

# 1 Linguistic modulation of the neural encoding of 2 phonemes

3 Seung-Goo Kim<sup>1</sup>, Federico De Martino<sup>2</sup>, Tobias Overath<sup>1,3,4</sup>

4 <sup>1</sup>Department of Psychology and Neuroscience, Duke University, Durham, N.C., U.S.A.

5 <sup>2</sup>Faculty of Psychology and Neuroscience, University of Maastricht, The Netherlands.

6 <sup>3</sup>Duke Institute for Brain Sciences, Duke University, Durham, N.C., U.S.A.

7 <sup>4</sup>Center for Cognitive Neuroscience, Duke University, Durham, N.C., U.S.A.

8

9 **Corresponding authors:**

10 [solleo@gmail.com](mailto:solleo@gmail.com) (S.-G. K.), [t.overath@duke.edu](mailto:t.overath@duke.edu) (T. O.)

11

12 **Keywords:** speech perception, phonemes, acoustic analysis, linguistic analysis, human  
13 auditory cortex, fMRI

14

## 15 **Abstract**

16 Speech comprehension entails the neural mapping of the acoustic speech signal onto learned linguistic  
17 units. This acousto-linguistic transformation is bi-directional, whereby higher-level linguistic processes  
18 (e.g., semantics) modulate the acoustic analysis of individual linguistic units. Here, we investigated the  
19 cortical topography and linguistic modulation of the most fundamental linguistic unit, the phoneme. We  
20 presented natural speech and ‘phoneme quilts’ (pseudo-randomly shuffled phonemes) in either a familiar  
21 (English) or unfamiliar (Korean) language to native English speakers while recording fMRI. This design  
22 dissociates the contribution of acoustic and linguistic processes towards phoneme analysis. We show  
23 that (1) the four main phoneme classes (vowels, nasals, plosives, fricatives) are differentially and  
24 topographically encoded in human auditory cortex, and that (2) their acoustic analysis is modulated by  
25 linguistic analysis. These results suggest that the linguistic modulation of cortical sensitivity to phoneme  
26 classes minimizes prediction error during natural speech perception, thereby aiding speech  
27 comprehension in challenging listening situations.

## 28 **1 Introduction**

29 Speech comprehension relies on the neural mapping of the acoustic speech signal onto linguistic  
30 categories (Hickok and Poeppel, 2007; Kleinschmidt and Jaeger, 2015; Poeppel et al., 2008). As such,  
31 the acoustic speech waveform that reaches our ears is converted into a neural code in the inner ear,  
32 which is then processed along the ascending auditory system and subsequently matched to learned  
33 linguistic categories (Friederici, 2011; Hickok and Poeppel, 2007). However, while this acousto-linguistic  
34 transformation is the basis for successful speech comprehension, many aspects of it still remain unknown.  
35 Here, we asked (1) whether the acousto-linguistic transformation of the most fundamental linguistic unit,  
36 the phoneme, is organized topographically by phoneme class in human auditory cortex, and (2) whether  
37 this transformation is malleable to top-down linguistic information.

38 The phoneme is the smallest perceptual unit capable of determining the meaning of a word (e.g., the  
39 words *pin* and *chin* differ only with respect to their initial phonemes) (Stevens, 2000). Of the upward of  
40 100 phonemes in use world-wide, approximately 44 phonemes make up the English language and these  
41 are categorized primarily based on articulatory features into four main classes: vowels, nasals and  
42 sonorants, plosives, fricatives and affricates (Ladefoged, 2001; Ladefoged and Johnstone, 2015). Each  
43 phoneme class has characteristic acoustic features; for example, while vowel sounds display a sustained  
44 period of harmonicity, plosives are characterized by a brief period of silence followed by a short  
45 broadband noise burst. Individual phonemes and the phoneme classes to which they belong have distinct  
46 temporal neural correlates: each phoneme class has a unique time-locked neural response, or phoneme-  
47 related potential (PRP; Khalighinejad et al. (2017); Lee and Overath (in revision)). The phoneme-class-  
48 specific PRPs likely reflect the neural analysis of their acoustic characteristics (e.g., timing of energy onset,  
49 harmonicity, etc.) in functionally separate parts of auditory cortex. While previous intracranial recording  
50 studies similarly revealed phonetic feature selectivity in the human superior temporal gyrus (Mesgarani et  
51 al., 2014; Yi et al., 2019), they found no topographical organization of phoneme classes.

52 Of course, phonemes do not occur in isolation, but instead form sequences to create syllables and  
53 words. The order in which phonemes can occur is governed by phonotactics, and is unique to each  
54 language (Chomsky and Halle, 1965). Apart from learning to recognize the language-specific phonemes  
55 themselves (Cheour et al., 1998), phonotactics is one of the first sets of rules infants need to learn during  
56 language acquisition (Friederici and Wessels, 1993; Jusczyk et al., 1994; Mattys and Jusczyk, 2001). This  
57 may be achieved via learning the likelihood of phoneme transitions: for example, in English certain  
58 phoneme transition probabilities are statistically unlikely (or even non-existent, e.g., /dla/) while others  
59 are statistically more likely (e.g., /gla/). A similar principle is thought to be employed for syllable transitions,  
60 where statistically improbable syllable transitions can indicate between-word boundaries (Saffran et al.,  
61 1996).

62 Thus, while the initial analysis of phonemes is based on their acoustic features (Khalighinejad et al.,  
63 2017; Lee and Overath, in revision; Mesgarani et al., 2014; Yi et al., 2019), subsequent processing stages  
64 are likely more linguistic in nature, such as those identifying language-specific phonemes or phonotactics,  
65 or even higher-level processes underlying the analysis of syntax, semantics, or lexical access (Friederici  
66 et al., 1993; Kocagoncu et al., 2017; Kutas and Hillyard, 1983). While decades of research have resulted  
67 in detailed speech/language models (Friederici, 2011; Hickok and Poeppel, 2007; Rauschecker and Scott,

68 2009), a clear demarcation between acoustic and linguistic analyses has largely remained elusive. One  
69 reason for this is that, in everyday listening situations, acoustic and linguistic analyses are difficult to  
70 separate and likely interact, e.g., via top-down modulation of acoustic feature analysis by linguistic  
71 processes (Anderson et al., 2003; Davis and Johnsrude, 2007; Díaz et al., 2008). In addition, previous  
72 studies that investigated phoneme processing in naturalistic contexts (Khalighinejad et al., 2017;  
73 Mesgarani et al., 2014), did so only in a familiar language: this approach is unable to dissociate the initial  
74 acoustic processes from the obligatory nature of linguistic processes that become engaged in a native,  
75 familiar language.

76 In contrast, Lee and Overath (in revision) were recently able to dissociate the acoustic and linguistic  
77 processes underlying phoneme analysis by comparing PRPs in familiar vs. foreign languages. They used  
78 a variant of a novel sound quilting algorithm (Overath et al., 2015; Overath and Paik, 2021) to create  
79 speech-based quilts in which linguistic units (phoneme, syllable, word) were pseudo-randomly ‘stitched  
80 together’ to form a new stimulus. This paradigm allowed the comparison of an acoustic stimulus  
81 manipulation (speech-based quilting) in a familiar vs. foreign language: if the processing of phonemes is  
82 affected by the acoustic manipulation (increasing linguistic unit size of speech quilts) in a familiar language  
83 only, then this would suggest that linguistic analysis in the familiar language influenced the acoustic  
84 analysis of phonemes. Put differently, if no phonemic repertoire or phonotactic rules are available to a  
85 listener (as is the case in a foreign language), the encoding of phonemes themselves should be  
86 independent of their ordering (phonotactics) or linguistic unit size in which they appear. Using EEG to  
87 investigate the PRP for different phoneme classes (Khalighinejad et al., 2017), Lee and Overath (in revision)  
88 found that vowels in particular are amenable to such top-down linguistic modulation. However, the limited  
89 spatial resolution of EEG did not allow inferences as to where in the auditory cortex (or beyond) such top-  
90 down modulation might originate, or act upon.

91 Recent advances in fMRI time-series analysis have demonstrated that the neural activity to natural  
92 speech stimuli can be predicted from fast-paced acoustic (e.g., envelope, spectrum), phonological, and  
93 semantic features via linearized encoding modeling (De Heer et al., 2017; Huth et al., 2016). Inspired by  
94 this approach, the current study employed linearized encoding modeling of fMRI data in human speech  
95 cortex in an effort to reveal the separate encoding of acoustic and linguistic features of speech.  
96 Specifically, we used speech-based quilting (phoneme quilts vs. original speech) in familiar (English) vs.  
97 foreign (Korean) languages to dissociate the neural correlates of the acoustic and linguistic processes  
98 that contribute to the analysis of a fundamental linguistic unit, the phoneme. We show, for the first time,  
99 (1) that individual phoneme classes are differentially and topographically encoded in fMRI data, and (2)  
100 that their acoustic analysis is modulated by linguistic processes.

## 101 **2 Results**

102 Linearized encoding models with predictors for the four phoneme classes (i.e., vowels, nasals and  
103 approximants, plosives, fricatives and affricatives; “Phonemes”) and the cochleogram envelope  
104 (“Envelope”) were used to predict the fMRI time series acquired from native English speakers without any  
105 knowledge in Korean while listening to speech stimuli in four conditions (phoneme quilts or original

## Linguistic Modulation of Phoneme Encoding

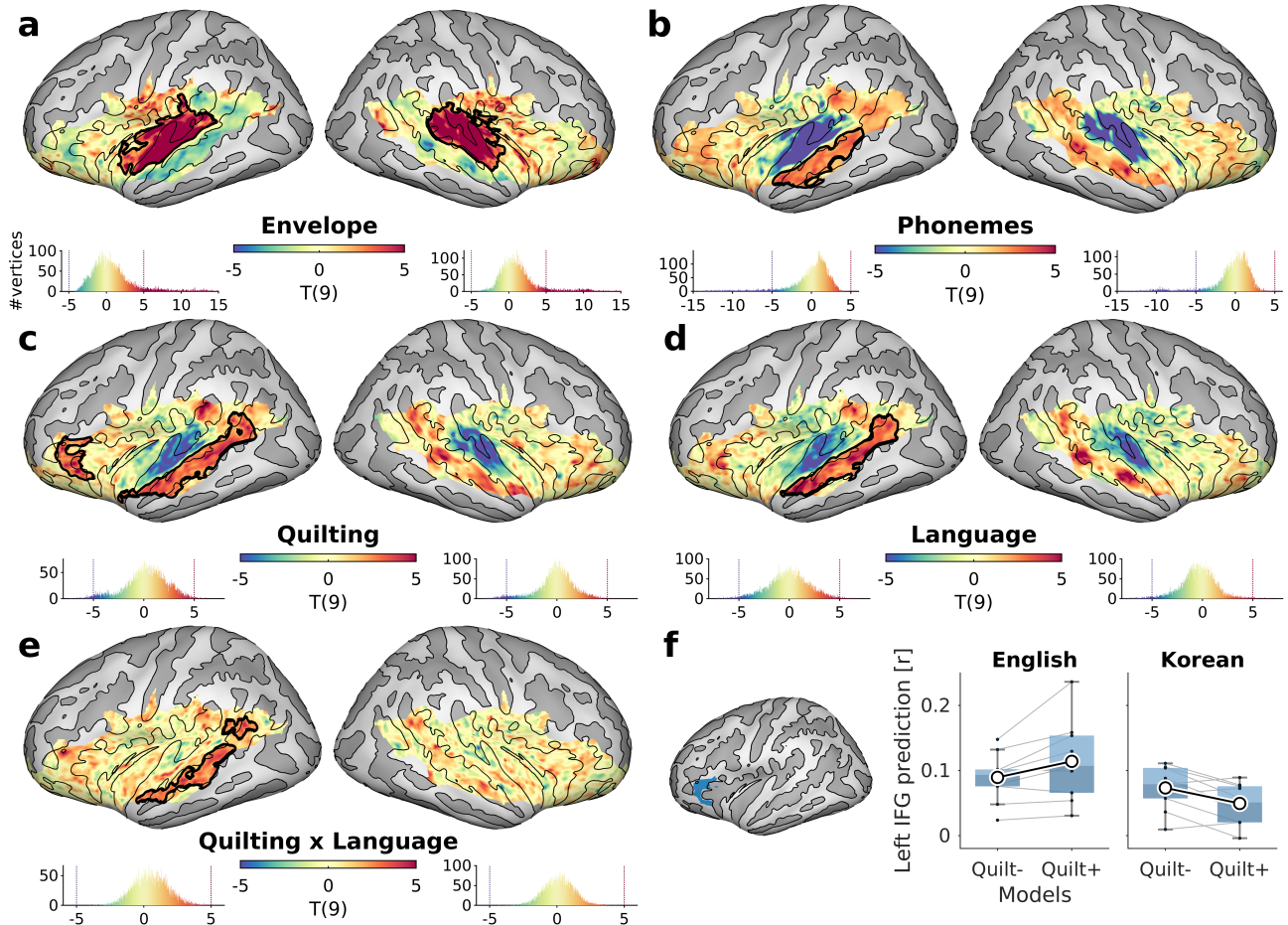
106 speech, in English or Korean). We used multiple ridge-regression models to infer the significance of the  
107 encoding contribution of a specific predictor or a condition from an increase in prediction accuracy (i.e.,  
108 Pearson's correlation coefficient) when adding the specific predictor or condition to a model (see **Figure**  
109 **4** for an overview of the analysis).

### 110 **2.1 Encoding of phonemes and conditions**

111 We first investigated whether the acoustic envelope and linguistic phoneme classes are encoded  
112 differently in auditory cortex. The inclusion of the Envelope predictor yielded a marked increase in  
113 prediction accuracy in the primary auditory cortex, mid-STG, and planum temporale (max  $t(9) = 19.816$ ;  
114 min cluster- $P < 0.0001$ ; max positive cluster size = 1649 vertices; max diff  $r = 0.0139$ , 95%CI = [0.012,  
115 0.015]; **Figure 1a**). Conversely, the addition of the Phoneme predictors revealed the strongest positive  
116 contribution in left STS (max  $t(9) = 5.540$ ; min cluster- $P = 0.032$ ; max cluster size = 678 vertices; max diff  
117  $r = 0.156$  [0.076, 0.253]; **Figure 1b**).

118 After establishing that both Envelope and Phonemes have significant contributions (and should  
119 therefore be included in the full model), we investigated the effects of the factors Quilting and Language.  
120 We found a main effect of Quilting in the left STS and left IFG (max  $t(9) = 7.298$ ; min cluster- $P = 0.003$ ;  
121 max cluster size = 1028 vertices; max diff  $r = 0.133$  [0.091, 0.171]; **Figure 1c**), and a main effect of  
122 Language in a similar location in the left STS, but not in IFG (max  $t(9) = 8.086$ ; min cluster- $P = 0.005$ ; max  
123 cluster size = 939 vertices; max diff  $r = 0.113$  [0.061, 0.166]; **Figure 1d**). Left STS also revealed an  
124 interaction, where the prediction accuracy change by modeling the Original-vs-Quilting conditions was  
125 positive for English conditions and negative for Korean (max  $t(9) = 5.465$ ; min cluster- $P = 0.003$ ; max  
126 cluster size = 517 vertices; max diff  $r = 0.181$  [0.103, 0.276]; **Figure 1e**).

127 A previous study from our group (Overath and Paik, 2021) used a similar paradigm to dissociate  
128 acoustic and linguistic processes by quilting speech (English vs. Korean) with various durations of  
129 temporal speech structure (set segment durations, e.g., 30 ms or 960 ms, as opposed to linguistic units  
130 with variable durations as in the current study); they found that activity in left IFG increased as a function  
131 of temporal speech structure only for English, but remained unaffected for increases in temporal speech  
132 structure in Korean. Since the acoustic manipulation of temporal speech structure was the same in both  
133 languages, this result was interpreted as evidence for an acousto-linguistic transformation in left IFG. We  
134 therefore tested for a similar interaction using anatomically defined ROIs of IFG (a combination of  
135 `G_front_inf-Orbital` and `G_front_inf-Triangul` in FreeSurfer's Destrieux "a2009s" Atlas). The  
136 permutation test revealed an interaction in the left IFG that was due to a selective increase in prediction  
137 accuracy in the English conditions ( $t[9] = 3.63$ ,  $P = 0.005$ ; **Figure 1f**).



138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

**Figure 1. Encoding of features and conditions.**  $t$ -statistic maps comparing prediction accuracies from models with vs. without (a) Envelope, (b) Phonemes, (c) Quilting, (d) Language, (e) and differences in Quilting encoding between languages (i.e., interaction).  $t$ -maps are truncated at  $|t| < 5$  for visualization; thick contours in black mark areas with cluster- $P < 0.05$ . Curvatures of the cortical surface are displayed in brighter (convex) and darker (concave) grays with an isocontour at the curvature of zero in black. Colored histograms of the  $t$ -values over the full range are displayed below each hemisphere. See Supplementary Figure S6 for effects at the subject-level. (f) Average prediction accuracies in the region-of-interest (ROI) in the left inferior frontal gyrus (IFG; marked in blue) are shown for English and Korean when modeling the quilting conditions (Quilt+) or not (Quilt-). Individual participants are marked as black dots and paired for identical participants. The means are marked by white circles and linked for comparison between models. Box plots mark the first three quartiles (top and bottom edges of a box and a shade) and the 1.5 interquartile range (whiskers).

154

## 2.2 Encoding of individual phoneme classes

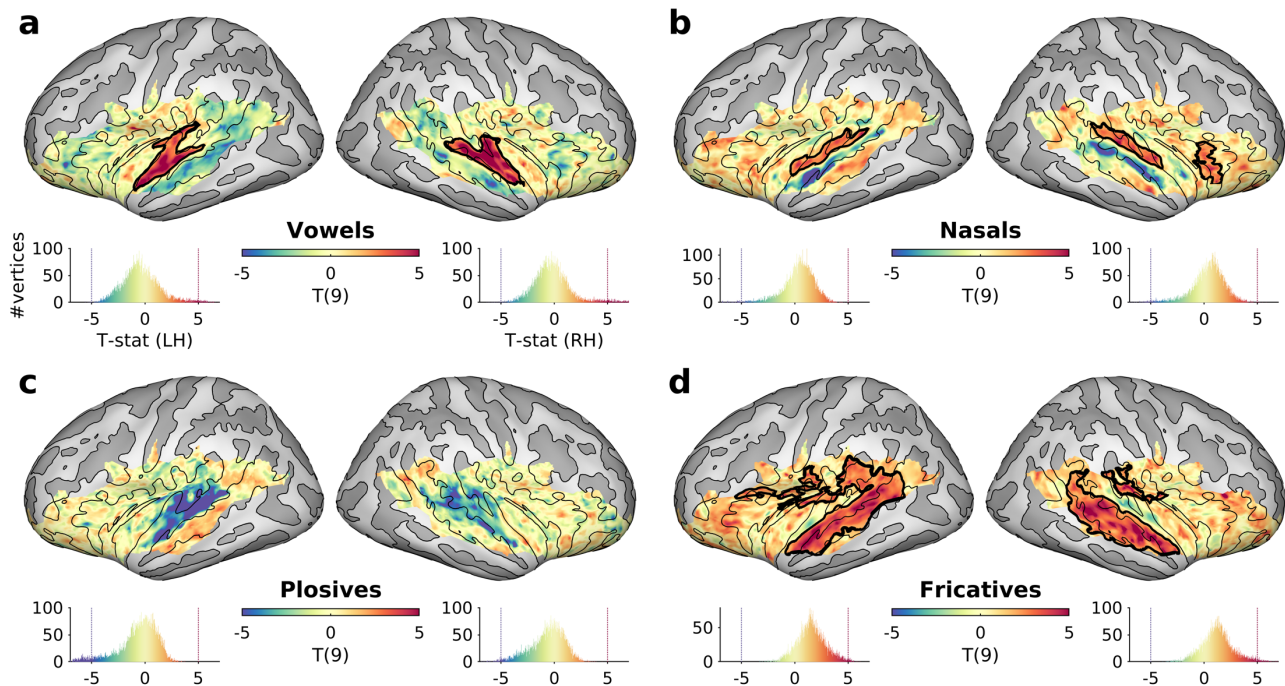
155

156

We further investigated the encoding of individual phoneme classes by comparing prediction accuracies between a full model and a reduced model without a specific phoneme class feature (Figure 2a-d). This

## Linguistic Modulation of Phoneme Encoding

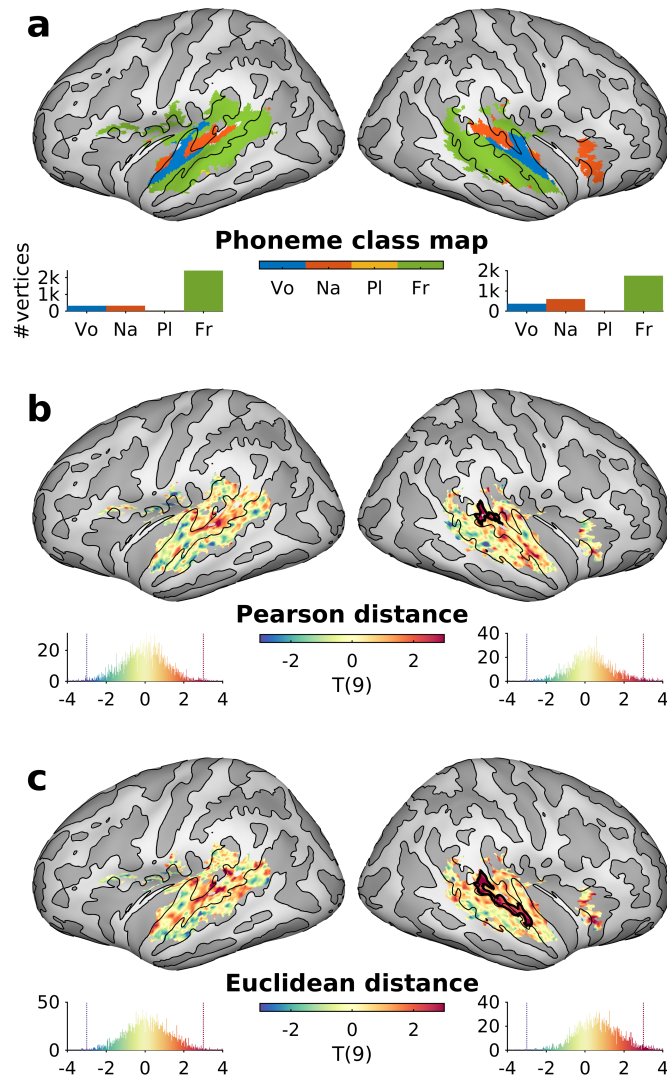
157 analysis revealed significant increases in prediction accuracies when adding vowels in the bilateral HG  
158 and STG, when adding nasals in the bilateral PT and lateral HG, and when adding fricatives in the bilateral  
159 STS.



160

161 **Figure 2. Unique encoding of individual phoneme classes. *t*-statistic maps comparing**  
162 **the prediction accuracy of a full model with that of a reduced model without a specific**  
163 **phoneme feature: (a) vowels, (b) nasals, (c) plosives, or (d) fricatives. Thick contours in**  
164 **black mark areas for which cluster-*P* < 0.05.**

165 Since adding phoneme classes had significant effects in distinct spatial patterns, we constructed a  
166 winner-take-all map of phoneme classes (thresholded for cluster-*P* < 0.05 for any phoneme class; **Figure**  
167 **3a**). The effects of adding phoneme classes were largely symmetrical between hemispheres: vowels in  
168 HG and lateral STG, nasals in anterior Heschl's sulcus (HS) and PT, and fricatives in lateral STG and STS  
169 (however, note the binary nature of the winner-take-all map; see **Figure S7** for low-dimensional  
170 projections of the phoneme encoding vectors). Consequently, we tested whether phoneme classes were  
171 encoded differently as a function of Quilting, and whether this effect was more pronounced in English  
172 than in Korean; such an interaction would suggest that linguistic processes (which are available in a  
173 familiar, or native language only) modulate the processing of phonemes. We reasoned that this would be  
174 the case if the encoding of phoneme classes was affected by quilting more in English than in Korean; in  
175 other words, if, for a given vertex, the difference between the phoneme class vector of quilted vs. original  
176 speech was more dissimilar in English than in Korean (see **Figure 5** for a schematic cartoon of the  
177 multivariate analysis). Both Pearson and Euclidian distances between phoneme class vectors in quilted  
178 vs. original speech conditions were significantly greater in English than in Korean in the right STG  
179 (Pearson distance, cluster-*P* = 0.0469, **Figure 3b**; Euclidian distance, cluster-*P* = 0.0029, **Figure 3c**).



180

181

182

183

184

185

186

187

**Figure 3. Encoding patterns of phoneme classes. (a) Winner-take-all map for the four phoneme classes. Vertices were selected for significant encoding for any phoneme class (cluster- $P < 0.05$ ). The histograms denote the number of 'winner' vertices per phoneme class in the left and right hemisphere. (b-c) Multivariate distance metrics were greater in English than Korean in the right posterior lateral STG (cluster- $P < 0.05$ ), when computed with respect to (b) Pearson distance and (c) Euclidean distance.**

188

### 3 Discussion

189

190

191

192

193

194

The phoneme is the fundamental linguistic unit that determines the meaning of words. We show that the four main phoneme classes are encoded in fMRI data recorded from continuous speech signals, revealing a distinct phoneme class topography in human auditory cortex. Moreover, the acoustic processes underlying this phoneme analysis are modulated by linguistic analysis, whereby the acoustic manipulation (phoneme quilts vs. original speech) affected phoneme encoding more in a familiar language than in a foreign language. The results also reveal relatively stronger neural correlates for lower-level acoustic



## Linguistic Modulation of Phoneme Encoding

195 attributes (e.g., speech envelope) of the speech signal in early auditory cortex, and for higher-level  
196 linguistic features (phoneme classes) in STS.

### 197 **3.1 Distinct encoding of individual phoneme classes**

198 To our knowledge, we present the first evidence for topographically organized phoneme-class sensitive  
199 responses in human auditory cortex using fMRI data measured from listening to continuous, natural  
200 speech signals. More specifically, while some previous studies investigated phoneme processing by  
201 presenting individual phonemes in isolation (e.g., Formisano et al. (2008); Obleser et al. (2010); see also  
202 review in DeWitt and Rauschecker (2012)), the power of the current approach rests on the fact that it  
203 delineates phoneme-class sensitive responses within a more ecologically valid environment of  
204 continuous speech (Hamilton and Huth, 2020). As such, the phoneme-class map can be regarded as the  
205 hemodynamic equivalent of the electrophysiological phoneme-related potential (Khalighinejad et al., 2017;  
206 Mesgarani et al., 2014).

207 **Figure 3a** revealed a clear topography for different phoneme classes, which was remarkably  
208 homologous between the left and right hemispheres, down to the cluster for nasals in antero-medial HS.  
209 Vowels showed the strongest sensitivity in early auditory cortex in HG and parts of STG. Since one of the  
210 main features of vowels is their harmonicity, which is a defining characteristic of pitch (Plack et al., 2005),  
211 this matches well with studies that have shown pitch-sensitive responses in these early cortical areas  
212 (Bendor and Wang, 2005; Griffiths and Hall, 2012; Hall and Plack, 2009; Kim et al., in revision). Of the four  
213 phoneme classes, fricatives 'won' the majority of vertices, which were mostly located in non-primary  
214 cortex (STG and STS). While the large extent was a somewhat unexpected finding, its location in non-  
215 primary cortex agrees with a previously proposed functional hierarchy in auditory cortex, whereby  
216 spectral filter width increases (e.g., from sinusoids via narrowband to broadband noise) from primary to  
217 non-primary auditory cortex (Carrasco and Lomber (2009); Chevillet et al. (2011); Pandya et al. (2007);  
218 Rauschecker and Tian (2004); Rauschecker et al. (1995); Wessinger et al. (2001) though see Overath et  
219 al. (2012); Wang et al. (2012)). Since high-frequency broadband noise is a defining characteristic of  
220 fricatives, the strong response in STG and STS could reflect such spectral sensitivity. Nasals revealed  
221 highest prediction accuracies in the non-primary auditory cortex (e.g., lateral HG, PP, and PT). These  
222 areas roughly correspond to regions that display sensitivity to slow temporal modulations (Baumann et  
223 al., 2015; Santoro et al., 2014). In particular, Schönwiesner and Zatorre (2009) revealed a preference for  
224 low spectral density (with slow temporal modulation rates) in these regions. Nasals are characterized by  
225 a 'murmur', related to the nasal resonances if the oral tract is closed, which effectively amounts to a low-  
226 pass filter (Qi and Fox, 1992). An animal model (ferrets) showed that nasals tend to excite primary auditory  
227 cells that are tuned for slow temporal modulations (Mesgarani et al., 2008). Thus, it is conceivable that  
228 the spectrotemporal modulation preference in the non-primary auditory cortex is reflected in the preferred  
229 encoding of the narrow-band filtered fluctuation of nasals. Finally, plosives, which are characterized by a  
230 complete occlusion followed a subsequent broadband burst, were the only phoneme class that did not  
231 reveal areas with maximal prediction accuracy compared to the other phoneme classes. As hinted in  
232 **Figure 2c**, there was a slight (subthreshold) increase in prediction accuracies in the left STS for plosives,  
233 which is similar to that of fricatives but at weaker magnitudes. This is in line with a similarity between  
234 responses to plosives and to fricatives (Khalighinejad et al., 2017), which suggests similar neural

235 generators. Plosives and fricatives are both characterized by noise bursts; the brevity of the burst in  
236 plosives compared to the more sustained noise burst in fricatives might have resulted in an overall lower  
237 encoding prediction for plosives in similar regions as those encoding fricatives, which in turn might  
238 explain the absence of plosive phoneme class ‘winner’ regions in the current study.

239 Mesgarani et al. (2014) measured phoneme-related potentials from continuous speech using  
240 intracranial ECoG recordings from six epileptic patients undergoing pre-surgery evaluation, but did not  
241 find a consistent topographic organization of phoneme classes across patients. However, upon closer  
242 inspection, their Figure S6, which depicts electrodes in a winner-take-all manner along the STG for one  
243 patient, is compatible to the results we report here with fMRI: vowel-sensitive electrodes were located  
244 near lateral HG and adjacent STG, while fricative-sensitive electrodes were found over STS and posterior  
245 STG. Interestingly, plosive-sensitive electrodes were also located over STS and posterior STG. Four of  
246 the six nasal-sensitive electrodes were located on STG either anterior or posterior to HG, which generally  
247 matches the current topography. Given that the current results are based on ~160 minutes of stimuli per  
248 participant (see Methods), compared to ~17 minutes of stimuli in Mesgarani et al. (2014), it is conceivable  
249 that the collection of more ECoG data would have revealed a topography such as the one we report here  
250 (though the higher signal-to-noise ratio of ECoG data likely reduces this order of magnitude difference).

251 While **Figure 3a** suggests clear topographic distinctions between phoneme classes, it is important to  
252 note that this is partly an artifact of winner-take-all maps and does not necessarily implicate that acoustic  
253 features of, for example, vowels are not processed beyond HG (just that their prediction accuracy was  
254 not maximal elsewhere). In fact, **Figure S7b** highlights the underlying high-dimensionality of responses  
255 at each vertex.

### 256 ***3.2 Linguistic modulation of acoustic phoneme-class analysis***

257 One of our aims was to dissociate acoustic from linguistic processes, which would enable us to determine  
258 their interaction, i.e., whether linguistic processes modulate the acoustic analysis of phonemes. To this  
259 end, we found that the acoustic manipulation (phoneme quilts vs. natural speech) had a larger effect on  
260 phoneme processing in a familiar language (English) than in a foreign language (Korean): phoneme class  
261 encoding was more dissimilar between phoneme-quilt and natural speech conditions in English than in  
262 Korean. Since the acoustic manipulation was the same for each language, this suggests that the greater  
263 difference between acoustic contexts was due to linguistic processes becoming engaged in a familiar  
264 language. Linguistic processes such as phonotactic, syntactic, as well as semantic analyses might  
265 therefore modulate the acoustic processing of phonemes, e.g., via hierarchical predictive coding or  
266 minimizing top-down prediction errors (Friston and Kiebel, 2009; Rao and Ballard, 1999). To our  
267 knowledge, this is the first demonstration of such linguistic modulation of a fundamental linguistic unit  
268 using fMRI. However, these results align with Lee and Overath (in revision), who found similar evidence  
269 for top-down linguistic modulation of phonemic analysis using a different recording modality (EEG).

270 Perhaps the best-known example of the modulatory influence of linguistic information is that of  
271 phonemic restoration (Samuel, 1981; Warren, 1970). In phonemic restoration, a phoneme is still  
272 ‘perceived’ even if it is masked or replaced completely by noise. This is often interpreted as an  
273 advantageous adaptation to speech perception in noisy environments, where it is common for

## Linguistic Modulation of Phoneme Encoding

274 interrupting or masking sounds to last only for a few tens or hundreds of milliseconds (i.e., on a temporal  
275 scale that is commensurate with that of phonemes). The top-down predictive nature of this phenomenon  
276 is further highlighted by the fact that, if the acoustic information is ambiguous, a ‘best guess’ phoneme  
277 is perceived (Leonard et al., 2016; Samuel, 1987). In fact, there is a wealth of evidence for such restorative  
278 processes in speech perception, for example from studies using noise-vocoded stimuli (Giraud et al.,  
279 2004; Narain et al., 2003; Obleser et al., 2008; Scott et al., 2000; Shannon et al., 1995; Wild et al., 2012)  
280 or other methods to distort the speech signal (Davis et al., 2011; Eckert et al., 2016), while the most  
281 common explanation for restorative effects refers to top-down, predictive (Friston and Kiebel, 2009)  
282 linguistic processes.

283 The locus of phonemic restoration, i.e. the region in which linguistic modulation is strongest, was  
284 recently shown to be situated in bilateral STG, likely due to receiving modulatory signals from left IFG  
285 (Leonard et al., 2016). This aligns remarkably well with the current study, where we found the strongest  
286 effect of linguistic modulation along right STG. Note that this region along STG touches upon areas of all  
287 three phoneme-class ‘winners’ (vowels, nasals, fricatives; cf. **Figure 3a**) and is therefore ideally situated  
288 to modulate the neural analysis of these phoneme classes. The apparent right-lateralization may in part  
289 be a consequence of the cluster-forming threshold, which penalizes or disregards smaller activation  
290 clusters, since similar but smaller peaks along left STG are also visible for both Pearson and Euclidean  
291 distance metrics (**Figure 3b–c**).

292 The STG is a reasonable locus for such linguistic modulation, since it represents an intermediary  
293 processing stage in the language network that receives bottom-up information from primary auditory  
294 cortex and PT, as well as top-down information from higher-order auditory and frontal regions (Friederici,  
295 2009, 2011; Hickok and Poeppel, 2007; Rauschecker and Scott, 2009). For example, the analysis of  
296 spectral shape (a necessary computation to differentiate between the formant structures of different  
297 vowels) relies on bottom-up changes in effective connectivity between HG to PT, as well as PT to  
298 STG/STS regions (Kumar et al., 2007; Warren et al., 2005). In contrast, top-down signals from frontal  
299 cortex (e.g., left IFG) have been shown to modulate speech processing in auditory cortex (Cope et al.,  
300 2017; Overath and Paik, 2021; Park et al., 2015; Sohoglu et al., 2012).

301 In the domain of electrophysiological measurements of speech perception, there is currently  
302 disagreement as to the extent that neural indices (such as speech-envelope entrainment, or phoneme  
303 encoding) can be interpreted as markers of linguistic processes that are necessary for speech  
304 comprehension (Di Liberto et al., 2015; Ding and Simon, 2013; Luo and Poeppel, 2007; Vanthornhout et  
305 al., 2018), or whether a more parsimonious explanation of these indices is that they reflect the analysis  
306 of characteristic acoustic properties of the speech signal (Baltzell et al., 2017; Daube et al., 2019; Howard  
307 and Poeppel, 2010; Millman et al., 2015; Verschueren et al., 2021). Our study is able to shed new light on  
308 this controversy by directly comparing the encoding of acoustic properties of phonemes in either a  
309 familiar language or in a foreign language, in which no higher-level linguistic analysis takes place. Based  
310 on the current results, we suggest that both accounts have merit: an (initial) analysis is likely driven by  
311 the acoustic properties of phonemes, while a (subsequent) linguistic analysis modulates this acoustic  
312 analysis.

313 We should note that the current study did not address or measure linguistic processes explicitly. For  
314 example, participants did not perform a linguistic task (e.g., speech comprehension), but were simply

315 asked to detect a change in speaker, a task that is largely orthogonal to linguistic processing (see also  
316 Overath and Paik (2021) for a similar task). Therefore, we interpret the linguistic modulation of phoneme  
317 class analysis as obligatory linguistic processes that become engaged as soon as familiar linguistic  
318 templates (e.g., phonotactics, syntax, lexicon, semantics) are detected in the signal. Future studies will  
319 need to determine whether, and to what extent, these obligatory linguistic processes for phoneme  
320 analysis are malleable to various tasks that engage specific linguistic processes. For example, the neural  
321 processing of acoustic features in speech sounds has been shown to be enhanced or sharpened if they  
322 are task-relevant, attended to vs. ignored, or primed (Holdgraf et al., 2016; Leonard et al., 2016;  
323 Mesgarani and Chang, 2012; Rutten et al., 2019), and similar processes might become engaged for  
324 phoneme class encoding.

### 325 ***3.3 Encoding of envelope and phoneme classes in the BOLD time series***

326 One of our preliminary aims was to confirm that rapid acoustic and phonetic features can be shown to  
327 be encoded in a hemodynamic response that is approximately two orders of magnitude slower (tens of  
328 milliseconds vs. seconds). Encoding of these features had previously been demonstrated using  
329 electrophysiological methods, which afford commensurate millisecond temporal resolution (Di Liberto et  
330 al., 2015; Khalighinejad et al., 2017; Mesgarani et al., 2014; Yi et al., 2019). Nevertheless, the novel use  
331 of linearized ridge-regression modeling of fMRI BOLD signal time series was recently employed to  
332 successfully (and separably) predict acoustic and phonetic features: De Heer et al. (2017) collected fMRI  
333 data while presenting continuous, natural speech, and were able to reveal that the acoustic speech  
334 envelope predicted the BOLD time series best in HG, whereas articulatory phonetic features were  
335 predicted most accurately in higher-level auditory cortex such as STG. The current study is in broad  
336 agreement with these findings: while areas in early auditory cortex best encoded the acoustic speech  
337 envelope, higher-level areas in STG and STS did so for phoneme classes.

338 More broadly, our study confirms that neural responses to rapid speech features, which are temporally  
339 integrated over several hundreds of milliseconds in the BOLD time series, can be revealed using linearized  
340 encoding modeling. Such models take advantage of the spatially separated functional organization of  
341 auditory cortex, for example with respect to prominent acoustic features such as frequency, spectro-  
342 temporal modulations, or spectral bandwidth (Baumann et al., 2015; Moerel et al., 2018; Rauschecker  
343 and Tian, 2004; Saenz and Langers, 2014; Santoro et al., 2014). This should encourage the future use of  
344 more naturalistic stimulus paradigms that allow the investigation of the complex dynamics of linguistic  
345 processes (Hamilton and Huth, 2020).

### 346 ***3.4 Modulation of acoustic and linguistic contexts***

347 The analyses of the two factors Quilting and Language were motivated by previous studies that  
348 investigated the processing of temporal speech structure using segment-based speech quilting. In  
349 particular, these studies showed sensitivity in STS to temporal speech structure in either only a foreign  
350 language (Overath et al., 2015), or both familiar and foreign languages (Overath and Paik, 2021), which is  
351 comparable to a main effect of Quilting here. In addition, activity in left IFG revealed an interaction  
352 between Quilting and Language and increased as a function of temporal speech structure only in the

## Linguistic Modulation of Phoneme Encoding

353 familiar language (Overath and Paik, 2021). In the current study, Quilting and Language both had greater  
354 prediction accuracies in left STS, while their interaction in the same area (as well as left IFG) was due to  
355 larger prediction accuracy differences between the two Quilting conditions in English than in Korean.

356 For successful speech comprehension, the temporal dynamics of speech necessitate analyses at  
357 multiple scales that are commensurate with the average durations of phonemes, syllables, words,  
358 sentences, etc. This temporal hierarchy is thought to be reflected in a cortical processing hierarchy in  
359 which the neuronal temporal window of integration (Theunissen and Miller, 1995) increases from primary  
360 auditory cortex via non-primary auditory cortex to frontal cortex (e.g., Lerner et al. (2011); Norman-  
361 Haignere et al. (2020); though see Blank and Fedorenko (2020) for a recent counterargument against the  
362 hierarchy). The current results of greater prediction accuracy in STS as a function of Quilting largely  
363 support this view. A novel finding is the left-hemispheric lateralization. However, it is possible that this  
364 was driven by the interaction between Quilting and Language.

365 It is important to note that the segment-based quilting in previous studies disrupted the speech signal  
366 to a larger degree than the speech-based quilting employed here. The shortest segment length (30 ms)  
367 used in the previous studies, together with their placement irrespective of linguistic units, likely resulted  
368 in no phonemes being left intact in the resulting speech quilt. In contrast, the current speech-based  
369 quilting procedure preserved the phonemes (though likely still disrupted co-articulation cues).

### 370 **3.5 Future directions**

371 The current study makes a number of predictions for future studies investigating the acousto-linguistic  
372 transformation of speech. We show evidence for linguistic modulation of a fundamental linguistic unit,  
373 the phoneme, in native English speakers when listening to English speech, but not when listening to a  
374 foreign language for which participants had no linguistic repertoire. Therefore, while it is unlikely that the  
375 current results are specific to English phonemes, future studies should confirm this interaction in native  
376 Korean participants who have no knowledge of English. Similarly, people who are perfectly bilingual in  
377 English and Korean should show evidence for linguistic modulation in both languages as a function of  
378 quilting, while those for whom both languages are foreign should not.

379 In addition, the fact that the linguistic modulation of the acoustic speech signal operates at such an  
380 early stage of linguistic analysis likely reflects its significance: if linguistic modulation starts at the level of  
381 phonemes, its ability to impact a later word processing stage is conceivably greater than if linguistic  
382 modulation only started at the word processing stage. Given the highly predictive nature of speech  
383 processing (see **Section 3.2** above), such modulation might be particularly helpful in situations in which  
384 the speech signal is compromised (e.g., in noisy conditions such as in a restaurant or bar). People with  
385 hearing loss (e.g., presbycusis) are a clinical population that is known to struggle in such situations, even  
386 with the help of hearing aids (Moore, 1996; Shinn-Cunningham and Best, 2008). It is therefore possible  
387 that (at least) one reason for their exacerbated speech comprehension difficulties in noisy situations is  
388 that the linguistic modulation of phonemes has deteriorated, thereby reducing the effectiveness of  
389 predictive speech processes. A similar argument might be made for people suffering from ‘hidden hearing  
390 loss’: i.e., hearing difficulties without detectable deficits in routine audiometry tests (Kujawa and Liberman,

391 2009; Ruggles et al., 2011). We predict that linguistic modulation of phoneme analysis is reduced in these  
392 populations (particularly in situations with background noise) and might thus serve as a clinical marker.

393 On a more technical note: we investigated the relative contributions of several predictors in our models.  
394 However, it is possible that the addition of predictors leads to a decrease in prediction accuracy, which  
395 can be explained by over-penalization originating from using a single regularizing parameter for all  
396 features. During optimization, introducing irrelevant features into the model could require greater  
397 regularization to minimize prediction errors. This would shrink predictions from “well-predicting”  
398 predictors, resulting in an overall decrease in prediction accuracy. In such instances, banded ridge  
399 regression, where independent regularization parameters for multiple groups of features are optimized  
400 together, has recently been proposed as a solution to avoid over-penalization (Nunez-Elizalde et al., 2019).  
401 Future studies will need to explore the benefits of such an approach in paradigms that attempt to explain  
402 multiple features from fMRI data.

### 403 **3.6 Conclusions**

404 In conclusion, the current study demonstrates for the first time that individual phoneme classes derived  
405 from continuous speech signals are encoded differentially in the BOLD signal time series. The phoneme  
406 class topography reflects the processing of the different acoustic features that characterize the four  
407 phoneme classes. As such, it represents a significant step in our understanding of the functional  
408 organization of human auditory cortex. Moreover, by using a design that dissociates acoustic from  
409 linguistic processes, we show that the acoustic processing of a fundamental linguistic unit, the phoneme,  
410 is modulated by linguistic analysis. The fact that this modulation already operates at such an early stage  
411 likely enhances its ability to impact subsequent, higher-level processing stages, and as such might  
412 represent an important mechanism that facilitates speech comprehension in challenging listening  
413 situations.

## 414 **4 Methods**

### 415 **4.1 Participants**

416 Ten native English speakers without any knowledge or experience in Korean participated in the current  
417 study (mean age =  $24.0 \pm 2.2$  years; 6 females). Eight participants volunteered in three sessions consisting  
418 of 8 runs each on separate days (intervals in days: mean = 8.5, standard deviation = 16.6, min = 1, max  
419 = 70) and two other participants in a single session each (6 runs and 8 runs, respectively), resulting in a  
420 total of 24 scanning sessions. This is on par with similar approaches that maximize intra-subject reliability  
421 over intra-subject variability in the data (Breedlove et al., 2020; Huth et al., 2016; Kay et al., 2008; Moerel  
422 et al., 2013; Naselaris et al., 2015; Norman-Haignere et al., 2015; Santoro et al., 2017).

423 All participants were recruited via the Brain Imaging and Analysis Center (BIAC) at Duke University  
424 Medical Center, NC, USA after safety screening for MRI (e.g., free of metal implants and claustrophobia).  
425 All reported to have normal hearing and no history or presence of neurological or psychiatric disorders.

## Linguistic Modulation of Phoneme Encoding

426 Informed written consent was obtained from all participants prior to the study in compliance with the  
427 protocols approved by the Duke University Health System Institutional Review Board.

### 428 **4.2 Stimuli**

429 Speech stimuli were created from recordings (44,100 Hz sampling rate, 16-bit precision) of four female  
430 bilingual (Korean and English) speakers reading textbooks in either language as in previous studies (Lee  
431 and Overath, in revision; Overath and Paik, 2021). Native English and Korean speakers judged the  
432 recordings as coming from native English and Korean speakers, respectively. Korean was chosen  
433 because of its dissimilarity to English: it shares no etymological roots with English and has different  
434 syntactic and phonetic structures (Sohn, 2001).

435 We used a modified version of the quilting algorithm (Lee and Overath, in revision; Overath et al., 2015)  
436 where we pseudorandomized the order of phonemes (instead of set segment lengths). First, phonemes  
437 were extracted from the recordings and corresponding transcripts using the Penn Phonetic Lab Forced  
438 Aligner<sup>1</sup> (Yuan and Liberman, 2008) for English speech and the Korean Phonetic Aligner<sup>2</sup> (Yoon and Kang,  
439 2013) for Korean speech. The phoneme segmentation output was a Praat TextGrid, which was then  
440 imported to MATLAB<sup>3</sup> via the mPraat toolbox (Bořil and Skarnitzl, 2016). The alignment was manually  
441 validated by a native English and Korean speaker, respectively (Lee and Overath, in revision; Overath and  
442 Paik, 2021). The durations of phonemes in the recordings of natural speech in milliseconds were as  
443 follows (see **Supplementary Figure S1** for histograms): min = 4.3, max = 396.2, mean = 72.8, median =  
444 63.8, standard deviation = 41.7, skewness = 1.2 in English ( $N = 10,514$ ); min = 8.9, max = 308.3, mean =  
445 71.9, median = 63.7, standard deviation = 36.0, skewness = 1.3 in Korean ( $N = 10,894$ ). The average  
446 durations were similar between languages (0.9 ms longer in English,  $t[21406] = 1.67$ ,  $P = 0.094$ ) while the  
447 distributions were slightly different for that English had more instances of short (e.g., < 20 ms) phonemes  
448 (Kolmogorov-Smirnov statistic = 0.1413,  $P = 10^{-93}$ ).

449 The phoneme segments were pseudorandomly rearranged to create novel phoneme quilts. For each  
450 stimulus, a random initial phoneme was chosen; subsequent phonemes were selected such that the  
451 acoustic change at the boundary was as close as possible to the acoustic change in the original source  
452 signal (using the L2-norm metric of an ERB-spaced cochleogram; see (Overath et al., 2015)). In addition,  
453 we applied the following exclusion criteria: (a) the phoneme duration needed to be at least 20 ms, (b) two  
454 identical phonemes could not occur next to each other, (c) for a given phoneme, its subsequent phoneme  
455 could not be the same as in the original source signal. We used the pitch-synchronous overlap-add  
456 (PSOLA) algorithm (Moulines and Charpentier, 1990) to further minimize abrupt changes in pitch at  
457 phoneme boundaries. Overall alterations due to the quilting algorithm were quantified by the Kullback-  
458 Leibler divergence ( $D_{KL}$ ) between L2-norm acoustic change distributions in the original source and the  
459 created phoneme quilt (median  $D_{KL} = 0.6873$  bits for English, 0.6004 bits for Korean; Wilcoxon rank sum  
460 equal median test:  $Z = 0.5913$ ,  $P = 0.5543$ ). In the phoneme quilts, the durations of phonemes in  
461 milliseconds were as follows (see **Supplementary Figure S1** for histograms): min = 20.0, max = 351.0,

---

<sup>1</sup> [https://babel.ling.upenn.edu/phonetics/old\\_website\\_2015/p2fa/index.html](https://babel.ling.upenn.edu/phonetics/old_website_2015/p2fa/index.html)

<sup>2</sup> <https://korean.utsc.utoronto.ca/kpa/>

<sup>3</sup> <https://github.com/bbTomas/mPraat>

462 mean = 72.3, median = 63.0, standard deviation = 39.4, skewness = 1.4 in English ( $N = 10,467$ ); min =  
463 20.0, max = 383.0, mean = 69.7, median = 60.0, standard deviation = 36.3, skewness = 1.5 in Korean ( $N$   
464 = 11,213). There were slight differences between languages in means (2.6 ms longer in English,  $t[21678]$   
465 = 5.08,  $P = 10^{-7}$ ) and distributions (KS-stat = 0.0657,  $P = 10^{-21}$ ), however, the mean difference of 2.6 ms  
466 is much shorter than the modeled cochlear integration time-window of 20 ms.

467 For both languages (English and Korean), the 33-s long stimuli in the two experimental conditions  
468 (Original and Phoneme Quilts) were created by concatenating six 5.5-s stimuli (24 unique exemplars per  
469 condition and language). Subsequent 5.5-s stimuli were either from the same or a different speaker  
470 (participants were asked to detect changes in the speaker, see **Section 4.3**). The overall sound intensity  
471 was normalized by equalizing the root-mean-square (RMS) signal intensity across stimuli. At the  
472 beginning and the end of the 33-s stimuli, 10-ms cosine ramps were applied to avoid abrupt intensity  
473 changes.

### 474 **4.3 Experimental procedure**

475 Functional MRI data were acquired while participants listened to the speech stimuli (either Original or  
476 Phoneme Quilts in either language) and performed a task to maintain attention to the stimuli. A 33-s trial  
477 consisted of six 5.5-s stimuli of multiple speakers in a given condition. Silent inter-stimulus intervals (ISIs)  
478 were uniformly varied between 5.6 s and 10.4 s (mean = 8 s). One run consisted of twelve 33-s trials, and  
479 one session consisted of eight 8.5-min runs (except for one participant, who only completed six runs).  
480 For one of the eight participants with three sessions, one run was prematurely terminated after 9 of the  
481 12 trials due to technical difficulties (the intact 9 trials from the run were still used in the analysis). In total,  
482 fMRI data corresponding to ~203 min/participant were obtained for the 8 participants with 3 full sessions  
483 (average of ~174 min/participant for all 10 participants); this corresponds to ~158 minutes of stimuli  
484 (excluding the ISI) per participant with 3 full sessions (average of ~137 min/participant for all 10  
485 participants).

486 The stimulus presentation timing was controlled via the Psychophysics Toolbox (v3.0.11<sup>4</sup>). Each run  
487 was triggered by the TTL signal from the MRI scanner mediated by a counter. Digital auditory signals at  
488 44,100 Hz sampling rate and 16-bit precision from a Windows desktop were converted to analog signals  
489 by an external digital amplifier (Sony, Tokyo, Japan) and delivered to participants via MRI-compatible  
490 insert earphones (S14, Sensimetrics, MA, USA) at a comfortable listening level (~75 dB SPL). Participants  
491 wore protective earmuffs on top of the earphones to further reduce acoustic noise from the MRI scanner.

492 The task was to indicate a change in speaker (i.e., a 5.5. s stimulus of one speaker followed by a  
493 different speaker) via a button press on an MRI-compatible four-button pad (average speaker changes  
494 per trial = 3.5, between 1 and 4). The performance was assessed via d-prime  $d' = \Phi^{-1}(\Pr(Y|s)) -$   
495  $\Phi^{-1}(\Pr(Y|n))$  where  $\Pr(Y|s)$  is the hit rate in “signal” trials and  $\Pr(Y|n)$  is the false alarm rate in “noise”  
496 trials and  $\Phi^{-1}(\cdot)$  is the inverse cumulative distribution function of the zero-mean, unit-variance Gaussian  
497 distribution (Macmillan and Kaplan, 1985). Responses were classified as a hit if they occurred within 3 s  
498 following a change in speaker (and otherwise classified as false alarm). In the case of multiple responses

---

<sup>4</sup> <http://psycho toolbox.org/>



## Linguistic Modulation of Phoneme Encoding

499 within one 5.5-s stimulus segment, only the first response was counted. For extreme values of hit/false  
500 alarm rates (i.e., 0 or 1), an adjustment (i.e., adding  $0.5/n$  to zero or subtracting  $0.5/n$  from one for  $n$  trials)  
501 was made to avoid infinite values of  $d'$  (Macmillan and Kaplan, 1985).

502 After each 33-s trial, participants received visual feedback about their performance ( $D' = d'/\max d'$   
503 where  $\max d'$  is a  $d'$  for a perfect performance, ranging between [-100%, 100%]) with a description  
504 (“POOR” for  $D' < 0$ , “FAIR” for  $0 \leq D' < 50\%$ , “GOOD” for  $50\% \leq D' < 100\%$ , “PERFECT!” for  $D' = 100\%$ )  
505 to encourage continued attention. While multiple button presses were discarded from computing  $d'$ , an  
506 alerting message was presented to the participants (“NO KEY PRESSED!” or “TOO MANY KEYS  
507 PRESSED!”) instead of the performance feedback when the button presses were too many ( $> 5$ ) or none  
508 (2.5% of total 2,397 trials from 9 participants; participant 1 was excluded from the  $d'$ -prime analysis due  
509 to a technical fault of the in-scanner response device). The average  $D'$  was  $61.1\% \pm 38.4\%$  points ( $d' =$   
510  $1.14 \pm 0.72$ ), without a significant difference between languages (repeated-measures ANOVA,  $\eta_p^2 = 0.43$ ,  
511  $F[1,7] = 5.46$ ,  $P = 0.21$ ), but between original speech and phoneme quilts ( $\eta_p^2 = 0.70$ ,  $F[1,7] = 16.37$ ,  $P =$   
512  $0.02$ ).

### 513 **4.4 Image acquisition**

514 All images were acquired using a GE MR 750 3.0 Tesla scanner (General Electric, Milwaukee, WI, USA)  
515 with a 32-channel head coil system at the Duke University Hospital, NC, USA. For blood-oxygen-level-  
516 dependent (BOLD) contrast, gradient-echo echo-planar imaging (GE-EPI) with a simultaneous multi-slice  
517 (SMS) acceleration factor of 3 (i.e., 3 slices acquired in parallel with aliasing of FOV/3 shifts) was used (in-  
518 plane pixel size =  $2 \times 2$  mm<sup>2</sup>, slice thickness = 2 mm, FOV = 256 mm, matrix size =  $128 \times 128$ , TE = 30  
519 ms, flip angle =  $73^\circ$ , TR = 1200 ms, phase-encoding direction = posterior-to-anterior). A total of 39 slices  
520 were acquired for each volume (13 slices per band) in an interleaved ascending sequence. At the  
521 beginning of a run, the volume was centered on the supratemporal plane, covering from the inferior  
522 colliculus to the inferior frontal gyrus. To correct for magnetic inhomogeneity artifacts, an additional GE-  
523 EPI image of 3 volumes with a reversed phase encoding direction (posterior-to-anterior) was acquired  
524 after each run except for the first participant.

525 For T1-weighted contrast, a magnetization prepared rapid gradient echo (MP-RAGE) scan covering  
526 the whole-brain (in-plane pixel size =  $1 \times 1$  mm<sup>2</sup>, slice thickness = 1 mm, FOV = 256 mm, matrix size =  
527  $256 \times 256$ , TE = 3.2 ms, flip angle =  $8^\circ$ , TR = 2264 ms, number of slices = 156) was acquired at the end  
528 of each session.

### 529 **4.5 Image processing**

#### 530 **4.5.1 Anatomical images**

531 T1-weighted images were segmented using SPM (SPM12; v7487<sup>5</sup>) to obtain tissue probability maps  
532 (`spm.spatial.preproc`), which were used for anatomical CompCor regressors (Behzadi et al., 2007).

---

<sup>5</sup> <https://www.fil.ion.ucl.ac.uk/spm/>

533 High-resolution cortical surfaces were fully automatically constructed using FreeSurfer (v6.0.0<sup>6</sup>) for  
534 surface-based analysis.

## 535 **4.5.2 Functional images**

536 The displacement due to inhomogeneity in the B0 field (i.e., susceptibility artifacts) was corrected using  
537 topup in FSL (v5.0.11<sup>7</sup>) with the reversed phase-encoding images. The first 6 volumes (i.e., “dummy  
538 scans”) were subsequently discarded from the analyses. Temporal and spatial realignments were  
539 achieved using SPM: the slices were first temporally aligned to the center of the TR using sinc-  
540 interpolation (`spm.temporal.st`), and then the volumes were spatially aligned to the mean volume using  
541 4-th degree B-spline interpolation (`spm.spatial.realignunwarp`). Since we used a multiband sequence  
542 (i.e., 3 slices were acquired simultaneously), slice acquisition time and reference time were provided  
543 (instead of slice order) for slice-timing correction.

544 Anatomical CompCor regressors were extracted from realigned EPI volumes. On concatenated time  
545 series from voxels with > 99% probability for white matter and cerebrospinal fluid, principal component  
546 analysis (PCA) was applied to extract principal components. Six components with highest eigenvalues  
547 were used as “CompCor” regressors in the GLM denoising procedure (see **Section 4.5.3**).

548 Next, the EPI volumes were projected onto individual cortical surfaces (~150,000 vertices per  
549 hemisphere) at the middle depth of cortices by averaging samples at the 40%, 50%, and 60% of cortical  
550 thickness to avoid aliasing (`mri_vol2surf` in FreeSurfer). Surface-mapped functional data were  
551 normalized to ‘fsaverage6’ surfaces (40,962 vertices per hemisphere) via spherical surface registration,  
552 and then smoothed with a 2-D Gaussian kernel with the full-width-at-half-maximum (FWHM) of 6 mm (i.e.,  
553 3 pixels of the EPI slices) via iterative nearest-neighbor averaging (`mri_surf2surf` in FreeSurfer).

## 554 **4.5.3 Surface-based GLM denoising**

555 We applied a model-based denoising technique for task-based fMRI data (“GLMdenoise” v1.4<sup>8</sup>) to the  
556 surface-mapped data (Kay et al., 2013). The algorithm extracts ‘noise’ regressors from the data that would  
557 increase prediction accuracy in leave-one-run-out-cross-validation. This is achieved by first defining  
558 ‘noise pool’ voxels with negative R<sup>2</sup> values for a given design matrix (i.e., voxels that are irrelevant to the  
559 task of interest), extracting principal components from the noise pool, and then determining an optimal  
560 number of components to remove as a minimal number where the improvement in cross-validation  
561 prediction decays. We used box-car functions to represent the four conditions in the design matrix. On  
562 average,  $4.5 \pm 2.1$  noise regressors were regressed out. These improved reliability in estimation (mean  
563 over standard errors ratio of coefficients estimates across CV folds: median increase = 0.82; mean  
564 increase = 1.12) but only slightly increased prediction accuracy (cross-validation R<sup>2</sup>: median increase =  
565 0.25% points; mean increase = 0.56% points). In addition to the noise regressors, the 4-th order  
566 polynomial fits to slow drifts in BOLD time series, the six CompCor regressors, and the button-press

---

<sup>6</sup> <http://freesurfer.net/>

<sup>7</sup> <https://fsl.fmrib.ox.ac.uk/>

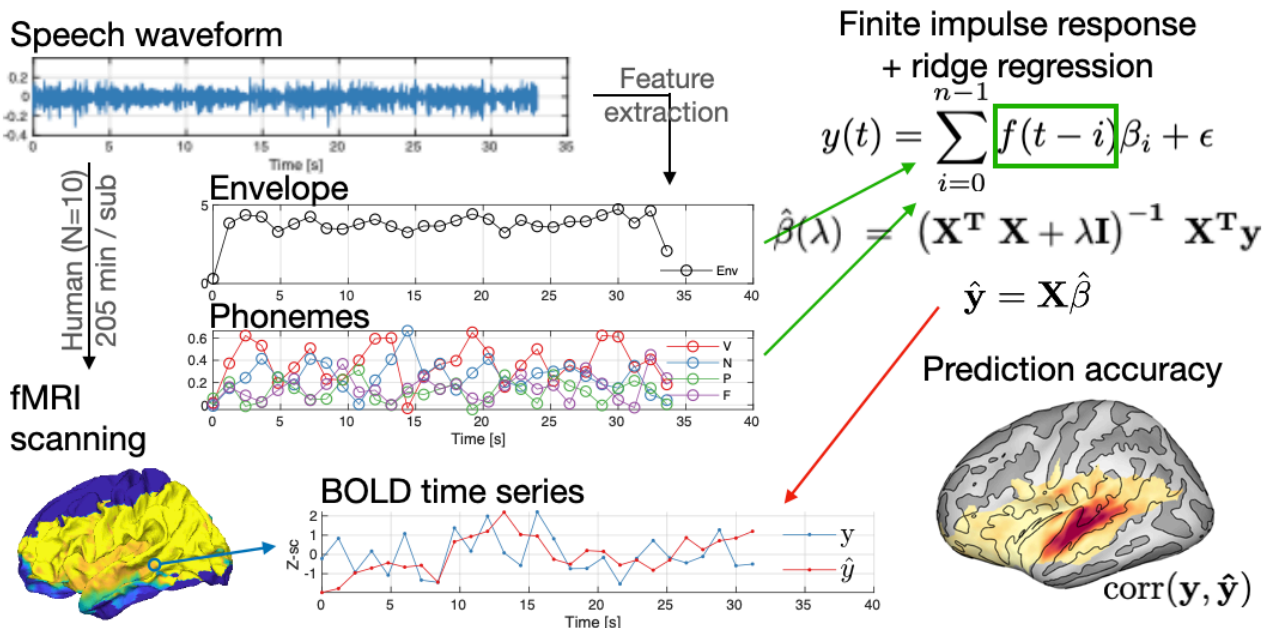
<sup>8</sup> <https://kendrickkay.net/GLMdenoise/>

## Linguistic Modulation of Phoneme Encoding

567 regressors convoluted with a canonical HRF were regressed out from the residuals (i.e., prediction from  
568 the design matrix subtracted from the data).

### 569 **4.6 Voxel-wise linearized encoding analysis**

570 We predicted BOLD time series at each voxel in response to speech sounds using a linearized encoding  
571 model based on finite-impulse response (FIR) functions. Multiple lags were used to model the variable  
572 hemodynamic responses in different cortical areas (De Heer et al., 2017; Huth et al., 2016). In order to  
573 account for the collinearity of predictors representing acoustic and phonetic information, we used ridge  
574 regression to fit the model (i.e., FIR weights) and evaluated the prediction via cross-validation. The  
575 procedures are explained in detail in the following subsections.



576

577 **Figure 4. Linearized encoding analysis overview. Functional MRI data was acquired**  
578 **from 10 human participants while listening to unmanipulated, or phoneme-scrambled**  
579 **speech stimuli in either English or Korean. From the speech waveform, cochleogram**  
580 **envelope and the duration of phoneme classes were extracted and down-sampled at**  
581 **the fMRI sampling rate (1/1.2 Hz). After preprocessing, surface-mapped BOLD time**  
582 **series was predicted using a linear method (finite impulse response modeling with**  
583 **ridge regression). The prediction accuracy was measured by Pearson correlation**  
584 **between actual and predicted BOLD time series.**

#### 585 **4.6.1 Vertex selection**

586 For our interest in auditory and linguistic processing, we restricted our analysis to vertices in cortical  
587 regions that are previously known to be involved in speech processing so as to avoid unnecessary  
588 computations. Specifically, from the automatic parcellation based on the Desikan-Killiany cortical atlas  
589 (Desikan et al., 2006), the following 19 labels were included: 'bankssts', 'caudalmiddlefrontal',

590 'inferiorparietal', 'inferiortemporal', 'lateralorbitofrontal', 'middletemporal', 'parsopercularis',  
591 'parsorbitalis', 'parstriangularis', 'postcentral', 'precentral', 'rostralmiddlefrontal', 'superiorparietal',  
592 'superiortemporal', 'supramarginal', 'frontalpole', 'temporalpole', 'transversetemporal', 'insula'. Selected  
593 regions are visualized in **Supplementary Figure S2**. Vertices with BOLD time series varied across  
594 participants due to the variability of head sizes, individual acquisition volumes at each session, and  
595 movements across runs during sessions. **Supplementary Figure S3** shows the overlap of selected  
596 vertices across participants. On average,  $28,297 \pm 3,748$  vertices were selected per participant.

#### 597 **4.6.2 Predictors**

598 We included as predictors (i) the durations of phoneme classes (vowels, nasals and approximants,  
599 plosives, fricatives and affricatives; Vo, Na, Pl, Fr, respectively), and (ii) the speech envelope (En). For (i),  
600 the onset time and duration of each phoneme were determined and then grouped according to phoneme  
601 class (Ladefoged and Johnstone, 2015; Shin, 2015) (see **Supplementary Table S1**). Bigram transition  
602 probabilities between phoneme classes (**Supplementary Figure S4**) were effectively altered by the  
603 quilting algorithm (Hotelling's  $T^2$  between Original and Phoneme-quilts = 1563,  $P < 10^{-6}$  for English;  
604 Hotelling's  $T^2 = 1258$ ,  $P < 10^{-6}$  for Korean). The durations of phoneme classes were modelled as box-car  
605 functions at the audio sampling rate (44.1 kHz) and were then down-sampled to  $1/TR$  ( $1/1.2 = 0.833$  Hz)  
606 following anti-aliasing low-pass filtering. To align with the slice timing correction applied to the BOLD  
607 time series, the resampled time points were also at the center of the TR. For (ii), the speech envelope was  
608 computed from a cochleogram (30 filters from 20 to 10,000 Hz, equally spaced on an equivalent  
609 rectangular bandwidth [ERB] scale) by raising the Hilbert envelope of the resulting cochleogram to a  
610 power of 0.3 to simulate cochlear compression and summing energy across all 30 ERB channels  
611 (McDermott and Simoncelli, 2011; Overath et al., 2015). The speech envelope was then down-sampled  
612 as for the phoneme class durations.

613 The down-sampled predictors showed strong collinearity; the square root of the maximum eigenvalue  
614 divided by the minimum eigenvalue of the design matrix (i.e., the "condition index") was 35, which is  
615 higher than a "diagnostic" criterion ( $> 30$ ) for a "moderate" multicollinearity (Belsley, 1991). This was due  
616 to the high dependency between the vowel and the envelope predictors; the proportions of explained  
617 variance by the corresponding eigenvector (i.e., variance decomposition proportion; VDP) were 0.93 and  
618 0.99 for the vowels and the envelope, respectively. The collinearity patterns were similar across  
619 conditions (**Supplementary Figure S5**). The existence of multicollinearity motivated the use of a  
620 penalized regression.

#### 621 **4.6.3 Finite-impulse response modelling**

622 A FIR model was used to predict the BOLD time series at each voxel. In this approach, we modelled the  
623 neural response as a convolution of the predictors and a linear FIR filter, which is a commonly used  
624 approach in receptive field mapping of neural populations (Ringach et al., 1997; Wu et al., 2006).

625 Consider a linear model for  $t$  time points and  $p$  predictors,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

## Linguistic Modulation of Phoneme Encoding

626 where  $\mathbf{X}$  is a  $(t \times p)$  design matrix (i.e., a FIR model),  $\mathbf{y}$  is a  $(t \times 1)$  data vector (i.e., BOLD time series at  
 627 a certain voxel),  $\boldsymbol{\beta}$  is a  $(t \times p)$  unknown coefficient vector, and  $\boldsymbol{\varepsilon}$  is a noise vector from a zero-mean  
 628 Gaussian distribution with a serial correlation  $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Omega})$  where  $\boldsymbol{\Omega}$  is a  $(t \times t)$  unknown covariance  
 629 matrix and  $\sigma^2$  is a scale factor. For the FIR modeling, the design matrix  $\mathbf{X}$  consists of matrices of delayed  
 630 features as:

$$631 \quad \mathbf{X} = [f_1 \quad f_2 \quad \cdots \quad f_p] * \mathbf{H}(n),$$

632 for  $p$  features and  $n$  delays as implemented in a convolutional kernel  $\mathbf{H}(n)$ , while  $*$  denotes the  
 633 convolution operation. The actual features tested are explained in Model Comparison (**Section 4.6.6**). A  
 634 Toeplitz matrix can be constructed for delayed features between time point  $t_1$  and  $t_2$  with  $n$  delays for the  
 635  $i$ -th feature as:

$$636 \quad f_i(t_1, t_2) * \mathbf{H}(n) = \begin{bmatrix} f_i(t_1) & f_i(t_1 - 1) & \cdots & f_i(t_1 - (n - 1)) \\ f_i(t_1 + 1) & f_i(t_1) & \cdots & f_i(t_1 - n) \\ \vdots & \vdots & \cdots & \vdots \\ f_i(t_2) & f_i(t_2 - 1) & \cdots & f_i(t_2 - (n - 1)) \end{bmatrix},$$

637 where  $f_i(t)$  is the scalar value of the  $i$ -th predictor at time point  $t$ . In the current study, we delayed the  
 638 predictors by 0, 1, ..., 20 TRs (0, 1.2, ..., 24 s). Once unknown coefficients (or weights) are estimated, an  
 639 inner product  $\mathbf{X}\hat{\boldsymbol{\beta}}$  is effectively a convolution of the  $i$ -th feature and the estimated filter.

### 640 4.6.4 Model estimation

641 While it is standard to pre-whiten the data when modeling autocorrelated noise for a Generalized Least  
 642 Squares (GLS) solution (Aitken, 1936), here we did not pre-whiten the model. This is because even with  
 643 autocorrelated noise, an Ordinary Least Squares (OLS) solution is still an unbiased estimator (only its  
 644 efficiency is suboptimal) and because our goal was to estimate (predict) responses, not to infer  
 645 significance. In particular, for the current data, GLS often yielded worse cross-validation prediction than  
 646 OLS. Therefore, we empirically determined not to pre-whiten the model.

647 As we detected a strong collinearity among the predictors, we applied  $L_2$ -norm regularization to the  
 648 OLS estimation of Equation (1), which is known as a ridge solution (Hoerl and Kennard, 1970):

$$\hat{\boldsymbol{\beta}}(\lambda) = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}, \quad (2)$$

649 where  $\hat{\boldsymbol{\beta}}(\lambda)$  is a vector of penalized estimates,  $\mathbf{I}$  is an identity matrix, and  $\lambda$  is a ridge penalty term. Note  
 650 that predictors and responses were standardized (i.e., Z-scored) prior to fitting so that estimated weights  
 651 (betas) could be compared across models with different lambdas (Santoro et al., 2014). For the  
 652 optimization of the hyperparameter  $\lambda$ , we used a method called “ridge trace” (Hoerl and Kennard, 1970),  
 653 which finds the smallest  $\lambda$  that returns “stabilized” normalized coefficients so that the bias introduced by  
 654  $\lambda$  can be minimized (i.e., an optimal point in bias-variance tradeoff). To define stability, we used a criterion  
 655 from Santoro et al. (2014). That is, for a given equation with  $p$  coefficients  $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_p]^T$ , we  
 656 determined an optimal lambda  $\lambda^*$  from a set of non-zero, incremental lambdas  $\{\lambda_1, \lambda_2, \dots, \lambda_k\}$  as the  
 657 smallest  $\lambda$  such that an increment of  $\lambda$  results in changes in all coefficient  $\hat{\beta}_i(\lambda)$  smaller than 20% of their  
 658 initial values  $\hat{\beta}_i(\lambda_1)$  for all  $p$  coefficients:

$$\lambda^* = \lambda: \Delta \hat{\beta}_i(\lambda) / \hat{\beta}_i(\lambda_1) < 0.2 \quad \forall \lambda \geq \lambda^*, \forall i \in \{1, 2, \dots, p\}. \quad (3)$$

659 Since we used five features for four conditions and 21 lags, the total number of predictors  $p$  was 420  
660 in the current study. We used a range of lambdas from  $10^{0.5}$  to  $10^{11}$  with multiplicative increment of  $10^{1/3}$ ,  
661 scaled by the number of predictors for the ridge trace method.

#### 662 4.6.5 Model validation

663 We assessed the predictive performance of vertex-wise linearized encoding models via cross-validation  
664 (CV). The runs per session were split into a training set and a test set. Then the coefficients (betas) in  
665 Equation (2) and the hyperparameters (lambdas) in Equation (3) were estimated with the training set. We  
666 predicted the test set with the predictor weights estimated from the training set as:

$$\hat{\mathbf{y}}_{te} = \mathbf{X}_{te} \hat{\boldsymbol{\beta}}_{tr}(\lambda) \quad (4)$$

667 where subscripts ‘tr’ and ‘te’ indicate the training set and test set, respectively.

668 We avoided leave-one-out-CV because of concerns regarding high variance due to the under-  
669 representativeness of test sets (Hastie et al., 2009; Poldrack et al., 2020). Instead, we used 2-fold CV  
670 (odd runs and even runs were training and test sets in one fold, and vice versa in another fold). For each  
671 fold, training and test sets consisted of about 4,000 time points (12 runs for each), except participants 1  
672 and 8, whose runs were a total of 6 and 8, respectively. Time points between trials were excluded. We  
673 used Pearson’s correlation ( $r$ ) between the predicted and measured BOLD time series as the performance  
674 metric.

#### 675 4.6.6 Model comparison

676 Our first objective was to demonstrate that phoneme class information is encoded in the BOLD time  
677 series. This was achieved by comparing the prediction accuracies of models with and without phoneme  
678 classes. The design matrix of the full model with phoneme classes (*Phon*) and envelope (*Env*) while  
679 encoding all four conditions can be written as:

$$\mathbf{X}_{Phon+Env,C} = [\mathbf{C} \otimes \mathbf{F}_{Phone+Env}] * \mathbf{H}(n), \quad (5)$$

680 where  $\mathbf{C}$  is a  $(t \times 4)$  matrix of dummy predictors of four conditions as  $\mathbf{C} = [\mathbf{c}_{EP} \quad \mathbf{c}_{EO} \quad \mathbf{c}_{KP} \quad \mathbf{c}_{KO}]$ ,  $\otimes$  is the  
681 Kronecker product operator,  $\mathbf{F}$  is a  $(t \times 5)$  matrix of predictors of five features as  $\mathbf{F}_{PE} =$   
682  $[\mathbf{f}_{Vo} \quad \mathbf{f}_{Na} \quad \mathbf{f}_{PI} \quad \mathbf{f}_{Fr} \quad \mathbf{f}_{En}]$ , and  $\mathbf{H}(n)$  is a kernel that imposes  $n$  delays. The subscripts denote the four  
683 experimental conditions (EP, English-Phoneme quilt; EO, English-Original; KP, Korean-Phoneme quilt;  
684 KO, Korean-Original) and features (Vo, vowels; Na, nasals and approximants; PI, plosives; Fr, fricatives  
685 and affricatives; En, envelope). The Kronecker product with dummy predictors for conditions creates  
686 condition-specific predictors (5 features in each of four conditions, 20 in total), and the convolution  
687 produces a total of 420 predictors (20 predictors x 21 lags).

## Linguistic Modulation of Phoneme Encoding

688 For model comparisons, we created a reduced model without the Phoneme predictors and only with  
 689 the Envelope predictor (84 predictors = 1 feature × 4 conditions × 21 lags) by replacing the feature matrix  
 690  $\mathbf{F}$  in Eq. (5), which can then be written as:

$$\mathbf{X}_{Env,C} = [\mathbf{C} \otimes \mathbf{F}_{Env}] * \mathbf{H}(\mathbf{n}),$$

691 where  $\mathbf{F}_{Env} = [f_{En}]$ . The null and alternative hypotheses for the encoding of phonemes can be formulated  
 692 as:

$$\begin{cases} H_0: \mathbb{E}(r_{Phon+Env,C+}) \leq \mathbb{E}(r_{Env,C}) \\ H_A: \mathbb{E}(r_{Phon+Env,C+}) > \mathbb{E}(r_{Env,C}) \end{cases}'$$

693 where  $\mathbb{E}$  is the expectation and  $r$  is the model performance with respect to Pearson's correlation  
 694 coefficient. In other words,  $H_0$  would be rejected if the prediction accuracy is greater in the full model  
 695 than in the reduced model without phonemes, since this indicates that the addition of phonemes in the  
 696 full model improves model performance. Otherwise,  $H_0$  can not be rejected.

697 Similarly, we constructed another reduced model without the Envelope predictor and only with the  
 698 Phoneme predictors (336 predictors = 4 features × 4 conditions × 21 lags) by replacing the feature matrix  
 699  $\mathbf{F}$  in Eq. (5):

$$\mathbf{X}_{Phon,C} = [\mathbf{C} \otimes \mathbf{F}_{Phon}] * \mathbf{H}(\mathbf{n}),$$

700 where  $\mathbf{F}_{Phon} = [f_{Vo} \ f_{Na} \ f_{Pl} \ f_{Fr}]$ . The hypotheses for the encoding of the speech envelope can then  
 701 be formulated similarly as outlined above. Note that in cross-validation, since the noise in the training and  
 702 test sets is independent, an increase in model complexity by additional predictors does not necessarily  
 703 lead to an increase of  $r$  (unlike overfitting to noise in the training set), unless the newly added predictors  
 704 capture certain activity that is common to both training and test data sets (Hastie et al., 2009; Kriegeskorte  
 705 et al., 2009; Varoquaux et al., 2017).

706 Our second objective was to test the contributions of the Language (English, Korean) and Quilting  
 707 (Phoneme quilt, Original) factors. This was again achieved by comparing the prediction accuracy  $r$  of a  
 708 full model with a reduced model. As for the main effect of Language, we would need to compare the full  
 709 model with a reduced model that does not capture the effect of language (i.e., the main effect of Language  
 710 = (Language + Quilting) - (Quilting) = Language). Specifically, a reduced model only with quilting-specific  
 711 predictors ("Quilting-only" model) was created by replacing the condition matrix  $\mathbf{C}$  in Eq. (5) as:

$$\mathbf{X}_{Phon+Env,Q} = [\mathbf{C}_Q \otimes \mathbf{F}_{Phon+Env}] * \mathbf{H}(\mathbf{n}),$$

712 where  $\mathbf{C}_Q = [c_P \ c_O]$  with dummy predictors for Phoneme quilt ( $c_P$ ) and Original ( $c_O$ ). Note that in  $\mathbf{C}_Q$ , the  
 713 English-Phoneme and Korean-Phoneme conditions would be modeled by a single dummy variable  $c_P$ ,  
 714 and the English-Original and Korean-Original would be collapsed into  $c_O$ . Similarly, for the main effect of  
 715 Quilting, a reduced model only with language-specific predictors ("Language-only" model) was  
 716 constructed by replacing the condition matrix  $\mathbf{C}$  in Eq. (5) as:

$$\mathbf{X}_{Phon+Env,L} = [\mathbf{C}_L \otimes \mathbf{F}_{Phon+Env}] * \mathbf{H}(\mathbf{n}),$$

717 where  $\mathbf{C}_L = [\mathbf{c}_E \ \mathbf{c}_K]$  with dummy predictors for English ( $\mathbf{c}_E$ ) and Korean ( $\mathbf{c}_K$ ). The logic is again that, if the  
718 full model performs better than an alternative reduced model (e.g., Language-ignored model), the  
719 improvement in information encoding can be attributed to the ignored factor (e.g., Language). More  
720 specifically, if the true FIR kernel for the phoneme class vowels was different when listening to Korean vs.  
721 English, modeling them together with a common predictor would result in the loss of predictive power.  
722 Therefore, the difference between the full model and the Language-ignored model can be interpreted as  
723 a main effect of Language, while the difference between the full model and the Quilting-ignored model  
724 can be interpreted as a main effect of Quilting. We further estimated the interaction between Language  
725 and Quilting by comparing the effect of Quilting estimated from subsets of the data: English conditions  
726 and Korean conditions, separately.

727 Statistical inference was computed via a non-parametric paired  $t$ -test using a cluster-based  
728 permutation test at group-level (Maris and Oostenveld, 2007). Specifically,  $r$  values of both models were  
729 calculated for each participant ( $N = 10$ ), and then the difference between two models at each vertex was  
730 calculated. Next, the signs of differences across participants were flipped over all possible permutations  
731 ( $2^{10} = 1,024$ ) to form a null distribution. One-tailed  $P$ -values were computed from the null distribution as  
732 we would regard a decrease of prediction accuracy as a non-significant encoding of the information as  
733 well as non-significant changes of prediction accuracy. Note that the inference was computed at the  
734 group-level, not the subject-level. Even with overlapping models in nested models, it is possible that the  
735 prediction could worsen due to the penalization introduced by additional variables. Bootstrapped 95%  
736 confidence intervals were computed for  $r$  differences (10,000 bootstrapping). Vertex-wise multiple  
737 comparisons correction was applied via a cluster-based permutation test as implemented in  
738 `ft_statistics_montecarlo.m` in FieldTrip (v20180903)<sup>9</sup> with a custom modification of `clusterstat.m`  
739 for a faster cluster identification through parallelization. In an earlier fMRI methodological study (Eklund  
740 et al., 2016), it was shown that a liberal cluster-forming threshold (CFT) in a cluster-level inference based  
741 on the random field theory resulted in a severely inflated family-wise error rate (FWER), whereas the  
742 permutation test showed a consistent, proper control of the FWER regardless of the choice of a CFT. A  
743 recent study formally showed that a CFT in permutation tests does not affect the FWER, but only the  
744 sensitivity (Maris, 2019). Thus, in the current study, clusters were defined by an arbitrary threshold of the  
745 alpha-level of 0.05 (for vertex-wise  $P$ -values) to improve the sensitivity, and the cluster-wise  $P$ -values are  
746 thresholded at the alpha-level of 0.05 to control the FWER at 0.05.

#### 747 **4.6.7 Phoneme-class-specific effects**

748 We further investigated the specific contribution of each phoneme class by comparing a full model with  
749 all phoneme classes (Eq. 5) with reduced models without a particular phoneme class. For example, a  
750 model without nasals would be  $\mathbf{X}_{Phon_{Na-+Env,C+}} = [\mathbf{C} \otimes \mathbf{F}_{Phon_{Na-+Env}}] * \mathbf{H}(\mathbf{n})$ , where  $\mathbf{F}_{Phon_{Na-+Env}} =$   
751  $[\mathbf{f}_{Vo} \ \mathbf{f}_{Pl} \ \mathbf{f}_{Fr} \ \mathbf{f}_{En}]$ . From such comparisons, we constructed a  $(1 \times 4)$  vector (“phoneme encoding  
752 vector”) of the changes of prediction accuracies for the four phoneme classes  $\mathbf{d} =$   
753  $[\Delta r_{Vo} \ \Delta r_{Pl} \ \Delta r_{Na} \ \Delta r_{Fr}]$  at each vertex. From these vertex-wise vectors, a winner-take-all phoneme

---

<sup>9</sup> <http://www.fieldtriptoolbox.org/>



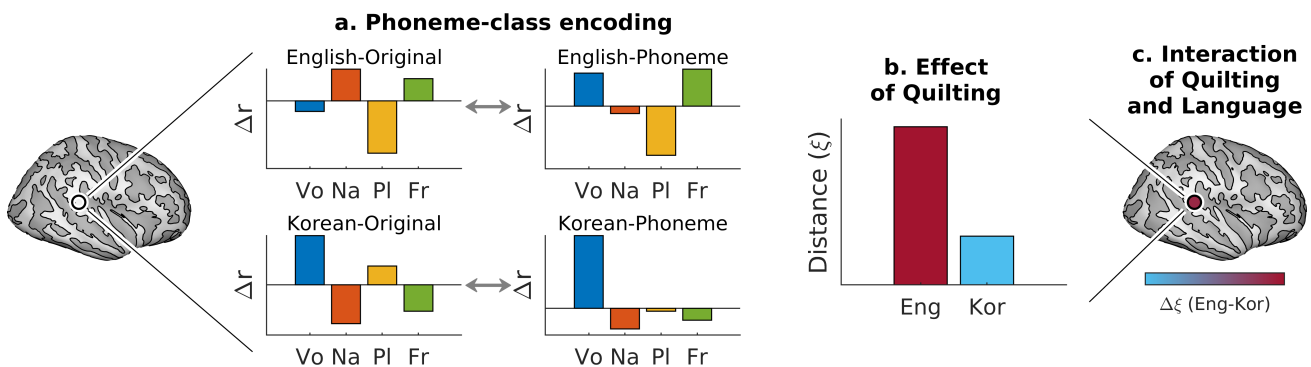
## Linguistic Modulation of Phoneme Encoding

754 class map was created (at each vector with a significant prediction accuracy increase [cluster- $P < 0.05$ ]  
 755 for at least one phoneme class).

756 Subsequently, we tested whether the phoneme encoding vectors were modulated differentially by  
 757 Language (English, Korean) and Quilting (Phoneme-quilt, Original). Specifically, for an interaction between  
 758 Language and Quilting, we expected that a multivariate dissimilarity of phoneme encoding patterns  
 759 between English-Original and English-Phoneme levels would be greater than that between Korean-  
 760 Original and Korean-Phoneme levels. Put differently, we expected the phoneme encoding patterns to be  
 761 more similar (i.e., constant) between the Korean pairs than the English pairs. The corresponding null and  
 762 alternative hypotheses can be expressed formally as:

$$\begin{cases} H_0: \mathbb{E}(\xi(\mathbf{d}_{EP}, \mathbf{d}_{EO})) \leq \mathbb{E}(\xi(\mathbf{d}_{KP}, \mathbf{d}_{KO})) \\ H_A: \mathbb{E}(\xi(\mathbf{d}_{EP}, \mathbf{d}_{EO})) > \mathbb{E}(\xi(\mathbf{d}_{KP}, \mathbf{d}_{KO})) \end{cases}$$

763 where  $\mathbb{E}$  is the expectation and  $\xi$  is a non-directional distance measure, either Pearson distance (1- $r$ ) or  
 764 Euclidian distance (see **Figure 5** for a schematic cartoon). Pearson distance is sensitive to (normalized)  
 765 relative patterns of the vectors, but insensitive to absolute magnitudes. We therefore incorporated the  
 766 Euclidian distance metric to characterize differences in magnitudes as well. We tested the difference  
 767 between Language pairs via one-sample  $t$ -tests using the cluster-based permutation test as described  
 768 above (**Section 4.6.6**).



769

770 **Figure 5. Schematic of the multivariate analysis on phoneme-class encoding vectors.**

771 **(a) For each vertex, its [1x4] phoneme-class encoding vector was defined by the**  
 772 **prediction accuracy changes when adding a particular phoneme class to the model**  
 773 **for each of four conditions (two languages x two quilting conditions). (b) Distance**  
 774 **metrics (Pearson or Euclidian) were computed within each language between the**  
 775 **Original and Phoneme quilt conditions. (c) The difference in this distance metric**  
 776 **between languages (English-minus-Korean; i.e., the interaction of Quilting and**  
 777 **Language) was mapped back to the corresponding vertex (see Figure 3b-c).**

778

779 For visualization, the 4-D phoneme encoding vectors were projected to a 3-D eigenspace using  
 780 principal component analysis (PCA). The first three principal component scores (i.e., eigenvariates)  
 781 were re-scaled between zero and one and mapped to RGB values, respectively. The vertex-mapped RGB  
 782 values were interpolated on the triangular faces of the cortical meshes using MALTAB's patch function.

783 Note that the visualization is intended as an intuitive representation only; the actual comparisons were  
784 made on the scalar distance values comparing 4-D vectors.

#### 785 **4.7 Data availability**

786 The data that support the findings of this study are available from the corresponding authors upon  
787 reasonable request.

#### 788 **4.8 Code availability**

789 The computer code that was used for this study is available on the Open Science Framework repository<sup>10</sup>.

### 790 **5 Acknowledgements**

791 The authors would like to thank Frankie Pennington and Joon Hyun Paik for manually checking phoneme  
792 onsets and offsets for the forced alignment of English and Korean recordings, respectively. The authors  
793 also thank Dr. Thomas E Cope for helpful comments on the earlier version of the manuscript. This work  
794 was supported by US National Institutes of Health grant R21DC016386 to T.O.

### 795 **6 References**

- 796 Aitken, A.C., 1936. On least squares and linear combination of observations. Proceedings of the  
797 Royal Society of Edinburgh. Section B: Biology 55, 42-48.
- 798 Anderson, J.L., Morgan, J.L., White, K.S., 2003. A statistical basis for speech sound discrimination.  
799 Language and Speech 46, 155-182.
- 800 Baltzell, L.S., Srinivasan, R., Richards, V.M., 2017. The effect of prior knowledge and intelligibility on  
801 the cortical entrainment response to speech. Journal of Neurophysiology 118, 3144-3151.
- 802 Baumann, S., Joly, O., Rees, A., Petkov, C.I., Sun, L., Thiele, A., Griffiths, T.D., 2015. The topography  
803 of frequency and time representation in primate auditory cortices. Elife 4, e03256.
- 804 Behzadi, Y., Restom, K., Liau, J., Liu, T.T., 2007. A component based noise correction method  
805 (compcor) for bold and perfusion based fmri. Neuroimage 37, 90-101.
- 806 Belsley, D.A., 1991. A guide to using the collinearity diagnostics. Computer Science in Economics  
807 and Management 4, 33-50.
- 808 Bendor, D., Wang, X., 2005. The neuronal representation of pitch in primate auditory cortex. Nature  
809 436, 1161-1165.

---

<sup>10</sup> [https://osf.io/zgj3m/?view\\_only=cd4942f9ea674d79a5644796d5498e3c](https://osf.io/zgj3m/?view_only=cd4942f9ea674d79a5644796d5498e3c)

## Linguistic Modulation of Phoneme Encoding

- 810 Blank, I.A., Fedorenko, E., 2020. No evidence for differences among language regions in their  
811 temporal receptive windows. *Neuroimage* 219, 116925.
- 812 Bořil, T., Skarnitzl, R., 2016. Tools rPraat and mPraat. *Text, Speech, and Dialogue. TSD 2016.*  
813 *Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 367-374.*
- 814 Breedlove, J.L., St-Yves, G., Olman, C.A., Naselaris, T., 2020. Generative feedback explains distinct  
815 brain activity codes for seen and mental images. *Current Biology* 30, 2211-2224.e2216.
- 816 Carrasco, A., Lomber, S.G., 2009. Evidence for hierarchical processing in cat auditory cortex:  
817 Nonreciprocal influence of primary auditory cortex on the posterior auditory field. *Journal of*  
818 *Neuroscience* 29, 14323-14333.
- 819 Cheour, M., Ceponiene, R., Lehtokoski, A., Luuk, A., Allik, J., Alho, K., Näätänen, R., 1998.  
820 Development of language-specific phoneme representations in the infant brain. *Nature*  
821 *Neuroscience* 1, 351-353.
- 822 Chevillet, M., Riesenhuber, M., Rauschecker, J.P., 2011. Functional correlates of the anterolateral  
823 processing hierarchy in human auditory cortex. *Journal of Neuroscience* 31, 9345-9352.
- 824 Chomsky, N., Halle, M., 1965. Some controversial questions in phonological theory. *Journal of*  
825 *Linguistics* 1, 97-138.
- 826 Cope, T.E., Sohoglu, E., Sedley, W., Patterson, K., Jones, P.S., Wiggins, J., Dawson, C., Grube, M.,  
827 Carlyon, R.P., Griffiths, T.D., Davis, M.H., Rowe, J.B., 2017. Evidence for causal top-down  
828 frontal contributions to predictive processes in speech perception. *Nature Communications*  
829 8, 2154.
- 830 Daube, C., Ince, R.A.A., Gross, J., 2019. Simple Acoustic Features Can Explain Phoneme-Based  
831 Predictions of Cortical Responses to Speech. *Current Biology* 29, 1924-1937.e1929.
- 832 Davis, M.H., Ford, M.A., Kherif, F., Johnsrude, I.S., 2011. Does semantic context benefit speech  
833 understanding through “top-down” processes? Evidence from time-resolved sparse fmri.  
834 *Journal of Cognitive Neuroscience* 23, 3914-3932.
- 835 Davis, M.H., Johnsrude, I.S., 2007. Hearing speech sounds: Top-down influences on the interface  
836 between audition and speech perception. *Hearing Research* 229, 132-147.
- 837 De Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., Theunissen, F.E., 2017. The hierarchical  
838 cortical organization of human speech processing. *Journal of Neuroscience* 37, 6539-6557.
- 839 Desikan, R.S., Segonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale,  
840 A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling  
841 system for subdividing the human cerebral cortex on mri scans into gyral based regions of  
842 interest. *Neuroimage* 31, 968-980.
- 843 DeWitt, I., Rauschecker, J.P., 2012. Phoneme and word recognition in the auditory ventral stream.  
844 *Proceedings of the National Academy of Sciences of the United States of America* 109, E505-  
845 E514.
- 846 Di Liberto, Giovanni M., O’Sullivan, James A., Lalor, Edmund C., 2015. Low-frequency cortical  
847 entrainment to speech reflects phoneme-level processing. *Current Biology* 25, 2457-2465.

- 848 Díaz, B., Baus, C., Escera, C., Costa, A., Sebastián-Gallés, N., 2008. Brain potentials to native  
849 phoneme discrimination reveal the origin of individual differences in learning the sounds of a  
850 second language. *Proceedings of the National Academy of Sciences of the United States of*  
851 *America* 105, 16083-16088.
- 852 Ding, N., Simon, J.Z., 2013. Adaptive temporal encoding leads to a background-insensitive cortical  
853 representation of speech. *The Journal of Neuroscience* 33, 5728-5735.
- 854 Eckert, M.A., Teubner-Rhodes, S., Vaden, K.I., Jr., 2016. Is listening in noise worth it? The  
855 neurobiology of speech recognition in challenging listening conditions. *Ear and Hearing* 37,  
856 101S-110S.
- 857 Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent  
858 have inflated false-positive rates. *Proceedings of the National Academy of Sciences of the*  
859 *United States of America*, 201602413.
- 860 Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. " Who" is saying" what"? Brain-based  
861 decoding of human voice and speech. *Science* 322, 970-973.
- 862 Friederici, A.D., 2009. Pathways to language: Fiber tracts in the human brain. *Trends in Cognitive*  
863 *Sciences* 13, 175-181.
- 864 Friederici, A.D., 2011. The brain basis of language processing: From structure to function.  
865 *Physiological Reviews* 91, 1357-1392.
- 866 Friederici, A.D., Pfeifer, E., Hahne, A., 1993. Event-related brain potentials during natural speech  
867 processing: Effects of semantic, morphological and syntactic violations. *Cognitive Brain*  
868 *Research* 1, 183-192.
- 869 Friederici, A.D., Wessels, J.M.I., 1993. Phonotactic knowledge of word boundaries and its use in  
870 infant speech perception. *Perception and Psychophysics* 54, 287-295.
- 871 Friston, K., Kiebel, S., 2009. Predictive coding under the free-energy principle. *Philosophical*  
872 *Transactions of the Royal Society of London. Series B: Biological Sciences* 364, 1211-1221.
- 873 Giraud, A.L., Kell, C., Thierfelder, C., Sterzer, P., Russ, M.O., Preibisch, C., Kleinschmidt, A., 2004.  
874 Contributions of sensory input, auditory search and verbal comprehension to cortical activity  
875 during speech processing. *Cerebral Cortex* 14, 247-255.
- 876 Griffiths, T.D., Hall, D.A., 2012. Mapping pitch representation in neural ensembles with fmri. *Journal*  
877 *of Neuroscience* 32, 13343-13347.
- 878 Hall, D.A., Plack, C.J., 2009. Pitch processing sites in the human auditory brain. *Cerebral Cortex* 19,  
879 576-585.
- 880 Hamilton, L.S., Huth, A.G., 2020. The revolution will not be controlled: Natural stimuli in speech  
881 neuroscience. *Language, Cognition and Neuroscience* 35, 573-582.
- 882 Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining,*  
883 *Inference, and Prediction.* Springer Science & Business Media.
- 884 Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. *Nature Reviews*  
885 *Neuroscience* 8, 393-402.

## Linguistic Modulation of Phoneme Encoding

- 886 Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems.  
887 *Technometrics* 12, 55-67.
- 888 Holdgraf, C.R., de Heer, W., Pasley, B., Rieger, J., Crone, N., Lin, J.J., Knight, R.T., Theunissen, F.E.,  
889 2016. Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nature*  
890 *Communications* 7, 13654.
- 891 Howard, M.F., Poeppel, D., 2010. Discrimination of Speech Stimuli Based on Neuronal Response  
892 Phase Patterns Depends on Acoustics But Not Comprehension. *Journal of Neurophysiology*  
893 104, 2500-2511.
- 894 Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals  
895 the semantic maps that tile human cerebral cortex. *Nature* 532, 453-458.
- 896 Jusczyk, P.W., Luce, P.A., Charles-Luce, J., 1994. Infants' sensitivity to phonotactic patterns in the  
897 native language. *Journal of Memory and Language* 33, 630-645.
- 898 Kay, K., Rokem, A., Winawer, J., Dougherty, R., Wandell, B., 2013. Glimdenoise: A fast, automated  
899 technique for denoising task-based fmri data. *Frontiers in Neuroscience* 7.
- 900 Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008. Identifying natural images from human  
901 brain activity. *Nature* 452, 352-355.
- 902 Khalighinejad, B., Cruzatto da Silva, G., Mesgarani, N., 2017. Dynamic encoding of acoustic features  
903 in neural responses to continuous speech. *Journal of Neuroscience* 37, 2176-2185.
- 904 Kim, S.-G., Overath, T., Sedley, W., Kumar, S., Teki, S., Patterson, D.R., Griffiths, T.D., in revision.  
905 MEG correlates of periodicity relevant to pitch perception in human auditory cortex.
- 906 Kleinschmidt, D.F., Jaeger, T.F., 2015. Robust speech perception: Recognize the familiar, generalize  
907 to the similar, and adapt to the novel. *Psychological Review* 122, 148-203.
- 908 Kocagoncu, E., Clarke, A., Devereux, B.J., Tyler, L.K., 2017. Decoding the cortical dynamics of  
909 sound-meaning mapping. *Journal of Neuroscience* 37, 1312-1319.
- 910 Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I., 2009. Circular analysis in systems  
911 neuroscience: The dangers of double dipping. *Nature Neuroscience* 12, 535.
- 912 Kujawa, S.G., Liberman, M.C., 2009. Adding insult to injury: Cochlear nerve degeneration after  
913 "temporary" noise-induced hearing loss. *Journal of Neuroscience* 29, 14077-14085.
- 914 Kumar, S., Stephan, K.E., Warren, J.D., Friston, K.J., Griffiths, T.D., 2007. Hierarchical processing of  
915 auditory objects in humans. *PLoS Computational Biology* 3, e100.
- 916 Kutas, M., Hillyard, S.A., 1983. Event-related brain potentials to grammatical errors and semantic  
917 anomalies. *Memory and Cognition* 11, 539-550.
- 918 Ladefoged, P., 2001. *Vowels and Consonants : An Introduction to the Sounds of Languages*. Wiley-  
919 Blackwell.
- 920 Ladefoged, P., Johnstone, K., 2015. *A Course in Phonetics*, Seventh edition. ed. Cengage Learning,  
921 Stamford, CT.
- 922 Lee, J., Overath, T., in revision. The neural analysis of phonemes is shaped by linguistic analysis.

- 923 Leonard, M.K., Baud, M.O., Sjerps, M.J., Chang, E.F., 2016. Perceptual restoration of masked  
924 speech in human cortex. *Nature Communications* 7, 13619.
- 925 Lerner, Y., Honey, C.J., Silbert, L.J., Hasson, U., 2011. Topographic mapping of a hierarchy of  
926 temporal receptive windows using a narrated story. *Journal of Neuroscience* 31, 2906-2915.
- 927 Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in  
928 human auditory cortex. *Neuron* 54, 1001-1010.
- 929 Macmillan, N.A., Kaplan, H.L., 1985. Detection theory analysis of group data: Estimating sensitivity  
930 from average hit and false-alarm rates. *Psychological Bulletin* 98, 185-199.
- 931 Maris, E., 2019. Enlarging the scope of randomization and permutation tests in neuroimaging and  
932 neuroscience. *BioRxiv*, p. 685560.
- 933 Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. *Journal of*  
934 *Neuroscience Methods* 164, 177-190.
- 935 Mattys, S.L., Jusczyk, P.W., 2001. Phonotactic cues for segmentation of fluent speech by infants.  
936 *Cognition* 78, 91-121.
- 937 McDermott, J.H., Simoncelli, E.P., 2011. Sound texture perception via statistics of the auditory  
938 periphery: Evidence from sound synthesis. *Neuron* 71, 926-940.
- 939 Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-  
940 talker speech perception. *Nature* 485, 233-236.
- 941 Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F., 2014. Phonetic feature encoding in human  
942 superior temporal gyrus. *Science* 343, 1006-1010.
- 943 Mesgarani, N., David, S.V., Fritz, J.B., Shamma, S.A., 2008. Phoneme representation and  
944 classification in primary auditory cortex. *Journal of the Acoustical Society of America* 123,  
945 899-909.
- 946 Millman, R.E., Johnson, S.R., Prendergast, G., 2015. The Role of Phase-locking to the Temporal  
947 Envelope of Speech in Auditory Perception and Speech Intelligibility. *Journal of Cognitive*  
948 *Neuroscience* 27, 533-545.
- 949 Moerel, M., De Martino, F., Kemper, V.G., Schmitter, S., Vu, A.T., Uğurbil, K., Formisano, E., Yacoub,  
950 E., 2018. Sensitivity and specificity considerations for fmri encoding, decoding, and mapping  
951 of auditory cortex at ultra-high field. *Neuroimage* 164, 18-31.
- 952 Moerel, M., De Martino, F., Santoro, R., Ugurbil, K., Goebel, R., Yacoub, E., Formisano, E., 2013.  
953 Processing of natural sounds: Characterization of multipeak spectral tuning in human  
954 auditory cortex. *Journal of Neuroscience* 33, 11888-11898.
- 955 Moore, B.C.J., 1996. Perceptual consequences of cochlear hearing loss and their implications for  
956 the design of hearing aids. *Ear and Hearing* 17, 133-161.
- 957 Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-  
958 speech synthesis using diphones. *Speech Communication* 9, 453-467.

## Linguistic Modulation of Phoneme Encoding

- 959 Narain, C., Scott, S.K., Wise, R.J.S., Rosen, S., Leff, A., Iversen, S.D., Matthews, P.M., 2003. Defining  
960 a left-lateralized response specific to intelligible speech using fmri. *Cerebral Cortex* 13, 1362-  
961 1368.
- 962 Naselaris, T., Olman, C.A., Stansbury, D.E., Ugurbil, K., Gallant, J.L., 2015. A voxel-wise encoding  
963 model for early visual areas decodes mental images of remembered scenes. *Neuroimage*  
964 105, 215-228.
- 965 Norman-Haignere, S., Kanwisher, Nancy G., McDermott, Josh H., 2015. Distinct cortical pathways  
966 for music and speech revealed by hypothesis-free voxel decomposition. *Neuron* 88, 1281-  
967 1296.
- 968 Norman-Haignere, S.V., Long, L.K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E.M., Feldstein,  
969 N.A., McKhann, G.M., Schevon, C.A., Flinker, A., 2020. Multiscale integration organizes  
970 hierarchical computation in human auditory cortex. *BioRxiv*.
- 971 Nunez-Elizalde, A.O., Huth, A.G., Gallant, J.L., 2019. Voxelwise encoding models with non-spherical  
972 multivariate normal priors. *Neuroimage* 197, 482-492.
- 973 Obleser, J., Eisner, F., Kotz, S.A., 2008. Bilateral speech comprehension reflects differential  
974 sensitivity to spectral and temporal features. *Journal of Neuroscience* 28, 8116-8123.
- 975 Obleser, J., Leaver, A., VanMeter, J., Rauschecker, J., 2010. Segregation of vowels and consonants  
976 in human auditory cortex: Evidence for distributed hierarchical organization. *Frontiers in*  
977 *Psychology* 1, 232.
- 978 Overath, T., McDermott, J.H., Zarate, J.M., Poeppel, D., 2015. The cortical analysis of speech-  
979 specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience* 18,  
980 903-911.
- 981 Overath, T., Paik, J.H., 2021. From acoustic to linguistic analysis of temporal speech structure:  
982 Acousto-linguistic transformation during speech perception using speech quilts.  
983 *Neuroimage*, 117887.
- 984 Overath, T., Zhang, Y., Sanes, D.H., Poeppel, D., 2012. Sensitivity to temporal modulation rate and  
985 spectral bandwidth in the human auditory system: Fmri evidence. *Journal of Neurophysiology*  
986 107, 2042-2056.
- 987 Pandya, P.K., Rathbun, D.L., Moucha, R., Engineer, N.D., Kilgard, M.P., 2007. Spectral and temporal  
988 processing in rat posterior auditory cortex. *Cerebral Cortex* 18, 301-314.
- 989 Park, H., Ince, Robin A.A., Schyns, Philippe G., Thut, G., Gross, J., 2015. Frontal top-down signals  
990 increase coupling of auditory low-frequency oscillations to continuous speech in human  
991 listeners. *Current Biology* 25, 1649-1653.
- 992 Plack, C.J., Oxenham, A.J., Fay, R.R., Popper, A., 2005. *Pitch: Neural Coding and Perception*.  
993 Springer, New York.
- 994 Poeppel, D., Idsardi, W.J., Wassenhove, V.v., 2008. Speech perception at the interface of  
995 neurobiology and linguistics. *Philosophical Transactions of the Royal Society B: Biological*  
996 *Sciences* 363, 1071-1086.
- 997 Poldrack, R.A., Huckins, G., Varoquaux, G., 2020. Establishment of best practices for evidence for  
998 prediction: A review. *JAMA Psychiatry* 77, 534-540.

- 999 Qi, Y., Fox, R.A., 1992. Analysis of nasal consonants using perceptual linear prediction. *Journal of*  
1000 *the Acoustical Society of America* 91, 1718-1726.
- 1001 Rao, R.P.N., Ballard, D.H., 1999. Predictive coding in the visual cortex: A functional interpretation of  
1002 some extra-classical receptive-field effects. *Nature Neuroscience* 2, 79-87.
- 1003 Rauschecker, J.P., Scott, S.K., 2009. Maps and streams in the auditory cortex: Nonhuman primates  
1004 illuminate human speech processing. *Nature Neuroscience* 12, 718-724.
- 1005 Rauschecker, J.P., Tian, B., 2004. Processing of band-passed noise in the lateral auditory belt cortex  
1006 of the rhesus monkey. *Journal of Neurophysiology* 91, 2578-2589.
- 1007 Rauschecker, J.P., Tian, B., Hauser, M., 1995. Processing of complex sounds in the macaque  
1008 nonprimary auditory cortex. *Science* 268, 111-114.
- 1009 Ringach, D.L., Sapiro, G., Shapley, R., 1997. A subspace reverse-correlation technique for the study  
1010 of visual neurons. *Vision Research* 37, 2455-2464.
- 1011 Ruggles, D., Bharadwaj, H., Shinn-Cunningham, B.G., 2011. Normal hearing is not enough to  
1012 guarantee robust encoding of suprathreshold features important in everyday communication.  
1013 *Proceedings of the National Academy of Sciences of the United States of America* 108,  
1014 15516-15521.
- 1015 Rutten, S., Santoro, R., Hervais-Adelman, A., Formisano, E., Golestani, N., 2019. Cortical encoding  
1016 of speech enhances task-relevant acoustic information. *Nature Human Behaviour* 3, 974-  
1017 987.
- 1018 Saenz, M., Langers, D.R.M., 2014. Tonotopic mapping of human auditory cortex. *Hearing Research*  
1019 307, 42-52.
- 1020 Saffran, J.R., Newport, E.L., Aslin, R.N., 1996. Word segmentation: The role of distributional cues.  
1021 *Journal of Memory and Language* 35, 606-621.
- 1022 Samuel, A.G., 1981. Phonemic restoration: Insights from a new methodology. *Journal of*  
1023 *Experimental Psychology: General* 110, 474-494.
- 1024 Samuel, A.G., 1987. Lexical uniqueness effects on phonemic restoration. *Journal of Memory and*  
1025 *Language* 26, 36-56.
- 1026 Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., Formisano, E., 2014.  
1027 Encoding of natural sounds at multiple spectral and temporal resolutions in the human  
1028 auditory cortex. *PLoS Computational Biology* 10.
- 1029 Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., Formisano, E., 2017.  
1030 Reconstructing the spectrotemporal modulations of real-life sounds from fmri response  
1031 patterns. *Proceedings of the National Academy of Sciences of the United States of America*  
1032 114, 4799-4804.
- 1033 Schönwiesner, M., Zatorre, R.J., 2009. Spectro-temporal modulation transfer function of single  
1034 voxels in the human auditory cortex measured with high-resolution fmri. *Proceedings of the*  
1035 *National Academy of Sciences of the United States of America* 106, 14611-14616.
- 1036 Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J.S., 2000. Identification of a pathway for intelligible  
1037 speech in the left temporal lobe. *Brain* 123, 2400-2406.



## Linguistic Modulation of Phoneme Encoding

- 1038 Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with  
1039 primarily temporal cues. *Science* 270, 303-304.
- 1040 Shin, J., 2015. Vowels and consonants. In: Brown, L., Yeon, J. (Eds.), *The handbook of Korean*  
1041 *linguistics*. Wiley-Blackwell, UK, pp. 3-21.
- 1042 Shinn-Cunningham, B.G., Best, V., 2008. Selective attention in normal and impaired hearing. *Trends*  
1043 *in Amplification* 12, 283-299.
- 1044 Sohn, H.-M., 2001. *The Korean Language*. Cambridge University Press, NY.
- 1045 Sohoglu, E., Peelle, J.E., Carlyon, R.P., Davis, M.H., 2012. Predictive top-down integration of prior  
1046 knowledge during speech perception. *Journal of Neuroscience* 32, 8443-8453.
- 1047 Stevens, K.N., 2000. *Acoustic Phonetics*. MIT press.
- 1048 Theunissen, F., Miller, J., 1995. Temporal encoding in nervous systems: A rigorous definition. *Journal*  
1049 *of Computational Neuroscience* 2, 149-162.
- 1050 Vanthornhout, J., Decruy, L., Wouters, J., Simon, J.Z., Francart, T., 2018. Speech Intelligibility  
1051 Predicted from Neural Entrainment of the Speech Envelope. *Journal of the Association for*  
1052 *Research in Otolaryngology* 19, 181-191.
- 1053 Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017.  
1054 Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage*  
1055 145, 166-179.
- 1056 Verschueren, E., Vanthornhout, J., Francart, T., 2021. The effect of stimulus intensity on neural  
1057 envelope tracking. *Hearing Research* 403, 108175.
- 1058 Wang, Y., Ding, N., Ahmar, N., Xiang, J., Poeppel, D., Simon, J.Z., 2012. Sensitivity to temporal  
1059 modulation rate and spectral bandwidth in the human auditory system: Meg evidence.  
1060 *Journal of Neurophysiology* 107, 2033-2041.
- 1061 Warren, J.D., Jennings, A.R., Griffiths, T.D., 2005. Analysis of the spectral envelope of sounds by the  
1062 human brain. *Neuroimage* 24, 1052-1057.
- 1063 Warren, R.M., 1970. Perceptual restoration of missing speech sounds. *Science* 167, 392-393.
- 1064 Wessinger, C.M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., Rauschecker, J.P., 2001. Hierarchical  
1065 organization of the human auditory cortex revealed by functional magnetic resonance  
1066 imaging. *Journal of Cognitive Neuroscience* 13, 1-7.
- 1067 Wild, C.J., Davis, M.H., Johnsrude, I.S., 2012. Human auditory cortex is sensitive to the perceived  
1068 clarity of speech. *Neuroimage* 60, 1490-1502.
- 1069 Wu, M.C.-K., David, S.V., Gallant, J.L., 2006. Complete functional characterization of sensory  
1070 neurons by system identification. *Annual Review of Neuroscience* 29, 477-505.
- 1071 Yi, H.G., Leonard, M.K., Chang, E.F., 2019. The encoding of speech sounds in the superior temporal  
1072 gyrus. *Neuron* 102, 1096-1110.
- 1073 Yoon, T.-J., Kang, Y., 2013. *The Korean Phonetic Aligner Program Suite*.

1074 Yuan, J., Liberman, M., 2008. Speaker identification on the scotus corpus. Journal of the Acoustical  
1075 Society of America 123, 3878.

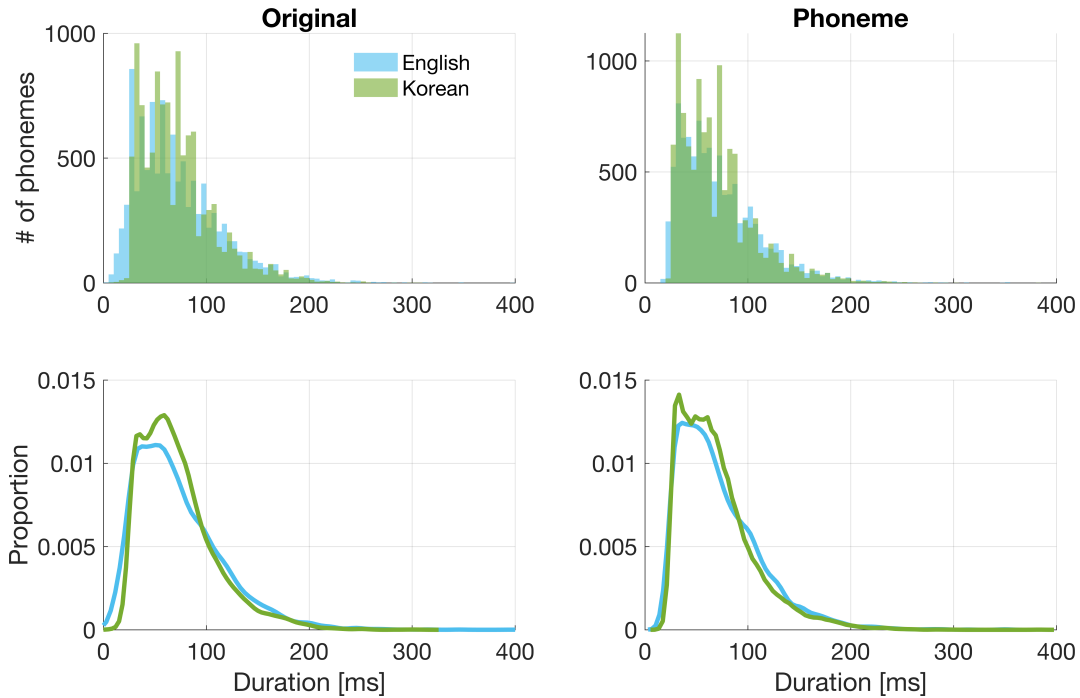
1076

1077

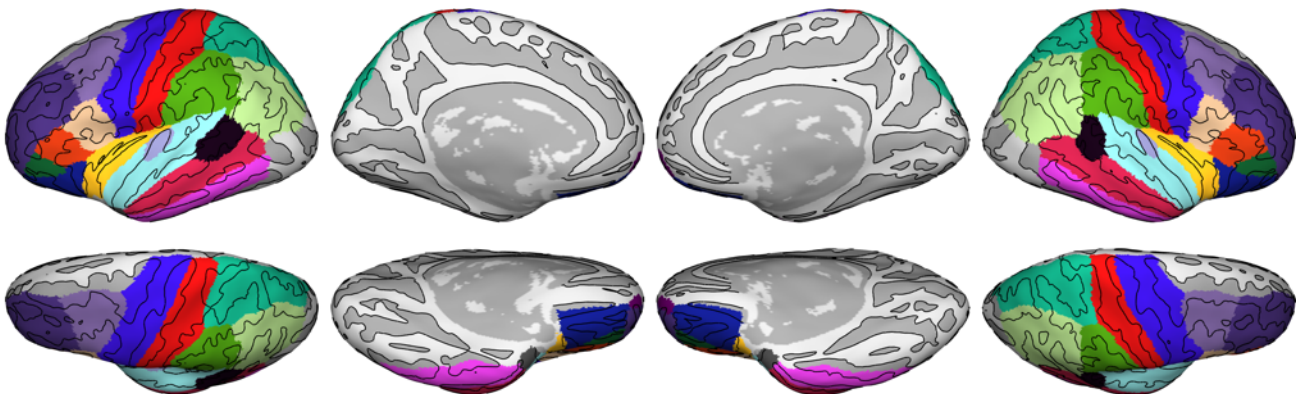
## Linguistic Modulation of Phoneme Encoding

# 1078 7 Supplementary materials

## 1079 7.1 Supplementary figures

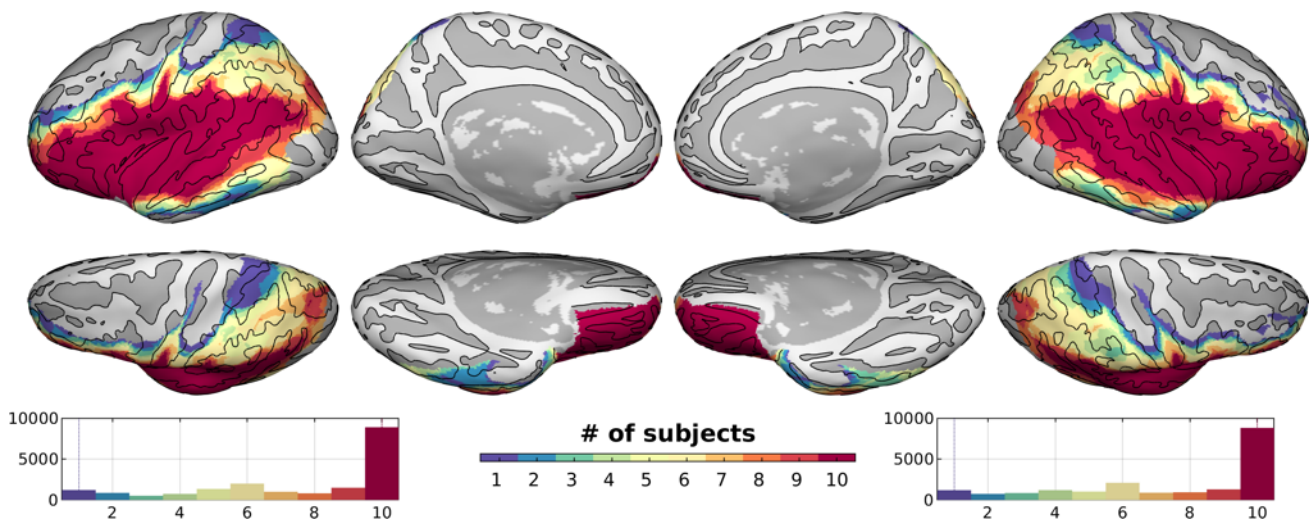


1080  
1081 **Figure S1. Histograms of phoneme durations. The distributions of phoneme durations**  
1082 **in the Original natural speech (left) and the Phoneme quilts (right) are shown for**  
1083 **English (light blue) and Korean (lime green) in histograms (top) and smoothed density**  
1084 **functions (bottom). Non-linguistic segments (e.g., short pauses) and the last segment**  
1085 **of each stimulus file (could have been cropped to equalize durations of stimuli) were**  
1086 **discarded from calculation.**

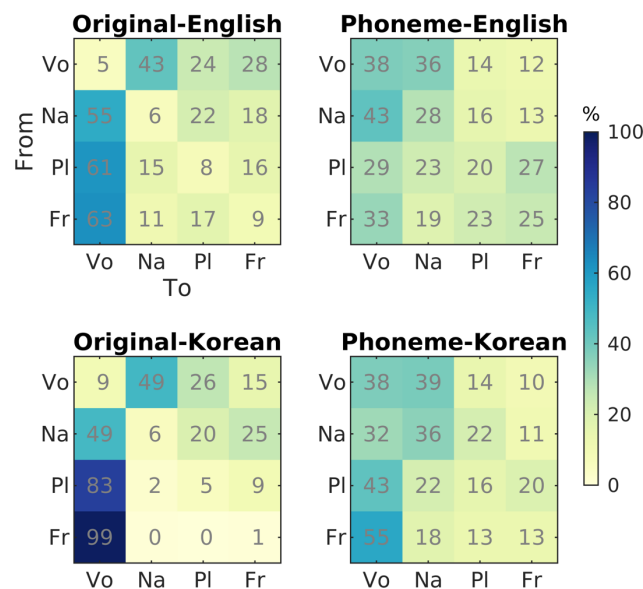


1088  
1089 **Figure S2. Regions of interest. From the automatic parcellation based on the Desikan-**  
1090 **Killiany cortical atlas (Desikan et al., 2006) in FreeSurfer, the following 19 labels were**

1091 included: 'bankssts', 'caudalmiddlefrontal', 'inferiorparietal', 'inferiortemporal',  
 1092 'lateralorbitofrontal', 'middletemporal', 'parsopercularis', 'parsorbitalis',  
 1093 'parstriangularis', 'postcentral', 'precentral', 'rostralmiddlefrontal', 'superiorparietal',  
 1094 'superiortemporal', 'supramarginal', 'frontalpole', 'temporalpole',  
 1095 'transversetemporal', 'insula'.



1097  
 1098 **Figure S3. Overlap of selected vertices across participants (N = 10). The colored**  
 1099 **histograms below display the number of vertices over the number of participants.**

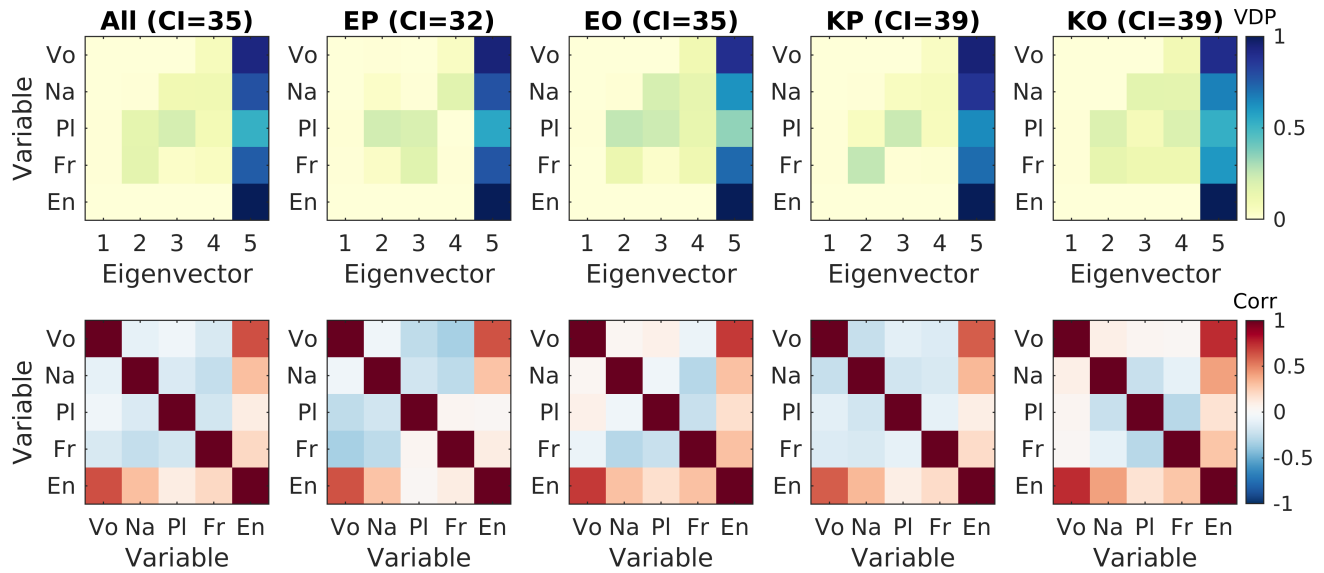


1101  
 1102 **Figure S4. Transition probability between phoneme classes. Phoneme transitions**  
 1103 **were counted as consecutive occurrences of four phoneme classes without taking**  
 1104 **word boundaries into account, and cumulated over all stimuli for visualization.**

Linguistic Modulation of Phoneme Encoding

1105 **Transition probabilities are displayed from the  $i$ -th phoneme class (Vo, vowel; Na,**  
 1106 **nasal; Pl, plosive; Fr, fricative) in rows to the  $j$ -th phoneme class in columns for the**  
 1107 **four main conditions ( $\Pr(j|i) = T_{i,j}/\sum_j T_{i,j}$  where  $T_{i,j}$  is the number of transitions from  $i$**   
 1108 **to  $j$ ).**

1109

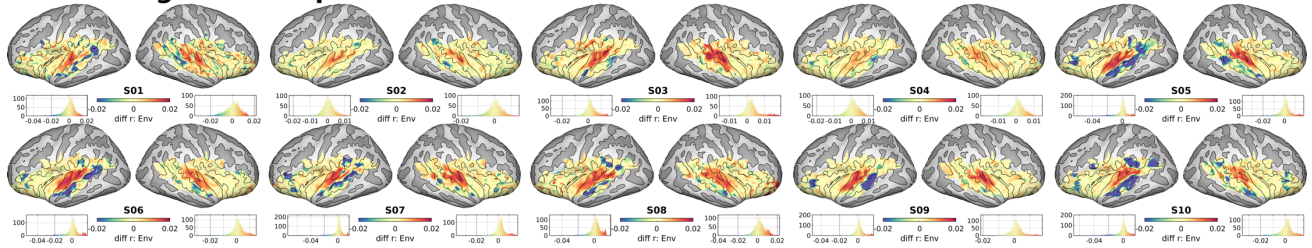


1110

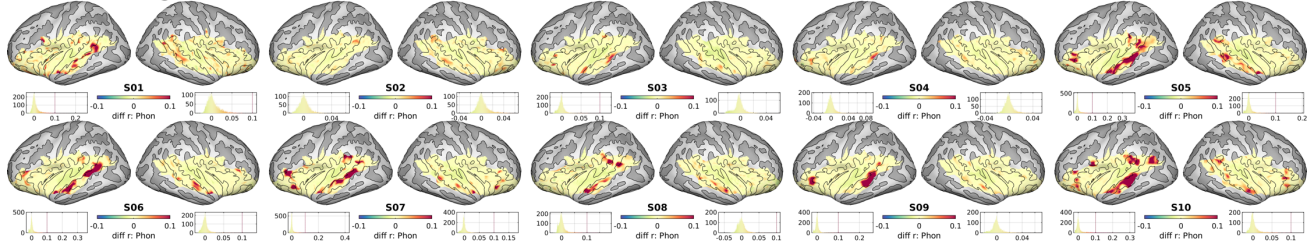
1111 **Figure S5. Collinearity of predictors. Variance decomposition proportion (VDP, upper)**  
 1112 **and pair-wise Pearson correlation (corr, lower) between predictive variables (Vo,**  
 1113 **vowel; Na, nasal; Pl, plosive; Fr, fricative; En, envelope) are shown for all conditions**  
 1114 **together (left-most column) and for each condition (EP, English-Phoneme; EO,**  
 1115 **English-Original, KP, Korean-Phoneme; KO, Korean-Original) with conditional indices**  
 1116 **(CI).**

1117

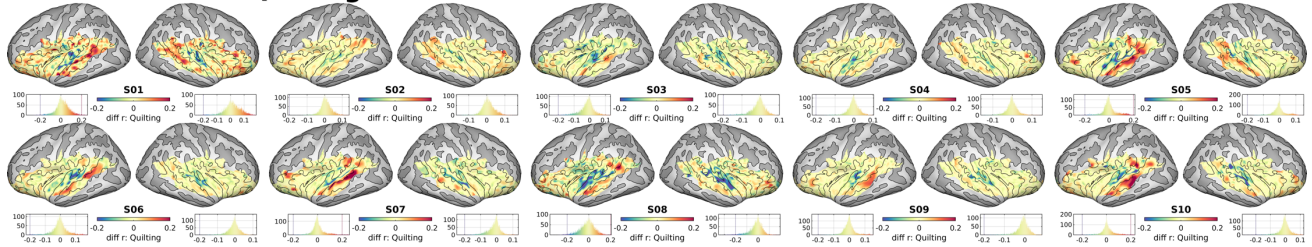
### a. Encoding of Envelope



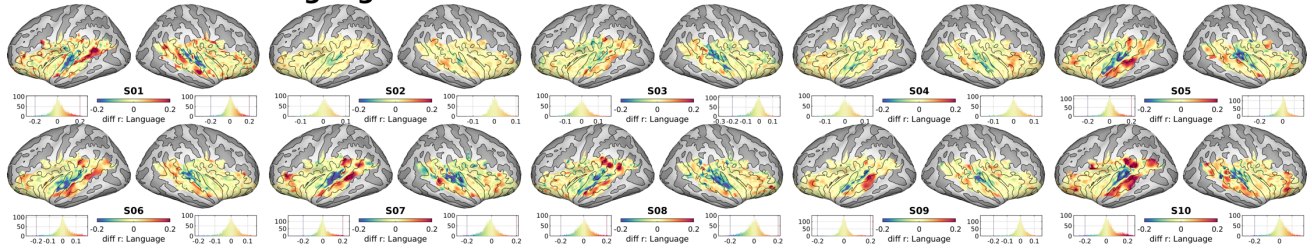
### b. Encoding of Phonemes



### c. Main effect of Quilting



### d. Main effect of Language



### f. Interaction of Quilting and Language

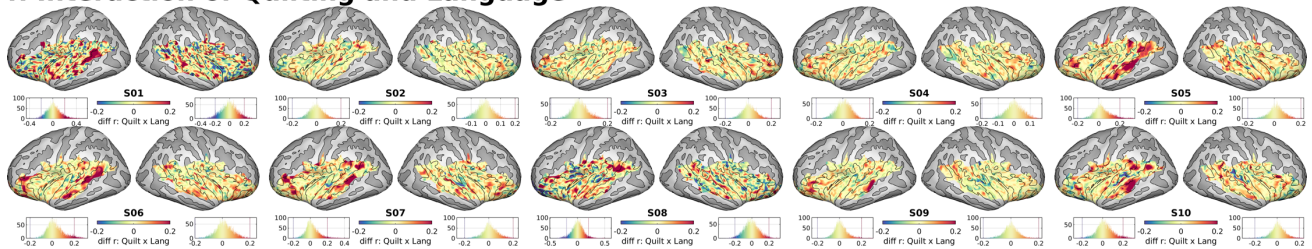


Figure S6. Encoding of features and conditions in individual participants.

Unthresholded effect size maps (differences in Pearson correlation) are shown for (a) Envelope, (b) Phonemes, (c) Quilting, (d) Language, and (e) the interaction of Quilting and Language.

1118

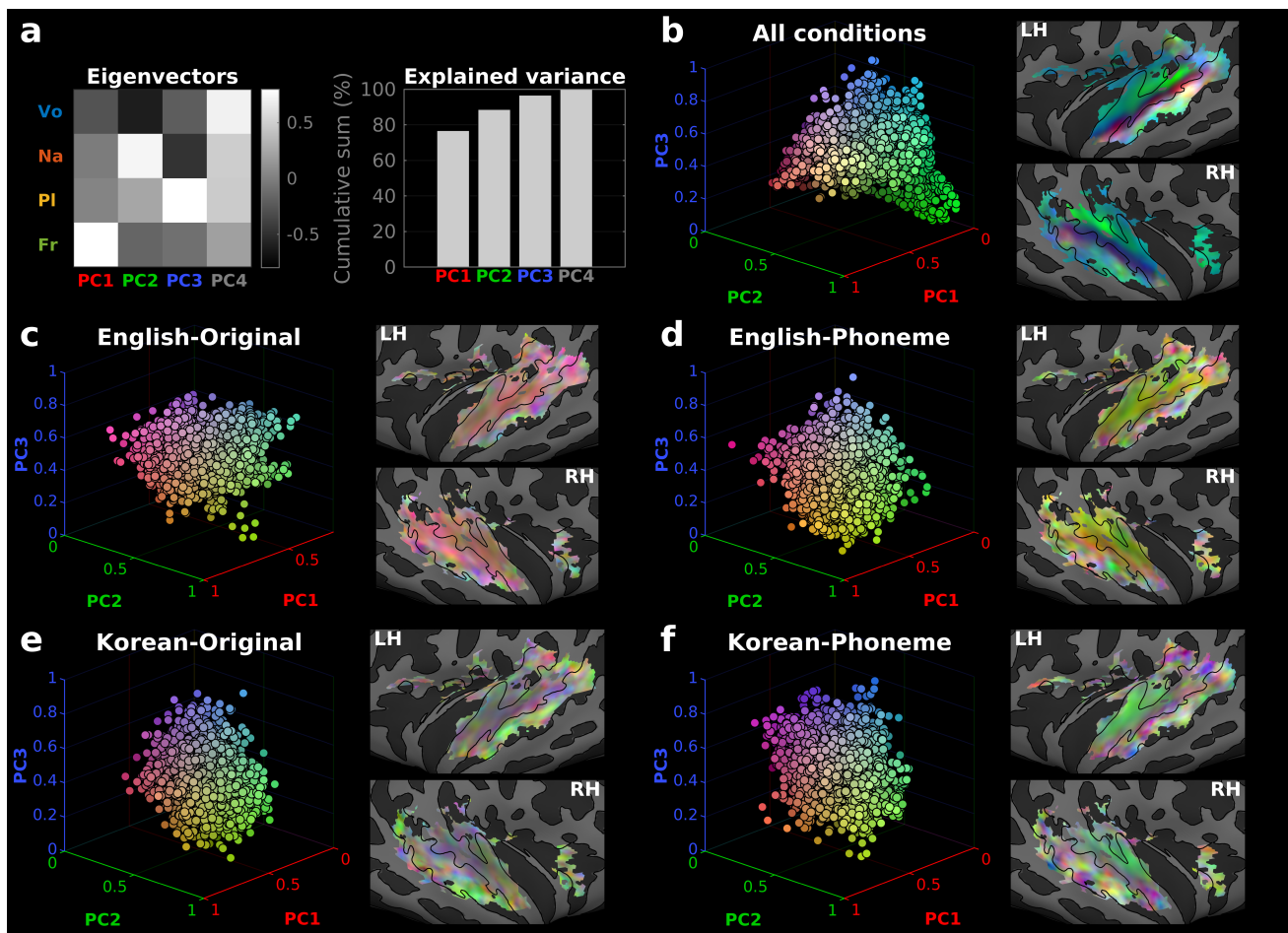
1119

1120

1121

1122

## Linguistic Modulation of Phoneme Encoding



1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133  
1134  
1135

**Figure S7. Low dimensional projection (3-D; RGB) of the phoneme-class encoding vectors (4-D) using principal component analysis (PCA). (a) Eigenvectors (left) and the explained variance (right) of the principal components. The first three components explained >98% of the variance. (b) RGB visualization of the distribution of the first three eigenvariates (i.e., factor loadings) in the 3-D eigenspace (left; each circle represents a vertex) and the anatomical space (right). Eigenvariates were scaled from 0 to 1 to enable the use of RGB values for visualization. Vertices were selected for significant encoding for any phoneme class (cluster- $P < 0.05$ ). (d-f) RGB visualization of the distribution of the first three eigenvariates (i.e., factor loadings) phoneme-class encoding vectors in each of the four conditions: (c) English-Original, (d) English-Phoneme quilt, (e) Korean-Original, (f) Korean-Phoneme quilt.**

1136 **7.2 Supplementary tables**

	<b>Vowel</b>	<b>Nasal/Approximant</b>	<b>Plosive</b>	<b>Fricative</b>
<b>English</b>	AA, AE, AH, AO, AW, AY, EH, ER, EY, IH, IY, OY, OW, UH, UW	L, M, N, NG, R, W, Y	B, D, G, K, P, T	CH, DH, F, JH, S, SH, TH, V, Z
<b>Korean</b>	A, AE, E, EO, EU, I, O, OE, U, WA, WAE, WE, WEO, WI, YA, YAE, YE, YEO, YI, YO, YU	L, M, N, NG, R	B, BB, D, DD, G, GG, K, P, T	C, H, J, JJ, S, SS

1137 **Table S1. Individual phonemes included in analysis for each articulatory phoneme**  
 1138 **class.**

1139