



**HAL**  
open science

## Neural manifolds carry reactivation of phonetic representations during semantic processing

Pavo Orepic, Wilson Truccolo, Eric Halgren, Sydney Cash, Anne-Lise Giraud Mamessier, Timothée Proix

► **To cite this version:**

Pavo Orepic, Wilson Truccolo, Eric Halgren, Sydney Cash, Anne-Lise Giraud Mamessier, et al.. Neural manifolds carry reactivation of phonetic representations during semantic processing. 2024. pasteur-04627390

**HAL Id: pasteur-04627390**

**<https://pasteur.hal.science/pasteur-04627390v1>**

Preprint submitted on 27 Jun 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

1 Neural manifolds carry reactivation of phonetic  
2 representations during semantic processing

3 Pavo Orepic<sup>1</sup>, Wilson Truccolo<sup>2,3</sup>, Eric Halgren<sup>4</sup>,  
4 Sydney S. Cash<sup>5</sup>, Anne-Lise Giraud<sup>1,6</sup>, Timothée Proix<sup>1\*</sup>

5 <sup>1</sup>Department of Basic Neurosciences, Faculty of Medicine, University of  
6 Geneva, Geneva, Switzerland.

7 <sup>2</sup>Department of Neuroscience, Brown University, Providence, Rhode  
8 Island, United States of America.

9 <sup>3</sup>Carney Institute for Brain Science, Brown University, Providence,  
10 Rhode Island, United States of America.

11 <sup>4</sup>Department of Neuroscience & Radiology, University of California San  
12 Diego, La Jolla, California, United States of America.

13 <sup>5</sup>Department of Neurology, Massachusetts General Hospital, Harvard  
14 Medical School, Boston, Massachusetts, United States of America.

15 <sup>6</sup>Institut Pasteur, Université Paris Cité, Hearing Institute, Paris, France.

16 \*Corresponding author(s). E-mail(s): [timothee.proix@unige.ch](mailto:timothee.proix@unige.ch);

17 **Abstract**

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

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

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

## 5 Introduction

6 How does the brain transform sounds into meanings? Most cognitive models of speech  
7 perception propose that speech sounds are processed sequentially by distinct neural  
8 modules. Semantic and conceptual representations emerge at the end of the ventral  
9 stream (i.e. the "what" stream) ([Hickok and Poeppel, 2007](#)), where perceived speech is  
10 sequentially transformed from spectro-temporal encoding of sounds in Heschl's gyrus,  
11 over phonetic features in the superior temporal gyrus (STG) and sulcus, to lexical  
12 and combinatorial semantics in the anterior temporal lobe (ATL) ([Mesgarani et al,](#)  
13 [2014](#); [Pylkkänen, 2020](#)). For instance, the ATL is considered a specialized semantic  
14 module, connected to modality-specific sources of information ([Patterson et al, 2007](#);  
15 [Ralph et al, 2017](#)), with causal semantic impairments occurring after bilateral ATL  
16 atrophy or damage due to viral infection ([Noppeney et al, 2006](#); [Lambon Ralph et al,](#)  
17 [2006](#); [Schwartz et al, 2009](#); [Cope et al, 2020](#)). This modular and sequential perspective  
18 implies that a complex neural process transforms speech features from one functional  
19 brain region to the next up to lexical and multimodal conceptual representations.

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

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

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

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

1 in the power of different local field potential (LFP) frequency bands: bottom-up pro-  
2 cesses have been associated with the low-gamma and theta bands, and top-down ones  
3 with the beta band ([Arnal and Giraud, 2012](#); [Fontolan et al, 2014](#); [van Kerkoerle et al,](#)  
4 [2014](#); [Michalareas et al, 2016](#); [Chao et al, 2018](#); [Giraud and Arnal, 2018](#)).

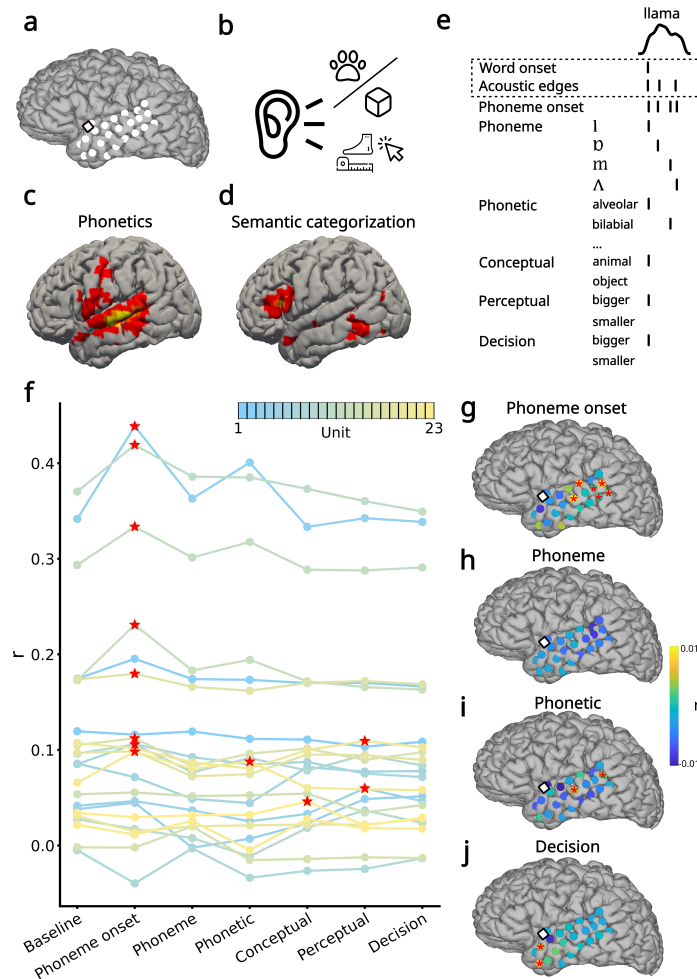
5 To identify the fine-grained mechanisms underlying speech neural transformations,  
6 it might be necessary to investigate speech processing at a smaller spatial scale giving  
7 access to the neuronal spiking level. Recent findings suggest that complex cognitive  
8 processes and behavioral features are encoded in low-dimensional neuronal spaces, also  
9 called *neural manifolds* ([Pillai and Jirsa, 2017](#); [Gallego et al, 2017](#); [Jazayeri and Ostojic,](#)  
10 [2021](#); [Vyas et al, 2020](#); [Chung and Abbott, 2021](#); [Truccolo, 2016](#); [Aghagolzadeh](#)  
11 [and Truccolo, 2016](#)). In the primate cortex, the dynamics on the neural manifolds  
12 characterize functionally distinct behaviors and conditions, such as sensorimotor com-  
13 putations, decision making, or working memory ([Mante et al, 2013](#); [Remington et al,](#)  
14 [2018](#); [Markowitz et al, 2015](#)). In humans, this endeavor is largely hindered by the  
15 invasiveness of single-neuron recordings, which are rarely performed. Although speech  
16 encoding has never been characterized at the level of collective dynamics of action  
17 potentials recorded from neuronal ensembles, it seems plausible that different aspects  
18 of speech, including transformations across the speech hierarchy, are encoded in such  
19 low-dimensional manifold representations. The condition- and history-dependent orga-  
20 nization of neuronal trajectories on the manifold could serve the purpose of integrating  
21 phonetic features into higher-level representations ([Pulvermüller, 2018](#); [Yi et al, 2019](#)).  
22 The handful of studies that have so far reported single unit activity associated with  
23 speech processing have highlighted the tuning of one or a few single units to phonetic  
24 features ([Chan et al, 2014, 2011](#); [Ossmy et al, 2015](#); [Lakretz et al, 2021](#)). By contrast,  
25 we focused here on the collective dynamics of tens of neurons in the light of the neu-  
26 ral manifold framework to address the mechanisms underlying phonetic-to-semantic  
27 transformations.

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

## 19 Results

### 20 Semantic and phonetic neuronal encodings in the aSTG

21 Microelectrode array and ECoG signals were recorded in a 31-year-old patient with  
22 pharmacologically resistant epilepsy who was implanted for clinical purposes (Fig.  
23 1a). The 10x10 MEA was located in the left anterior superior temporal gyrus (aSTG)  
24 (square on Fig. 1a). 23 ECoG electrodes of interest covered a large portion of the  
25 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.

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

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

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

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



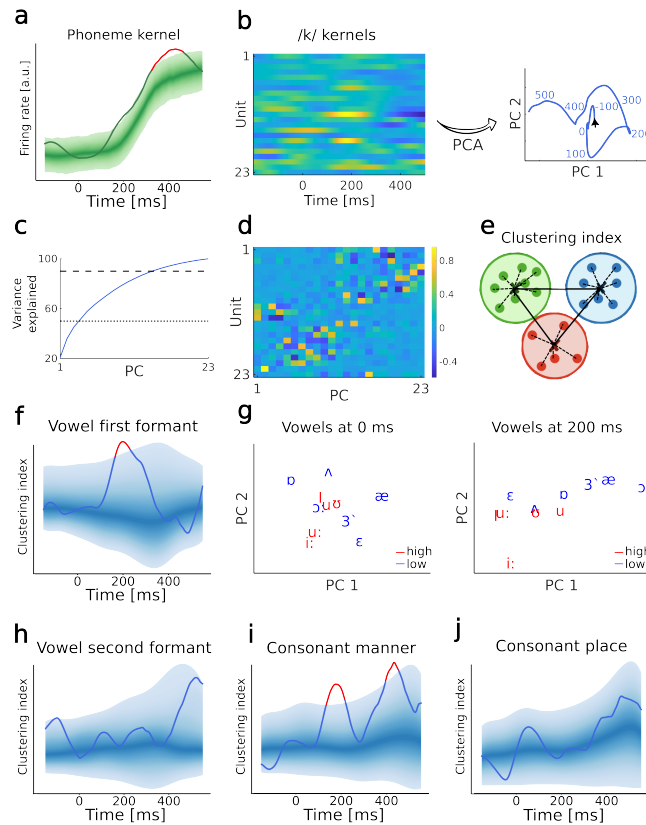
1 We fitted the mTRF encoding models to the soft-normalized firing rate of single  
2 units recorded with the MEA (Methods). We considered the 23 spike-sorted single  
3 units (over a total of 176) with an average firing rate higher than 0.3 spikes/second  
4 (mean: 1.6, sd: 1.39, range 0.43 - 5.27 spikes/second,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

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

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

## 24 **Distributed phonetic encoding**

25 We showed that only a few single units, when taken independently, encoded phonetic  
26 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.

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

8 To address this hypothesis, we performed principal component analysis (PCA)  
9 on all concatenated phoneme kernels obtained with mTRF (Fig. 2b). The first four  
10 principal components (PCs) accounted for about 50% of phoneme feature variance (Fig  
11 2c), and were distributed across different units (Fig 2d). Having half of the variance  
12 explained by the low-dimensional and correlated activity of several units indicates that  
13 phonemes are dynamically represented within a manifold.

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

23 We observed significant clustering within the vowel first formant group between  
24 the two phonetic features that reflect the high and low position of the tongue (Fig.  
25 2f). The clustering specifically occurred at about 200 ms, which was apparent in the  
26 position of the vowels in the two-dimensional PC space (Fig. 2g). We replicated these  
27 findings using higher dimensional PC spaces (Supplementary Fig. 3). Contrary to

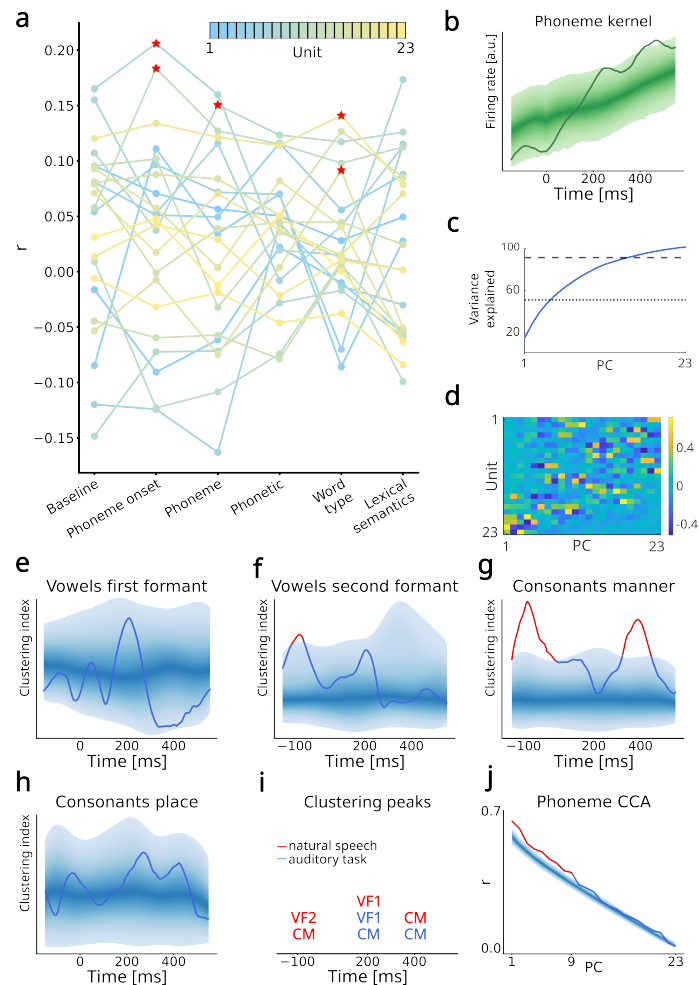
1 the vowel first formant group, we observed no significant clustering between phonetic  
2 features of the vowel second formant group (front and back position of the tongue,  
3 Fig. 2h, Supplementary Fig. 3 and 4). These findings indicate selective clustering of  
4 phoneme kernels based on vowel first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

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

13 We replicated the observed clustering by vowels first formant and consonant man-  
14 ner of articulation, as well as the lack of clustering by vowel second formant and  
15 consonant place of articulation with additional control analyses. Linear discriminant  
16 analysis (LDA) classifier revealed separability of these two phonetic categories in the  
17 same time windows (i.e. at about 200 ms for vowels first formant and both at about  
18 200 and 400 ms for consonants manner of articulation) (Supplementary Fig. 5). Simi-  
19 larly, a rank-regression approach, where the ranks indicate the first or second formant  
20 value, showed significant ordering of the vowels along the first formant at 200 ms, and  
21 no ordering along the second formant (Supplementary Fig. 6). (This analysis cannot  
22 be performed for consonants, where no ranking is possible across the different pho-  
23 netic groups). Finally, we observed similar organizational patterns through k-means  
24 clustering, an unsupervised data-driven approach (Supplementary Fig. 7).

25 To summarize, we observed a significant clustering of phoneme trajectories on a  
26 low-dimensional neural manifold: vowel trajectories clustered along first formants at  
27 200 ms, while consonant trajectories clustered by manner of articulation at 200 and  
28 400 ms.

## 1 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.

1 To investigate whether the results generalized to natural speech perception, we  
2 analyzed neuronal recordings from the same intracortical MEA during a spontaneous  
3 conversation between the participant and another person recorded in a separate exper-  
4 imental session. As for the semantic task, we performed spike sorting and selected the  
5 23 units with the highest spiking rate (mean: 0.58 spikes/second, sd: 0.89, range 0.1  
6 - 3.39) out of 212 clustered single units. We identified 664 words (272 unique) pro-  
7 nounced by the other person in the recording. From those words, we segmented the  
8 same 32 phonemes as in the task. We designed two high-level features, namely word  
9 class (e.g. noun, adjective, conjunction, etc), and lexical semantics, computed as the  
10 Lancaster sensorimotor norms of each word (Lynott et al, 2019) (see Supplementary  
11 Tables 4 and 5 for an overview). We then fitted the mTRF encoding models as in the  
12 semantic task dataset, and likewise observed that only a few units significantly cor-  
13 related with speech features (Fig. 3a), and that the average phoneme kernel became  
14 significant compared to its surrogate distribution at about 200 and 400 ms, suggesting  
15 again a population effect (Fig. 3b).

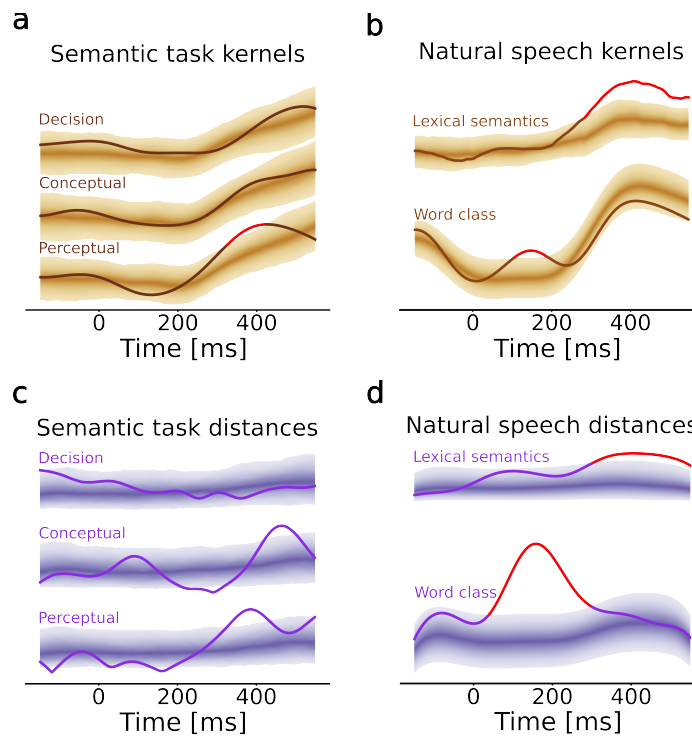
16 We thus proceeded as before by performing PCA on the kernels. Similar to the  
17 task, 50% of the variance was explained by a few PCs (Fig. 3c), and the processing  
18 of phonemes was distributed across units (Fig. 3d). We further computed the clus-  
19 tering index for phonetic features in the low-dimensional neural manifold spanned by  
20 the first two PCs. Remarkably, the clustering index profiles were similar to those in  
21 the semantic task. For the vowel first formants, we observed a peak of the clustering  
22 index at about 200 ms (Fig. 3e). Although this peak was not significant for the two-  
23 dimensional space, it was present for all dimensions, and reached significance for the  
24 one-dimensional space (Supplementary Fig. 8). This peak was also significant in all  
25 control analyses (LDA classifier, rank regression, and k-means, Supplementary Fig.  
26 9-11). For vowel second formants, a trend consistent across dimensions was observed  
27 at 200 ms (Fig. 3f). The second formant phonetic features also clustered before the

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

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

## 21 **Encoding of semantic features**

22 Considering that the MEA is implanted in a cortical area involved in auditory semantic  
23 processing (Ralph et al, 2017), we repeated the same analyses for the semantic kernels,  
24 both for the semantic task and natural speech perception. We first considered the  
25 time course of the semantic kernels averaged across all units, and compared it to the  
26 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.**

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

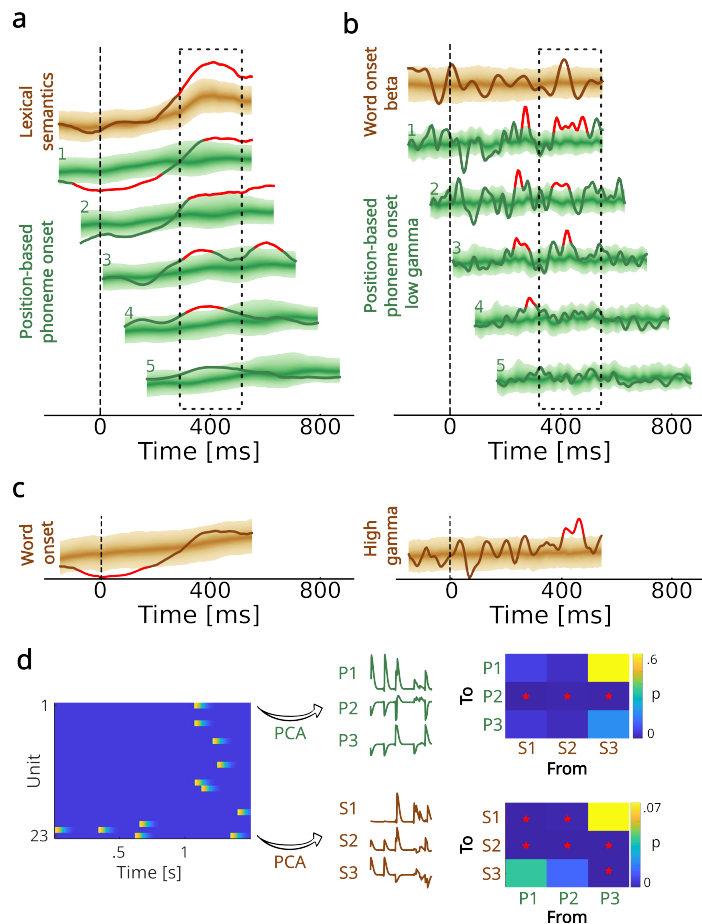
6 We then turned to the neuronal population dynamics. We projected the three  
7 semantic kernels of the semantic task to the low-dimensional neural manifold. Because  
8 there are only two categories for each of the probed semantic features, we could not



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

## 15 **Concurrent encoding of bottom-up phonetic and top-down** 16 **semantic features**

17 To further investigate the mechanisms underlying the integration of phonetic and  
18 semantic features, we explored the timing of their representations in more detail.  
19 Inspired by the analysis-by-synthesis framework, we hypothesized that an early  
20 bottom-up phonetic processing enables a later top-down word-level semantic guess,  
21 which is compared to actual phonetic features (Halle and Stevens, 1959; Bever and  
22 Poeppel, 2010). For the late semantic-phonetic comparison to occur, phonetic fea-  
23 tures should be (re)encoded concurrently with semantic processing at about 400 ms  
24 after word onset. To challenge this hypothesis, we created additional mTRF features  
25 that regressed phoneme onsets at different positions within each word (i.e., phoneme  
26 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).

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

10 To further test the compatibility with the analysis-by-synthesis framework, we  
11 asked whether the encoding reflected top-down predictive or bottom-up cumulative  
12 processes during natural speech perception. The occurrence of top-down processes is  
13 typically accompanied by beta peaks in the LFP power, while bottom-up processes  
14 are reflected in LFP low-gamma power peaks (Arnal and Giraud, 2012; Fontolan et al,  
15 2014). To identify the timing of top-down and bottom-up processes, we thus extracted  
16 and averaged the LFP power across the MEA contacts for both the beta (12-30 Hz)  
17 and low-gamma (30-70 Hz) bands. We then fitted an mTRF model to these two vari-  
18 ables using word onset features for the beta band (reflecting word-level top-down  
19 processing) and position-based phoneme onsets for the low-gamma band (reflecting  
20 position-dependent bottom-up processing). For the beta band, we found two signifi-  
21 cant positive peaks: at word onset and 400 ms after word onset (Fig. 5b, brown trace).  
22 In the low-gamma band, we aligned position-based phoneme onset kernels based on  
23 word onset as before, and observed negative and positive significant peaks before  
24 400 ms in the low-gamma band that were not accompanied by a beta peak, sug-  
25 gesting early bottom-up processes. We also found an alignment of positive significant  
26 peaks at 400 ms after word onset (Fig. 5b, green traces), showing that both top-  
27 down semantic and bottom-up phonetic processing occurred at 400 ms. Following the

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

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

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

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

## 1 Discussion

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

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

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

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

1 of bottom-up phoneme representations highlighted by concurrent low-gamma peaks.  
2 This top-down semantic guess could result from a cumulative build-up of bottom-up  
3 phonetic processes reflected by low-gamma peaks occurring before the 400-ms align-  
4 ment, in agreement with most models of spoken-word recognition (Marslen-Wilson,  
5 1987; Norris and McQueen, 2008). The specific mechanisms underlying the emergence  
6 of such phonemic-to-semantic interface remain to be uncovered. For example, distinct  
7 phonetic features could be represented either sequentially (Hickok and Poeppel, 2007;  
8 Dehaene et al, 2015) or persistently (Perdikis et al, 2011; Yi et al, 2019; Martin, 2020)  
9 with more resolved temporal mechanisms (Fontolan et al, 2014; Hovsepyan et al, 2023)  
10 before being pooled together in a semantic representation.

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

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

4 Using the natural speech dataset, we contrasted neuronal signatures underlying  
5 semantic and syntactic encoding. Syntactic and semantic encoding occurred at about  
6 150 - 200 ms and 400 ms respectively in their corresponding neural manifolds. The  
7 early syntactic encoding might reflect a fast combinatorial process, as reported with  
8 MEG ([Pylkkänen, 2020](#)) and/or early effects contributing to conceptual and percep-  
9 tual categorization in the ATL at about 200 ms ([Borghesani et al, 2019](#); [Chan et al,](#)  
10 [2011](#); [Chen et al, 2016](#); [Dehaene, 1995](#); [Hinojosa et al, 2001](#)). The semantic categoriza-  
11 tion effect about 400 ms is reminiscent of the N400 component frequently reported in  
12 the ATL, for both semantic composition and lexical decision ([Kutas and Federmeier,](#)  
13 [2011](#); [Dikker et al, 2020](#); [López Zunini et al, 2020](#); [Bentin et al, 1985](#); [Barber et al,](#)  
14 [2013](#); [Vignali et al, 2023](#); [Rahimi et al, 2022](#); [Lau et al, 2008](#); [Kutas and Federmeier,](#)  
15 [2000](#)).

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



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

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

## 12 **Methods**

### 13 **Participant**

14 The study participant, a male in his thirties with pharmacologically resistant epilepsy,  
15 underwent intracranial electrode implantation as part of his clinical epilepsy treat-  
16 ment. He was a native English speaker with normal sensory and cognitive functions and  
17 demonstrated left-hemisphere language dominance through a WADA test. The patient  
18 experienced partial complex seizures originating from mesial temporal region electrode  
19 contacts. The surgical intervention involved the removal of the left anterior temporal  
20 lobe, along with the microelectrode implantation site, left parahippocampal gyrus, left  
21 hippocampus, and left amygdala. The patient achieved seizure-free status one year  
22 post-surgery, with no significant changes in language functions observed in formal neu-  
23ropsychological testing conducted at that time. Informed consent was obtained, and  
24 the study was conducted under the oversight of the Massachusetts General Hospital  
25 Institutional Review Board (IRB). The study included both the intracortical implan-  
26 tation of a microelectrode array (MEA) and the performance of a semantic task and

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

### 3 **Neural recordings**

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

11 Single-unit action potentials were recorded with a 10-by-10, 400  $\mu\text{m}$  electrode dis-  
12 tance, microelectrode array (Utah array, Blackrock Neurotech) surgically implanted  
13 within the left anterior superior temporal gyrus (aSTG). Electrodes were 1.5 mm long  
14 and contained a 20- $\mu\text{m}$  platinum tip. The implantation site was excised and the subse-  
15 quent histological analysis revealed the spatial orientation of the electrode tips within  
16 the depths of cortical layer III, proximal to layer IV, with no notable histological  
17 abnormalities in the neighboring cortical environment. Data acquisition was acquired  
18 with a Blackrock NeuroPort system, with a sampling frequency of 30 kilosamples per  
19 second, and an analog bandpass filter ranging from 0.3 Hz to 7.5 kHz for antialiasing.

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

### 23 **Auditory stimuli**

24 Neural data was recorded during two separate experimental sessions, that took place  
25 on the same day, two hours apart ([Chan et al, 2014](#)). In the first session, the partici-  
26 pant performed an auditory semantic categorization task. The stimuli were standalone

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

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

## 19 **Signal preprocessing**

20 Spike detection and sorting were performed with the semi-automatic wave.clus algo-  
21 rithm (Quiroga et al, 2004). Across 96 active electrodes, we identified 176 and 212  
22 distinct units for the sessions with the semantic task and the natural speech, respec-  
23 tively. For the semantic task, we considered units with firing rate higher than 0.3  
24 spikes/second, resulting in 23 units (0.43 – 5.27 spikes/s). For a fair comparison across  
25 both sessions, we also selected the 23 most spiking units in the natural speech (0.1 –  
26 3.39 spikes/s).

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

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

## 15 **Phoneme segmentation and categorization**

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

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

1 back), consonants articulation place (bilabial, labiodental, alveolar, velar, uvular, glot-  
2 tal), consonants articulation manner (plosive, nasal, fricative, approximant, lateral  
3 approximant).

#### 4 **mTRF features**

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

12 Word onset was marked by a value of one located at the onset of each word.  
13 Acoustic edges were defined as local maxima in the derivative of the speech envelope  
14 ([Oganian and Chang, 2019](#)). The speech envelope was computed as the logarithmi-  
15 cally scaled root mean square of the audio signal using the MATLAB `mTRFenvelope`  
16 function. Phoneme onset feature indicates onsets of all phonemes in a word. Phoneme  
17 identity feature was multivariable with 32 regressors, each indicating onsets of a differ-  
18 ent phoneme, as defined by the IPA table. Phonetic category feature was multivariable,  
19 and included four phonetic groups (vowel first and second formant, consonant man-  
20 ner, and place of articulations). All other features were multivariables with a value  
21 located at the corresponding word onset. Perceptual category, conceptual category,  
22 and semantic decision had two regressors, defined respectively as bigger and smaller,  
23 animal and object, and decision on whether the object/animal was bigger or smaller  
24 than a foot. Word class had 13 regressors, indicating different word classes (noun,  
25 verb, adjective, adverb, article, auxiliary, demonstrative, quantifier, preposition, pro-  
26 noun, conjunction, interjection, number). Finally, the lexical semantics feature was

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

## 5 **mTRF estimation**

6 mTRFs were estimated using the mTRF MATLAB toolbox (Crosse et al, 2016). All  
7 mTRFs were always of encoding type, relating the stimulus features to neural data,  
8 with resulting kernels in the time range between -200 and 600 ms. The first and last  
9 50 ms were not considered in the analysis, to avoid possible edge effects. Both for  
10 semantic task and natural speech, the baseline model contained the word onset and  
11 the acoustic onset edge features. All other models included the baseline features and  
12 one of the additional target features defined above. Estimation was performed by a  
13 ridge regression, using a nested cross-validation procedure (see below). The goodness  
14 of fit was defined as Pearson's correlation between model prediction and neural data.

## 15 **Cross-validation**

16 For the semantic task, we performed a nested cross-validation. In the outer cross-  
17 validation loop, we split the 400 words randomly into 5 sets of 80 (20% of the words  
18 each). Thus, in each fold, 80 trials belonged to a hold-out test set, while the remaining  
19 320 words belonged to the train set. The five folds were identical across all models. For  
20 a given outer fold, among the 320 train-set trials, we performed another 8-fold inner  
21 cross-validation loop for the ridge regression hyperparameter tuning, with  $\lambda$  ranging  
22 from  $10^{-6}$  to  $10^6$ . The optimal lambda was then used to retrain the model on the 320  
23 words of the training set of the outer cross-validation loop fold. The model predictions  
24 and Pearson's correlation with the neural data were then computed for the 80 words  
25 of the test set. Across the 5 folds, we thus obtained five correlation values, of which  
26 we report the average.

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

## 7 **Surrogate distributions and statistical significance**

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

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

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

### 3 **Averaged kernels**

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

### 8 **Correcting for multiple comparisons**

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

### 17 **Clustering of phoneme kernels in the PC space**

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



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

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

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

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

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

#### 4 **Separability of semantic feature kernels in the PC space**

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

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

#### 18 **Comparison of trajectories between the semantic task and** 19 **natural speech**

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

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

#### 4 **Position-based phoneme onset kernels**

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

#### 20 **LFP beta, low-gamma, and BHA power analysis**

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

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

## 7 **Granger causality**

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

## 24 **fMRI databases**

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

1 2020). It uses text mining and meta-analysis techniques to automatically produce  
2 large-scale mappings between fMRI brain activity and a cognitive process of interest.  
3 We were primarily interested in observing the proximity of phonetic and semantic  
4 processing close to the MEA implantation site. As keywords, we used "phonetics" and  
5 "semantic categorization".

6 **Acknowledgments.** This work was funded by Swiss National Science Founda-  
7 tion career grant 193542 (T.P.), EU FET-BrainCom project (A.G.), NCCR Evolving  
8 Language, Swiss National Science Foundation Agreement #51NF40\_180888 (A.G.),  
9 Swiss National Science Foundation project grant 163040 (A.G.), National Institutes  
10 of Health (NIH), National Institute of Neurological Disorders and Stroke (NINDS),  
11 grants R01NS079533 (WT) and R01NS062092 (SSC), and the Pablo J. Salame Gold-  
12 man Sachs endowed Associate Professorship of Computational Neuroscience at Brown  
13 University (WT).

## 1   **References**

- 2   Aghagolzadeh M, Truccolo W (2016) Inference and Decoding of Motor Cortex Low-  
3   Dimensional Dynamics via Latent State-Space Models. *IEEE Transactions on*  
4   *Neural Systems and Rehabilitation Engineering* 24(2):272–282. [https://doi.org/10.](https://doi.org/10.1109/TNSRE.2015.2470527)  
5    [1109/TNSRE.2015.2470527](https://doi.org/10.1109/TNSRE.2015.2470527)
- 6   Arnal LH, Giraud AL (2012) Cortical oscillations and sensory predictions. *Trends in*  
7   *Cognitive Sciences* 16(7):390–398. <https://doi.org/10.1016/j.tics.2012.05.003>
- 8   Barber HA, Otten LJ, Kousta ST, et al (2013) Concreteness in word processing: ERP  
9   and behavioral effects in a lexical decision task. *Brain and Language* 125(1):47–53.  
10   <https://doi.org/10.1016/j.bandl.2013.01.005>
- 11   Barnett L, Seth AK (2014) The MVGC multivariate Granger causality toolbox: A new  
12   approach to Granger-causal inference. *Journal of Neuroscience Methods* 223:50–68.  
13   <https://doi.org/10.1016/j.jneumeth.2013.10.018>
- 14   Barnett L, Seth AK (2015) Granger causality for state-space models. *Physical Review*  
15    E 91(4):040101. <https://doi.org/10.1103/PhysRevE.91.040101>
- 16   Bentin S, McCARTHY GREGORY, Wood CC (1985) Event-related potentials, lexical  
17   decision and semantic priming. *Electroencephalography and clinical Neurophysiol-*  
18    *ogy* 60:343–355
- 19   Bever TG, Poeppel D (2010) Analysis by Synthesis: A (Re-)Emerging Program of  
20   Research for Language and Vision. *Biolinguistics* 4(2-3):174–200. [https://doi.org/](https://doi.org/10.5964/bioling.8783)  
21    [10.5964/bioling.8783](https://doi.org/10.5964/bioling.8783)
- 22   Boersma P (2001) Praat, a system for doing phonetics by computer. *Glott Int* 5(9):341–  
23    345

- 1 Borghesani V, Buiatti M, Eger E, et al (2019) Conceptual and Perceptual Dimensions  
2 of Word Meaning Are Recovered Rapidly and in Parallel during Reading. *Journal*  
3 *of Cognitive Neuroscience* 31(1):95–108. [https://doi.org/10.1162/jocn\\_a.01328](https://doi.org/10.1162/jocn_a.01328)
- 4 Caucheteux C, Gramfort A, King JR (2022) Deep language algorithms predict seman-  
5 tic comprehension from brain activity. *Scientific Reports* 12(1):16327. [https://doi.](https://doi.org/10.1038/s41598-022-20460-9)  
6 [org/10.1038/s41598-022-20460-9](https://doi.org/10.1038/s41598-022-20460-9)
- 7 Chan AM, Baker JM, Eskandar E, et al (2011) First-Pass Selectivity for Seman-  
8 tic Categories in Human Anteroventral Temporal Lobe. *Journal of Neuroscience*  
9 31(49):18119–18129. <https://doi.org/10.1523/JNEUROSCI.3122-11.2011>
- 10 Chan AM, Dykstra AR, Jayaram V, et al (2014) Speech-Specific Tuning of Neurons  
11 in Human Superior Temporal Gyrus. *Cerebral Cortex* 24(10):2679–2693. [https://](https://doi.org/10.1093/cercor/bht127)  
12 [doi.org/10.1093/cercor/bht127](https://doi.org/10.1093/cercor/bht127)
- 13 Chang EF, Raygor KP, Berger MS (2015) Contemporary model of language orga-  
14 nization: An overview for neurosurgeons. *Journal of Neurosurgery* 122(2):250–261.  
15 <https://doi.org/10.3171/2014.10.JNS132647>
- 16 Chao ZC, Takaura K, Wang L, et al (2018) Large-Scale Cortical Networks for  
17 Hierarchical Prediction and Prediction Error in the Primate Brain. *Neuron*  
18 100(5):1252–1266.e3. <https://doi.org/10.1016/j.neuron.2018.10.004>
- 19 Chen Y, Shimotake A, Matsumoto R, et al (2016) The ‘when’ and ‘where’ of semantic  
20 coding in the anterior temporal lobe: Temporal representational similarity analysis  
21 of electrocorticogram data. *Cortex* 79:1–13. [https://doi.org/10.1016/j.cortex.2016.](https://doi.org/10.1016/j.cortex.2016.02.015)  
22 [02.015](https://doi.org/10.1016/j.cortex.2016.02.015)
- 23 Chung S, Abbott L (2021) Neural population geometry: An approach for under-  
24 standing biological and artificial neural networks. *Current Opinion in Neurobiology*

- 1     70:137–144. <https://doi.org/10.1016/j.conb.2021.10.010>
- 2     Churchland MM, Cunningham JP, Kaufman MT, et al (2012) Neural popula-  
3     tion dynamics during reaching. *Nature* 487(7405):51–56. [https://doi.org/10.1038/](https://doi.org/10.1038/nature11129)  
4     [nature11129](https://doi.org/10.1038/nature11129)
- 5     Cope TE, Shtyrov Y, MacGregor LJ, et al (2020) Anterior temporal lobe is neces-  
6     sary for efficient lateralised processing of spoken word identity. *Cortex* 126:107–118.  
7     <https://doi.org/10.1016/j.cortex.2019.12.025>
- 8     Crone NE, Boatman D, Gordon B, et al (2001) Induced electrocorticographic gamma  
9     activity during auditory perception. *Clinical Neurophysiology* 112(4):565–582. [https://doi.org/10.1016/S1388-2457\(00\)00545-9](https://doi.org/10.1016/S1388-2457(00)00545-9)
- 10
- 11    Crosse MJ, Di Liberto GM, Bednar A, et al (2016) The multivariate temporal  
12    response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals  
13    to continuous stimuli. *Frontiers in human neuroscience* 10:604
- 14    Damera SR, Chang L, Nikolov PP, et al (2023) Evidence for a Spoken Word Lexicon  
15    in the Auditory Ventral Stream. *Neurobiology of Language* 4(3):420–434. <https://doi.org/10.1162/nol.a.00108>
- 16
- 17    Dehaene S (1995) Electrophysiological evidence for category-specific word processing  
18    in the normal human brain. *NeuroReport* 6(16):2153
- 19    Dehaene S, Meyniel F, Wacongne C, et al (2015) The Neural Representation of  
20    Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees.  
21    *Neuron* 88(1):2–19. <https://doi.org/10.1016/j.neuron.2015.09.019>
- 22    DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. *Trends in*  
23    *Cognitive Sciences* 11(8):333–341. <https://doi.org/10.1016/j.tics.2007.06.010>



- 1 Dikker S, Assaneo MF, Gwilliams L, et al (2020) Magnetoencephalography and Lan-  
2 guage. *Neuroimaging Clinics of North America* 30(2):229–238. [https://doi.org/10.](https://doi.org/10.1016/j.nic.2020.01.004)  
3 [1016/j.nic.2020.01.004](https://doi.org/10.1016/j.nic.2020.01.004)
- 4 Dockès J, Poldrack RA, Primet R, et al (2020) NeuroQuery, comprehensive meta-  
5 analysis of human brain mapping. *eLife* 9:e53385. [https://doi.org/10.7554/eLife.](https://doi.org/10.7554/eLife.53385)  
6 [53385](https://doi.org/10.7554/eLife.53385)
- 7 Dykstra AR, Chan AM, Quinn BT, et al (2012) Individualized localization and cortical  
8 surface-based registration of intracranial electrodes. *NeuroImage* 59(4):3563–3570.  
9 <https://doi.org/10.1016/j.neuroimage.2011.11.046>
- 10 Fontolan L, Morillon B, Liegeois-Chauvel C, et al (2014) The contribution of  
11 frequency-specific activity to hierarchical information processing in the human audi-  
12 tory cortex. *Nature Communications* 5:4694. <https://doi.org/10.1038/ncomms5694>
- 13 Friederici AD, Kotz SA (2003) The brain basis of syntactic processes: Functional  
14 imaging and lesion studies. *NeuroImage* 20:S8–S17. [https://doi.org/10.1016/j.](https://doi.org/10.1016/j.neuroimage.2003.09.003)  
15 [neuroimage.2003.09.003](https://doi.org/10.1016/j.neuroimage.2003.09.003)
- 16 Gallego JA, Perich MG, Miller LE, et al (2017) Neural Manifolds for the Control of  
17 Movement. *Neuron* 94(5):978–984. <https://doi.org/10.1016/j.neuron.2017.05.025>
- 18 Giraud AL, Arnal LH (2018) Hierarchical Predictive Information Is Channeled by  
19 Asymmetric Oscillatory Activity. *Neuron* 100(5):1022–1024. [https://doi.org/10.](https://doi.org/10.1016/j.neuron.2018.11.020)  
20 [1016/j.neuron.2018.11.020](https://doi.org/10.1016/j.neuron.2018.11.020)
- 21 Gwilliams L, King JR (2020) Recurrent processes support a cascade of hierarchical  
22 decisions. *eLife* 9:1–20. <https://doi.org/10.7554/eLife.56603>

- 1 Halle M, Stevens K (1959) Analysis by synthesis. In: Proceeding of the Seminar on  
2 Speech Compression and Processing, Paper D7, vol II. W. Wathen-Dunn & L. E.  
3 Woods (eds.)
- 4 Hamilton LS, Oganian Y, Hall J, et al (2021) Parallel and distributed encoding of  
5 speech across human auditory cortex. *Cell* 184(18):4626–4639.e13. [https://doi.org/  
6 10.1016/j.cell.2021.07.019](https://doi.org/10.1016/j.cell.2021.07.019)
- 7 Heilbron M, Armeni K, Schoffelen JM, et al (2022) A hierarchy of linguistic predictions  
8 during natural language comprehension. *Proceedings of the National Academy of  
9 Sciences* 119(32):e2201968119. <https://doi.org/10.1073/pnas.2201968119>
- 10 Hickok G, Poeppel D (2007) The cortical organization of speech processing. *Nature  
11 Reviews Neuroscience* 8(5):393–402. <https://doi.org/10.1038/nrn2113>
- 12 Hinojosa JA, Martín-Loeches M, Muñoz F, et al (2001) Electrophysiological evidence  
13 of a semantic system commonly accessed by animals and tools categories. *Cognitive  
14 Brain Research* 12(2):321–328. [https://doi.org/10.1016/S0926-6410\(01\)00039-8](https://doi.org/10.1016/S0926-6410(01)00039-8)
- 15 Hovsepyan S, Olasagasti I, Giraud AL (2023) Rhythmic modulation of prediction  
16 errors: A top-down gating role for the beta-range in speech processing. *PLOS Com-  
17 putational Biology* 19(11):e1011595. <https://doi.org/10.1371/journal.pcbi.1011595>
- 18 Jazayeri M, Ostojic S (2021) Interpreting neural computations by examining intrinsic  
19 and embedding dimensionality of neural activity. *Current Opinion in Neurobiology*  
20 70:113–120. <https://doi.org/10.1016/j.conb.2021.08.002>
- 21 Keshishian M, Akkol S, Herrero J, et al (2023) Joint, distributed and hierarchically  
22 organized encoding of linguistic features in the human auditory cortex. *Nature  
23 Human Behaviour* 7(5):740–753. <https://doi.org/10.1038/s41562-023-01520-0>

- 1 Kisler T, Reichel U, Schiel F (2017) Multilingual processing of speech via web services.  
2 Computer Speech & Language 45:326–347. [https://doi.org/10.1016/j.csl.2017.01.](https://doi.org/10.1016/j.csl.2017.01.005)  
3 [005](https://doi.org/10.1016/j.csl.2017.01.005)
- 4 Kutas M, Federmeier KD (2000) Electrophysiology reveals semantic memory use in  
5 language comprehension. Trends in Cognitive Sciences 4(12):463–470. [https://doi.](https://doi.org/10.1016/S1364-6613(00)01560-6)  
6 [org/10.1016/S1364-6613\(00\)01560-6](https://doi.org/10.1016/S1364-6613(00)01560-6)
- 7 Kutas M, Federmeier KD (2011) Thirty Years and Counting: Finding Meaning in the  
8 N400 Component of the Event-Related Brain Potential (ERP). Annual Review of  
9 Psychology 62(1):621–647. <https://doi.org/10.1146/annurev.psych.093008.131123>
- 10 Lakretz Y, Ossmy O, Friedmann N, et al (2021) Single-cell activity in human STG  
11 during perception of phonemes is organized according to manner of articulation.  
12 NeuroImage 226:117499. <https://doi.org/10.1016/j.neuroimage.2020.117499>
- 13 Lambon Ralph MA, Lowe C, Rogers TT (2006) Neural basis of category-specific  
14 semantic deficits for living things: Evidence from semantic dementia, HSVE and  
15 a neural network model. Brain 130(4):1127–1137. [https://doi.org/10.1093/brain/](https://doi.org/10.1093/brain/awm025)  
16 [awm025](https://doi.org/10.1093/brain/awm025)
- 17 Lau EF, Phillips C, Poeppel D (2008) A cortical network for semantics:  
18 (de)constructing the N400. Nature Reviews Neuroscience 9(12):920–933. [https:](https://doi.org/10.1038/nrn2532)  
19 [//doi.org/10.1038/nrn2532](https://doi.org/10.1038/nrn2532)
- 20 Leszczyński M, Barczak A, Kajikawa Y, et al (2020) Dissociation of broadband  
21 high-frequency activity and neuronal firing in the neocortex. Science Advances  
22 6(33):eabb0977. <https://doi.org/10.1126/sciadv.abb0977>

- 1 López Zunini RA, Baart M, Samuel AG, et al (2020) Lexical access versus lexical  
2 decision processes for auditory, visual, and audiovisual items: Insights from behav-  
3 ioral and neural measures. *Neuropsychologia* 137:107305. [https://doi.org/10.1016/  
4 j.neuropsychologia.2019.107305](https://doi.org/10.1016/j.neuropsychologia.2019.107305)
- 5 Lynott D, Connell L, Brysbaert M, et al (2019) The Lancaster Sensorimotor Norms:  
6 Multidimensional measures of Perceptual and Action Strength for 40,000 English  
7 words. <https://doi.org/10.31234/osf.io/ktjwp>
- 8 Mante V, Sussillo D, Shenoy KV, et al (2013) Context-dependent computation by  
9 recurrent dynamics in prefrontal cortex. *Nature* 503(7474):78–84. [https://doi.org/  
10 10.1038/nature12742](https://doi.org/10.1038/nature12742)
- 11 Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-  
12 data. *Journal of Neuroscience Methods* 164(1):177–190. [https://doi.org/10.1016/j.  
13 jneumeth.2007.03.024](https://doi.org/10.1016/j.jneumeth.2007.03.024)
- 14 Markowitz DA, Curtis CE, Pesaran B (2015) Multiple component networks support  
15 working memory in prefrontal cortex. *Proceedings of the National Academy of  
16 Sciences* 112(35):11084–11089. <https://doi.org/10.1073/pnas.1504172112>
- 17 Marslen-Wilson WD (1987) Functional parallelism in spoken word-recognition. *Cog-  
18 nition* 25(1-2):71–102. [https://doi.org/10.1016/0010-0277\(87\)90005-9](https://doi.org/10.1016/0010-0277(87)90005-9)
- 19 Martin AE (2020) A Compositional Neural Architecture for Language. *Journal of  
20 Cognitive Neuroscience* 32(8):1407–1427. <https://doi.org/10.1162/jocn.a.01552>
- 21 Mesgarani N, Cheung C, Johnson K, et al (2014) Phonetic Feature Encoding in Human  
22 Superior Temporal Gyrus. *Science* 343(6174):1006–1010. [https://doi.org/10.1126/  
23 science.1245994](https://doi.org/10.1126/science.1245994)

- 1 Michalareas G, Vezoli J, van Pelt S, et al (2016) Alpha-Beta and Gamma Rhythms  
2 Subserve Feedback and Feedforward Influences among Human Visual Cortical  
3 Areas. *Neuron* 89(2):384–397. <https://doi.org/10.1016/j.neuron.2015.12.018>
- 4 Noppeney U, Patterson K, Tyler LK, et al (2006) Temporal lobe lesions and seman-  
5 tic impairment: A comparison of herpes simplex virus encephalitis and semantic  
6 dementia. *Brain* 130(4):1138–1147. <https://doi.org/10.1093/brain/awl344>
- 7 Norris D, McQueen JM (2008) Shortlist B: A Bayesian model of continuous  
8 speech recognition. *Psychological Review* 115(2):357–395. [https://doi.org/10.1037/  
9 0033-295X.115.2.357](https://doi.org/10.1037/0033-295X.115.2.357)
- 10 Oganian Y, Chang EF (2019) A speech envelope landmark for syllable encoding in  
11 human superior temporal gyrus. *Science advances* 5(11):eaay6279
- 12 Ossmy O, Fried I, Mukamel R (2015) Decoding speech perception from single cell activ-  
13 ity in humans. *NeuroImage* 117:151–159. [https://doi.org/10.1016/j.neuroimage.  
14 2015.05.001](https://doi.org/10.1016/j.neuroimage.2015.05.001)
- 15 Patterson K, Nestor PJ, Rogers TT (2007) Where do you know what you know?  
16 The representation of semantic knowledge in the human brain. *Nature Reviews  
17 Neuroscience* 8(12):976–987. <https://doi.org/10.1038/nrn2277>
- 18 Perdikis D, Huys R, Jirsa VK (2011) Time Scale Hierarchies in the Functional  
19 Organization of Complex Behaviors. *PLoS Computational Biology* 7(9):e1002198.  
20 <https://doi.org/10.1371/journal.pcbi.1002198>
- 21 Pesaran B, Vinck M, Einevoll GT, et al (2018) Investigating large-scale brain dynamics  
22 using field potential recordings: Analysis and interpretation. *Nature Neuroscience*  
23 <https://doi.org/10.1038/s41593-018-0171-8>

- 1 Pillai AS, Jirsa VK (2017) Symmetry Breaking in Space-Time Hierarchies Shapes  
2 Brain Dynamics and Behavior. *Neuron* 94(5):1010–1026. [https://doi.org/10.1016/j.  
3 neuron.2017.05.013](https://doi.org/10.1016/j.neuron.2017.05.013)
- 4 Pulvermüller F (2018) Neural reuse of action perception circuits for language, concepts  
5 and communication. *Progress in Neurobiology* 160:1–44. [https://doi.org/10.1016/j.  
6 pneurobio.2017.07.001](https://doi.org/10.1016/j.pneurobio.2017.07.001)
- 7 Pylykkänen L (2020) Neural basis of basic composition: What we have learned from the  
8 red-boat studies and their extensions. *Philosophical Transactions of the Royal Soci-  
9 ety B: Biological Sciences* 375(1791):20190299. [https://doi.org/10.1098/rstb.2019.  
10 0299](https://doi.org/10.1098/rstb.2019.0299)
- 11 Quiroga RQ, Nadasdy Z, Ben-Shaul Y (2004) Unsupervised spike detection and sorting  
12 with wavelets and superparamagnetic clustering. *Neural computation* 16(8):1661–  
13 1687
- 14 Rahimi S, Farahibozorg SR, Jackson R, et al (2022) Task modulation of spatiotem-  
15 poral dynamics in semantic brain networks: An EEG/MEG study. *NeuroImage*  
16 246:118768. <https://doi.org/10.1016/j.neuroimage.2021.118768>
- 17 Ralph MAL, Jefferies E, Patterson K, et al (2017) The neural and computational bases  
18 of semantic cognition. *Nature Reviews Neuroscience* 18(1):42–55. [https://doi.org/  
19 10.1038/nrn.2016.150](https://doi.org/10.1038/nrn.2016.150)
- 20 Remington ED, Narain D, Hosseini EA, et al (2018) Flexible Sensorimotor Computa-  
21 tions through Rapid Reconfiguration of Cortical Dynamics. *Neuron* 98(5):1005–  
22 1019.e5. <https://doi.org/10.1016/j.neuron.2018.05.020>

- 1 Rutten S, Santoro R, Hervais-Adelman A, et al (2019) Cortical encoding of speech  
2 enhances task-relevant acoustic information. *Nature Human Behaviour* 3(9):974–  
3 987. <https://doi.org/10.1038/s41562-019-0648-9>
- 4 Schwartz MF, Kimberg DY, Walker GM, et al (2009) Anterior temporal involvement  
5 in semantic word retrieval: Voxel-based lesion-symptom mapping evidence from  
6 aphasia. *Brain* 132(12):3411–3427. <https://doi.org/10.1093/brain/awp284>
- 7 Scott SK (2000) Identification of a pathway for intelligible speech in the left temporal  
8 lobe. *Brain* 123(12):2400–2406. <https://doi.org/10.1093/brain/123.12.2400>
- 9 Sjerps MJ, Fox NP, Johnson K, et al (2019) Speaker-normalized sound representations  
10 in the human auditory cortex. *Nature Communications* 10(1). <https://doi.org/10.1038/s41467-019-10365-z>
- 11
- 12 Truccolo W (2016) From point process observations to collective neural dynamics:  
13 Nonlinear Hawkes process GLMs, low-dimensional dynamics and coarse graining.  
14 *Journal of Physiology-Paris* 110(4):336–347. <https://doi.org/10.1016/j.jphysparis.2017.02.004>
- 15
- 16 van Kerkoerle T, Self MW, Dagnino B, et al (2014) Alpha and gamma oscillations  
17 characterize feedback and feedforward processing in monkey visual cortex. *Proceed-*  
18 *ings of the National Academy of Sciences* 111(40):14332–14341. <https://doi.org/10.1073/pnas.1402773111>
- 19
- 20 Vignali L, Xu Y, Turini J, et al (2023) Spatiotemporal dynamics of abstract and  
21 concrete semantic representations. *Brain and Language* 243:105298. <https://doi.org/10.1016/j.bandl.2023.105298>
- 22
- 23 Visser M, Lambon Ralph MA (2011) Differential Contributions of Bilateral Ventral  
24 Anterior Temporal Lobe and Left Anterior Superior Temporal Gyrus to Semantic

- 1 Processes. *Journal of Cognitive Neuroscience* 23(10):3121–3131. [https://doi.org/10.](https://doi.org/10.1162/jocn.a.00007)  
2 [1162/jocn.a.00007](https://doi.org/10.1162/jocn.a.00007)
- 3 Vyas S, Golub MD, Sussillo D, et al (2020) Computation Through Neural Population  
4 Dynamics. *Annual Review of Neuroscience* 43(1):249–275. [https://doi.org/10.1146/](https://doi.org/10.1146/annurev-neuro-092619-094115)  
5 [annurev-neuro-092619-094115](https://doi.org/10.1146/annurev-neuro-092619-094115)
- 6 Yi HG, Leonard MK, Chang EF (2019) The Encoding of Speech Sounds in the Superior  
7 Temporal Gyrus. *Neuron* 102(6):1096–1110. [https://doi.org/10.1016/j.neuron.2019.](https://doi.org/10.1016/j.neuron.2019.04.023)  
8 [04.023](https://doi.org/10.1016/j.neuron.2019.04.023)
- 9 Zhang Y, Ding Y, Huang J, et al (2021) Hierarchical cortical networks of “voice  
10 patches” for processing voices in human brain. *Proceedings of the National Academy*  
11 *of Sciences* 118(52):e2113887118. <https://doi.org/10.1073/pnas.2113887118>