



HAL
open science

Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content

Laura Gwilliams, Jean-Rémi King, Alec Marantz, David Poeppel

► **To cite this version:**

Laura Gwilliams, Jean-Rémi King, Alec Marantz, David Poeppel. Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content. 2020. hal-03089733

HAL Id: hal-03089733

<https://hal.science/hal-03089733>

Preprint submitted on 29 Dec 2020

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.

Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content

*Laura Gwilliams^{1,2}, Jean-Remi King^{1,3}, †Alec Marantz^{1,2,4} & †David Poeppel^{1,5}

¹Department of Psychology, New York University,

²NYU Abu Dhabi Institute,

³École normale supérieure,

⁴Department of Linguistics, New York University,

⁵Max Planck Institute

Abstract

Listeners experience speech as a sequence of discrete words. However, the real input is a continuously varying acoustic signal that blends words and phonemes into one another. Here we recorded two-hour magnetoencephalograms from 21 subjects listening to stories, in order to investigate how the brain concurrently solves three competing demands: 1) processing overlapping acoustic-phonetic information while 2) keeping track of the relative order of phonemic units and 3) maintaining individuated phonetic information until successful word recognition. We show that the human brain transforms speech input, roughly at the rate of phoneme duration, along a temporally-defined representational trajectory. These representations, absent from the acoustic signal, are active earlier when phonemes are predictable than when they are surprising, and are sustained until lexical ambiguity is resolved. The results reveal how phoneme sequences in natural speech are represented and how they interface with stored lexical items.

Keywords: phonology, MEG, magnetoencephalography, auditory sequences, brain, language

One sentence summary

The human brain keeps track of the relative order of speech sound sequences by jointly encoding content and elapsed processing time

* Correspondence to: leg5@nyu.edu

† Denotes equal contribution

1 Speech comprehension involves mapping non-stationary, highly variable and continuous
2 acoustic signals onto discrete linguistic representations [1]. Although the human experience is
3 typically one of effortless understanding, the computational infrastructure underpinning speech
4 processing remains a major challenge for neuroscience [2] and artificial intelligence systems [3]
5 alike.

6 Existing cognitive models primarily serve to explain the recognition of words in isolation
7 [4, 5, 6]. Predictions of these models have gained empirical support in terms of neural encoding
8 of phonetic features [7, 8, 9, 10], and interactions between phonetic and (sub)lexical units of
9 representation [11, 12, 13, 14, 15]. What is not well understood, and what such models largely
10 ignore, however, is how sequences of acoustic-phonetic signals (e.g. the phonemes *k-a-t*) are
11 mapped to lexical items (e.g. *cat*) during comprehension of naturalistic continuous speech.

12 One substantial challenge is that naturalistic language does not come pre-parsed: there are,
13 e.g. no reliable cues for word boundaries, and adjacent speech sounds (phonemes) acoustically
14 overlap both within and across words due to co-articulation [1]. In addition, the same sequence
15 of phonemes can form completely different words (e.g. *pets* versus *pest*), so preserving phoneme
16 order is critical. Furthermore, phonemes elicit a *cascade* of neural responses, which long surpass
17 the duration of the phonemes themselves [16, 17, 9]). This means, concretely, that a given
18 phoneme_{*i*} is still present in both the acoustic and neural signals while subsequent phonemes
19 stimulate the cochlea. Such signal complexity presents serious challenges for the key goals of
20 achieving invariance and perceptual constancy in spoken language comprehension.

21 Based on decoding analyses of acoustic and neural data we show how the brain orchestrates
22 these overlapping inputs and overlapping neural processes, without confusing either the content
23 or order of the phoneme sequences. We address how the language system (i) simultaneously
24 processes acoustic-phonetic information of overlapping inputs; (ii) keeps track of the relative
25 order of those inputs; and (iii) maintains information sufficiently long enough to interface with
26 (sub)lexical representations.

27 1. Results

28 Neural responses were recorded with magnetoencephalography (MEG) while 21 participants
29 listened to four short stories. Each subject completed two one-hour recording sessions, yielding
30 brain responses to ~50,000 phonemes, ~10,000 words and ~1,000 sentences per subject (see
31 Figure 1A).

32 1.1. Phonetic feature encoding in acoustic and neural signals

33 First we tested when and how linguistic features of the speech input are encoded in acoustic
34 (spectro-temporal) and neural (MEG) signals. To this aim, we fit a ridge regression to decode
35 14 phonetic features (one-hot encoding of three distinctive features *place*, *manner* and *voicing*),
36 either from 50 frequency bands of the acoustic spectrogram (acoustic analysis) or from the 208
37 MEG channels (neural analysis). Using a 25-split cross validation loop, the model was trained
38 on responses to all phonemes from the training set, and then tested as a function of their relative
39 position in the words.

40 Figure 1B shows the outcome of these analyses. Although the average phoneme duration is
41 less than 80 ms (mean duration = 78 ms; SD = 34 ms), phonetic features (averaged over position)
42 can be decoded from the acoustic signal between 0-300 ms ($p < .001$; $\hat{t} = 9.56$), and between
43 50-300 ms in the neural signal ($p < .001$; $\hat{t} = 3.61$). This confirms that featural cues extend

44 to neighbouring phonemes. The ability to decode such phonetic features from the spectrum of
 45 the acoustics shows the existence of putatively invariant acoustic cues, which sufficiently generalise
 46 across phoneme locations [18, 19]. Furthermore, phonetic features that were more strongly
 47 encoded in the acoustic signal were also better decoded from the neural signal (Spearman correlation
 48 $r = .59$; $p = .032$; note that the large difference in decoding accuracy between acoustic and
 49 brain signals is expected given the signal-to-noise ratio of single-trial MEG recordings).

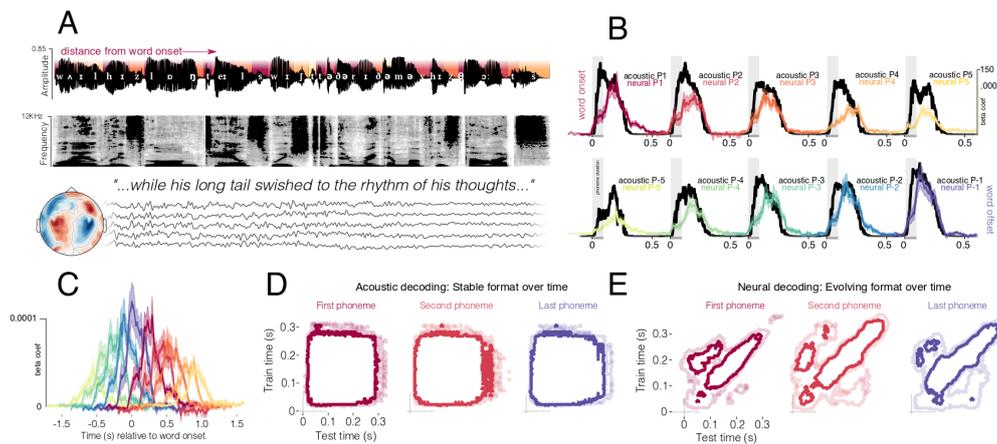


Figure 1: Experimental design and acoustic-phonetic analysis. A: Example sentence from the stories, with the parse into phonological units superimposed on the acoustic waveform. Colours of the segments at the top half of the waveform indicate the phoneme's distance from the beginning of a word (darker red at word onset). The spectrogram of the same sentence appears below the waveform. Five example sensor time-courses are shown below, indicating that all recordings of the continuous stories were recorded with continuous concurrent MEG. B: Time-course of phonetic-feature decoding accuracy. Black lines show accuracy of decoding features from the acoustic spectrogram. Coloured lines show results when decoding the same features from the MEG sensors. Shading in the neural data corresponds to the standard error of the mean across subjects. Results are plotted separately for 10 different phoneme positions, where P1:P5 indicates distance from word onset and P-1:P-5 distance from word offset. All plots share the same y-axis scales, which are different for neural and acoustic analyses (top right). C: The same neural decoding data are here overlaid, relative to the average duration between one phoneme and the next (around 80 ms). Multiple phonemes can be read out from the neural signal at the same time. D: Results of the temporal generalisation (TG) analysis on the acoustic data. The y-axis corresponds to the time that the decoder was trained, relative to phoneme onset; the x-axis corresponds to the time that the decoder was tested, relative to phoneme onset. The results are shown separately for three different phoneme positions. Contours represent 95% and 90% percentile decoding accuracy. E: Results of the same TG analysis applied to the MEG data, showing a very different dynamic profile from the acoustic analysis. Contours represent 95% and 90% percentile decoding accuracy.

50 1.2. Rapidly evolving neural representations

51 On average, phonetic features were linearly decodable for three times longer than the duration
 52 of the phoneme itself. This suggests that, at any one time, three phonemes are being processed
 53 concurrently (Figure 1C). How does the brain implement this set of parallel computations and
 54 prevent interference between the resulting content?

55 We tested whether the pattern of neural activity (from the MEG analysis) or the combination
 56 of spectro-temporal features (from the acoustic analysis) remained stable with respect to discrim-
 57 inability using temporal generalisation analysis [20]. This reveals whether a given representation
 58 evolves or is transformed during processing.

59 For the acoustic analysis (Figure 1D), there was significant generalisation, leading to no sta-
60 tistical differences between the accuracy time-course of a single decoder, as compared to inde-
61 pendent decoders at each time sample ($p = .51$; $\hat{\tau} = -.67$). This ‘square’ temporal generalisation
62 suggests that although the acoustic signals are transient and dynamic, they contain stationary
63 cues for acoustic-phonetic features. By contrast, the underlying representations of neural in-
64 formation evolved rapidly over time (Figure 1E). Concretely, any particular topographic pattern
65 was informative to read out a phonetic feature for around 80 ms, whereas the duration of the
66 entire dynamic process lasted around 300 ms. This was confirmed using an independent samples
67 t-test, comparing diagonal and horizontal decoding performance ($p < .001$; $\hat{\tau} = 7.54$). Neural and
68 acoustic dynamics did not change as a function of phoneme position.
69 Practically speaking, that the neural responses show a diagonal rather than square generali-
70 sation pattern means the underlying activations are evolving over time: activity supporting the
71 processing of a particular phonetic feature is either moving across cortical regions or evolving or
72 transforming within a particular cortical region.

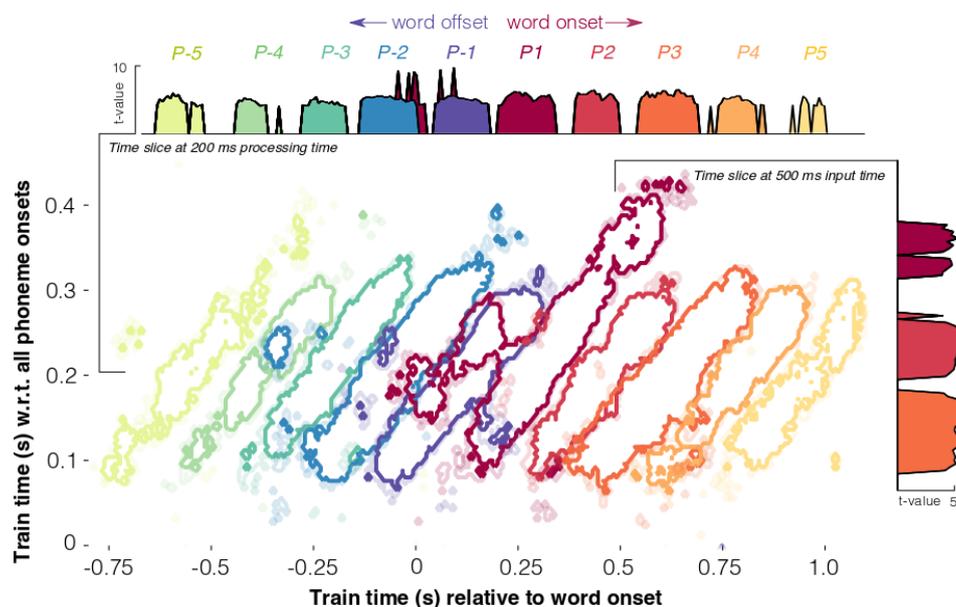


Figure 2: **Phonetic feature processing across the sequence.** Temporal generalisation (TG) results superimposed for 10 phoneme positions. From word onset (P1, dark red), and from word offset (P-1, dark blue). The result for each phoneme position is shifted by the average duration from one phoneme to the next. The y-axis corresponds to the time that the decoder was trained, relative to phoneme onset. The x-axis corresponds to the time that the decoder was tested, relative to (superimposed) word onset. Contours represent a t-value threshold of 4 (darker lines) and 3.5 (lighter lines). The panel at the top shows the t-values for each phoneme at a training time of 200 ms, showing that there is no representational overlap between neighbouring phonemes. The panel on the right shows a slice of time at 500 ms after the onset of the first phoneme. This shows that at a single moment relative to input time, multiple (at least three) phonemes can be read out of the neural responses.

73 1.3. Phonetic sequence processing

74 The above results show that while multiple phonemes are represented simultaneously in the
75 brain, their dynamic encoding scheme may minimise the overlap between neighbouring repre-
76 sentations. To test this hypothesis, we aligned the temporal generalisation matrices relative to the
77 average latency between two adjacent phonemes. For example, relative to word onset, phoneme
78 P1 is plot at $t=0$, P2 at $t=80$, P3 at $t=160$, etc. We extracted the time-samples that exceeded a $p <$
79 $.05$ threshold, Bonferroni-corrected across the 201 time-samples of a single processing time. We
80 then computed the relative overlap between the time-samples of one phonemic unit and another.
81 As shown in Figure 2, there is virtually no overlap when the data are examined at a particular
82 processing time (horizontal axis). Crucially, this suggests that although multiple phonemes are
83 processed in parallel, any given pattern of neural activity only represents one phoneme at a time,
84 allowing each phoneme an individuated representation.

85 1.4. Representations are stable for the phoneme duration

86 Although the results show a clear evolution of representational format, each underlying neural
87 pattern remained stable for ~ 80 ms, i.e. average phoneme duration. To test whether this
88 maintenance scales with phoneme duration, we grouped trials into quartiles, and analysed brain
89 responses to the shortest and longest phonemes (~ 4500 trials in each bin; mean duration 45 and
90 135 ms). Phoneme duration correlated with the duration of temporal generalisation across training
91 time: longer phonemes generalised for an average 56 ms longer than shorter phonemes ($p =$
92 $.005$; $\hat{t} = -2.6$) (Figure 3A).

93 1.5. Representations are shared across phoneme positions: invariance

94 Next we tested whether the same representational transformation is applied regardless of
95 phoneme position. For this, we trained a classifier on the phonetic features of word onset
96 phonemes and then tested this decoder on responses to the second, third and last phonemes
97 (Figure 3B). We could read out the features of all three phoneme positions from 20-270 ms ($p <$
98 $.001$; $\hat{t} = 3.3$), with comparable performance, thus supporting the position-invariant encoding of
99 phonetic features.

100 1.6. Phonetic processing is modulated by word boundaries

101 How does the brain interface phonemic sequences with (sub)lexical representations (mor-
102 phemes or words)? To address this issue, we evaluated decoding performance at word bound-
103 aries: word onset (position P1) and word offset (position P-1) separately for each family of
104 phonetic features (place of articulation, manner, and voicing) (Figure 3C).

105 Phonetic features were decodable earlier at word onset than offset, yielding a significant
106 difference during the first 250 ms (place: $p = .03$, $\hat{t} = 2.77$, 84-112 ms; $p < .001$, $\hat{t} = -2.8$, 156-
107 240 ms; manner $p < .001$, $\hat{t} = 3.03$, 72-196 ms; $p = .004$, $\hat{t} = 2.75$, 220-300ms). The latency
108 between average neural and acoustic maximum accuracy was 136 ms (SD = 13 ms) at word onset
109 and 4 ms (SD = 13 ms) at word offset (see Figure 1B), leading to a significant difference between
110 onset and offset phonemes averaged over phonetic features ($t = -3.08$; $p = .002$). Furthermore,
111 place and voicing features were sustained in the neural signal significantly longer for phonemes
112 at the beginning of words as compared to the end (place: $p = .009$, $\hat{t} = -3.05$, 302-418 ms;
113 voicing: $p < .001$, $\hat{t} = -3.76$, 328-428 ms). This was also true when averaging over all features (p
114 $< .001$, $\hat{t} = -3.79$, 328-396 ms) (see Figures 1B and 2).

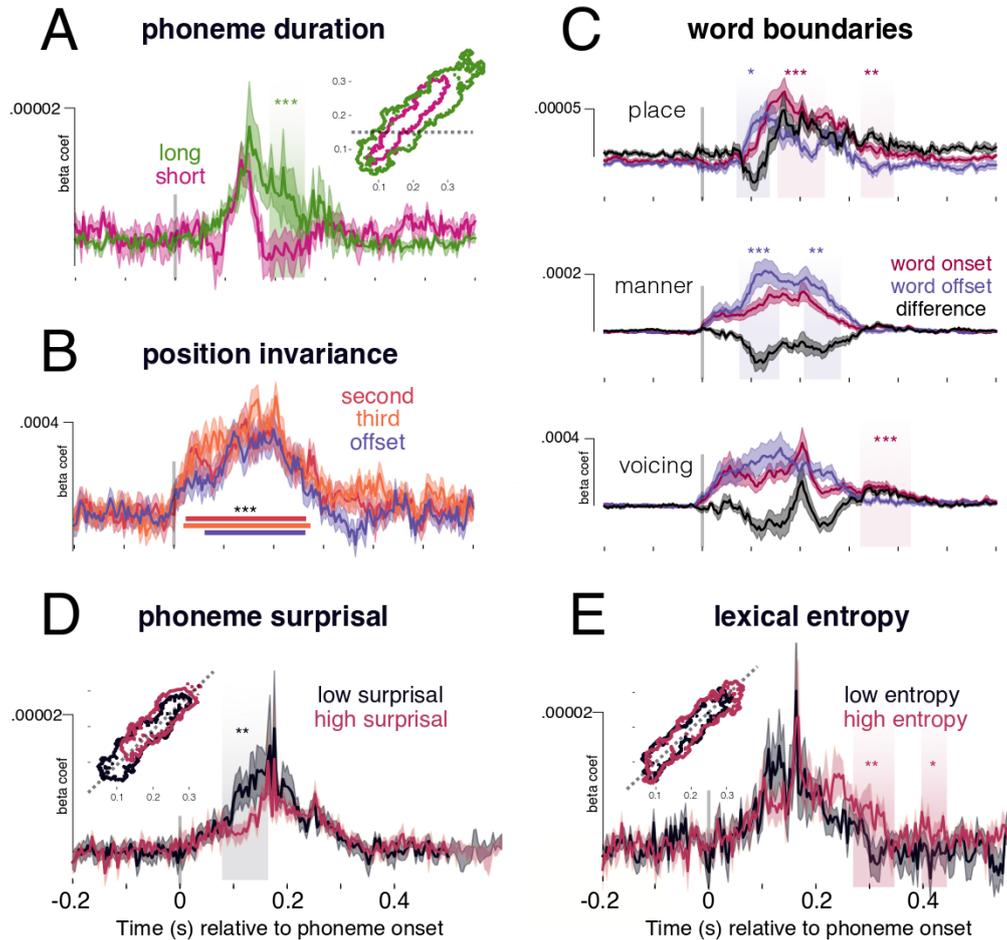


Figure 3: **Elucidating sequence dynamics.** A: TG analysis median split into short (average 45 ms) and long (average 135 ms) phonemes. Contour inlay represents the borders of the significant clusters at $p < .001$. Waveforms represent a horizontal slice at 140 ms (shown as a dashed line in the contour plot). B: Decoding performance when training on responses to word onset, and evaluating on second, third and last phoneme in the word. Lines represent against-chance temporal clusters exceeding $p < .001$ for the three phoneme positions. C: Analysing responses along the diagonal plane for phonemes at word onset (dark red) and offset (dark blue) and their subtraction (black), split into the three families of phonetic features. Coloured shading corresponds to significant clusters found using applying a temporal cluster test to the difference wave. D: Analysis on all non-onset phonemes split into median surprisal, along the diagonal plane (slice shown in contour plot). Highlighted areas show significant temporal clusters between low and high surprisal. E: Analysis on all non-onset phonemes split into median cohort entropy, also along the diagonal plane. Highlighted areas show significant temporal clusters between low and high entropy. Shading on the waveform of all plots represents standard error of the mean across subjects. * = $p < .05$; ** = $p < .01$; *** = $p < .001$.

115 *1.7. Predictable phonemes are processed earlier*

116 We hypothesised that the latency shift observed at word boundaries may be due to the pre-
 117 dictability of each phoneme. Specifically, we tested whether expected phonemes could be de-

118 coded earlier from the neural signal than surprising ones. To control for co-linearity between
119 word-boundaries and surprisal, we selected all phonemes that were *not* at word onset, and tested
120 decoding accuracy as a function of quartile phoneme surprisal (see Methods for details on how
121 this variable was computed). Each analysis bin contained ~4500 trials, with a mean surprisal of
122 0.12, 1.38, 2.8 and 5.16 bits.

123 There was a systematic latency shift as a function of non-onset phoneme surprisal (Figure
124 3D): more predictable phonemes were decoded earlier than less predictable ones, leading to a
125 significant difference between low and high surprisal from 120-132 ms ($p = .007$). Surprisal did
126 not significantly modulate peak decoding accuracy (all uncorrected p -values $> .2$).

127 1.8. *Phonetic features are maintained until lexical identification*

128 Finally, to investigate whether phonetic representations are maintained until word recogni-
129 tion, we test two hypotheses: (i) phonetic features are maintained until the identification of a
130 word boundary; or (ii) they are maintained until certainty about word identity surpasses a partic-
131 ular threshold. We evaluated whether decoding performance between 300-420 ms (the window
132 that showed the word onset/offset effect) was better explained by word length or by word cer-
133 tainty (entropy over possible lexical candidates).

134 For (i), we compared decoding performance of word-onset phonemes, grouped into med-
135 ian word length (shorter mean length = 2.54 phonemes; 4058 trials; longer mean length =
136 5.2 phonemes; 2841 trials). No significant differences between groups were found (all clusters p
137 $> .2$, duration < 2 ms). For (ii), we grouped trials based on non-onset cohort entropy, and ran the
138 analysis on all phonemes that did *not* occur at word onset (~4500 trials per bin, mean values of
139 0.03, 0.77, 2.04 and 4.48 bits). In the window of interest, higher entropy phonemes we decoded
140 with significantly higher performance (304-328 ms, $p = .002$, $\hat{t} = -2.12$) (see Figure 3E). This
141 suggests that phonetic information is maintained for longer in cases of higher lexical uncertainty.

142 2. Discussion

143 How is the rapidly unfolding speech signal transformed into a sequence of discrete linguistic
144 units? We analysed MEG responses to continuous speech, as a function of phonetic features
145 and position in the phoneme sequence. Our results show that although both acoustic cues and
146 neural processes overlap in time (lasting around 300 ms), the underlying representation evolves
147 at the phonemic unit rate (around 80 ms), thus ensuring non-overlapping representations to avoid
148 interference. Furthermore, we demonstrate that features are processed earlier when the phoneme
149 is predictable – and then maintained until lexical identity is resolved. Taken together, our results
150 show that a highly dynamic and adaptive language system underpins phonological and lexical
151 processing during naturalistic listening.

152 The stationarity of the acoustic signal *versus* the dynamics of the corresponding neural rep-
153 resentations highlight that speech driven-responses are more than just a reflection of the acoustic
154 signals [21]. These dynamical representations allow the brain to process multiple (at least three)
155 successive phonemes simultaneously, without blending them within a common activity pattern.
156 This grants two computational advantages. First, it serves to avoid interference between phonetic
157 units, by ensuring an orthogonal format in processing space. This answers the key question of
158 how overlapping sequences are maintained without confusing the content of the signal. Second,
159 relative position is implicitly coded in the representational format of each phoneme at a given
160 input time. This allows the system to keep track of the order of speech sounds, i.e. to know that
161 you were asked to *teach* and not *cheat*, or that you are eating *melons* and not *lemons*.

162 Building on this observation, we found that the representational trajectory is consistent across
163 phoneme positions, thus leading to significant generalisation from one phoneme position to an-
164 other, with comparable magnitude. Although a simple result, this rules out a number of compet-
165 ing hypotheses. First, it is hard to reconcile these results with an explicit sequence representation.
166 For example, if the brain represents a sequence of all elapsed phonemes, the representation of
167 phoneme *X* at word onset would generalise poorly to third position *ABX* and even worse to sixth
168 position *ABCDEX*. Second, under the same logic, this result rules out the idea that phonemes have
169 a context-dependent encoding scheme, such as being represented along with their co-articulatory
170 neighbours [22]. In that case, phoneme *X* would have a different representation in the context
171 *AXB* and *VXY*. Finally, generalisability is inconsistent with position-specific encoding accounts,
172 such as edge-based schemes [23, 24], which would posit that *X* is encoded differently in *ABX* and
173 *XBC*. Instead, our results support a context-independent account, which encodes distance from
174 phoneme onset, regardless of lexical edges.

175 If all phonemes follow a common representational trajectory relative to phoneme onset, how
176 can we describe the transformational space? One possibility links to articulatory phonology
177 [25]. Although we showed the input acoustic space to be (surprisingly) static, the articulatory
178 gestures which produce those acoustics are inherently dynamic [26]. It is plausible, therefore,
179 that speech sounds are processed via articulatory commands, which are believed to jointly encode
180 both the sound being produced and the temporal delay relative to articulatory onset. This idea
181 of joint content-temporal coding resonates with recent findings of sequence encoding – finding
182 evidence for dedicated temporal codes in rat hippocampus [27]. Future work will need to further
183 delineate the temporal code used for phonological processing, and how the spatial location of
184 these responses changes as a function of processing time.

185 A critical finding is that the representational trajectory gets systematically delayed as a func-
186 tion of phonological uncertainty (surprisal) and systematically sustained as a function of lexical
187 uncertainty (cohort entropy). This suggests that the language system continuously adapts its
188 processes based on information across multiple levels of linguistic description simultaneously.

189 The latency shift for more predictable phonemes straightforwardly aligns with models of
190 predictive coding [28, 29] and analysis-by-synthesis [30]: when predictability for a phoneme is
191 strong, processes can be initiated earlier (perhaps in some cases before the sensory input) than
192 when the phoneme identity is unknown. Although previous work has shown that processing of
193 the speech signal is sensitive to phoneme probability within a word [31, 11, 13, 14, 32], this is
194 the first study quantifying the *consequences* this has for encoding the content of those phonemes.
195 Interestingly, we did not observe an effect of predictability on overall decoding performance,
196 suggesting that processing delays may serve as a compensatory mechanism to allow more in-
197 formation to be accumulated in order to reach the same strength of encoding [33]. Future work
198 should test whether this local (within-word) predictability metric has similar consequences to
199 global (across-word) metrics.

200 The finding that phonetic features are maintained longer in the face of lexical ambiguity is
201 a critical piece of the puzzle for understanding the interface between acoustic-phonetic repre-
202 sentations and the mental lexicon. This result not only highlights the adaptivity of the speech
203 processing system but also demonstrates the online bi-directional interaction between hierarchi-
204 cal levels of processing. Our results suggest that acoustic-phonetic information is maintained
205 until the (sub)lexical identity reaches a confidence threshold. To our knowledge, this is the first
206 evidence for active maintenance of phonetic information until statistically-defined boundaries,
207 and has clear processing advantages in the face of phonological ambiguity and lexical revision
208 [15].

209 Overall, our results reveal that the brain implements an elegant computational solution to
210 the processing of rapid, overlapping phoneme sequences. Namely, that the phonetic content of
211 the unfolding speech signal is jointly encoded with elapsed processing time. Future work will
212 need to assess the generality of this computational framework, and whether it subserves sequence
213 processing across other modalities and domains.

214 **3. Method**

215 *3.1. Participants*

216 Twenty-one native English participants were recruited from the NYU Abu Dhabi community
217 (13 female; age: $M=24.8$, $SD=6.4$). All provided their informed consent and were compensated
218 for their time. Participants reported having normal hearing and no history of neurological disorders.
219 Each subject participated in the experiment twice. Time between sessions ranged from 1
220 day to 2 months.

221 *3.2. Stimulus development*

222 Four fictional stories were selected from the Open American National Corpus: Cable spool
223 boy (about two bothers playing in the woods); LW1 (sci-fi story about an alien spaceship trying
224 to find home); Black willow (about an author struggling with writer's block); Easy money (about
225 two old friends using magic to make money).

226 Stimuli were annotated for phoneme boundaries and labels using the 'gentle aligner' from
227 the Python module *lowerquality*. Some prior testing provided better results than the Penn Forced
228 Aligner.

229 Each of the stories were synthesised using the Mac OSX text-to-speech application. Three
230 synthetic voices were used (Ava, Samantha, Allison). Voices changed every 5-20 sentences. The
231 speech rate of the voices ranged from 145-205 words per minute, which also changed every 5-20
232 sentences. The silence between sentences randomly varied between 0-1000 ms.

233 *3.3. Procedure*

234 Before the experiment proper, the participant was exposed to 20 seconds of each speaker
235 explaining the structure of the experiment. This was designed to help the participants attune to
236 the synthetic voices.

237 The order of stories was fully crossed using a Latin-square design. Participants heard the
238 stories in the same order during both the first and second sessions.

239 Participants answered a two-choice question on the story content every ~3 minutes. For ex-
240 ample, one of the questions was "what was the location of the bank that they robbed"? The
241 purpose of the questions was to keep participants attentive and to have a formal measure of en-
242 gagement. All participants performed this task at ceiling, with an accuracy of 98%. Participants
243 responded with a button press.

244 Stimuli were presented binaurally to participants through tube earphones (Aero Technolo-
245 gies), at a mean level of 70 dB SPL. The stories ranged from 8-25 minutes, with a total running
246 time of ~1 hour.

247 *3.4. MEG acquisition*

248 Marker coils were placed at five positions to localise each participant's skull relative to the
249 sensors. These marker measurements were recorded just before and after the experiment in order
250 to track the degree of movement during the recording.

251 MEG data were recorded continuously using a 208 channel axial gradiometer system (Kanazawa
252 Institute of Technology, Kanazawa, Japan), with a sampling rate of 1000 Hz and applying an on-
253 line low-pass filter of 200 Hz.

254 *3.5. Preprocessing MEG*

255 The raw MEG data were noise reduced using the Continuously Adjusted Least Squares
256 Method (CALM: (Adachi et al., 2001)), with MEG160 software (Yokohawa Electric Corporation
257 and Eagle Technology Corporation, Tokyo, Japan).

258 The data were bandpass-filtered between 0.1 and 50 Hz using MNE-Python's default param-
259 eters with firwin design [34] and downsampled to 250 Hz. Epochs were segmented from 200
260 ms pre-phoneme onset to 600 ms post-phoneme onset. No baseline correction was applied. No
261 other data cleaning was performed.

262 *3.6. Preprocessing auditory signals*

263 We computed a time-frequency decomposition of the auditory signals by applying a 100-
264 sample Hamming window to the auditory waveform. This resulted in a power estimate at each of
265 50 linearly spaced frequency bands from 1-11250 Hz. These data were then also downsampled
266 to 250 Hz, and segmented from 200-600 ms in order to match the dimensionality and size of the
267 MEG epochs.

268 *3.7. Modeled features*

269 We investigated whether single-trial sensor responses varied as a function of fourteen binary
270 phonetic features, as derived from the multi-value feature system reported in [35]. Note that
271 this feature system is sparse relative to the full set of distinctive features that can be identified
272 in English; however, it serves as a reasonable approximation of the phonemic inventory for our
273 purposes.

274 *Voicing.* This refers to whether the vocal chords vibrate during production. For example, this is
275 the difference between *b* versus *p* and *z* versus *s*.

276 *Manner of articulation.* Manner refers to the way by which air is allowed to pass through the
277 articulators during production. Here we tested five manner features: fricative, nasal, plosive,
278 approximant, and vowel.

279 *Place of articulation.* Place refers to where the articulators (teeth, tongue, lips) are positioned
280 during production. For vowels, this consists of: central vowel, low vowel, mid vowel, high
281 vowel. For consonants, this consists of: coronal, glottal, labial and velar.

282 *Nuisance variables.* In the same model, we also accounted for variance explained by 'nuisance
283 variables' – i.e. structural and statistical co-variables of the phonemes. Though we were not
284 interested in interpreting the results of these features, we included them in the model to be sure
285 that they did not account for our main analysis on the phonetic features. These features included:
286 primary stress, secondary stress, frequency of the sequence, suffix onset, prefix onset, root onset,
287 syllable location in the word, and syllable onset. These features were extracted from the English
288 Lexicon Project [36].

289 *Subset variables.* Throughout the analysis, we subset trials based on their relationship to: word
290 onset, word offset, surprisal, entropy, distance from onset, distance from offset.

291 Surprisal is given as:

$$P(w|C) = \frac{f(w)}{\sum_{w \in C} f(w)} \quad (1)$$

292 and cohort entropy is given as:

$$-\sum_{w \in C} P(w|C) \log_2 P(w|C) \quad (2)$$

293 where C is the set of all words consistent with the heard sequence of phonemes thus far, and
294 $f(w)$ is the frequency of the word w . Measures of spoken word frequency were extracted from
295 the English Lexicon Project [36].

296 3.8. Decoding

297 Decoding analyses were performed separately on the acoustic signal and on the neural signal.
298 For the acoustic decoding, the input features were the power estimates at each of the 50 frequency
299 bands from 1-1125 Hz. For the neural decoding, the input features were the magnitude of activity
300 at each of the 208 MEG sensors. This approach allows us to decode from multiple, potentially
301 overlapping, neural representations, without relying on gross modulations in activation strength
302 [37].

303 Because some of the features in our analysis are correlated with one another, we need to
304 jointly evaluate the accuracy of each decoding model relative to its performance in predicting
305 all modelled features, not just the target feature of interest. This is because, if evaluating each
306 feature independently, we will not be able to dissociate the decoding of feature f from the de-
307 coding of the correlated feature \hat{f} . The necessity to use decoding over encoding models here,
308 though (which, do not suffer so harshly from the problem of co-variance in the stimulus space)
309 is one of signal to noise: we expect any signal related to linguistic processes to be contained in
310 low-amplitude responses that are distributed over multiple sensors. Our chances of uncovering
311 reliable responses to these features is boosted by using multi-variate models [37].

312 To overcome the issue of co-variance, but still to capitalise on the advantages of decoding
313 approaches, we implement a back-to-back ridge regression model [38]. This involves a two
314 stage process. First, a ridge regression model was fit on a random (shuffled) half of the data,
315 at a single time-point. The mapping was learnt between the multivariate input (either activity
316 across sensors or power over frequency bands) and the univariate stimulus feature (one of the 31
317 features described above). All decoders were provided with data normalised by the mean and
318 standard deviation in the training set:

$$\operatorname{argmin}_{\beta} \sum_i (y_i \beta^T X_i)^2 + \alpha \|\beta\|^2 \quad (3)$$

319 where $y_i \in \{\pm 1\}$ is the feature to be decoded at trial i and X_i is the multivariate acoustic or neural
320 measure. The l2 regularisation parameter α was also fit, testing 20 log-spaced values from 1^{-5} to
321 1^5 . This was implemented using the *RidgeCV* function in *scikit-learn* [39].

322 Then, we use the other half of the acoustic or neural responses to generate a prediction for
323 each of the 31 features corresponding to the test set. However, because the predictions are cor-
324 related, we need to jointly-evaluate the accuracy of decoding each feature, to take into account

325 the variance explained by correlated non-target features. To do this, we fit another ridge regression model, this time learning the beta coefficients that map the matrix of *true* feature values to
326 *predicted* feature values:
327

$$\underset{\beta}{\operatorname{argmin}} \sum_i (y_i \beta^T \hat{Y}_i)^2 + \alpha \|\beta\|^2 \quad (4)$$

328 where $y_i \in \{\pm 1\}$ is the ground truth of a particular stimulus feature at trial i and \hat{Y}_i is the prediction
329 for all stimulus features. A new regularisation parameter α was learnt for this stage. By including
330 all stimulus features in the model, this accounts for the correlation between the feature of interest
331 and the other features. From this, we use the beta-coefficients that map the true stimulus feature
332 to the predicted stimulus feature.

333 The train/test split was performed 100 times, and the beta-coefficients were averaged across
334 iterations. This circumvents the issue of unstable coefficients when modelling correlated variables.
335 These steps were applied to each subject independently.

336 3.9. Temporal generalisation decoding

337 Temporal generalization (TG) consists of testing whether a temporal decoder fit on a training
338 set at time t can decode a testing set at time t' [20]. This means that rather than evaluating
339 decoding accuracy just at the time sample that the model was trained on, we evaluate its accuracy
340 across all possible train/testing time combinations.

341 TG can be summarised with a square training time \times testing time decoding matrix. To quantify
342 the stability of neural representations, we measured the duration of above-chance generalization
343 of each temporal decoder. To quantify the dynamics of neural representations, we compared
344 the mean duration of above-chance generalization across temporal decoders to the duration of
345 above-chance temporal decoding (i.e. the diagonal of the matrix versus its rows). These two
346 metrics were assessed within each subject and tested with second-level statistics across subjects.

347 3.10. Comparing decoding performance between trial subsets

348 We apply analyses that rely on comparing decoding performance for different subsets of
349 trials (e.g. between high/low surprisal, or beginning/end of word). We conduct this analysis
350 by first training our decoding models on responses to all phonemes, thus yielding a set of fit
351 model weights (a *topographic pattern*) at each millisecond relative to phoneme onset. We then
352 separately evaluate the performance of these decoders on the subset trials of interest. This yields
353 a time-course or generalisation matrix for each group of trials that we evaluate on.

354 3.11. Group statistics

355 In order to evaluate whether decoding performance is better than chance, we perform second-
356 order statistics. This involves testing whether the distribution of beta coefficients across subjects
357 significantly differs from chance (zero) across time using a one-sample permutation cluster test
358 with default parameters specified in the MNE-Python package [34].

359 4. Acknowledgements

360 We thank Graham Flick for help with data collection. We are very grateful to William Id-
361 sardi, Arianna Zuanazzi, Joan Opella, Pablo Ripolles-Vidal and Omri Raccach for their feedback
362 on an earlier version of the manuscript. **Funding:** This project received funding from the Abu
363 Dhabi Institute G1001 (AM); NIH R01DC05660 (DP), European Union's Horizon 2020 research
364 and innovation program under grant agreement No 660086, the Bettencourt-Schueller Founda-
365 tion, the Fondation Roger de Spoelberch, the Philippe Foundation (JRK) and The William Orr
366 Dingwall Dissertation Fellowship (LG). **Author contributions:** LG: conceptualisation; method-
367 ology; software; validation; formal analysis; investigation; data curation; writing - original draft
368 preparation and review and editing; visualisation. JRK: conceptualisation; methodology; soft-
369 ware; supervision. AM: conceptualisation; writing - review and editing; supervision; funding
370 acquisition. DP: conceptualisation; writing - review and editing; supervision; funding acquisi-
371 tion. **Competing interests:** The authors declare no competing interests. **Data and materials**
372 **availability:** Preprocessed data will be made available after publication.

373 References

- 374 [1] Pisoni, D. B. & Luce, P. A. Acoustic-phonetic representations in word recognition. *Cognition* **25**, 21–52 (1987).
375 [2] Wöstmann, M., Fiedler, L. & Obleser, J. Tracking the signal, cracking the code: Speech and speech comprehension
376 in non-invasive human electrophysiology. *Language, Cognition and Neuroscience* **32**, 855–869 (2017).
377 [3] Benzeghiba, M. *et al.* Automatic speech recognition and speech variability: A review. *Speech communication* **49**,
378 763–786 (2007).
379 [4] Marslen-Wilson, W. D. & Welsh, A. Processing interactions and lexical access during word recognition in contin-
380 uous speech. *Cognitive psychology* **10**, 29–63 (1978).
381 [5] McClelland, J. L. & Elman, J. L. The trace model of speech perception. *Cognitive psychology* **18**, 1–86 (1986).
382 [6] Norris, D. Shortlist: A connectionist model of continuous speech recognition. *Cognition* **52**, 189–234 (1994).
383 [7] Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal
384 gyrus. *Science* **343**, 1006–1010 (2014).
385 [8] Chang, E. F. *et al.* Categorical speech representation in human superior temporal gyrus. *Nature neuroscience* **13**,
386 1428 (2010).
387 [9] Khalighinejad, B., da Silva, G. C. & Mesgarani, N. Dynamic encoding of acoustic features in neural responses to
388 continuous speech. *Journal of Neuroscience* **37**, 2176–2185 (2017).
389 [10] Yi, H. G., Leonard, M. K. & Chang, E. F. The encoding of speech sounds in the superior temporal gyrus. *Neuron*
390 **102**, 1096–1110 (2019).
391 [11] Gwilliams, L. & Marantz, A. Non-linear processing of a linear speech stream: The influence of morphological
392 structure on the recognition of spoken arabic words. *Brain and language* **147**, 1–13 (2015).
393 [12] Leonard, M. K., Baud, M. O., Sjerps, M. J. & Chang, E. F. Perceptual restoration of masked speech in human
394 cortex. *Nature communications* **7**, 1–9 (2016).
395 [13] Gwilliams, L., Poeppel, D., Marantz, A. & Linzen, T. Phonological (un) certainty weights lexical activation. *arXiv*
396 *preprint arXiv:1711.06729* (2017).
397 [14] Brodbeck, C., Hong, L. E. & Simon, J. Z. Rapid transformation from auditory to linguistic representations of
398 continuous speech. *Current Biology* **28**, 3976–3983 (2018).
399 [15] Gwilliams, L., Linzen, T., Poeppel, D. & Marantz, A. In spoken word recognition, the future predicts the past.
400 *Journal of Neuroscience* **38**, 7585–7599 (2018).
401 [16] Picton, T. W., Woods, D. L., Baribeau-Braun, J. & Healey, T. M. Evoked potential audiometry. *J Otolaryngol* **6**,
402 90–119 (1977).
403 [17] Näätänen, R. & Picton, T. The n1 wave of the human electric and magnetic response to sound: a review and an
404 analysis of the component structure. *Psychophysiology* **24**, 375–425 (1987).
405 [18] Stevens, K. N. & Blumstein, S. E. Invariant cues for place of articulation in stop consonants. *The Journal of the*
406 *Acoustical Society of America* **64**, 1358–1368 (1978).
407 [19] Blumstein, S. E. & Stevens, K. N. Acoustic invariance in speech production: Evidence from measurements of
408 the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America* **66**, 1001–1017
409 (1979).
410 [20] King, J. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method.
411 *Trends in cognitive sciences* **18**, 203–210 (2014).
412 [21] Daube, C., Ince, R. A. & Gross, J. Simple acoustic features can explain phoneme-based predictions of cortical
413 responses to speech. *Current Biology* **29**, 1924–1937 (2019).
414 [22] Wickelgren, W. A. Short-term memory for phonemically similar lists. *The American Journal of Psychology* **78**,
415 567–574 (1965).
416 [23] Glasspool, D. W. & Houghton, G. Serial order and consonant–vowel structure in a graphemic output buffer model.
417 *Brain and language* **94**, 304–330 (2005).
418 [24] Fischer-Baum, S. A common representation of serial position in language and memory. In *Psychology of Learning*
419 *and Motivation*, vol. 68, 31–54 (Elsevier, 2018).
420 [25] Ohala, J. J., Browman, C. P. & Goldstein, L. M. Towards an articulatory phonology. *Phonology* **3**, 219–252 (1986).
421 [26] Browman, C. P. & Goldstein, L. Articulatory gestures as phonological units. *Phonology* **6**, 201–251 (1989).
422 [27] MacDonald, C. J., Lepage, K. Q., Eden, U. T. & Eichenbaum, H. Hippocampal “time cells” bridge the gap in
423 memory for discontinuous events. *Neuron* **71**, 737–749 (2011).
424 [28] Sohoglu, E., Peelle, J. E., Carlyon, R. P. & Davis, M. H. Predictive top-down integration of prior knowledge during
425 speech perception. *Journal of Neuroscience* **32**, 8443–8453 (2012).
426 [29] Bendixen, A., Scharinger, M., Strauss, A. & Obleser, J. Prediction in the service of comprehension: modulated
427 early brain responses to omitted speech segments. *Cortex* **53**, 9–26 (2014).
428 [30] Halle, M. & Stevens, K. Speech recognition: A model and a program for research. *IRE transactions on information*
429 *theory* **8**, 155–159 (1962).

- 430 [31] Gagnepain, P., Henson, R. N. & Davis, M. H. Temporal predictive codes for spoken words in auditory cortex.
431 *Current Biology* **22**, 615–621 (2012).
- 432 [32] Di Liberto, G. M., Wong, D., Melnik, G. A. & de Cheveigné, A. Low-frequency cortical responses to natural
433 speech reflect probabilistic phonotactics. *Neuroimage* **196**, 237–247 (2019).
- 434 [33] Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annual review of neuroscience* **30** (2007).
- 435 [34] Gramfort, A. *et al.* Mne software for processing meg and eeg data. *Neuroimage* **86**, 446–460 (2014).
- 436 [35] King, S. & Taylor, P. Detection of phonological features in continuous speech using neural networks (2000).
- 437 [36] Balota, D. A. *et al.* The english lexicon project. *Behavior research methods* **39**, 445–459 (2007).
- 438 [37] King, J.-R. *et al.* Encoding and decoding neuronal dynamics: Methodological framework to uncover the algorithms
439 of cognition (2018).
- 440 [38] King, J.-R., Charton, F., Lopez-Paz, D. & Oquab, M. Discriminating the influence of correlated factors from
441 multivariate observations: the back-to-back regression. *bioRxiv* (2020).
- 442 [39] Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *Journal of machine learning research* **12**, 2825–2830
443 (2011).