

1 Latent neural dynamics encode temporal context in speech

2 Emily P Stephen^{1,2}, Yuanning Li¹, Sean Metzger¹, Yulia Oganian¹, Edward F Chang^{1,*}

3
4 ¹ University of California San Francisco, Department of Neurological Surgery, San Francisco, CA 94143

5 ² Boston University, Department of Mathematics and Statistics, Boston, MA 02215

6 * Corresponding Author: Edward.Chang@ucsf.edu

7 8 Abstract

9 Direct neural recordings from human auditory cortex have demonstrated encoding for acoustic-
10 phonetic features of consonants and vowels. Neural responses also encode distinct acoustic
11 amplitude cues related to timing, such as those that occur at the onset of a sentence after a silent
12 period or the onset of the vowel in each syllable. Here, we used a group reduced rank regression
13 model to show that distributed cortical responses support a low-dimensional latent state
14 representation of temporal context in speech. The timing cues each capture more unique
15 variance than all other phonetic features and exhibit rotational or cyclical dynamics in latent
16 space from activity that is widespread over the superior temporal gyrus. We propose that these
17 spatially distributed timing signals could serve to provide temporal context for, and possibly bind
18 across time, the concurrent processing of individual phonetic features, to compose higher-order
19 phonological (e.g. word-level) representations.

20 Introduction

21 Natural speech is a continuous stream of complex acoustic features, and listeners build representations
22 of auditory objects at multiple levels, from phonemes, to syllables, words, and phrases (Berwick et al.,
23 2013; Chomsky, 1985). The cortical basis of these dynamic compositional operations is an active area of
24 research. There is evidence that the superior temporal gyrus (STG) performs speech-specific extraction
25 of acoustic-phonetic features (Mesgarani et al., 2014), but where and how these segmental features are
26 composed into longer units like words is less understood. Since the cascade of neural activity evoked by
27 a given acoustic-phonetic feature can last longer than the feature itself (Gwilliams et al., 2020;
28 Khalighinejad et al., 2017; Mesgarani et al., 2014; Näätänen and Picton, 1987; Norman-Haignere et al.,
29 2020), there is potential for overlap in the neural representations over time. Hence the neural
30 computations underlying speech comprehension should have a way to keep track of the temporal
31 context of the individual phonetic units in order to compose them into a higher order unit such as a word
32 (Fischer-Baum, 2018; Gwilliams et al., 2020).

33
34 We hypothesized that the mechanisms underlying temporal context tracking and composition in
35 auditory cortex would be reflected in low-dimensional latent dynamics of electrocorticography (ECoG)-

36 scale neural recordings. As neural recordings have grown in dimension, latent state models describing
37 lower-dimensional summaries of populations of neurons have become more popular as the explanatory
38 framework for understanding neural computation. In particular, there is a growing trend to map out
39 geometric characteristics of latent states that could be indicative of the computational roles that are
40 being played by the network (Russo et al., 2020, 2018; Seely et al., 2016; Vyas et al., 2020). One such
41 geometrical motif is rotational dynamics (Churchland et al., 2012), which have been implicated in
42 coordinating movements over time in the motor system (Buonomano and Laje, 2010; Cannon and Patel,
43 2021; Russo et al., 2020, 2018) (see Discussion). While the neural activity underlying speech perception
44 is likely to be very different from that underlying motor sequencing, low-dimensional dynamics across
45 the speech-responsive network in STG could reflect similar computational strategies to coordinate
46 temporal context during speech perception.

47
48 There is already reason to believe that STG encodes information about timing: some STG populations
49 respond to amplitude onset events found at the beginning of a sentence after silence period, or the
50 acoustic edges that occur at the onset of vowels in syllables (called ‘peak rate’) (Hamilton et al., 2018;
51 Oganian and Chang, 2019). If these signals are strong (representing a large proportion of the variance),
52 temporally similar across different populations, and spatially widespread, they could constitute a low-
53 dimensional latent state. In fact, Hamilton and colleagues (Hamilton et al., 2018) were able to find low-
54 dimensional dynamics tied to sentence onsets using unsupervised linear dimensionality reduction.
55 Unfortunately, due to the complex nature of the task (with a high-dimensional stimulus space and
56 relevant stimulus features occurring closely in time), unsupervised methods have trouble uncovering
57 dynamics related to other stimulus features, whose neural responses may overlap temporally and
58 spatially with sentence onset responses. This makes it difficult to describe latent dynamics related to
59 peak rate events, which are more closely aligned in timescale to the low-level compositional operations
60 that we seek to describe. Supervised models, on the other hand, have historically focused on individual
61 electrodes and as a result fail to describe latent dynamics that may reflect computational principles on
62 a larger spatial scale.

63
64 Here we use a multivariate supervised approach to model the activity across all speech-responsive STG
65 electrodes. Using integrative reduced rank regression (iRRR) (Li et al., 2019), we simultaneously estimate
66 a separate low-dimensional latent state for each stimulus feature, including sentence onsets, peak rate
67 events, and acoustic-phonetic features based on the place and manner of articulation. We find that iRRR
68 outperforms models that treat each electrode individually, indicating that substantial feature-related
69 information is shared across electrodes. The sentence onset and peak rate features explain more of the
70 variance than phonetic features, reaffirming the importance of these timing-related features for
71 encoding in STG. Furthermore, the latent states for the onset and peak rate are low-dimensional (5 and
72 6 dimensional, respectively) and distributed over centimeters of cortex, indicating a widespread signal
73 that would be available to coordinate local and downstream processing. Geometrically, the latent
74 dynamics contain a large proportion of rotational dynamics. Projections of the neural responses onto

75 these low-dimensional spaces can be used to decode the time relative to the most recent sentence onset
76 or peak rate event, with performance that is better than decoding from the full high-dimensional
77 responses across all electrodes. We propose that the sentence onset response is an initialization signal
78 and the peak rate latent states encode the time relative to acoustic events at the sentence and syllable
79 scales. For peak rate, this spatially distributed timing signal could be used in local and downstream
80 processing when composing word-level representations from low-level acoustic features.

81 Results

82 Model motivation and design

83 We modeled the high gamma (70-150 Hz) amplitude recorded on 331 speech-responsive electrodes
84 located over the left superior temporal gyrus (STG) in 11 participants while they passively listened to 438
85 naturally spoken sentences from the Texas Instruments and Massachusetts Institute of Technology
86 (TIMIT) acoustic-phonetic corpus (Garofolo et al., 1993). High gamma amplitudes in neural voltage
87 recordings are known to correlate with the firing rates (Dubey and Ray, 2020; Manning et al., 2009; Ray
88 et al., 2008; Ray and Maunsell, 2011; Scheffer-Teixeira et al., 2013) and dendritic processes (Bédard et
89 al., 2006; Leszczyński et al., 2020; Miller et al., 2009; Suzuki and Larkum, 2017) of neurons near the
90 electrode (Buzsáki et al., 2012), and we use them here as a proxy for the level of population activity
91 under the ECoG electrodes. Using our model, we show that high gamma responses to speech stimuli
92 across hundreds of electrodes can be parsimoniously represented as a combination of a few low-
93 dimensional latent state responses to specific feature events in the stimulus. Two latent states in
94 particular, corresponding to the sentence onset and peak rate features, reflect a large proportion of the
95 explained variance in the model, and their dynamic properties suggest specific computational roles in
96 the speech perception network.

97
98 Successful previous models of high gamma activity over STG have taken two different approaches: using
99 supervised regression to model single-electrode responses as a function of spectral or linguistic
100 characteristics in the audio speech signal (Aertsen and Johannesma, 1981; Holdgraf et al., 2017;
101 Mesgarani et al., 2014; Oganian and Chang, 2019; Theunissen et al., 2001), and using unsupervised
102 dimensionality reduction to infer latent states without reference to the characteristics of the audio
103 stimulus (Hamilton et al., 2018).

104
105 The advantage of the single-electrode regression models is that they characterize the relationship
106 between the neural responses and acoustic features in the speech signal. In the models, the high gamma
107 responses on individual electrodes are considered to be the result of a convolution of time-dependent
108 receptive fields with corresponding time series of acoustic features. The classic spectrotemporal
109 receptive field (STRF) model, for example, uses a mel spectrogram of the stimulus as the acoustic feature
110 representation, resulting in a framework where the neural receptive fields act as a linear filter on the

111 speech spectrogram (Theunissen et al., 2001). Based on the observation that electrode activity over STG
112 reflects information at the level of phonetic features rather than individual phonemes (Mesgarani et al.,
113 2014), Oganian and Chang (Oganian and Chang, 2019) used an event-based feature representation to
114 capture these effects and to show that some electrodes additionally have responses triggered by
115 sentence onsets and sharp transients in the acoustic envelope of the speech signal, called peak rate
116 events. While these models have been instrumental in describing the response patterns on individual
117 electrodes, they fail to capture latent dynamics that are shared across multiple electrodes, which could
118 uncover computational principles at work at a larger spatial scale.

119
120 An alternative approach uses unsupervised dimensionality reduction to investigate latent structure in
121 neural responses to speech (e.g. (Hamilton et al., 2018)). Using convex nonnegative matrix factorization,
122 they showed that electrodes can be naturally classified into two groups, “onset” electrodes that have a
123 short increase in high gamma activity at the onset of a sentence, and “sustained” electrodes that show
124 increased high gamma activity throughout the stimulus. This observation is also apparent using principal
125 component analysis, in which the first component has characteristic sustained profile, and the second
126 component has the onset profile (See Supplementary Figure S1). Note that the high gamma signals are
127 not intrinsically low-dimensional: 2 dimensions capture only 24% of the variance in speech responsive
128 electrodes (comparable to 16.9% of the variance in all electrodes captured in the first two clusters of
129 (Hamilton et al., 2018)) and 189 dimensions are necessary to capture 80% of the variance. This could be
130 related to the high-dimensional nature of the task: in an unsupervised framework in which the system
131 responds to stimulus features, the response dimensionality needs to be at least as high-dimensional as
132 the task itself (Gao et al., 2017; Stringer et al., 2019). Furthermore, both of these components are time-
133 locked to sentence onset, and it is difficult to connect them or higher components to other speech
134 features, possibly because the dynamics related to other features are not orthogonal to the sentence-
135 onset subspace or to each other. In particular, the dependence of the neural responses on the peak rate
136 events is not apparent from this analysis, and a model that could capture latent dynamics related to peak
137 rate would be valuable for describing population encoding of shorter timescales.

138
139 We chose to use a model that combines the advantages of the regression and dimensionality reduction
140 approaches, using a multivariate integrative reduced rank regression model (iRRR) (Li et al., 2019) to
141 estimate the latent dynamics attributed to each speech feature separately. This group-reduced-rank
142 model partitions the expected neural activity into a separate latent state for each feature, choosing the
143 best latent dimensionality for each feature while penalizing the total dimensionality across all features.
144 The model uses a multivariate adaptation of the event-based regression framework of Oganian and
145 Chang (Oganian and Chang, 2019). In matrix form, the model has the following structure:

$$146 \quad Y = \sum_{f=1}^F X_f B_f + E \quad (1)$$

147 Where Y is the $T \times N$ matrix of high gamma amplitude values across electrodes and timepoints, each X_f
148 ($T \times D$) represents the delayed feature events for feature f , and E ($T \times N$) is Gaussian noise, assumed
149 to be uncorrelated across electrodes (T : number of timepoints; N : number of electrodes, D : number of

150 delays, F : number of features). In Y , X , and E , the timepoints corresponding to subsequent sentence
151 stimuli are stacked together. The coefficient matrices B_f ($D \times N$) are the multivariate temporal response
152 functions (MTRFs), representing the responses of each electrode to the given feature across electrodes
153 and delays (up to 750ms).

154

155 Only speech responsive electrodes over STG were used for this analysis, defined using single-electrode
156 fits to a linear spectrotemporal model (see electrode selection in Methods). Figure 1A shows the
157 electrodes that were used, colored by the testing r^2 value of the fitted spectrotemporal model: STG
158 electrodes with $r^2 > 0.05$ were used for subsequent analyses ($N = 331$).

159

160 Figure 1B shows the feature events for an example sentence stimulus, “They’ve never met, you know”.
161 The top two panels show the stimulus waveform and mel spectrogram, respectively, with the times of
162 sentence onset and peak rate events indicated with vertical lines (solid and dashed, respectively). The
163 features fall into two categories: timing (sentence onset and peak rate) and acoustic-phonetic (dorsal,
164 coronal, labial, high, low, front, back, plosive, fricative, nasal). With the exception of peak rate, all of the
165 feature events were encoded as binary time series with a 1 representing an event occurring, and 0
166 otherwise. For peak rate, the time series contained continuous values representing the slope of the
167 acoustic amplitude signal at the time of maximal change, and 0 at all other times (in Figure 1B, red lines
168 indicate peak rate event times and red numbers indicate the peak rate magnitude). We chose to include
169 magnitude for peak rate events, because it is known to correlate very well with stressed syllables, i.e.
170 syllables with higher stress will have higher peak rate magnitude.

171

172 We fit the regression model using integrative reduced-rank regression (iRRR) (Li et al., 2019), which
173 applies a penalty based on a weighted sum of the nuclear norms of the feature matrices (see Methods
174 for more detail):

$$175 \{\hat{B}_f, iRRR\}_{f=1}^F = \underset{B_f \in \mathbb{R}^{D \times N}}{\operatorname{argmin}} \frac{1}{2T} \|Y - \sum_{f=1}^F X_f B_f\|_{\mathcal{F}}^2 + \lambda \sum_{f=1}^F w_f \|B_f\|_* \quad (2)$$

176 where $\|\cdot\|_{\mathcal{F}}$ represents the Frobenius (L2) norm, the w_f s are chosen to balance the regularization
177 across features, and λ is a regularization parameter. The notation $\|\cdot\|_*$ represents the nuclear norm, or
178 the sum of the singular values of the bracketed matrix. The nuclear norm penalty acts as an L1 penalty
179 on the singular values of each feature matrix, so the regression tends to find solutions where the feature
180 matrices are low-rank (i.e. sparse in the singular values). Because many of the singular values will be
181 zero, the fitted feature matrices can be represented using a low-dimensional singular value
182 decomposition:

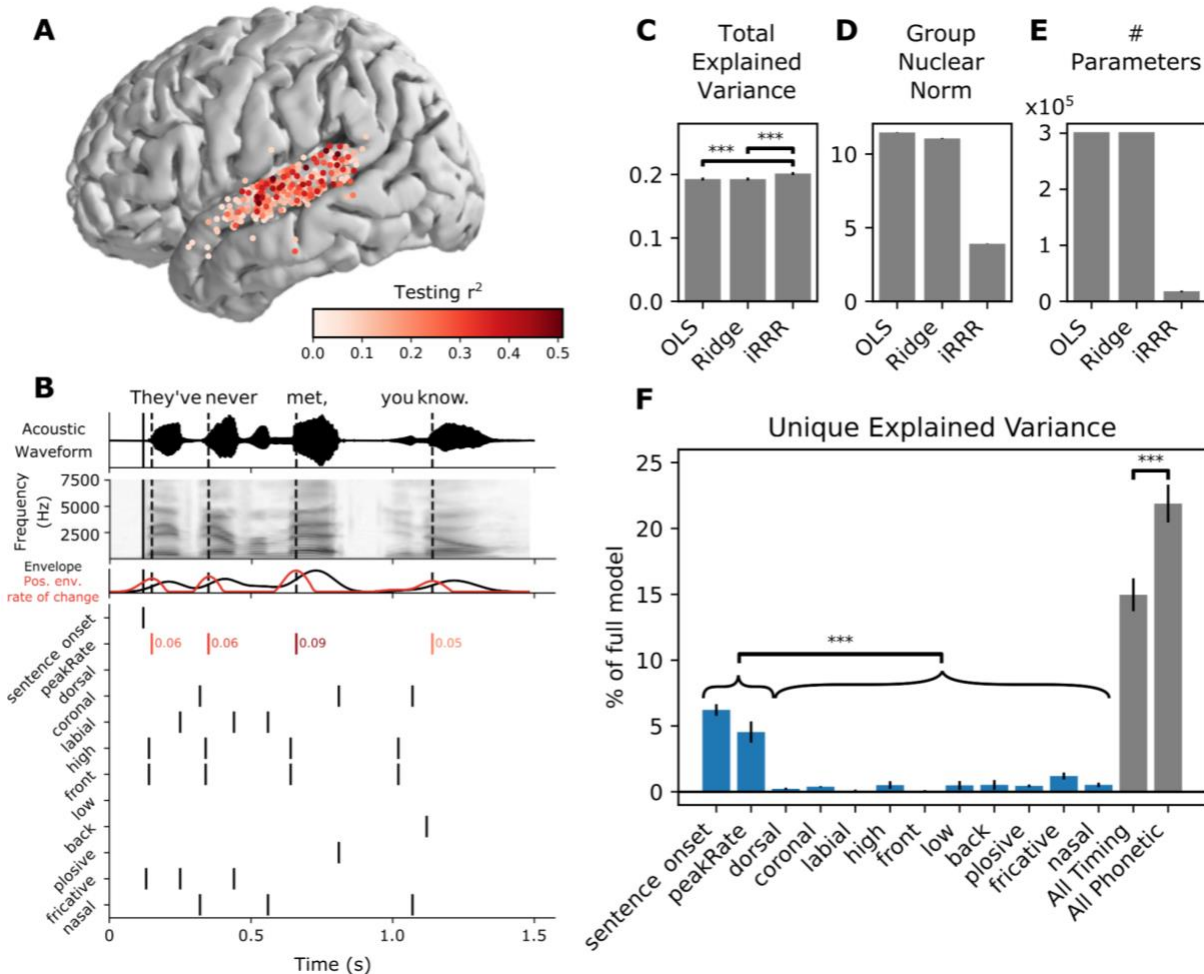
$$183 \hat{B}_f = U_f S_f V_f^T \quad (3)$$

184 where U_f is $D \times k$, S_f is $k \times k$, and V_f^T is $k \times N$, for some $k < N$. In other words, the full multivariate
185 feature receptive fields can be represented with a small number of patterns across time (columns of U_f),
186 patterns across electrodes (rows of V_f^T), and corresponding weights (values on the diagonal of S_f). The
187 number of dimensions k can be different for each feature, and it comes from balancing the contribution

188 of the feature to the first term (the mean squared error) with the contribution of the feature to the
 189 second term (the nuclear norm penalty), relative to other features. Increasing the tuning parameter λ
 190 will tend to increase the total number of dimensions used across all features.

191

192 For comparison, we also fit the same model using ordinary least squares (OLS) and ridge regression
 193 where a separate regularization parameter was chosen for each electrode. All models were fit using 10-
 194 fold cross validation. For the iRRR and ridge models, the regularization parameters were fit with an
 195 additional level of 5-fold cross-validation nested within the outer cross-validation.



196

197

198

199

200

201

202

203

204

205

206

207

208

Figure 1: iRRR outperforms models that treat each electrode individually, and sentence onset and peak rate capture more of the variance than phonetic features. A: Electrodes used for model fitting, colored according to the testing r^2 of the linear spectrotemporal (STRF) model (electrodes were selected for subsequent analysis if they were located over STG and if their testing r^2 for the spectrotemporal model was greater than 0.05). B: Features used for feature temporal receptive field modeling. Top: the acoustic waveform of an example sentence. The solid vertical line shows the sentence onset event, and the dashed vertical lines show the times of the peak rate events. Second panel: the corresponding mel-band spectrogram. Third panel: the envelope of the acoustic waveform (black) and the positive rate of change of the envelope (red). The peaks in the positive envelope rate of change are the peak rate events. Bottom: the feature time series. White space represents no event (encoded by 0 in the feature matrix), black lines represent event times (encoded by 1), and red lines indicate peak rate event times with their corresponding magnitude indicated to the right. C, D, E: Performance of the iRRR model in comparison to ordinary least squares (OLS) and ridge

209 regression (Ridge). 95% confidence intervals were estimated using the standard error of the mean across
210 cross-validation folds (see Methods). Significance was assessed for comparisons using two-sided paired t-
211 tests across cross-validation folds, *** $p < 0.0005$. C: Total explained variance, computed as the testing r^2
212 computed over all speech-responsive electrodes. D: Group nuclear norm, meaning the penalty term from
213 the iRRR model (see Equation 2). E: The effective number of parameters for the fitted models. F: Unique
214 explained variance for each feature (over all speech-responsive electrodes), expressed as a percentage of
215 the variance captured by the full model. Comparing individual features, both timing features have
216 significantly more unique explained variance than all phonetic features, after Bonferroni correction over
217 pairs (left). Also shown is the unique explained variance for the combined timing features (sentence onset
218 and peak rate) and the combined phonetic features (right). When the features are grouped, the phonetic
219 features capture more unique explained variance than the timing features.

220 iRRR outperforms models that treat each electrode individually, and sentence onset and
221 peak rate capture more of the variance than phonetic features

222 Figure 1C-E compare the three different fitting frameworks: OLS, ridge regression, and iRRR. Because the
223 regression framework is the same for all three, the fitted models have very similar total explained
224 variance (r^2 computed over all electrodes, Figure 1C), but iRRR by design achieves a much smaller nuclear
225 norm (Figure 1D), which results in solutions that can be described with 94% fewer parameters than OLS
226 and ridge regression (Figure 1E). The fact that the iRRR model captures as much information as the single-
227 electrode models using far fewer parameters suggests that substantial feature-related information is
228 shared across electrodes.

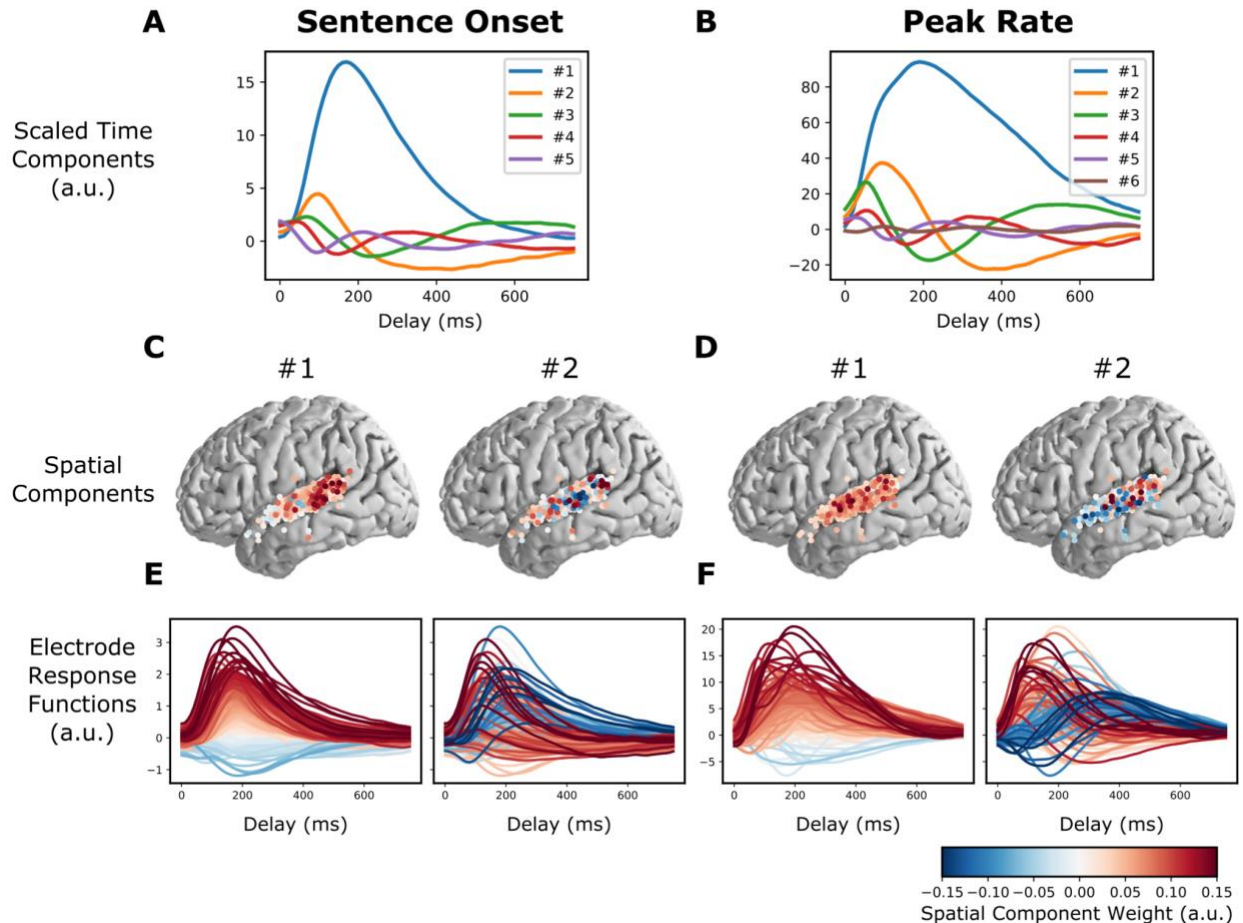
229
230 In order to compare the contribution to the model of the different features, we fit reduced versions of
231 the iRRR model with each feature left out. From there, we could compute the percent of explained
232 variance by comparing the r^2 of the full model (r_{Full}^2) to the r^2 of the model without feature f (r_{-f}^2):

$$233 \quad 100 \times (r_{Full}^2 - r_{-f}^2) / r_{Full}^2 \quad (4)$$

234 Figure 1F shows the result of this analysis: sentence onset and peak rate explain a larger percentage of
235 the full model variance than each of the phonetic features ($p < 0.0005$ for all comparisons using a two-
236 sided paired t-test after Bonferroni correction). This suggests that these two timing features reflect a
237 substantial amount of the speech-induced response across STG.

238
239 When the features are grouped into timing (sentence onset and peak rate) and phonetic (all other
240 features) groups, both groups explain a large proportion of the variance (15% and 22%, respectively).
241 Comparing the groups, however, the phonetic features explain more of the unique variance than the
242 timing features ($p < 0.0005$, two-sided paired t-test). This could be surprising in light of the individual
243 feature comparisons: while timing features capture more explained variance than phonetic features
244 when compared individually, when combined they capture less explained variance. This is likely due to
245 (1) correlations between individual phonetic features that lead to lower individual unique explained
246 variance and (2) the fact that more electrodes respond to sentence onset and peak rate than individual
247 phonetic features (Oganian and Chang, 2019), meaning that sentence onset and peak rate have more
248 widespread spatial support than the more spatially localized phonetic features. This more widespread

249 spatial support means that the iRRR model is better able to consolidate the activity patterns across
 250 multiple electrodes, i.e. capture the latent dynamics, for the sentence onset and peak rate features than
 251 for the phonetic features. Accordingly, the following two sections describe the latent state
 252 representations for the sentence onset and peak rate features in more detail.



253

254

255

256

257

258

259

260

261

262

263

264

Figure 2: The model fit captures known response differences between pSTG and mSTG. A and B: Time components for the sentence onset and peak rate response matrices, scaled by their singular value (all panels of this figure use the fit from the first cross-validation fold). C: The first two spatial components (across electrodes) for sentence onset. E: The electrode responses to sentence onset events (rows of the sentence onset response matrix), colored by the first (left) or second (right) peak rate spatial component. The first spatial component for sentence onset shows that electrodes with large sentence onset responses (red lines in the left plot of E) tend to be in posterior STG (red circles in the left plot of C). D and F: (like C and E, but for peak rate). The second spatial component divides electrodes into fast and slow peak rate responses (red and blue lines in the right plot of F), which tend to occur over pSTG and mSTG, respectively (red and blue circles in the right plot of D).

265

The model fit captures known response differences between pSTG and mSTG

266

In Hamilton and colleagues' (Hamilton et al., 2018) unsupervised model, the "onset" cluster of electrodes was found to occur primarily over the posterior portion of STG (pSTG). This observation led them to

267

268 propose that pSTG may play a role in detecting temporal landmarks at the sentence and phrase level,
269 because the short-latency, short-duration responses to sentence onsets in pSTG would be able to encode
270 the event time with high temporal resolution. This idea fits well within a long history of evidence that
271 stimulus responses in mSTG have longer latencies and longer durations than those in pSTG (Hamilton et
272 al., 2020; Jasmin et al., 2019; Yi et al., 2019). Here, the model fits recapitulate these known differences
273 between mSTG and pSTG.

274

275 As discussed above (Equation 3), the feature response matrices that are fitted by the iRRR model can be
276 decomposed into a small number of components across time (“time components”, columns of U_f),
277 components across electrodes (“spatial components”, rows of V_f^T), and corresponding weights (values
278 on the diagonal of S_f). Figure 2 shows the Sentence Onset and Peak Rate fitted feature matrices
279 decomposed in this way (Since U_f and V_f are orthonormal, their columns are unit vectors: as a result,
280 their units are arbitrary and can be best interpreted in relative terms).

281

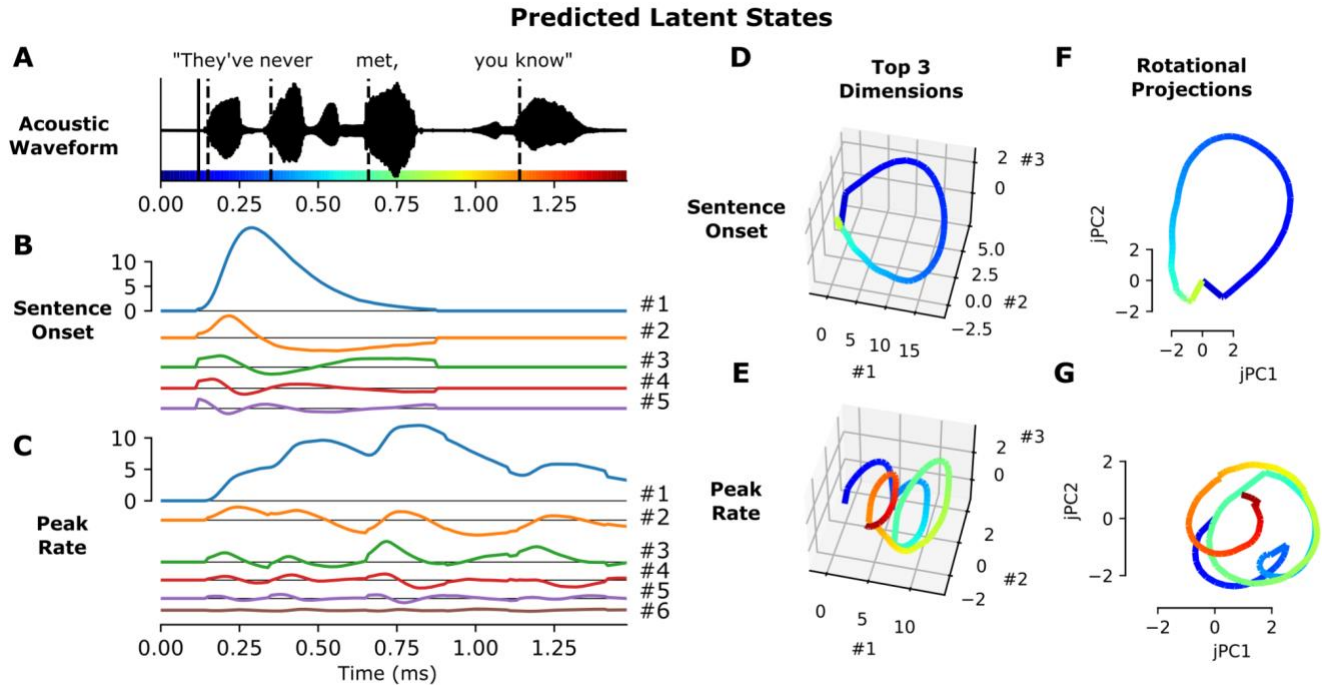
282 Figures 2A and B show the time components scaled by their corresponding weights, and Figures 2C and
283 D show the first two spatial components. To illustrate how the low dimensional components map back
284 to the response functions for individual neurons, Figures 2E and F show the individual electrode response
285 functions (rows of \widehat{B}_f), colored by the spatial component from Figures 2C and D.

286

287 Looking at the left panel of Figures 2C and 2E, we can see that electrodes that have large values in the
288 first spatial component (red circles in Figure 2C, left) have relatively larger overall responses to sentence
289 onset events (red lines in Figure 2E, left). These electrodes occur primarily over pSTG, which is in line
290 with previous findings (Hamilton et al., 2018).

291

292 For peak rate, the first component plays the same role: electrodes that have larger values in the first
293 spatial component (Figure 2D, left) have relatively larger overall responses to peak rate events (Figure
294 2F, left). Electrodes with large peak rate responses are not limited to pSTG like sentence onset
295 electrodes: rather, they are distributed over all of STG. In other words, the encoding of peak rate in STG
296 is not focal but is distributed over centimeters of cortex, suggesting a representation on a large spatial
297 scale. Interestingly, the second component does appear to have a spatial distinction between pSTG and
298 mSTG: electrodes with positive values for the second component tend to occur over pSTG, while
299 electrodes with negative values for the second component tend to occur over mSTG (Figure 2D, right).
300 The negative and positive values distinguish response functions by their temporal response profile:
301 positive values correspond to electrodes that have an early peak rate response, while negative values
302 correspond to electrodes that have a late peak rate response (Figure 2F, right). This suggests that peak
303 rate responses over pSTG are faster than peak rate responses over mSTG.



304
305
306
307
308
309
310
311
312
313
314
315

Figure 3: Feature latent states have rotational dynamics that capture continuous relative timing information. A: Acoustic waveform of the stimulus. Solid and dashed vertical lines indicate the timing of the sentence onset and peak rate events, respectively. Colors along the x-axis are used to indicate time parts D-G. B, C: Predicted latent states for the sentence onset and peak rate features corresponding to the given stimulus. D, E: Top three dimensions of the predicted sentence onset and peak rate latent states (the top three dimensions capture 98.7% and 98.8% of the variance in the sentence onset and peak rate coefficient matrices, respectively). F, G: Projection of the predicted sentence onset and peak rate latent states onto the plane of fastest rotation (identified using jPCA). The displayed jPCA projections capture 31.8% and 20.3% of the variance in the sentence onset and peak rate coefficient matrices, respectively. All panels of this figure use the fit from the first cross-validation fold.

316
317

Feature latent states have rotational dynamics that capture continuous relative timing information

318
319
320
321

To show how the latent states behave during the presentation of a stimulus, we used the fitted model to predict the dynamics in each latent state during the presentation of the sentence "They've never met, you know" (Figure 3). Predictions from the model can be computed in latent space using the decomposition defined in Equation 3:

322

$$\hat{Y}_{f;latent} = X_f U_f S_f \quad (5)$$

323
324
325
326
327
328

The sentence onset latent space has 5 dimensions and the peak rate latent space has 6 dimensions. While the sentence onset feature only occurs once at the beginning of the stimulus, evoking a single response across the sentence onset dimensions, the peak rate feature occurs several times, and the dynamics of the peak rate latent state do not go back to baseline in between peak rate events (Figure 3B and C). Plotting the top three dimensions, which capture more than 98% of the variance in the coefficient matrices (\hat{B}_f), shows cyclical dynamics for both sentence onset and peak rate (Figure 3D and

329 E): the sentence onset state rotates once at the beginning of the sentence, and the peak rate latent state
330 rotates 3-4 times, once after each peak rate event.

331

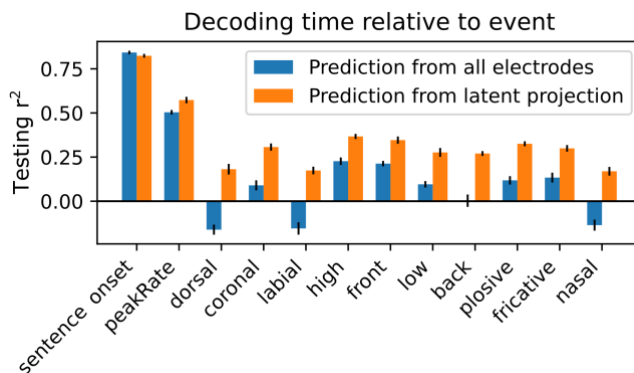
332 To quantify this effect, we used jPCA (Churchland et al., 2012) to identify the most rotational 2
333 dimensional subspace within the top three components of \hat{B}_f . These planes capture 31.8% and 20.3% of
334 the variance in the sentence onset and peak rate coefficient matrices, respectively, and they highlight
335 the cyclical dynamics that were visible in the top 3 dimensions (Figure 3F and G).

336

337 Note that seeing cyclical dynamics in the latent states is not necessarily surprising: the coefficient
338 matrices \hat{B}_f describe smooth multivariate evoked responses that will tend to start and end at the same
339 baseline. We highlight them here to motivate a geometrical argument for the computational role of the
340 peak rate responses (see Discussion) and to make the case that the structure of the peak rate responses
341 enables them to act as a temporal context signal against which other features are organized. In order for
342 the peak rate latent state to play this role, the trajectories should be sufficiently spread out in latent
343 space to enable downstream areas to decode the time relative to the most recent peak rate event using
344 just the instantaneous latent state. We investigate whether this is true in the next section.

345

346



347

348

349

350

351

352

353

Figure 4: Latent states from the model can be used to decode time relative to feature events. Performance of a perceptron model trained to decode the time relative to the most recent feature event, for each feature. The models were trained either using the full high-dimensional set of high gamma responses across electrodes (blue bars) or using the projection of those responses onto the subspaces spanned by the feature latent states (orange bars). Performance is quantified using the testing set r^2 .

354 Latent states from the model can be used to decode time relative to feature events

355 So far, we have described how the model is fit using known feature event times, and how the fitted
356 model can be used to predict responses given new feature events. We also wanted to know whether the
357 model fit could be used to decode the timing of events, which would indicate that sufficient information
358 is contained in the feature responses for downstream areas to use them as temporal context signals.

359

360 The set of spatial components for each feature defines a feature-specific subspace of the overall
361 electrode space. The projection of the observed high gamma time series onto this subspace is an
362 approximation of the feature latent state (note that it is not exact, because the different feature
363 subspaces are not orthogonal to each other):

$$364 \tilde{Y}_{f,proj} = YV_f \quad (6)$$

365 We asked whether this latent projection time series could be used to decode the time since the most
366 recent feature event.

367

368 Figure 4 shows the result of this analysis: a perceptron model was trained to decode the time since the
369 most recent feature event up to 750 ms, given either the activity on the full set of electrodes or the
370 projection of the electrode activity onto the corresponding feature subspace (see Methods). The
371 decoder for sentence onset performs slightly better when using all electrodes, which may be due to the
372 large proportion of the overall activity that is time-locked to sentence onsets (see Supplementary Figure
373 S1). For all other features, however, decoder performance using the reduced-dimensional latent
374 subspaces performs even better than decoding using the full dimensional activity across electrodes
375 (paired t-test over 10 cross validation folds, $p < 0.05$ with Bonferroni correction across features). Because
376 no information is gained in the projection operation, this is an indication that projecting onto the latent
377 subspaces increases the signal to noise ratio, i.e. removes activity that is irrelevant to decoding relative
378 time.

379 Discussion

380 We have shown that a low dimensional regression model, iRRR, performs as well as classic models in
381 representing high-gamma responses to timing and phonetic features of auditory stimuli, while using far
382 fewer parameters. It accomplishes this compression by capturing similarities in feature responses that
383 are shared across electrodes, which enables a low-dimensional latent state interpretation of the
384 dynamics of high gamma responses to stimulus features. The sentence onset and peak rate features
385 capture more unique variance than the other (phonetic) features, their responses are spread over both
386 mSTG and pSTG, and their latent states show rotational dynamics that repeat after each event. Based
387 on the geometry, duration, and spatial extent of the latent dynamics, we make the case that the
388 sentence onset response could act as an initialization signal to kick the network into a speech-encoding
389 state, while the peak rate response could provide a widespread temporal context signal that could be
390 used to compose word-level representations from low-level acoustic and phonetic features.

391

392 The large magnitude of sentence onset responses in ECoG high gamma responses has been reported
393 before (Hamilton et al., 2018): here, we confirm their large contribution to STG responses both using our
394 iRRR model (Figure 1) and using PCA (Supplementary Figure S1). Importantly, the latent dynamics related
395 to sentence onset last about 600 ms (Figure 2a). Since sentences in English often last longer than 600 ms
396 (e.g. the sentences in the TIMIT corpus used here ranged from 900 ms to 2.6 s), these onset-related

397 dynamics are unsuited to encode temporal context on an entire sentence level. Furthermore, sentence
398 boundaries in continuous natural speech are rarely indicated with pauses or silence (Yoon et al., 2007),
399 meaning that neural responses to acoustic onsets are unlikely to code sentence transitions. Rather, the
400 latent dynamics in response to onsets may serve as a non-speech specific temporal indicator of the
401 transition from silence to sound, occurring during perception of any auditory stimulus. During speech
402 perception, the speech-related cortical networks could use this non-specific event as a reset or
403 initialization signal. The idea that a large transient in the latent state could act to transition a network
404 between states is also thought to occur in the motor system, where condition-invariant movement onset
405 responses in the latent state mark the transition from motor preparation to motor behavior (Kaufman
406 et al., 2016).

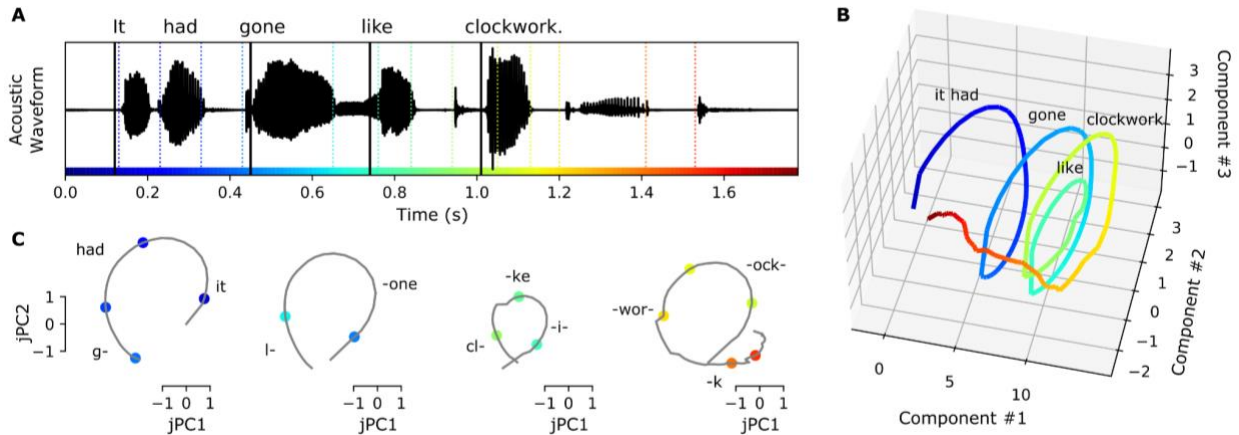
407

408 With regard to the peak rate dynamics, we propose that the computational role of the peak rate feature
409 response is to keep track of word-level temporal context using a clock-like representation. The idea that
410 structured latent state dynamics can act as clocks has been proposed in several different cognitive
411 domains, most commonly in the motor system (Buonomano and Laje, 2010; Churchland et al., 2012;
412 Remington et al., 2018; Vyas et al., 2020) (c.f. (Lebedev et al., 2020)) and in temporal interval estimation
413 and perception (Cannon and Patel, 2021; Gámez et al., 2019; Mauk and Buonomano, 2004; Wang et al.,
414 2018). In the motor system, Russo and colleagues (Russo et al., 2020) describe population dynamics in
415 primary motor cortex (M1) and supplementary motor area (SMA) while a monkey performed a cyclic
416 motor action. The population dynamics in M1 were rotational, exhibiting one rotation for each motor
417 cycle, while the dynamics in SMA were shaped like a spiral, where 2-dimensional rotations for each
418 motor cycle were translated along a third dimension. They proposed that this structure would be well-
419 suited to keep track of progress through multi-cycle actions: each rotation encodes a single action, and
420 translation along the third dimension encodes progress through the motor sequence. The rotational
421 component of SMA population trajectories has also been suggested to operate as a time-keeping signal
422 in auditory beat perception, where rotations through latent space keep track of the interval between
423 beats (Cannon and Patel, 2021).

424

425 The peak rate latent state in STG could similarly be playing a computational role in auditory speech
426 perception: the rotations in the peak rate subspace could serve to keep track of the time relative to the
427 peak rate event, chunking time into intervals starting at the onset of a vowel. These intervals could then
428 be used by downstream processing to give temporal context to the fine-grained phonetic feature
429 information conveyed by other subpopulations. In other words, the rotational peak rate latent state
430 could provide a temporal scaffolding on which individual phonetic features can be organized. Figure 5
431 illustrates this idea: when hearing the sentence “It had gone like clockwork,” the peak rate latent state
432 partitions the sentence into four rotations, each one capturing the time since the most recent peak rate
433 event. Downstream processing streams could combine this information with the phonetic feature
434 information to put the phonetic feature events into their local context, here at the level of words or
435 small sets of words (Figure 5C). Peak rate is in a unique position to play this role: it is the only feature

436 that repeats within the linguistic structure of speech at the level of syllables/words, without reference
 437 to the linguistic contents. In addition, the peak rate responses are distributed over centimeters of cortex
 438 (Figure 2D) so the temporal context information would be widely available to local and downstream
 439 processing.
 440



441 Figure 5: Peak rate rotational latent states could provide a temporal scaffolding on which individual acoustic
 442 features can be organized. A: The acoustic waveform for the stimulus “It had gone like clockwork”. Solid
 443 vertical lines indicate the times of peak rate events, and colored dashed vertical lines indicate the times of
 444 phonetic feature events. Colors are used to indicate time in all panels. B: The predicted peak rate latent
 445 state follows a spiral trajectory in the top 3 dimensions. C: Projected onto the plane of greatest rotation
 446 (jPC1 and 2), the predicted peak rate latent state divides the sentence into four intervals, each consisting
 447 of a rotation through state space that captures the time since the peak rate event occurred. Downstream
 448 processing could combine the relative time information encoded in the peak rate subspace (grey traces)
 449 with the feature identities encoded in the feature subspaces (colored points) to compose higher-order
 450 representations of words or small groups of words. Text in panels B and C indicates the approximate timing
 451 of the words in the stimulus.
 452

453
 454 In order for the peak rate latent state to play this role, it should have a couple of properties. First, there
 455 should be a mapping from points in state space to different relative times. As we showed in Figure 3, the
 456 rotational dynamics cause different relative times to be encoded in different locations of the latent
 457 space. Second, the trajectories in latent space should be consistent enough to support decoding of
 458 relative time in the presence of noise. In Figure 4, we showed that the projections of the neural activity
 459 onto the subspaces spanned by the feature latent states support decoding of the time relative to the
 460 most recent feature event. Note that while the latent state projections support decoding better than
 461 decoding from the full high-dimensional signal, the actual performance for peak rate is somewhat low
 462 (~50%). A possible reason for this could be that some peak rate events are more effective at driving the
 463 latent state than others (even after accounting for peak rate magnitude, as the model does), resulting in
 464 inconsistent decoding of the time since the most recent peak rate event.
 465

466 Beyond the two-dimensional rotational dynamics, the peak rate latent trajectory forms a spiral in 3
 467 dimensions (Figure 5B), similar to population trajectories in SMA during motor sequences (Russo et al.,

468 2020). This suggests that the peak rate subpopulation may additionally encode the ordering of the word-
469 level intervals within a larger linguistic context, such as the phrase level.

470

471 Furthermore, the representation of these intervals does not require top-down predictive coding
472 (Hovsepyan et al., 2020; Lewis and Bastiaansen, 2015; Park et al., 2015; Pefkou et al., 2017) or
473 entrainment of ongoing oscillations (Canolty, 2007; Ghitza, 2011; Giraud and Poeppel, 2012; Hovsepyan
474 et al., 2020; Martin, 2020; Pittman-Polletta et al., 2020): in our model they are implemented via event-
475 related potentials triggered by discrete acoustic (peak rate) events. While top-down and oscillatory
476 mechanisms may play important roles in speech perception, our model demonstrates that some speech
477 segmentation and context processing can be performed without them.

478

479 The events that we focus on for speech segmentation are peak rate events, moments of sharp increases
480 in the acoustic envelope. The peak rate events in the model are coded with their magnitude (the slope
481 of the rise in the acoustic envelope), which allows the model dynamics to change proportionally to the
482 size of the event. This is important because peak rate events, also called auditory onset edges (Biermann
483 and Heil, 2000; Doelling et al., 2014; Heil and Neubauer, 2001), differ in magnitude based on the stress
484 level of the corresponding syllable (Oganian and Chang, 2019). This means that the dynamics triggered
485 by peak rate events are sensitive to prosodic structure, both stressed syllables within words and stressed
486 words within phrases. To investigate this further, it would be helpful to use a speech stimulus corpus
487 with more complex prosodic structure than the TIMIT corpus used here.

488

489 In summary, our model (iRRR) represents STG high gamma responses to natural speech stimuli as a
490 superposition of responses to individual phonetic and timing features, where each feature has a
491 corresponding low-dimensional latent state that is shared across electrodes. It performs as well as single
492 electrode models while using far fewer parameters, indicating that substantial feature-related
493 information is shared across electrodes. Sentence onset and peak rate events, features representing
494 timing at the sentence and syllable scales, capture more unique variance than phonetic features. The
495 latent dynamics for sentence onset and peak rate contain information about the time since the most
496 recent (sentence onset or peak rate) event, and the information is distributed across centimeters of
497 cortex. We make the case that for peak rate, this relative timing information could play a role in
498 composing word-level representations from low-level acoustic features, without requiring oscillatory or
499 top-down mechanisms.

500

501

502

503

504

505 Author Contributions

506 Conceptualization: E.P.S, E.F.C., Y.O., Y.L., S.M.; Data Curation: Y.O.; Formal Analysis: E.P.S., Y.O., Y.L.;
507 Funding acquisition: E.F.C.; Supervision: E.F.C.; Resources: E.F.C.; Software: E.P.S., Y.O.; Visualization:
508 E.P.S., Y.L., S.M., Y.O.; Writing: - original draft: E.P.S.; Writing - review & editing: E.P.S., E.F.C., Y.O., Y.L.,
509 S.M..

510 Acknowledgements

511 This work was supported by grants from the NIH (R01-DC012379 and U01-NS117765 to EFC). This
512 research was also supported by Bill and Susan Oberndorf, The Joan and Sandy Weill Foundation, and The
513 William K. Bowes Foundation. The authors would also like to thank the members of the Chang lab at
514 UCSF as well as James Hieronymus and Benjamin Antin for valuable feedback.

515 Declaration of Interests

516 The authors declare no competing interests.

517 Bibliography

- 518 Aertsen, A.M.H.J., Johannesma, P.I.M., 1981. The Spectro-Temporal Receptive Field: A functional
519 characteristic of auditory neurons. *Biol. Cybern.* 42, 133–143. <https://doi.org/10.1007/BF00336731>
- 520 Antin, B., Shenoy, K., Linderman, S., 2021. Probabilistic jPCA: a constrained model of neural dynamics.,
521 in: *Cosyne Abstracts 2021*. Presented at the Cosyne21, Online.
- 522 Aoi, M.C., Mante, V., Pillow, J.W., 2020. Prefrontal cortex exhibits multidimensional dynamic encoding
523 during decision-making. *Nat. Neurosci.* 23, 1410–1420. <https://doi.org/10.1038/s41593-020-0696-5>
- 524 Aoi, M.C., Pillow, J.W., 2019. Model-based targeted dimensionality reduction for neuronal population
525 data 15.
- 526 Austern, M., Zhou, W., 2020. Asymptotics of Cross-Validation. *ArXiv200111111 Math Stat*.
- 527 Bates, S., Hastie, T., Tibshirani, R., 2021. Cross-validation: what does it estimate and how well does it
528 do it? *ArXiv210400673 Math Stat*.
- 529 Bédard, C., Kröger, H., Destexhe, A., 2006. Does the $1/f$ Frequency Scaling of Brain Signals Reflect Self-
530 Organized Critical States? *Phys. Rev. Lett.* 97, 118102.
531 <https://doi.org/10.1103/PhysRevLett.97.118102>
- 532 Bengio, Y., Grandvalet, Y., 2004. No unbiased estimator of the variance of k-fold cross-validation. *J.*
533 *Mach. Learn. Res.* 5, 1089–1105.
- 534 Berwick, R.C., Friederici, A.D., Chomsky, N., Bolhuis, J.J., 2013. Evolution, brain, and the nature of
535 language. *Trends Cogn. Sci.* 17, 89–98. <https://doi.org/10.1016/j.tics.2012.12.002>
- 536 Biermann, S., Heil, P., 2000. Parallels between timing of onset responses of single neurons in cat and of
537 evoked magnetic fields in human auditory cortex. *J. Neurophysiol.* 84, 2426–2439.
538 <https://doi.org/10.1152/jn.2000.84.5.2426>
- 539 Buonomano, D.V., Laje, R., 2010. Population clocks: motor timing with neural dynamics. *Trends Cogn.*
540 *Sci.* 14, 520–527. <https://doi.org/10.1016/j.tics.2010.09.002>

- 541 Buzsáki, G., Anastassiou, C.A., Koch, C., 2012. The origin of extracellular fields and currents — EEG,
542 ECoG, LFP and spikes. *Nat. Rev. Neurosci.* 13, 407–420. <https://doi.org/10.1038/nrn3241>
- 543 Cannon, J.J., Patel, A.D., 2021. How Beat Perception Co-opts Motor Neurophysiology. *Trends Cogn. Sci.*
544 25, 137–150. <https://doi.org/10.1016/j.tics.2020.11.002>
- 545 Canolty, R.T., 2007. Spatiotemporal dynamics of word processing in the human brain. *Front. Neurosci.*
546 1, 185–196. <https://doi.org/10.3389/neuro.01.1.1.014.2007>
- 547 Chomsky, N., 1985. *Syntactic structures*, 14. printing. ed, Janua Linguarum Series minor. Mouton, The
548 Hague.
- 549 Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., Shenoy, K.V.,
550 2012. Neural population dynamics during reaching. *Nature* 1–8.
551 <https://doi.org/10.1038/nature11129>
- 552 Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M.,
553 Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for
554 subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*
555 31, 968–980. <https://doi.org/10.1016/j.neuroimage.2006.01.021>
- 556 Dietterich, T.G., 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning
557 Algorithms. *Neural Comput.* 10, 1895–1923. <https://doi.org/10.1162/089976698300017197>
- 558 Doelling, K., Arnal, L., Ghitza, O., Poeppel, D., 2014. Acoustic landmarks drive delta-theta oscillations to
559 enable speech comprehension by facilitating perceptual parsing. *NeuroImage* 85.
560 <https://doi.org/10.1016/j.neuroimage.2013.06.035>
- 561 Dubey, A., Ray, S., 2020. Comparison of tuning properties of gamma and high-gamma power in local
562 field potential (LFP) versus electrocorticogram (ECoG) in visual cortex. *Sci. Rep.* 10, 5422.
563 <https://doi.org/10.1038/s41598-020-61961-9>
- 564 Fischer-Baum, S., 2018. A Common Representation of Serial Position in Language and Memory, in:
565 *Psychology of Learning and Motivation*. Elsevier, pp. 31–54.
566 <https://doi.org/10.1016/bs.plm.2018.08.002>
- 567 Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D.H., Busa, E., Seidman, L.J.,
568 Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., Dale, A.M., 2004. Automatically
569 Parcellating the Human Cerebral Cortex. *Cereb. Cortex* 14, 11–22.
570 <https://doi.org/10.1093/cercor/bhg087>
- 571 Gámez, J., Mendoza, G., Prado, L., Betancourt, A., Merchant, H., 2019. The amplitude in periodic neural
572 state trajectories underlies the tempo of rhythmic tapping. *PLOS Biol.* 17, e3000054.
573 <https://doi.org/10.1371/journal.pbio.3000054>
- 574 Gao, P., Trautmann, E., Yu, B.M., Santhanam, G., Ryu, S., Shenoy, K., Ganguli, S., 2017. A theory of
575 multineuronal dimensionality, dynamics and measurement 1–50. <https://doi.org/10.1101/214262>
- 576 Garofolo, J.S., Lamel, L.F., Fisher, W.M., Pallett, D.S., Dahlgren, N.L., Zue, V., Fiscus, J.G., 1993. TIMIT
577 Acoustic-Phonetic Continuous Speech Corpus. <https://doi.org/10.35111/17GK-BN40>
- 578 Ghitza, O., 2011. Linking Speech Perception and Neurophysiology: Speech Decoding Guided by
579 Cascaded Oscillators Locked to the Input Rhythm. *Front. Psychol.* 2.
580 <https://doi.org/10.3389/fpsyg.2011.00130>
- 581 Giraud, A.-L., Poeppel, D., 2012. Cortical oscillations and speech processing: emerging computational
582 principles and operations. *Nat. Neurosci.* 15, 511–517. <https://doi.org/10.1038/nn.3063>
- 583 Gwilliams, L., King, J.-R., Marantz, A., Poeppel, D., 2020. Neural dynamics of phoneme sequencing in
584 real speech jointly encode order and invariant content (preprint). *Neuroscience*.
585 <https://doi.org/10.1101/2020.04.04.025684>

- 586 Hamilton, L.S., Chang, D.L., Lee, M.B., Chang, E.F., 2017. Semi-automated Anatomical Labeling and
587 Inter-subject Warping of High-Density Intracranial Recording Electrodes in Electrocorticography.
588 *Front. Neuroinformatics* 11. <https://doi.org/10.3389/fninf.2017.00062>
- 589 Hamilton, L.S., Edwards, E., Chang, E.F., 2018. A Spatial Map of Onset and Sustained Responses to
590 Speech in the Human Superior Temporal Gyrus. *Curr. Biol.* 28, 1860-1871.e4.
591 <https://doi.org/10.1016/j.cub.2018.04.033>
- 592 Hamilton, L.S., Oganian, Y., Chang, E.F., 2020. Topography of speech-related acoustic and phonological
593 feature encoding throughout the human core and parabelt auditory cortex. *bioRxiv*
594 2020.06.08.121624. <https://doi.org/10.1101/2020.06.08.121624>
- 595 Heil, P., Neubauer, H., 2001. Temporal Integration of Sound Pressure Determines Thresholds of
596 Auditory-Nerve Fibers. *J. Neurosci.* 21, 7404–7415. <https://doi.org/10.1523/JNEUROSCI.21-18-07404.2001>
- 598 Holdgraf, C.R., Rieger, J.W., Micheli, C., Martin, S., Knight, R.T., Theunissen, F.E., 2017. Encoding and
599 Decoding Models in Cognitive Electrophysiology. *Front. Syst. Neurosci.* 11.
600 <https://doi.org/10.3389/fnsys.2017.00061>
- 601 Hovsepyan, S., Olasagasti, I., Giraud, A.-L., 2020. Combining predictive coding and neural oscillations
602 enables online syllable recognition in natural speech. *Nat. Commun.* 11, 3117.
603 <https://doi.org/10.1038/s41467-020-16956-5>
- 604 Jasmin, K., Lima, C.F., Scott, S.K., 2019. Understanding rostral–caudal auditory cortex contributions to
605 auditory perception. *Nat. Rev. Neurosci.* 20, 425–434. <https://doi.org/10.1038/s41583-019-0160-2>
- 606 Kaufman, M.T., Seely, J.S., Sussillo, D., Ryu, S.I., Shenoy, K.V., Churchland, M.M., 2016. The Largest
607 Response Component in the Motor Cortex Reflects Movement Timing but Not Movement Type.
608 *eneuro* 3, ENEURO.0085-16.2016. <https://doi.org/10.1523/ENEURO.0085-16.2016>
- 609 Khalighinejad, B., Cruzatto da Silva, G., Mesgarani, N., 2017. Dynamic Encoding of Acoustic Features in
610 Neural Responses to Continuous Speech. *J. Neurosci.* 37, 2176–2185.
611 <https://doi.org/10.1523/JNEUROSCI.2383-16.2017>
- 612 Lebedev, M.A., Ninenko, I., Ossadtchi, A., 2020. Rotational dynamics versus sequence-like responses
613 (preprint). *Neuroscience*. <https://doi.org/10.1101/2020.09.16.300046>
- 614 Leszczyński, M., Barczak, A., Kajikawa, Y., Ulbert, I., Falchier, A.Y., Tal, I., Haegens, S., Melloni, L., Knight,
615 R.T., Schroeder, C.E., 2020. Dissociation of broadband high-frequency activity and neuronal firing in
616 the neocortex. *Sci. Adv.* 6, eabb0977. <https://doi.org/10.1126/sciadv.abb0977>
- 617 Lewis, A.G., Bastiaansen, M., 2015. A predictive coding framework for rapid neural dynamics during
618 sentence-level language comprehension. *CORTEX* 68, 155–168.
619 <https://doi.org/10.1016/j.cortex.2015.02.014>
- 620 Li, G., Liu, X., Chen, K., 2019. Integrative multi-view regression: Bridging group-sparse and low-rank
621 models. *Biometrics* 75, 593–602. <https://doi.org/10.1111/biom.13006>
- 622 Manning, J.R., Jacobs, J., Fried, I., Kahana, M.J., 2009. Broadband Shifts in Local Field Potential Power
623 Spectra Are Correlated with Single-Neuron Spiking in Humans. *J. Neurosci.* 29, 13613–13620.
624 <https://doi.org/10.1523/JNEUROSCI.2041-09.2009>
- 625 Martin, A.E., 2020. A Compositional Neural Architecture for Language. *J. Cogn. Neurosci.* 32, 1407–
626 1427. https://doi.org/10.1162/jocn_a_01552
- 627 Mauk, M.D., Buonomano, D.V., 2004. THE NEURAL BASIS OF TEMPORAL PROCESSING. *Annu. Rev.*
628 *Neurosci.* 27, 307–340. <https://doi.org/10.1146/annurev.neuro.27.070203.144247>
- 629 Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F., 2014. Phonetic Feature Encoding in Human
630 Superior Temporal Gyrus. *Science* 343, 1006–1010. <https://doi.org/10.1126/science.1245994>

- 631 Miller, K.J., Sorensen, L.B., Ojemann, J.G., Den Nijs, M., 2009. Power-law scaling in the brain surface
632 electric potential. *PLoS Comput. Biol.* 5, e1000609.
633 <https://doi.org/10.1371/journal.pcbi.1000609.g005>
- 634 Nätänen, R., Picton, T., 1987. The N1 Wave of the Human Electric and Magnetic Response to Sound: A
635 Review and an Analysis of the Component Structure. *Psychophysiology* 24, 375–425.
636 <https://doi.org/10.1111/j.1469-8986.1987.tb00311.x>
- 637 Norman-Haignere, S.V., Long, L.K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E.M., Feldstein, N.A.,
638 McKhann, G.M., Schevon, C.A., Flinker, A., Mesgarani, N., 2020. Multiscale integration organizes
639 hierarchical computation in human auditory cortex (preprint). *Neuroscience*.
640 <https://doi.org/10.1101/2020.09.30.321687>
- 641 Oganian, Y., Chang, E.F., 2019. A speech envelope landmark for syllable encoding in human superior
642 temporal gyrus. *Sci. Adv.* 14.
- 643 Park, H., Ince, R.A.A., Schyns, P.G., Thut, G., Gross, J., 2015. Frontal Top-Down Signals Increase Coupling
644 of Auditory Low-Frequency Oscillations to Continuous Speech in Human Listeners. *Curr. Biol.* 25,
645 1649–1653. <https://doi.org/10.1016/j.cub.2015.04.049>
- 646 Pefkou, M., Arnal, L.H., Fontolan, L., Giraud, A.-L., 2017. θ -Band and β -Band Neural Activity Reflects
647 Independent Syllable Tracking and Comprehension of Time-Compressed Speech. *J. Neurosci.* 37,
648 7930–7938. <https://doi.org/10.1523/JNEUROSCI.2882-16.2017>
- 649 Pittman-Polletta, B.R., Wang, Y., Stanley, D.A., Schroeder, C.E., Whittington, M.A., Kopell, N.J., 2020.
650 Differential contributions of synaptic and intrinsic inhibitory currents to speech segmentation via
651 flexible phase-locking in neural oscillators (preprint). *Neuroscience*.
652 <https://doi.org/10.1101/2020.01.11.902858>
- 653 Ray, S., Crone, N.E., Niebur, E., Franaszczuk, P.J., Hsiao, S.S., 2008. Neural Correlates of High-Gamma
654 Oscillations (60-200 Hz) in Macaque Local Field Potentials and Their Potential Implications in
655 Electroencephalography. *J. Neurosci. Off. J. Soc. Neurosci.* 28, 11526–11536.
656 <https://doi.org/10.1523/JNEUROSCI.2848-08.2008>
- 657 Ray, S., Maunsell, J.H.R., 2011. Different Origins of Gamma Rhythm and High-Gamma Activity in
658 Macaque Visual Cortex. *PLoS Biol.* 9, e1000610. <https://doi.org/10.1371/journal.pbio.1000610.g008>
- 659 Remington, E.D., Egger, S.W., Narain, D., Wang, J., Jazayeri, M., 2018. A Dynamical Systems Perspective
660 on Flexible Motor Timing. *Trends Cogn. Sci.* 22, 938–952. <https://doi.org/10.1016/j.tics.2018.07.010>
- 661 Russo, A.A., Bittner, S.R., Perkins, S.M., Seely, J.S., London, B.M., Lara, A.H., Miri, A., Marshall, N.J.,
662 Kohn, A., Jessell, T.M., Abbott, L.F., Cunningham, J.P., Churchland, M.M., 2018. Motor Cortex Embeds
663 Muscle-like Commands in an Untangled Population Response. *Neuron* 97, 953-966.e8.
664 <https://doi.org/10.1016/j.neuron.2018.01.004>
- 665 Russo, A.A., Khajeh, R., Bittner, S.R., Perkins, S.M., Cunningham, J.P., Abbott, L.F., Churchland, M.M.,
666 2020. Neural Trajectories in the Supplementary Motor Area and Motor Cortex Exhibit Distinct
667 Geometries, Compatible with Different Classes of Computation. *Neuron* 107, 745-758.e6.
668 <https://doi.org/10.1016/j.neuron.2020.05.020>
- 669 Scheffer-Teixeira, R., Belchior, H., Leão, R.N., Ribeiro, S., Tort, A.B.L., 2013. On high-frequency field
670 oscillations (>100 Hz) and the spectral leakage of spiking activity. *J. Neurosci.* 33, 1535–1539.
671 <https://doi.org/10.1523/JNEUROSCI.4217-12.2013>
- 672 Seely, J.S., Kaufman, M.T., Ryu, S.I., Shenoy, K.V., Cunningham, J.P., Churchland, M.M., 2016. Tensor
673 Analysis Reveals Distinct Population Structure that Parallels the Different Computational Roles of
674 Areas M1 and V1. *PLOS Comput. Biol.* 12, e1005164. <https://doi.org/10.1371/journal.pcbi.1005164>

- 675 Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., Harris, K.D., 2019. High-dimensional geometry
676 of population responses in visual cortex. *Nature* 571, 361–365. [https://doi.org/10.1038/s41586-019-](https://doi.org/10.1038/s41586-019-1346-5)
677 1346-5
- 678 Suzuki, M., Larkum, M.E., 2017. Dendritic calcium spikes are clearly detectable at the cortical surface.
679 *Nat. Commun.* 8, 276. <https://doi.org/10.1038/s41467-017-00282-4>
- 680 Theunissen, F.E., David, S.V., Singh, N.C., Hsu, A., Vinje, W.E., Gallant, J.L., 2001. Estimating spatio-
681 temporal receptive fields of auditory and visual neurons from their responses to natural stimuli.
682 *Netw. Comput. Neural Syst.* 12, 289–316. <https://doi.org/10.1080/net.12.3.289.316>
- 683 Vyas, S., Golub, M.D., Sussillo, D., Shenoy, K.V., 2020. Computation Through Neural Population
684 Dynamics. *Annu. Rev. Neurosci.* 43, 249–275. [https://doi.org/10.1146/annurev-neuro-092619-](https://doi.org/10.1146/annurev-neuro-092619-094115)
685 094115
- 686 Wang, J., Narain, D., Hosseini, E.A., Jazayeri, M., 2018. Flexible timing by temporal scaling of cortical
687 responses. *Nat. Neurosci.* 21, 102–110. <https://doi.org/10.1038/s41593-017-0028-6>
- 688 Yi, H.G., Leonard, M.K., Chang, E.F., 2019. The Encoding of Speech Sounds in the Superior Temporal
689 Gyrus. *Neuron* 102, 1096–1110. <https://doi.org/10.1016/j.neuron.2019.04.023>
- 690 Yoon, T.-J., Cole, J., Hasegawa-Johnson, M., 2007. On the edge: Acoustic cues to layered prosodic
691 domains, in: *Proceedings of ICPhS*. Citeseer, pp. 1264–1267.
- 692
- 693

694 Methods

695 Participants

696 Participants included 11 patients (6M/5F, age 31 +/- 12 years) undergoing treatment for intractable
697 epilepsy. As a part of their clinical evaluation for epilepsy surgery, high-density intracranial electrode
698 grids (AdTech 256 channels, 4mm center-to-center spacing and 1.17mm diameter) were implanted
699 subdurally over the left peri-Sylvian cortex. All procedures were approved by the University of
700 California, San Francisco Institutional Review Board, and all patients provided informed written
701 consent to participate. Data used in this study was previously reported in (Hamilton et al., 2018).

702 Experimental Stimuli

703 Stimuli consisted of 499 English sentences from the TIMIT acoustic-phonetic corpus (Garofolo et al.,
704 1993), spoken by male and female speakers with a variety of North American accents. Stimuli were
705 presented through free-field Logitech speakers at comfortable ambient loudness (~70 dB), controlled
706 by a custom MATLAB script. Participants passively listened to the sentences in 4 blocks, each lasting
707 about 4 minutes. A subset of 438 sentences were selected for analysis that were heard once by all 11
708 subjects. The sentences had durations between 0.9 and 2.6s, with a 400ms intertrial interval.

709 Neural recordings and electrode localization

710 Neural recordings were acquired at a sampling rate of 3051.8 Hz using a 256-channel PZ2 amplifier or
711 512-channel PZ5 amplifier connected to an RZ2 digital acquisition system (Tucker-Davis Technologies,
712 Alachua, FL, USA).

713

714 Electrodes were localized by coregistering a preoperative T1 MRI scan of the individual subject's brain
715 with a postoperative CT scan of the electrodes in place. Freesurfer was used to create a 3d model of
716 the individual subjects' pial surfaces, run automatic parcellation to get individual anatomical labels,
717 and warp the individual subject surfaces into the cvs_avg35_inMNI152 average template (Desikan et
718 al., 2006; Fischl et al., 2004). More detailed procedures are described in (Hamilton et al., 2017).

719 Preprocessing

720 For each electrode, the high gamma amplitude time series were extracted from the broadband neural
721 recordings as follows (Hamilton et al., 2018; Oganian and Chang, 2019). First, the signals were
722 downsampled to 400 Hz, rereferenced to the common average in blocks of 16 channels (blocks shared
723 the same connector to the preamplifier), and notch filtered at 60, 120, and 180 Hz to remove line noise
724 and its harmonics. These LFP signals were then filtered using a bank of 8 Gaussian filters with center
725 frequencies logarithmically spaced between 70 and 150 Hz. Using the Hilbert transform, the amplitude

726 of the analytic signal was computed for each of these frequency bands, and for each electrode the high
727 gamma amplitude was defined as the first principal component across these 8 frequency bands.
728 Finally, the high gamma amplitude was further downsampled to 100Hz and z-scored based on the
729 mean and standard deviation across each experimental block.

730 Electrode selection

731 In order select speech-responsive electrodes over STG, electrodes were included (1) if they were
732 located over the STG, as identified in the Freesurfer anatomical parcellation of the individual subject
733 cortical surface, and (2) if their high gamma activity was well-predicted by a linear spectrotemporal
734 model (Hamilton et al., 2018).

735
736 For this single electrode analysis, the model had the form of a spectrotemporal receptive field (STRF):
737
$$y(t) = \sum_f \sum_{\tau} S(f, t - \tau) \beta(\tau, f) + e(t) \quad (7)$$

738 where y is the high gamma amplitude on a single electrode, S is the mel spectrogram of the speech
739 audio signal over frequencies between 75Hz and 8kHz, coefficients β vary across frequencies and
740 delays between 0 and 500ms, and e is the zero-mean Gaussian error term. Ridge regression was used
741 to fit the models (see Model fitting below for details of the ridge regression framework): the data were
742 split into 80% training and 20% testing data sets, the training data was used to choose the alpha
743 parameter according to a 5-fold cross-validation, the full training data was fit using the chosen α
744 parameter, and the r^2 was assessed on the testing data (see Explained Variance Calculation below for
745 computation of r^2). Electrodes with $r^2 > 0.05$ were included in subsequent analyses. The selected
746 electrodes and their corresponding r^2 values are shown in Figure 1A.

747 Regression model setup

748 The multivariate temporal receptive field model has the following structure:

$$749 Y = \sum_{f=1}^F X_f B_f + E \quad (8)$$

750

751 Where:

- 752 • Y is the $T \times N$ matrix of z-scored high gamma amplitude values across electrodes and timepoints.
753 The time dimension represents a concatenation of all 438 sentence stimuli that were heard by
754 every subject, from 500 ms before sentence onset until 500 ms after sentence offset (132,402
755 timepoints, later split for cross validation, see Model Fitting below). The electrode dimension
756 includes speech-responsive electrodes from all subjects (331 electrodes).
- 757 • Each X_f ($T \times D$) represents the delayed feature events for feature f . The first column contains
758 the feature events across time (1 representing an event occurring, 0 otherwise. For peak rate,
759 events were coded by a real-valued magnitude, see Figure 1B). Following columns contain the
760 same time series, offset by time-delays between 10 ms and 750 ms (76 delays). There were 12

761 features: sentence onset, peak rate, dorsal, coronal, labial, high, front, low, back, plosive,
762 fricative, and nasal (described below).

- 763 • E ($T \times N$) is Gaussian noise, assumed to be uncorrelated across electrodes
- 764 • B_f ($D \times N$) are the coefficient matrices, i.e. the multivariate temporal response functions
765 (MTRFs), representing the responses of each electrode to the given feature across electrodes and
766 delays
- 767 • T : number of timepoints; N : number of electrodes, D : number of delays, F : number of features.

768

769 Sentence onset was defined as the sound onset time for the sentence stimulus. Peak rate was
770 extracted by taking the derivative of the analytic envelope of the speech signal: the peak rate event
771 times were the times when the derivative reached a maximum, and the peak rate magnitude was the
772 value of the derivative at that time point (Oganian and Chang, 2019). Phonetic feature event times
773 (dorsal, coronal, labial, high, front, low, back, plosive, fricative, nasal) were extracted from time-aligned
774 phonetic transcriptions of the TIMIT corpus, which were timed to the onset of the respective
775 phonemes in the speech signal (Garofolo et al., 1993).

776 Model fitting

777 The model was fit using ordinary least squares (OLS), ridge regression, and iRRR. The difference
778 between the three is the objective function that is minimized to choose the fitted coefficient matrices:

779

$$780 \{\hat{B}_{f,OLS}\}_{f=1}^F = \operatorname{argmin}_{B_f \in \mathbb{R}^{D \times N}} \frac{1}{2T} \|Y - \sum_{f=1}^F X_f B_f\|_{\mathcal{F}^2} \quad (9a)$$

$$781 \{\hat{B}_{f,ridge}\}_{f=1}^F = \operatorname{argmin}_{B_f \in \mathbb{R}^{D \times N}} \frac{1}{2T} \|Y - \sum_{f=1}^F X_f B_f\|_{\mathcal{F}}^2 + \alpha \sum_{f=1}^F \|B_f\|_{\mathcal{F}}^2 \quad (9b)$$

$$782 \{\hat{B}_{f,iRRR}\}_{f=1}^F = \operatorname{argmin}_{B_f \in \mathbb{R}^{D \times N}} \frac{1}{2T} \|Y - \sum_{f=1}^F X_f B_f\|_{\mathcal{F}}^2 + \lambda \sum_{f=1}^F w_f \|B_f\|_* \quad (9c)$$

783

784 The weights used for the iRRR model were chosen to balance the different features (Li et al 2019):

$$785 w_f = \sigma(X_f, 1) \left\{ \sqrt{N} + \sqrt{r(X_f)} \right\} / T \quad (10)$$

786 where $\sigma(X_f, 1)$ is the first singular value of the matrix X_f and $r(X_f) = D$ is the rank of matrix X_f . All
787 predictors X_f and responses Y were column-centered before fitting the models.

788

789 In order to compute confidence intervals for model performance metrics, models were fit using 10-fold
790 cross validation, using group cross validation to keep time points corresponding to the same sentence
791 stimulus in the same fold. For ridge regression and iRRR, an additional nested 5-fold cross validation
792 was used to choose the α and λ parameters within each fold of the outer cross-validation.

793

794 Note that the approach of using a regression framework to fit a group-reduced rank model of neural
795 activity has been used before (Aoi et al., 2020; Aoi and Pillow, 2019): the iRRR framework differs in that
796 it uses an L1 relaxation, resulting in a convex optimization formulation that can be fit efficiently using
797 alternating direction method of multipliers.

798 Model performance metrics

799 Total explained variance (Figure 1C) was calculated as:

$$800 \quad r^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (11)$$

801 where the SS_{res} is the residual sum of squares computed on the testing dataset:

$$802 \quad SS_{res} = \|Y - \sum_{f=1}^F X_f B_f\|_{\mathcal{F}}^2 \quad (12)$$

803 and SS_{tot} is the total sum of squares computed on the testing dataset:

$$804 \quad SS_{tot} = \|Y\|_{\mathcal{F}}^2 \quad (13)$$

805

806 The group nuclear norm (Figure 1D) was computed as the penalty term in the iRRR model:

$$807 \quad \sum_{f=1}^F w_f \|B_f\|_* \quad (14)$$

808

809 Because OLS and ridge regression yield full-rank coefficient matrices, the number of parameters
810 (Figure 1E) used for both is DN . For iRRR, the number of parameters is $k(D + N + 1)$, based on the
811 singular value decomposition described in Equation 3, reproduced here:

$$812 \quad \hat{B}_f = U_f S_f V_f^T \quad (15)$$

813 Unique explained variance for each feature (Figure 1F) was computed by fitting a reduced iRRR model
814 without the feature f , and then comparing the total explained variance of the full model r_{Full}^2 to the
815 total explained variance of the reduced model r_{-f}^2 . The reduced iRRR model was fit using the same λ
816 value as the full model, chosen using nested cross validation on the full model as described above. For
817 the “all timing” category, the reduced model was fit without sentence onset and peak rate, and for the
818 “all phonetic” category, the reduced model was fit without the phonetic features. The unique
819 explained variance was expressed as a percentage of the full model:

$$820 \quad 100 \times \frac{r_{Full}^2 - r_{-f}^2}{r_{Full}^2} \quad (16)$$

821

822 All metrics are reported in terms of the mean across the 10 folds of the cross validation, and 95%
823 confidence intervals are $\pm t_{9,0.975} s / \sqrt{10}$, where s is the sample standard deviation across the 10 cross
824 validation folds. Note that these confidence intervals do not account for the dependence between
825 cross-validation folds due to reuse of samples in training and testing sets, and may therefore be
826 smaller than the true intervals (Austern and Zhou, 2020; Bates et al., 2021; Bengio and Grandvalet,
827 2004).

828

829 Significant differences between conditions were assessed using paired two-tailed t-tests across cross-
830 validation folds (Dietterich, 1998) for the following comparisons (with the resulting p-value ranges):

- 831 1. Total explained variance for OLS vs Ridge ($p > 0.05$), OLS vs iRRR ($p < 0.0005$), and Ridge vs iRRR
832 ($p < 0.0005$).
- 833 2. Unique explained variance of sentence onset vs each acoustic-phonetic feature and peak rate
834 vs each acoustic-phonetic feature. Here the p-values were Bonferroni corrected across the (2
835 timing features times 10 acoustic-phonetic features) 20 comparisons. After correction, all
836 comparisons were significant with $p < 0.0005$.
- 837 3. Unique explained variance of the combined timing features vs the combined acoustic-phonetic
838 features ($p < 0.0005$).

839 Similar to the confidence intervals described above, the significance tests did not account for the
840 dependence between cross-validation folds and may therefore have an inflated type II error (Austern
841 and Zhou, 2020; Bates et al., 2021; Bengio and Grandvalet, 2004).

842 Computing predicted responses

843 Given a fitted model, the predicted latent response to a stimulus matrix X_f is (reproduced from
844 Equation 5):

$$845 \hat{Y}_{f;latent} = X_f U_f S_f \quad (17)$$

846 where, as before, X_f ($T \times D$) represents the delayed feature events for feature f , U_f is the $D \times k$ time
847 components for feature f , and S_f is a diagonal matrix containing the weights for each component
848 ($k \times k$). $\hat{Y}_{f;latent}$ is a $T \times k$ matrix representing the predicted response within the k -dimensional
849 latent space of the feature. Figure 3 shows the predicted sentence onset and peak rate responses to
850 the sentence “They’ve never met, you know”.

851 jPCA

852 The plane of fastest rotation for the sentence onset and peak rate latent states (Figure 3C) was
853 identified by applying jPCA (Churchland et al., 2012) to the feature coefficient matrices \hat{B}_f . Using jPCA,
854 we modeled the temporal receptive fields in the coefficient matrix as a linear dynamical system
855 evolving over delays:

$$856 \frac{d\hat{B}_f(t)}{dt} = M \hat{B}_f(t) \quad (18)$$

857 where t indexes the delay dimension of \hat{B}_f , so the dynamical system describes the evolution of an N -
858 dimensional dynamical system over D timepoints. By approximating the derivative on the left hand
859 side using first differences, the transition matrix M can be fit using regression. Furthermore, the purely
860 rotational component of the transition matrix can be isolated by constraining the matrix M to be skew-
861 symmetric, having purely imaginary eigenvalues that come in complex conjugate pairs. The pair of
862 eigenvectors with the largest magnitude eigenvalues describes the plane with the fastest rotations.

863

864 It is important to note that jPCA identifies planes with fast rotational dynamics, regardless of whether
865 they capture a large proportion of the variance of the dynamics in the original dynamical system.
866 Classic jPCA uses PCA in preprocessing in order to confine the analysis to six dimensions of largest
867 variance. Here, the iRRR model chooses k dimensions for each feature that are most valuable to the
868 overall fit of the model. Hence there was no need to perform additional PCA to reduce the
869 dimensionality. However, because the coefficient matrices had dimensions capturing very little
870 variance, we did subselect components to capture 98% of the variance of the coefficient matrices. For
871 both sentence onset and peak rate, this corresponded to the top 3 components. Hence the jPCA plane
872 represents the plane of maximal rotation within a 3-dimensional subspace capturing 98% of the
873 variance in the 5-dimensional (or 6-dimensional) coefficient matrix for sentence onset (or peak rate). If
874 we had used more components for the jPCA computation, the rotational dynamics would be stronger
875 but they would capture much less of the variance (using k dimensions vs using 3 dimensions: 2.8% vs
876 31.8% for sentence onset and 4.8% vs 20.3% peak rate), making them less informative about the
877 overall population dynamics.

878

879 Once the jPCs were computed using the coefficient matrices, the predicted trajectory for a given
880 stimulus (Figure 3F and G) is calculated as:

$$881 \hat{Y}_{f,jPCA} = X_f J_f \quad (19)$$

$$882 J_f = [E_1 + E_2, j(E_1 - E_2)]$$

883 where E_1 and E_2 are the eigenvectors with largest eigenvalues of the skew-symmetric matrix M
884 defined above. J_f is therefore the $N \times 2$ projection matrix from electrode space onto the plane of
885 highest rotation from jPCA.

886 Event latency decoding

887 For the decoding analysis (Figure 4), a perceptron model was trained to predict the time relative to the
888 most recent feature event (up to 750 ms). The model was designed using the MLPRegressor class of
889 the sklearn package, with one hidden layer with 20 hidden units using a logistic activation function. We
890 used a simple perceptron model in order to account for possible nonlinearities in the mapping from
891 electrode space / feature latent space to relative times.

892

893 Using the same cross-validation framework that was used for iRRR model fitting, the perceptron model
894 was trained using the training data (high gamma amplitudes) either across all electrodes Y or using the
895 projected data onto the latent state subspace (reproduced from Equation 6):

$$896 \tilde{Y}_{f,proj} = Y V_f \quad (20)$$

897 where V_f is the $N \times k$ matrix of electrode components for feature f , as above. The $T \times k$ matrix

898 $\tilde{Y}_{f,proj}$ is an approximation of the latent state across time, but it may be contaminated by activity from

899 other features because the V_f matrices do not describe orthogonal subspaces. It also contains activity
900 from noise.

901

902 Performance of the models was assessed using r^2 (Equation 11) on the held-out testing data for the
903 cross-validation fold. The 95% confidence intervals were computed using the t distribution as described
904 above, and the performance of the models trained on all-electrodes was compared to the performance
905 of the models trained on the latent projections using a two-sided paired t test, as described above
906 (Model Performance Metrics), Bonferroni corrected across the 12 features. The sentence onset model
907 performed better using all electrodes than the latent projection (corrected $p < 0.05$), while the models
908 for all other features performed better using the latent projection than using all electrodes (corrected
909 $p < 0.05$).

910 Code availability

911 Custom Python code to perform the iRRR fits is available online
912 (https://github.com/emilyps14/iRRR_python), which is a port of the Matlab implementation by the
913 original authors (<https://github.com/reagan0323/iRRR>, Li et al 2019). Python code for the analysis
914 pipeline described above is also available (https://github.com/emilyps14/mtrf_python). We thank
915 Antin and colleagues (Antin et al., 2021) for their implementation of jPCA in the Python programming
916 language (<https://github.com/bantin/jPCA>), which we used to perform the jPCA.

917

918

919

920 **Supplementary Material**

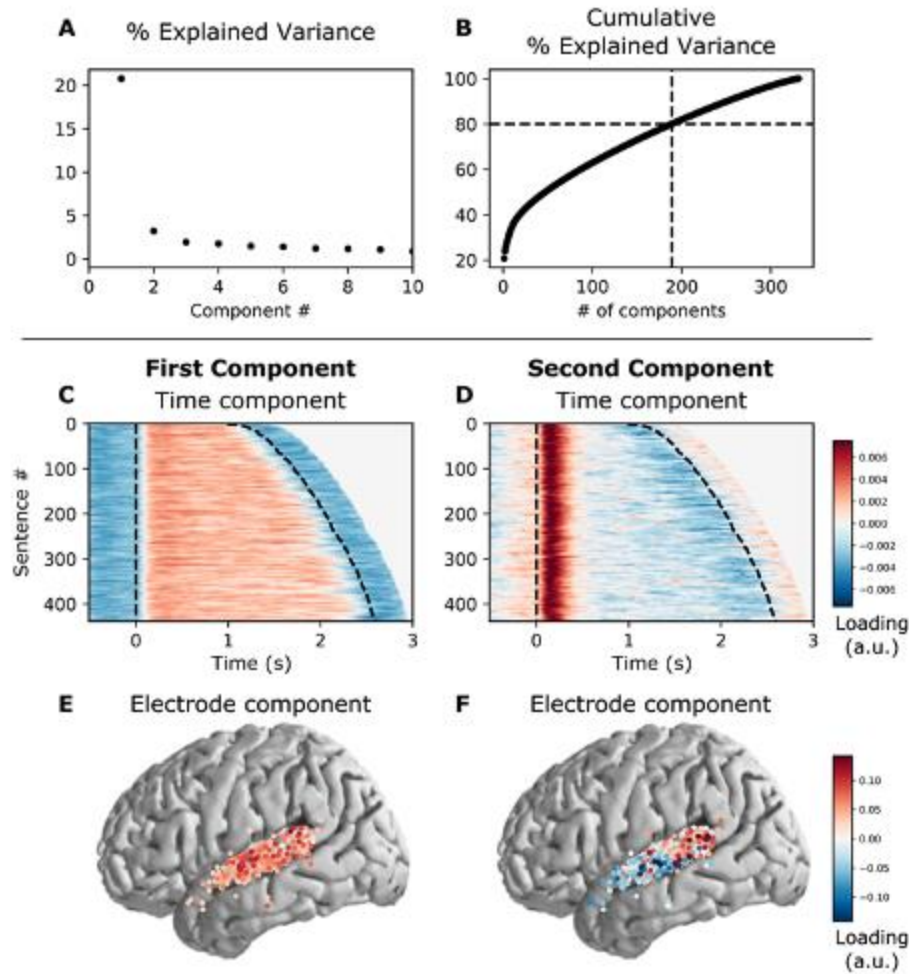
921 **Table S1. Clinical and demographic details for subjects. Hem = hemisphere of implantation.**

Subject ID	Hem	Age	Sex	Handedness	Language dominance	Epilepsy focus
SL01	L	44	M	R	L	Left posterior STG
SL02	L	19	M	R	L	Left anterior frontal lobe
SL03	L		M		L	Left anterior temporal lobe
SL04	L	32	F	R	L	Left anterior temporal lobe
SL05	L	25	F	L	L	Left medial temporal lobe
SL06	L	31	F	R	L	Left hippocampus/anterior lateral temporal
SL07	L	20	F	R	L	Left hippocampus
SL08	L	60	M	R (converted from L)	L	Left mesial temporal structures
SL09	L	26	M	R	L	Left mesial and anterior lateral temporal cortex
SL10	L	22	M	R	L	Left anterior temporal lobe
SL11	L	31	F	R	L	Left hippocampus/amygdala

922

923

924



925
926
927
928
929
930
931
932
933
934
935
936
937
938

Figure S1: PCA partitions the high gamma activity across speech-responsive electrodes into a posterior onset response and a spatially widespread sustained response. A: The percent explained variance of the principal components. B: The cumulative percent explained variance. Note that 189 dimensions are required to capture 80% of the variance in the high gamma activity. C: The timecourse of the first component, aligned to sentence onset. Dashed lines indicate the start and end of the sentence stimulus, and the sentences have been ordered by their duration. This component has sustained responses, in the sense that the activity is high during the entire stimulus. D: The timecourse of the second component, aligned to sentence onset. This component has onset responses, in the sense that there is a short positive transient immediately after sentence onset. E: The spatial support of the first component. This component is spatially spread out over all of STG. F: The spatial support of the second component. This component is spatially divided, with strong positive weights over posterior STG.