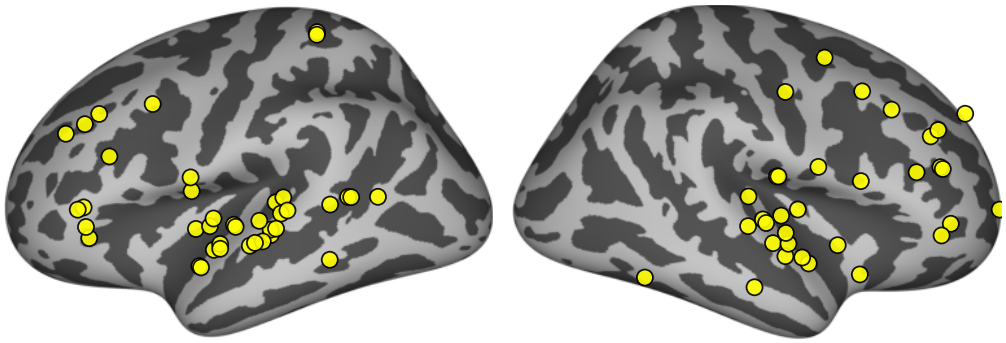


Figure H.1: **Electrodes found by statistical testing for peakedness ($N = 77$).** Statistical testing for peakedness by comparing electrodes' effective dimensions trajectories to a template exponentially modified Gaussian resulted in $N = 77$ electrodes ($\alpha = 0.05$, FDR-corrected over all 1688 electrodes). These 77 electrodes are largely distributed in frontal and temporal areas that broadly correspond to speech processing (Fedorenko et al., 2024), and include the set of 51 electrodes that we manually annotated as displaying a pronounced post-word-onset peak.

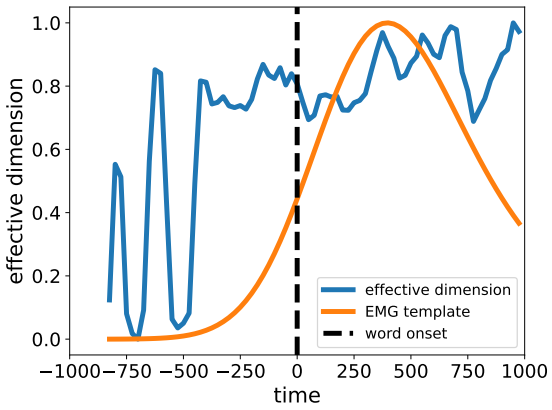


Figure H.2: **Example of an electrode identified by significance testing but that does not show a post-word-onset peak**. Electrode P2Ie16 from Participant 5's dimensionality trajectory is shown (blue), against the template exponentially-modified Gaussian (orange). The significance test returned $p < 0.05$. However, the effective dimension trajectory does not resemble a pronounced peak after word onset—this trajectory moreover contrasts with the peaked electrode trajectories shown in Figure 4A. We excluded this and similar electrodes from the peaked subset.

areas. We note, however, that the 26 electrodes found by the quantitative criterion but not manually annotated to be in the peaked set, while significantly similar to the Gaussian template, did not correspond to visual intuitions of a post-onset peak (see Figure H.2 for an example). For this reason, we privileged the manually selected subset in analysis, prioritizing precision over recall.

Appendix H.1. Peaked-dimensionality electrodes do not peak for non-speech sounds

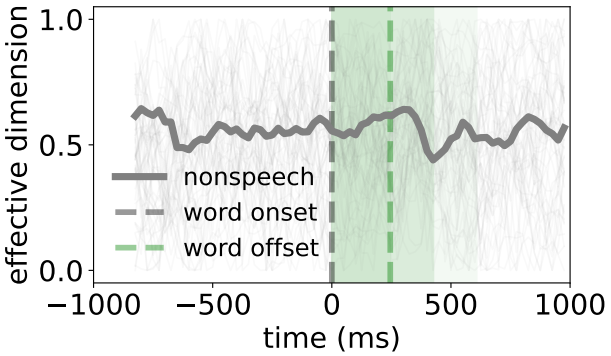


Figure H.3: **Average dimensionality trajectories in peaked electrodes over non-speech segments.** The dimensionality trajectory over non-speech segments is shown for the 51 electrodes whose dimensionality during speech segments showed a post-word onset peak. In the non-speech segments, there is no peak, indicating that the peak is due to the presence of speech.

Appendix I. Quantifying representational stability across time

We wanted to quantify how stable the representational space of a single electrode is across processing time. To do so, we first demeaned speech representations across the time dimension. Then, we computed *representational dissimilarity matrices* (RDMs) between the demeaned speech representations at subsequent timesteps, using the same representations (128ms timechunks) and step size (25ms) as in previous analyses. In brief, our RDMs are given by an $N \times N$ matrix where each of the N rows and columns corresponds to a word, and each entry represents the Euclidean distance between a given word's representation at time step t and a different word's representation at the same time step. Then, we quantify the change in representational geometry from time step t to $t+1$ via Representational Similarity Analysis (RSA) (Kriegeskorte & Kievit, 2013; Kriegeskorte et al., 2008), which takes the Spearman correlation between the lower-triangular part of the RDM at time step t and that at time step $t+1$. An RSA score close to 1

indicates that the representational geometry (as captured by pairwise cosine similarity) remains stable from one time step to the next; a score close to 0 indicates no correlation between subsequent representational geometries, and a score close to -1 indicates that representations close together at time step t are now far apart at time step $t+1$.

Appendix I.1. An RSA baseline based on random time lags

We wanted to compare the RSA scores from the above procedure (representations of 125ms and stepsize of 25ms), which compare representation spaces that close together in time, to representations further apart in time. To do so, for each electrode of each participant, we sampled $N = 10$ random time lags $t \sim \mathcal{N}(\mu, \sigma)$ per movie, where $\mu = 0$ and $\sigma = \frac{\text{length of movie}}{\sqrt{6}}$ were computed analytically for each movie as the theoretical expectation and standard deviation of the *difference* between two i.i.d. uniform time samples. Empirically, the *absolute difference* between time samples was $\mathbb{E}[T_1 - T_2] \approx 40$ minutes and $\sigma(T_1 - T_2) \approx 35$ minutes for $T_i \sim \text{Unif}[0, \text{length of movie (s)}]$, $i = 1, 2$. For a random representation in word onset ± 1 second, we computed the RSA between that representation and the representation at a time lag t away. Finally, we averaged the resulting RSA scores over all representation pairs to obtain the final random baseline score discussed in Section 4.3.

Appendix I.2. Absolute change in dimensionality correlates less to representational change

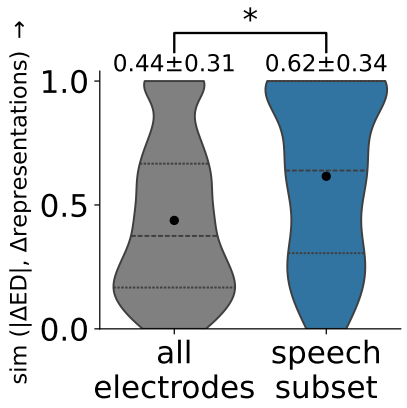


Figure I.1: **The absolute change in effective dimension correlates to speech representational instability for a moderate fraction of the word time course.**

Appendix J. A closer look at top-PC subspaces

Appendix J.1. Each PC may encode many features, and each feature is encoded by many PCs

Figure J.1A shows, for example electrode O1aIb9 of Participant 3, the evolution of the effective dimension (top) along with the evolution of the top six principal components (PCs) over word processing (bottom). Representational "snapshots" are shown in the heatmaps for word onset $-500, -375, \ldots, +500$ms, plotting the Spearman correlation between speech features (y-axis) and the top six PCs, from left to right in *decreasing* order by explained variance. In the figure, colorful squares denote a significant correlation at $\alpha = 0.05$, and gray squares otherwise. We remark that individual PCs, especially the top PC, encode multiple features at once: for example, at word onset $+250$ms, PC1 simultaneously encodes volume, iconicity, and word position in the sentence. Similarly, each feature may be encoded by more than one PC: for instance, *volume* at word onset $+250$ms is spread across PCs 1 and 2. Interestingly, each PC broadly encoded features throughout the speech hierarchy, from low-level acoustic information to lexical semantic and syntactic information. This observation holds true despite the fact that this electrode is located in the left superior temporal gyrus near sensory processing. Table J.1 shows that PCs encoding multiple features at once *generalized* across electrodes, where the average number of speech features simultaneously encoded at any time point for each PC was greater than one. These results align with existing work, which has found—on the level of the entire signal, not individual PCs—that local electrode sites dynamically encode information throughout the speech processing hierarchy (Gwilliams, Marantz, et al., 2025; Keshishian et al., 2023).

Appendix J.2. Representational dynamics are dominated by a single slower mode

A closer look at the temporal evolution of the top ($k = 6$) principal components (PCs) of example electrode O1aIb9, see Figure J.1B, reveals that the top PC accounts for most representational stability across time, while *smaller* PCs decay more quickly. In particular, each line in the plot reports absolute Spearman correlations between each PC and itself at subsequent time lags $[-500, \ldots, 500]$ms around word onset, with a stepsize of 125ms (darker blue means a *lower* PC).[6] Temporal variation in how top PCs at subsequent timesteps correlate to each other

[6] We report the *absolute value* of the Spearman correlation ρ between subsequent time steps, as whether a PC linearly represents information is robust to sign flips.

Electrode set		# speech features encoded	Min	Max
speech subset	PC1	2.66 ± 1.42	1	10
all	PC1	1.84 ± 1.14	1	10
speech subset	PC2	1.61 ± 0.93	1	8
all	PC2	1.44 ± 0.75	1	8
speech subset	PC3	1.48 ± 0.78	1	7
all	PC3	1.42 ± 0.74	1	8
speech subset	PC4	1.43 ± 0.73	1	6
all	PC4	1.40 ± 0.70	1	8
speech subset	PC5	1.43 ± 0.73	1	8
all	PC5	1.39 ± 0.69	1	8
speech subset	PC6	1.42 ± 0.72	1	8
all	PC6	1.39 ± 0.69	1	8

Table J.1: **Number of speech features encoded above chance by individual PCs.** We summarize how many speech features were significantly encoded by each of the top six principal component (PC) across all electrodes of all participants, timestamps, and trials. For each PC, we computed a distribution over samples in which each sample corresponds to a single electrode–trial–time point. A feature was counted as encoded if its correlation with the PC was statistically significant ($p < 0.05$). We then summarized the resulting distributions, *restricting analysis to data points where at least one feature was encoded*, by the mean number of features $\pm$1SD (third column), minimum, and maximum number of encoded features per sample. This analysis was performed separately for all electrodes and the speech subset, where, on average, PCs in the speech subset encode more distinct features than the average electrode (top row higher than bottom row for each PC).

is reflected in the content they encode. Word onset elicits a period of dimensionality expansion and rapid local change in PC1 (Figure J.1B), and this change is reflected in Figure J.1A by the fading of syntactic information (distance to head and word position) from -250ms to 125ms and the appearance of acoustic information that peaks around $+250$ms.

Similar dynamics generalize across electrodes. Figure J.1C shows that across participants, movies, and electrodes, PC1 modestly autocorrelates across timesteps: the average absolute Spearman correlation over trials, electrodes, and timesteps was $|\rho(\text{PC1}_t, \text{PC1}_{t+125ms})| \approx 0.25$ (SD 0.19). Instead, PCs 2-6 showed a near-zero average autocorrelation: $|\rho| = 0.06, 0.06, 0.05, 0.04, 0.04$, respectively for PCs 2-6. Interestingly, the *average* autocorrelation for each PC remained consistent across time steps within word onset $\pm$1s, even when crossing word boundaries. However, this pattern differed across electrode type; in particular, PC1 in speech-responsive electrodes showed a clear dip in autocorrelation after word onset, see Figure J.1C. Despite the low autocorrelations in the higher PCs, it is not necessarily the case that they encode noise. The average percent-variance explained by the PC1 through PC6 across timesteps, electrodes, trials, and participants is 23.1 ± 18.6, 10.4 ± 3.5, 8.6 ± 2.4, 7.5 ± 2.0, 6.6 ± 1.8, and $5.8 \pm 1.6\%$, respectively, where PCs 2-4, for instance, still encode a sizeable amount of the total variance.

Finally, we verified that a given speech representation PC and the same PC index of a random lagged representation during the movie had a correlation of ≈ 0; this implies that correlations thus far indeed reflect locally autocorrelated speech processing rather than some computational artifact.

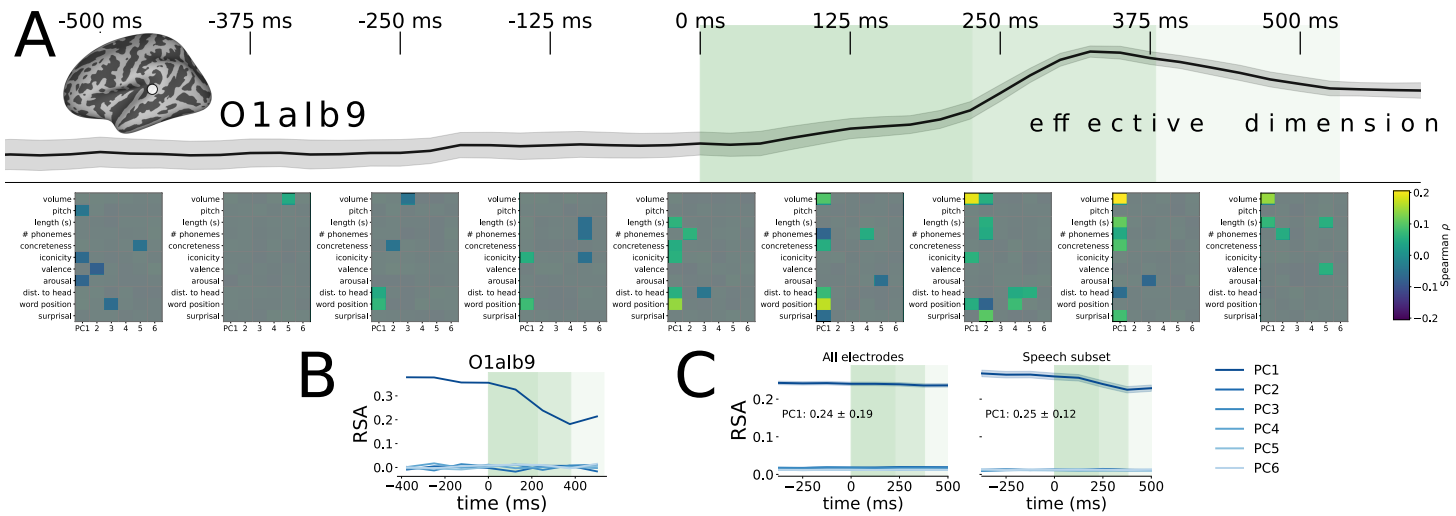


Figure J.1: **Representational dynamics are dominated by the top principal component. (A) Evolution of top-PC space for electrode O1aIb9 of Participant 3.** For each time in $[-500, -375, \cdots, 375, 500]$ms around word onset, a "representational snapshot" shows the Spearman correlations between speech features (y-axis) and individual PCs (x-axis). Gray cells denote correlations that were not significant at $\alpha = 0.05$. Over the time course of a word, word-relevant information emerges and fades, notably syntactic information (word position and distance to syntactic head) around word onset and acoustic information (volume) between roughly 125 to 500ms. **(B) PC autocorrelations of electrode O1aIb9 by top principal component.** PC*k*-to-PC*k* autocorrelations computed on time lags of 125ms corresponding to the representational snapshots. Higher PCs (lighter blue) do not autocorrelate at a time scale of 125ms, but PC1 shows an autocorrelation around $\rho = 0.3$. Notably, the PC1 autocorrelation *decreases* after word onset. **(C) Top-PC autocorrelations averaged across electrodes.** Across the set of all electrodes (left) and speech subset (right), the top PC shown in dark blue is most autocorrelated at the considered time lag, while higher PCs are not autocorrelated. The top PC in speech-responsive electrodes, however, shows a post-word–onset dip, suggesting that this PC may process new information relevant to the current word.

References

Abbe, E., Bengio, S., Boix-Adserà, E., Littwin, E., & Susskind, J. M. (2023). Transformers learn through gradual rank increase. *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=qieeNlO3C7

Abbott, L. F., Rajan, K., & Sompolinsky, H. (2011, January). Interactions between intrinsic and stimulus-evoked activity in recurrent neural networks. In M. Ding PhD & D. Glanzman PhD (Eds.), *The dynamic brain: An exploration of neuronal variability and its functional significance* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195393798.003.0004

Antonello, R., Turek, J. S., Vo, V. A., & Huth, A. (2021). Low-dimensional structure in the space of language representations is reflected in brain responses. In A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems*. https://openreview.net/forum?id=UYI6Sk_3Nox

Badre, D., Bhandari, A., Keglovits, H., & Kikumoto, A. (2021). The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, *38*, 20–28. https://doi.org/10.1016/j.cobeha.2020.07.002

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, *183*(4), 954–967.e21. https://doi.org/10.1016/j.cell.2020.09.031

Botch, T. L., & Finn, E. S. (2024). Neural representations of concreteness and concrete concepts are specific to the individual. *Journal of Neuroscience*, *44*(45).

Brincat, S. L., Siegel, M., von Nicolai, C., & Miller, E. K. (2018). Gradual progression from sensory to task-related processing in cerebral cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(30), E7202–E7211. https://doi.org/10.1073/pnas.1717075115

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, *28*(24), 3976–3983.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Cai, X., Huang, J., Bian, Y., & Church, K. (2021). Isotropy in the contextual embedding space: Clusters and manifolds. *International Conference on Learning Representations*.

Canatar, A., Feather, J., Wakhloo, A., & Chung, S. (2023). A spectral theory of neural prediction and alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (pp. 47052–47080, Vol. 36). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/9308d1b7d4ae2d3e2e67ae94b1078bf7-Paper-Conference.pdf

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134. https://doi.org/10.1038/s42003-022-03036-1

Chen, C., Dupré la Tour, T., Gallant, J. L., Klein, D., & Deniz, F. (2024). The cortical representation of language timescales is shared between reading and listening. *Communications Biology*, *7*(1), 284. https://doi.org/10.1038/s42003-024-05909-z

Chen, Z., Isik, L., & Bonner, M. F. (2026). Multidimensional dynamics of object representations in the human visual system, 2026.04.27.720701. https://doi.org/10.64898/2026.04.27.720701

Cheng, E., Doimo, D., Kervadec, C., Macocco, I., Yu, J., Laio, A., & Baroni, M. (2025). Emergence of a high-dimensional abstraction phase in language transformers. *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=0fD3iIBhlV

Cheng, E., Vaidya, A., & Antonello, R. (2026). Abstraction induces the brain alignment in language and speech models. *arXiv*.

Chung, S., Lee, D. D., & Sompolinsky, H. (2018). Classification and geometry of general perceptual manifolds. *Phys. Rev. X*, *8*, 031003.

Cohen, U., Chung, S., Lee, D. D., & Sompolinsky, H. (2020). Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, *11*(1), 746. https://doi.org/10.1038/s41467-020-14578-5

Cunningham, J. P., & Yu, B. M. (2014). Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, *17*(11), 1500–1509. https://doi.org/10.1038/nn.3776

de Vries, I. E. J., & Wurm, M. F. (2023). Predictive neural representations of naturalistic dynamic input. *Nature Communications*, *14*(1), 3858. https://doi.org/10.1038/s41467-023-39355-y

Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., & King, J.-R. (2023). Decoding speech perception from non-invasive brain recordings. *Nature Ma-*

chine Intelligence, *5*(10), 1097–1107. https://doi.org/10.1038/s42256-023-00714-5

Del Giudice, M. (2021). Effective dimensionality: A tutorial. *Multivariate Behavioral Research*, *56*, 527–542. https://doi.org/10.1080/00273171.2020.1743631

Desbordes, T., Lakretz, Y., Chanoine, V., Oquab, M., Badier, J.-M., Trébuchon, A., Carron, R., Bénar, C.-G., Dehaene, S., & King, J.-R. (2023). Dimensionality and ramping: Signatures of sentence integration in the dynamics of brains and deep language models. *The Journal of Neuroscience*, *43*(29), 5350–5364. https://doi.org/10.1523/JNEUROSCI.1163-22.2023

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341. https://doi.org/10.1016/j.tics.2007.06.010

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, *25*(5), 289–312. https://doi.org/10.1038/s41583-024-00802-4

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, *113*(41), E6256–E6262.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013, August). Word surprisal predicts n400 amplitude during reading. In H. Schuetze, P. Fung, & M. Poesio (Eds.), *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 878–883). Association for Computational Linguistics. https://aclanthology.org/P13-2152/

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The erp response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Friederici, A. D. (2011). The brain basis of language processing: From structure to function [PMID: 22013214]. *Physiological Reviews*, *91*(4), 1357–1392. https://doi.org/10.1152/physrev.00006.2011

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010

Gadonneix, J., Zhang, M., Rapin, J., Evanson, L., Bourdillon, P., & King, J.-R. (2026). Temporal structure of the language hierarchy within small cortical patches. https://arxiv.org/abs/2604.03021

Galella, S., Wehrheim, M., & Kaschube, M. (2025). Dimensionality mismatch between brains and artificial neural networks. *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=fyp34w19N2

Ganguli, S., & Sompolinsky, H. (2012). Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annual review of neuroscience*, *35*, 485–508. https://api.semanticscholar.org/CorpusID:13860379

Gao, P., & Ganguli, S. (2015). On simplicity and complexity in the brave new world of large-scale neuroscience. *Current Opinion in Neurobiology*, *32*, 148–155. https://doi.org/https://doi.org/10.1016/j.conb.2015.04.003

Gao, P., Trautmann, E. M., Yu, B. M., Santhanam, G., Ryu, S. I., Shenoy, K. V., & Ganguli, S. (2017). A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*. https://api.semanticscholar.org/CorpusID:19938440

Gauthaman, R. M., Ménard, B., & Bonner, M. F. (2025). Universal scale-free representations in human visual cortex. *PLOS Computational Biology*, *21*(11), e1013714. https://doi.org/10.1371/journal.pcbi.1013714

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Gwilliams, L., Bhaya-Grossman, I., Zhang, Y., Scott, T., Harper, S., & Levy, D. (2025). Computational architecture of speech comprehension in the human brain. *Annual Review of Linguistics*, *11*(1), 209–226.

Gwilliams, L., Marantz, A., Poeppel, D., & King, J.-R. (2025). Hierarchical dynamic coding coordinates speech comprehension in the human brain. *Proceedings of the National Academy of Sciences*, *122*(42), e2422097122. https://doi.org/10.1073/pnas.2422097122

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. https://aclanthology.org/N01-1021/

Harvey, S. E., Lipshutz, D., & Williams, A. H. (2024). What representational similarity measures imply about decodable information. *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*. https://openreview.net/forum?id=hqfzH6GCYj

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *28*(10), 2539–2550. https://doi.org/10.1523/JNEUROSCI.5487-07.2008

Jazayeri, M., & Ostojic, S. (2021). Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Current Opinion in Neurobiology*, *70*, 113–120. https://doi.org/10.1016/j.conb.2021.08.002

Keshishian, M., Akkol, S., Herrero, J., Bickel, S., Mehta, A. D., & Mesgarani, N. (2023). Joint, distributed and hierarchically organized encoding of linguistic features in the human auditory cortex. *Nature Human Behaviour*, *7*(5), 740–753. https://doi.org/10.1038/s41562-023-01520-0

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://doi.org/10.3389/neuro.06.004.2008

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Litwin-Kumar, A., Harris, K. D., Axel, R., Sompolinsky, H., & Abbott, L. F. (2017). Optimal degrees of synaptic connectivity. *Neuron*, *93*(5), 1153–1164.e7. https://doi.org/10.1016/j.neuron.2017.01.030

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature neuroscience*, *25*(8), 1014–1019.

Mohammad, S. M. (2025). Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms. https://arxiv.org/abs/2503.23547

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S. S., Naccache, L., Hale, J. T., Pallier, C., et al. (2017). Neurophysiologi-

cal dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, *114*(18), E3669–E3678.

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *neuron*, *88*(6), 1281–1296.

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature neuroscience*, *18*(6), 903–911.

Parthasarathy, A., Tang, C., Herikstad, R., Cheong, L. F., Yen, S.-C., & Libedinsky, C. (2019). Time-invariant working memory representations in the presence of code-morphing in the lateral prefrontal cortex. *Nature Communications*, *10*(1), 4995. https://doi.org/10.1038/s41467-019-12841-y

Posani, L., Wang, S., Muscinelli, S. P., Paninski, L., & Fusi, S. (2025). Rarely categorical, always high-dimensional: How the neural code changes along the cortical hierarchy. *bioRxiv*, 2024.11.15.623878. https://doi.org/10.1101/2024.11.15.623878

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Ray, S., & Maunsell, J. H. R. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS biology*, *9*(4), e1000610. https://doi.org/10.1371/journal.pbio.1000610

Recanatesi, S., Farrell, M., Advani, M., Moore, T., Lajoie, G., & Shea-Brown, E. (2019). Dimensionality compression and expansion in deep neural networks [arXiv:1906.00443 [cs, stat]]. https://doi.org/10.48550/arXiv.1906.00443

Regev, T. I., Casto, C., Hosseini, E. A., Adamek, M., Ritaccio, A. L., Willie, J. T., Brunner, P., & Fedorenko, E. (2024). Neural populations in the language network differ in the size of their temporal receptive windows. *Nature Human Behaviour*, *8*(10), 1924–1942. https://doi.org/10.1038/s41562-024-01944-2

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590. https://doi.org/10.1038/nature12160

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: A defense of functional localizers. *Neuroimage*, *30*(4), 1088–1096.

Schaeffer, R., Khona, M., Chandra, S., Ostrow, M., Miranda, B., & Koyejo, S. (2024). Does maximizing neural regression scores teach us about the brain?

UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models. https://openreview.net/forum?id=vbtj05J68r

Schönmann, I., Szewczyk, J., de Lange, F. P., & Heilbron, M. (2026). Stimulus dependencies—rather than next-word prediction—can explain pre-onset brain encoding in naturalistic listening designs (N. Ding & H. Luo, Eds.). *eLife*, *14*, RP106543. https://doi.org/10.7554/eLife.106543

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), e2307876121. https://doi.org/10.1073/pnas.2307876121

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., & Harris, K. D. (2019). High-dimensional geometry of population responses in visual cortex. *Nature*, *571*(7765), 361–365. https://doi.org/10.1038/s41586-019-1346-5

Tjuka, A., Forkel, ·., & List, J.-M. (2022). Linking norms, ratings, and relations of words and concepts across multiple language varieties. *Behavior Research Methods*, *54*, 864–884. https://doi.org/10.3758/s13428-021-01650-1

Tuckute, G., Feather, J., Boebinger, D., & McDermott, J. H. (2023). 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), 1–70. https://doi.org/10.1371/journal.pbio.3002366

Tuckute, G., Lee, E. J., Ou, Y., Fedorenko, E., & Kay, K. (2025). A two-dimensional space of linguistic representations shared across individuals, 2025.05.21.655330. https://doi.org/10.1101/2025.05.21.655330

Valeriani, L., Doimo, D., Cuturello, F., Laio, A., ansuini, A., & Cazzaniga, A. (2023). The geometry of hidden representations of large transformer models. *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=cCYvakU5Ek

Wang, C., Subramaniam, V., Yaari, A. U., Kreiman, G., Katz, B., Cases, I., & Barbu, A. (2023). BrainBERT: Self-supervised representation learning for intracranial recordings. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=xmcYx_reUn6

Wang, C., Yaari, A., Singh, A. K., Subramaniam, V., Rosenfarb, D., DeWitt, J., Misra, P., Madsen, J. R., Stone, S., Kreiman, G., Katz, B., Cases, I., &

Barbu, A. (2024). Brain treebank: Large-scale intracranial recordings from naturalistic language stimuli. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in neural information processing systems* (pp. 96505–96540, Vol. 37). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2024/file/aefa2385b3f33abf1526ae4e2c208cd9-Paper-Datasets_and_Benchmarks_Track.pdf

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470. https://doi.org/10.1162/tacl_a_00612

Winter, B., Lupyan, G., Perry, L. K., Dingemanse, M., & Perlman, M. (2024). Iconicity ratings for 14,000+ english words. *Behavior Research Methods*, *56*(3), 1640–1655.

Woolnough, O., Donos, C., Murphy, E., Rollo, P. S., Roccaforte, Z. J., Dehaene, S., & Tandon, N. (2023). Spatiotemporally distributed frontotemporal networks for sentence reading. *Proceedings of the National Academy of Sciences*, *120*(17), e2300252120.

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron*, *102*(6), 1096–1110.

Zahorodnii, A., Wang, C., Stankovits, B., Moraitaki, C., Chau, G., Barbu, A., Katz, B., & Fiete, I. R. (2025). Neuroprobe: Evaluating intracranial brain responses to naturalistic stimuli. https://arxiv.org/abs/2509.21671

Zhang, J., Li, H., Qu, J., Liu, X., Feng, X., Fu, X., & Mei, L. (2024). Language proficiency is associated with neural representational dimensionality of semantic concepts. *Brain and Language*, *258*, 105485. https://doi.org/10.1016/j.bandl.2024.105485