Low-dimensional neuronal population dynamics in anterior superior temporal gyrus reactivate phonetic representations during semantic processing

Pavo Orepic[1], Wilson Truccolo[2,3], Sydney S. Cash[4],
Anne-Lise Giraud[1,5], 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 Sciences, Brown University, Providence, Rhode Island, United States of America.
[4]Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, United States of America.
[5]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 sequentially encodes speech through a hierarchy of cognitive processes, from early representations of acoustic and phonetic features to late semantic encoding. Yet the mechanisms by which neural representations are transformed across the speech hierarchy remain poorly specified. 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 auditory processing, during a semantic categorization task and natural speech perception. In both conditions, low-dimensional neuronal population dynamics revealed a distributed and parallel encoding of phonetic and semantic representations. During natural speech, the low-dimensional dynamics were simultaneous to a power increase in the beta and low-gamma local field potentials, reflecting concurrent instantiation of top-down predictive and bottom-up cumulative processes. Our results support a mechanism for phonetic-to-semantic transformations encoded at the neuronal population level.

Keywords: speech perception, phonetic, semantic, neuronal population dynamics, neural manifold

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 a sequence along 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). The ATL is seen as a conceptual semantic hub, connected to modality-specific sources of information (Patterson et al, 2007; Ralph et al, 2017). 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.

However, recent neuroimaging findings suggested that these transformations across brain regions might be much less modular and sequential than predicted by the classical modular view of language processing (Ralph et al, 2017). Fine-grained electrocorticography (ECoG) recordings from the middle and posterior STG reveal encoding of phonetic features without strict spatial segregation, but rather mixed interleaved representations (Hamilton et al, 2021). At the semantic level, the anterior superior temporal gyrus, 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 (Visser and Lambon Ralph, 2011). In functional MRI (fMRI) data, neural activity recorded through incrementally higher cognitive brain regions correlates with increasingly deeper layers of large language models (Caucheteux et al, 2022). These findings suggest that some brain regions represent multiple speech features, making them suitable candidates for housing the transformations across the speech hierarchy.

Studies using time-resolved recording techniques such as intracranial EEG, EEG, or MEG, additionally show simultaneous encoding of features across the speech hierarchy, organized by increasingly larger time scales (Heilbron et al, 2022; Gwilliams and King, 2020; Keshishian et al, 2023). For instance, phonetic features are short-lived and typically processed around 100-200 ms in the STG (Mesgarani et al, 2014; Hamilton et al, 2021), while syntactic and semantic composition activity appears in the ATL as early as 200-250 ms (Pylkkänen, 2020; Friederici and Kotz, 2003). Other semantic aspects are encoded later around 400 ms with a large integration time, as reflected in EEG and MEG recordings by the well-characterized N400 event-related potential for both semantic composition and lexical decision (Borghesani et al, 2019; 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). Those neuroimaging methods, however, have not so far elucidated the precise mechanisms of how phonetic features are transformed into semantic representations.

A prominent mechanism for the instantiation of these transformations is the long-standing analysis-by-synthesis framework. It suggests that speech comprehension combines several sequential stages (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 proposition corresponding to this semantic representation is then generated and directly compared to the actual acoustic input. Depending on the comparison, the proposition is either accepted or rejected in favor of new updated hypotheses. These hierarchical interactions are typically reflected in the power of different 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).

To identify the mechanisms underlying speech transformations, it might be necessary to investigate speech processing at a smaller spatial scale giving access to the neuronal spiking level, an endeavor that is hindered by the invasiveness of single neuron sampling in humans. Speech encoding has thus never been characterized at the level of neuronal population dynamics. However, recent findings suggest that complex cognitive processes are encoded at a local scale by low-dimensional functional spaces, also called *neural manifolds*, whose main property is to encode behavioral features in a highly compact way (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, neuronal dynamics trajectories on the 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). It is thus plausible that different aspects of speech, including transformations across the speech hierarchy, are encoded in such local multidimensional neuronal population dynamics. Moreover, the condition- and history-dependent organization of these neuronal trajectories on the manifold could ideally serve the purpose of integrating phonetic features into higher-level representations (Pulvermüller, 2018; Yi et al, 2019). To date, however, only a few studies have reported single unit activity associated with speech processing (Chan et al, 2014, 2011; Ossmy et al, 2015; Lakretz et al, 2021). One of those reported the tuning of single units for phonetic features (Chan et al, 2014). We re-analyzed those data in the light of the neuronal manifold framework to address the mechanisms underlying phonetic-to-semantic transformations.

These unique recordings come 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. 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 as probed by the MEA. Phoneme-related neuronal dynamics were organized by their corresponding phonetic content during specific periods of speech processing. 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 in parallel with the encoding of semantic features, both culminating around 400 ms after word onset, and with peaks in low-gamma and beta power. These findings suggest that phonetic features directly interact with semantic representations by the simultaneous instantiation of predictive bottom-up and top-down mechanisms, in agreement with the analysis-by-synthesis framework (Bever and Poeppel, 2010).

Results

Semantic and phonetic neuronal encodings in the aSTG

We recorded microelectrode array and ECoG signals of a 31-year-old patient with pharmacologically resistant epilepsy (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

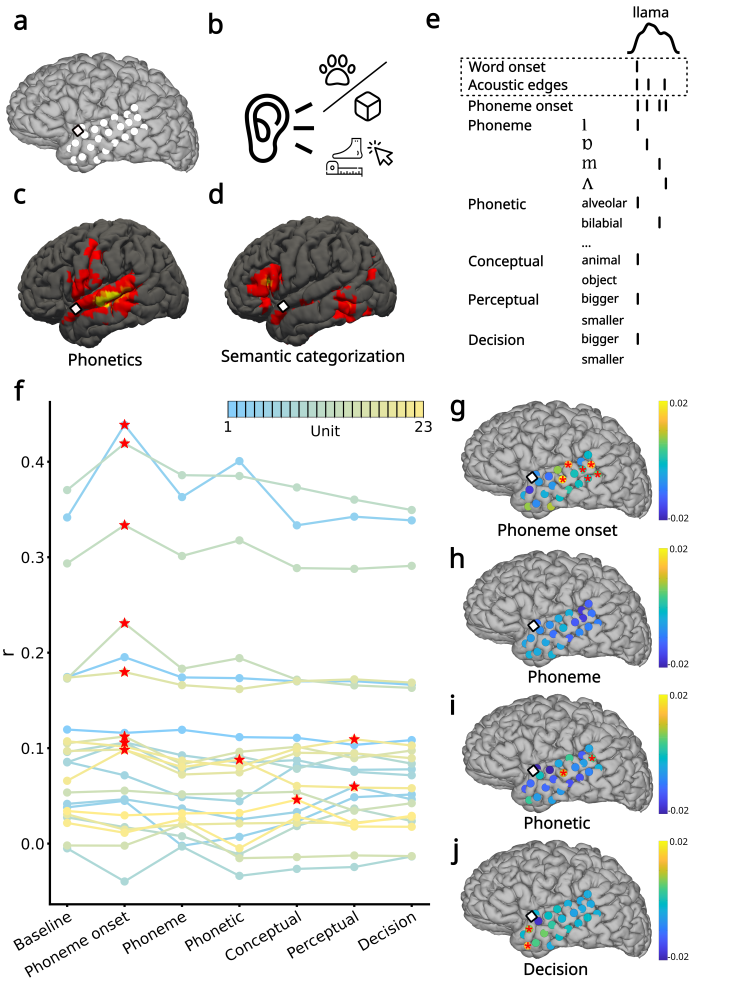

Fig. 1 | **Semantic and phonetic encoding at the single-unit level in the aSTG. a.** Locations of the microelectrode array (square) and the ECoG array (circles) on the cortical surface. **b.** Experimental design of the auditory semantic categorization task. The participant heard 400 nouns that denoted either animals or objects and was instructed to press a button if the item he heard was bigger than a foot in size. **c.** fMRI correlates of phonetic processing obtained from fMRI large-scale databases. **d.** fMRI correlates of semantic categorization. **e.** mTRF features for the example 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) for each model and 23 single units with the highest firing rates in the ensemble. The traces for different units are color-coded based on their firing rate, from lower (blue) to higher (yellow). Red stars indicate model-unit pairs that were significantly higher than the baseline in a permutation test. **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 ECoG channels that were significantly higher than the baseline in a permutation test.

Fig. 1a). The participant performed an auditory semantic categorization task, and later engaged in a spontaneous conversation. In the auditory task, the participant was instructed to indicate, by pressing a button, whether the word he heard 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 between two categories, either bigger or smaller than a foot, resulting in a balanced 2-by-2 design.

We used multivariate temporal response function (mTRF) models to contrast the encoding of different linguistic processes and speech features (Methods). To obtain an initial list of relevant linguistic processes for the mTRF analysis, we also used large-scale databases of fMRI and lesion data to identify the cognitive processes associated with the aSTG where the intracortical MEA was located (Yarkoni et al, 2011a; Dockès et al, 2020). Nearby brain regions have been indicated to process both phonetics (middle STG) (Fig. 1c) and semantic categorization (anterior superior temporal sulcus, anterior middle temporal gyrus) (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 and 2). The semantic decision feature was regressed separately from the perceptual category feature as the participant responded correctly in 80.25% of the trials. We constructed a baseline model based on the 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 and 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, signal processing). We considered the 23 (over a total of 176) spike-sorted units) 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). We observed that only a few units significantly responded to different features. Specifically, eight units responded significantly to phoneme onset ($p < 0.05$ based on chance level performance of a surrogate distribution, see Methods), one to phonetic features, one to conceptual category, 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 features. Six ECoG electrodes significantly correlated to phoneme onset (Fig. 1g) and two for phonetic features (Fig. 1i) in the posterior middle and superior temporal gyrus, and two electrodes in the aTL significantly activated to semantic decision (Fig. 1j). No other electrodes showed significant correlations (Supplementary Fig. 3).

Additionally, we investigated the encoding of phonetic features. To this aim, we separated phonetic features into four different groups (vowel first formant, vowel second formant, consonant manner of articulation, and consonant place of articulation) and separately contrasted a model for each phonetic feature group with a base model including only phoneme onsets and acoustic features. We found that two units were significant for consonant manner and vowel first formant models, and none for the other two models (Supplementary Fig. 3g). None of the individual phonetic categories were significantly correlated at the ECoG electrode adjacent to the MEA (Supplementary Fig. 14).

Distributed phonetic encoding

We showed that only a few single units, when taken independently, encoded phonetic features. However, several other units showed some non-significant changes (permutation test, Fig. 1f) that might have contributed to the overall encoding through neuronal population dynamics. Indeed, 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, suggesting a mean-field population effect (Fig. 2a). We therefore hypothesized that a read-out of phonetic encoding emerged at the level of the neuronal population dynamics and could be described by a low-dimensional neural manifold.

To confirm our 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 five different units (Fig 2d). This confirmed that a large part of the variance is explained by the low-dimensional and correlated activity of several units, indicating the presence of a manifold supporting phoneme neuronal representations.

We then asked whether the obtained low-dimensional neural manifold carried a functional read-out 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 (vowels first formant, vowels second formant, consonants manner, consonants 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).

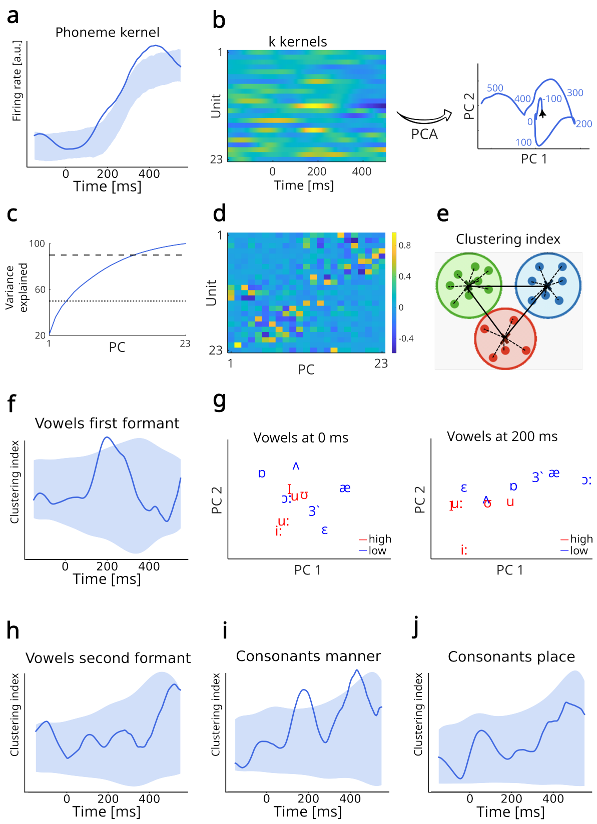

Fig. 2 | Distributed encoding and clustering of phonemes along phonetic categories. **a.** Average mTRF kernel for phoneme feature. The shaded area indicates the 95% confidence region of the surrogate (chance-level) distribution. The phoneme kernel was significant at two time periods, centered around 200 and 400 ms. **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 then projected into a two-dimensional space using principal component analysis (PCA), resulting in one phoneme trajectory. Numbers along the trajectory indicate the time at which the trajectory reached the corresponding location. **c.** PCA variance explained. Four PCs explain 50% of variance (dotted line) and 15 explain 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.** The schematics show how the clustering index is computed as intra-cluster distance (full lines) subtracted from between-cluster distances (bold lines). **f.** Clustering index for vowels grouped by the first formant (high and low tongue position). The shaded area represents the 95% confidence region of the surrogate (chance-level) distribution. This clustering index is 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 (high and low tongue position). Mirroring the increase in the clustering index, the separation between those two phonetic features in the PC space is absent at time 0, becoming evident at 200 ms. **h.** Clustering index for vowels grouped by the second formant. **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.

We observed significant clustering within the vowel first formant group between the two phonetic features that reflects the high and low position of the tongue (Fig. 2f). The clustering specifically occurred at a time window centered around 200 ms, which was apparent in the position of the vowels in the two-dimensional PC space (Fig. 2g). We replicated our findings using higher dimensional PC spaces (Supplementary Fig. 4). 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, Figure 2h, Supplementary Fig. 5 and Fig. 6). 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 independently of the chosen number of dimensions (Supplementary Fig. 7). Clustering using consonant place of articulations (bilabial, labiodental, dental, alveolar, velar, uvular, and glottal features) was not significant in any low-dimensional space (Fig 2l, Supplementary Fig. 8).

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 (Supplementary Material, illustrated in Supplementary Figures 7-10, and summarized in Supplementary Table 3). Linear discriminant analysis (LDA) classifier revealed linear separability of the same two phonetic categories in the same time windows (i.e., around 200 ms for vowels first formant and both around 200 and 400 ms for consonants manner of articulation, Supplementary Figure 7). 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. (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.

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

To investigate whether our results generalize 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 auditory task, we performed spike sorting and selected the 23 most spiking units (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). We then fitted the mTRF encoding models as in the auditory task dataset. Similarly to the task, we observed that only a few units were significantly correlated with speech features (Fig. 3a), and that the average phoneme kernel became significant compared to its surrogate distribution around 200 and 400 ms, suggesting again a population effect (Fig. 3b).

We thus proceeded as before by performing PCA on the kernels. As in the task, 50% of the variance was explained by 5 PCs (Fig. 3c), and the processing of phonemes was distributed across units (Figure 3d). We further computed the clustering index for phonetic features in the low-dimensional neural manifold spanned by the first two PCs. Remarkably, we observed clustering index profiles similar to the auditory task.

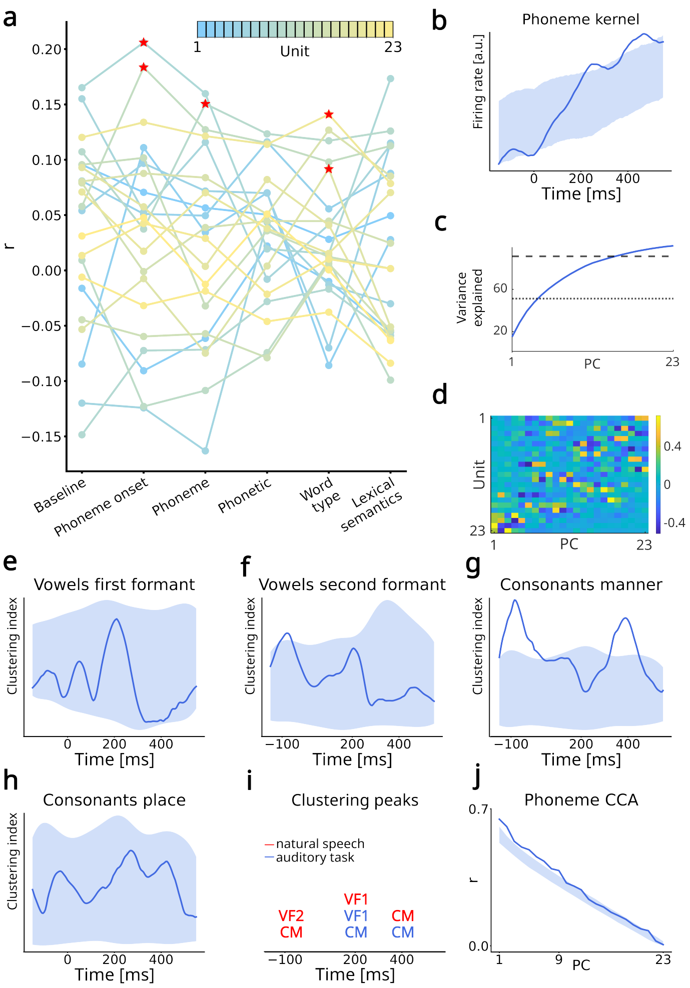

Fig. 3 | **Generalization to natural speech perception a.** r values for each model and 23 most-spiking units. The units are color-coded based on their firing rate, from blue (lower) to yellow (higher). Red stars indicate model-unit pairs significantly higher than the baseline in a permutation test. **b.** Average mTRF kernel for phoneme feature. The shaded area indicates the 95% confidence region of the surrogate (chance-level) distribution. Phoneme kernel was significant at two time periods, centered around 250 and 450 ms. **c.** PCA variance explained. Four PCs explain 50% of variance (dotted line) and 15 explain 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. **i.** Overview of the significant clustering peaks observed both during the auditory task and natural speech perception. **j.** Canonical correlation analysis (CCA) between the PC projections of phoneme kernels obtained during the auditory task and natural speech perception. The shaded area indicates the 95% confidence region of the surrogate distribution. CCA revealed significant correlations for the first nine PCs.

For vowel first formants, we observed a peak of the clustering index at around 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. 15). This peak was also significant in all control analyses (LDA classifier, rank regression, and k-means, Supplementary Fig. 16). 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 related to prediction mechanisms present in natural speech perception but not during the auditory task. For consonant manner of articulation, we observed significant clustering indices around 400 ms, as well as an additional peak around -100 ms that was possibly related to prediction mechanisms (Fig. 3g). The peak at 200 ms that was observed in the task did not occur here. Finally, for consonants place of articulation, no significant peak of clustering index was observed (Fig. 3h). We performed the same control analyses as before (LDA classifier, rank regression, and k-means clustering), and could replicate all the findings described here (Supplementary Material, illustrated in Supplementary Figures 12-19, and summarized in Supplementary Table 4). Timeline on Fig. 3i summarizes the significant clustering peaks across the two datasets.

Finally, we compared the similarity of low-dimensional phoneme trajectories obtained in the auditory 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 auditory 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 in a similar way - individual phonemes trace highly similar trajectories in the low-dimensional neural manifold.

Encoding of semantic features

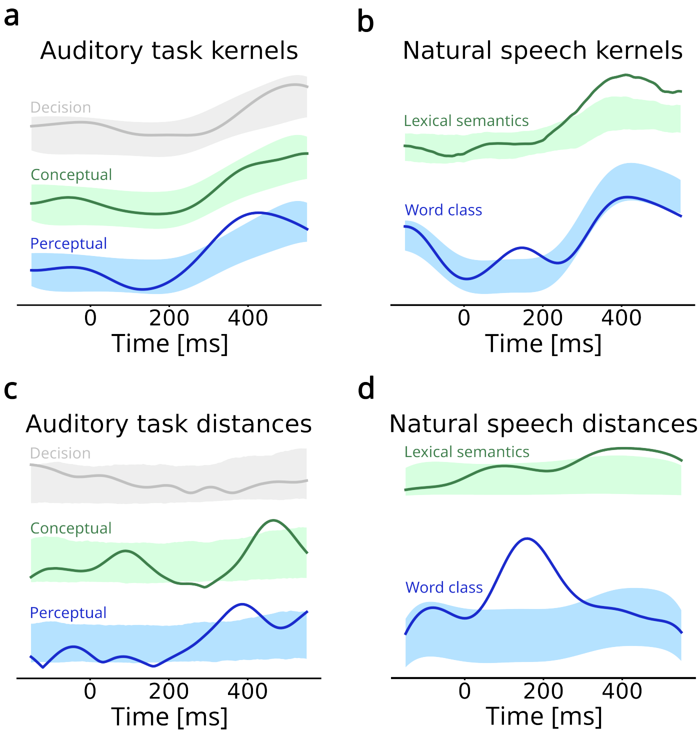

Fig. 4 | **Encoding of semantic features. a.** Auditory task semantic feature kernels. The shaded areas indicate the 95% confidence regions of the surrogate (chance-level) distribution. The perceptual category feature is significantly encoded during a time window centered around 400 ms. **b.** Euclidean distances in the PC space between the auditory task feature kernels. The perceptual semantic feature kernels are separated in the PC space slightly before the categorical semantic feature kernels (around 400 and 450 ms, respectively). **c.** Natural speech perception semantic kernels. Word class is processed around 150 ms, and lexical semantics is processed in a wide window around 400 ms. **d.** Euclidean distances in the PC space between the semantic features during natural speech perception are maximal around 150 ms and 400 ms respectively.

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 auditory 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 period at around 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 auditory task.

We then turned to the neuronal population dynamics. We projected the three semantic kernels of the auditory task to the low-dimensional neural manifold. Because we only have two categories for each of the 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. 24). 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 around 450 ms, even though the averaged perceptual 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), we found a strong separation between word classes at 150-200 ms, possibly reflecting syntactic processing, and a significant separation between lexical semantics kernels around 400 ms.

Parallel 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. Specifically, we hypothesized that phonetic features might be represented with shorter delays as the phoneme position in the word increases, resulting in simultaneous

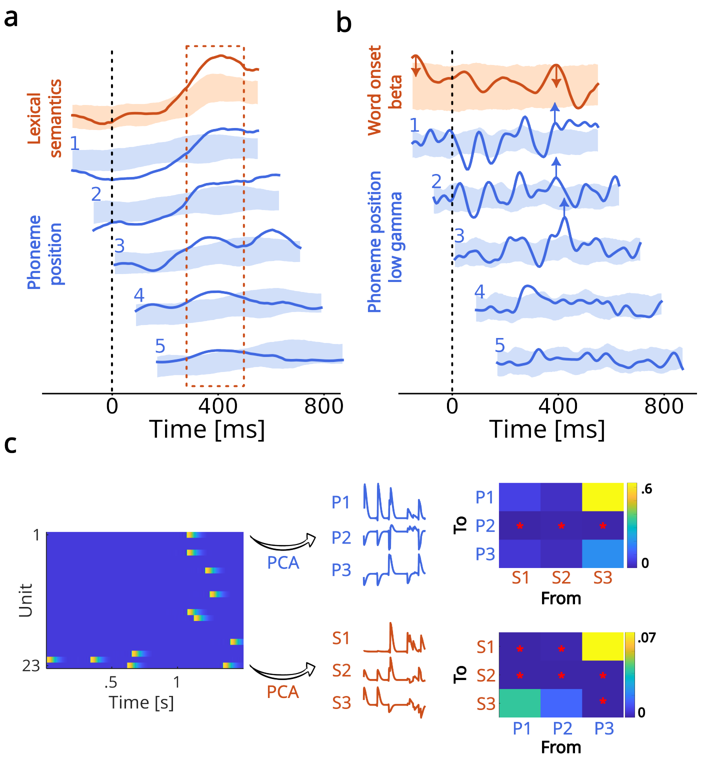

Fig. 5 | **Parallel encoding of bottom-up phonetic and top-down semantic features a.** Kernels for phoneme positions (blue) within a word shifted by average phoneme duration (80 ms) and aligned with the lexical semantic kernel at word onset (red). Aligned phoneme position kernels peaked simultaneously with the onset of the lexical semantic kernel peak. **b.** Beta band kernel for word onset is indicated in red, and low-gamma kernels for phoneme positions are indicated in blue and shifted by average phoneme duration. The late peak in the beta word-onset kernel occurred simultaneously with the aligned low-gamma phoneme onset peaks. **c.** 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).

activation of all past phonemes at 400 ms after word onset, in parallel to semantic processing. To verify this hypothesis, we created additional mTRF features that regressed phoneme positions within each word (i.e., phoneme order). We then aligned phoneme position kernels with word onset by shifting each position for the corresponding multiple of the average phoneme duration (80 ms). If our hypothesis is true, a significant period of activations should align across the shifted kernels. For the auditory task, we did not find strong significant peaks for the distinct phoneme positions, which precluded making strong conclusions about the validity of our hypothesis (Supp. Fig. 25). For natural speech perception, however, we did observe such an alignment of significant peaks for phoneme positions 1-5 at 400 ms, simultaneously to semantic encoding (Fig. 5a). This suggests that our hypothesis might hold true for natural speech perception.

We then asked whether this parallel encoding reflected the bottom-up and top-down predictive processes of the analysis-by-synthesis framework (Bever and Poeppel, 2010). For this, we focused on natural speech perception, where more semantic top-down processes are expected, and where there are no predictability biases related to the task design, such as fixed stimulus onset timing. The occurrence of bottom-up processes is typically accompanied by low-gamma peaks in the LFP power, while top-down processes are reflected in LFP beta power peaks (Arnal and Giraud, 2012). To identify the timing of bottom-up and top-down processes, we thus extracted and averaged the LFP power across the MEA contacts for both the low-gamma (30-50 Hz) and beta (15-25 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 phoneme positions for the low-gamma band (reflecting position-dependent bottom-up processing). For the beta band, we found two significant positive peaks at 100 before word onset and 400 ms after word onset (Fig. 5b, red traces). In the low-gamma band, we aligned phoneme position kernels based on word onset as before, and found an alignment of positive significant peaks at 400 ms after word onset (Fig. 5b, blue traces). This finding suggests that both bottom-up phonetic and top-down semantic processing occur at 400 ms, in agreement with the analysis-by-synthesis framework. We observed other negative and positive significant peaks before 400 ms in the low-gamma band that were not accompanied by a beta peak. Following the analysis-by-synthesis framework, these peaks might reflect bottom-up-only processes that contribute to information build-up leading to the 400-ms beta top-down peak.

Moreover, they might correspond to the early significant negative peaks observed for phoneme positions one and two at the neuronal level (compare Fig. 5a and b).

Finally, we investigated whether the phonetic low-dimensional neural dynamics were time causally predictive of the semantic low-dimensional dynamics (reflecting a bottom-up process) or, inversely, 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. 5c).

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

Discussion

In this study, we showed that the low-dimensional neuronal population dynamics recorded in the aSTG during speech processing encoded phonetic and semantic features during both a semantic categorization auditory task and natural speech perception. We identified a neural manifold for both features and observed a functional separation of their corresponding trajectories across time. Specifically, phoneme trajectories clustered according to their phonetic features in both conditions, while semantic trajectories separated based on their perceptual and conceptual features during the auditory task, and on lexical semantic and word class features during natural speech. Moreover, the low-dimensional phoneme representations traced highly correlated feature-specific trajectories during both conditions. During natural speech, semantic and phonetic encoding occurred in parallel at 400 ms after word onset. Specifically, the timing of phonetic encoding was increasingly shorter for successive phoneme positions within the word, such that processing of all phonemes simultaneously culminated at 400 ms after word onset. This parallel encoding of phonemes and semantics was mirrored in the bottom-up and top-down processes as measured by peaks in the low-gamma and beta power, and there was a bidirectional causal relationship between their low-dimensional representations.

We observed simultaneous encoding of phonetic and semantic features in the same focal cortical area of the aSTG. This finding extends the traditional models of speech processing (Hickok and Poeppel, 2007) that posit a distinct modular processing of different levels across the speech hierarchy. Specifically, we observed that the average activity of the same neuronal population encoded both phonemes (time-locked to each phoneme onset), and semantic information (time-locked to each word onset). On average, phonemes were encoded in two time windows centered around 200 and 400 ms after the phoneme onset (Figure 2a), and semantic information around 400 ms after the word onset (Figure 4a). To investigate whether there is a relationship between these two effects, we then time-locked phonemes to the word onset as well and observed that the later they occurred within the word, the sooner they were processed. Moreover, when aligned to the word onset, the processing of phonemes overlapped with the semantic processing window at 400 ms (Fig. 5a). This shows that the two processing peaks observed on the average phoneme kernel in fact represent one peak that is divided in time as a function of phoneme position within the word and suggests that there might be a functional relationship between these two processes.

Previous studies suggest that interactions between phonetic and semantic processing could be instantiated through recurrent neural networks, where phonetic activations persist or are reactivated at the timing of semantic processing (Bever and Poeppel, 2010). Others suggest that phonetic representations might successively change a unique representation that correlates with each previous phonetic feature at the time of semantic processing (Perdikis et al, 2011; Yi et al, 2019; Martin, 2020).

Finally, according to the prominent analysis-by-synthesis framework, the two levels are related through dynamic bottom-up and top-down predictive mechanisms (Arnal and Giraud, 2012). One feature of this framework's proposal is that supportive evidence can be obtained by investigating the power increase in different LFP frequency bands - namely, top-down processes have been shown to be associated with the beta band, and bottom-up with the low-gamma increases (Arnal and Giraud, 2012). Here, we followed the same approach and investigated the increase in the beta and the (shifted) low-gamma band corresponding to word onset. Replicating our firing rate findings, we observed significant alignments of top-down and bottom-up processes, as indicated by transients in these two frequency bands, around 400 ms post-word-onset (Fig. 5b). Moreover, we observed several significant low-gamma increases occurring before the 400-ms alignment. This suggests that, following the word onset, the population activity first encodes phonemes in a bottom-up fashion, which might lead to a build-up of information necessary for semantic processing that occurs at 400 ms. Once enough bottom-up phonetic information is accumulated, a top-down semantic process is initiated at 400 ms, which then initiates a new wave of bottom-up processes at the phoneme level. We further confirmed the existence of a bidirectional relationship between phonetic and semantic processing through Granger causality analysis (Fig. 5c).

Our findings in a controlled auditory semantic categorization task generalized to a natural conversation. This is particularly remarkable, as the two datasets have several important differences. In the controlled task, the participant heard isolated word recordings, all of which were nouns of similar duration and normalized for sound intensity, and focused on a simple cognitive task – assessing the size of the heard objects and animals. Contrary to the controlled task, natural speech perception involves many other cognitive and perceptual mechanisms. First, the natural conversation included sentences and all word types, not just isolated nouns. This allowed us to generalize our semantic findings from the narrow and carefully-defined task scenario (i.e., animals-vs-objects) to a much larger category of lexical semantics, as well as to word class encoding that is suggestive of syntactic processing. Second, even though speech was uttered by another speaker, we still retrieved invariant phonetic representations, both in time and in the shape of the low-dimensional trajectories. This indicates that the phonetic representations we found are speaker-normalized, similar to previous results at the LFP level (Sjerps et al, 2019). Third, sounds were much more varied in their durations and intensities as compared to the controlled task, likely entailing other variations, 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, including context effects, which were only present for natural speech. These strong predictive processes were revealed by our LFP analysis, showing significant beta power bands at -100 ms and 400 ms, in agreement with the parallel encoding periods we found in the neural manifold.

Another interesting aspect of natural speech is that it allows us to contrast semantic and syntactic features. While we found that syntactic and semantic are both encoded by the same neuronal population dynamics, they were differentiated by their timings, with the syntactic processing occurring at around 150 - 200 ms, while the semantic processing occurred later, at around 400 ms. This syntactic processing might be a sign of early combinatorial processing, as reported with MEG (Pylkkänen, 2020). It might also help in participating in early conceptual and perceptual categorization in the ATL at around 200 ms (Borghesani et al, 2019; Chan et al, 2011; Chen et al, 2016; Dehaene, 1995; Hinojosa et al, 2001). We found that other semantic categorization occurred later, around 400 ms, which is reminiscent of the N400 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). In particular, the perceptual aspect of semantics occurred slightly before the conceptual one, in accordance with previous results (Borghesani et al, 2019)

Our work extends and experimentally supports the neural manifold framework for neural speech processing. Indeed, we found that a large amount of the neuronal variance related to phonetic and semantic encoding could be accounted for by a low-dimensional space, where the functional read-out emerges at the level of the neuronal population dynamics (Vyas et al, 2020; Jazayeri and Ostojic, 2021; Chung and Abbott, 2021). The neural manifold hypothesis is an efficient encoding method that allows rich dynamics to be encoded in noise-resilient latent dynamics. We provided evidence that speech encoding at the phonetic and semantic level operates at the level of neuronal population dynamics, akin to many other cognitive processes, such as sensorimotor processing, decision making, or object recognition (Gallego et al, 2017; Mante et al, 2013; DiCarlo and Cox, 2007). In particular, we observed that 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 explain why we did not observe such encoding in aSTG ECoG signals, which roughly correspond to 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 (auditory task and natural speech) up to the 9th PC (Figure 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, the common population patterns remain preserved. Collectively, our findings illustrate that the neural manifold framework is particularly suitable for investigating neuronal dynamics during speech processing, as it allows for a subtle characterization of intricate dynamical processes spanned across the speech hierarchy.

Another important finding that was facilitated through the neural manifold framework is that low-dimensional phoneme trajectories cluster according to their phonetic features. Specifically, vowels clustered based on their first formant feature and consonants based on their manner of articulation. This extends previous works that described the encoding of phonetic features across the temporal lobe (Mesgarani et al, 2014; Hamilton et al, 2021) by demonstrating that these features can be constructed at a more fine-grained level through the coordinated dynamics of individual phoneme representations. Moreover, the time windows in which these clusterings occurred matched the periods in which the average phoneme kernel was significant (200 and 400 ms), again pointing to the importance of fine-tuned coordinated activity of individual neurons. Importantly, we found this effect to be very robust, as it generalized from auditory task to natural speech perception and was replicated through different analysis approaches in both conditions (i.e., linear discriminant analysis, rank regression, k-means clustering).

There are a few limitations and open questions related to our findings. First, our dataset is limited to one participant and we thus cannot assess the generalizability of these findings to other participants. However, human intracortical MEA recordings are extremely rare and such datasets represent an invaluable opportunity to investigate cognitive mechanisms at the level of ensembles of single-unit action potentials recorded from neuronal populations. In fact, we observed speech encodings that were not detectable even at the most adjacent ECoG contact, demonstrating the unique advantages of this approach. Second, such high precision comes with the price of low spatial generalizability. Namely, 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 possible that aSTG is organized into small functional subdivisions and that we would observe significant clustering based on other phonetic features if the array was implanted elsewhere. Finally, the population of recorded neurons is also limited by the design of the MEA. Interestingly, only a few tens of units over around two hundred identified responded to different speech features, showing that the neuronal population encoding is relatively sparse. However, this further demonstrates the importance of investigating the coordinated activity of neuronal population firing, as it is able to reveal patterns that are absent at the level of isolated neuronal activity.

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. 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

The data used in this study appeared first in (Chan et al, 2014).

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

Neural recordings

Cortical local field potentials were recorded with an 8-by-8, 1-cm electrode distance, subdural ECoG array (Adtech Medical) 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 on 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.

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 auditory task and the natural conversation respectively. In both sessions and for a fair comparison, the analysis was done on the 23 most-spiking units (task: $0.43 - 5.27$ spikes/second; speech: $0.1 - 3.39$ spikes/second). For the auditory task, we considered the units with hiring rates higher than 0.3 spikes/second. For the natural speech, we kept the same number of units as in the auditory task.

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 task session, 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 conversation session, we selected 91 segments of the firing rate where a person was talking to the participant (see below).

The signals from the ECoG grid were first filtered to remove line noise using a notch filter at 60 Hz and harmonics (120, 180 Hz, 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 semilogarithmically respectively. The resulting BHA was downsampled to 100 Hz.

Auditory stimuli

Neuronal data was recorded during two separate experimental sessions, that took place on the same day, two hours apart. 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 noun words 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. In order 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 natural conversation with another person present in the room. The conversation 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 conversation.

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: vowels first formant (open, close), vowels second formant (front, back), consonants articulation place (bilabial, labiodental, alveolar, velar, uvular, glottal), consonants articulation manner (plosive, nasal, fricative, approximant, lateral approximant). See also Supplementary Table 1 for vowels and Supplementary Table 2 for consonants).

mTRF features

For both sessions, we extracted the following features: word onsets, acoustic edges (envelope rate peaks), phoneme onsets, phonemes, and phonetic categories. For the auditory task, we additionally computed the following semantic features: perceptual category, conceptual category, and semantic decision. For the natural conversation, we additionally created word class and lexical semantics features. All stimuli were designed as Dirac functions centered at the onset of the corresponding feature.

Word onset was marked by a Dirac function centered at the onset of each word. Acoustic edges were defined as local maxima in the derivative of the speech envelope (Oganian and Chang, 2019). Speech envelope was computed as the logarithmically scaled root mean square of the audio signal using the MATLAB mTRFenvelope function. Phoneme onset feature indicated onsets of all phonemes in a word. Phoneme identity was multivariable with 32 regressors, each indicating onsets of a different phoneme, as defined by the IPA table. Phonetic categories was multivariable, and defined as described in the Secion 5. All other features are multivariables with a Dirac function centered 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 neuronal data (firing rates or BHA), with resulting kernels in the time range between -200 and 600 ms. The first and last 50 ms were not considered in the analysis, in order to avoid possible edge effects resulting from regressing Dirac stimuli to continuous neuronal data. For both task and conversation sessions, 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 neuronal data.

Cross-validation

For the auditory 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 5 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, chosen among the following 13 lambda values: (10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}, 10^{4}, 10^{5} ,10^{6}). The optimal lambda is then used to retrain the model on the 320 words of the training set of the outer cross-validation loop fold. The model predictions are then computed for the 80 words of the test set. Pearson's correlation with the neuronal data is then computed for these 80 worlds. Across the 5 folds, we thus obtain 5 values of Pearson's correlation, of which we report the average.

For the natural conversation, 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 sec; sd: 6.73), chosen through random shuffling across the five folds until the standard deviation 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 auditory 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 contains the phoneme onset feature together with the baseline features, the surrogate model is created by randomly shuffling the 400 phoneme onset features across 400 words independently, while keeping the order of the baseline features (e.g. word onset and acoustic edges) constant. In this way, the baseline features are properly regressed to the neuronal data, while the target feature (e.g. phoneme onset) is randomly assigned to neuronal data.

For the natural speech, it was not possible to shuffle trials in the same way, as each auditory segment was of a different duration. This poses a problem because the mTRF features have to be of the same length as the neuronal data, which is not the case for natural speech, contrary to the auditory task where each word was standardized for the duration. Instead, we used the following method: for multivariable features, the surrogates were computed by randomly assigning each Dirac to a particular regressor (e.g., for the phoneme 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 has only one regressor, we would randomly split the trace into 2 parts and switch the order of the parts.

A model was considered to be statistically significant if the original model performed better compared to the 95th percentile of the surrogate distribution. The same statistical principle was used to determine the significant periods of the shape of the feature's kernel.

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 is defined as the average Euclidean distance between all pairs of cluster centroids. Intra-cluster distance is defined as the Euclidean distance of each cluster element to the corresponding centroid. By subtracting intra- from between-cluster distance, our index rewards cluster separability (between-cluster distance) and penalizes spatial dispersion (intra-cluster distance).

We performed clustering in the two-dimensional PC space (Results section 5), but systematically confirmed our results in the PC spaces using up to six dimensions (Supplementary Material).

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

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 it 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 (Supplementary Figure 11 c). 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 first ran K-means clustering 1000 times, as the clustering results slightly differ based 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, linguistically-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

Semantic features only have two regressors, 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 if it was higher than the 95th percentile of the surrogate distribution.

Comparison between the shapes of auditory task and natural conversation trajectories

To investigate whether the trajectories of the phoneme kernels projected to the low-dimensional space were similar between the auditory task and the natural conversation, we computed the canonical correlation between each kernel pair (e.g. kernel of phoneme /k/ extracted during the auditory task and the /k/ kernel from the natural conversation). 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 kernel in the natural conversation, rendering the test more difficult to pass. The correlation is considered significant if it is higher than the value of the 95th percentile of the surrogate correlation distribution.

Phoneme position kernels

To investigate whether there are any interactions between the encoding of phonetic and semantic features, we first created mTRF kernels for each phoneme position within the word and then aligned the resulting kernels with the lexical semantics kernel. Phoneme position was thus a multivariable with five regressors, each indicating the onset of the corresponding phoneme position across all words. For instance, phoneme position 2 indicates the onset of all 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 one for position three by 160 ms, etc. Surrogate kernel distributions were computed as described above. This allowed us to observe when significant peaks for each phoneme occurred with respect to the same time reference: word onset.

LFP low-beta and gamma 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 phoneme position as stimulus, and either beta (30-50 Hz) or low-gamma (15-25 Hz) LFP power as response. Word onset and phoneme position 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 phoneme-order low-gamma kernels with the same aligning procedure as described above.

Granger causality

To investigate the causality between the low-dimensional phonetic and semantic projections, we used the Multivariate Granger causality toolbox that is based on state-space Granger causal analysis (Barnett and Seth, 2014; Pesaran et al, 2018). 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 rate was then projected into the 3D 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 3 PCs of one feature (e.g. phonetic) from all 3 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 for semantic-to-phonetic predictions. Autoregression model parameters are estimated from data with the Levinson-Wiggins-Robinson algorithm and the Granger's F-test was used to assess the model's significance.

fMRI databases

Functional classification of the cortical surface surrounding the MEA with respect to different linguistic processes was performed using NeuroSynth (Yarkoni et al, 2011b) and NeuroQuery (Dockès et al, 2020) databases. They use 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 microelectrode 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Yarkoni T, Poldrack RA, Nichols TE, et al (2011a) Large-scale automated synthesis of human functional neuroimaging data. Nature Methods 8(8):665–670. https://doi.org/10.1038/nmeth.1635

Yarkoni T, Poldrack RA, Nichols TE, et al (2011b) NeuroSynth: A new platform for large-scale automated synthesis of human functional neuroimaging data. Frontiers in Neuroinformatics 5. https://doi.org/10.3389/conf.fninf.2011.08.00058

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