Article

Current Biology

Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing

Highlights

- EEG reflects categorical processing of phonemes within continuous speech
- EEG is best modeled when representing speech as acoustic signal plus phoneme labels
- Neural delta and theta bands reflect this speech-specific cortical activity
- Specific speech articulatory features are discriminable in EEG responses

Authors

Giovanni M. Di Liberto, James A. O'Sullivan, Edmund C. Lalor

Correspondence

diliberg@tcd.ie (G.M.D.L.), edlalor@tcd.ie (E.C.L.)

In Brief

Di Liberto et al. show that EEG responses to natural speech are best modeled when that speech is represented in terms of its low-level acoustics plus a categorical labeling of phonetic features. This suggests that EEG reflects categorical phoneme-level speech processing and provides a new framework for studying such processing.

Di Liberto et al., 2015, Current Biology *25*, 2457–2465
October 5, 2015
http://dx.doi.org/10.1016/j.cub.2015.08.030

CellPress

Current Biology
Article

Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing

Giovanni M. Di Liberto,[1,*] James A. O'Sullivan,[1] and Edmund C. Lalor[1,*]
[1]Trinity College Institute of Neuroscience, School of Engineering, and Trinity Centre for Bioengineering Trinity College Dublin, Dublin 2, Ireland
*Correspondence: diliberg@tcd.ie (G.M.D.L.), edlalor@tcd.ie (E.C.L.)
http://dx.doi.org/10.1016/j.cub.2015.08.030

SUMMARY

The human ability to understand speech is underpinned by a hierarchical auditory system whose successive stages process increasingly complex attributes of the acoustic input. It has been suggested that to produce categorical speech perception, this system must elicit consistent neural responses to speech tokens (e.g., phonemes) despite variations in their acoustics. Here, using electroencephalography (EEG), we provide evidence for this categorical phoneme-level speech processing by showing that the relationship between continuous speech and neural activity is best described when that speech is represented using both low-level spectrotemporal information and categorical labeling of phonetic features. Furthermore, the mapping between phonemes and EEG becomes more discriminative for phonetic features at longer latencies, in line with what one might expect from a hierarchical system. Importantly, these effects are not seen for time-reversed speech. These findings may form the basis for future research on natural language processing in specific cohorts of interest and for broader insights into how brains transform acoustic input into meaning.

INTRODUCTION

Humans effortlessly parse the spectrotemporally complex acoustic patterns of continuous speech into coherent, categorical, semantic representations. This is true despite enormous variations in the low-level spectrotemporal dynamics of speech across listening conditions, including different speaker accents, coarticulation effects, and prosodic fluctuations. Although the precise neurophysiological mechanisms and neuroanatomic infrastructure underpinning this ability are not well understood [1], it has been proposed that robust speech perception is the product of a hierarchical auditory processing system whose successive stages process increasingly complex attributes of the audio input [2–4]. In particular, it has been suggested that, while earlier areas of the auditory system undoubtedly respond to acoustic differences in speech tokens, later areas must exhibit consistent neural responses to those tokens in order to produce a categorical perception of words and phonemes [2–4].

fMRI [5, 6], nonhuman primate electrophysiology [7], and electrocorticography (ECoG) [4, 8, 9] have all made important contributions to our understanding of hierarchical speech encoding in the brain. However, all of these methods have their shortcomings. Electroencephalography (EEG) and magnetoencephalography (MEG), as macroscopic non-invasive technologies, may offer important opportunities for further progress. Although these approaches have been used for years to study the processing of discrete syllables [10], recent research has also shown that both EEG and MEG index cortical entrainment to the low-frequency amplitude envelope of natural speech [11–13]. This has proved useful for investigating the mechanisms underlying speech processing [14], how such processing is affected by attention [13, 15, 16], and how audio and visual speech interact [17, 18]. However, it remains unclear to what extent these EEG/MEG indices reflect higher-level speech-specific processing versus lower-level processing of the spectrotemporal/acoustic stimulus dynamics [19].

There has been somewhat equivocal evidence that speech intelligibility affects these envelope entrainment measures, suggesting that they may indeed index speech-specific processing [20–22]. These findings have led to the suggestion that different neural populations, having different functional roles in receptive speech processing, may simultaneously contribute to envelope entrainment measures [19]. Support for this notion comes from ECoG research focusing on low-frequency entrainment to speech that has shown differential effects of "cocktail party"-type attention in low- and higher-level auditory cortical areas [9]. However, there has been no definitive evidence to date that low-frequency EEG or MEG entrainment reflects processing at the level of categorical speech perception.

Here we investigated the degree to which low-level (envelope, spectrogram) and higher-level (phonemic, phonetic feature) characteristics of natural speech are reflected in EEG activity. In doing so, we provide evidence that EEG not only reflects passive neural following of the acoustic energy of speech but also indexes the categorical perception of phonemes in the human brain. Furthermore, we sought to determine whether processing of different phonetic features can be discriminated in EEG responses and found that this discriminative power varies as a function of response latency, in line with what one might expect of a hierarchical system.

RESULTS

128-channel EEG was recorded from ten subjects as they listened to segments of an audiobook and ten subjects who listened to the same audiobook played in reverse (five subjects undertook both experiments). To identify neural indices of lower- and higher-level speech processing, we investigated mappings

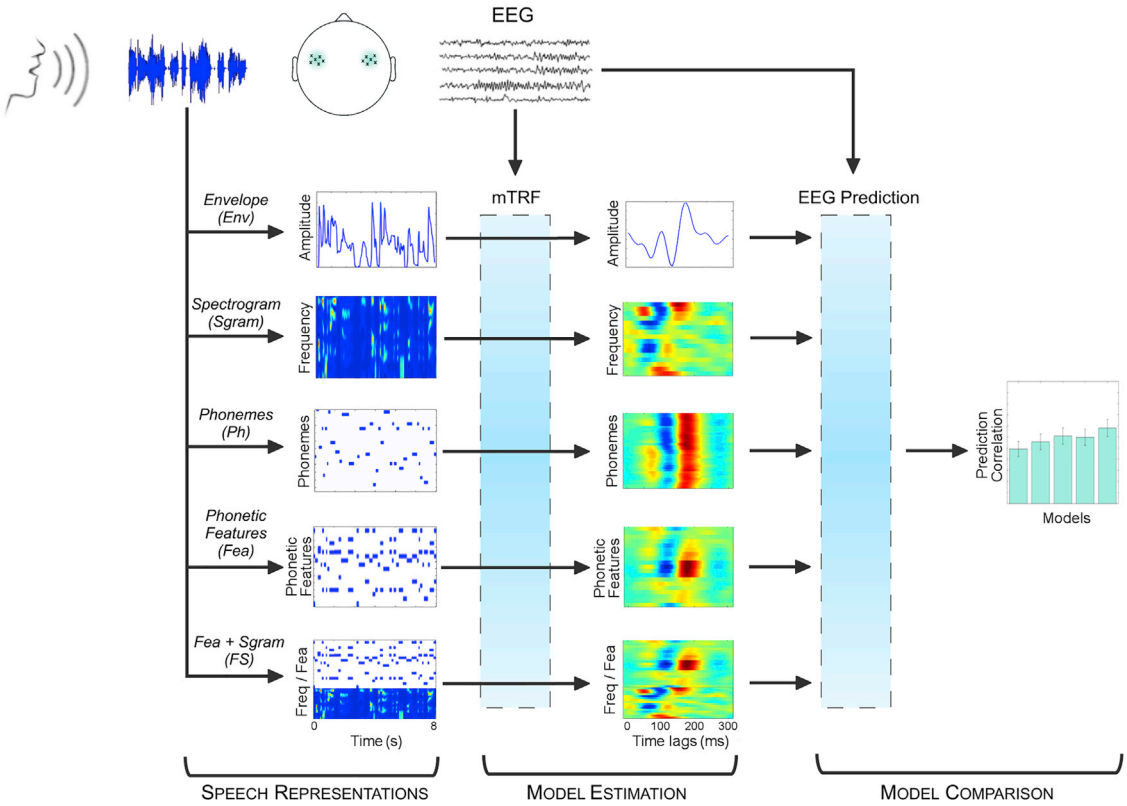

Figure 1. Assessing the Encoding of Speech Features in EEG

128-channel EEG data were recorded while subjects listened to continuous, natural speech consisting of a male speaker reading from a novel or its time-reversed complement. Linear regression was used to fit multivariate temporal response functions (mTRFs) between the low-frequency (1–15 Hz) EEG data and five different representations of the speech stimulus. Each mTRF model was then tested for its ability to predict EEG using leave-one-out cross-validation.

between different representations of the speech and the low-frequency (1–15 Hz) EEG (Figure 1). Specifically, we did this by using linear regression to model the relationship between each speech representation and the data from each EEG channel (Figure 1). This approach has been used previously to describe the relationship between the speech envelope and EEG [12], MEG [23], and ECoG [9] data. The resulting models are commonly referred to as temporal response functions (TRFs). Here, as we will be representing speech using multiple variables, we refer to our models as multivariate temporal response functions (mTRFs).

Neural Evidence for Phonetic Processing

We employed a cross-validation approach to quantify how well each speech representation related to the neural data. Specifically, we fit our mTRF models using a subset of the speech segments for each subject and used these models to predict the data corresponding to the remaining segments. The quality of the prediction was assessed using correlation (Pearson's r). The overarching rationale was to use variations in these EEG prediction scores across speech representations as a dependent measure for assessing how well the EEG reflects the processing of lower- and higher-level speech features. We focused our analysis on the EEG data from six bilateral pairs of frontotemporal electrodes in order to investigate auditory cortical activity bilaterally (see Supplemental Experimental Procedures and Figure S1).

We tested five speech representations (Figure 1; see Supplemental Experimental Procedures): (1) broadband amplitude envelope, *Env*; (2) spectrogram, *Sgram*; (3) time-aligned sequence of phonemes, *Ph*; (4) time-aligned sequence of phonetic features, *Fea*; and (5) a combination of time-aligned phonetic features and spectrogram, *FS*. Neural entrainment to speech envelopes is well established and, as such, performance of the *Env* model acted as a baseline with which to compare the performance of the other models. Robust mappings between speech spectrograms and high-gamma-frequency ECoG have been previously shown [24]. However it is unknown whether this richer representation can be accurately indexed using low-frequency EEG, something we address with the *Sgram* model. Similarly, the relationship between high-frequency ECoG and a categorical phoneme representation of speech has been examined before [8]. However, no such relationship has been investigated for EEG (or MEG), hence the *Ph* model. Transforming phonemes into a lower-dimensional phonetic-feature representation [25] frames our results in terms of the articulatory and acoustic properties of each phoneme and has advantages for the efficiency of this type of modeling. This motivated our *Fea* model.

An important issue when considering the spectrogram representation and the phonemic/phonetic-feature representations is that they are mutually highly redundant. This is because, on average, each phoneme will have a particular characteristic spectrotemporal profile. So if each phoneme were always

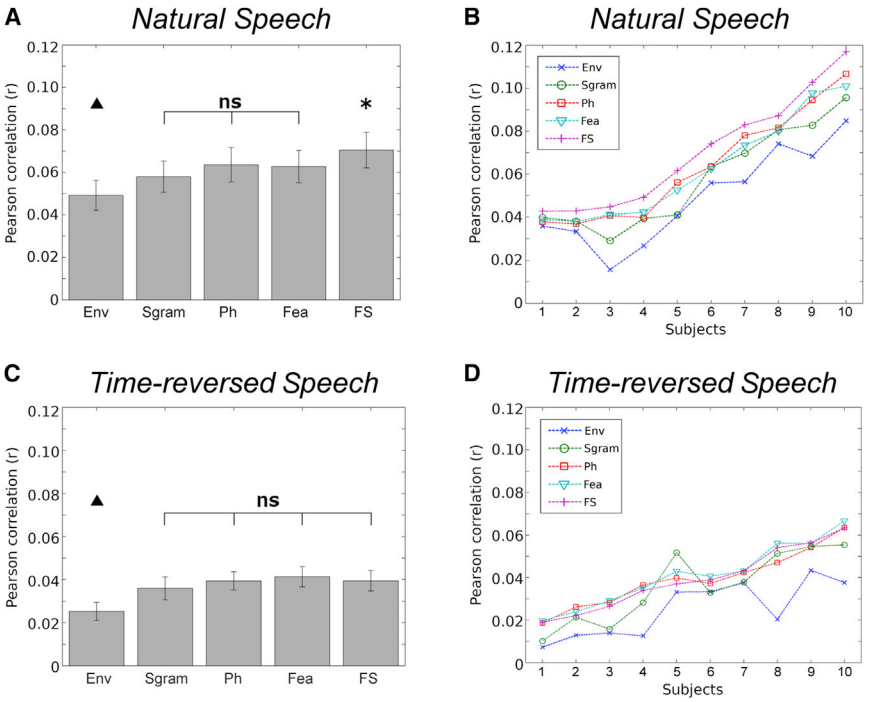

Figure 2. EEG Responses to Forward Speech, but Not Time-Reversed Speech, Are Best Predicted When Speech Is Represented as a Combination of Spectrotemporal Features and Phonetic-Feature Labels

(A) Grand-average EEG prediction correlations (Pearson's r) for each speech model (mean ± SEM). While there is no statistical difference in prediction performance between the spectrogram (*Sgram*), phonetic-features (*Fea*), and phonemic (*Ph*) models ($p > 0.05$), all of these models are better predictors of the EEG than that based on the envelope (*Env*; ▲$p < 0.01$). Importantly, the model based on the combination of phonetic features and spectrogram (*FS*) outperforms all other models (*$p < 0.01$).

(B) Correlation values between recorded EEG and that predicted by each mTRF model for individual subjects. The subjects are sorted according to the prediction correlations of the *FS* model. Although the results show variability across subjects, the *FS* model outperforms all the other models for every subject. The model based on the speech envelope (*Env*) performs worse than every other model for every subject.

(C) Grand-average EEG prediction correlations for the time-reversed speech condition (mean ± SEM). Prediction correlations using the model based on the envelope (*Env*) are lower than those for all other models (▲$p < 0.05$). As with normal speech, there is no statistical difference in prediction performance between the spectrogram (*Sgram*), phonetic-features (*Fea*), and phonemic (*Ph*) models ($p > 0.05$). Importantly in this case, there is also no difference between the performance of those models and that based on the combination of phonetic features and spectrogram (*FS*; $p > 0.05$).

(D) Correlation values between recorded EEG and that predicted by each mTRF model for individual subjects for time-reversed speech. The subjects are sorted according to the prediction correlations of the *FS* model. The model based on the speech envelope (*Env*) performs worse than every other model for every subject.

spoken in the same way, then the two representations would be equivalent. However, in natural speech this is not the case, with significant variation in the spectrotemporal profile of a given phoneme across instances. One might thus expect that our *Ph* model, which is ignorant of these variations, would underperform relative to the *Sgram* model. However, it is also true that human listeners categorically perceive phonemes despite spectrotemporal variations, a fact that is presumably underpinned by consistent neural responses to those phonemes [2, 3]. Such consistent responses would be captured by our *Ph* model, potentially leading to it outperforming the *Sgram* model, which is ignorant of the categorical nature of these utterances. Indeed, given their mutual redundancy and complementary strengths, both models may perform similarly. To attempt to reveal their complementary strengths, we also derived a model based on combining the time-aligned phonetic features and the corresponding speech spectrogram (the *FS* model). Improved performance of this model over the others would suggest that the EEG is indexing the processing of both low-level acoustic fluctuations and higher-level phonetic features.

In line with this hypothesis, the average performance of the *FS* model across our 12 chosen electrodes was better than all other models (ANOVA: $F(1.5,13.6) = 29.1$, $p = 2.9 \times 10^{-5}$; post hoc paired t test comparisons of *FS* with all other models: $p = 0.001$, $p = 0.002$, $p = 8.2 \times 10^{-5}$, $p = 0.001$ for *Env*, *Sgram*, *Ph*, and *Fea* respectively; Figure 2A). Indeed, *FS* was best for all ten subjects (Figure 2B). The fact that the *Ph* and *Fea* models are simple transformations of one another was reflected in the lack of any performance difference between them ($p > 0.05$). There was also no difference between the *Ph* or *Fea* and *Sgram* models ($p = 0.24$ and $p = 0.34$, respectively), which, as mentioned previously, was always a possibility. Importantly, given the reliance on envelope in many studies of speech neurophysiology [9, 14, 26], the *Env* model underperformed relative to all other models ($p < 0.01$). Furthermore, we found no lateralization effects in the performance of any model ($p > 0.05$, ANOVA). Indeed, model performances were qualitatively similar at other scalp locations (Figure S2).

While we contend that the improved performance of the *FS* model is evidence for the encoding of both low-level acoustic variations and higher-level phonetic features, it remained possible that this result was driven by the *FS* model having more free parameters than the other models. We sought to test whether or not this was the case by investigating the performance of several other high-dimensional models. Combining *Ph* and *Sgram* did not outperform the *FS* model ($p > 0.05$), even though it has 16 additional dimensions. Also, combining the *Ph* and *Fea* models did not outperform either the *Ph* or *Fea* models alone ($p > 0.05$). These results suggest that the greater number of parameters in the *FS* model does not explain our finding.

However, to further establish the validity of our interpretation, we performed the same analyses on the data from the subjects who listened to time-reversed speech. Because the same speech segments were used, the same *Env*, *Sgram*, *Ph*, *Fea*, and *FS* models could be used in a time-reversed fashion. The key manipulation here is that time-reversed speech has the

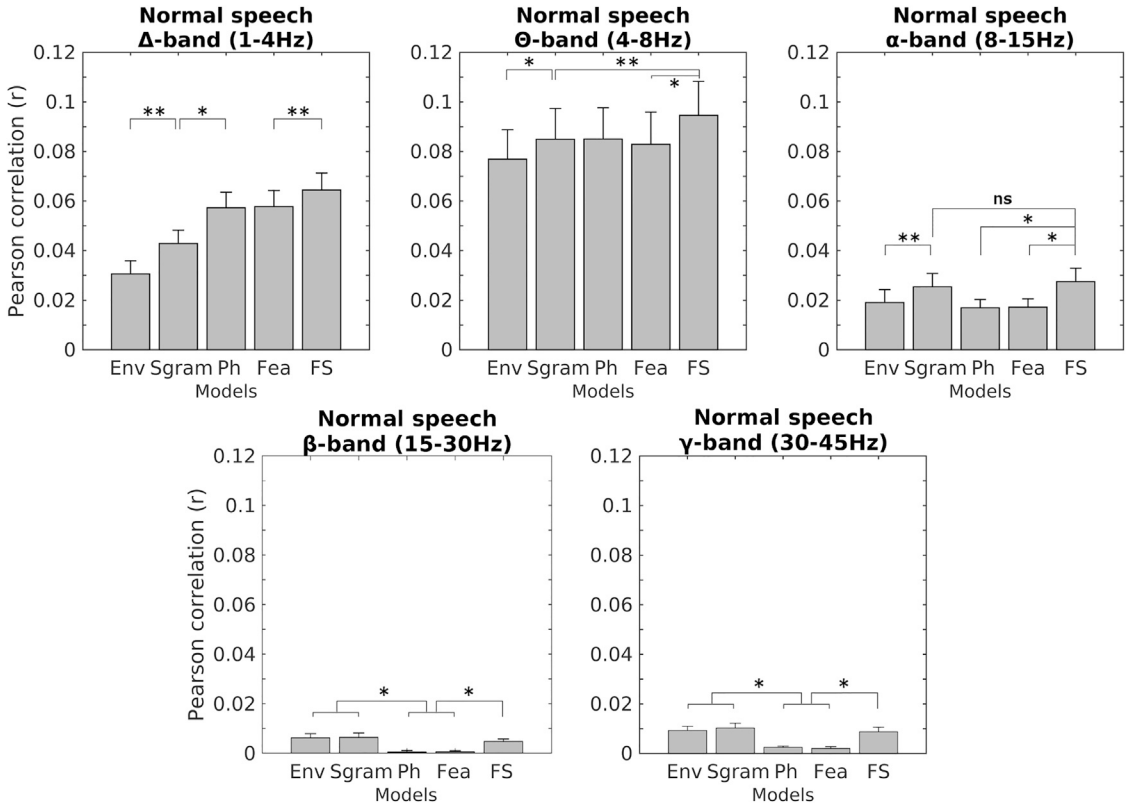

Figure 3. EEG Response Prediction for Different EEG Frequency Bands

Grand-average EEG prediction correlations (Pearson's r) for each speech model (mean ± SEM) for delta (A), theta (B), alpha (C), beta (D), and low-gamma (E) EEG frequencies. * indicates prediction differences at the level of $p < 0.05$ and ** at the level of $p < 0.01$, both using planned paired t tests.

same long-term amplitude spectrum as natural speech but is not perceived as intelligible speech. Overall, the prediction values were lower than for forward speech, likely as a result of differences in top-down attention and also consistent with previous research showing weaker neural entrainment to unintelligible speech [20]. However, crucially, whereas the *Env* model again performed more poorly than the others ($p < 0.05$), the *FS* model in this case showed no improvement over the *Sgram*, *Ph*, or *Fea* models ($p > 0.05$; Figures 2C and 2D). This supports our contention that the *FS* model in the case of forward speech indexes the neural processing of speech features at the level of phonemes.

Phonetic Processing across Different EEG Frequency Bands

Given previous research positing different functional roles in speech processing for different cortical oscillations [27] and in particular differential encoding of speech features by delta- and theta-band entrainment [19], we examined the improved model performances for different frequency bands. Phoneme-level processing (i.e., *FS* outperforming all other models) was evident only in the delta and theta bands (Figure 3). The *Ph* and *Fea* models outperformed the *Sgram* model for the delta band while the *Sgram* model outperformed the *Ph* and *Fea* models for the alpha, beta, and low-gamma bands, possibly evincing the differential sensitivity of these bands to detailed acoustic information (*Sgram*) and categorical phonemic processing (*Ph* and *Fea*). Relative model performances in the theta band are qualitatively similar to the results obtained above with the broadband (1–15 Hz) signal. EEG prediction scores are very low for beta and low gamma, in keeping with the generally low signal-to-noise ratio for EEG at these frequencies. Given the higher prediction scores for delta, theta, and alpha and the phonetic processing effects visible using the broadband (1–15 Hz) EEG signal, we continue to analyze this broader representation of the EEG in the following section.

Insights into Hierarchical Speech Encoding Based on Multivariate Temporal Response Functions

Examination of the various mTRFs gives insight into the factors driving the performance of each model. Contrasting the *Env* model (Figure 4A) with the *Sgram* model (Figure 4B) explains why the former leads to poorer prediction, as it discards a wealth of response variability across frequency by reducing the speech signal to a scalar value.

The *Ph* mTRF shows variable dynamics across phonemes (Figure 4C). To reveal groups of phonemes with similar responses, we performed hierarchical clustering on the *Ph* mTRFs at the 12 electrodes of interest (see Supplemental Experimental Procedures). In doing so, we found that the model could accurately discriminate consonants and vowels (32 of 35 phonemes classified correctly). For visualization purposes (Figure 4C), we present the phonemes grouped as vowels, diphthongs, semi-vowels,

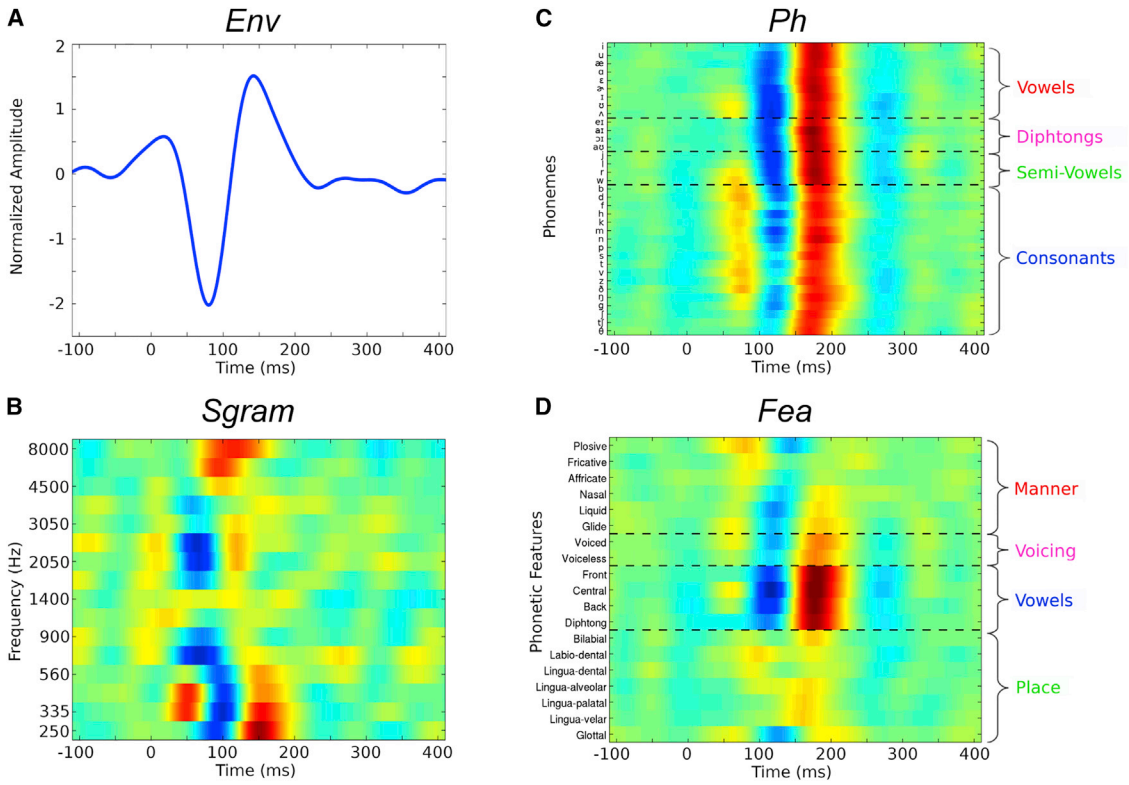

Figure 4. mTRF Models for Natural Speech Reflect Sensitivity to Different Speech Features

mTRFs plotted for envelope (*Env*, A), spectrogram (*Sgram*, B), phonemic (*Ph*, C), and phonetic-features (*Fea*, D) models at peri-stimulus time lags from −100 to 400 ms for natural speech, averaged over 12 frontotemporal electrodes (see Figure S1). The phonemes were sorted based on a hierarchical clustering analysis on the average mTRF after grouping them into vowels, diphthongs, semi-vowels, and consonants. Horizontal dashed lines separate distinct categories of phonemes and phonetic features.

and consonants, and we sort within each group according to the hierarchical clustering distances. It can be seen that the mTRFs for consonants show activation by around 50 ms, while those for vowels do not generally show a significant response before ~100 ms. In contrast to these early differences, all phonemes show a similar response between ~150 and 200 ms. This timing pattern is consistent with previous research showing that acoustic-phonetic features of speech modulate activity in non-primary auditory cortex from 50–100 ms onward, with language-specific phonetic-phonological analysis starting by 100–200 ms [10]. Interestingly, in the case of time-reversed speech, the *Ph* mTRF amplitude is noticeably lower than for forward speech, particularly during the 150–200 ms interval (Figure S3).

When considering the mTRF for the *Fea* model, it should be remembered that each phoneme is simply a combination of phonetic features. Indeed, a linear mapping from the *Fea* mTRF to the phonemic space produced a model that is highly correlated with the *Ph* model ($r = 0.93$, $p = 1.6 \times 10^{-5}$; two-tailed t test). We therefore consider these two models to be essentially equivalent. However, while the mTRF for the *Ph* model highlights differences between vowels and consonants, the mTRF for the *Fea* model allows us to visualize sensitivity to different articulatory speech features (Figure 4D). Again, the vowels stand out strongly from the features associated with consonants. But within each of the consonant-related features, a considerable degree of variability is evident across the specific distinctions.

Sensitivity of EEG to Phonetic Features as a Function of Latency

We wished to test the hypothesis that the sensitivity of our neural responses to different acoustic and phonetic features would increase as a function of response latency in line with what one might expect of a hierarchical system. To do this, we applied unsupervised multi-dimensional scaling (see Experimental Procedures) to the mTRFs in the time intervals 50–100 ms, 100–150 ms, and 150–200 ms, which correspond approximately to the three main peaks in the *Ph* mTRF (Figure 4C). This approach allowed us to build a geometric space in which the Euclidean distance between phonemes (or phonetic features) corresponds to the similarity of their neural responses. Furthermore, this allowed us to examine how sensitive the neural responses were to different phonetic features by quantifying how well the responses clustered according to the different groups of phonetic features that produced them. We did this by performing k-means clustering, where k is the number of groups under consideration, and then calculating the corresponding F-scores (the harmonic mean of precision and recall) between the actual grouping and the result of the clustering (see Supplemental Experimental Procedures). All statistical tests were performed using a jackknifed

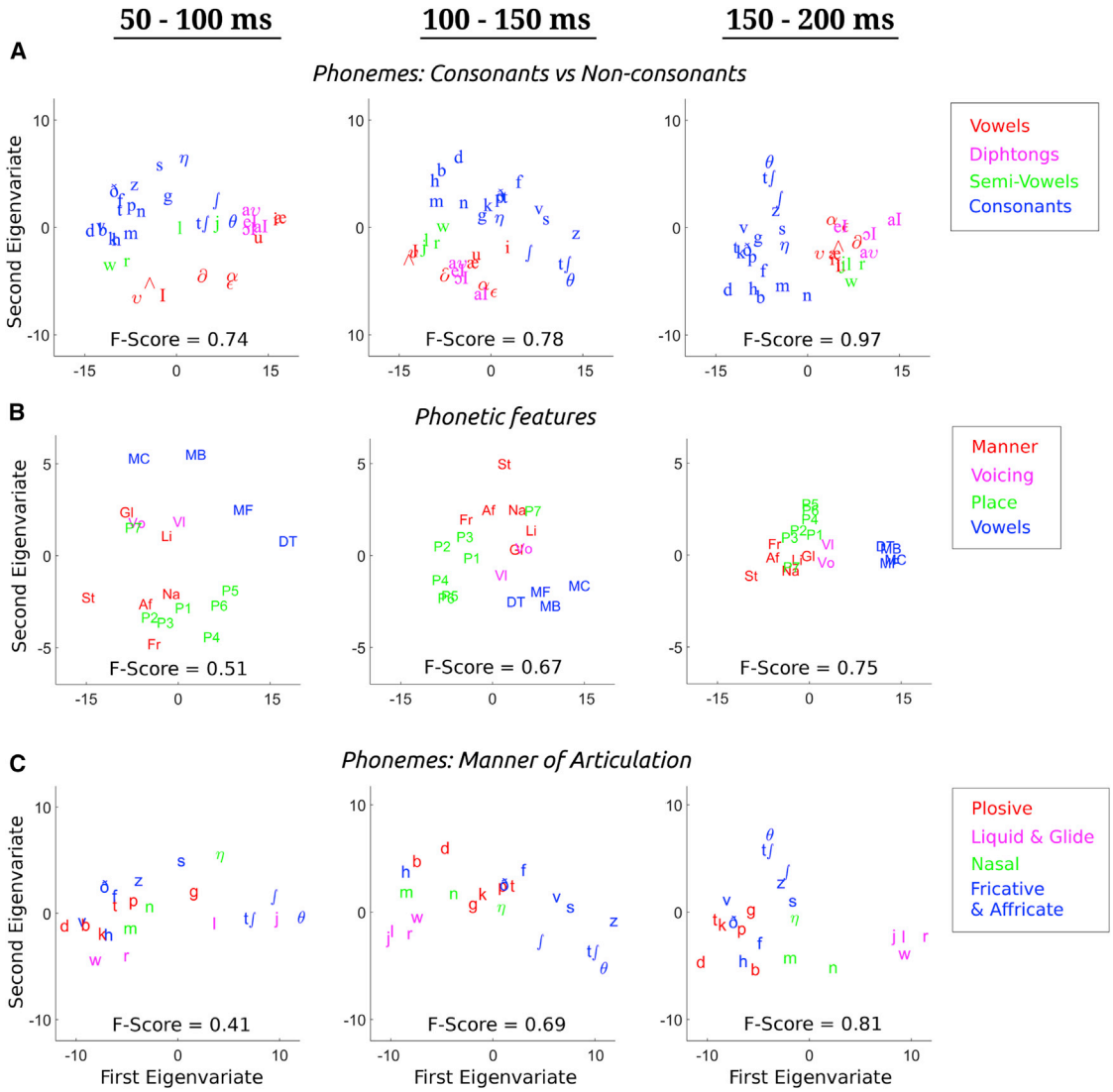

Figure 5. Sensitivity of EEG to Speech Features Increases with Response Latency

Multidimensional scaling (MDS) on the phonetic-features and phonemic mTRFs as a function of peri-stimulus time lag. By carrying out repeated k-means classification, we derive F-score measures that represent the discriminability of our mTRFs in each of the three time intervals 50–100 ms, 100–150 ms, and 150–200 ms, which correspond approximately to the three main peaks and troughs of the phonemic mTRF (Figure 3C).

(A) MDS on the phonetic-features mTRF. The F-scores indicate the differential sensitivity of responses to manner of articulation, voicing, backness of a vowel, and place of articulation features. These F-scores show significant increase with response latency.

(B) MDS on the phoneme mTRF. The F-scores are a measure of the binary classification of responses to consonants and non-consonants (vowels, diphthongs, and semi-vowels), which, again, significantly increase with latency. Although the classification performed was binary, the distinction between the four main classes of phonemes is evident.

(C) MDS on the phoneme mTRF. Here the F-scores are computed for the four classes: plosive, liquid & glide, nasal, and fricative & affricate. The four categories are progressively more separable across the three time intervals (jackknife method, $p < 0.0005$).

one-way repeated-measures ANOVA with a Greenhouse-Geisser correction if the assumption of sphericity was not met.

The increasing F-scores as a function of latency for the *Ph* mTRF show that the responses become more discriminative between consonants and non-consonants at longer latencies ($F(2.0, 18.0) = 3 \times 10^9$, $p < 0.0005$; Figure 5A). Similarly, the F-scores for the *Fea* mTRF show that the responses become more sensitive to different groups of phonetic features as a function of latency ($F(1.3, 11.4) = 10^5$, $p < 0.0005$; Figure 5B). Again, it can be seen that responses to vowels are clearly separable from those to consonant-related features at longer latencies. Analysis within each phonetic-feature group revealed no sensitivity in our mTRFs for place of articulation, for voicing, or for different vowels (data not shown). However, the mTRFs did discriminate manner of articulation (Figure 5C), especially at longer latencies ($F(2.0, 18.0) = 215.0$, $p < 0.0005$). These results show that noninvasive neural responses to speech are sensitive to specific phonetic features and that this sensitivity increases as a function of latency.

The lack of response sensitivity to different specific vowels above, combined with the high degree of discriminability between vowels and consonants, caused us to wonder whether our model performance (Figure 2A) was mostly driven by this between-class response sensitivity. We tested this by randomly relabeling the consonants in our time-aligned phoneme model (*Ph*) with other consonants and by relabeling the vowels with randomly chosen vowels. This led to a marked drop in EEG prediction performance (mean ± SD, $r = 0.0247 \pm 0.0009$, shuffled over 50 randomly relabeled versions of the stimulus, compared with 0.0635 for the correct *Ph* labeling). This suggests that while the neural responses strongly discriminate between vowels and consonants, the data are also sensitive to differences within these two classes.

Finally, we repeated the above analyses for the time-reversed speech (Figure S4). In this case, consonants and non-consonants could still be discriminated in the *Ph* mTRF ($F(1.1, 20.3) = 42.9$, $p < 0.0005$). In addition, phonetic features ($F(1.3, 23.1) = 148.0$, $p < 0.0005$) and manner of articulation ($F(2.0, 36.0) = 147.7$, $p < 0.0005$) could also be discriminated. However, importantly, unlike for forward speech, there was no significant relationship between discriminability and latency for either phonetic features ($F(1.3, 11.6) = 0.1$, $p = 0.79$) or manner of articulation ($F(2.0, 18.0) = 0.46$, $p = 0.64$), and discriminability for consonants and non-consonants did not monotonically increase with latency.

DISCUSSION

For humans to successfully process natural speech, they must parse complex and variable acoustic inputs into categorical units and correctly encode those units as particular phonemes [4]. Here, in the context of natural speech, we have shown that low-frequency, noninvasively recorded EEG indexes this categorical phoneme-level processing. Furthermore, we have shown that the articulatory features of speech can best be discriminated by responses at longer latencies, in line with what one might expect of a hierarchical system.

Our findings have important implications for current theories on cortical entrainment to the envelope of speech [11, 14, 19, 27]. In particular, we have shown that the processing of different speech features that covary with the envelope can be dissociated according to the neural responses they elicit. Therefore, neural measures based on the envelope alone are likely to include contributions from neural populations at different levels of the speech processing hierarchy. Given the relatively modest difference in modeling performance between our *FS* and *Sgram* models, it is entirely possible that the speech-specific contribution to measures of cortical entrainment is relatively small in comparison to the more general response to the stimulus acoustics. One brain region that could be responsible for such a contribution is the superior temporal sulcus (STS). It has been suggested that STS is involved in phonological-level processing bilaterally [28], a finding that fits with the lack of any lateralization effects in our prediction performances (Figures S1 and S2). While this region has been implicated in many other cognitive domains [29], recent neuroimaging work has suggested that it may represent a special locus of speech analysis that is distinct from lexical, semantic, or syntactic processes [30]. The notion that speech-specific effects in EEG may derive from a relatively small contribution from a specific brain region such as STS would partly explain why it has been so difficult to definitively say whether envelope entrainment measures reflect anything more than low-level processing of the acoustics of speech [19].

Recently it has been suggested that there may be different functional roles for entrainment at different frequencies, with theta-band entrainment (4–8 Hz) encoding speech features critical for intelligibility and delta-band entrainment (1–4 Hz) being related to the perceived, non-speech-specific acoustic rhythm [19]. Our finding that *FS* outperforms all other models for delta and theta bands (Figure 3) suggests that both of these bands may reflect speech-specific processing. One attempt to reconcile these views is to suggest that relying on envelope tracking as a dependent measure, particularly for delta band where it performs poorly, results in a lower sensitivity to subtle speech-specific effects. While we have argued that these speech-specific effects can be seen by comparing *FS* and *Sgram* performances, it is also worth noting how the *Sgram* and *Ph* model performances differentially vary across frequency bands (e.g., compare the relative model performances for delta and alpha). This variation potentially provides another way to disambiguate lower- and higher-level speech processing effects, something we will investigate in future work.

Our findings align well with recent invasive ECoG research investigating the encoding of natural speech in the human brain [4, 8]. Specifically, based on recordings from the superior temporal gyrus (STG) in epilepsy patients, high gamma frequency (75–150 Hz) activity was shown to encode an acoustic-phonetic representation of speech. Based on this, it has been suggested that the STG may be a transitional stage in the auditory processing hierarchy, early enough to still encode the acoustic features of speech but high enough to exhibit response selectivity to complex spectrotemporal patterns [31]. The fact that the ECoG recordings were shown to be optimally sensitive to intermediate acoustic-phonetic speech features at an intermediate response time lag of around 150 ms [8] agrees reasonably well with the increased discriminative power of our EEG responses at this latency. While we have speculated that our findings may have specific contributions from STS, the concordance with ECoG from STG suggests that the analysis framework we have outlined may represent an important mechanism for applying findings from the ECoG community into research with a wider variety of subjects including infants [32], children with developmental difficulties [33], the elderly [34], and patients with psychiatric disorders [35]. This is particularly important because much of the EEG/MEG research in these cohorts relies on stimuli composed of discrete syllables, leading to a literature that is limited in what it can say about the parsing and processing of continuous speech (e.g., [32]).

Developing further insights using our approach would benefit from an ability to disentangle the activity from the many neural sources that are concurrently active during speech processing. Although this issue is often seen as a shortcoming of EEG and MEG, it can also be seen as a strength in terms of the global view of hierarchical processing that these methods provide. But it will still be necessary to further characterize how different speech representations map to different neural responses and to determine which specific neural populations are responsible for

those responses. Furthermore, it will be necessary to disentangle how much the cortical entrainment of speech is driven by additive evoked activity and how much by the entrainment of ongoing oscillations [9, 27, 36]. One potentially fruitful approach to address these questions is to manipulate the relative amount of low- and high-level information that is available in the speech stimuli, with a view toward disambiguating the information contained within our *Sgram* and *Ph* models (e.g., [37]). Indeed, this is already possible to an extent by considering the difference between the *FS* and *Sgram* model performances, which we contend is likely to reflect phoneme-level processing in relative isolation. Importantly, this difference was positive for each and every subject. As such, it has the potential to act as a dependent measure in research aimed at understanding speech processing in particular populations. The sensitivity of response functions to different phonetic features and how that sensitivity varies with latency also represent potentially useful dependent measures of speech-specific processing.

EXPERIMENTAL PROCEDURES

All experimental procedures were approved by the Ethics Committee of the School of Psychology at Trinity College Dublin. In the first experiment, ten subjects undertook 28 trials, each of ~155 s in length, in which they were presented with an audiobook version of a classic work of fiction read by a male American English speaker. The second experiment involved the presentation of the same trials in the same order, but with each of the 28 speech stimuli played in reverse. All stimuli were presented monophonically using headphones in a dark room while subjects fixated on a crosshair centered on a screen. High-density EEG was recorded and digitally filtered between 1 and 15 Hz and referenced to the average of the two mastoid channels. Linear regression was used to determine multivariate temporal response functions (mTRFs) describing a mapping between the EEG and five speech representations. The broadband amplitude envelope representation (*Env*) was calculated using the Hilbert transform. The spectrogram representation (*Sgram*) was obtained by first filtering the speech stimulus into 16 frequency bands between 250 Hz and 8 kHz and then using the Hilbert transform to compute the amplitude envelope for each band. The phoneme representation (*Ph*) was computed using the forced-alignment software Prosodylab-Aligner [38], which returns the starting and ending time points for each phoneme. This representation was a multivariate time series composed of mutually exclusive binary arrays (one for each phoneme). The average acoustic envelope produced by this alignment can be seen in Figure S5. The phonetic-features representation was obtained through a linear mapping of the phonemic representation into a space of 19 features [8], which are a distinctive subset of those defined by Chomsky and Halle [25] to describe the articulatory and acoustic properties of the phonetic content of speech. Finally, we propose a model that combines *Fea* and *Sgram* (*FS*), which was obtained by concatenating *Fea* and *Sgram* into a single matrix.

The sensitivity of EEG to different speech features was assessed by using leave-one-out cross-validation to see how well each model could predict EEG data, as indexed by Pearson's correlation. This was done using mTRFs computed over a range of time lags from 0 to 250 ms, as no visible response was present outside this range. While this was done for all electrodes, we focused our analysis on a set of 12 electrodes from the two areas of the scalp with the highest prediction correlations (six symmetric pairs on the left and right scalp; see Supplemental Experimental Procedures for more details), without biasing any of the mTRF models. This subset of electrodes was used to obtain the prediction correlations presented in Figures 2 and 3. The average of the mTRFs across these 12 electrodes is presented in Figures 4 and S3. To examine the dissimilarity of neural responses to different speech tokens, a multi-dimensional scaling (MDS) analysis was applied to the phonemic and phonetic-features mTRF models on all 12 of these channels collectively (i.e., without averaging; Figures 5 and S4). Similar to previous research [4], we employed a non-metric MDS that minimizes the reconstruction error measured by Kruskal stress [39]. The MDS was calculated in five dimensions (eigenvariates), which was enough to allow the reconstruction of the original dissimilarities with an accuracy in excess of 90% in all cases (stress $\leq$ 0.1) [39]. 100 repetitions of k-means unsupervised classification with prior knowledge of the number of classes k [40] was performed to classify phonemes and phonetic features. For each repetition, this was performed based on the phonemic or phonetic-features mTRF model at all electrodes of interest for every subject. Sensitivity to different phonemes and features was quantified by calculating F-scores on the output of the k-means clustering [41]. The values reported in Figures 5 and S4 are the averages of these repetitions.

A more detailed description of the subjects, data collection, and analysis can be found in the Supplemental Information.

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2015.08.030.

AUTHOR CONTRIBUTIONS

The study was conceived and the experiments were designed by E.C.L. G.M.D.L. programmed the tasks and collected the data. G.M.D.L. and J.A.O. analyzed the data. E.C.L., G.M.D.L., and J.A.O. wrote the manuscript.

ACKNOWLEDGMENTS

This study was supported by a grant from Science Foundation Ireland (09-RFP-NES2382) and by an Irish Research Council Government of Ireland Postgraduate Scholarship. The authors thank S. Shamma, N. Mesgarani, J. Butler, S. Kelly, J. Murphy, M. Crosse, and G. Loughnane for useful discussions and comments on the manuscript.

Received: May 14, 2015
Revised: July 27, 2015
Accepted: August 13, 2015
Published: September 24, 2015

REFERENCES

1. Poeppel, D. (2014). The neuroanatomic and neurophysiological infrastructure for speech and language. Curr. Opin. Neurobiol. *28*, 142–149.
2. Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.H., Saberi, K., Serences, J.T., and Hickok, G. (2010). Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cereb. Cortex *20*, 2486–2495.
3. Peelle, J.E., Johnsrude, I.S., and Davis, M.H. (2010). Hierarchical processing for speech in human auditory cortex and beyond. Front. Hum. Neurosci. *4*, 51.
4. Chang, E.F., Rieger, J.W., Johnson, K., Berger, M.S., Barbaro, N.M., and Knight, R.T. (2010). Categorical speech representation in human superior temporal gyrus. Nat. Neurosci. *13*, 1428–1432.
5. DeWitt, I., and Rauschecker, J.P. (2012). Phoneme and word recognition in the auditory ventral stream. Proc. Natl. Acad. Sci. USA *109*, E505–E514.
6. Obleser, J., Zimmermann, J., Van Meter, J., and Rauschecker, J.P. (2007). Multiple stages of auditory speech perception reflected in event-related FMRI. Cereb. Cortex *17*, 2251–2257.
7. Rauschecker, J.P., and Scott, S.K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat. Neurosci. *12*, 718–724.
8. Mesgarani, N., Cheung, C., Johnson, K., and Chang, E.F. (2014). Phonetic feature encoding in human superior temporal gyrus. Science *343*, 1006–1010.
9. Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon,

J.Z., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron *77*, 980–991.

10. Salmelin, R. (2007). Clinical neurophysiology of language: the MEG approach. Clin. Neurophysiol. *118*, 237–254.
11. Aiken, S.J., and Picton, T.W. (2008). Human cortical responses to the speech envelope. Ear Hear. *29*, 139–157.
12. Lalor, E.C., and Foxe, J.J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur. J. Neurosci. *31*, 189–193.
13. Ding, N., and Simon, J.Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. USA *109*, 11854–11859.
14. Luo, H., and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron *54*, 1001–1010.
15. Power, A.J., Foxe, J.J., Forde, E.J., Reilly, R.B., and Lalor, E.C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. Eur. J. Neurosci. *35*, 1497–1503.
16. Kerlin, J.R., Shahin, A.J., and Miller, L.M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". J. Neurosci. *30*, 620–628.
17. Zion Golumbic, E., Cogan, G.B., Schroeder, C.E., and Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". J. Neurosci. *33*, 1417–1426.
18. Luo, H., Liu, Z., and Poeppel, D. (2010). Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. PLoS Biol. *8*, e1000445.
19. Ding, N., and Simon, J.Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. Front. Hum. Neurosci. *8*, 311.
20. Peelle, J.E., Gross, J., and Davis, M.H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb. Cortex *23*, 1378–1387.
21. Howard, M.F., and Poeppel, D. (2010). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. J. Neurophysiol. *104*, 2500–2511.
22. Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., and Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biol. *11*, e1001752.
23. Ding, N., and Simon, J.Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J. Neurophysiol. *107*, 78–89.
24. Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N.E., Knight, R.T., and Chang, E.F. (2012). Reconstructing speech from human auditory cortex. PLoS Biol. *10*, e1001251.
25. Chomsky, N., and Halle, M. (1968). The Sound Pattern of English (Harper & Row).
26. Ding, N., Chatterjee, M., and Simon, J.Z. (2013). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. Neuroimage *88C*, 41–46.
27. Giraud, A.L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. *15*, 511–517.
28. Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nat. Rev. Neurosci. *8*, 393–402.
29. Hein, G., and Knight, R.T. (2008). Superior temporal sulcus–It's my area: or is it? J. Cogn. Neurosci. *20*, 2125–2136.
30. Overath, T., McDermott, J.H., Zarate, J.M., and Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat. Neurosci. *18*, 903–911.
31. Shamma, S. (2014). How phonetically selective is the human auditory cortex? Trends Cogn. Sci. *18*, 391–392.
32. Kuhl, P.K. (2010). Brain mechanisms in early language acquisition. Neuron *67*, 713–727.
33. Gervais, H., Belin, P., Boddaert, N., Leboyer, M., Coez, A., Sfaello, I., Barthélémy, C., Brunelle, F., Samson, Y., and Zilbovicius, M. (2004). Abnormal cortical voice processing in autism. Nat. Neurosci. *7*, 801–802.
34. Ruggles, D., Bharadwaj, H., and Shinn-Cunningham, B.G. (2012). Why middle-aged listeners have trouble hearing in everyday settings. Curr. Biol. *22*, 1417–1422.
35. Li, X., Branch, C.A., and DeLisi, L.E. (2009). Language pathway abnormalities in schizophrenia: a review of fMRI and other imaging studies. Curr. Opin. Psychiatry *22*, 131–139.
36. Schroeder, C.E., and Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci. *32*, 9–18.
37. Zoefel, B., and VanRullen, R. (2015). Selective perceptual phase entrainment to speech rhythm in the absence of spectral energy fluctuations. J. Neurosci. *35*, 1954–1964.
38. Gorman, K., Howell, J., and Wagner, M. (2011). Prosodylab-Aligner: A tool for forced alignment of laboratory speech. Can. Acoust. *39*, 192–193.
39. Kruskal, J.B., and Wish, M. (1978). Multidimensional Scaling (Sage Publications).
40. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, *Volume I: Theory of Statistics*, L. Lecam, and J. Neyman, eds. (University of California Press), pp. 281–297.
41. Rijsbergen, C.J.V. (1979). Information Retrieval (Butterworth-Heinemann).

Current Biology

Supplemental Information

Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing

Giovanni M. Di Liberto, James A. O'Sullivan, and Edmund C. Lalor

Supplemental Figures

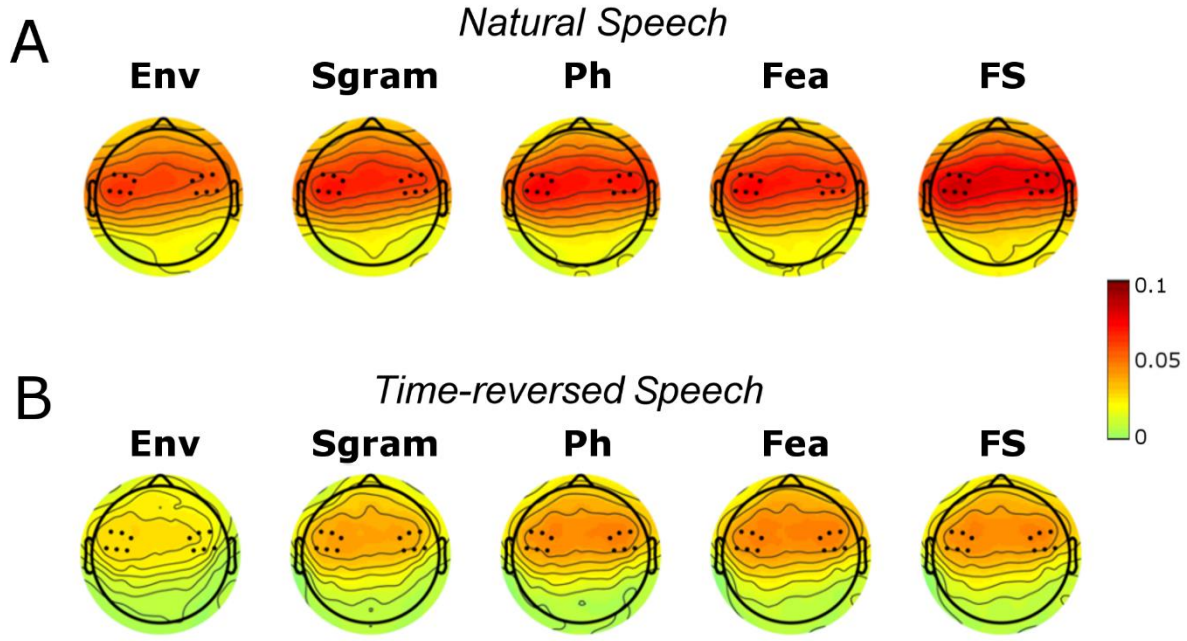

Figure S1. Topographical distribution of the EEG prediction accuracies (related to Figure 2, 3, 4, 5). The topographical distributions of the prediction accuracies are shown for the 5 models (*Env*, *Sgram*, *Ph*, *Fea*, and *FS*) and for the 2 conditions: natural **(a)** and time-reversed **(b)** speech. A nonparametric test was performed to test for differences between these topographies (T-ANOVA). No significant differences were found ($P > 0.05$). The electrodes that were chosen for all analyses are represented by black dots.

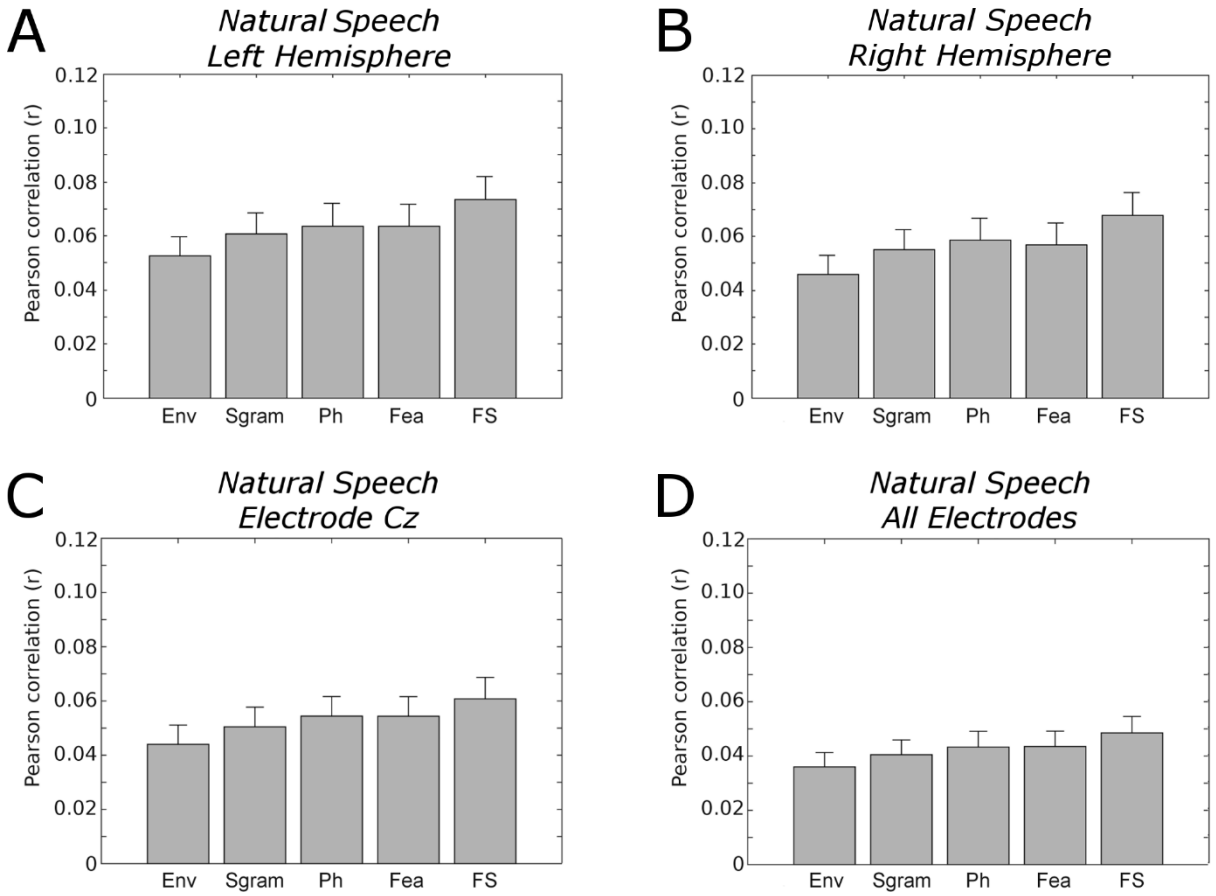

Figure S2. Model performance at different scalp locations (related to Figure 2). Grand-average EEG prediction correlations (Pearson's *r*) for each speech model (mean ± SEM) on **(a)** 6 electrodes over left scalp, **(b)** 6 electrodes over right scalp, **(c)** midline electrode Cz, and **(d)** all 128 electrodes. It can be seen that for left, right and midline Cz electrodes, the relative performance of all models is qualitatively similar. Average performance across all 128 electrodes **(d)** is noticeably lower given that it includes efforts to predict EEG on channels that do not strongly represent auditory cortical activity (e.g., occipital channels), especially when referenced to the mastoids as we have done.

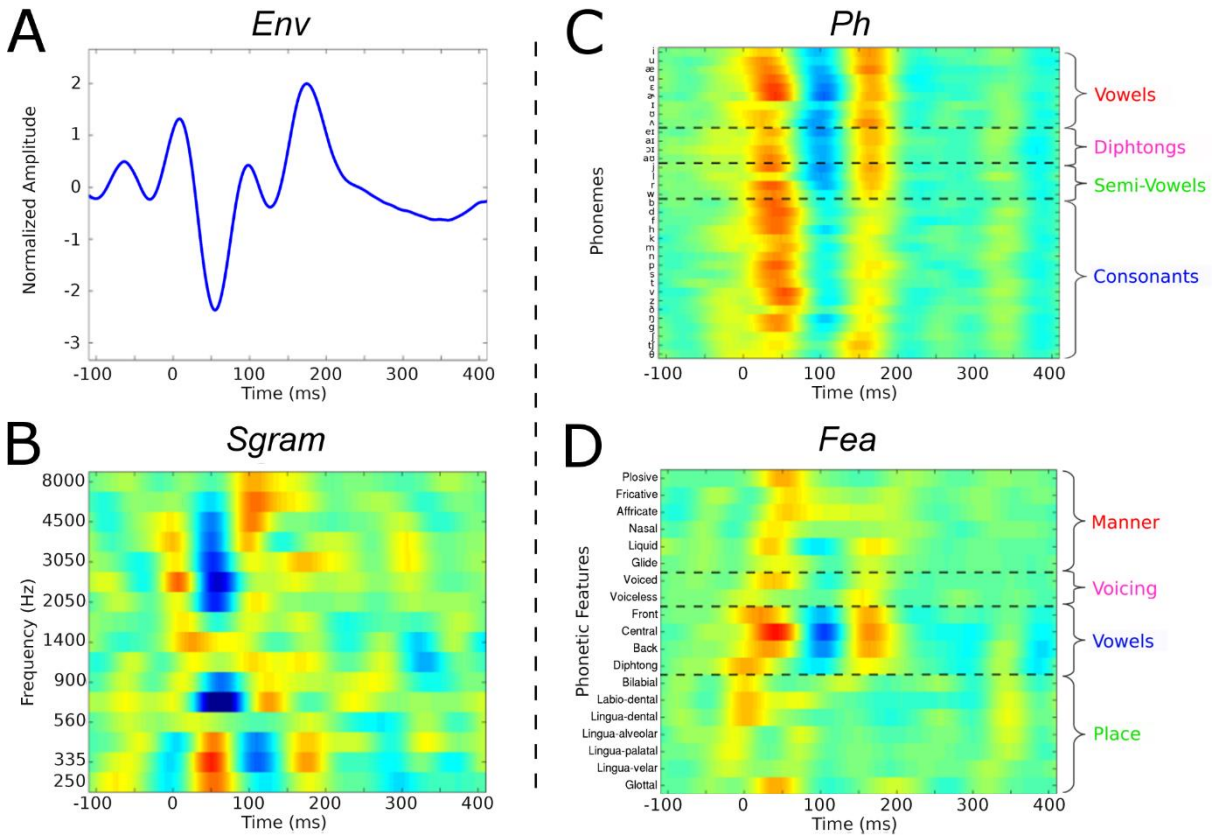

Figure S3. mTRF models for time-reversed speech (related to Figure 4). mTRFs plotted for **(a)** envelope, **(b)** spectrogram, **(c)** phoneme and **(d)** phonetic feature models at peri-stimulus time-lags from -100 to 400 ms for time-reversed speech. The phonemes were sorted based on a hierarchical clustering analysis on the average TRF after grouping them into vowels, diphthongs, semi-vowels, and consonants. Horizontal dashed lines separated distinct categories of phonemes and phonetic-features.

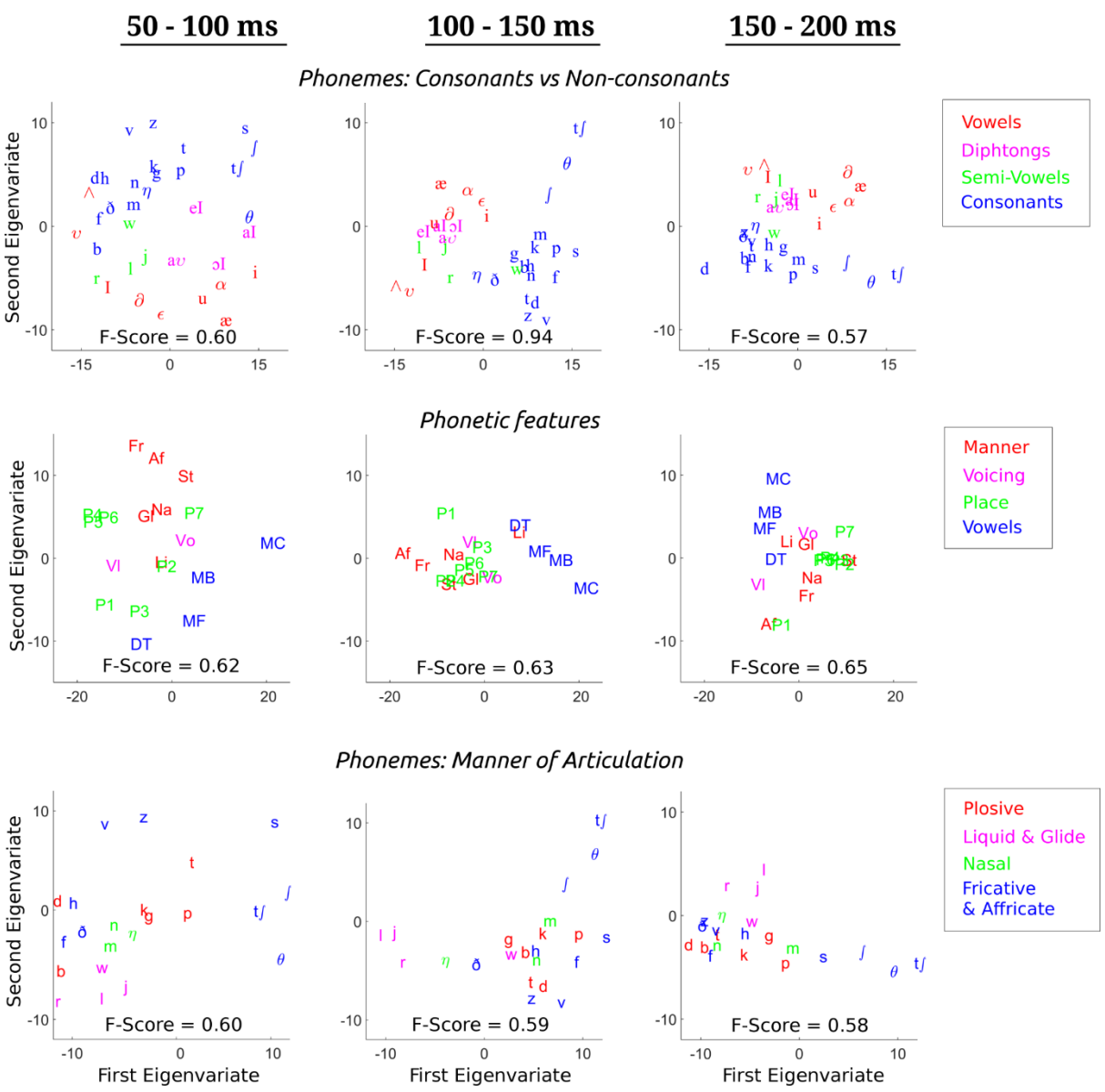

Figure S4. Discriminability of speech features in EEG for time-reversed speech (related to Figure 5). Multidimensional scaling (MDS) on the phonetic-features and phoneme mTRF as a function of peri-stimulus time-lag. *F*-Scores measures are derived with the same procedure used in Figure 5. **(a)** MDS on the phonetic-features mTRF. The *F*-scores indicate the differential sensitivity of responses to manner of articulation, voicing, backness of a vowel, and place of articulation features. These *F*-scores show no significant increase with response latency. **(b)** MDS on the phoneme mTRF. The *F*-scores are a measure of the binary classification of responses to consonants and non-consonants (vowels, diphthongs, and semi-vowels). As in **Figure 5**, the figures show a distinction between the four main classes of phonemes. However, the same monotonic increase with latency is not seen. **(c)** MDS on the phoneme mTRF. Here, the *F*-scores are computed for the four classes: plosive; liquid and glide; nasal; fricative and affricate. There is no significant difference in *F*-score with latency.

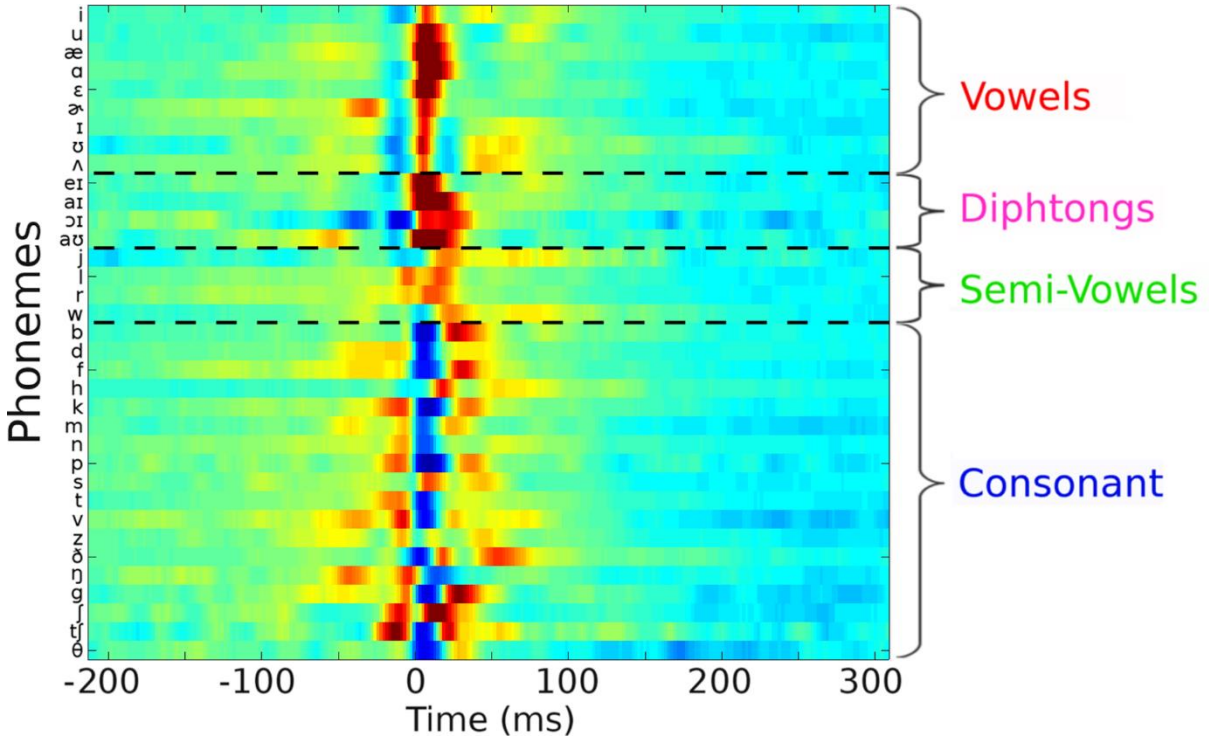

Figure S5. Average acoustic envelope of phonemes (related to Figure 4). The average temporal profile of the envelope of each phoneme aligned using the forced-alignment software [38]. While some variance is evident in phoneme onset time across different phonemes, the key feature of our *Ph* and *Fea* models is to ignore *within* phoneme variance across different utterances of that phoneme by labeling those events categorically.

Supplemental Experimental Procedures

Subjects

This study consisted of two experiments. Ten healthy subjects (7 male) aged between 23 and 38 years old participated in the first experiment, and ten subjects (7 male) aged between 21 and 32 years old participated in the second (5 subjects participated in both experiments). The study was undertaken in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the School of Psychology at Trinity College Dublin. Each subject provided written informed consent. Subjects reported no history of hearing impairment or neurological disorder.

Stimuli and Experimental Procedure

In the first experiment, subjects undertook 28 trials, each of the same length (just under 155 seconds), where they were presented with a professional audio-book version of a popular mid-20th century American work of fiction written in an economical and understated style and read by a single male American speaker. The trials preserved the storyline, with neither repetitions nor discontinuities. The average speech rate was ~ 210 words/min. Similarly, the second experiment involved the presentation of the same trials in the same order, but with each of the 28 speech segments played in reverse. All stimuli were presented monophonically at a sampling rate of 44,100 Hz using Sennheiser HD650 headphones and Presentation software from Neurobehavioral Systems (http://www.neurobs.com). Testing was carried out in a dark room and subjects were instructed to maintain visual fixation for the duration of each trial on a crosshair centered on the screen, and to minimize eye blinking and all other motor activities.

Data Acquisition and Preprocessing

Electroencephalographic (EEG) data were recorded from 128 scalp electrodes for 18 subjects, and 160 scalp electrodes for 2 subjects. The data acquired using the 160-electrode system were mapped to the same 128 electrode positions used for all other subjects using a spline interpolation algorithm

(EEGLAB; [S1]) resulting in a coherent dataset with identical channel configuration for all subjects. Data were filtered over the range 0 - 134 Hz, and digitized with a sampling frequency of 512 Hz using a BioSemi Active Two system. Data were analyzed offline using MATLAB 2014 software (The Mathworks Inc.). EEG data were digitally filtered between 1 and 15 Hz using a Chebyshev Type 2 filter in both a forwards and backwards direction to remove phase-distortion. In order to reduce the processing time required, all EEG data were then down-sampled to 128 Hz. Excessively noisy EEG channels were rejected based on several criteria [S2], and the data on these channels were estimated using spherical spline interpolation (EEGLAB; [S1]). Independent component analysis (ICA) was performed independently for each subject using the Infomax algorithm [S3]. Components constituting artefacts were removed via visual inspection of their topographical distribution and frequency content and the remaining components were back-projected to EEG electrode space. All channels were then referenced to the average of the two mastoid channels.

TRF Computation

The method used here to analyze the mapping between the various speech representations and the recorded EEG data is commonly known as a temporal response function (TRF). A TRF can be interpreted as a filter that describes the brain's linear transformation of a stimulus feature, *S(t),* to the continuous neural response *R(t)*, i.e.,

$$R(t) = TRF * S(t)$$

where * represents the convolution operator. The TRFs were calculated by performing regularized linear regression between our stimulus variables and our EEG. Specifically we perform ridge regression wherein a parameter (lambda) is set to control overfitting (see [S4] for a detailed description of this step). Given that the stimulus here is often represented as a multivariate feature vector, we refer to our TRFs as multivariate TRFs (mTRFs). mTRFs were calculated using custom written, publicly available software (http://www.mee.tcd.ie/lalorlab/resources.html; v1.1).

In previous work, we have attempt to cast our TRF functions with μV as their unit of measure [S5, S6]. However, this relies on a decision to normalize the input stimulus values between some limits and, as such, has been somewhat arbitrary. In addition, in the present work, the mTRFs are multivariate which further complicates the issue of precise units. For these reasons, and in line with previous work from other groups (e.g., [13]), the mTRFs are presented here in arbitrary units. The colors in the mTRF plots can be interpreted as follows: red at a particular latency indicates that, at that poststimulus lag, the EEG voltage is driven in a positive direction by the presentation of that particular stimulus (e.g., phoneme or frequency). And blue means the EEG voltage at that poststimulus lag is driven negative by that stimulus. Thus, given the same normalization strategy for forward and time-reversed speech, the mTRF responses can be compared in terms of their amplitudes, despite their description in terms of arbitrary units.

Speech Representations

We estimated mTRFs based on five distinct representations of the speech stimulus:

1. Broadband amplitude envelope **(*Env*)**: This was calculated as:

$$Env = |x_a(t)|, \quad x_a(t) = x(t) + j\hat{x}(t),$$

where $x_a(t)$is the complex analytic signal obtained by the sum of the original speech $x(t)$ and its Hilbert transform $\hat{x}(t)$. *Env* was defined as the absolute value of $x_a(t)$. This was then downsampled to the same sampling frequency as the EEG data, after applying a zero-phase shift anti-aliasing filter.

2. Spectrogram **(*Sgram*)**: This was obtained by first filtering the speech stimulus into 16 frequency bands between 250 Hz and 8 kHz according to Greenwood's equation [S7], and then computing the amplitude envelope (as above) for each frequency band.

3. Phonemes **(*Ph*)**: This representation was computed using the *Prosodylab-Aligner* [38] which, given a speech file and the corresponding textual orthographical transcription, automatically partitions each word into phonemes from the American English International Phonetic Alphabet

(IPA) and performs forced-alignment [S8], returning the starting and ending time-points for each phoneme. This information was then converted into a multivariate time-series composed of indicator variables, which are binary arrays (one for each phoneme). These are active for the time-points in which phonemes occurred. The phonemes are mutually exclusive, so that only one can be active at each sample point. We selected a subset of the IPA composed of the 35 most frequent phonemes in the presented speech stimuli (3 of 38 IPA phonemes were excluded as being outliers in terms of how rare they were). *Ph* is a language dependent representation of speech.

4. Phonetic features **(*Fea*):** This representation was obtained through a linear mapping of the phonemic representation into a space of 19 features (based on the University of Iowa's phonetics project http://www.uiowa.edu/~acadtech/phonetics/english/english.html/) and using an approach similar to [8]. This set of phonetic features were a distinctive subset of those defined by Chomsky and Halle [S9] to describe the articulatory and acoustic properties of the phonetic content of speech. In particular, the chosen features are related to the manner of articulation, to the voicing of a consonant, to the backness of a vowel, and to the place of articulation. Each phoneme consists of a combination of distinct features; therefore this is a set of non-mutually exclusive descriptors. *Fea* is a language independent representation of speech.

5. Finally, we propose a model that combines *Fea* and *Sgram* (***FS***): This was obtained by concatenating *Fea* and *Sgram* into a single data matrix. This representation consists of 19 phonetic features and 16 frequency bands; therefore *FS* has 35 dimensions. The rationale for combining these particular two representations was that the better performance of the *Sgram* model relative to the *Env* model suggested it as a more optimal way to capture processing of the low-level acoustics. Choosing between *Ph* and *Fea* was simply done for efficiency given that *Fea* is essentially a more concise representation of the same information as that contained in *Ph*.

Model Evaluation

We wished to compare how each speech representation mapped to the EEG. To do this, we used a leave-one-out cross-validation approach, whereby, for each representation, an mTRF was trained on 27 trials, and used to predict the EEG data from the remaining trial. This process was repeated until the data from all trials were predicted. Prediction accuracies were evaluated by determining a correlation coefficient (Pearson's *r*) between the actual and predicted EEG data on each electrode channel. Note that silent time intervals were removed from the correlation evaluation (the same intervals were removed from all speech representations). When analyzing the characteristics of the mTRFs (**Figure 4** and **Figure S3**), the results were obtained using the whole dataset, with no division between training and testing data.

Electrode Selection

The model evaluation procedure revealed a specific distribution of EEG prediction correlations across the scalp. Importantly, there was no statistical difference in the distribution of these predictions between the 5 models (**Figure S1**, *Env*, *Sgram*, *Ph*, *Fea*, and *FS*; $P > 0.05$ T-ANOVA [S10]). A set of 12 electrodes from the 2 areas of the scalp with high prediction correlations were selected (6 on the left side of the scalp, and their symmetrical counterparts on the right), without biasing any of the mTRF models. Each of the selected electrodes was among the 12 electrodes with highest prediction correlations for over 90% of the cross-validation steps for every model. This subset of electrodes was used to obtain the prediction correlations presented in **Figure 2**. The average of the mTRFs across these 12 electrodes is presented in **Figure 4** and **Figure S3**. The MDS analysis was conducted on the mTRFs obtained from all 12 channels collectively (i.e., without averaging).

Time-lag Selection

The presented mTRFs were first computed on a broad time-window from -150 to 450 ms. Based on visual inspection, this time interval was then restricted to lags from 0 to 250 ms as no visible response was present outside this range. This was confirmed by a quantitative search using the maximization of the EEG prediction correlation as the objective function.

Terms and Definitions

Low- and high-level features of speech are not well-defined in the literature [34]. In this study, we refer to *Env* and *Sgram* as low-level representations of speech, as they describe physical variations of the waveform that reach the cochlea. In contrast, we refer to *Ph* and *Fea* as high-level representations, as they describe categorical properties that can be obtained from the speech only after several neural computations.

Multi-Dimensional Scaling

Figure 5 and **Figure S4** display the results of a multi-dimensional scaling (MDS) analysis applied to the phonemic and phonetic-features mTRF models (i.e., on the multidimensional array of weights produced by the linear regression). Given a set of 'objects', MDS is an analytic vehicle which transforms each object into a point in a multi-dimensional space. Importantly, the distances between the objects reproduce an empirical matrix of dissimilarities $\boldsymbol{D_i}$. In our case, the objects are phonemes (or phonetic-features) and the dissimilarities are standardized Euclidean distances between their neural responses. In particular, a phoneme object is composed of the linear regression weights at all of the 12 EEG channels of interest for each subject. As such, it incorporates spatial information across channels as well as temporal information across the mTRF allowing insights into the spatiotemporal activation differences between different phoneme objects. That said, the differences we report are most likely driven by temporal information (**Figure S1**).

Previous research has suggested that MDS is a useful analysis tool for the categorical perception of phonetic stimuli [4, S11, S12], as well as other areas of research. Similar to previous

research [4], we employed a non-metric MDS that minimizes the reconstruction error measured by Kruskall Stress [39]. Finally, the MDS was calculated in 5 dimensions (eigenvariates), which was enough to allow the reconstruction of the original dissimilarities with an accuracy in excess of 90% in all cases (stress ≤ 0.1) [39].

F-Scores

In a classification task, given a set of 'objects' grouped in a meaningful set of classes, the main goal of a classifier is to predict the class to which each object belongs. The *F*-Score (or F_1-Score) is a measure of the quality of such predictions, obtained as the harmonic mean of precision and recall [41]. The *F*-Score can be used, for example, to compare distinct classification algorithms when using the same data, to compare different conditions or, as in this case, different datasets (i.e., TRFs for particular phonemes or features) when using the same classifier.

We performed 100 repetitions of the randomized classification algorithm *k*-means (unsupervised classification, with prior knowledge of the number of classes *k* [40]) to classify phonemes (and phonetic-features) and to study which classes of features are represented in the mTRFs. For each repetition, given the phonemic (or phonetic-features) mTRF model at all electrodes of interest for every subject, *k*-means returns a set of predictions for which an *F*-Score is evaluated. The values reported in **Figure 5** and **Supplemental Figure S4** are the averages of these repetitions. In order to work with *k* classes with 'roughly' the same number of elements, the classification for consonants/non-consonants and for manner of articulation was performed on a revised set of classes, where the smallest classes were merged with the most similar ones.

Statistical Analyses

All statistical analyses were performed using a repeated measures ANOVA to compare distributions of Pearson correlation values across models and speech direction (forward and time-reversed speech) and to compare *F*-Score classifications across response intervals and speech direction. In the latter analysis, we used the *jackknife* method on the *F*-Scores of the 10 subjects. The values reported

use the convention $F(df, df_{error})$. Greenhouse-Geisser corrections were made if Mauchly's test of sphericity was not met. All post-hoc model comparisons were performed using Bonferroni corrected paired t-tests.

References

S1. Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of neuroscience methods *134*, 9-21.

S2. Junghöfer, M., Elbert, T., Tucker, D.M., and Rockstroh, B. (2000). Statistical control of artifacts in dense array EEG/MEG studies. Psychophysiology *37*, 523-532.

S3. Makeig, S., Jung, T.-P., Ghahremani, D., and Sejnowski, T.J. (1996). Independent component analysis of simulated ERP data. Institute for Neural Computation, University of California: technical report INC-9606.

S4. O'Sullivan, J.A., Shamma, S.A., and Lalor, E.C. (2015). Evidence for Neural Computations of Temporal Coherence in an Auditory Scene and Their Enhancement during Active Listening. J Neurosci *35*, 7256-7263.

S5. Lalor, E.C., Pearlmutter, B.A., Reilly, R.B., McDarby, G., and Foxe, J.J. (2006). The VESPA: a method for the rapid estimation of a visual evoked potential. Neuroimage *32*, 1549-1561.

S6. Lalor, E.C., Power, A.J., Reilly, R.B., and Foxe, J.J. (2009). Resolving precise temporal processing properties of the auditory system using continuous stimuli. J Neurophysiol *102*, 349-359.

S7. Greenwood, D.D. (1961). Auditory Masking and the Critical Band. The Journal of the Acoustical Society of America *33*, 484-502.

S8. Yuan, J., and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. J. Acoust. Soc. Am. *123*, 3878.

S9. Chomsky, N., and Halle, M. (1968). The sound pattern of English.

S10. Lehmann, D., and Skrandies, W. (1980). Reference-free identification of components of checkerboard-evoked multichannel potential fields. Electroencephalography and clinical neurophysiology *48*, 609-621.

S11. Shepard, R.N. (1980). Multidimensional scaling, tree-fitting, and clustering. Science *210*, 390-398.

S12. Iverson, P., and Kuhl, P.K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. J. Acoust. Soc. Am. *97*, 553-562.