Linguistic modulation of the neural encoding of phonemes

Seung-Goo Kim[1], Federico De Martino[2], Tobias Overath[1,3,4]
[1]Department of Psychology and Neuroscience, Duke University, Durham, N.C., U.S.A.
[2]Faculty of Psychology and Neuroscience, University of Maastricht, The Netherlands.
[3]Duke Institute for Brain Sciences, Duke University, Durham, N.C., U.S.A.
[4]Center for Cognitive Neuroscience, Duke University, Durham, N.C., U.S.A.

Corresponding authors:
solleo@gmail.com (S.-G. K.), t.overath@duke.edu (T. O.)

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

Abstract

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

1 Introduction

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

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

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

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

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

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

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

2 Results

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

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

2.1 Encoding of phonemes and conditions

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

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

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

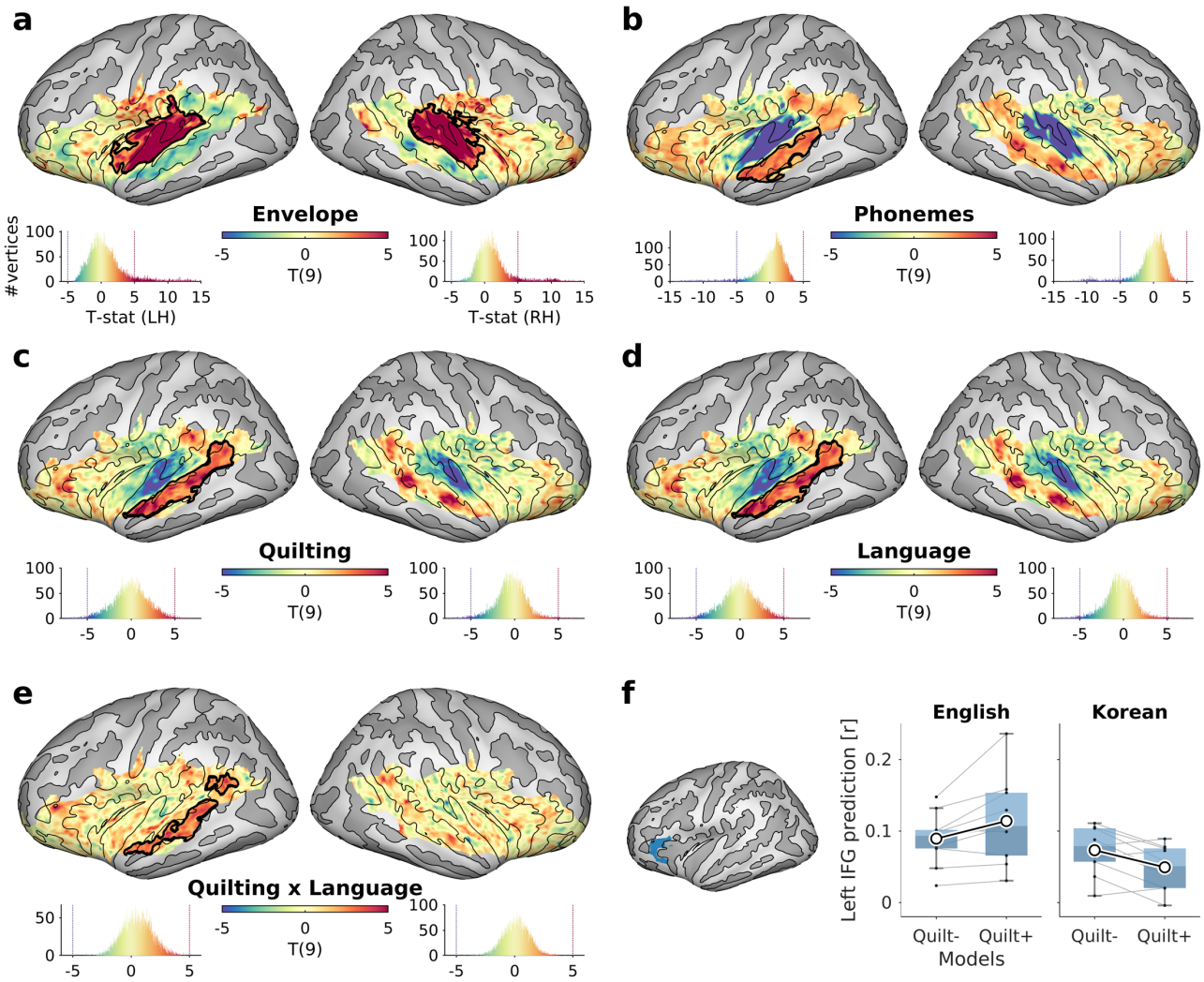

Figure 1. Encoding of features and conditions. *t*-statistic maps comparing prediction accuracies from models with vs. without (a) Envelope, (b) Phonemes, (c) Quilting, (d) Language, (e) and differences in Quilting encoding between languages (i.e., interaction). *t*-maps are truncated at $|t| < 5$ for 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).

2.2 Encoding of individual phoneme classes

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

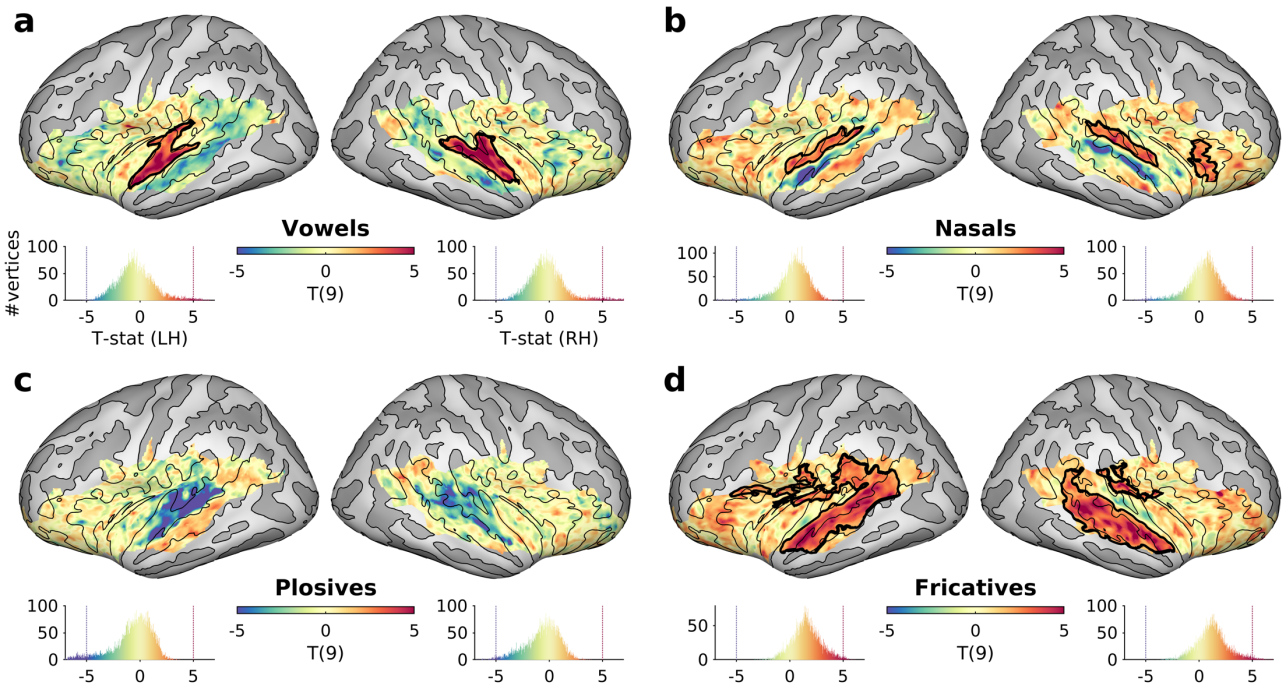

Figure 2. Unique encoding of individual phoneme classes. *t*-statistic maps comparing the prediction accuracy of a full model with that of a reduced model without a specific phoneme feature: (a) vowels, (b) nasals, (c) plosives, or (d) fricatives. Thick contours in black mark areas for which cluster-*P*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

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

multivariate analysis). Both Pearson and Euclidian distances between phoneme class vectors in quilted vs. original speech conditions were significantly greater in English than in Korean in the right STG (Pearson distance, cluster-*P* = 0.0469, **Figure 3b**; Euclidian distance, cluster-*P* = 0.0029, **Figure 3c**).

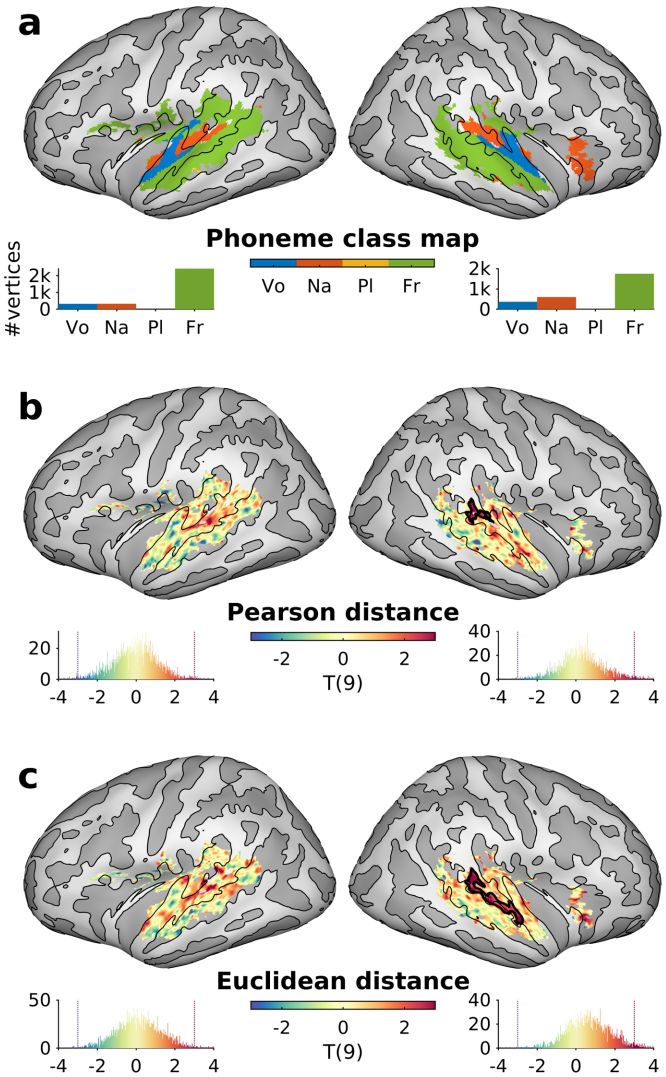

Figure 3. Encoding patterns of phoneme classes. (a) Winner-take-all map for the four phoneme classes. Vertices were selected for significant encoding for any phoneme class (cluster-*P* < 0.05). The histograms denote the number of 'winner' vertices per phoneme class in the left and right hemisphere. (b-c) 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.

3 Discussion

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

3.1 Distinct encoding of individual phoneme classes

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

Figure 3a revealed a clear topography for different phoneme classes, which was remarkably homologous between the left and right hemispheres, down to the cluster for nasals in antero-medial HS. Vowels showed the strongest sensitivity in early auditory cortex in HG and parts of STG. Since one of the main features of vowels is their harmonicity, which is a defining characteristic of pitch (Plack et al., 2005), this matches well with studies that have shown pitch-sensitive responses in these early cortical areas (Bendor and Wang, 2005; Griffiths and Hall, 2012; Hall and Plack, 2009; Kim et al., in revision). Of the four phoneme classes, fricatives ‘won’ the majority of vertices, which were mostly located in non-primary cortex (STG and STS). While the large extent was a somewhat unexpected finding, its location in non-primary cortex agrees with a previously proposed functional hierarchy in auditory cortex, whereby spectral filter width increases (e.g., from sinusoids via narrowband to broadband noise) from primary to non-primary auditory cortex (Carrasco and Lomber (2009); Chevillet et al. (2011); Pandya et al. (2007); Rauschecker and Tian (2004); Rauschecker et al. (1995); Wessinger et al. (2001) though see Overath et al. (2012); Wang et al. (2012)). Since high-frequency broadband noise is a defining characteristic of fricatives, the strong response in STG and STS could reflect such spectral sensitivity. Nasals revealed highest prediction accuracies in the non-primary auditory cortex (e.g., lateral HG, PP, and PT). These areas roughly correspond to regions that display sensitivity to slow temporal modulations (Baumann et al., 2015; Santoro et al., 2014). In particular, Schönwiesner and Zatorre (2009) revealed a preference for low spectral density (with slow temporal modulation rates) in these regions. Nasals are characterized by a 'murmur', related to the nasal resonances if the oral tract is closed, which effectively amounts to a low-pass filter (Qi and Fox, 1992). An animal model (ferrets) showed that nasals tend to excite primary auditory

cells that are tuned for slow temporal modulations (Mesgarani et al., 2008). Thus, it is conceivable that the spectrotemporal modulation preference in the non-primary auditory cortex is reflected in the preferred encoding of the narrow-band filtered fluctuation of nasals. Finally, plosives, which are characterized by a complete occlusion followed a subsequent broadband burst, were the only phoneme class that did not reveal areas with maximal prediction accuracy compared to the other phoneme classes. As hinted in **Figure 2c**, there was a slight (subthreshold) increase in prediction accuracies in the left STS for plosives, which is similar to that of fricatives but at weaker magnitudes. This is in line with a similarity between responses to plosives and to fricatives (Khalighinejad et al., 2017), which suggests similar neural generators. Plosives and fricatives are both characterized by noise bursts; the brevity of the burst in plosives compared to the more sustained noise burst in fricatives might have resulted in an overall lower encoding prediction for plosives in similar regions as those encoding fricatives, which in turn might explain the absence of plosive phoneme class 'winner' regions in the current study.

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

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

3.2 Linguistic modulation of acoustic phoneme-class analysis

One of our aims was to dissociate acoustic from linguistic processes, which would enable us to determine their interaction, i.e., whether linguistic processes modulate the acoustic analysis of phonemes. To this end, we found that the acoustic manipulation (phoneme quilts vs. natural speech) had a larger effect on phoneme processing in a familiar language (English) than in a foreign language (Korean): phoneme class encoding was more dissimilar between phoneme-quilt and natural speech conditions in English than in Korean. Since the acoustic manipulation was the same for each language, this suggests that the greater difference between acoustic contexts was due to linguistic processes becoming engaged in a familiar language. Linguistic processes such as phonotactic, syntactic, as well as semantic analyses might therefore modulate the acoustic processing of phonemes, e.g., via hierarchical predictive coding or minimizing top-down prediction errors (Friston and Kiebel, 2009; Rao and Ballard, 1999). To our

knowledge, this is the first demonstration of such linguistic modulation of a fundamental linguistic unit using fMRI. However, these results align with Lee and Overath (in revision), who found similar evidence for top-down linguistic modulation of phonemic analysis using a different recording modality (EEG).

Perhaps the best-known example of the modulatory influence of linguistic information is that of phonemic restoration (Samuel, 1981; Warren, 1970). In phonemic restoration, a phoneme is still 'perceived' even if it is masked or replaced completely by noise. This is often interpreted as an advantageous adaptation to speech perception in noisy environments, where it is common for interrupting or masking sounds to last only for a few tens or hundreds of milliseconds (i.e., on a temporal scale that is commensurate with that of phonemes). The top-down predictive nature of this phenomenon is further highlighted by the fact that, if the acoustic information is ambiguous, a 'best guess' phoneme is perceived (Leonard et al., 2016; Samuel, 1987). In fact, there is a wealth of evidence for such restorative processes in speech perception, for example from studies using noise-vocoded stimuli (Giraud et al., 2004; Narain et al., 2003; Obleser et al., 2008; Scott et al., 2000; Shannon et al., 1995; Wild et al., 2012) or other methods to distort the speech signal (Davis et al., 2011; Eckert et al., 2016), while the most common explanation for restorative effects refers to top-down, predictive (Friston and Kiebel, 2009) linguistic processes.

The locus of phonemic restoration, i.e. the region in which linguistic modulation is strongest, was recently shown to be situated in bilateral STG, likely due to receiving modulatory signals from left IFG (Leonard et al., 2016). This aligns remarkably well with the current study, where we found the strongest effect of linguistic modulation along right STG. Note that this region along STG touches upon areas of all three phoneme-class 'winners' (vowels, nasals, fricatives; cf. **Figure 3a**) and is therefore ideally situated to modulate the neural analysis of these phoneme classes. The apparent right-lateralization may in part be a consequence of the cluster-forming threshold, which penalizes or disregards smaller activation clusters, since similar but smaller peaks along left STG are also visible for both Pearson and Euclidean distance metrics (Error! Reference source not found.**e-f**).

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

In the domain of electrophysiological measurements of speech perception, there is currently disagreement as to the extent that neural indices (such as speech-envelope entrainment, or phoneme encoding) can be interpreted as markers of linguistic processes that are necessary for speech comprehension (Di Liberto et al., 2015; Ding and Simon, 2013; Luo and Poeppel, 2007; Vanthornhout et al., 2018), or whether a more parsimonious explanation of these indices is that they reflect the analysis of characteristic acoustic properties of the speech signal (Baltzell et al., 2017; Daube et al., 2019; Howard and Poeppel, 2010; Millman et al., 2015; Verschueren et al., 2021). Our study is able to shed new light on

this controversy by directly comparing the encoding of acoustic properties of phonemes in either a familiar language or in a foreign language, in which no higher-level linguistic analysis takes place. Based on the current results, we suggest that both accounts have merit: an (initial) analysis is likely driven by the acoustic properties of phonemes, while a (subsequent) linguistic analysis modulates this acoustic analysis.

We should note that the current study did not address or measure linguistic processes explicitly. For example, participants did not perform a linguistic task (e.g., speech comprehension), but were simply asked to detect a change in speaker, a task that is largely orthogonal to linguistic processing (see also Overath and Paik (2021) for a similar task). Therefore, we interpret the linguistic modulation of phoneme class analysis as obligatory linguistic processes that become engaged as soon as familiar linguistic templates (e.g., phonotactics, syntax, lexicon, semantics) are detected in the signal. Future studies will need to determine whether, and to what extent, these obligatory linguistic processes for phoneme analysis are malleable to various tasks that engage specific linguistic processes. For example, the neural processing of acoustic features in speech sounds has been shown to be enhanced or sharpened if they are task-relevant, attended to vs. ignored, or primed (Holdgraf et al., 2016; Leonard et al., 2016; Mesgarani and Chang, 2012; Rutten et al., 2019), and similar processes might become engaged for phoneme class encoding.

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

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

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

3.4 Modulation of acoustic and linguistic contexts

The analyses of the two factors Quilting and Language were motivated by previous studies that investigated the processing of temporal speech structure using segment-based speech quilting. In particular, these studies showed sensitivity in STS to temporal speech structure in either only a foreign language (Overath et al., 2015), or both familiar and foreign languages (Overath and Paik, 2021), which is comparable to a main effect of Quilting here. In addition, activity in left IFG revealed an interaction between Quilting and Language and increased as a function of temporal speech structure only in the familiar language (Overath and Paik, 2021). In the current study, Quilting and Language both had greater prediction accuracies in left STS, while their interaction in the same area (as well as left IFG) was due to larger prediction accuracy differences between the two Quilting conditions in English than in Korean.

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

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

3.5 Future directions

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

In addition, the fact that the linguistic modulation of the acoustic speech signal operates at such an early stage of linguistic analysis likely reflects its significance: if linguistic modulation starts at the level of phonemes, its ability to impact a later word processing stage is conceivably greater than if linguistic modulation only started at the word processing stage. Given the highly predictive nature of speech processing (see **Section 3.2** above), such modulation might be particularly helpful in situations in which

the speech signal is compromised (e.g., in noisy conditions such as in a restaurant or bar). People with hearing loss (e.g., presbycusis) are a clinical population that is known to struggle in such situations, even with the help of hearing aids (Moore, 1996; Shinn-Cunningham and Best, 2008). It is therefore possible that (at least) one reason for their exacerbated speech comprehension difficulties in noisy situations is that the linguistic modulation of phonemes has deteriorated, thereby reducing the effectiveness of predictive speech processes. A similar argument might be made for people suffering from 'hidden hearing loss': i.e., hearing difficulties without detectable deficits in routine audiometry tests (Kujawa and Liberman, 2009; Ruggles et al., 2011). We predict that linguistic modulation of phoneme analysis is reduced in these populations (particularly in situations with background noise) and might thus serve as a clinical marker.

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

3.6 Conclusions

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

4 Methods

4.1 Participants

Ten native English speakers without any knowledge or experience in Korean participated in the current study (mean age = 24.0 ± 2.2 years; 6 females). Eight participants volunteered in three sessions consisting of 8 runs each on separate days (intervals in days: mean = 8.5, standard deviation = 16.6, min = 1, max

= 70) and two other participants in a single session each (6 runs and 8 runs, respectively), resulting in a total of 24 scanning sessions. This is on par with similar approaches that maximize intra-subject reliability over intra-subject variability in the data (Breedlove et al., 2020; Huth et al., 2016; Kay et al., 2008; Moerel et al., 2013; Naselaris et al., 2015; Norman-Haignere et al., 2015; Santoro et al., 2017).

All participants were recruited via the Brain Imaging and Analysis Center (BIAC) at Duke University Medical Center, NC, USA after safety screening for MRI (e.g., free of metal implants and claustrophobia). All reported to have normal hearing and no history or presence of neurological or psychiatric disorders. Informed written consent was obtained from all participants prior to the study in compliance with the protocols approved by the Duke University Health System Institutional Review Board.

4.2 Stimuli

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

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

The phoneme segments were pseudorandomly rearranged to create novel phoneme quilts. For each stimulus, a random initial phoneme was chosen; subsequent phonemes were selected such that the acoustic change at the boundary was as close as possible to the acoustic change in the original source signal (using the L2-norm metric of an ERB-spaced cochleogram; see (Overath et al., 2015)). In addition, we applied the following exclusion criteria: (a) the phoneme duration needed to be at least 20 ms, (b) two

[1] https://babel.ling.upenn.edu/phonetics/old_website_2015/p2fa/index.html
[2] https://korean.utsc.utoronto.ca/kpa/
[3] https://github.com/bbTomas/mPraat

identical phonemes could not occur next to each other, (c) for a given phoneme, its subsequent phoneme could not be the same as in the original source signal. We used the pitch-synchronous overlap-add (PSOLA) algorithm (Moulines and Charpentier, 1990) to further minimize abrupt changes in pitch at phoneme boundaries. Overall alterations due to the quilting algorithm were quantified by the Kullback-Leibler divergence (D_{KL}) between L2-norm acoustic change distributions in the original source and the created phoneme quilt (median D_{KL} = 0.6873 bits for English, 0.6004 bits for Korean; Wilcoxon rank sum equal median test: Z = 0.5913, P = 0.5543). In the phoneme quilts, the durations of phonemes in milliseconds were as follows (see **Supplementary Figure S1** for histograms): min = 20.0, max = 351.0, mean = 72.3, median = 63.0, standard deviation = 39.4, skewness = 1.4 in English (N = 10,467); min = 20.0, max = 383.0, mean = 69.7, median = 60.0, standard deviation = 36.3, skewness = 1.5 in Korean (N = 11,213). There were slight differences between languages in means (2.6 ms longer in English, t[21678] = 5.08, $P = 10^{-7}$) and distributions (KS-stat = 0.0657, $P = 10^{-21}$), however, the mean difference of 2.6 ms is much shorter than the modeled cochlear integration time-window of 20 ms.

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

4.3 Experimental procedure

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

The stimulus presentation timing was controlled via the Psychophysics Toolbox (v3.0.11[4]). Each run was triggered by the TTL signal from the MRI scanner mediated by a counter. Digital auditory signals at 44,100 Hz sampling rate and 16-bit precision from a Windows desktop were converted to analog signals by an external digital amplifier (Sony, Tokyo, Japan) and delivered to participants via MRI-compatible

[4] http://psychtoolbox.org/

insert earphones (S14, Sensimetrics, MA, USA) at a comfortable listening level (~75 dB SPL). Participants wore protective earmuffs on top of the earphones to further reduce acoustic noise from the MRI scanner.

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

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

4.4 Image acquisition

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

For T1-weighted contrast, a magnetization prepared rapid gradient echo (MP-RAGE) scan covering the whole-brain (in-plane pixel size = 1 × 1 mm^2, slice thickness = 1 mm, FOV = 256 mm, matrix size = 256 × 256, TE = 3.2 ms, flip angle = 8°, TR = 2264 ms, number of slices = 156) was acquired at the end of each session.

4.5 Image processing

4.5.1 Anatomical images

T1-weighted images were segmented using SPM (SPM12; v7487[5]) to obtain tissue probability maps (`spm.spatial.preproc`), which were used for anatomical CompCor regressors (Behzadi et al., 2007). High-resolution cortical surfaces were fully automatically constructed using FreeSurfer (v6.0.0[6]) for surface-based analysis.

4.5.2 Functional images

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

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

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

4.5.3 Surface-based GLM denoising

We applied a model-based denoising technique for task-based fMRI data ("GLMdenoise" v1.4[8]) to the surface-mapped data (Kay et al., 2013). The algorithm extracts 'noise' regressors from the data that would increase prediction accuracy in leave-one-run-out-cross-validation. This is achieved by first defining 'noise pool' voxels with negative R^2 values for a given design matrix (i.e., voxels that are irrelevant to the task of interest), extracting principal components from the noise pool, and then determining an optimal

[5] https://www.fil.ion.ucl.ac.uk/spm/
[6] http://freesurfer.net/
[7] https://fsl.fmrib.ox.ac.uk/
[8] https://kendrickkay.net/GLMdenoise/

number of components to remove as a minimal number where the improvement in cross-validation prediction decays. We used box-car functions to represent the four conditions in the design matrix. On average, 4.5 ± 2.1 noise regressors were regressed out. These improved reliability in estimation (mean over standard errors ratio of coefficients estimates across CV folds: median increase = 0.82; mean increase = 1.12) but only slightly increased predication accuracy (cross-validation R^2: median increase = 0.25% points; mean increase = 0.56% points). In addition to the noise regressors, the 4-th order polynomial fits to slow drifts in BOLD time series, the six CompCor regressors, and the button-press regressors convoluted with a canonical HRF were regressed out from the residuals (i.e., prediction from the design matrix subtracted from the data).

4.6 Voxel-wise linearized encoding analysis

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

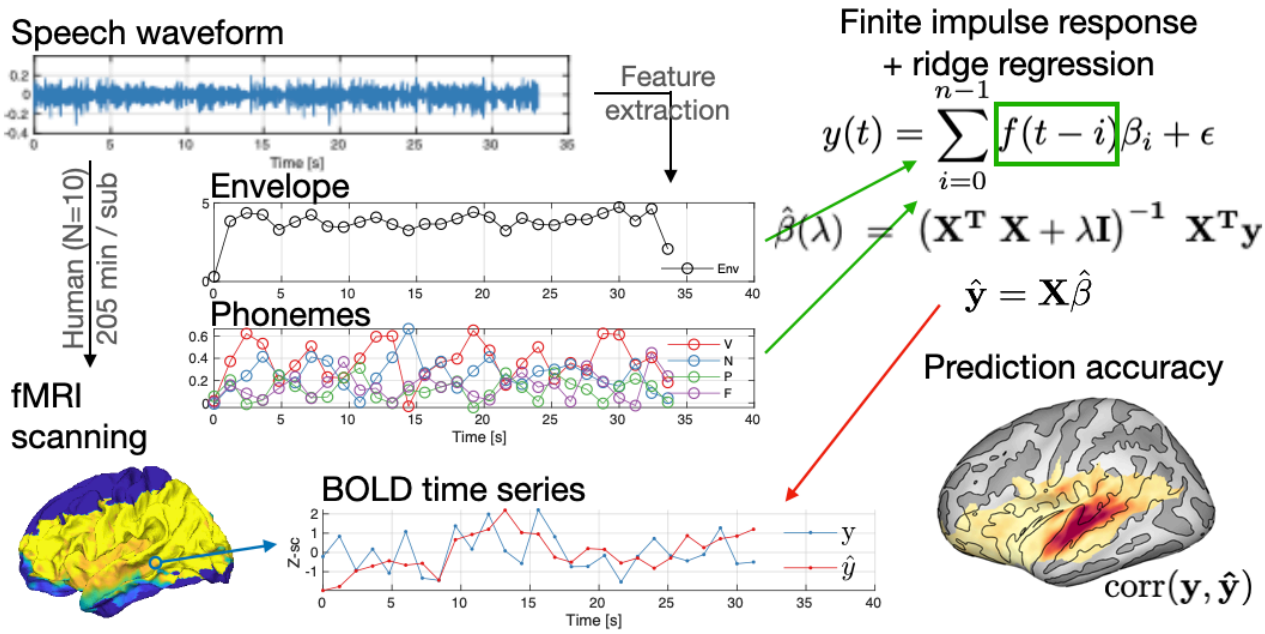

Figure 4. Linearized encoding analysis overview. Functional MRI data was acquired from 10 human participants while listening to unmanipulated, or phoneme-scrambled speech stimuli in either English or Korean. From the speech waveform, cochleogram envelope and the duration of phoneme classes were extracted and down-sampled at the fMRI sampling rate (1/1.2 Hz). After preprocessing, surface-mapped BOLD time series was predicted using a linear method (finite impulse response modeling with

ridge regression). The prediction accuracy was measured by Pearson correlation between actual and predicted BOLD time series.

4.6.1 Vertex selection

For our interest in auditory and linguistic processing, we restricted our analysis to vertices in cortical regions that are previously known to be involved in speech processing so as to avoid unnecessary computations. Specifically, from the automatic parcellation based on the Desikan-Killiany cortical atlas (Desikan et al., 2006), the following 19 labels were included: 'bankssts', 'caudalmiddlefrontal', 'inferiorparietal', 'inferiortemporal', 'lateralorbitofrontal', 'middletemporal', 'parsopercularis', 'parsorbitalis', 'parstriangularis', 'postcentral', 'precentral', 'rostralmiddlefrontal', 'superiorparietal', 'superiortemporal', 'supramarginal', 'frontalpole', 'temporalpole', 'transversetemporal', 'insula'. Selected regions are visualized in **Supplementary Figure S2**. Vertices with BOLD time series varied across participants due to the variability of head sizes, individual acquisition volumes at each session, and movements across runs during sessions. **Supplementary Figure S3** shows the overlap of selected vertices across participants. On average, 28,297 ± 3,748 vertices were selected per participant.

4.6.2 Predictors

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

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

4.6.3 Finite-impulse response modelling

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

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

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 a certain voxel), $\boldsymbol{\beta}$ is a $(t \times p)$ unknown coefficient vector, and $\boldsymbol{\varepsilon}$ is a noise vector from a zero-mean Gaussian distribution with a serial correlation $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Omega})$ where $\boldsymbol{\Omega}$ is a $(t \times t)$ unknown covariance matrix and σ^2 is a scale factor. For the FIR modeling, the design matrix **X** consists of matrices of delayed features as:

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

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

$$f_i(t1, t2) * \mathbf{H}(n) = \begin{bmatrix} f_i(t_1) & f_i(t_1 - 1) & \dots & f_i(t_1 - (n-1)) \\ f_i(t_1 + 1) & f_i(t_1) & \dots & f_i(t_1 - n) \\ \vdots & \vdots & \dots & \vdots \\ f_i(t_2) & f_i(t_2 - 1) & \dots & f_i(t_2 - (n-1)) \end{bmatrix},$$

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

4.6.4 Model estimation

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

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

$$\widehat{\boldsymbol{\beta}}(\lambda) = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}, \tag{2}$$

where $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ is a vector of penalized estimates, $\mathbf{I}$ is an identity matrix, and λ is a ridge penalty term. Note 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 λ, we used a method called "ridge trace" (Hoerl and Kennard, 1970),
653 which finds the smallest λ that returns "stabilized" normalized coefficients so that the bias introduced by
654 λ 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 $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]^T$, we
656 determined an optimal lambda λ^* from a set of non-zero, incremental lambdas $\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ as the
657 smallest λ such that an increment of λ results in changes in all coefficient $\hat{\beta}_{\mathrm{i}}(\lambda)$ smaller than 20% of their
658 initial values $\hat{\beta}_{\mathrm{i}}(\lambda_1)$ for all *p* coefficients:

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

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:

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

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:

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

680 where **C** is a $(t \times 4)$ matrix of dummy predictors of four conditions as $\mathbf{C} = [\boldsymbol{c}_{\mathrm{EP}} \quad \boldsymbol{c}_{\mathrm{EO}} \quad \boldsymbol{c}_{\mathrm{KP}} \quad \boldsymbol{c}_{\mathrm{KO}}]$, $\otimes$ is the
681 Kronecker product operator, **F** is a $(t \times 5)$ matrix of predictors of five features as $\mathbf{F}_{\boldsymbol{PE}} =$

682 $[\boldsymbol{f}_{\mathrm{Vo}} \quad \boldsymbol{f}_{\mathrm{Na}} \quad \boldsymbol{f}_{\mathrm{Pl}} \quad \boldsymbol{f}_{Fr} \quad \boldsymbol{f}_{En}]$, and $\mathbf{H}(\mathrm{n})$ 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; Pl, 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).

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 **F** in Eq. (5), which can then be written as:

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

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

$$\begin{cases} H_0: & \mathbb{E}(r_{Phon+Env,C+}) \leq \mathbb{E}(r_{Env,C}) \\ H_{\mathrm{A}}: & \mathbb{E}(r_{Phon+Env,C+}) > \mathbb{E}(r_{Env,C}) \end{cases},$$

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 **F** in Eq. (5):

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

700 where $\mathbf{F}_{Phon} = [f_{\mathrm{Vo}} \quad f_{\mathrm{Na}} \quad f_{Pl} \quad f_{\mathrm{Fr}}]$. 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 **C** in Eq. (5) as:

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

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

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

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

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

4.6.7 Phoneme-class-specific effects

[9] http://www.fieldtriptoolbox.org/

We further investigated the specific contribution of each phoneme class by comparing a full model with all phoneme classes (Eq. 5) with reduced models without a particular phoneme class. For example, a model without nasals would be $\mathbf{X}_{Phon_{Na-}+Env,C+} = \left[\mathbf{C} \otimes \mathbf{F}_{Phon_{Na-}+Env}\right] * \mathbf{H}(\mathbf{n})$, where $\mathbf{F}_{Phon_{Na-}+Env} = [\boldsymbol{f}_{Vo} \quad \boldsymbol{f}_{Pl} \quad \boldsymbol{f}_{Fr} \quad \boldsymbol{f}_{En}]$. From such comparisons, we constructed a (1×4) vector ("phoneme encoding vector") of the changes of prediction accuracies for the four phoneme classes $\boldsymbol{d} = [\Delta r_{Vo} \quad \Delta r_{Pl} \quad \Delta r_{Na} \quad \Delta r_{Fr}]$ at each vertex. From these vertex-wise vectors, a winner-take-all phoneme class map was created (at each vector with a significant prediction accuracy increase [cluster-$P < 0.05$] for at least one phoneme class).

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

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

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

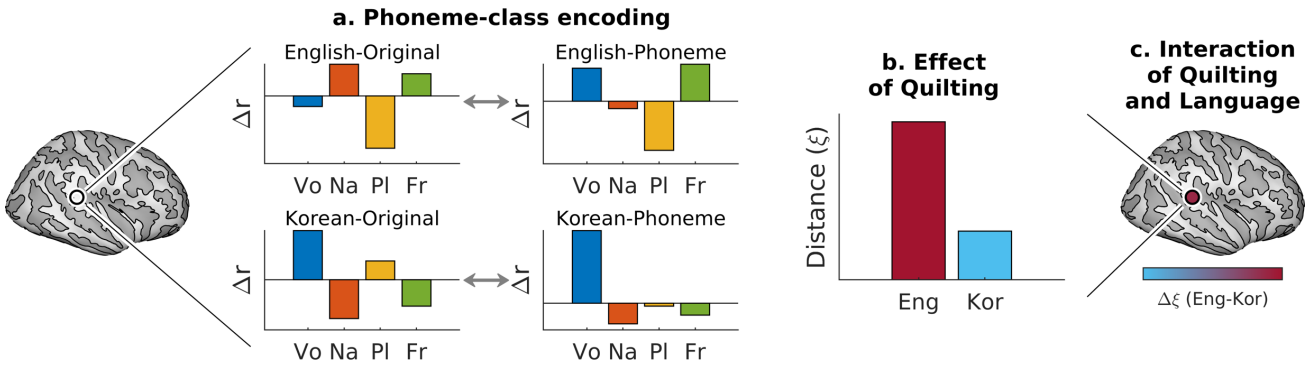

Figure 5. Schematic of the multivariate analysis on phoneme-class encoding vectors. (a) For each vertex, its [1x4] phoneme-class encoding vector was defined by the prediction accuracy changes when adding a particular phoneme class to the model for each of four conditions (two languages x two quilting conditions). (b) Distance metrics (Pearson or Euclidian) were computed within each language between the Original and Phoneme quilt conditions. (c) The difference in this distance metric between languages (English-minus-Korean; i.e., the interaction of Quilting and Language) was mapped back to the corresponding vertex (see Figure 3b-c).

For visualization, the 4-D phoneme encoding vectors were projected to a 3-D eigenspace using principal component analysis (PCA). The first three principal component scores (i.e., eigenvariates) were re-scaled between zero and one and mapped to RGB values, respectively. The vertex-mapped RGB values were interpolated on the triangular faces of the cortical meshes using MALTAB's `patch` function. Note that the visualization is intended as an intuitive representation only; the actual comparisons were made on the scalar distance values comparing 4-D vectors.

4.7 Data availability

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

4.8 Code availability

The computer code that was used for this study is available on the Open Science Framework repository[10].

5 Acknowledgements

The authors would like to thank Frankie Pennington and Joon Hyun Paik for manually checking phoneme onsets and offsets for the forced alignment of English and Korean recordings, respectively. This work was supported by US National Institutes of Health grant R21DC016386 to T.O.

6 References

Aitken, A.C., 1936. On least squares and linear combination of observations. Proceedings of the Royal Society of Edinburgh. Section B: Biology 55, 42-48.

Anderson, J.L., Morgan, J.L., White, K.S., 2003. A statistical basis for speech sound discrimination. Language and Speech 46, 155-182.

Baltzell, L.S., Srinivasan, R., Richards, V.M., 2017. The effect of prior knowledge and intelligibility on the cortical entrainment response to speech. Journal of Neurophysiology 118, 3144-3151.

Baumann, S., Joly, O., Rees, A., Petkov, C.I., Sun, L., Thiele, A., Griffiths, T.D., 2015. The topography of frequency and time representation in primate auditory cortices. Elife 4, e03256.

Behzadi, Y., Restom, K., Liau, J., Liu, T.T., 2007. A component based noise correction method (compcor) for bold and perfusion based fmri. Neuroimage 37, 90-101.

[10] https://osf.io/zgj3m/?view_only=cd4942f9ea674d79a5644796d5498e3c

Belsley, D.A., 1991. A guide to using the collinearity diagnostics. Computer Science in Economics and Management 4, 33-50.

Bendor, D., Wang, X., 2005. The neuronal representation of pitch in primate auditory cortex. Nature 436, 1161-1165.

Blank, I.A., Fedorenko, E., 2020. No evidence for differences among language regions in their temporal receptive windows. Neuroimage 219, 116925.

Bořil, T., Skarnitzl, R., 2016. Tools rPraat and mPraat. Text, Speech, and Dialogue. TSD 2016. Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 367-374.

Breedlove, J.L., St-Yves, G., Olman, C.A., Naselaris, T., 2020. Generative feedback explains distinct brain activity codes for seen and mental images. Current Biology 30, 2211-2224.e2216.

Carrasco, A., Lomber, S.G., 2009. Evidence for hierarchical processing in cat auditory cortex: Nonreciprocal influence of primary auditory cortex on the posterior auditory field. Journal of Neuroscience 29, 14323-14333.

Cheour, M., Ceponiene, R., Lehtokoski, A., Luuk, A., Allik, J., Alho, K., Näätänen, R., 1998. Development of language-specific phoneme representations in the infant brain. Nature Neuroscience 1, 351-353.

Chevillet, M., Riesenhuber, M., Rauschecker, J.P., 2011. Functional correlates of the anterolateral processing hierarchy in human auditory cortex. Journal of Neuroscience 31, 9345-9352.

Chomsky, N., Halle, M., 1965. Some controversial questions in phonological theory. Journal of Linguistics 1, 97-138.

Cope, T.E., Sohoglu, E., Sedley, W., Patterson, K., Jones, P.S., Wiggins, J., Dawson, C., Grube, M., Carlyon, R.P., Griffiths, T.D., Davis, M.H., Rowe, J.B., 2017. Evidence for causal top-down frontal contributions to predictive processes in speech perception. Nature Communications 8, 2154.

Daube, C., Ince, R.A.A., Gross, J., 2019. Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. Current Biology 29, 1924-1937.e1929.

Davis, M.H., Ford, M.A., Kherif, F., Johnsrude, I.S., 2011. Does semantic context benefit speech understanding through “top–down” processes? Evidence from time-resolved sparse fmri. Journal of Cognitive Neuroscience 23, 3914-3932.

Davis, M.H., Johnsrude, I.S., 2007. Hearing speech sounds: Top-down influences on the interface between audition and speech perception. Hearing Research 229, 132-147.

De Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., Theunissen, F.E., 2017. The hierarchical cortical organization of human speech processing. Journal of Neuroscience 37, 6539-6557.

Desikan, R.S., Segonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage 31, 968-980.

DeWitt, I., Rauschecker, J.P., 2012. Phoneme and word recognition in the auditory ventral stream. Proceedings of the National Academy of Sciences of the United States of America 109, E505-E514.

Di Liberto, Giovanni M., O'Sullivan, James A., Lalor, Edmund C., 2015. Low-frequency cortical entrainment to speech reflects phoneme-level processing. Current Biology 25, 2457-2465.

Díaz, B., Baus, C., Escera, C., Costa, A., Sebastián-Gallés, N., 2008. Brain potentials to native phoneme discrimination reveal the origin of individual differences in learning the sounds of a second language. Proceedings of the National Academy of Sciences of the United States of America 105, 16083-16088.

Ding, N., Simon, J.Z., 2013. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. The Journal of Neuroscience 33, 5728-5735.

Eckert, M.A., Teubner-Rhodes, S., Vaden, K.I., Jr., 2016. Is listening in noise worth it? The neurobiology of speech recognition in challenging listening conditions. Ear and Hearing 37, 101S-110S.

Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences of the United States of America, 201602413.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. " Who" is saying" what"? Brain-based decoding of human voice and speech. Science 322, 970-973.

Friederici, A.D., 2009. Pathways to language: Fiber tracts in the human brain. Trends in Cognitive Sciences 13, 175-181.

Friederici, A.D., 2011. The brain basis of language processing: From structure to function. Physiological Reviews 91, 1357-1392.

Friederici, A.D., Pfeifer, E., Hahne, A., 1993. Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. Cognitive Brain Research 1, 183-192.

Friederici, A.D., Wessels, J.M.I., 1993. Phonotactic knowledge of word boundaries and its use in infant speech perception. Perception and Psychophysics 54, 287-295.

Friston, K., Kiebel, S., 2009. Predictive coding under the free-energy principle. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 364, 1211-1221.

Giraud, A.L., Kell, C., Thierfelder, C., Sterzer, P., Russ, M.O., Preibisch, C., Kleinschmidt, A., 2004. Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. Cerebral Cortex 14, 247-255.

Griffiths, T.D., Hall, D.A., 2012. Mapping pitch representation in neural ensembles with fmri. Journal of Neuroscience 32, 13343-13347.

Hall, D.A., Plack, C.J., 2009. Pitch processing sites in the human auditory brain. Cerebral Cortex 19, 576-585.

Hamilton, L.S., Huth, A.G., 2020. The revolution will not be controlled: Natural stimuli in speech neuroscience. Language, Cognition and Neuroscience 35, 573-582.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nature Reviews Neuroscience 8, 393-402.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55-67.

Holdgraf, C.R., de Heer, W., Pasley, B., Rieger, J., Crone, N., Lin, J.J., Knight, R.T., Theunissen, F.E., 2016. Rapid tuning shifts in human auditory cortex enhance speech intelligibility. Nature Communications 7, 13654.

Howard, M.F., Poeppel, D., 2010. Discrimination of Speech Stimuli Based on Neuronal Response Phase Patterns Depends on Acoustics But Not Comprehension. Journal of Neurophysiology 104, 2500-2511.

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453-458.

Jusczyk, P.W., Luce, P.A., Charles-Luce, J., 1994. Infants' sensitivity to phonotactic patterns in the native language. Journal of Memory and Language 33, 630-645.

Kay, K., Rokem, A., Winawer, J., Dougherty, R., Wandell, B., 2013. Glmdenoise: A fast, automated technique for denoising task-based fmri data. Frontiers in Neuroscience 7.

Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008. Identifying natural images from human brain activity. Nature 452, 352-355.

Khalighinejad, B., Cruzatto da Silva, G., Mesgarani, N., 2017. Dynamic encoding of acoustic features in neural responses to continuous speech. Journal of Neuroscience 37, 2176-2185.

Kim, S.-G., Overath, T., Sedley, W., Kumar, S., Teki, S., Patterson, D.R., Griffths, T.D., in revision. MEG correlates of periodicity relevant to pitch perception in human auditory cortex.

Kleinschmidt, D.F., Jaeger, T.F., 2015. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. Psychological Review 122, 148-203.

Kocagoncu, E., Clarke, A., Devereux, B.J., Tyler, L.K., 2017. Decoding the cortical dynamics of sound-meaning mapping. Journal of Neuroscience 37, 1312-1319.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I., 2009. Circular analysis in systems neuroscience: The dangers of double dipping. Nature Neuroscience 12, 535.

Kujawa, S.G., Liberman, M.C., 2009. Adding insult to injury: Cochlear nerve degeneration after “temporary” noise-induced hearing loss. Journal of Neuroscience 29, 14077-14085.

Kumar, S., Stephan, K.E., Warren, J.D., Friston, K.J., Griffiths, T.D., 2007. Hierarchical processing of auditory objects in humans. PLoS Computational Biology 3, e100.

Kutas, M., Hillyard, S.A., 1983. Event-related brain potentials to grammatical errors and semantic anomalies. Memory and Cognition 11, 539-550.

Ladefoged, P., 2001. Vowels and Consonants : An Introduction to the Sounds of Languages. Wiley-Blackwell.

Ladefoged, P., Johnstone, K., 2015. A Course in Phonetics, Seventh edition. ed. Cengage Learning, Stamford, CT.

Lee, J., Overath, T., in revision. The neural analysis of phonemes is shaped by linguistic analysis.

Leonard, M.K., Baud, M.O., Sjerps, M.J., Chang, E.F., 2016. Perceptual restoration of masked speech in human cortex. Nature Communications 7, 13619.

Lerner, Y., Honey, C.J., Silbert, L.J., Hasson, U., 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. Journal of Neuroscience 31, 2906-2915.

Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54, 1001-1010.

Macmillan, N.A., Kaplan, H.L., 1985. Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. Psychological Bulletin 98, 185-199.

Maris, E., 2019. Enlarging the scope of randomization and permutation tests in neuroimaging and neuroscience. BioRxiv, p. 685560.

Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods 164, 177-190.

Mattys, S.L., Jusczyk, P.W., 2001. Phonotactic cues for segmentation of fluent speech by infants. Cognition 78, 91-121.

McDermott, J.H., Simoncelli, E.P., 2011. Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. Neuron 71, 926-940.

Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485, 233-236.

Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F., 2014. Phonetic feature encoding in human superior temporal gyrus. Science 343, 1006-1010.

Mesgarani, N., David, S.V., Fritz, J.B., Shamma, S.A., 2008. Phoneme representation and classification in primary auditory cortex. Journal of the Acoustical Society of America 123, 899-909.

Millman, R.E., Johnson, S.R., Prendergast, G., 2015. The Role of Phase-locking to the Temporal Envelope of Speech in Auditory Perception and Speech Intelligibility. Journal of Cognitive Neuroscience 27, 533-545.

Moerel, M., De Martino, F., Kemper, V.G., Schmitter, S., Vu, A.T., Uğurbil, K., Formisano, E., Yacoub, E., 2018. Sensitivity and specificity considerations for fmri encoding, decoding, and mapping of auditory cortex at ultra-high field. Neuroimage 164, 18-31.

Moerel, M., De Martino, F., Santoro, R., Ugurbil, K., Goebel, R., Yacoub, E., Formisano, E., 2013. Processing of natural sounds: Characterization of multipeak spectral tuning in human auditory cortex. Journal of Neuroscience 33, 11888-11898.

Moore, B.C.J., 1996. Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids. Ear and Hearing 17, 133-161.

Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication 9, 453-467.

Narain, C., Scott, S.K., Wise, R.J.S., Rosen, S., Leff, A., Iversen, S.D., Matthews, P.M., 2003. Defining a left-lateralized response specific to intelligible speech using fmri. Cerebral Cortex 13, 1362-1368.

Naselaris, T., Olman, C.A., Stansbury, D.E., Ugurbil, K., Gallant, J.L., 2015. A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. Neuroimage 105, 215-228.

Norman-Haignere, S., Kanwisher, Nancy G., McDermott, Josh H., 2015. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88, 1281-1296.

Norman-Haignere, S.V., Long, L.K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E.M., Feldstein, N.A., McKhann, G.M., Schevon, C.A., Flinker, A., 2020. Multiscale integration organizes hierarchical computation in human auditory cortex. BioRxiv.

Nunez-Elizalde, A.O., Huth, A.G., Gallant, J.L., 2019. Voxelwise encoding models with non-spherical multivariate normal priors. Neuroimage 197, 482-492.

Obleser, J., Eisner, F., Kotz, S.A., 2008. Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. Journal of Neuroscience 28, 8116-8123.

Obleser, J., Leaver, A., VanMeter, J., Rauschecker, J., 2010. Segregation of vowels and consonants in human auditory cortex: Evidence for distributed hierarchical organization. Frontiers in Psychology 1, 232.

Overath, T., McDermott, J.H., Zarate, J.M., Poeppel, D., 2015. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nature Neuroscience 18, 903-911.

Overath, T., Paik, J.H., 2021. From acoustic to linguistic analysis of temporal speech structure: Acousto-linguistic transformation during speech perception using speech quilts. Neuroimage, 117887.

Overath, T., Zhang, Y., Sanes, D.H., Poeppel, D., 2012. Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: Fmri evidence. Journal of Neurophysiology 107, 2042-2056.

Pandya, P.K., Rathbun, D.L., Moucha, R., Engineer, N.D., Kilgard, M.P., 2007. Spectral and temporal processing in rat posterior auditory cortex. Cerebral Cortex 18, 301-314.

Park, H., Ince, Robin A.A., Schyns, Philippe G., Thut, G., Gross, J., 2015. Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. Current Biology 25, 1649-1653.

Plack, C.J., Oxenham, A.J., Fay, R.R., Popper, A., 2005. Pitch: Neural Coding and Perception. Springer, New York.

Poeppel, D., Idsardi, W.J., Wassenhove, V.v., 2008. Speech perception at the interface of neurobiology and linguistics. Philosophical Transactions of the Royal Society B: Biological Sciences 363, 1071-1086.

Poldrack, R.A., Huckins, G., Varoquaux, G., 2020. Establishment of best practices for evidence for prediction: A review. JAMA Psychiatry 77, 534-540.

Qi, Y., Fox, R.A., 1992. Analysis of nasal consonants using perceptual linear prediction. Journal of the Acoustical Society of America 91, 1718-1726.

Rao, R.P.N., Ballard, D.H., 1999. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience 2, 79-87.

Rauschecker, J.P., Scott, S.K., 2009. Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. Nature Neuroscience 12, 718-724.

Rauschecker, J.P., Tian, B., 2004. Processing of band-passed noise in the lateral auditory belt cortex of the rhesus monkey. Journal of Neurophysiology 91, 2578-2589.

Rauschecker, J.P., Tian, B., Hauser, M., 1995. Processing of complex sounds in the macaque nonprimary auditory cortex. Science 268, 111-114.

Ringach, D.L., Sapiro, G., Shapley, R., 1997. A subspace reverse-correlation technique for the study of visual neurons. Vision Research 37, 2455-2464.

Ruggles, D., Bharadwaj, H., Shinn-Cunningham, B.G., 2011. Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. Proceedings of the National Academy of Sciences of the United States of America 108, 15516-15521.

Rutten, S., Santoro, R., Hervais-Adelman, A., Formisano, E., Golestani, N., 2019. Cortical encoding of speech enhances task-relevant acoustic information. Nature Human Behaviour 3, 974-987.

Saenz, M., Langers, D.R.M., 2014. Tonotopic mapping of human auditory cortex. Hearing Research 307, 42-52.

Saffran, J.R., Newport, E.L., Aslin, R.N., 1996. Word segmentation: The role of distributional cues. Journal of Memory and Language 35, 606-621.

Samuel, A.G., 1981. Phonemic restoration: Insights from a new methodology. Journal of Experimental Psychology: General 110, 474-494.

Samuel, A.G., 1987. Lexical uniqueness effects on phonemic restoration. Journal of Memory and Language 26, 36-56.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., Formisano, E., 2014. Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Computational Biology 10.

Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., Formisano, E., 2017. Reconstructing the spectrotemporal modulations of real-life sounds from fmri response patterns. Proceedings of the National Academy of Sciences of the United States of America 114, 4799-4804.

Schönwiesner, M., Zatorre, R.J., 2009. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fmri. Proceedings of the National Academy of Sciences of the United States of America 106, 14611-14616.

Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J.S., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123, 2400-2406.

Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270, 303-304.

Shin, J., 2015. Vowels and consonants. In: Brown, L., Yeon, J. (Eds.), The handbook of Korean linguistics. Wiley-Blackwell, UK, pp. 3-21.

Shinn-Cunningham, B.G., Best, V., 2008. Selective attention in normal and impaired hearing. Trends in Amplification 12, 283-299.

Sohn, H.-M., 2001. The Korean Language. Cambridge University Press, NY.

Sohoglu, E., Peelle, J.E., Carlyon, R.P., Davis, M.H., 2012. Predictive top-down integration of prior knowledge during speech perception. Journal of Neuroscience 32, 8443-8453.

Stevens, K.N., 2000. Acoustic Phonetics. MIT press.

Theunissen, F., Miller, J., 1995. Temporal encoding in nervous systems: A rigorous definition. Journal of Computational Neuroscience 2, 149-162.

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J.Z., Francart, T., 2018. Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope. Journal of the Association for Research in Otolaryngology 19, 181-191.

Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. Neuroimage 145, 166-179.

Verschueren, E., Vanthornhout, J., Francart, T., 2021. The effect of stimulus intensity on neural envelope tracking. Hearing Research 403, 108175.

Wang, Y., Ding, N., Ahmar, N., Xiang, J., Poeppel, D., Simon, J.Z., 2012. Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: Meg evidence. Journal of Neurophysiology 107, 2033-2041.

Warren, J.D., Jennings, A.R., Griffiths, T.D., 2005. Analysis of the spectral envelope of sounds by the human brain. Neuroimage 24, 1052-1057.

Warren, R.M., 1970. Perceptual restoration of missing speech sounds. Science 167, 392-393.

Wessinger, C.M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., Rauschecker, J.P., 2001. Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. Journal of Cognitive Neuroscience 13, 1-7.

Wild, C.J., Davis, M.H., Johnsrude, I.S., 2012. Human auditory cortex is sensitive to the perceived clarity of speech. Neuroimage 60, 1490-1502.

Wu, M.C.-K., David, S.V., Gallant, J.L., 2006. Complete functional characterization of sensory neurons by system identification. Annual Review of Neuroscience 29, 477-505.

Yi, H.G., Leonard, M.K., Chang, E.F., 2019. The encoding of speech sounds in the superior temporal gyrus. Neuron 102, 1096-1110.

Yoon, T.-J., Kang, Y., 2013. The Korean Phonetic Aligner Program Suite.

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

7 Supplementary materials

7.1 Supplementary figures

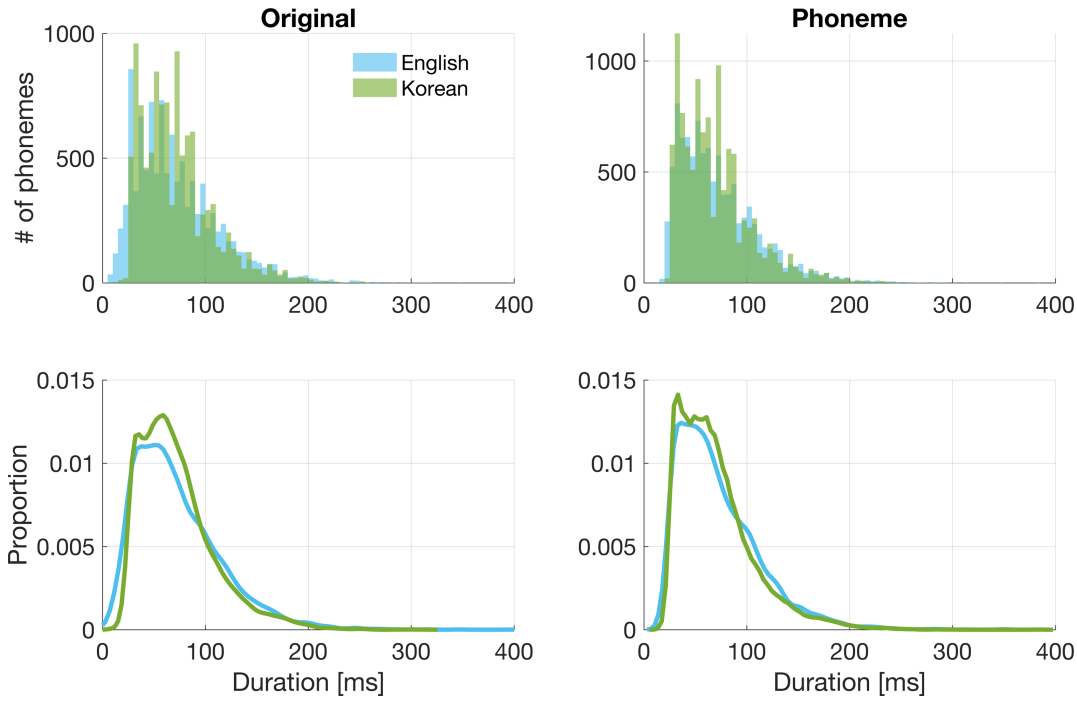

Figure S1. Histograms of phoneme durations. The distributions of phoneme durations in the Original natural speech (left) and the Phoneme quilts (right) are shown for English (light blue) and Korean (lime green) in histograms (top) and smoothed density functions (bottom). Non-linguistic segments (e.g.,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

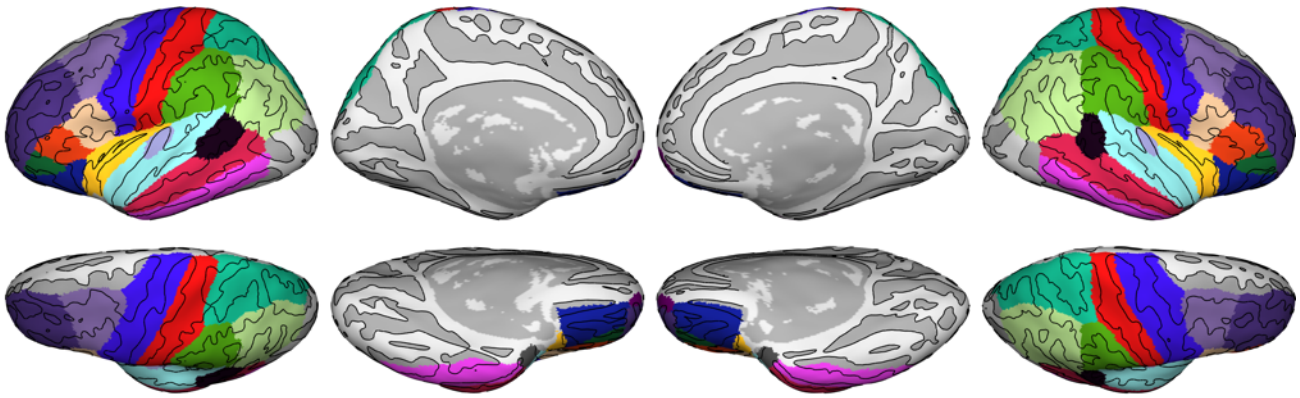

Figure S2. Regions of interest. From the automatic parcellation based on the Desikan-Killiany cortical atlas (Desikan et al., 2006) in FreeSurfer, the following 19 labels were

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

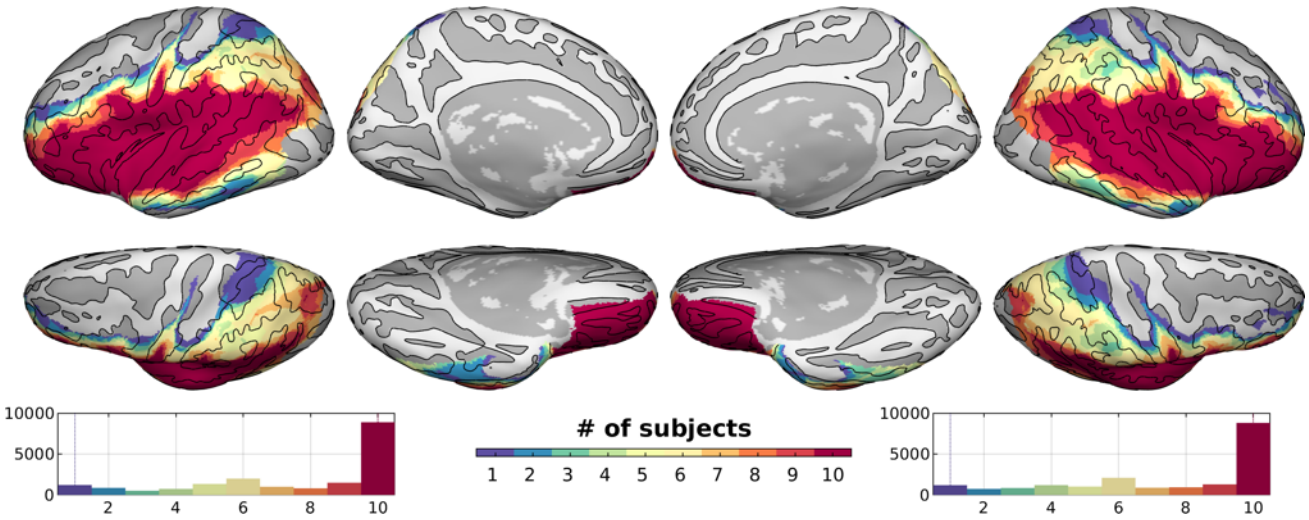

Figure S3. Overlap of selected vertices across participants (N = 10). The colored histograms below display the number of vertices over the number of participants.

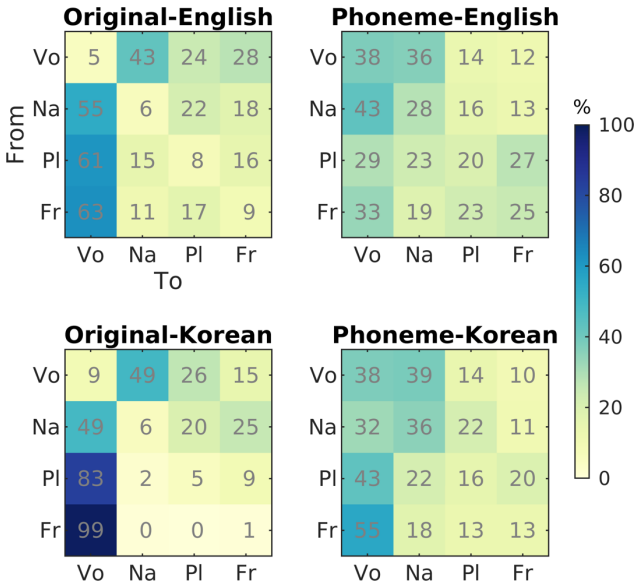

Figure S4. Transition probability between phoneme classes. Phoneme transitions were counted as consecutive occurrences of four phoneme classes without taking word boundaries into account, and cumulated over all stimuli for visualization.

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

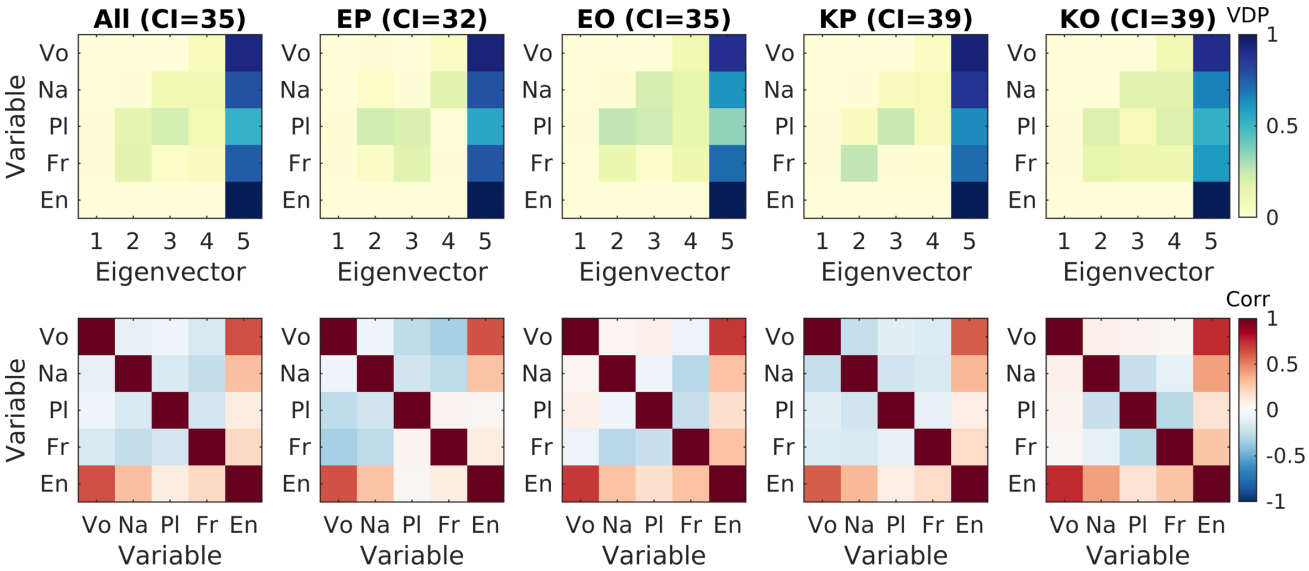

Figure S5. Collinearity of predictors. Variance decomposition proportion (VDP, upper) and pair-wise Pearson correlation (corr, lower) between predictive variables (Vo, vowel; Na, nasal; Pl, plosive; Fr, fricative; En, envelope) are shown for all conditions together (left-most column) and for each condition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

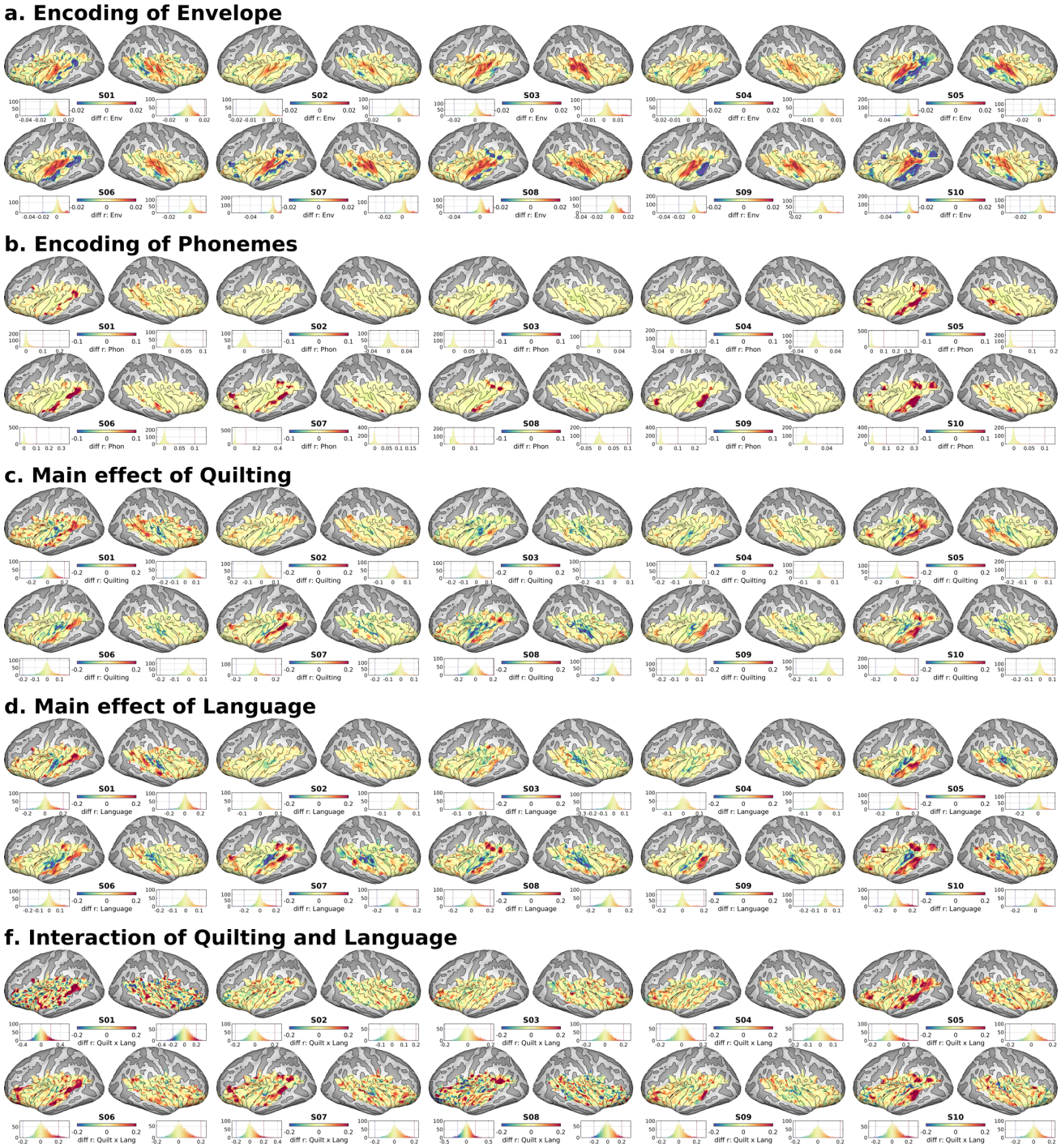

Figure S6. Encoding of features and conditions in individual participants. Unthresholded effect size maps (differences in Pearson correlation) are shown for (a) Envelope, (b) Phonemes, (c) Quilting, (d) Language, and (e) the interaction of Quilting and Language.

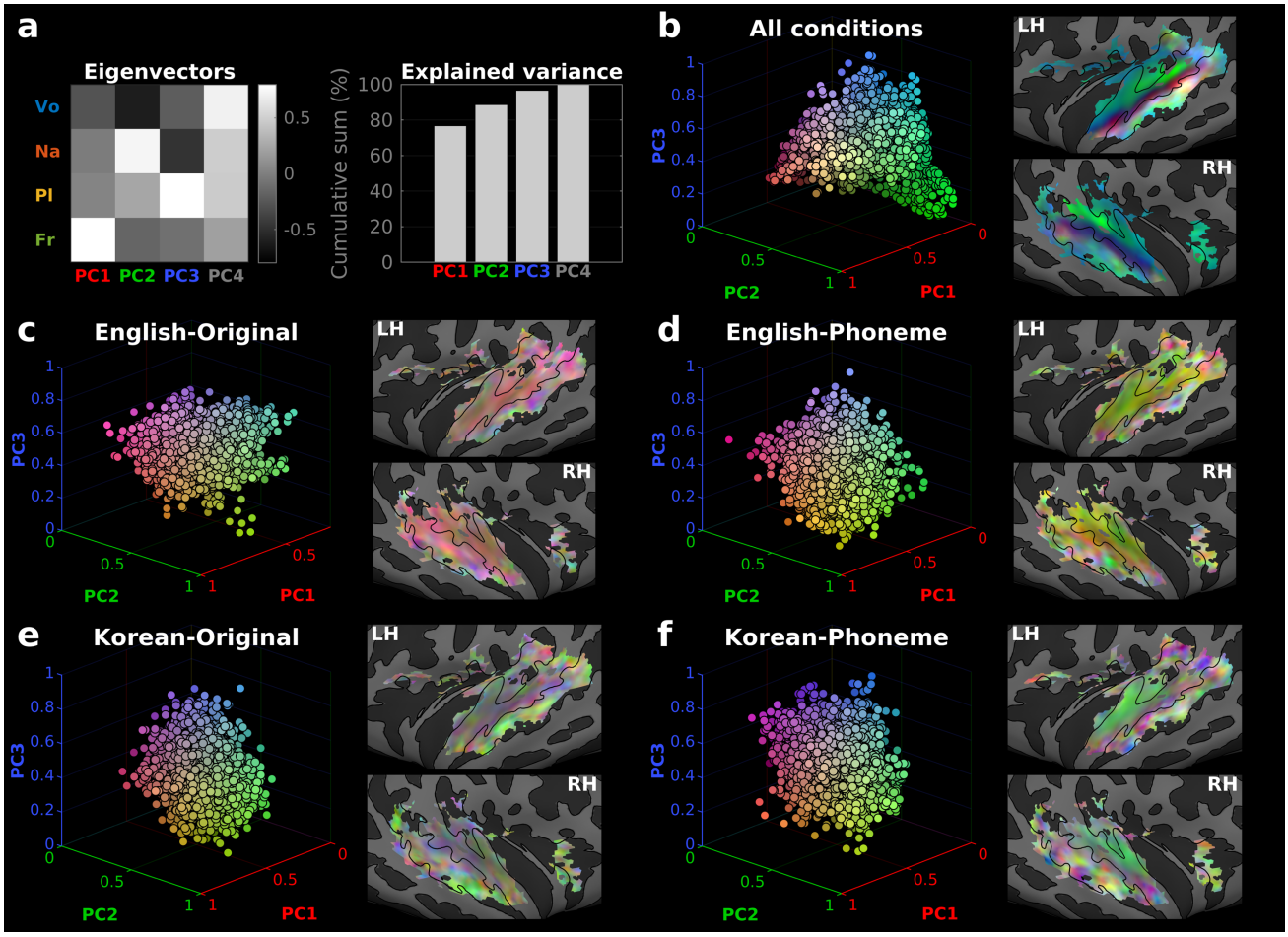

Figure S7. Low dimensional projection (3-D; RGB) of the phoneme-class encoding vectors (4-D) using principal component analysis (PCA). (a) Eigenvectors (left) and the explained variance (right) of the principal components. The first three components explained >98% of the variance. (b) RGB visualization 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.

7.2 Supplementary tables

	Vowel	Nasal/Approximant	Plosive	Fricative
English	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
Korean	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

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