Neural manifolds carry reactivation of phonetic representations during semantic processing

Pavo Orepic[1], Wilson Truccolo[2,3], Eric Halgren[4], Sydney S. Cash[5], Anne-Lise Giraud[1,6], Timothée Proix[1*]

[1]Department of Basic Neurosciences, Faculty of Medicine, University of Geneva, Geneva, Switzerland.
[2]Department of Neuroscience, Brown University, Providence, Rhode Island, United States of America.
[3]Carney Institute for Brain Science, Brown University, Providence, Rhode Island, United States of America.
[4]Department of Neuroscience & Radiology, University of California San Diego, La Jolla, California, United States of America.
[5]Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, United States of America.
[6]Institut Pasteur, Université Paris Cité, Hearing Institute, Paris, France.

*Corresponding author(s). E-mail(s): timothee.proix@unige.ch;

Abstract

Traditional models of speech perception posit that neural activity encodes speech through a hierarchy of cognitive processes, from low-level representations of acoustic and phonetic features to high-level semantic encoding. Yet it remains unknown how neural representations are transformed across levels of the speech hierarchy. Here, we analyzed unique microelectrode array recordings of neuronal spiking activity from the human left anterior superior temporal gyrus, a brain region at the interface between phonetic and semantic speech processing, during a semantic categorization task and natural speech perception. We identified distinct neural manifolds for semantic and phonetic features, with a functional separation of the corresponding low-dimensional trajectories. Moreover, phonetic and semantic representations were encoded concurrently and reflected in power increases in the beta and low-gamma local field potentials, suggesting top-down predictive and bottom-up cumulative processes. Our results are the first to demonstrate

mechanisms for hierarchical speech transformations that are specific to neuronal population dynamics.

Keywords: speech perception, neural manifold, neuronal population dynamics, phonetic, semantic, neural oscillations, analysis-by-synthesis

Introduction

How does the brain transform sounds into meanings? Most cognitive models of speech perception propose that speech sounds are processed sequentially by distinct neural modules. Semantic and conceptual representations emerge at the end of the ventral stream (i.e. the "what" stream) (Hickok and Poeppel, 2007), where perceived speech is sequentially transformed from spectro-temporal encoding of sounds in Heschl's gyrus, over phonetic features in the superior temporal gyrus (STG) and sulcus, to lexical and combinatorial semantics in the anterior temporal lobe (ATL) (Mesgarani et al, 2014; Pylkkänen, 2020). For instance, the ATL is considered a specialized semantic module, connected to modality-specific sources of information (Patterson et al, 2007; Ralph et al, 2017), with causal semantic impairments occurring after bilateral ATL atrophy or damage due to viral infection (Noppeney et al, 2006; Lambon Ralph et al, 2006; Schwartz et al, 2009; Cope et al, 2020). This modular and sequential perspective implies that a complex neural process transforms speech features from one functional brain region to the next up to lexical and multimodal conceptual representations.

Recent neuroimaging findings suggest, however, that these transformations across brain regions might be less modular and sequential than predicted by the classical hierarchical view of language processing (Ralph et al, 2017; Yi et al, 2019; Caucheteux et al, 2022). Fine-grained electrocorticography (ECoG) recordings from the middle and posterior STG reveal that phonetic features are encoded without strict spatial segregation, but rather via mixed interleaved representations (Hamilton et al, 2021). At the semantic level, the anterior superior temporal gyrus (aSTG), as a part of the

ATL, responds more specifically to semantic decisions from heard speech, without a clear anatomo-functional separation from the middle STG (Scott, 2000; Visser and Lambon Ralph, 2011; Chang et al, 2015; Zhang et al, 2021; Damera et al, 2023). In functional MRI (fMRI) data, neural activity recorded through incrementally higher cognitive brain regions correlates with increasingly deeper layers in large language models (Caucheteux et al, 2022). These findings suggest that some brain regions represent multiple speech features, making them suitable candidates for housing transformations across the speech hierarchy.

Studies using time-resolved recording techniques such as EEG, MEG, or intracranial EEG, additionally showed simultaneous encoding of features across the speech hierarchy, organized in increasingly larger time scales (Heilbron et al, 2022; Gwilliams and King, 2020; Keshishian et al, 2023). While phonetic features are short-lived and typically processed around 100-200 ms in the STG (Mesgarani et al, 2014; Hamilton et al, 2021), semantic processing, e.g. semantic composition and lexical decisions, is associated with a longer-lasting compound event at 400 ms as reflected by the N400 component (Kutas and Federmeier, 2011; Dikker et al, 2020). Yet, syntactic and semantic composition activity can be detected in the ATL as early as 200-250 ms (Pylkkänen, 2020; Friederici and Kotz, 2003), suggesting an early use of phonetic features to access meaning, a process that has so far not been described.

Such early transformations are envisaged in the analysis-by-synthesis framework, which posits that speech comprehension results from a series of sequential loops (Halle and Stevens, 1959; Bever and Poeppel, 2010): incoming speech inputs are first sequentially processed, e.g. at the phonological level, and combined into a first semantic guess based on prior knowledge and context. A word prior elicited by this semantic representation is then generated and directly compared to the actual acoustic input. Depending on the amount of error thus generated, the prior is either accepted or rejected in favor of newly updated hypotheses. These hierarchical interactions are typically reflected

in the power of different local field potential (LFP) frequency bands: bottom-up processes have been associated with the low-gamma and theta bands, and top-down ones with the beta band (Arnal and Giraud, 2012; Fontolan et al, 2014; van Kerkoerle et al, 2014; Michalareas et al, 2016; Chao et al, 2018; Giraud and Arnal, 2018).

To identify the fine-grained mechanisms underlying speech neural transformations, it might be necessary to investigate speech processing at a smaller spatial scale giving access to the neuronal spiking level. Recent findings suggest that complex cognitive processes and behavioral features are encoded in low-dimensional neuronal spaces, also called *neural manifolds* (Pillai and Jirsa, 2017; Gallego et al, 2017; Jazayeri and Ostojic, 2021; Vyas et al, 2020; Chung and Abbott, 2021; Truccolo, 2016; Aghagolzadeh and Truccolo, 2016). In the primate cortex, the dynamics on the neural manifolds characterize functionally distinct behaviors and conditions, such as sensorimotor computations, decision making, or working memory (Mante et al, 2013; Remington et al, 2018; Markowitz et al, 2015). In humans, this endeavor is largely hindered by the invasiveness of single-neuron recordings, which are rarely performed. Although speech encoding has never been characterized at the level of collective dynamics of action potentials recorded from neuronal ensembles, it seems plausible that different aspects of speech, including transformations across the speech hierarchy, are encoded in such low-dimensional manifold representations. The condition- and history-dependent organization of neuronal trajectories on the manifold could serve the purpose of integrating phonetic features into higher-level representations (Pulvermüller, 2018; Yi et al, 2019). The handful of studies that have so far reported single unit activity associated with speech processing have highlighted the tuning of one or a few single units to phonetic features (Chan et al, 2014, 2011; Ossmy et al, 2015; Lakretz et al, 2021). By contrast, we focused here on the collective dynamics of tens of neurons in the light of the neural manifold framework to address the mechanisms underlying phonetic-to-semantic transformations.

We used unique recordings coming from a microelectrode array (MEA) implanted in the left anterior STG of a patient with pharmacologically resistant epilepsy, while he performed an auditory semantic categorization task and engaged in a spontaneous natural conversation (Chan et al, 2014). The MEA was placed at the intersection of areas traditionally associated with the processing of phonetic and semantic information, while an ECoG grid simultaneously recorded LFP activity along the temporal lobe. Despite the absence of detectable power effects on the most proximal ECoG channels, we observed a distributed encoding of phonemes at the local neuronal population scale. The dynamics of individual phoneme trajectories were organized according to their corresponding phonetic features only at specific time periods. Critically, the same phonetic organization generalized to natural speech, with different speakers and a variety of complex linguistic and predictive processes. During natural speech processing, population encoding of phonetic features occurred concurrently with the encoding of semantic features, both culminating at about 400 ms after word onset. This was reflected in the peaks in low-gamma and beta power, in agreement with the analysis-by-synthesis framework (Bever and Poeppel, 2010). These findings suggest that phonetic features interact with semantic representations by the simultaneous instantiation of predictive bottom-up and top-down mechanisms.

Results

Semantic and phonetic neuronal encodings in the aSTG

Microelectrode array and ECoG signals were recorded in a 31-year-old patient with pharmacologically resistant epilepsy who was implanted for clinical purposes (Fig. 1a). The 10x10 MEA was located in the left anterior superior temporal gyrus (aSTG) (square on Fig. 1a). 23 ECoG electrodes of interest covered a large portion of the left temporal cortical surface (circles on Fig. 1a). The participant performed first an

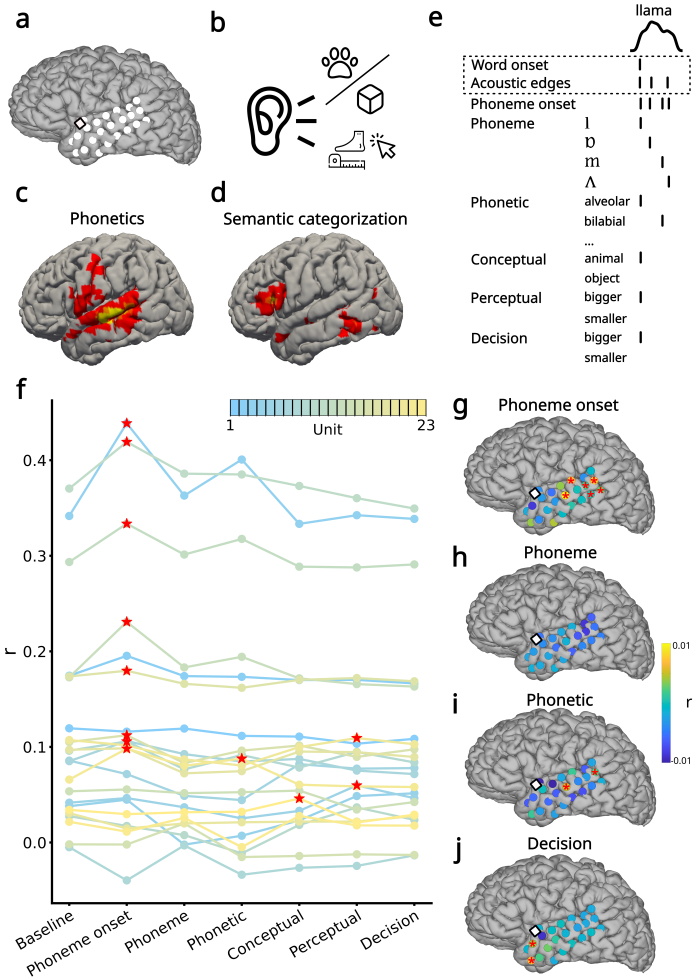

Fig. 1 | **Semantic and phonetic encoding at the single-unit level in the aSTG. a.** Locations of the microelectrode array (square) and the ECoG array (circles) on the cortical surface. **b.** Experimental design of the auditory semantic categorization task. The participant heard 400 nouns that denoted either animals or objects and was instructed to press a button if the item was bigger than a foot in size. **c.** fMRI correlates of phonetic processing obtained from fMRI large-scale database (Dockès et al, 2020). **d.** fMRI correlates of semantic categorization. **e.** mTRF features for the example English word "llama". The time series at the top indicates the speech envelope. The features used in the baseline model are indicated by a dotted square. **f.** Pearson correlation coefficient (r values) for each model fitted to each of the 23 single units with the highest firing rates in the ensemble. The lines for different units are color-coded based on firing rate, from lower (blue) to higher (yellow). Red stars indicate units for which r values of the fitted models are significantly higher ($p < 0.05$) than the chance level r values of a surrogate distribution. **g-j** Encoding of phoneme onsets, phonemes, phonetic categories, and semantic decision across ECoG channels. Colors indicate differences in r values compared to the baseline model. Red stars indicate significance as above.

auditory semantic categorization task, and then engaged in a spontaneous conversation. In the semantic task, the participant was instructed to indicate, by button press,

whether the heard word was smaller or bigger than a foot (Fig 1b). 400 unique nouns were presented, half of which indicated objects (e.g. chair), and the other half animals or body parts (e.g. donkey, eyebrow). In both groups (object, animal), words were equally divided in two categories, either bigger or smaller than a foot, resulting in a balanced 2-by-2 design.

We used multivariable temporal response function (mTRF) models to contrast the encoding of different linguistic processes and speech features (see Methods). To obtain an initial list of relevant linguistic processes for the mTRF analysis, we also used a large-scale fMRI database to specify the cognitive processes associated with the exact location in the aSTG where the intracortical MEA was positioned (Dockès et al, 2020). Nearby brain regions are indicated to process both phonetics (middle STG) (Fig. 1c) and semantic categorization (anterior superior temporal sulcus) (Fig. 1d).

Next, for each word, we identified the following speech features (Fig 1e): (i) acoustic, including word onset and acoustic edges (envelope rate peaks) (Oganian and Chang, 2019); (ii) phonemic, including phoneme onset and each phoneme identity; (iii) phonetic, including features based on vowel first and second formants, and consonant manner and place of articulation; (iv) semantic, including the word's conceptual category (object vs. animal), perceptual category (bigger or smaller than a foot), and semantic decision (participant's response about whether the object was bigger or smaller than a foot) (Supplementary tables 1, 2, and 3). The semantic decision feature was regressed separately from the perceptual category feature based on the fact that the participant only responded correctly in 80.25% of the trials. We constructed a baseline model based on acoustic features, and for each feature of interest, we compared the baseline model performance with the performance of a mTRF model containing the feature of interest together with the baseline features. The model performance was computed as Pearson's correlation between the model prediction and the neuronal activity in a 5-fold nested cross-validation procedure (Methods).

We fitted the mTRF encoding models to the soft-normalized firing rate of single units recorded with the MEA (Methods). We considered the 23 spike-sorted single units (over a total of 176) with an average firing rate higher than 0.3 spikes/second (mean: 1.6, sd: 1.39, range 0.43 - 5.27 spikes/second, Methods). Only a few units significantly responded to different features. Eight of them correlated significantly to phoneme onsets ($p < 0.05$ based on the chance level performance of a surrogate distribution, see Methods), one to phonetic features, one to conceptual categories, and two to perceptual categories (Fig. 1f).

By contrast, the broadband high-frequency activity (BHA, as a proxy for local firing rate activity (Crone et al, 2001)) of the ECoG electrode in the immediate vicinity of the MEA did not correlate significantly to any probed features. Six other ECoG electrodes significantly correlated to phoneme onset (Fig. 1g), two for phonetic features (Fig. 1i) in the posterior middle and superior temporal gyrus, and two for semantic decision in the aTL (Fig. 1j). No other models showed significant correlations (Supplementary Fig. 1).

Our next goal was to further explore the encoding of phonetic features. To this aim, we separated them into four different groups (vowel first formant, vowel second formant, consonant manner of articulation, and consonant place of articulation), and contrasted a model for each phonetic feature group with a base model including only phoneme onsets and acoustic features. Two units showed a significant spiking increase for consonant manner and vowel first formant models, and none for the other two models (Supplementary Fig. 2). None of the individual phonetic groups were distinguished by the ECoG electrode adjacent to the MEA (Supplementary Fig. 2).

Distributed phonetic encoding

We showed that only a few single units, when taken independently, encoded phonetic features. However, several other units showed non-significant changes (permutation

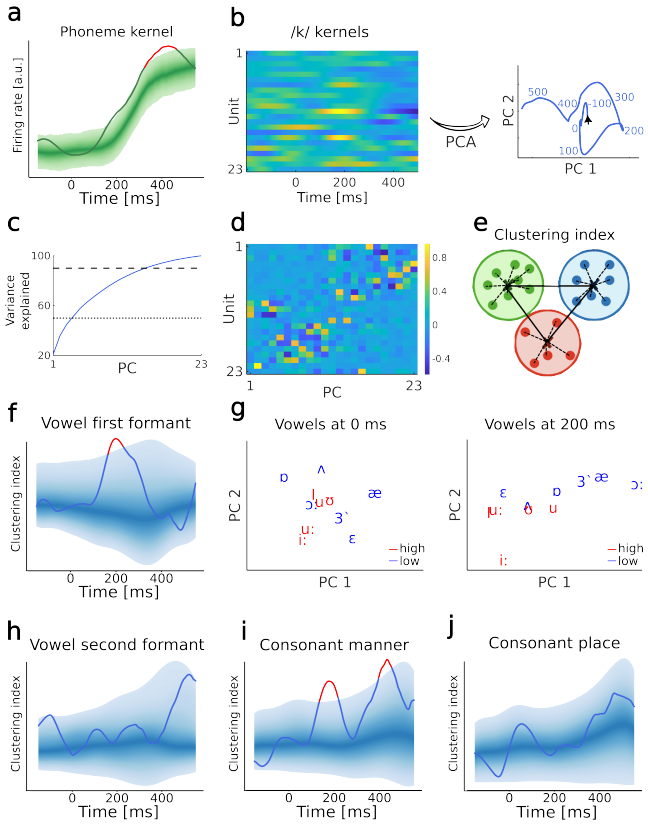

Fig. 2 | **Distributed encoding and clustering of phonemes along phonetic categories. a.** Average across mTRF phoneme kernels. The 95% confidence region of the surrogate (chance-level) distribution is shown with brighter shading for increasingly peripheral percentiles. Red segments indicate significant periods after multiple comparison correction (cluster-based test). The averaged phoneme kernel was significant at two time periods, centered around 200 and 400 ms, with the second period surviving multiple comparisons. **b.** mTRF kernels identified on each single unit for one example phoneme /k/. The units are sorted increasingly by their mean firing rate. Brighter colors indicate higher values of the kernel. These kernels are projected into a two-dimensional space using principal component analysis (PCA), resulting in a single phoneme trajectory. Numbers along the trajectory indicate the time points at which the trajectory reached the corresponding location. **c.** PCA variance explained. Four PCs explained 50% of variance (dotted line) and 15 explained 90% (dashed line). **d.** Principal component (PC) coefficients of isolated single units. Several units are represented in the first PCs, indicating distributed encoding of phonemes. **e.** Schematics showing how the clustering index is computed. Intra-cluster distance (full lines) was subtracted from between-cluster distances (bold lines). **f.** Clustering index for vowels grouped by the first formant (high and low tongue position). Shading and red segments are as in **a**. This clustering index was significant during a time interval centered around 200 ms. **g.** Distribution of vowels in the two-dimensional PC space at 0 and 200 ms. Vowels are color-coded based on their first formant. Mirroring the increase in the clustering index, the separation between those two phonetic features in the PC space was absent at 0 ms, and became evident at 200 ms. **h.** Clustering index for vowels grouped by the second formant did not reveal any significant periods. **i.** Clustering index for consonants grouped by the manner of articulation revealed two significant periods, centered around 200 and 400 ms. **j.** Clustering index for consonants grouped by the place of articulation did not reveal any significant periods.

test, Fig. 1f), which nevertheless might have contributed to the overall encoding through neuronal population dynamics. The time course of the phoneme kernels averaged across all units compared to the distribution of surrogate (chance-level) kernels showed two significant periods, centered around 200 and 400 ms post phoneme onset, suggesting a mean-field population effect (Fig. 2a). We thus hypothesized that a readout of phonetic encoding emerged at the level of the neuronal population dynamics, and that these dynamics could be captured by a low-dimensional neural manifold.

To address this hypothesis, we performed principal component analysis (PCA) on all concatenated phoneme kernels obtained with mTRF (Fig. 2b). The first four principal components (PCs) accounted for about 50% of phoneme feature variance (Fig 2c), and were distributed across different units (Fig 2d). Having half of the variance explained by the low-dimensional and correlated activity of several units indicates that phonemes are dynamically represented within a manifold.

We then asked whether the obtained low-dimensional neural manifold carried a functional relevance for phonetic encoding. For this, we analyzed the time course of phoneme kernels projected to the neural manifold (Fig 2d) and investigated whether the phonemes grouped according to phonetic categories. Thus, for each group of phonetic features (vowel first formant, vowel second formant, consonant manner, consonant place), we clustered the corresponding phoneme kernel trajectories at each time point in the low-dimensional space spanned by the first two PCs. Then, we computed a clustering index as the difference of between- and intra-cluster distances (Fig 2e) and compared it against the surrogate distribution (Methods).

We observed significant clustering within the vowel first formant group between the two phonetic features that reflect the high and low position of the tongue (Fig. 2f). The clustering specifically occurred at about 200 ms, which was apparent in the position of the vowels in the two-dimensional PC space (Fig. 2g). We replicated these findings using higher dimensional PC spaces (Supplementary Fig. 3). Contrary to

the vowel first formant group, we observed no significant clustering between phonetic features of the vowel second formant group (front and back position of the tongue, Fig. 2h, Supplementary Fig. 3 and 4). These findings indicate selective clustering of phoneme kernels based on vowel first formant at 200 ms.

We next investigated the organization of consonant trajectories in the low-dimensional space. The clustering index based on the manner of articulation (plosive, nasal, fricative, approximant, and lateral approximant features) indicated two significant periods centered around 200 ms and 400 ms (Fig. 2h). The same two peaks persisted irrespective of the number of PCs selected (Supplementary Fig. 3). There was no observable clustering along the consonant place of articulation (bilabial, labiodental, dental, alveolar, velar, uvular, and glottal features) in any low-dimensional space (Fig 2l, Supplementary Fig. 3).

We replicated the observed clustering by vowels first formant and consonant manner of articulation, as well as the lack of clustering by vowel second formant and consonant place of articulation with additional control analyses. Linear discriminant analysis (LDA) classifier revealed separability of these two phonetic categories in the same time windows (i.e. at about 200 ms for vowels first formant and both at about 200 and 400 ms for consonants manner of articulation) (Supplementary Fig. 5). Similarly, a rank-regression approach, where the ranks indicate the first or second formant value, showed significant ordering of the vowels along the first formant at 200 ms, and no ordering along the second formant (Supplementary Fig. 6). (This analysis cannot be performed for consonants, where no ranking is possible across the different phonetic groups). Finally, we observed similar organizational patterns through k-means clustering, an unsupervised data-driven approach (Supplementary Fig. 7).

To summarize, we observed a significant clustering of phoneme trajectories on a low-dimensional neural manifold: vowel trajectories clustered along first formants at 200 ms, while consonant trajectories clustered by manner of articulation at 200 and 400 ms.

Generalization to natural speech perception

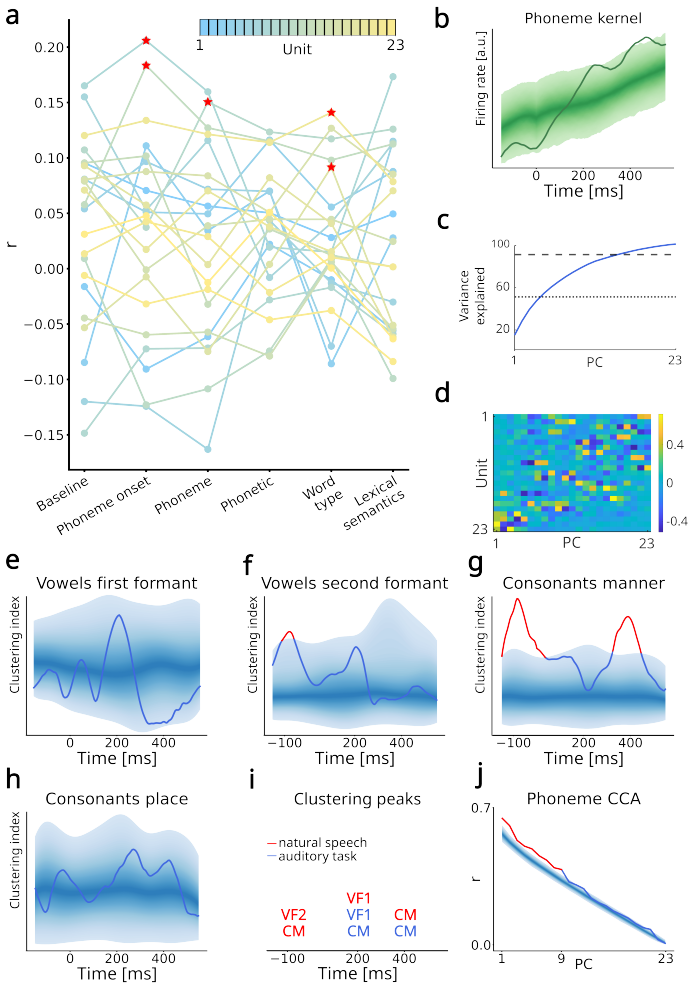

Fig. 3 | **Generalization to natural speech perception a.** r values for each model and the 23 most-spiking units. The units are color-coded based on their firing rate, from blue (lower) to yellow (higher). Red stars indicate model-unit pairs significantly higher than the baseline in a permutation test. **b.** Average across mTRF phoneme kernels. The 95% confidence region of the surrogate (chance-level) distribution is shown with brighter shading for increasingly peripheral percentiles. The averaged phoneme kernel was significant at two time periods, centered around 200 and 400 ms. **c.** PCA variance explained. Five PCs explained 50% of variance (dotted line) and 15 explained 90% (dashed line). **d.** Principal component (PC) coefficients of identified units. Several units are represented in the first PCs, indicating distributed encoding of phonemes also during natural speech perception. **e-h.** Clustering index for the four groups of phonetic features. Shading as in **b**. Red segments indicate significant periods after multiple comparison correction (cluster-based test). **i.** Overview of the significant clustering peaks observed both during the semantic task and natural speech perception. **j.** Canonical correlation analysis (CCA) between the PC projections of phoneme kernels obtained during the semantic task and natural speech perception. Shading and red segments as in **b**. CCA revealed significant correlations for the first nine PCs.

To investigate whether the results generalized to natural speech perception, we analyzed neuronal recordings from the same intracortical MEA during a spontaneous conversation between the participant and another person recorded in a separate experimental session. As for the semantic task, we performed spike sorting and selected the 23 units with the highest spiking rate (mean: 0.58 spikes/second, sd: 0.89, range 0.1 - 3.39) out of 212 clustered single units. We identified 664 words (272 unique) pronounced by the other person in the recording. From those words, we segmented the same 32 phonemes as in the task. We designed two high-level features, namely word class (e.g. noun, adjective, conjunction, etc), and lexical semantics, computed as the Lancaster sensorimotor norms of each word (Lynott et al, 2019) (see Supplementary Tables 4 and 5 for an overview). We then fitted the mTRF encoding models as in the semantic task dataset, and likewise observed that only a few units significantly correlated with speech features (Fig. 3a), and that the average phoneme kernel became significant compared to its surrogate distribution at about 200 and 400 ms, suggesting again a population effect (Fig. 3b).

We thus proceeded as before by performing PCA on the kernels. Similar to the task, 50% of the variance was explained by a few PCs (Fig. 3c), and the processing of phonemes was distributed across units (Fig. 3d). We further computed the clustering index for phonetic features in the low-dimensional neural manifold spanned by the first two PCs. Remarkably, the clustering index profiles were similar to those in the semantic task. For the vowel first formants, we observed a peak of the clustering index at about 200 ms (Fig. 3e). Although this peak was not significant for the two-dimensional space, it was present for all dimensions, and reached significance for the one-dimensional space (Supplementary Fig. 8). This peak was also significant in all control analyses (LDA classifier, rank regression, and k-means, Supplementary Fig. 9-11). For vowel second formants, a trend consistent across dimensions was observed at 200 ms (Fig. 3f). The second formant phonetic features also clustered before the

phoneme onset, specifically at -100 ms, possibly reflecting prediction mechanisms present in natural speech perception but not during the semantic task. For consonant manner of articulation, we observed significant clustering indices at about 400 ms, as well as an additional peak at about -100 ms that was possibly related to prediction mechanisms (Fig. 3g). The peak at 200 ms that was observed in the semantic task did not occur here. Finally, for consonant place of articulation, no significant peak of clustering index was observed (Fig. 3h). We performed the same control analyses as before (higher dimensional PC spaces, LDA classifier, rank regression, and k-means clustering), and replicated all the findings (Supplementary Fig. 9-11). The timeline on Fig. 3i summarizes the significant clustering peaks across both datasets.

Finally, we compared the similarity of low-dimensional phoneme trajectories obtained in the semantic task versus natural speech. To that aim, we performed canonical correlation analysis (CCA) between the projected phoneme kernels of the two datasets, and compared it against the surrogate distribution of canonical correlations obtained by shuffling the kernels for natural speech. We observed a significant canonical correlation between the individual phoneme kernels for the first nine PC dimensions (Fig 3j). This shows that although the phonemes in the semantic task and natural speech perception are encoded by different units (as both recordings are separated by a few hours), the neuronal population dynamics are encoded similarly, with individual phonemes tracing highly similar trajectories in the low-dimensional neural manifold.

Encoding of semantic features

Considering that the MEA is implanted in a cortical area involved in auditory semantic processing (Ralph et al, 2017), we repeated the same analyses for the semantic kernels, both for the semantic task and natural speech perception. We first considered the time course of the semantic kernels averaged across all units, and compared it to the surrogate distribution (Fig 4a). The perceptual category kernel showed a significant

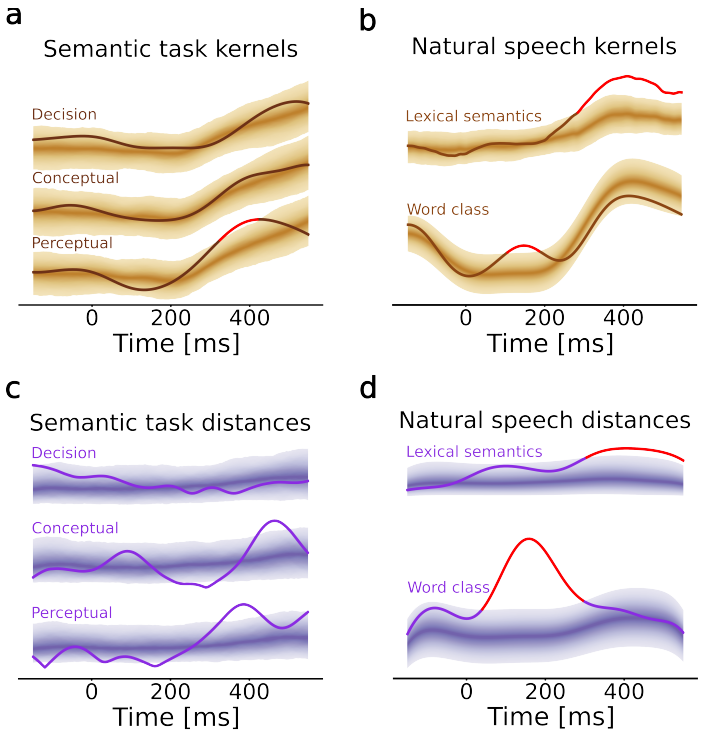

Fig. 4 | Encoding of semantic features. a. Semantic feature kernels of the semantic task. The 95% confidence region of the surrogate (chance-level) distribution is shown with brighter shading for increasingly peripheral percentiles. Red segments indicate significant periods after multiple comparison correction (cluster-based test). The perceptual category feature was significantly encoded during a time window centered at about 400 ms. **b.** Semantic kernels of natural speech perception. Shading and red segments as in **a**. Word class was processed at about 150 ms, and lexical semantics was processed in a broad window at about 400 ms. **c.** Euclidean distances in the PC space between the semantic task feature kernels. The perceptual semantic feature kernels were separated in the PC space slightly before the categorical semantic feature kernels (at about 400 and 450 ms, respectively). **d.** Euclidean distances in the PC space between the semantic features during natural speech perception were maximal at about 150 ms and 400 ms respectively. Shading and red segments as in **a**.

period at about 400 ms after word onset, while other kernels were not significant. We also repeated the same analysis for natural speech perception. The word class kernel (nouns, verbs, etc.) averaged across units showed a significant period at 150 ms (Fig 4b). The lexical semantic kernel averaged across units also showed a significant activation at 400 ms, consistent with our findings on the semantic task.

We then turned to the neuronal population dynamics. We projected the three semantic kernels of the semantic task to the low-dimensional neural manifold. Because there are only two categories for each of the probed semantic features, we could not

perform clustering analysis as for the phonemic features. Instead, at each time point, we computed the Euclidean distance between the projected trajectories and compared the resulting distance against its surrogate distribution (Fig 4c, Supplementary Fig. 12). We observed a significant separation of perceptual category kernels (bigger vs. smaller) at 400 ms. Interestingly, we also observed a significant separation for the conceptual category kernels (object vs. animal) at about 450 ms, even though the averaged conceptual category kernel was never significant, highlighting that the relevant functional read-out of certain aspects of semantic processing might be emerging only at the level of the neuronal population dynamics. Finally, the Euclidean distance for semantic decision was not significant. For natural speech perception, projecting those kernels to the low-dimensional neural manifold (Fig 4d, Supplementary Fig. 13), we found a strong separation between word classes at 150-200 ms, possibly reflecting syntactic processing, and a significant separation between lexical semantics kernels at about 400 ms.

Concurrent encoding of bottom-up phonetic and top-down semantic features

To further investigate the mechanisms underlying the integration of phonetic and semantic features, we explored the timing of their representations in more detail. Inspired by the analysis-by-synthesis framework, we hypothesized that an early bottom-up phonetic processing enables a later top-down word-level semantic guess, which is compared to actual phonetic features (Halle and Stevens, 1959; Bever and Poeppel, 2010). For the late semantic-phonetic comparison to occur, phonetic features should be (re)encoded concurrently with semantic processing at about 400 ms after word onset. To challenge this hypothesis, we created additional mTRF features that regressed phoneme onsets at different positions within each word (i.e., phoneme order). We then aligned these position-based phoneme onset kernels with word onset

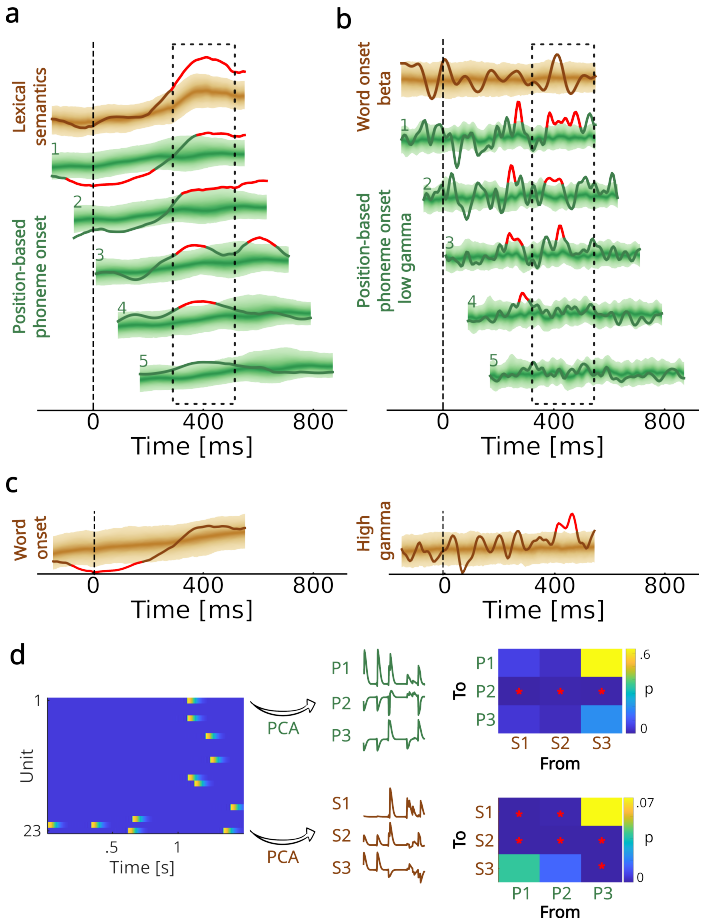

Fig. 5 | Concurrent encoding of bottom-up phonetic and top-down semantic features a. Kernels for position-based phoneme onsets (green) within a word shifted by average phoneme duration (80 ms) and aligned with the lexical semantic kernel at word onset (brown). The 95% confidence region of the surrogate (chance-level) distribution is shown with brighter shading for increasingly peripheral percentiles. Red segments indicate significant periods after multiple comparison correction (cluster-based test). Aligned phoneme onset kernels peaked simultaneously with the lexical semantic kernel peak (dashed rectangle). **b.** Beta band kernel for word onset is indicated in brown, and low-gamma kernels for position-based phoneme onsets are indicated in green and shifted by average phoneme duration. Shading and red segments as in **a**. The late peak in the beta word-onset kernel at about 400 ms occurred simultaneously with the aligned low-gamma phoneme onset peaks (dashed rectangle). **c.** Word onset kernels for firing rate (left) and broadband high-frequency activity (right). Despite the absence of the late effects on the firing rate kernel, significant broadband high-frequency activity was observed after 400 ms. **d.** Granger causality between phonetic and semantic low-dimensional representations. The left plot shows a 1.5-second snippet of the spiking data recorded during natural speech perception. These are projected onto three PCs to obtain time-varying phonetic (P) and semantic (S) features, which are then used in the Granger causality analysis. Matrices indicate p-values of Granger's F-test for a causal relationship from the dimensions indicated on the x-axis to the dimensions indicated on the y-axis. The stars indicate significant relationships at the 0.05 threshold. There was a significant causal relationship in both directions (i.e. phonetic to semantic and vice versa).

by shifting each phoneme onset for the corresponding multiple of the average phoneme duration (80 ms). If the hypothesis is true, a significant period of activation should align across the shifted kernels. While no significant peaks for the distinct phoneme onsets were observed in the semantic task (Supplementary Fig. 14), we found an alignment of significant peaks for phoneme onsets at positions 1-5 at about 400 ms during natural speech perception, simultaneously to semantic encoding (Fig. 5a). The analysis-by-synthesis framework might thus hold true for natural speech perception, where more semantic top-down processing is expected, and where there are no predictability biases related to the task design, such as fixed stimulus onset timing.

To further test the compatibility with the analysis-by-synthesis framework, we asked whether the encoding reflected top-down predictive or bottom-up cumulative processes during natural speech perception. The occurrence of top-down processes is typically accompanied by beta peaks in the LFP power, while bottom-up processes are reflected in LFP low-gamma power peaks (Arnal and Giraud, 2012; Fontolan et al, 2014). To identify the timing of top-down and bottom-up processes, we thus extracted and averaged the LFP power across the MEA contacts for both the beta (12-30 Hz) and low-gamma (30-70 Hz) bands. We then fitted an mTRF model to these two variables using word onset features for the beta band (reflecting word-level top-down processing) and position-based phoneme onsets for the low-gamma band (reflecting position-dependent bottom-up processing). For the beta band, we found two significant positive peaks: at word onset and 400 ms after word onset (Fig. 5b, brown trace). In the low-gamma band, we aligned position-based phoneme onset kernels based on word onset as before, and observed negative and positive significant peaks before 400 ms in the low-gamma band that were not accompanied by a beta peak, suggesting early bottom-up processes. We also found an alignment of positive significant peaks at 400 ms after word onset (Fig. 5b, green traces), showing that both top-down semantic and bottom-up phonetic processing occurred at 400 ms. Following the

analysis-by-synthesis framework, early low-gamma peaks might reflect early bottom-up-only processes that contribute to information build-up, possibly also observed in the early significant negative peaks of the first and second phoneme at the neuronal level (Fig. 5a). These bottom-up processes then lead to the 400-ms concurrent beta top-down and low-gamma bottom-up peaks, simultaneously to the significant positive neuronal peaks at 400 ms (compare Fig. 5a and b).

As an additional confirmation of the 400 ms top-down process, we extracted the broadband high-frequency activity (BHA, 70-150 Hz) across the MEA, as it has recently been shown to originate from both local neuronal firing and feedback predictive information (Leszczyński et al, 2020). We found a significant peak of the BHA kernel at about 400 ms (Fig. 5c). By contrast, the firing rate kernel aligned to word onset did not show any late significant peaks. This finding further confirms that top-down processes occurred at about 400 ms, as suggested by the corresponding beta peak (Fig. 5b).

Finally, we investigated whether the phonetic low-dimensional neuronal dynamics were time causally predictive of the semantic low-dimensional dynamics (reflecting a bottom-up process) or, inversely, whether the semantic trajectories predicted the phonetic ones (for a top-down process). For this, we projected the neuronal firing rate during natural speech to the first three dimensions of the phonetic and semantic PC spaces and investigated Granger causality between the projected time series. We found significant effects in both directions, indicating a bidirectional causal relationship between low-dimensional semantic and phonetic processing (Fig. 5d), which again supports the analysis-by-synthesis framework.

Together, these findings suggest that neuronal population dynamics in the aSTG encode phonetic and semantic features through concurrent bottom-up and top-down processes.

Discussion

Using both an auditory semantic categorization task and natural speech perception, we showed that the low-dimensional neuronal population dynamics recorded in the aSTG during speech processing concurrently encode phonetic and semantic features. We identified a neural manifold for both feature groups and observed a functional separation of their corresponding trajectories across time. Specifically, phoneme trajectories consistently clustered according to phonetic features and were highly correlated across task and natural speech conditions, suggesting invariant representations across different contexts. Similarly, semantic trajectories separated along their perceptual and conceptual features during the semantic task, and along lexical semantic and word class features during natural speech. During natural speech, semantic and phonetic encoding occurred in parallel at 400 ms after word onset with a bidirectional causal relationship between their low-dimensional representations. Phonetic encoding occurred significantly earlier for successive phoneme positions within words, such that processing of all phonemes culminated simultaneously at 400 ms after word onset. In agreement with the analysis-by-synthesis framework, this parallel encoding of phonemes and semantics was mirrored in concurrent bottom-up and top-down processes as measured by peaks in the low-gamma and beta power following early bottom-up-only processes.

A large amount of the neuronal variance related to phonetic and semantic encoding was accounted for by low-dimensional neural manifolds. Low-dimensional phoneme trajectories clustered according to phonetic (e.g. the first formant for vowels and the manner of articulation for consonants) and semantic features (e.g. perceptual and lexical semantics), denoting a large representational flexibility. These findings experimentally support the existence of neural manifolds for neural speech processing, extending recent findings in other cognitive domains, such as sensorimotor processing, decision making, or object recognition (Gallego et al, 2017; Mante et al, 2013; DiCarlo

and Cox, 2007). We observed that latent dynamics on the neural manifold constitute a robust functional read-out for speech processing. Specifically, the encoding of speech features became more prominent when considering the coordinated dynamics of the PCs, as opposed to the kernel activity simply averaged across units. This might account for the lack of detectable phonetic and semantic encoding at the same aSTG location in the ECoG signals, reflecting the average firing rate over a large population of neurons (Leszczyński et al, 2020). Further, we found that the low-dimensional phoneme features traced highly correlated trajectories across two conditions (semantic task and natural speech) up to the 9th PC (Fig. 3j). This is remarkable considering that the two conditions were separated by more than two hours and that the spike sorting procedure identified different units on the same array. This shows that despite the difference in the underlying units, common population patterns remain preserved, suggesting that the functional read-out emerges at the level of the neuronal population dynamics (Rutten et al, 2019; Vyas et al, 2020; Jazayeri and Ostojic, 2021; Chung and Abbott, 2021).

Encoding of phonetic and semantic features overlapped both in space and time. Both were encoded in the same focal cortical area of the aSTG, which contrasts with the modular partitioning of phonetic and semantic encoding traditionally observed throughout the temporal lobe (Hickok and Poeppel, 2007). Temporally, the processing of phonemes aligned to word onset concurred with the semantic processing window at about 400 ms, with a bidirectional causal relationship (Fig. 5). This advocates for the analysis-by-synthesis framework, where the distinct levels are related through dynamic bottom-up and top-down predictive loops (Halle and Stevens, 1959; Bever and Poeppel, 2010). Following this approach, we found significant alignments of top-down and bottom-up processes, as indicated by transient increases in beta and low-gamma power at about 400 ms post-word-onset (Fig. 5b). The beta peak might indicate a top-down semantic guess which is compared to a reinstantiation

of bottom-up phoneme representations highlighted by concurrent low-gamma peaks. This top-down semantic guess could result from a cumulative build-up of bottom-up phonetic processes reflected by low-gamma peaks occurring before the 400-ms alignment, in agreement with most models of spoken-word recognition (Marslen-Wilson, 1987; Norris and McQueen, 2008). The specific mechanisms underlying the emergence of such phonemic-to-semantic interface remain to be uncovered. For example, distinct phonetic features could be represented either sequentially (Hickok and Poeppel, 2007; Dehaene et al, 2015) or persistently (Perdikis et al, 2011; Yi et al, 2019; Martin, 2020) with more resolved temporal mechanisms (Fontolan et al, 2014; Hovsepyan et al, 2023) before being pooled together in a semantic representation.

The findings obtained in a controlled auditory semantic categorization task generalized to natural conversation. This is particularly remarkable, as the two datasets differ in important ways. First, in the controlled task, the participant heard isolated word recordings, all of which were nouns normalized for duration and sound intensity, while focusing on a simple cognitive task – assessing the size of the heard objects and animals. Natural speech, on the other hand, involves many other cognitive and perceptual mechanisms, including connected speech that amounts to sentences containing all word types. The intrinsic difference between datasets thus allowed us to generalize the semantic findings from a carefully designed semantic categorization task to a larger category of lexical semantics, and to word class encoding hence approaching syntactic processing. Second, even though natural speech was uttered by another speaker, we could retrieve invariant phonetic representations, both in the timing and the shape of the low-dimensional trajectories. Therefore the low-dimensional representations underpin speaker-normalization of phonetic representations, a notion that was suggested by LFP-level findings (Sjerps et al, 2019). Third, in the natural speech task, sounds were more variable in duration and intensity than in the controlled task, entailing other processing dimensions, e.g. prosody, accents, intonation, etc. Finally, natural speech

involves a whole range of predictive processes spanning the entire speech hierarchy from low-level acoustics to syntax and semantics. Together, these findings suggest that low-dimensional neuronal population encoding is invariant to speech context.

Using the natural speech dataset, we contrasted neuronal signatures underlying semantic and syntactic encoding. Syntactic and semantic encoding occurred at about 150 - 200 ms and 400 ms respectively in their corresponding neural manifolds. The early syntactic encoding might reflect a fast combinatorial process, as reported with MEG (Pylkkänen, 2020) and/or early effects contributing to conceptual and perceptual categorization in the ATL at about 200 ms (Borghesani et al, 2019; Chan et al, 2011; Chen et al, 2016; Dehaene, 1995; Hinojosa et al, 2001). The semantic categorization effect about 400 ms is reminiscent of the N400 component frequently reported in the ATL, for both semantic composition and lexical decision (Kutas and Federmeier, 2011; Dikker et al, 2020; López Zunini et al, 2020; Bentin et al, 1985; Barber et al, 2013; Vignali et al, 2023; Rahimi et al, 2022; Lau et al, 2008; Kutas and Federmeier, 2000).

These rare human MEA recordings provide a unique opportunity to investigate the neuronal population effects, at a spatial resolution that was never approached before for speech, in particular dynamical effects that were not detectable even at the most adjacent ECoG contact. However, the conclusions must be taken with caution, as should any data coming from a single participant. Further, high-resolution data come with the price of a restricted spatial sampling. Our findings are specific to a small 4-by-4 mm area of aSTG encompassed by the implantation site of the MEA. This might explain why we observed very specific phonetic effects - e.g. vowels clustering according to the first and not the second format. It is thus possible that the aSTG is organized into small functional subdivisions and that we would observe other clustering principles based on other phonetic features if the array was implanted in nearby regions. Finally, the population of recorded neurons is also limited by the design of the MEA. Only a

few tens of units over around two hundred identified responded to the probed speech features, suggesting that coordinated neuronal population encoding is relatively sparse.

To conclude, our study provides evidence for a parallel, distributed, and low-dimensional encoding of phonetic and semantic features, that is specific to neuronal population firing patterns in a focal region in the aSTG. These local dynamic population effects are part of bottom-up and top-down dynamics involving oscillatory bidirectional activity potentially involving higher- and lower-tier regions of the language hierarchy. Extending the rapidly emerging neural manifold framework to speech processing, these findings shed new light on the brain mechanisms underlying phonetic and semantic integration and pave the way toward the elucidation of the intricacies behind the complex transformations across the speech processing hierarchy.

Methods

Participant

The study participant, a male in his thirties with pharmacologically resistant epilepsy, underwent intracranial electrode implantation as part of his clinical epilepsy treatment. He was a native English speaker with normal sensory and cognitive functions and demonstrated left-hemisphere language dominance through a WADA test. The patient experienced partial complex seizures originating from mesial temporal region electrode contacts. The surgical intervention involved the removal of the left anterior temporal lobe, along with the microelectrode implantation site, left parahippocampal gyrus, left hippocampus, and left amygdala. The patient achieved seizure-free status one year post-surgery, with no significant changes in language functions observed in formal neuropsychological testing conducted at that time. Informed consent was obtained, and the study was conducted under the oversight of the Massachusetts General Hospital Institutional Review Board (IRB). The study included both the intracortical implantation of a microelectrode array (MEA) and the performance of a semantic task and

natural speech. The MEA recordings were used only for scientific research purposes and played no role in the clinical assessments and decisions.

Neural recordings

Cortical local field potentials were recorded with an 8-by-8 subdural ECoG array (Adtech Medical), 1-cm electrode distance, implanted above the left lateral cortex, including frontal, temporal, and anterior parietal areas. For the purpose of this article, only the electrodes covering the lateral temporal lobe were included in the analysis. The signal was recorded with a sampling rate of 500 Hz, with a bandpass filter spanning 0.1 to 200 Hz. All electrode positions were accurately localized relative to the participant's reconstructed cortical surface (Dykstra et al, 2012).

Single-unit action potentials were recorded with a 10-by-10, 400 µm electrode distance, microelectrode array (Utah array, Blackrock Neurotech) surgically implanted within the left anterior superior temporal gyrus (aSTG). Electrodes were 1.5 mm long and contained a 20-µm platinum tip. The implantation site was excised and the subsequent histological analysis revealed the spatial orientation of the electrode tips within the depths of cortical layer III, proximal to layer IV, with no notable histological abnormalities in the neighboring cortical environment. Data acquisition was acquired with a Blackrock NeuroPort system, with a sampling frequency of 30 kilosamples per second, and an analog bandpass filter ranging from 0.3 Hz to 7.5 kHz for antialiasing.

The location of the recording arrays was based on clinical considerations. In particular, the MEA was placed in the superior temporal gyrus because this was a region within a larger area anticipated to be resected based on prior imaging data.

Auditory stimuli

Neural data was recorded during two separate experimental sessions, that took place on the same day, two hours apart (Chan et al, 2014). In the first session, the participant performed an auditory semantic categorization task. The stimuli were standalone

audio files of 400 words pronounced by a male speaker and normalized for intensity and duration (500 ms). The participant was presented with 800 nouns in a randomized order, and with 2.2 s stimulus onset asynchrony. Out of 800 words, 400 were presented only once, while the remaining 400 consisted of 10 words repeated 40 times each. To avoid biasing effects, in our analysis we considered only the 400 words that were repeated only once. Specifically, the inclusion of repeated words leads to the overrepresentation of a few phonemes compared to other phonemes, biasing the regression analyses. Half of the 400 unique words referred to objects and half to animals. Following a word presentation, the participant was instructed to press a button if the referred item was bigger than a foot in size. Half of the items in each group (animals, objects) were bigger than a foot, resulting in a balanced 2-by-2 design.

In the second session, the participant engaged in a conversation with another person present in the room. The natural speech was recorded using a far-field microphone and manually transcribed. The recording was split into 91 segments that contained clear speech recordings of the other person talking (i.e. without overlapping speech or other background sounds). Each segment was cleaned for background noise and amplified to 0 dBFS in Audacity software. We used a total of 664 words (272 unique) across all trials of the natural speech (Supplementary Table 4).

Signal preprocessing

Spike detection and sorting were performed with the semi-automatic wave_clus algorithm (Quiroga et al, 2004). Across 96 active electrodes, we identified 176 and 212 distinct units for the sessions with the semantic task and the natural speech, respectively. For the semantic task, we considered units with firing rate higher than 0.3 spikes/second, resulting in 23 units (0.43 – 5.27 spikes/s). For a fair comparison across both sessions, we also selected the 23 most spiking units in the natural speech (0.1 – 3.39 spikes/s).

The spike train of each unit was smoothed with a 25 ms wide Gaussian kernel to obtain the firing rate time series. Firing rate time series were then soft-normalized by the range of the unit increased with a constant 5, following previous studies (Churchland et al, 2012), and downsampled at 200 Hz. For the semantic task, firing rate time series were split into 400 trials. Each trial lasted 1.5 seconds and included 0.5-second periods before and after the word presentation. For the natural speech, we selected 91 segments of the firing rate where a person was talking to the participant (see above).

The signals from the ECoG grid were first filtered to remove line noise using a notch filter at 60 Hz and harmonics (120, 180, and 240 Hz). We then applied common-average referencing. For each channel, we extracted broadband high-frequency activity (BHA) in the 70-150 Hz range (Crone et al, 2001). BHA was computed as the average z-scored amplitude of eight band-pass Gaussian filters with center frequencies and bandwidth increasing logarithmically and semi-logarithmically respectively. The resulting BHA was downsampled to 100 Hz.

Phoneme segmentation and categorization

Audio files and corresponding transcripts were segmented both into words and phonemes by creating PRAAT TextGrid files (Boersma, 2001) through WebMAUS software (Kisler et al, 2017). Phonetic symbols in the resulting files were encoded in X-SAMPA, a phonetic alphabet designed to cover the entire range of characters in the 1993 version of the International Phonetic Alphabet (IPA) in a computer-readable format. All TextGrids were manually inspected and converted into tabular formats using the TEICONVERT tool. Diphtongues and phonemes that occurred less than 5 times throughout the entire session were removed from the analysis.

We used 32 segmented phonemes, divided into 11 vowels and 21 consonants, and further labeled according to the standard IPA phonetic categorizations (Supplementary Tables 1 and 2): vowels first formant (high, low), vowels second formant (front,

back), consonants articulation place (bilabial, labiodental, alveolar, velar, uvular, glottal), consonants articulation manner (plosive, nasal, fricative, approximant, lateral approximant).

mTRF features

For both sessions, we extracted the following features: word onset, acoustic edge (envelope rate peaks), phoneme onset, phoneme identity, and phonetic category. For the semantic task, we additionally computed the following semantic features: perceptual category, conceptual category, and semantic decision. For the natural speech, we additionally created word class and lexical semantics features. All features were designed as values located at the onset of the corresponding stimuli, all other values being set to zeros.

Word onset was marked by a value of one located at the onset of each word. Acoustic edges were defined as local maxima in the derivative of the speech envelope (Oganian and Chang, 2019). The speech envelope was computed as the logarithmically scaled root mean square of the audio signal using the MATLAB mTRFenvelope function. Phoneme onset feature indicates onsets of all phonemes in a word. Phoneme identity feature was multivariable with 32 regressors, each indicating onsets of a different phoneme, as defined by the IPA table. Phonetic category feature was multivariable, and included four phonetic groups (vowel first and second formant, consonant manner, and place of articulations). All other features were multivariables with a value located at the corresponding word onset. Perceptual category, conceptual category, and semantic decision had two regressors, defined respectively as bigger and smaller, animal and object, and decision on whether the object/animal was bigger or smaller than a foot. Word class had 13 regressors, indicating different word classes (noun, verb, adjective, adverb, article, auxiliary, demonstrative, quantifier, preposition, pronoun, conjunction, interjection, number). Finally, the lexical semantics feature was

multivariable, designed by regressing each of the 11 sensorimotor norms at the corresponding word onset (Lynott et al, 2019). All stimuli were smoothed with a 25-ms-wide Gaussian kernel and downsampled to either 200 Hz (to match single unit firing rates) or 100 Hz (to match BHA from ECoG channels) before fitting the mTRF models.

mTRF estimation

mTRFs were estimated using the mTRF MATLAB toolbox (Crosse et al, 2016). All mTRFs were always of encoding type, relating the stimulus features to neural data, with resulting kernels in the time range between -200 and 600 ms. The first and last 50 ms were not considered in the analysis, to avoid possible edge effects. Both for semantic task and natural speech, the baseline model contained the word onset and the acoustic onset edge features. All other models included the baseline features and one of the additional target features defined above. Estimation was performed by a ridge regression, using a nested cross-validation procedure (see below). The goodness of fit was defined as Pearson's correlation between model prediction and neural data.

Cross-validation

For the semantic task, we performed a nested cross-validation. In the outer cross-validation loop, we split the 400 words randomly into 5 sets of 80 (20% of the words each). Thus, in each fold, 80 trials belonged to a hold-out test set, while the remaining 320 words belonged to the train set. The five folds were identical across all models. For a given outer fold, among the 320 train-set trials, we performed another 8-fold inner cross-validation loop for the ridge regression hyperparameter tuning, with λ ranging from 10^{-6} to 10^{6}. The optimal lambda was then used to retrain the model on the 320 words of the training set of the outer cross-validation loop fold. The model predictions and Pearson's correlation with the neural data were then computed for the 80 words of the test set. Across the 5 folds, we thus obtained five correlation values, of which we report the average.

For natural speech, we also used nested cross-validation. The 5-fold outer cross-validation loop is performed by splitting the 91 segments into five folds of approximately similar duration (mean: 393.28 s; sd: 6.73 s), chosen through random shuffling across the five folds until the standard deviation of the duration distribution across folds was smaller than 10 seconds. We applied a similar procedure for the inner cross-validation loop in each of the five folds.

Surrogate distributions and statistical significance

For each model of the semantic task, we created a distribution of 1000 surrogate models by shuffling the target feature across words, and keeping the baseline features constant. For instance, in the model that contained the phoneme onset feature together with the baseline features (word onset and acoustic edges), the surrogate model was created by randomly shuffling the 400 phoneme onset features across 400 words independently, while keeping the order of the baseline features constant. In this way, the baseline features were properly regressed to the neural data, while the target feature (e.g. phoneme onset) was randomly assigned to neural data.

For the natural speech condition, it was not possible to shuffle trials in the same way, as each auditory segment was of a different duration. This posed a problem because the mTRF features have to be of the same length as the neural data, which was not the case for natural speech, contrary to the semantic task where each word had the same duration. Instead, we used the following method: for multivariable features, the surrogates were computed by randomly assigning each non-zero value to a particular regressor (e.g. for the phoneme identity feature, the first phoneme is randomly assigned to any of the 32 regressors, the second to any of the remaining 31, etc). For features with a single variable, surrogates were computed by performing a circular shift with a random onset. For instance, for the phoneme onset feature, which had only one regressor, we randomly split the trace into 2 parts and switched the order of the parts.

A model was considered statistically significant if the original model performed better than the 95th percentile of the surrogate distribution (one-tailed test).

Averaged kernels

To obtain averaged kernels, we took the mean of kernels across all units (and phonemes for the phoneme kernels). An averaged kernel was considered statistically significant if higher or lower than the 97.5th or 2.5th percentile respectively of the surrogate distribution (two-tailed test).

Correcting for multiple comparisons

To control for multiple comparisons, we used a nonparametric approach based on the cluster mass test (Maris and Oostenveld, 2007). From each permutation, we subtracted the mean value of the surrogate distribution and clustered consecutive time points that were outside the 95% confidence region of the surrogate distribution. Cluster-level statistics was defined as the sum of the mean-centered values within a cluster. For negative clusters, we considered the absolute value of the sum. Clusters of the observed data that were higher than the 95th percentile of the distribution of maximum cluster surrogates were considered as passing the multiple comparison correction.

Clustering of phoneme kernels in the PC space

Clustering was performed by assigning each phoneme to a particular phonetic class and computing the clustering index, defined as the difference between between-cluster and intra-cluster distances. Specifically, for each cluster, we first found the location of the centroid by averaging the coordinates of all cluster elements. Between-cluster distance was defined as the average Euclidean distance between all pairs of cluster centroids. Intra-cluster distance was defined as the Euclidean distance of each cluster element to the corresponding centroid. By subtracting intra- from between-cluster distance, our

index rewarded cluster separability (between-cluster distance) and penalized spatial dispersion (intra-cluster distance).

We performed clustering in the two-dimensional PC space (Results), but systematically confirmed our results in the PC spaces using up to six dimensions (Supplementary Material). Clustering was considered statistically significant if the original model had a higher clustering index than the 95th percentile of the surrogate distribution (one-tailed test).

To confirm our clustering results, we performed several control analyses, described here shortly and in detail in Supplementary Material:

1. linear discriminant analysis (LDA) classifier: at each time point and for each phonetic feature group, we first ran an LDA classifier to compute the means of the multivariate normal distributions of phonemes sharing the same phonetic feature. Then, we computed the average Euclidean distance between all phonetic feature means and compared this distance against the distribution of 1000 surrogates.
2. rank regression for vowels: we additionally explored whether the actual first and second formant frequency values were encoded in the low-dimensional space. To that aim, we assigned a rank value (1-7) to each vowel, based on the formant values indicated in the standard IPA table. At each time point, the ranked order of vowels was correlated with their coordinates on the first three PCs, and compared against a distribution of 1000 surrogates.
3. correlation with K-means connectivity matrices: to investigate whether the same results would emerge in a data-driven fashion, for each time point, we ran K-means clustering 1000 times (clustering results slightly differed depending on the algorithm's random initialization). Then, we computed an average N-by-N connectivity matrix that indicated how often each of the N phonemes was clustered together. Finally, we correlated the resulting connectivity matrix with the connectivity

matrix of the actual, phonetic-based clusters (vowel first formant, vowel second formant, consonants manner, consonants place), and compared the correlation value against the distribution of 1000 surrogates.

Separability of semantic feature kernels in the PC space

Each semantic feature of the semantic task only had two regressors (e.g. animal vs object, bigger vs smaller), hence not allowing any clustering. To identify the periods during which the two kernels of each semantic feature were significantly separated in the PC space, we computed the average Euclidean distance between the two. The same process was repeated for each of the 1000 surrogate models, and the distance was considered significant when higher than the 95th percentile of the surrogate distribution (one-tailed test).

Semantic features of the natural speech were computed similarly. For the Word class feature, we averaged the Euclidean distances between all pairs of regressors (noun, verb, adjective, etc). For the Lexical semantic feature, we computed the distance between the visual Lancaster sensorimotor norm and the average of the 10 other norms (Lynott et al, 2019), as visual norm was the most present in our dataset (visual norm strength: 2.73; other norms, mean: 1.22, standard deviation: 0.6).

Comparison of trajectories between the semantic task and natural speech

To investigate whether the trajectories of the phoneme kernels projected to the low-dimensional space were similar between the semantic task and the natural speech, we computed the canonical correlation between each kernel pair (e.g. kernel of phoneme /k/ extracted during the semantic task and the /k/ kernel from the natural speech). The canonical correlation was compared against the distribution of surrogate model canonical correlations for all 23 dimensions. Particularly, the surrogate distribution

was computed by shuffling the kernels in the natural speech, which constitutes a strong surrogate test. The correlation was considered significant if it was higher than the value of the 95th percentile of the surrogate correlation distribution (one-tailed test).

Position-based phoneme onset kernels

To probe for interactions between the encoding of phonetic and semantic features, we first created mTRF kernels for phoneme onsets at each phoneme position within the word and then aligned the resulting kernels with the corresponding semantics kernel (lexical semantics for natural speech and perceptual semantics for the task). Position-based phoneme onset was thus a multivariable feature with five regressors, each indicating the onset of the corresponding phoneme position across all words. For instance, phoneme onset at position 2 indicates the onset of second phonemes in all words, regardless of the phoneme type. The resulting kernels were then shifted by the multiples of 80 ms, which is a rounded value of average phoneme duration (mean: 82.6 ms, sd: 3.1 ms). Thus, the kernel for position two was shifted by 80 ms, the kernel for position three by 160 ms, etc. Surrogate kernel distributions were computed as described above. Phoneme onset kernels are considered significant if they are higher or lower than the value of the 97.5th or 2.5th percentile respectively of their surrogate kernel distributions (two-tailed test). This allowed us to observe when significant peaks for each phoneme occurred in reference to the same time point: word onset.

LFP beta, low-gamma, and BHA power analysis

We further investigated the nature of these kernels within the analysis-by-synthesis framework, by fitting mTRF encoding-type models with either word onset or position-based phoneme onset as stimulus, and either beta (12 - 30 Hz), low-gamma (30 - 70 Hz), or BHA (70 - 150 Hz) LFP power as response. Word onset and phoneme onset stimuli were the same as the ones used before. For each microelectrode channel, LFP power bands were computed by first applying a 9-order bandpass Butterworth filter

with zero-phase forward and reverse digital filtering, then subtracting the mean from the resulting trace, and finally computing the absolute value of its Hilbert transform. The resulting LFP powers were then averaged across all channels and entered into the mTRF models. Finally, we compared the significant periods of the word-onset beta-power kernel and position-based phoneme-onset low-gamma kernels with the same aligning procedure as above.

Granger causality

We adopted a Granger causality measure (Pesaran et al, 2018) to explore the temporal causality between the low-dimensional phonetic and semantic representations. We used the multivariate Granger causality toolbox (MVGC v1.3, 2022), which is based on a state-space formulation of Granger causal analysis (Barnett and Seth, 2014, 2015). We first applied a half-Gaussian filter 25 ms wide to the spiking traces of individual units to obtain a causal firing rate estimate. The resulting firing rates were then projected into the three-dimensional phonetic and semantic PC spaces, constructed by performing PCA on the corresponding phonetic and semantic mTRF kernels as described above. The Granger models were created by separately predicting each of the three PCs of one feature (e.g. phonetic) from all three dimensions of another feature (e.g. semantic). We then combined the resulting Granger coefficients into two 3-by-3 matrices, one for phonetic-to-semantic and another one for semantic-to-phonetic predictions. Autoregression model parameters were estimated from data via the Levinson-Wiggins-Robinson algorithm. The order of AR models was selected via the Akaike Information Criterion (AIC). Statistical significance of estimated Granger causality measures was assessed via the corresponding F-test as implemented by the toolbox.

fMRI databases

Functional classification of the cortical surface surrounding the MEA with respect to different linguistic processes was performed using NeuroQuery database (Dockès et al,

2020). It uses text mining and meta-analysis techniques to automatically produce large-scale mappings between fMRI brain activity and a cognitive process of interest. We were primarily interested in observing the proximity of phonetic and semantic processing close to the MEA implantation site. As keywords, we used "phonetics" and "semantic categorization".

Acknowledgments. This work was funded by Swiss National Science Foundation career grant 193542 (T.P.), EU FET-BrainCom project (A.G.), NCCR Evolving Language, Swiss National Science Foundation Agreement #51NF40_180888 (A.G.), Swiss National Science Foundation project grant 163040 (A.G.), National Institutes of Health (NIH), National Institute of Neurological Disorders and Stroke (NINDS), grants R01NS079533 (WT) and R01NS062092 (SSC), and the Pablo J. Salame Goldman Sachs endowed Associate Professorship of Computational Neuroscience at Brown University (WT).

References

Aghagolzadeh M, Truccolo W (2016) Inference and Decoding of Motor Cortex Low-Dimensional Dynamics via Latent State-Space Models. IEEE Transactions on Neural Systems and Rehabilitation Engineering 24(2):272–282. https://doi.org/10.1109/TNSRE.2015.2470527

Arnal LH, Giraud AL (2012) Cortical oscillations and sensory predictions. Trends in Cognitive Sciences 16(7):390–398. https://doi.org/10.1016/j.tics.2012.05.003

Barber HA, Otten LJ, Kousta ST, et al (2013) Concreteness in word processing: ERP and behavioral effects in a lexical decision task. Brain and Language 125(1):47–53. https://doi.org/10.1016/j.bandl.2013.01.005

Barnett L, Seth AK (2014) The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. Journal of Neuroscience Methods 223:50–68. https://doi.org/10.1016/j.jneumeth.2013.10.018

Barnett L, Seth AK (2015) Granger causality for state-space models. Physical Review E 91(4):040101. https://doi.org/10.1103/PhysRevE.91.040101

Bentin S, McCARTHY GREGORY, Wood CC (1985) Event-related potentials, lexical decision and semantic priming. Electroencephalography and clinical Neurophysiology 60:343–355

Bever TG, Poeppel D (2010) Analysis by Synthesis: A (Re-)Emerging Program of Research for Language and Vision. Biolinguistics 4(2-3):174–200. https://doi.org/10.5964/bioling.8783

Boersma P (2001) Praat, a system for doing phonetics by computer. Glot Int 5(9):341–345

Borghesani V, Buiatti M, Eger E, et al (2019) Conceptual and Perceptual Dimensions of Word Meaning Are Recovered Rapidly and in Parallel during Reading. Journal of Cognitive Neuroscience 31(1):95–108. https://doi.org/10.1162/jocn_a_01328

Caucheteux C, Gramfort A, King JR (2022) Deep language algorithms predict semantic comprehension from brain activity. Scientific Reports 12(1):16327. https://doi.org/10.1038/s41598-022-20460-9

Chan AM, Baker JM, Eskandar E, et al (2011) First-Pass Selectivity for Semantic Categories in Human Anteroventral Temporal Lobe. Journal of Neuroscience 31(49):18119–18129. https://doi.org/10.1523/JNEUROSCI.3122-11.2011

Chan AM, Dykstra AR, Jayaram V, et al (2014) Speech-Specific Tuning of Neurons in Human Superior Temporal Gyrus. Cerebral Cortex 24(10):2679–2693. https://doi.org/10.1093/cercor/bht127

Chang EF, Raygor KP, Berger MS (2015) Contemporary model of language organization: An overview for neurosurgeons. Journal of Neurosurgery 122(2):250–261. https://doi.org/10.3171/2014.10.JNS132647

Chao ZC, Takaura K, Wang L, et al (2018) Large-Scale Cortical Networks for Hierarchical Prediction and Prediction Error in the Primate Brain. Neuron 100(5):1252–1266.e3. https://doi.org/10.1016/j.neuron.2018.10.004

Chen Y, Shimotake A, Matsumoto R, et al (2016) The 'when' and 'where' of semantic coding in the anterior temporal lobe: Temporal representational similarity analysis of electrocorticogram data. Cortex 79:1–13. https://doi.org/10.1016/j.cortex.2016.02.015

Chung S, Abbott L (2021) Neural population geometry: An approach for understanding biological and artificial neural networks. Current Opinion in Neurobiology

70:137–144. https://doi.org/10.1016/j.conb.2021.10.010

Churchland MM, Cunningham JP, Kaufman MT, et al (2012) Neural population dynamics during reaching. Nature 487(7405):51–56. https://doi.org/10.1038/nature11129

Cope TE, Shtyrov Y, MacGregor LJ, et al (2020) Anterior temporal lobe is necessary for efficient lateralised processing of spoken word identity. Cortex 126:107–118. https://doi.org/10.1016/j.cortex.2019.12.025

Crone NE, Boatman D, Gordon B, et al (2001) Induced electrocorticographic gamma activity during auditory perception. Clinical Neurophysiology 112(4):565–582. https://doi.org/10.1016/S1388-2457(00)00545-9

Crosse MJ, Di Liberto GM, Bednar A, et al (2016) The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. Frontiers in human neuroscience 10:604

Damera SR, Chang L, Nikolov PP, et al (2023) Evidence for a Spoken Word Lexicon in the Auditory Ventral Stream. Neurobiology of Language 4(3):420–434. https://doi.org/10.1162/nol_a_00108

Dehaene S (1995) Electrophysiological evidence for category-specific word processing in the normal human brain. NeuroReport 6(16):2153

Dehaene S, Meyniel F, Wacongne C, et al (2015) The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees. Neuron 88(1):2–19. https://doi.org/10.1016/j.neuron.2015.09.019

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

Dikker S, Assaneo MF, Gwilliams L, et al (2020) Magnetoencephalography and Language. Neuroimaging Clinics of North America 30(2):229–238. https://doi.org/10.1016/j.nic.2020.01.004

Dockès J, Poldrack RA, Primet R, et al (2020) NeuroQuery, comprehensive meta-analysis of human brain mapping. eLife 9:e53385. https://doi.org/10.7554/eLife.53385

Dykstra AR, Chan AM, Quinn BT, et al (2012) Individualized localization and cortical surface-based registration of intracranial electrodes. NeuroImage 59(4):3563–3570. https://doi.org/10.1016/j.neuroimage.2011.11.046

Fontolan L, Morillon B, Liegeois-Chauvel C, et al (2014) The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. Nature Communications 5:4694. https://doi.org/10.1038/ncomms5694

Friederici AD, Kotz SA (2003) The brain basis of syntactic processes: Functional imaging and lesion studies. NeuroImage 20:S8–S17. https://doi.org/10.1016/j.neuroimage.2003.09.003

Gallego JA, Perich MG, Miller LE, et al (2017) Neural Manifolds for the Control of Movement. Neuron 94(5):978–984. https://doi.org/10.1016/j.neuron.2017.05.025

Giraud AL, Arnal LH (2018) Hierarchical Predictive Information Is Channeled by Asymmetric Oscillatory Activity. Neuron 100(5):1022–1024. https://doi.org/10.1016/j.neuron.2018.11.020

Gwilliams L, King JR (2020) Recurrent processes support a cascade of hierarchical decisions. eLife 9:1–20. https://doi.org/10.7554/eLife.56603

Halle M, Stevens K (1959) Analysis by synthesis. In: Proceeding of the Seminar on Speech Compression and Processing, Paper D7, vol II. W. Wathen-Dunn & L. E. Woods (eds.)

Hamilton LS, Oganian Y, Hall J, et al (2021) Parallel and distributed encoding of speech across human auditory cortex. Cell 184(18):4626–4639.e13. https://doi.org/10.1016/j.cell.2021.07.019

Heilbron M, Armeni K, Schoffelen JM, et al (2022) A hierarchy of linguistic predictions during natural language comprehension. Proceedings of the National Academy of Sciences 119(32):e2201968119. https://doi.org/10.1073/pnas.2201968119

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nature Reviews Neuroscience 8(5):393–402. https://doi.org/10.1038/nrn2113

Hinojosa JA, Martín-Loeches M, Muñoz F, et al (2001) Electrophysiological evidence of a semantic system commonly accessed by animals and tools categories. Cognitive Brain Research 12(2):321–328. https://doi.org/10.1016/S0926-6410(01)00039-8

Hovsepyan S, Olasagasti I, Giraud AL (2023) Rhythmic modulation of prediction errors: A top-down gating role for the beta-range in speech processing. PLOS Computational Biology 19(11):e1011595. https://doi.org/10.1371/journal.pcbi.1011595

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, et al (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

Kisler T, Reichel U, Schiel F (2017) Multilingual processing of speech via web services. Computer Speech & Language 45:326–347. https://doi.org/10.1016/j.csl.2017.01.005

Kutas M, Federmeier KD (2000) Electrophysiology reveals semantic memory use in language comprehension. Trends in Cognitive Sciences 4(12):463–470. https://doi.org/10.1016/S1364-6613(00)01560-6

Kutas M, Federmeier KD (2011) Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). Annual Review of Psychology 62(1):621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Lakretz Y, Ossmy O, Friedmann N, et al (2021) Single-cell activity in human STG during perception of phonemes is organized according to manner of articulation. NeuroImage 226:117499. https://doi.org/10.1016/j.neuroimage.2020.117499

Lambon Ralph MA, Lowe C, Rogers TT (2006) Neural basis of category-specific semantic deficits for living things: Evidence from semantic dementia, HSVE and a neural network model. Brain 130(4):1127–1137. https://doi.org/10.1093/brain/awm025

Lau EF, Phillips C, Poeppel D (2008) A cortical network for semantics: (de)constructing the N400. Nature Reviews Neuroscience 9(12):920–933. https://doi.org/10.1038/nrn2532

Leszczyński M, Barczak A, Kajikawa Y, et al (2020) Dissociation of broadband high-frequency activity and neuronal firing in the neocortex. Science Advances 6(33):eabb0977. https://doi.org/10.1126/sciadv.abb0977

López Zunini RA, Baart M, Samuel AG, et al (2020) Lexical access versus lexical decision processes for auditory, visual, and audiovisual items: Insights from behavioral and neural measures. Neuropsychologia 137:107305. https://doi.org/10.1016/j.neuropsychologia.2019.107305

Lynott D, Connell L, Brysbaert M, et al (2019) The Lancaster Sensorimotor Norms: Multidimensional measures of Perceptual and Action Strength for 40,000 English words. https://doi.org/10.31234/osf.io/ktjwp

Mante V, Sussillo D, Shenoy KV, et al (2013) Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature 503(7474):78–84. https://doi.org/10.1038/nature12742

Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods 164(1):177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Markowitz DA, Curtis CE, Pesaran B (2015) Multiple component networks support working memory in prefrontal cortex. Proceedings of the National Academy of Sciences 112(35):11084–11089. https://doi.org/10.1073/pnas.1504172112

Marslen-Wilson WD (1987) Functional parallelism in spoken word-recognition. Cognition 25(1-2):71–102. https://doi.org/10.1016/0010-0277(87)90005-9

Martin AE (2020) A Compositional Neural Architecture for Language. Journal of Cognitive Neuroscience 32(8):1407–1427. https://doi.org/10.1162/jocn_a_01552

Mesgarani N, Cheung C, Johnson K, et al (2014) Phonetic Feature Encoding in Human Superior Temporal Gyrus. Science 343(6174):1006–1010. https://doi.org/10.1126/science.1245994

Michalareas G, Vezoli J, van Pelt S, et al (2016) Alpha-Beta and Gamma Rhythms Subserve Feedback and Feedforward Influences among Human Visual Cortical Areas. Neuron 89(2):384–397. https://doi.org/10.1016/j.neuron.2015.12.018

Noppeney U, Patterson K, Tyler LK, et al (2006) Temporal lobe lesions and semantic impairment: A comparison of herpes simplex virus encephalitis and semantic dementia. Brain 130(4):1138–1147. https://doi.org/10.1093/brain/awl344

Norris D, McQueen JM (2008) Shortlist B: A Bayesian model of continuous speech recognition. Psychological Review 115(2):357–395. https://doi.org/10.1037/0033-295X.115.2.357

Oganian Y, Chang EF (2019) A speech envelope landmark for syllable encoding in human superior temporal gyrus. Science advances 5(11):eaay6279

Ossmy O, Fried I, Mukamel R (2015) Decoding speech perception from single cell activity in humans. NeuroImage 117:151–159. https://doi.org/10.1016/j.neuroimage.2015.05.001

Patterson K, Nestor PJ, Rogers TT (2007) Where do you know what you know? The representation of semantic knowledge in the human brain. Nature Reviews Neuroscience 8(12):976–987. https://doi.org/10.1038/nrn2277

Perdikis D, Huys R, Jirsa VK (2011) Time Scale Hierarchies in the Functional Organization of Complex Behaviors. PLoS Computational Biology 7(9):e1002198. https://doi.org/10.1371/journal.pcbi.1002198

Pesaran B, Vinck M, Einevoll GT, et al (2018) Investigating large-scale brain dynamics using field potential recordings: Analysis and interpretation. Nature Neuroscience https://doi.org/10.1038/s41593-018-0171-8

Pillai AS, Jirsa VK (2017) Symmetry Breaking in Space-Time Hierarchies Shapes Brain Dynamics and Behavior. Neuron 94(5):1010–1026. https://doi.org/10.1016/j.neuron.2017.05.013

Pulvermüller F (2018) Neural reuse of action perception circuits for language, concepts and communication. Progress in Neurobiology 160:1–44. https://doi.org/10.1016/j.pneurobio.2017.07.001

Pylkkänen L (2020) Neural basis of basic composition: What we have learned from the red–boat studies and their extensions. Philosophical Transactions of the Royal Society B: Biological Sciences 375(1791):20190299. https://doi.org/10.1098/rstb.2019.0299

Quiroga RQ, Nadasdy Z, Ben-Shaul Y (2004) Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural computation 16(8):1661–1687

Rahimi S, Farahibozorg SR, Jackson R, et al (2022) Task modulation of spatiotemporal dynamics in semantic brain networks: An EEG/MEG study. NeuroImage 246:118768. https://doi.org/10.1016/j.neuroimage.2021.118768

Ralph MAL, Jefferies E, Patterson K, et al (2017) The neural and computational bases of semantic cognition. Nature Reviews Neuroscience 18(1):42–55. https://doi.org/10.1038/nrn.2016.150

Remington ED, Narain D, Hosseini EA, et al (2018) Flexible Sensorimotor Computations through Rapid Reconfiguration of Cortical Dynamics. Neuron 98(5):1005–1019.e5. https://doi.org/10.1016/j.neuron.2018.05.020

Rutten S, Santoro R, Hervais-Adelman A, et al (2019) Cortical encoding of speech enhances task-relevant acoustic information. Nature Human Behaviour 3(9):974–987. https://doi.org/10.1038/s41562-019-0648-9

Schwartz MF, Kimberg DY, Walker GM, et al (2009) Anterior temporal involvement in semantic word retrieval: Voxel-based lesion-symptom mapping evidence from aphasia. Brain 132(12):3411–3427. https://doi.org/10.1093/brain/awp284

Scott SK (2000) Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123(12):2400–2406. https://doi.org/10.1093/brain/123.12.2400

Sjerps MJ, Fox NP, Johnson K, et al (2019) Speaker-normalized sound representations in the human auditory cortex. Nature Communications 10(1). https://doi.org/10.1038/s41467-019-10365-z

Truccolo W (2016) From point process observations to collective neural dynamics: Nonlinear Hawkes process GLMs, low-dimensional dynamics and coarse graining. Journal of Physiology-Paris 110(4):336–347. https://doi.org/10.1016/j.jphysparis.2017.02.004

van Kerkoerle T, Self MW, Dagnino B, et al (2014) Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. Proceedings of the National Academy of Sciences 111(40):14332–14341. https://doi.org/10.1073/pnas.1402773111

Vignali L, Xu Y, Turini J, et al (2023) Spatiotemporal dynamics of abstract and concrete semantic representations. Brain and Language 243:105298. https://doi.org/10.1016/j.bandl.2023.105298

Visser M, Lambon Ralph MA (2011) Differential Contributions of Bilateral Ventral Anterior Temporal Lobe and Left Anterior Superior Temporal Gyrus to Semantic

Processes. Journal of Cognitive Neuroscience 23(10):3121–3131. https://doi.org/10.1162/jocn_a_00007

Vyas S, Golub MD, Sussillo D, et al (2020) Computation Through Neural Population Dynamics. Annual Review of Neuroscience 43(1):249–275. https://doi.org/10.1146/annurev-neuro-092619-094115

Yi HG, Leonard MK, Chang EF (2019) The Encoding of Speech Sounds in the Superior Temporal Gyrus. Neuron 102(6):1096–1110. https://doi.org/10.1016/j.neuron.2019.04.023

Zhang Y, Ding Y, Huang J, et al (2021) Hierarchical cortical networks of "voice patches" for processing voices in human brain. Proceedings of the National Academy of Sciences 118(52):e2113887118. https://doi.org/10.1073/pnas.2113887118