Published in final edited form as:
Nat Hum Behav. 2023 May ; 7(5): 740–753. doi:10.1038/s41562-023-01520-0.

Joint, distributed and hierarchically organized encoding of linguistic features in the human auditory cortex

Menoua Keshishian[1,2], **Serdar Akkol**[3], **Jose Herrero**[3,4], **Stephan Bickel**[3,4], **Ashesh D. Mehta**[3,4], **Nima Mesgarani**[1,2,✉]

[1]Department of Electrical Engineering, Columbia University, New York, NY, USA.

[2]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA.

[3]Feinstein Institutes for Medical Research, Northwell Health, Manhasset, NY, USA.

[4]Department of Neurosurgery, Hofstra-Northwell School of Medicine, Manhasset, NY, USA.

Abstract

The precise role of the human auditory cortex in representing speech sounds and transforming them to meaning is not yet fully understood. Here we used intracranial recordings from the auditory cortex of neurosurgical patients as they listened to natural speech. We found an explicit, temporally ordered and anatomically distributed neural encoding of multiple linguistic features, including phonetic, prelexical phonotactics, word frequency, and lexical–phonological and lexical–semantic information. Grouping neural sites on the basis of their encoded linguistic features revealed a hierarchical pattern, with distinct representations of prelexical and postlexical features distributed across various auditory areas. While sites with longer response latencies and greater distance from the primary auditory cortex encoded higher-level linguistic features, the encoding of lower-level features was preserved and not discarded. Our study reveals a cumulative mapping of sound to meaning and provides empirical evidence for validating neurolinguistic and psycholinguistic models of spoken word recognition that preserve the acoustic variations in speech.

Speech comprehension is the process of extracting meaning from a sound pressure waveform produced by a speaker. In each language, speech sounds can be abstracted at multiple levels of analysis, the smallest of which consists of a finite set of perceptually distinct phonetic features (for example, voicing and aspiration). Certain combinations of

Reprints and permissions information is available at www.nature.com/reprints.

✉**Correspondence and requests for materials** should be addressed to Nima Mesgarani. nima@ee.columbia.edu.

Author contributions
M.K. and N.M. designed the experiment. M.K., S.A., J.H., S.B., A.D.M. and N.M. recorded the neural data. M.K. and N.M. analysed the data and wrote the manuscript. All authors commented on the manuscript.

Reporting summary
Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Competing interests
The authors declare no competing interests.

Extended data is available for this paper at https://doi.org/10.1038/s41562-023-01520-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-023-01520-0.

these features form the phonemes—the smallest units of speech that can alter meaning (for example, /b/ in ‘bad’ versus /d/ in ‘dad’). Combinations of multiple phonemes form prelexical units that subsequently give rise to words, or the lexicon that conveys semantic meaning. The interactions between units at each level follow certain rules. For example, phonotactic probabilities describe the likelihood of certain phoneme combinations[1,2] (for example, /ba/ is more likely than /bu/ in the English language), whereas lexical–phonological probabilities describe these likelihoods in the context of word formation (for example, in the English lexicon, /ə·k·aʊ·n·t·ə·b·ɪ/ fully predicts the next phonemes of ‘accountabi.lity’)[3]. Words, in addition to their frequency of occurrence, are related to each other through phonological and semantic associations (for example, ‘cat’ is a phonetic neighbour of ‘cap’ and a semantic neighbour of ‘dog’)[4,5]. How and where this intricate mapping of speech sounds to meaning occurs in the human auditory cortex remains largely unclear.

Our understanding of the cognitive processes that extract meaning from speech has been shaped largely by psycholinguistic studies of spoken word recognition (SWR). Behavioural studies have established that the process of word recognition entails gradual integration across multiple phonemes[6–9]. The evidence suggests that spoken words are recognized relationally in the context of the mental lexicon[4]. This means that as a word is being heard, multiple candidate words are activated in parallel in proportion to their phonetic and semantic similarity and prior probability (for example, word frequency and contextual constraints), and these words then compete for recognition. Many computational models have been proposed to account for various SWR phenomena[10–12], including categorical perception of speech sounds[13], the influence of phonotactic probabilities on phoneme recognition[2], early and pre-offset identification of words[14], and the impacts of word frequency and phonological[15] and semantic[5,16–18] neighbourhoods on word recognition accuracy and speed. While these models can successfully predict many behavioural effects, they widely differ in the nature of the activation and competition mechanisms they employ[11]. This diversity can be seen particularly in the assumed intermediate representations that are used to map sounds to meaning, ranging from articulatory and acoustic–phonetic features[8,19,20], phonemes[21,22], allophones and probabilistic phonotactics[23], to distributed representations of abstract prelexical states[24,25]. This diversity underscores the insufficiency of behavioural outcomes alone in constraining and specifying the exact nature of the intermediate representations and interactions that are used as the brain makes sense of speech, which highlights the critical role of neurobiological studies of SWR.

In parallel to psycholinguistic research, neuroimaging studies have provided complementary evidence by investigating the anatomical and functional organization of language in the human brain. Researchers have hypothesized that various levels of processing occur hierarchically in the human brain[26–28]; the processing of low-level acoustic features is suggested to occur in subcortical areas and the primary auditory cortex (PAC)[29,30], whereas neural encoding of phonetic features[29,31–34] and units that require more extended temporal integration has been suggested to emerge in higher auditory areas[26,29,35–39]. This view is also supported by studies of focal brain injury, which can cause selective impairments at various linguistic levels between speech sounds and meaning. These anatomically specific selective impairments include difficulty in the identification of acoustic–phonetic cues and phonemes[40,41], lexical–phonological forms[42–44], and semantic and syntactic information[45–47]. The heterogeneity in impairment patterns raises questions regarding the specificity of processing stages across brain regions. Together with neuroimaging studies, these findings have produced a coarse anatomical map of where acoustic, phonetic and semantic processing occurs in the brain[26,48,49]. However, a more fundamental question than anatomical localization that has remained unanswered is how the various levels of linguistic features are encoded by cortical activity—that is, the precise relationship between different levels of linguistic representation and the neural responses in auditory cortical regions. Without such knowledge, the existence and the exact nature of phonological, prelexical, lexical and semantic representations in the human auditory cortex remain speculative. In particular, are the hypothesized intermediate linguistic features at different levels of abstraction explicitly encoded in the auditory cortex? Are the increasingly complex linguistic units formed successively and disjointly in different populations of neurons in the ascending auditory pathway[19,21], or are they represented jointly by the same population of neurons[24,50]? How are prelexical and postlexical linguistic features organized in the primary and non-primary auditory cortical areas?

To address these questions, we directly measured the neural responses to natural speech from the auditory cortex of neurosurgical patients undergoing invasive electrophysiological monitoring for epilepsy surgery. Using a ridge regression encoding model, we measured the encoding of five broad levels of linguistic features—phonetic, phonotactic, frequency, lexical–phonological and lexical–semantic. By characterizing neural sites on the basis of the pattern of their linguistic feature encoding, our results shed light on the nature and organization of the neural basis of speech perception at various levels of linguistic processing with implications for neurobiological and computational SWR models.

Results

We recorded intracranial electroencephalography (iEEG) data from 15 human participants implanted with subdural (electrocorticography) and depth (stereotactic EEG) electrodes (Fig. 1a and Extended Data Figs. 1 and 2). The participants listened to 30 minutes of continuous speech spoken by four speakers (two male). To ensure that the participants were engaged in the task, we paused the audio at random intervals and asked the participants to report the last sentence of the story before the pause. All participants were attentive and could correctly repeat the speech utterances.

We extracted the envelope of the high-gamma (70–150 Hz) band, which has been shown to correlate with neural firing in the proximity of the recording electrode[51,52], as the neural response measure of the recorded signals. We restricted our analyses to speech-responsive sites in the auditory cortex that had a higher response to speech than to silence (Methods). This criterion resulted in a total of 242 responsive neural sites, of which 113 were in Heschl’s gyrus (HG), 32 in the planum temporale (PT), 46 in the anterior superior temporal gyrus (aSTG) and 51 in the posterior STG (pSTG). The electrodes were split evenly between hemispheres ($N = 121$ each). Figure 1a shows the response latency of the electrodes, which shows a gradient of low to high latencies from medial to lateral auditory cortex (the mean and standard deviation of latency are 92 ± 36 ms in HG, 94 ± 22 ms in the PT, 112 ± 27 ms in the aSTG and 108 ± 48 ms in the pSTG), consistent with anatomical studies showing the primary auditory function of HG and the non-primary function of the PT and the STG[53]. We also used an anatomical measure of distance from PAC, choosing posteromedial HG (TE1.1) as a reference point for PAC (Fig. 1a)[54,55]. The relation between response latency and anatomical position was tested using a linear mixed-effects model trained to predict response latency from distance to PAC and hemisphere, with random intercepts for participants. Distance but not hemisphere had a significant effect on response latency (distance: $F_{1,239} = 27.79$; $P < 0.001$; $\beta_{\mathrm{dist}} = 1.47$ ms mm^{-1}; 95% CI, 0.92 to 2.01; hemisphere: $F_{1,239} = 0.26$; $P = 0.61$; $\beta_{\mathrm{hemi}} = -2.35$ ms/(left − right); 95% CI, −11.46 to 6.79).

Linguistic information of speech stimuli

To measure the encoding of different levels of linguistic information in the neural responses, we first needed to define and quantify these features in our speech stimuli. We chose a broad set of linguistic descriptors to represent different aspects of linguistic processing as put forward by psycholinguistic studies[4–8,56] (Fig. 1b).

Phonetic features (P): Our first level uses the smallest contrastive units of language, called phonemes. We represented each phoneme with 22 phonetic features corresponding to its distinctive features of voicing, manner of articulation and place of articulation[57,58] (Extended Data Fig. 3). This level represents the transformation and normalization of allophonic variations into a small set of perceptual categories.

Phonotactic frequency (T): For phonotactics, which represent the phoneme transition probabilities, we used the logarithm of phoneme transition frequencies (biphone frequencies), calculated from a large English corpus[59]. The frequency of a phoneme bigram represents the degree of exposure of an average native listener to that bigram and measures its probability in natural speech. We purposefully chose a non-position-specific measure of phonotactics (as opposed to the more common approach[2]) to maximally dissociate this effect from lexical processes. This level represents the expectation and the surprisal of the listener when hearing a new phoneme, based on the immediate past. This prelexical phonotactics feature could indicate predictive coding mechanisms that operate at the phonemic level[60–63].

Word frequency (F): Out of all possible phoneme sequences, a small subset forms the words of a language (lexicon). Some words are used more frequently than others, with behavioural influences that appear soon after the word onset and increase as more of the word is heard[64]. Word frequency biases the reaction time of a listener, where high-frequency words are detected faster. We quantified word frequency as the log-frequency of words calculated from a large English language corpus[59]. This level represents the listener’s context-independent expectation of hearing a particular word.

Lexical–phonological features (L): Because words in spoken language are heard gradually, each phoneme in a word conveys a varying amount of information towards the identification of the actual word[15]. Lexical–phonological features consist of lexical entropy and lexical surprisal (equivalent to cohort entropy and phoneme surprisal in ref.[63]), which are calculated for each prefix phoneme substring within a word (for example, /k/, /k·æ/ and /k·æ·t/ for the word ‘cat’). The lexical entropy for a phoneme prefix in a word is the Shannon entropy of the set of all words in the lexicon that start with that same phoneme sequence (‘onset competitors’), while the surprisal is the negative logarithm of the probability of hearing the final phoneme of the sequence given the preceding phonemes. This feature represents the gradual competition among phonologically similar words in the lexicon according to the cohort theory of SWR[65] as the target word is being heard one phoneme at a time. More specifically, entropy represents the residual uncertainty in identifying the current word depending on the remaining competitors, while surprisal represents a predictive coding process that operates at the word level. At the lexical level, the prior expectation is formed by the knowledge of the lexicon as opposed to the low-level phonemic statistics. The behavioural effects of both entropy and surprisal have been reported in prior studies[66,67].

Semantics (S): Finally, some words can have more similar meanings even though they may be very different acoustically and phonologically. We refer to this semantic relation between a word and the rest of the lexicon as lexical–semantic information. We chose semantic neighbourhood density (SND), obtained from the English Lexicon Project[56,68], to represent lexical–semantic features of words. This measure indicates the relative distance between a word and its closest semantic neighbours, obtained from a global co-occurrence model. This parameter influences behaviour during word recognition[5,56,69,70] and can represent the degree of activation of semantically related words when one hears a target word. This final level represents the next logical step in the word recognition process, which is the linking of words as acoustic objects to their conceptual meanings. Specifically, this parameter can capture the spreading of semantic activation. Notably, the behavioural studies of lexical access have shown that the semantic spreading effects may occur even before the word is fully recognized[5,56,69].

In analysing the neural data, we have distinguished between prelexical and postlexical features. We define this contrast by whether a feature is still meaningful in the absence of the lexicon or whether it can be defined only in the context of the lexicon. By this definition, phonetics and phonotactics are assumed to be prelexical, as they can be fully characterized without any knowledge of the lexicon. In contrast, word frequency defines the relative frequency of one word compared with those of the other words in the lexicon, lexical–phonological features define the competition dynamics between competing words in the lexicon and lexical semantics determines the relation between a target word and other semantically related words in the lexicon. A complete description of the features and their calculation procedures can be found in the Methods.

Linguistic encoding in neural data

Having quantified several levels of linguistic features in our speech stimuli, we examined the ability of these features to predict unseen neural data using a cross-validated ridge regression framework (Fig. 1c). The input to the regression model (predictors) was a 510 ms (51 time samples at 100 Hz) time course of all included features (acoustic and linguistic) stacked together (59 dimensions). The input was fed to the model such that the stimulus window corresponding to the time samples $[t - 50, t]$ was used to predict the neural responses at time t, making it a causal time-domain convolution. The window size was determined to maximize the cross-validated prediction accuracy of left-out data (Extended Data Fig. 4). For each electrode, we first fitted a model that contained all acoustic and linguistic features (P, T, F, L and S), computing a cross-validated prediction accuracy by fitting the model on $N-1$ trials and testing on the left-out trial, and then averaging the N correlation r values of the test predictions against the observed neural data. We then tested the predictive power of each feature one by one by replacing that feature with 100 randomly permuted control distributions (null hypotheses), refitting the regression model and computing the new average cross-validated r values, as well as the difference from the r value of the true model (Δr_c, $1 \le c \le 100$). We interpreted the magnitude of change of cross-validated prediction accuracy in the negative direction as each feature was replaced with a control (or, equivalently, the amount of increase in prediction accuracy of the true model over the control) as the degree to which that feature was encoded in the neural response[29,63,71–73]. To quantify this change, we performed a one-sample t-test on the distribution of Δr_c such that a positive t value denotes better prediction by the true model, and we chose the t statistic as the measure of encoding. This method is advantageous over an incremental model where features are added to the model one by one in a fixed order[63,71–73] due to the correlations that exist between different linguistic features (Extended Data Fig. 5). Moreover, our method does not necessitate any assumptions regarding the order, timing or hierarchy of these features.

The null condition simulated features with the same dimension, distribution and timing as the true features, except that the values were drawn from permuted linguistic distributions, such that for a given control $f_c(w_i) = F(w_j)$, where f_c is the permuted feature and F is the true feature, and w_i and w_j are two phonemes, biphones, phoneme sequences or words (depending on F). An important aspect of our control permutation tests is the consistency across tokens of the same type, where all occurrences of the same bigram, prefix or word are given the same value, albeit this value is not derived from the actual distribution in the English language. For example, the null condition for the phonotactic feature always assigned the same value to a particular phoneme sequence such as /ba/; however, this value was chosen randomly and did not correspond to the true probability of that sequence in the English language. These control conditions are stricter than the commonly used shuffling of the features[63,73] because unconstrained shuffling can artificially lower the out-of-sample prediction accuracy due to the added randomness (noise) in the regression predictors. Additionally, the speech auditory spectrogram and its half-wave rectified temporal derivative were included to account for the nonabstract acoustic information and the effect of acoustic edges in the neural response (Fig. 1c)[63,74–76].

Diversity in linguistic encoding across neural sites

Figure 1d shows six example electrodes chosen to represent the diversity in linguistic feature encoding across neural sites. The first example electrode (E1 in HG) did not show any significant improvement in prediction accuracy compared with the null model for any of the linguistic features. The second electrode (E2 in HG) showed a significant improvement only for phonetic features. The third to sixth electrodes (E3 in HG and E4–E6 in pSTG) showed significant encoding of phonetics based on different combinations of higher-level linguistic features. For example, E3 encodes phonotactics; E4, E5 and E6 encode lexical–phonological information; and E5 and E6 encode semantics. These six example electrodes exemplify the heterogeneity in linguistic feature encoding across the neural population. Motivated by this observed diversity, we then characterized the encoding patterns across the entire population of electrodes.

To uncover the patterns of linguistic encoding across all sites, we first quantified the encoding significance of each linguistic feature for each site and performed double agglomerative hierarchical clustering to group linguistic features and electrodes on the basis of the similarity of their encoding patterns. Figure 2a shows the results of the clustering analysis. Clustering the rows (linguistic feature dimension) groups different linguistic features by the similarity in their encoding patterns across the neural population. Clustering the columns (electrodes) groups neural sites by the similarity in their linguistic feature encoding. This analysis thus revealed the encoding patterns across both the feature and electrode dimensions. The correlation values between the encoding of linguistic features across the neural population (rows of the matrix in Fig. 2a) are also shown in Fig. 2b.

Clustering the linguistic features (rows) on the basis of their population encoding revealed a separate grouping of lexical features (L, F and S) from the prelexical features (P and T) (Fig. 2a, red versus black horizontal groups). This separate grouping confirmed a distinct encoding of prelexical and postlexical features across the neural population. This notion was also supported by the higher correlation between the encoding patterns of lexical features (F, L and S) across sites (Fig. 2b, red square) than the correlation between them and other prelexical features (P and T). The clustering of the features also suggested a cumulative encoding from phonetic (P) to prelexical (T) to lexical (L, F and S) representations.

Clustering the electrodes on the basis of their linguistic feature encoding also revealed a few notable patterns. First, phonetic features were encoded in most electrode sites (71.9%), in contrast to the other features (T, 21.1%; F, 30.2%; L, 22.7%; S, 13.2%). Most sites jointly encoded phonetics and other features, and there was a smaller group of sites that simultaneously encoded all linguistic features. Second, except for phonetic features, the encoding of all other features increased with response latency, as seen in the significant positive correlations shown in Fig. 2c. This positive correlation with latency indicates a progressive transformation of acoustic to linguistic features, where neural sites with higher response latencies also encoded higher levels of linguistic representation. This progression is consistent with the average distance of electrodes from PAC (Fig. 2d), where the representation of P/T to F/L to S increased as we moved further away from posteromedial HG. Third, the electrode sites that encoded prelexical phonotactics (T) and lexical–phonological (L) features only partially overlapped, which indicated that while a group of electrodes encoded both T and L, there were other electrodes that encoded only T or only L; this finding suggests that phoneme combination is a process that may start prelexically, and lexical influences on phoneme combinations may come later and in higher areas. This notion is supported by the higher correlation between L encoding and the response latencies in the neural population than between T encoding and response latencies (Fig. 2c) and by the higher average distance of sites that encode L features from PAC than the distance of sites that encode T features (Fig. 2d). The L sites were farther away from PAC[54,55] and hence probably in higher auditory cortical areas. This observed hierarchical encoding is consistent with the hypothesized linguistic processing that posits transitions from phonetic to prelexical to lexical to semantic features. Finally, projecting the linguistic feature encoding of electrode sites on a two-dimensional plane (Fig. 2e) shows a distinct encoding pattern across anatomical regions of HG, the PT and the STG, with apparent separation between the linguistic encoding patterns in HG (blue dots) and the STG (red dots) and an in-between encoding pattern for PT electrodes (yellow dots).

To further examine the encoding of linguistic features across different anatomical regions, we measured the proportion of neural sites in each region that significantly encoded a feature ($t > 19$, corresponding to the true feature being better than 97% of controls) (Fig. 2f). This analysis showed that phonetic features were encoded across all regions, yet the proportion of sites that encoded phonetic features was higher in HG and the PT. Prelexical phonotactic features (T) were encoded mostly in PT electrodes, at a level significantly more than in aSTG and pSTG. Lexical-level features (F, L and S) were not encoded in HG, but they were encoded in both the PT and pSTG electrodes. Anterior STG, on average, had a word frequency (F) encoding between HG and PT/pSTG, and no L or S encoding. Notably, the pSTG encoded only postlexical (L) but not prelexical (T) phoneme combinations. Together, these results reveal a hierarchical and distributed encoding of linguistic features where higher auditory cortical areas gradually represent higher-order linguistic information and show how features at different levels of granularity are simultaneously encoded across the auditory cortex.

Temporal dynamics of linguistic encoding

To examine the temporal characteristics of phonetic, phonotactic and lexical feature encoding in the auditory cortex, we analysed the β coefficients of the temporal response functions (TRFs) that predict the neural data from all features (the model in Fig. 1c). To find a representative TRF for each feature, we first selected the subset of electrode sites that showed a significant encoding of that feature (determined by $t > 19$—that is, the true distribution was better than 97% of the controls). We then computed the first principal component (PC) of the regression weights across recording channels to find the TRF that represented the maximum variance for the target feature (see Extended Data Fig. 6 for the explained variance of components). For features that have multiple dimensions (A1, A2 and P), we performed the principal component analysis (PCA) jointly across feature dimensions and channels. For a measure of robustness, we repeated this analysis for each feature by bootstrapping the electrode subset 1,000 times. Figure 3a shows the average and standard deviation of the first PCs computed from the TRF of each feature. Figure 3b shows the time of the TRF peak as an approximation of the processing delay using the same bootstrapping procedure. Taken together, Fig. 3 reveals a temporally ordered appearance of features. This successive temporal emergence of features from the basic acoustic representation (spectrogram) to prelexical (phono-tactics) and lexical surprisal to phonetics, frequency, and lexical–phonological and lexical–semantic representations is consistent with the suggested hierarchy of processing in the correlation results in Fig. 2b. Moreover, Fig. 3 underscores the gradual computation of linguistic features, which starts from processing the speech sound components as early as 100 ms after hearing the acoustic components until up to hundreds of milliseconds, which is needed for the accumulation of sounds and the extraction of their semantic content.

Anatomical organization of linguistic encoding

We quantified the relation between anatomical position and feature encoding using a linear mixed-effects model trained to predict the encoding of each feature (t values) from distance to PAC and hemisphere, with random intercepts for participants. Distance from PAC had a strong effect on F/L/S encoding such that farther populations are more likely to encode those features, and a weak effect for T encoding in the opposite direction (P: $F_{1,239} = 3.48$; $P = 0.063$; $\beta_{\text{dist}} = -0.65$; 95% CI, −1.33 to 0.04; T: $F_{1,239} = 6.01$; $P = 0.015$; $\beta_{\text{dist}} = -0.32$; 95% CI, −0.57 to 0.06; F: $F_{1,239} = 13.37$; $P < 0.001$; $\beta_{\text{dist}} = 0.75$; 95% CI, 0.35 to 1.16; L: $F_{1,239} = 21.22$; $P < 0.001$; $\beta_{\text{dist}} = 0.46$; 95% CI, 0.26 to 0.66; S: $F_{1,239} = 32.91$; $P < 0.001$; $\beta_{\text{dist}} = 0.47$; 95% CI, 0.31 to 0.63). In contrast, hemisphere only had a significant effect for semantic (S) encoding, where the left hemisphere was more likely to encode S (P: $F_{1,239} = 3.30$; $P = 0.071$; $\beta_{\text{hemi}} = -10.13$; 95% CI, −21.12 to 0.86; T: $F_{1,239} = 0.12$; $P = 0.73$; $\beta_{\text{hemi}} = -0.73$; 95% CI, −4.79 to 3.33; F: $F_{1,239} = 3.39$; $P = 0.067$; $\beta_{\text{hemi}} = 6.11$; 95% CI, −0.42 to 12.65; L: $F_{1,239} = 0.29$; $P = 0.59$; $\beta_{\text{hemi}} = 0.87$; 95% CI, −2.32 to 4.05; S: $F_{1,239} = 7.87$; $P = 0.005$; $\beta_{\text{hemi}} = 3.62$; 95% CI, 1.08 to 6.17).

To study the spatial organization of linguistic feature encoding in more detail, we show the distribution of t statistics for each feature across the medial–lateral and anterior–posterior axes of the auditory cortex (on the FreeSurfer average brain[77]), interpolated using k-nearest neighbours with $k = 5$ (Fig. 4). These plots show the widespread encoding of phonetics, as seen in Fig. 2a, and reveal the increasing encoding of lexical features as we move from medial HG to lateral STG. Additionally, we observe an asymmetry between the linguistic feature encoding in the left and right hemispheres, consistent with the hemisphere effect observed in the linear mixed-effects model.

Discussion

Direct neural recordings from the human auditory cortex revealed an explicit and distributed neural encoding of multiple levels of linguistic processing between the auditory stimulus and lexical semantics, meaning that these linguistic features could linearly predict the neural responses significantly better than null features. Grouping neural sites on the basis of the similarity of the linguistic features they expressed revealed distinct encoding patterns across neural sites, with contrasting representations of prelexical and postlexical features. Anatomical and functional localization of neural sites showed that the encoding of low- to high-level linguistic features appeared gradually from primary to non-primary auditory cortical areas. This anatomically distributed and temporally ordered appearance of various levels of linguistic features suggests a hierarchical processing scheme that enables the human auditory cortex to gradually transform speech sounds to decode meaning. Combining multilevel linguistic features and invasive electrophysiological recordings reveals a joint encoding of different levels of linguistic processing in relation to each other across different anatomical areas and times.

SWR models have been attempting to account for a variety of SWR phenomena[10–12], including categorical perception of speech sounds[13], the influence of phonotactic probabilities on phoneme recognition[2], early and pre-offset identification of words[14], and the impacts of word frequency and phonological and semantic neighbourhoods on word recognition accuracy and speed[5,15,18]. A major difference between these SWR models is in the assumed intermediate representations that are used in the mapping of sounds to meaning. Some models assume explicit abstractions such as articulatory and acoustic–phonetic features[8,19,20], phonemes[21,22] or allophonic patterns[23], while others assume distributed representations of abstract prelexical states[24,25]. Our results shed light on this question by showing the levels of linguistic processing that are explicitly encoded in various parts of the auditory cortex, including the acoustic manifestation of allophonic variations (spectrogram features), the ubiquitous appearance of phonetic distinctions in most recorded sites, prelexical phonotactics and postlexical features including lexical–phonological and lexical–semantic features. Moreover, while SWR models often assume a sequential change in the encoded features, our results show a joint encoding of multiple features in the same neural response, suggesting that the higher-level distinctions gradually accumulate and are jointly encoded, and the lower-level representations are not discarded. This finding is particularly important in the context of a major topic of scientific debate regarding SWR models, which is the degree of abstraction/normalization that is assumed as sounds are mapped to meaning. Some models of SWR assume discrete abstract representation of intermediate linguistic elements (for example, phonology and syllables) and word forms, treating individual variations in speech as noise[78,79]. Other models argue that such abstractions are ‘social objects’ learned from society rather than natural phenomena, that the rich acoustic representation of speech is faithfully encoded and stored (exemplars), and that word recognition entails comparing the new stimulus with the many stored exemplars. These models propose that abstractions emerge only at the retrieval stage[80–82]. This view is supported by the ability of a listener to recall and reproduce not only the linguistic content of a spoken utterance but also its paralinguistic features such as the speaker’s voice, prosody and emotional tone. More contemporary studies have argued that pure abstractionist or episodic approaches to lexical representation that sidestep any level of abstraction cannot fully explain the behavioural results. Instead, they suggest that both normalization and maintaining episodic memories of words are important parts of the process[83–85] and need not be mutually exclusive[78,86]. Our results support this view, as we found that fine-grained continuous acoustic–phonetic and low-level details of speech are not discarded as higher-order lexical representations emerge in downstream areas of the auditory cortex. Instead, general representations at higher levels of the auditory cortex (for example, the STG) can encode both acoustic variability and invariant prelexical and lexical abstractions, hence supporting the hypothesis that humans encode and store not only lexical information but also the unique attributes of each utterance, which can be used for retrieving earlier features such as prosody. It is worth noting that our observed joint encoding of low- and high-level features could be limited to the auditory cortex. Future research examining other parts of the speech cortex can further address this question.

Psycholinguistic studies of SWR have identified multiple levels of linguistic processing that directly and indirectly interact with each other. These linguistic levels include phonetics, phonotactics[87], word frequency[23,64], phonological neighbourhoods[23,65] and semantics[16–18]. Our results shed light on the neural mechanisms that underlie these behavioural findings by showing an explicit encoding of these distinct linguistic levels in the ascending auditory pathway. In particular, psycholinguistic studies have shown a ubiquitous effect for word frequency in SWR, which starts early after onset and gradually increases as a word unfolds over time[64]. Consistent with this behavioural result, our temporal analysis showed an early neural encoding of word frequency compared with the other lexical features. The TRF for mapping word frequency to the neural data started early and gradually built up after the word onset. Another notable example is the prelexical influence of probabilistic phonotactics on SWR, as opposed to the lexical influence of neighbourhood density[87]. Our results show that probabilistic phonotactics, both prelexical and lexical (in the form of lexical surprisal), were indeed encoded early, while lexical competition and neighbourhood density effects emerged considerably later in the neural responses. Phonotactics and lexical surprisal both reflect the expectedness of hearing the current phoneme given the preceding phonemes, although these expectations reflect different levels of representation: phonetic for the phonotactic feature and lexical for the lexical surprisal feature. Lexical entropy, in contrast, reflects the degree of lexical ambiguity that remains unresolved given the phonemes in a word that have been heard so far.

We found an anatomical gradient of change in the degree of linguistic representation where HG sites mostly encoded low-level linguistic features of phonetics and phonotactics. In contrast, the PT and STG were more responsive to higher-level lexical and semantic features. It is important to mention that these coarse anatomical boundaries do not necessarily conform to functional auditory fields, which is why we primarily used a functional clustering of neural responses rather than anatomical grouping. Nevertheless, cytoarchitectonic (cellular)[88–90] and myeloarchitectonic (fibre)[91] studies have shown a gradient of structural change from the PAC in HG to non-primary regions of the PT to non-primary areas in the lateral STG. Our findings are consistent with these studies that identify HG as the locus of the PAC[92,93], PT as an intermediary stage[94] and STG as the processing location for high-level speech units[32,58]. However, there is disagreement on whether pSTG is critical for word comprehension[95–97]. Some argue that pSTG is involved in the phonological representation of words and has a supportive rather than critical role in word comprehension, while aSTG and the temporal pole are crucial for word comprehension[96,97]. At first glance, this runs contrary to our finding in Fig. 2f, which shows lexical–semantic encoding in pSTG and not in aSTG. A possible explanation for this discrepancy could be sampling bias, since we do not have adequate sampling of the more anterior part of aSTG as opposed to the more posterior parts of pSTG. The few electrodes in the most anterior part of left aSTG in Fig. 4 that have a stronger lexical–semantic than lexical–phonological encoding suggest that semantic representation may be stronger in the more anterior parts of aSTG. Future studies with higher-density recordings from these areas and other brain regions implicated in speech processing[26,29,36] could further tease apart the linguistic response properties within each auditory field and provide critical information for fully describing the functional organization of the human speech cortex. Moreover, whether the observed linguistic feature encoding occurs at the single-neuron level or is an emergent property of population responses cannot be differentiated in our data because the high-gamma responses recorded from electrocorticography electrodes reflect the neural firing of a large population of neurons in the proximity of the recording sites[51,52]. For example, if neuron A is tuned to feature X, and neuron B to feature Y, the nearby electrode will exhibit a tuning to feature $X + Y$, which is not coded by any underlying individual neuron. Electrophysiological methods that allow recording from a smaller number of neurons[98,99] will help further clarify the representational and computational properties of linguistic feature encoding in the human auditory cortex to better define the functional and anatomical organization of the speech cortex.

Our findings could have direct implications for studies of speech communication impairment. The acquired patterns of impairment show great heterogeneity, particularly in the level of linguistic processing that is impacted when mapping an auditory stimulus onto lexical meaning. Studies of focal brain injury have shown selective impairments in acoustic phonetic cues, phonemic categories, lexical–phonological forms, and semantic and syntactic representations, demonstrating that there must be distinct intermediary processing stages. For example, cortical deafness causes impairments in all tasks that require phonetic or phonological processing[41]. Patients with pure word deafness show deficits at the subphonemic level—for example, in the identification of voicing or the place of articulation in stop consonants[42–44]. Other kinds of deficits implicate representations or processes further upstream from the acoustic–phonetic level—for example, in the failure to map an acoustic representation onto phonemic categories[40]. Impairments selective to the lexical–phonetic level are seen in conditions such as word-meaning deafness[45,47], where a participant can be better at repeating words than at repeating non-words, even though the participant is unable to comprehend spoken words[45]. A breakdown in the mapping from lexical–phonological forms to more abstract lexical–semantic representations also appears in patients with transcortical sensory aphasia[46] and Wernicke's aphasia, in which defective lexical comprehension occurs. These patients can repeat both words and non-words but exhibit impaired auditory comprehension. Finally, it has been shown that auditory lexical comprehension deficits in aphasic patients are not due to a perceptual deficit below the lexical level, as demonstrated by a weak correlation between comprehension measures and phoneme discrimination scores in these patients[100]. Our finding of anatomically distributed encoding of linguistic hierarchy, from prelexical to semantic levels, supports the notion that the route from sound to lexical meaning includes multiple intermediate processing stages that can be selectively disrupted by focal brain injury. However, the exact nature of these intermediate levels remains debated. In addition, with an encoding paradigm, we cannot prove causality; as such, we cannot rule out the possibility that a feature that first appears in a certain brain region is extracted in a higher region (inside or outside the auditory cortex) and then fed back upstream. To answer this question, we performed latency analysis comparing the latency of the same feature across different neural populations, but the results were inconclusive. Further research is needed to determine how these processes are impaired in various speech communication disorders—for example, by precisely mapping the anatomical distribution of the intermediate stages within each participant or distinguishing the feedforward and feedback mechanisms that contribute to the extraction of these linguistic features[101,102].

Due to the inherent correlations between linguistic elements (Extended Data Fig. 5), linguistic information at two different levels can overlap in their predictive power. This confounding factor motivated us to disregard the shared information and only study the independent effects of each variable by controlling for competing possibilities. To this end, we included spectrograms with multiple frequency bins instead of a simple speech envelope to account for subphonetic spectral information, we included acoustic edges to account for effects that have been shown to correlate with phonetics[76] and we chose a wide range of intermediate linguistic features to account for effects at different levels of coarseness. Nonetheless, we cannot entirely rule out the possibility that a neglected intermediate feature will be able to explain the effects of a higher-level feature included in our model.

We have intentionally limited the scope of our study to linguistic features that are insensitive to context. For example, in certain sentences there could be competing estimates of word boundaries from the perspective of the listener[103]. Since we choose firm boundaries for all words based on the speech transcript, we do not consider these competing estimates. The surrounding context can also alter a word’s meaning and recognition time. We presented the values for each feature over the entire duration of their corresponding units[72]—phoneme for P, T and L, and word for F and S—as opposed to presenting them only at the onset of each unit[63,71]. This choice allowed us to account for the natural variation in the duration of phonemes and words, but it does not account for different recognition times. An alternative is to present F and S features at each word’s uniqueness point as opposed to its onset. In the context of our study, it is not obvious that this is a better choice. For example, word frequency (F) has been shown to have early behavioural effects that start right after the word onset[64]. Additionally, the uniqueness point is more relevant to words that are heard in isolation, since different contexts can alter the exact time of recognition for a word and make it more difficult to pinpoint the uniqueness time. In either case, this phenomenon limits the temporal precision of any analysis performed on continuous speech data.

In summary, our results provide direct evidence for a sequential extraction of linguistic features in a hierarchy with a high degree of anatomical specificity. These findings shed light on the representational and computational organization of cortical speech processing and pave the way towards the construction of more comprehensive neurophysiological models of speech processing in the human speech cortex.

Methods

Participants and neural recording

Fifteen patients with pharmacoresistant focal epilepsy were included in this study (eight male, seven female; age mean, 36; s.d., 14; range, 19–58 years). All patients underwent chronic iEEG monitoring at North-shore University Hospital to identify epileptogenic foci in the brain for later removal. Twelve patients were implanted with stereoelectroencephalographic depth arrays only (2 mm or 1.3 mm platinum cylinders, 4.4 mm or 2.2 mm centre-to-centre distance, 0.8 mm diameter; PMT Corporation), and three were implanted with both depth electrodes and subdural grids and/or strips (2 mm or 3 mm platinum disks, 4 mm or 10 mm centre-to-centre distance; PMT Corporation). Intracranial EEG time series were manually inspected for signal quality and were free from interictal spikes. All research protocols were approved and monitored by the institutional review board at the Feinstein Institutes for Medical Research, and informed written consent to participate in research studies was obtained from each patient before electrode implantation. No statistical methods were used to pre-determine the number of participants or electrodes, but our sample size is similar to those reported in previous publications[58,104].

Intracranial EEG signals were continuously acquired with two different setups: nine patients were recorded at 3 kHz per channel (16-bit precision, range ±8 mV, DC) with a TDT data acquisition module (Tucker-Davis Technologies), and six patients were recorded at 1 kHz per channel with a Natus data acquisition module (XLTEK EMU128FS/NeuroLink IP 256 systems; Natus Medical Inc.). A subdural or subdermal electrode was used as a reference, determined by signal quality at the bedside after online visualization of the spectrogram of the signal. The envelope of the high-gamma response (75–150 Hz) was extracted by first filtering neural signals with a bandpass filter and then using the Hilbert transform to calculate the envelope. The high-gamma responses were z-scored and resampled to 100 Hz. Speech signals were simultaneously recorded with the iEEG to allow precise synchronization between the stimulus and the neural recording.

Electrode localization

Electrodes were localized using the iELVis toolbox[105]. Prior to the iEEG recordings, each patient underwent a T1-weighted 1 mm isometric structural magnetic resonance imaging (MRI) scan on a 3 T scanner. After the electrode implantation, a computed tomography (CT) scan together with a T1-weighted MRI scan at 1.5 T were acquired. The post-implantation CT and MRI scans were co-registered to the preoperative MRI scan using FSL's BET and FLIRT algorithms[106–108]. Afterwards, the artefacts of the contacts on the co-registered CT were identified manually in BioImageSuite[109]. Volumetric information was obtained by processing and reconstructing the T1 scan using FreeSurfer v.6.0 (recon-all command)[77]. For the anatomical analyses across participants, we mapped the coordinates of the electrodes for each participant to the FreeSurfer average brain (fsaverage), which is a template brain based on a combination of MRI scans of 40 real brains.

Stimuli

The stimulus consisted of continuous speech (two male and two female speakers). Half the content was selected from a children's storybook ('Hank the Cowdog'), and the other half comprised four short instructional monologues on how to perform different tasks (for example, how to make waffles). The total duration of the auditory material was 30 minutes, and it was sampled at 11,025 Hz. The 30-minute data were recorded in 53 segments of roughly equal duration, each corresponding to a few sentences. There were no long pauses within a segment.

Task

The audio segments were presented to the participants in a fixed order with a short pause between segments belonging to the same story (usually less than a minute) and a longer pause (possibly a few minutes) between different stories. At the end of some segments chosen at random, the participants were asked to repeat the last sentence of the segment.

This was done to ensure that they were paying meaningful attention to the stimulus and following the storyline.

Electrode selection

To determine whether an electrode site responds to speech, we compared its neural activity during pre-stimulus silence ([−1, 0] second period relative to segment onset) and speech ([0.5, 1.5] second period relative to segment onset). We concatenated the activity of each group across all 53 segments and performed a two-sided Wilcoxon rank-sum test between response time points in the speech and silence groups. Electrode sites with $|Z| \geq 10$ were defined as responsive to speech and included in the analysis. Of the total 4,186 electrodes, 535 passed this test. Since response to speech is compared with response to silence and not non-speech sounds, speech responsiveness in this case does not necessarily mean speech specificity.

Finally, we constrained our analyses to responsive sites within the auditory cortex—specifically, HG, which includes the PAC or core[110]; the PT, which coincides with the parabelt; and the STG. Electrodes coregistered to MRI images were labelled independently by two authors (S.A. and S.B.) as being located in HG, PT, aSTG or pSTG, and any discrepancy was then discussed and resolved. The border between aSTG and pSTG was defined as the crossing of a virtual line extending from the trans-verse temporal sulcus with the lateral surface of the STG[104,111]. Of the 535 responsive electrode sites, 242 passed this criterion.

Acoustic features

An auditory spectrogram representation of speech was calculated from a model of the peripheral auditory system[74]. This model consists of the following stages: (1) a cochlear filter bank consisting of 128 constant-Q filters equally spaced on a logarithmic axis, (2) a hair cell stage consisting of a lowpass filter and a nonlinear compression function, and (3) a lateral inhibitory network consisting of a first-order derivative along the spectral axis. Finally, the envelope of each frequency band was calculated to obtain a time-frequency representation simulating the pattern of activity on the auditory nerve. The final spectrogram had a sampling frequency of 100 Hz. The spectral dimension was downsampled from 128 frequency channels to 16 channels to reduce the model complexity. Acoustic edges were calculated per frequency bin as the half-wave rectified derivative of the spectrogram:

$$\text{onset}\left(t, \upsilon\right) = \begin{cases} x(t, \upsilon) - x(t-1, \upsilon), x(t, \upsilon) \geq x(t-1, \upsilon) \\ 0, \text{otherwise} \end{cases}$$

where $x(t, \nu)$ is the value of the spectrogram at time t for frequency band ν.

Phoneme and word alignment

We used the Prosodylab-Aligner[112] to align the speech stimuli to the words in the speech transcript and partition the words into phonemes of the International Phonetic Alphabet for American English. The estimated phoneme and word boundaries were then inspected to make sure the alignment succeeded for all stimuli.

Phonetic features

The phonetic features for each phoneme included 22 binary phoneme attributes defining the voicing, manner of articulation and place of articulation of each phoneme (for the complete list, see Extended Data Fig. 3). To generate the control data for phonetic features, we took the Carnegie Mellon University pronunciation dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict), grouped words by their length measured in phonemes and then shuffled the word-to-phoneme mapping within each group. As a result, each word has a consistent pronunciation at every occurrence, but words that share phonemes have independent pronunciations—for example, /kæt/ and /bæt/ no longer share two of their three phonemes. We constrained the reassociation to words of same length so that we kept the phoneme alignment information intact and because words of same length are more similar in frequency of occurrence (that is, shorter words tend to be more frequent). This is a rather strict control since shuffling pronunciations with other actual English words maintains the proper syllabic structure for English words.

Phonotactic features

As a measure of phonotactic probabilities, we used the logarithm of phoneme bigram (biphone) frequencies. To calculate the bigram frequencies, we used the Carnegie Mellon University dictionary to convert words to phoneme sequences and counted the total occurrence of each bigram using the SUBTLEX-US corpus, which is an English word frequency dataset calculated from movie subtitles[59]. We then computed a log-frequency metric for each bigram *ab*:

$$\text{logfreq}_{ab} = \log(\text{freq}_{ab}) .$$

Since the biphone frequencies were calculated from a word frequency dataset and without access to word transition probability information, we counted the first phoneme transition of words separately from non-first phonemes. For example, the biphones for the phrase ‘red hat’ are the following: /#r/, /re/, /ed/, /#h/ (not /dh/), /hæ/ and /æt/.

To generate controls for the phonotactic feature, we shuffled the bigram-to-frequency associations (that is, the look-up table for bigram frequencies), which means that each bigram was associated with the frequency of a randomly chosen bigram from the true distribution. This control scheme maintained consistency across multiple occurrences of the same bigram. To counter the effect of the separation caused by the first versus non-first phoneme grouping, we performed the above shuffling separately for first phones (ones starting with #) and non-first biphones, so that any first versus non-first effect would be maintained in the control and thus discounted.

Word frequency

For the word frequency feature, we simply used the log-frequency of words from the SUBTLEX-US dataset. For the control condition, we grouped words on the basis of their phoneme length and shuffled the word-to-frequency associations within each group.

Lexical–phonological features

To measure the lexical–phonological effect, we used lexical entropy and surprisal (equivalent to cohort entropy and phoneme surprisal in ref.[63]). These values were calculated for each phoneme within a word from the previous phonemes in that word. The surprisal caused by phoneme φ_i, $S(i)$, in word $w = \varphi_1 \ldots \varphi_K$ indicates the improbability of hearing phoneme φ_i based on the previous $i - 1$ phonemes that came before it in the word and is calculated as follows:

$$s\left(i\right) = -\log_2 \frac{\mathrm{freq}(\mathrm{cohort}_i)}{\mathrm{freq}(\mathrm{cohort}_{i-1})}$$

where freq(cohort$_i$) is the summed frequency of all words that start with the phoneme sequence $\varphi_1 \ldots \varphi_i$. The lexical entropy, $E(i)$, for phoneme φ_i is the entropy within all words that start with the phoneme sequence $\varphi_1 \ldots \varphi_i$ (the cohort)[63]:

$$E\left(i\right) = -\sum_{\mathrm{word} \in \{\mathrm{cohort}_i\}} p(\mathrm{word}) \log_2 p(\mathrm{word})$$

where p(word) indicates the relative frequency of the word within the cohort. These two parameters together encode the incremental lexical competition among all phonologically consistent candidates as a word is being heard, weighted by their frequency. To compute lexical surprisal for the word-initial phoneme, we assumed a transition from the entire lexicon—that is, how surprising it is to hear a word starting with phoneme φ given all the words in the lexicon.

To generate lexical–phonological controls, we grouped all cohorts on the basis of the length of their shared phoneme sequence and shuffled the cohort-to-frequency associations within each group. We used this constrained shuffling to keep the effect of secondary information such as the phoneme position in the word and word length unchanged. This control scheme also satisfies consistency—that is, if two words share their first k phonemes, the cohort information for their first k positions is the same because the same cohorts are mapped to the same information.

Lexical–semantic features

To study the encoding of semantic information, we represented each word with its SND obtained from the English Lexicon Project, which refers to the relative distance between a word and its closest neighbours based on a global cooccurrence model[56,68]. The neighbourhood density can encode the degree of activation of semantically related words in the lexicon upon hearing the target word. The control for the semantic condition was constructed by grouping words on the basis of their phoneme length and shuffling the word-to-SND associations within each group.

Fitting TRFs

Regularized linear TRF models were fitted using ridge regression with the multivariate TRF (mTRF) MATLAB toolbox[113]. A TRF is equivalent to a one-dimensional convolution along time or a finite impulse response filter. For each electrode, a causal model was trained to predict the neural response at each time point from the current and past 500 ms (time samples [$t - 50$, t]) of auditory stimulus (16 + 16 dimensions) and linguistic features (22 + 1 + 1 + 2 + 1 dimensions). Including a constant bias term, each model for an electrode has a total of 51 × 59 + 1 = 3,010 parameters. Phonetic, phonotactic and lexical–phonological information was specified for the full duration of each phoneme, while word frequency and lexical–semantic information were specified for the full duration of each word. All linguistic information was convolved with a Hanning window with a width of seven samples (±30 ms) to smooth the transitions between adjacent units while also capturing the coarticulation effect for phonemes. The optimal regularization parameter for each electrode was chosen by 53-fold cross-validation based on the experimental trials. We used the average prediction r value from the 53 left-out trials as our main performance measure.

For visualization purposes in Fig. 3, we fit slightly modified models to the data. First, we did not convolve the linguistic features with a Hanning window to keep transitions sharp, improving the temporal precision of the model TRF weights. Second, we expanded the temporal width of the TRF by 100 ms from each side (710 ms total), making the model non-causal. This means that predictors in the future 100 ms (ten samples) can influence the model prediction at the current time step. We discarded 50 ms (five samples) from each side of the resulting TRFs for the analyses in Fig. 3 to avoid regression artefacts at the extremes of the model[113]. The encoding model for each electrode was only fit with the optimal regularization parameter that maximizes the cross-validated out-of-sample prediction for that electrode. The final model weights were obtained by fitting $k = 53$ models for the k-fold data and averaging across folds.

Determining the significance of encoding

The significance of each linguistic feature's encoding was determined by comparing the encoding of the true distribution of that feature in the English language with 100 control conditions with permuted distributions, denoting significance as the t statistic of the one-sample one-tailed t-test testing whether $\Delta r > 0$ being greater than 19, where Δr is the distribution of the difference of the true model out-of-sample prediction score from those of the control models. This threshold on the statistic corresponds to a confidence of 97.1%. In theory, the dimensionality of a feature could artificially affect its measured significance of encoding, especially for phonetic features, whose dimensionality is an order of magnitude higher than that of the other linguistic features. We explored this matter by recomputing the encoding of dimensionality-reduced (PCA) phonetic features and found that the degree of encoding is correlated with the variance explained rather than the number of dimensions, and that the reduction in phonetic encoding only leads to an increase in phonotactic encoding.

Computing distance from PAC

We used posteromedial HG (TE1.1) as a reference for the PAC (Fig. 1a)[54,55]. We calculated the Euclidean distance in the fsaverage space between each electrode and the reference point on the corresponding hemisphere.

Measuring response latency

Electrode response latencies (Figs. 1a and 2c) were measured by fitting a spectro-temporal receptive field to each electrode (ridge regression from the time-lagged spectrogram to neural activity), taking the sum of squares of spectro-temporal receptive field coefficients across frequencies and finding the location of the peak of that curve. Feature-based latencies (Fig. 3b) were measured using the procedure described in the next section.

Computing the temporal response profiles of features

To compute the temporal response profile for a given feature f_i with a measure of robustness, we selected the group of electrodes that significantly encoded f_i and then randomly resampled that group 1,000 times with replacement (bootstrapping). For each subsample, we computed the eigenvector with the most explained variance (the first PC) of the model coefficients (TRFs) corresponding to feature f_i by performing a PCA using singular value decomposition across the sampled electrodes.

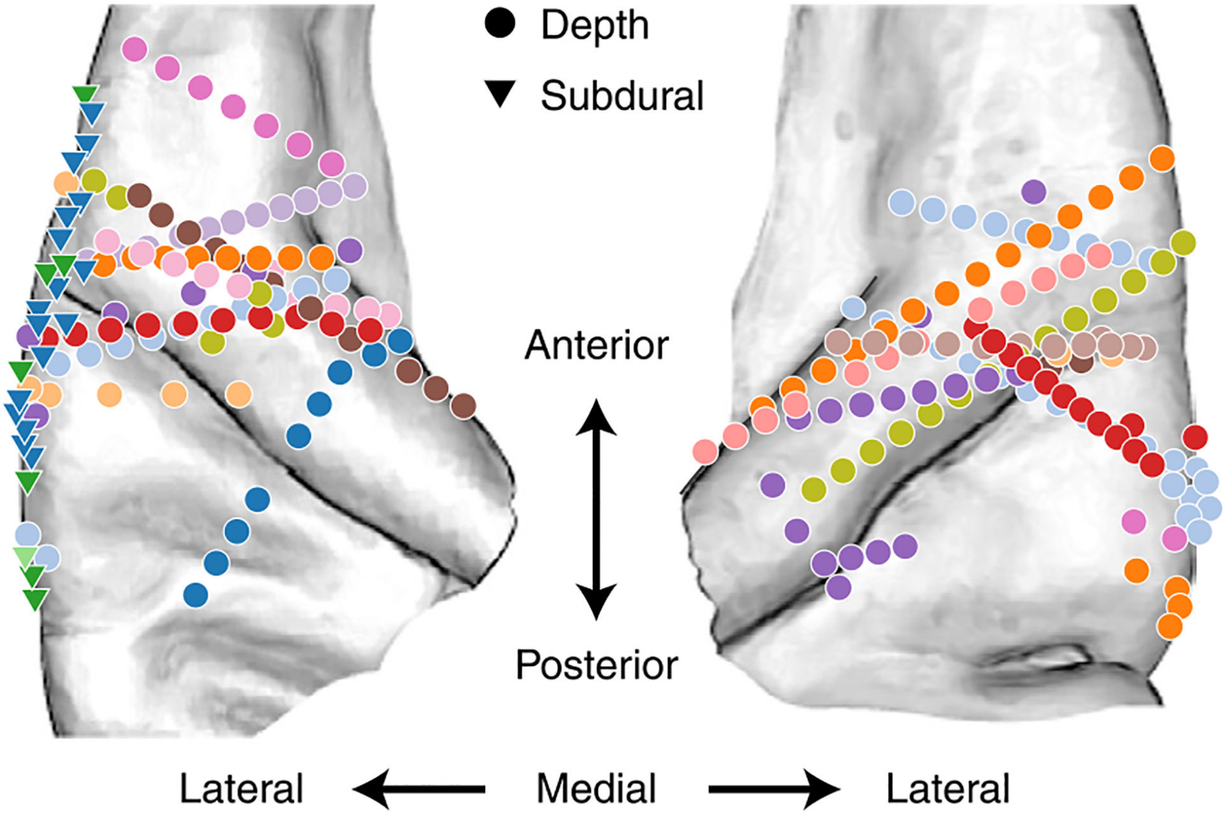
For Fig. 3a, to generate the temporal response profile of f_i, we took the mean and standard deviation of the 1,000 PCs, over all subsamples. For Fig. 3b, to measure the feature-based latency for f_i, we first computed the latency of the peak of the first PC and then took the mean and standard deviation of the 1,000 latencies, over all subsamples. Extended Data Fig. 6 shows on average how much of the total variance is explained by the first PC in the case of each feature. For the features that represent multi-dimensional embeddings (auditory spectrogram, acoustic edges and phonetic features), we performed the PCA across channels and feature dimensions simultaneously. In other words, for the $T \times D \times C$ weight tensor where T is the time window, D is the feature dimension (for example, 22 for phonetic features) and C is the size of the subgroup significantly encoding f_i, we first reshaped the tensor into a $T \times DC$ matrix and then performed PCA across the DC dimension. Note that since the time course displayed in Fig. 3a is that of a PC and not the direct weight associated with an electrode, its sign is ambiguous.

Clustering feature encoding effects

The input to the clustering analysis was a feature-by-electrode matrix of t values denoting the prediction improvement in the true model over the null distribution of models that had one feature replaced by a control feature for each electrode. Since the t values varied in range from feature to feature, we performed a non-standard transformation on the t values prior to clustering, since a large-valued feature (that is, phonetics) could easily dominate the clustering. All t values less than 2.5 were clipped to 2.5, t values between 2.5 and 7.5 remained linear, and t values greater than 7.5 were compressed through an $x^{0.47}$ transformation. The peculiar choice of parameters was aimed at achieving the most sensible automatic ordering of electrodes (x axis), since there are many valid orderings for the same clusters. Since our other (non-visual) analyses of t values are not sensitive to these outlier effects, we did not perform such transformations in other places.

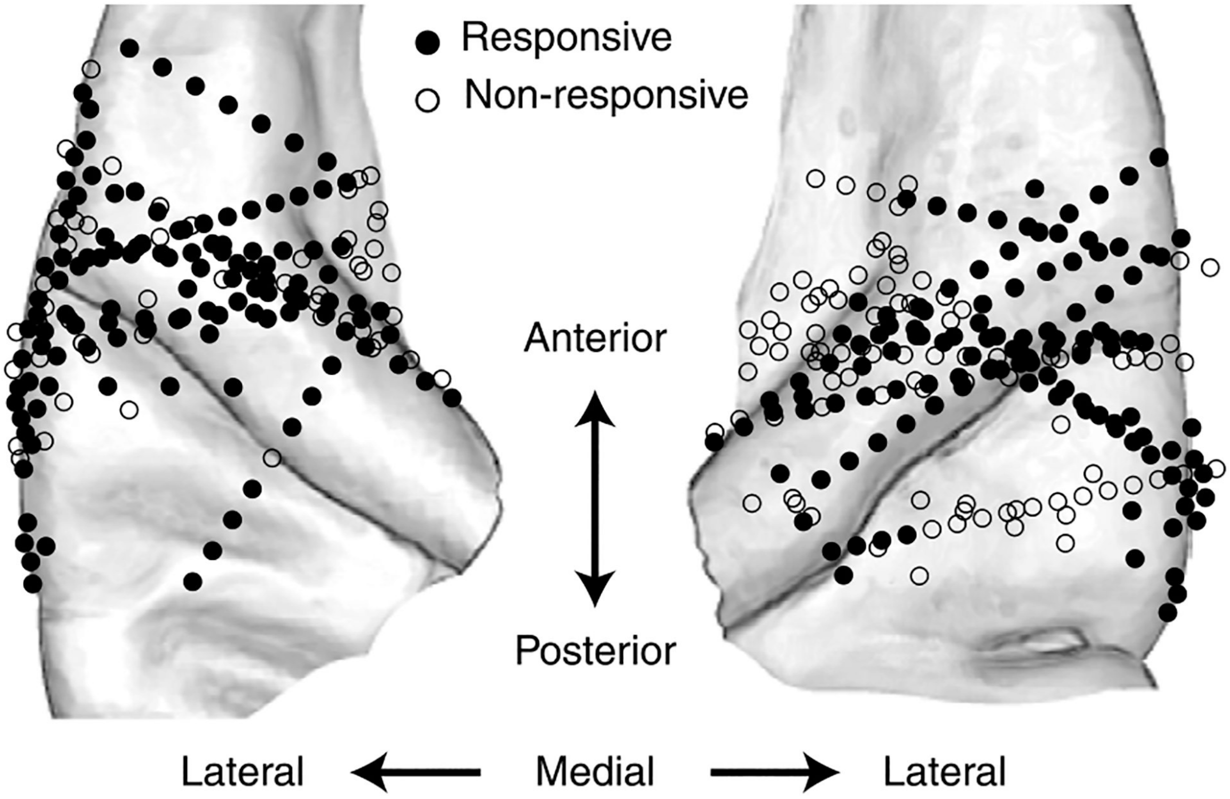
We performed agglomerative hierarchical clustering on the transformed data to group both linguistic features and electrodes on the basis of the similarity of their improved prediction accuracies. For clustering the features (rows), we used the correlation metric for distance and the average linkage clustering as the linkage criteria. We chose these parameters so that features that co-occur were grouped closer to each other. For electrode (column) clustering, we used the Euclidean metric as the distance and Ward's criterion as the linkage function[114]. We chose Euclidean for two reasons: the low number of features can be noisy if computing correlations, and we did not want to disregard the relative sizes of the effects. Ward's method is a linkage criterion suited for Euclidean distance when the number of elements being clustered is relatively large.

Extended Data



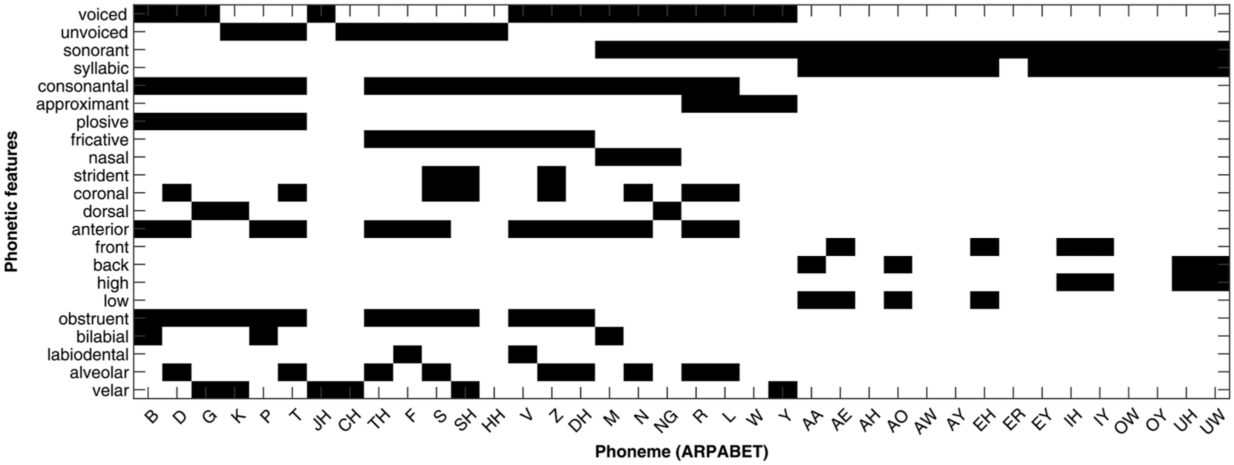
Extended Data Fig. 1 |. Electrode locations.

Electrodes are distributed across fifteen subjects and are either depth or subdural grids Cand/or strips. Shape indicates electrode type, where circles represent depth electrodes, and triangles represent subdural contacts. Shape colour indicates which of the fifteen subjects an electrode belongs to.



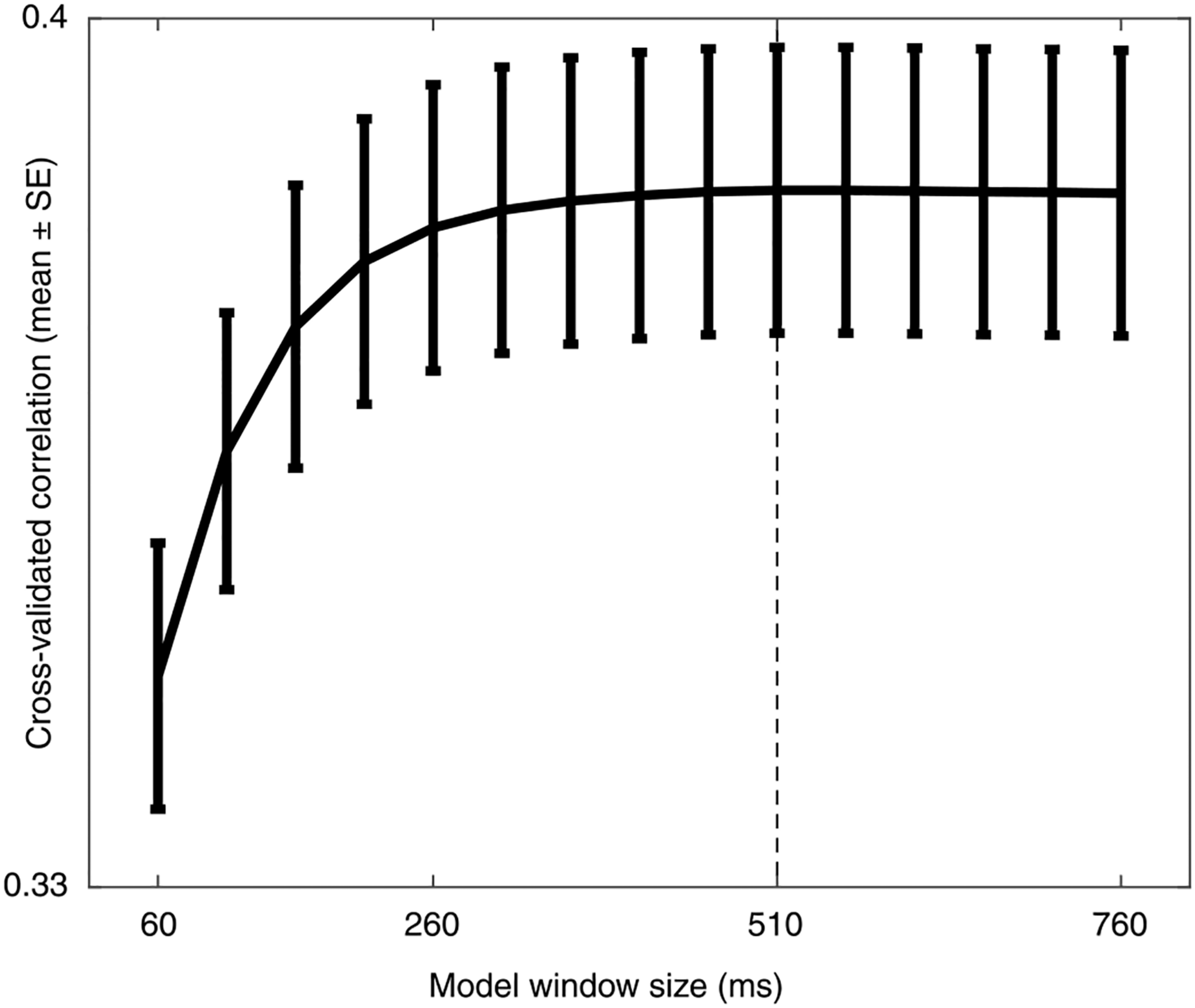
Extended Data Fig. 2 |. Speech-responsiveness.

Circles represent electrodes in the immediate vicinity of the auditory cortex, based on their 3D coordinates in the 'fsaverage' space. Filled circles indicate speech-responsiveness, meaning the electrode site responds significantly differently to speech compared to silence (see 'Electrode selection' in Methods). Non-responsiveness could mean the electrode is not sound-responsive, is sound-responsive but not speech-responsive, or has insufficient signal-to-noise ratio (SNR). The non-responsive electrodes were excluded from all analyses.



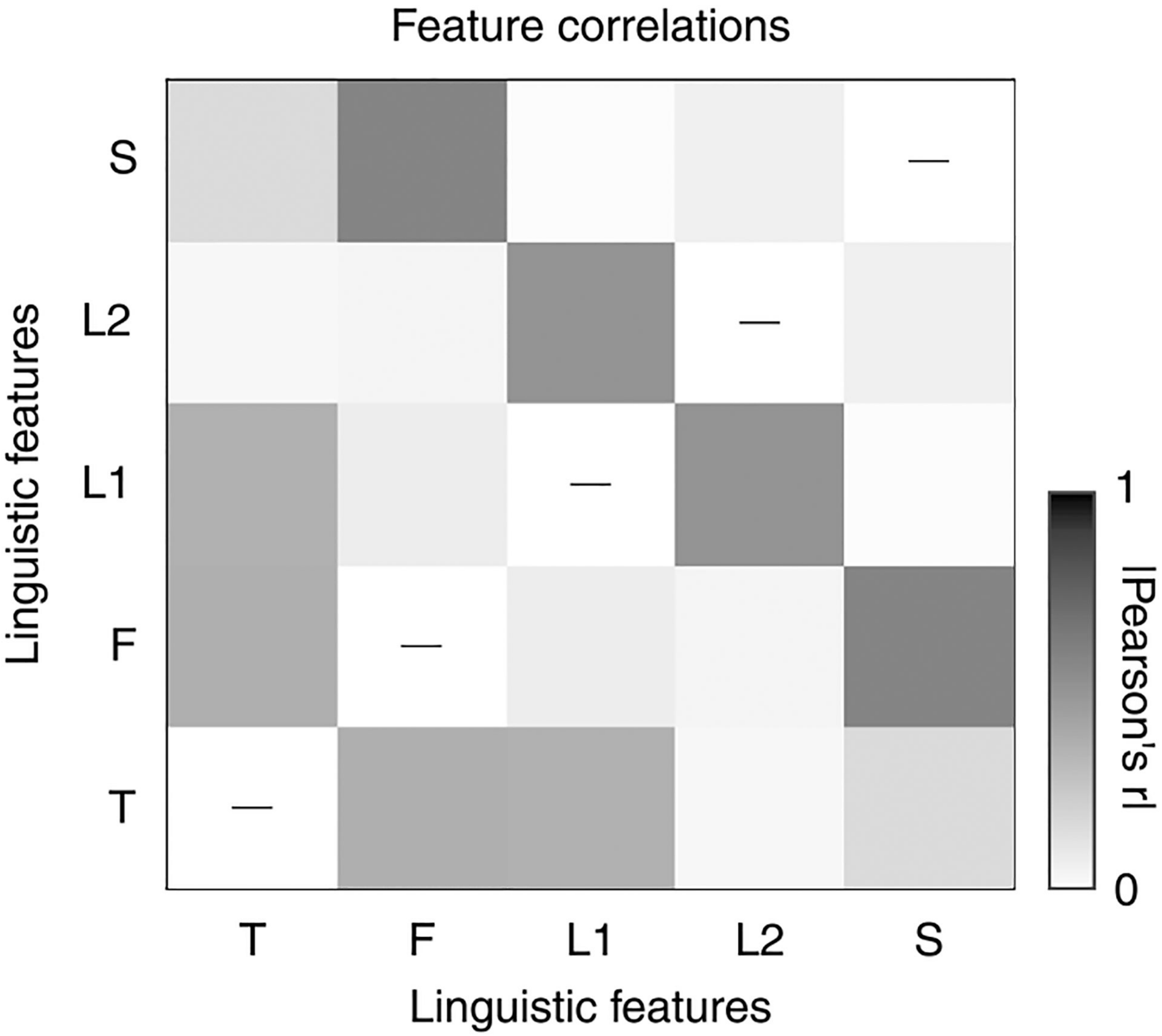
Extended Data Fig. 3 |. Phonetic features.

Each phoneme is represented by 22 binary features based on its voicing, place, and manner of articulation features.



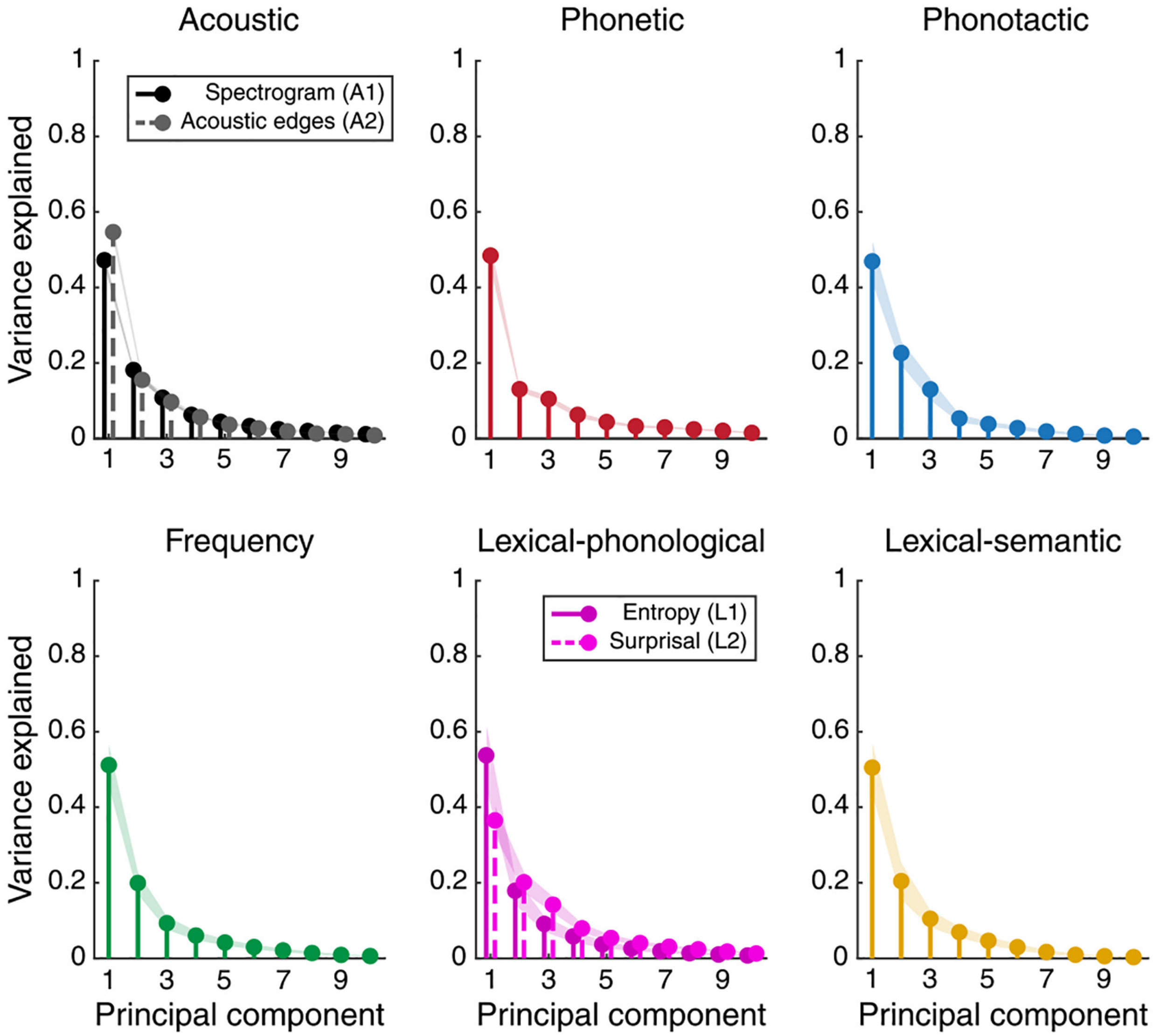
Extended Data Fig. 4 |. Selecting stimulus window length for prediction.

A window size of 510 ms was chosen to maximize linear model performance with the minimal number of parameters. We fit multiple models, each with a different number of time-lags (window size), from 60 ms to 760 ms. Each model was trained with the full list of predictors shown in Fig. 1c on all electrodes selected by the selection criteria described in Methods ($n = 242$), and only differed from the other models in the number of lags. Error bars indicate standard error (SE) over electrodes. To compare two different sizes, we perform a paired-sample one-tailed t-test on the cross-validated out-of-sample prediction r-values to determine whether the larger model improves upon the smaller one. The 510 ms model (dashed line) showed a significant improvement over all smaller models (60 ms – 410 ms, $p < 0.001$; 460 ms, $p = 0.023$). No larger model showed significant improvement over the 510 ms model ($p > 0.5$).



Extended Data Fig. 5 |. Correlations among features.

The linguistic features defined in this study are themselves correlated with each other. This plot shows the absolute value of the Pearson correlation coefficient for all pairs of 1-dimensional linguistic features (22-dimensional phonetic features excluded from figure; L1: lexical entropy, L2: lexical surprisal). The correlations are computed on the same 30-minute dataset used for all other analyses. All correlations are statistically significant ($p < 0.001$).



Extended Data Fig. 6 |. Explained variance of TRF diversity.

The bootstrapped (*n* = 1000) PCA analysis in Fig. 3 generates multiple eigenvectors at each bootstrap sample. We use the eigenvector that captures the most variance for computing the time courses (3a) and peak latencies (3b). This plot shows the mean and standard deviation of the explained variance for each of the top-10 principal components, computed using the same bootstrap procedure. In all cases, the first principal component captures roughly half of the total variance across all electrodes.

Acknowledgements

This work was funded by National Institutes of Health grant no. R01DC018805 (N.M.) and National Institute on Deafness and Other Communication Disorders grant no. R01DC014279 (N.M.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Data availability

Linguistic features were extracted from the SUBTLEX-US word frequency dataset[59] and the English Lexicon Project website (https://elexicon.wustl.edu/). The data that support the findings of this study are available upon request from the corresponding author (N.M.). The data are shared upon request due to the sensitive nature of human patient data.

References

1. Chomsky N & Halle M The Sound Pattern of English (Harper & Row, 1968).
2. Vitevitch MS & Luce PA Probabilistic phonotactics and neighborhood activation in spoken word recognition. J. Mem. Lang 40, 374–408 (1999).
3. Kiparsky P Word-formation and the lexicon. In Mid-America Linguistics Conference 3–29 (Mid-America Linguistics Conference, University of Kansas, Kansas, 1982).
4. Luce PA & Pisoni DB Recognizing spoken words: the neighborhood activation model. Ear Hear. 19, 1–36 (1998). [PubMed: 9504270]
5. Buchanan L, Westbury C & Burgess C Characterizing semantic space: neighborhood effects in word recognition. Psychon. Bull. Rev 8, 531–544 (2001). [PubMed: 11700905]
6. Grosjean F Spoken word recognition processes and the gating paradigm. Percept. Psychophys 28, 267–283 (1980). [PubMed: 7465310]
7. Marslen-Wilson WD Speech shadowing and speech comprehension. Speech Commun. 4, 55–73 (1985).
8. Marslen-Wilson WD Functional parallelism in spoken word-recognition. Cognition 25, 71–102 (1987). [PubMed: 3581730]
9. Allopenna PD, Magnuson JS & Tanenhaus MK Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. J. Mem. Lang 38, 419–439 (1998).
10. Dahan D & Magnuson JS in Handbook of Psycholinguistics (eds Traxler MJ & Gernsbacher MA) 249–283 (Elsevier, 2006); 10.1016/B978-012369374-7/50009-2
11. Magnuson JS, Mirman D & Harris HD in The Cambridge Handbook of Psycholinguistics (eds Spivey M et al.) 76–103 (Cambridge Univ. Press, 2012); 10.1017/cbo9781139029377.008
12. Pisoni DB & McLennan CT in Neurobiology of Language (eds Hickok G & Small SL) 239–253 (Elsevier, 2015); 10.1016/B978-0-12-407794-2.00020-1
13. Bidelman GM, Moreno S & Alain C Tracing the emergence of categorical speech perception in the human auditory system. NeuroImage 79, 201–212 (2013). [PubMed: 23648960]
14. Fernald A, Swingley D & Pinto JP When half a word is enough: infants can recognize spoken words using partial phonetic information. Child Dev. 72, 1003–1015 (2001). [PubMed: 11480931]
15. Magnuson JS, Dixon JA, Tanenhaus MK & Aslin RN The dynamics of lexical competition during spoken word recognition. Cogn. Sci 31, 133–156 (2007). [PubMed: 21635290]
16. Yee E & Sedivy JC Eye movements to pictures reveal transient semantic activation during spoken word recognition. J. Exp. Psychol. Learn. Mem. Cogn 32, 1–14 (2006). [PubMed: 16478336]
17. Tyler LK, Voice JK & Moss HE The interaction of meaning and sound in spoken word recognition. Psychon. Bull. Rev 7, 320–326 (2000). [PubMed: 10909140]
18. Mirman D & Magnuson JS Dynamics of activation of semantically similar concepts during spoken word recognition. Mem. Cogn 37, 1026–1039 (2009). 2009 37:7.
19. McClelland JL & Elman JL The TRACE model of speech perception. Cogn. Psychol 18, 1–86 (1986). [PubMed: 3753912]
20. Scharenborg O Modeling the use of durational information in human spoken-word recognition. J. Acoust. Soc. Am 127, 3758–3770 (2010). [PubMed: 20550274]
21. Norris D Shortlist: a connectionist model of continuous speech recognition. Cognition 52, 189–234 (1994).

22. Scharenborg O, Norris D, ten Bosch L & McQueen JM How should a speech recognizer work? Cogn. Sci 29, 867–918 (2005). [PubMed: 21702797]

23. Luce PA, Goldinger SD, Auer ET & Vitevitch MS Phonetic priming, neighborhood activation, and PARSYN. Percept. Psychophys 62, 615–625 (2000). [PubMed: 10909252]

24. Gaskell MG & Marslen-Wilson WD Integrating form and meaning: a distributed model of speech perception. Lang. Cogn. Process 12, 613–656 (1997).

25. Norris D in Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives (ed. Altmann GTM) 87–104 (MIT, 1990).

26. DeWitt I & Rauschecker JP Phoneme and word recognition in the auditory ventral stream. Proc. Natl Acad. Sci. USA 109, E505–E514 (2012). [PubMed: 22308358]

27. Poeppel D The neuroanatomic and neurophysiological infrastructure for speech and language. Curr. Opin. Neurobiol 28, 142–149 (2014). [PubMed: 25064048]

28. Price CJ The anatomy of language: a review of 100 fMRI studies published in 2009. Ann. N. Y. Acad. Sci 1191, 62–88 (2010). [PubMed: 20392276]

29. de Heer WA, Huth AG, Griffiths TL, Gallant JL & Theunissen FE The hierarchical cortical organization of human speech processing. J. Neurosci 37, 6539–6557 (2017). [PubMed: 28588065]

30. Langers DR, Backes WH & van Dijk P Spectrotemporal features of the auditory cortex: the activation in response to dynamic ripples. NeuroImage 20, 265–275 (2003). [PubMed: 14527587]

31. Chan AM et al. Speech-specific tuning of neurons in human superior temporal gyrus. Cereb. Cortex 24, 2679–2693 (2013). [PubMed: 23680841]

32. Chang EF et al. Categorical speech representation in human superior temporal gyrus. Nat. Neurosci 13, 1428–1432 (2010). [PubMed: 20890293]

33. Mesgarani N, David SV, Fritz JB & Shamma SA Phoneme representation and classification in primary auditory cortex. J. Acoust. Soc. Am 123, 899–909 (2008). [PubMed: 18247893]

34. Steinschneider M et al. Differential activation of human core, non-core and auditory-related cortex during speech categorization tasks as revealed by intracranial recordings. Front. Neurosci 8, 240 (2014). [PubMed: 25157216]

35. Ding N, Melloni L, Zhang H, Tian X & Poeppel D Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci 19, 158–164 (2015). [PubMed: 26642090]

36. Honey CJ et al. Slow cortical dynamics and the accumulation of information over long timescales. Neuron 76, 423–434 (2012). [PubMed: 23083743]

37. Leonard MK, Bouchard KE, Tang C & Chang EF Dynamic encoding of speech sequence probability in human temporal cortex. J. Neurosci 35, 7203–7214 (2015). [PubMed: 25948269]

38. Lerner Y, Honey CJ, Silbert LJ & Hasson U Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci 31, 2906–2915 (2011). [PubMed: 21414912]

39. Overath T, McDermott JH, Zarate JM & Poeppel D The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat. Neurosci 18, 903–911 (2015). [PubMed: 25984889]

40. Caramazza A, Berndt RS & Basili AG The selective impairment of phonological processing: a case study. Brain Lang. 18, 128–174 (1983). [PubMed: 6839129]

41. Engelien A et al. The neural correlates of ‘deaf-hearing’ in man: conscious sensory awareness enabled by attentional modulation. Brain 123, 532–545 (2000). [PubMed: 10686176]

42. Auerbach SH, Allard T, Naeser M, Alexander MP & Albert ML Pure word deafness: analysis of a case with bilateral lesions and a defect at the prephonemic level. Brain 105, 271–300 (1982). [PubMed: 7082991]

43. Wang E, Peach RK, Xu Y, Schneck M & Manry C II Perception of dynamic acoustic patterns by an individual with unilateral verbal auditory agnosia. Brain Lang. 73, 442–455 (2000). [PubMed: 10860565]

44. Poeppel D Pure word deafness and the bilateral processing of the speech code. Cogn. Sci 25, 679–693 (2001).

45. Franklin S, Turner J, Ralph MAL, Morris J & Bailey PJ A distinctive case of word meaning deafness? Cogn. Neuropsychol 13, 1139–1162 (1996).

46. Boatman D et al. Transcortical sensory aphasia: revisited and revised. Brain 123, 1634–1642 (2000). [PubMed: 10908193]

47. Kohn SE & Friedman RB Word-meaning deafness: a phonological–semantic dissociation. Cogn. Neuropsychol 3, 291–308 (1986).

48. Rauschecker JP & Scott SK Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat. Neurosci 12, 718–724 (2009). [PubMed: 19471271]

49. Rauschecker JP in The Senses: A Comprehensive Reference (ed. Fritzsch B) 791–811(Elsevier, 2020).

50. Gaskell MG & Marslen-Wilson WD Representation and competition in the perception of spoken words. Cogn. Psychol 45, 220–266 (2002). [PubMed: 12528902]

51. Ray S & Maunsell JHR Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLoS Biol. 9, e1000610 (2011). [PubMed: 21532743]

52. Buzsáki G, Anastassiou CA & Koch C The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. Nat. Rev. Neurosci 13, 407–420 (2012). [PubMed: 22595786]

53. Clarke S & Morosan P in The Human Auditory Cortex (eds Poeppel D, Overath T, Popper A & Fay R) 11–38 (Springer, 2012).

54. Norman-Haignere SV & McDermott JH Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. PLoS Biol. 16, e2005127 (2018). [PubMed: 30507943]

55. Baumann S, Petkov CI & Griffiths TD A unified framework for the organization of the primate auditory cortex. Front. Syst. Neurosci 0, 11 (2013).

56. Shaoul C & Westbury C Exploring lexical co-occurrence space using HiDEx. Behav. Res. Methods 42, 393–413 (2010). [PubMed: 20479171]

57. Ladefoged P & Johnson K A Course in Phonetics (Wadsworth Publishing Company,2010).

58. Mesgarani N, Cheung C, Johnson K & Chang EF Phonetic feature encoding in human superior temporal gyrus. Science 343, 1006–1010 (2014). [PubMed: 24482117]

59. Brysbaert M & New B Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behav. Res. Methods 41, 977–990 (2009). [PubMed: 19897807]

60. Ylinen S et al. Predictive coding of phonological rules in auditory cortex: a mismatch negativity study. Brain Lang. 162, 72–80 (2016). [PubMed: 27588355]

61. Friston K The free-energy principle: a rough guide to the brain? Trends Cogn. Sci 13, 293–301 (2009). [PubMed: 19559644]

62. Gagnepain P, Henson RN & Davis MH Temporal predictive codes for spoken words in auditory cortex. Curr. Biol 22, 615–621 (2012). [PubMed: 22425155]

63. Brodbeck C, Hong LE & Simon JZ Rapid transformation from auditory to linguistic representations of continuous speech. Curr. Biol 28, 3976–3983 (2018). [PubMed: 30503620]

64. Dahan D, Magnuson JS & Tanenhaus MK Time course of frequency effects in spoken-word recognition: evidence from eye movements. Cogn. Psychol 42, 317–367 (2001). [PubMed: 11368527]

65. Marslen-Wilson WD & Welsh A Processing interactions and lexical access during word recognition in continuous speech. Cogn. Psychol 10, 29–63 (1978).

66. Balling LW & Baayen RH Probability and surprisal in auditory comprehension of morphologically complex words. Cognition 125, 80–106 (2012). [PubMed: 22841290]

67. Wurm LH, Ernestus M, Schreuder R & Baayen RH Dynamics of the auditory comprehension of prefixed words. Ment. Lex 1, 125–146 (2006).

68. Balota DA et al. The English Lexicon Project. Behav. Res. Methods 39, 445–459 (2007). [PubMed: 17958156]

69. Danguecan AN & Buchanan L Semantic neighborhood effects for abstract versus concrete words. Front. Psychol 7, 1034 (2016). [PubMed: 27458422]

70. Mirman D & Magnuson JS The impact of semantic neighborhood density on semantic access. In Proc. 28th Annual Conference of the Cognitive Science Society (eds Sun R & Miyake N) 1823–1828 (2006).

71. Broderick MP, Anderson AJ, di Liberto GM, Crosse MJ & Lalor EC Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Curr. Biol 28, 803–809.e3 (2018). [PubMed: 29478856]

72. di Liberto GM, O'Sullivan JA & Lalor EC Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr. Biol 25, 2457–2465 (2015). [PubMed: 26412129]

73. Di Liberto GM, Wong D, Melnik GA & de Cheveigné A Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. NeuroImage 196, 237–247 (2019). [PubMed: 30991126]

74. Yang X-B, Wang K & Shamma SA Auditory representations of acoustic signals. IEEE Trans. Inf. Theory 38, 824–839 (1992).

75. Kluender KR, Coady JA & Kiefte M Sensitivity to change in perception of speech. Speech Commun. 41, 59–69 (2003). [PubMed: 28747807]

76. Daube C, Ince RAA & Gross J Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. Curr. Biol 29, 1924–1937 (2019). [PubMed: 31130454]

77. Fischl B et al. Automatically parcellating the human cerebral cortex. Cereb. Cortex 14, 11–22 (2004). [PubMed: 14654453]

78. Pisoni DB in Talker Variability in Speech Processing (eds Johnson K & Mullennix JW) 9–32 (Morgan Kaufmann, 1997).

79. Luce PA & McLennan CT in The Handbook of Speech Perception (eds Pisoni DB & Remez RE) 590–609 (Blackwell, 2008); 10.1002/9780470757024.ch24

80. Port RF Rich memory and distributed phonology. Lang. Sci 32, 43–55 (2010).

81. Nosofsky RM Attention, similarity, and the identification–categorization relationship. J. Exp. Psychol. Gen 115, 39–57 (1986). [PubMed: 2937873]

82. Kruschke JK ALCOVE: an exemplar-based connectionist model of category learning. Psychol. Rev 99, 22–44 (1992). [PubMed: 1546117]

83. Magnuson JS, Nusbaum HC, Akahane-Yamada R & Saltzman D Talker familiarity and the accommodation of talker variability. Atten. Percept. Psychophys 83, 1842–1860 (2021). [PubMed: 33398658]

84. McLaughlin D, Dougherty S, Lember R & Perrachione T Episodic memory for words enhances the language familiarity effect in talker identification. In Proc. 18th International Congress of Phonetic Sciences (ed. The Scottish Consortium for ICPhS 2015) 367.1–4 (University of Glasgow, Glasgow, 2015).

85. Choi JY, Hu ER & Perrachione TK Varying acoustic–phonemic ambiguity reveals that talker normalization is obligatory in speech processing. Atten. Percept. Psychophys 80, 784–797 (2018). [PubMed: 29417449]

86. Pisoni DB & Levi SV in The Oxford Handbook of Psycholinguistics (ed. Gaskell MG) 3–18 (Oxford Univ. Press, 2007); 10.1093/oxfordhb/9780198568971.013.0001

87. Vitevitch MS, Luce PA, Pisoni DB & Auer ET Phonotactics, neighborhood activation, and lexical access for spoken words. Brain Lang. 68, 306–311 (1999). [PubMed: 10433774]

88. von Economo CF & Koskinas GN Die Cytoarchitektonik der Hirnrinde des Erwachsenen Menschen (J. Springer, 1925).

89. Galaburda A & Sanides F Cytoarchitectonic organization of the human auditory cortex. J. Comp. Neurol 190, 597–610 (1980). [PubMed: 6771305]

90. Morosan P, Rademacher J, Palomero-Gallagher N & Zilles K in The Auditory Cortex (eds Heil P, Scheich H, Budinger E & Konig R) 45–68 (Psychology Press, 2005).

91. Hopf A Die Myeloarchitektonik des Isocortex Temporalis Beim Menschen (De Gruyter, 1951).

92. Moerel M, De Martino F & Formisano E An anatomical and functional topography of human auditory cortical areas. Front. Neurosci 8, 225 (2014). [PubMed: 25120426]

93. Nourski KV Auditory processing in the human cortex: an intracranial electrophysiology perspective. Laryngoscope Investig. Otolaryngol 2, 147–156 (2017).

94. Griffiths TD & Warren JD The planum temporale as a computational hub. Trends Neurosci. 25, 348–353 (2002). [PubMed: 12079762]

95. Hillis AE, Rorden C & Fridriksson J Brain regions essential for word comprehension: drawing inferences from patients. Ann. Neurol 81, 759–768 (2017). [PubMed: 28445916]

96. Mesulam M-M et al. Word comprehension in temporal cortex and Wernicke area. Neurology 92, e224–e233 (2019). [PubMed: 30578374]

97. Binder JR Current controversies on Wernicke's area and its role in language. Curr. Neurol. Neurosci. Rep 17, 58 (2017). [PubMed: 28656532]

98. Muller L, Hamilton LS, Edwards E, Bouchard KE & Chang EF Spatial resolution dependence on spectral frequency in human speech cortex electrocorticography. J. Neural Eng 13, 56013 (2016).

99. Khodagholy D et al. NeuroGrid: recording action potentials from the surface of the brain. Nat. Neurosci 18, 310–315 (2015). [PubMed: 25531570]

100. Blumstein SE, Baker E & Goodglass H Phonological factors in auditory comprehension in aphasia. Neuropsychologia 15, 19–30 (1977). [PubMed: 831150]

101. Norris D, McQueen JM & Cutler A Prediction, Bayesian inference and feedback in speech recognition. Lang. Cogn. Neurosci 31, 4–18 (2016). [PubMed: 26740960]

102. Magnuson JS, Mirman D, Luthra S, Strauss T & Harris HD Interaction in spoken word recognition models: feedback helps. Front. Psychol 9, 369 (2018). [PubMed: 29666593]

103. Norris D & McQueen JM Shortlist B: a Bayesian model of continuous speech recognition. Psychol. Rev 115, 357–395 (2008). [PubMed: 18426294]

104. Hamilton LS, Oganian Y, Hall J & Chang EF Parallel and distributed encoding of speech across human auditory cortex. Cell 184, 4626–4639.e13 (2021). [PubMed: 34411517]

105. Groppe DM et al. iELVis: an open source MATLAB toolbox for localizing and visualizing human intracranial electrode data. J. Neurosci. Methods 281, 40–48 (2017). [PubMed: 28192130]

106. Jenkinson M & Smith S A global optimisation method for robust affine registration of brain images. Med. Image Anal 5, 143–156 (2001). [PubMed: 11516708]

107. Jenkinson M, Bannister P, Brady M & Smith S Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841 (2002). [PubMed: 12377157]

108. Smith SM Fast robust automated brain extraction. Hum. Brain Mapp 17, 143–155 (2002). [PubMed: 12391568]

109. Papademetris X et al. BioImage Suite: an integrated medical image analysis suite: an update. Insight J. 2006, 209 (2006). [PubMed: 25364771]

110. Sweet RA, Dorph-Petersen K & Lewis DA Mapping auditory core, lateral belt, and parabelt cortices in the human superior temporal gyrus. J. Comp. Neurol 491, 270–289 (2005). [PubMed: 16134138]

111. Ozker M, Schepers IM, Magnotti JF, Yoshor D & Beauchamp MS A double dissociation between anterior and posterior superior temporal gyrus for processing audiovisual speech demonstrated by electrocorticography. J. Cogn. Neurosci 29, 1044–1060 (2017). [PubMed: 28253074]

112. Gorman K, Howell J & Wagner M Prosodylab-Aligner: a tool for forced alignment of laboratory speech. Can. Acoust 39, 192–193 (2011).

113. Crosse MJ, di Liberto GM, Bednar A & Lalor EC The Multivariate Temporal Response Function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. Front. Hum. Neurosci 10, 604 (2016). [PubMed: 27965557]

114. Ward JH Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc 58, 236–244 (1963).

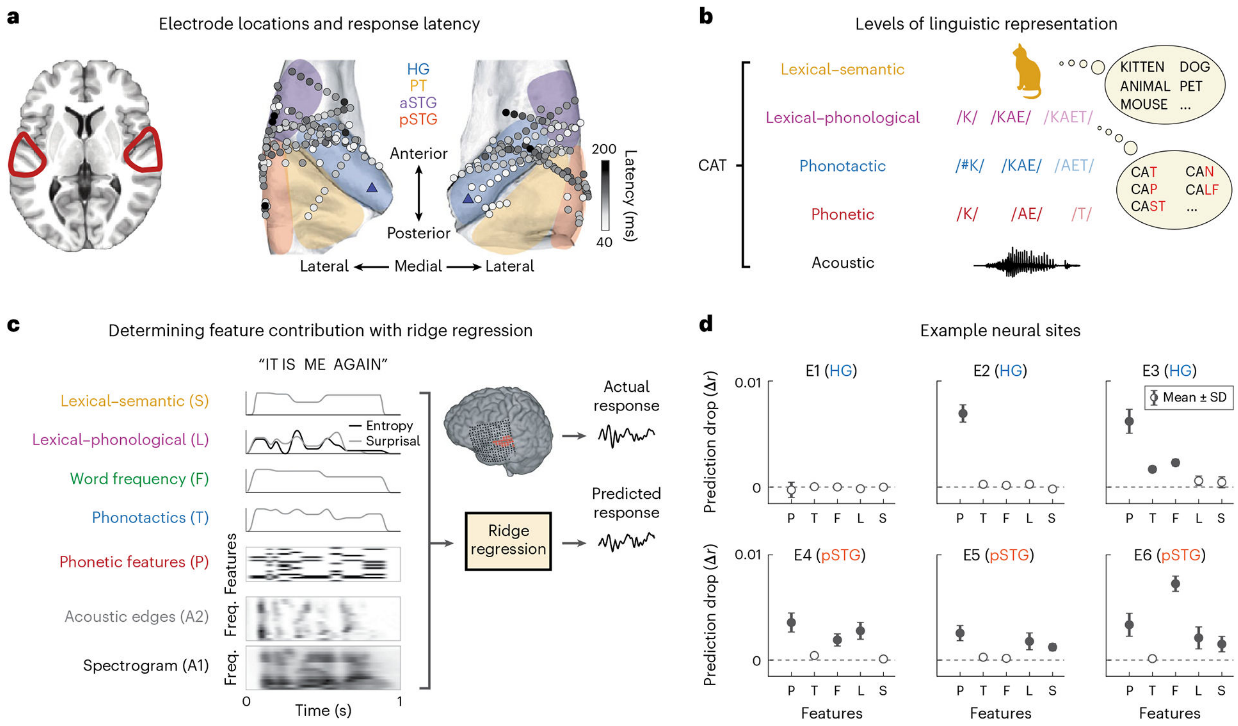


Fig. 1 |. Coverage and paradigm.

a, Electrode coverage and response latencies shown for 15 neurosurgical patients on the average FreeSurfer brain. The shaded regions indicate the general area of regions of interest. The blue triangles on the posteromedial HG sections indicate the reference point used for calculating distance from PAC. **b**, Different levels of linguistic features for the example word ‘cat’. **c**, Fitting TRFs using ridge regression for quantifying linguistic feature encoding in neural data. For each linguistic feature f, one at a time, we replaced that feature with a control variable $\widehat{f}$, generated by permuting values of that feature across the language. We then compared the accuracy of predicting the neural data with the true predictors versus control predictors where only f has been replaced with $\widehat{f}$. We repeated the process 100 times for each feature. **d**, Mean and standard deviation of the distribution of differences between true and control prediction accuracy (Δr_c) for six example electrodes. Zero indicates that the control model performed the same as the true model, while positive values indicate that the true model outperformed the control model. We performed a one-sample t-test for each of the distributions against zero (P/T/F/L/S t values: E1, −3.46/2.32/0.76/−7.25/−0.22; E2, 80.38/11.66/9.09/7.56/−13.21; E3, 52.63/54.36/64.16/11.72/7.96; E4, 37.58/18.00/30.59/32.63/5.11; E5, 33.24/10.15/9.61/20.96/31.93; E6, 29.34/4.99/93.65/19.31/19.50). The filled circles indicate that a feature was determined to be significantly encoded at the electrode site ($t > 19$).

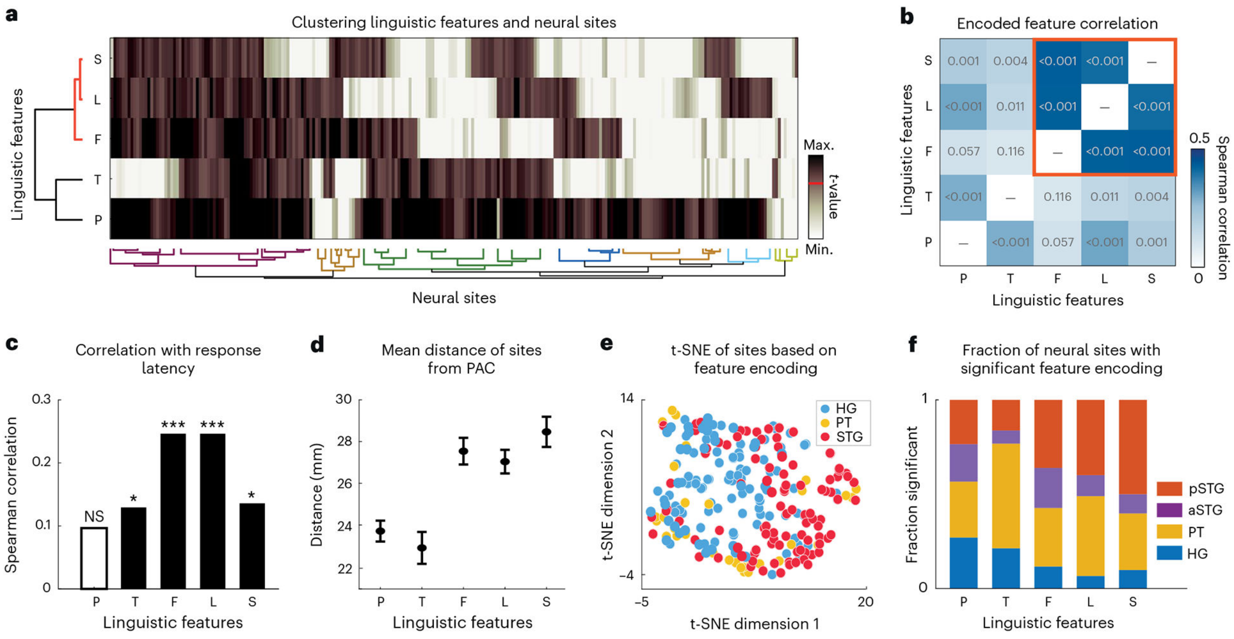


Fig. 2 |. Diversity in linguistic feature encoding.

a, Agglomerative clustering of the *t* values of prediction improvements over the corresponding control distribution (Fig. 1d) for all features and electrodes. Small values were clipped (white) and large values were compressed (dark brown) to reduce noise in clustering (Methods). The red horizontal line on the colour scale denotes the significance threshold ($t > 19$). **b**, Pairwise Spearman correlations between *t* values of different features computed across electrodes. The red square indicates lexical-level features (frequency, lexical–phonological and lexical–semantic) that are more highly correlated with each other. The values inside the boxes are the *P* values of the correlations. **c**, Spearman correlations of different feature encoding *t* values with neural response latencies of the electrodes (*P* values: P, 0.134; T, 0.044; F, <0.001; L, <0.001; S, 0.035). $*P < 0.05$; $***P < 0.001$; NS, not significant. **d**, Mean and standard error of electrode distance from posteromedial HG (TE1.1) as a reference for the PAC. For each feature, we measured the mean and s.e. over the subset of electrodes that significantly encode that feature ($t > 19$; sample sizes: P, 174; T, 51; F, 73; L, 55; S, 32 electrodes). **e**, *t*-distributed stochastic neighbour embedding (*t*-SNE) representation of feature encoding values from **a**. The colours indicate brain regions. **f**, The normalized fraction of electrodes in each brain region significantly encoding the corresponding feature ($t > 19$; sample sizes: HG, 113; PT, 32; aSTG, 46; pSTG, 51 electrodes). First, the fraction of neural sites in each region that significantly encode a particular feature is calculated. Next, this fraction is normalized by the sum of all fractions across regions.

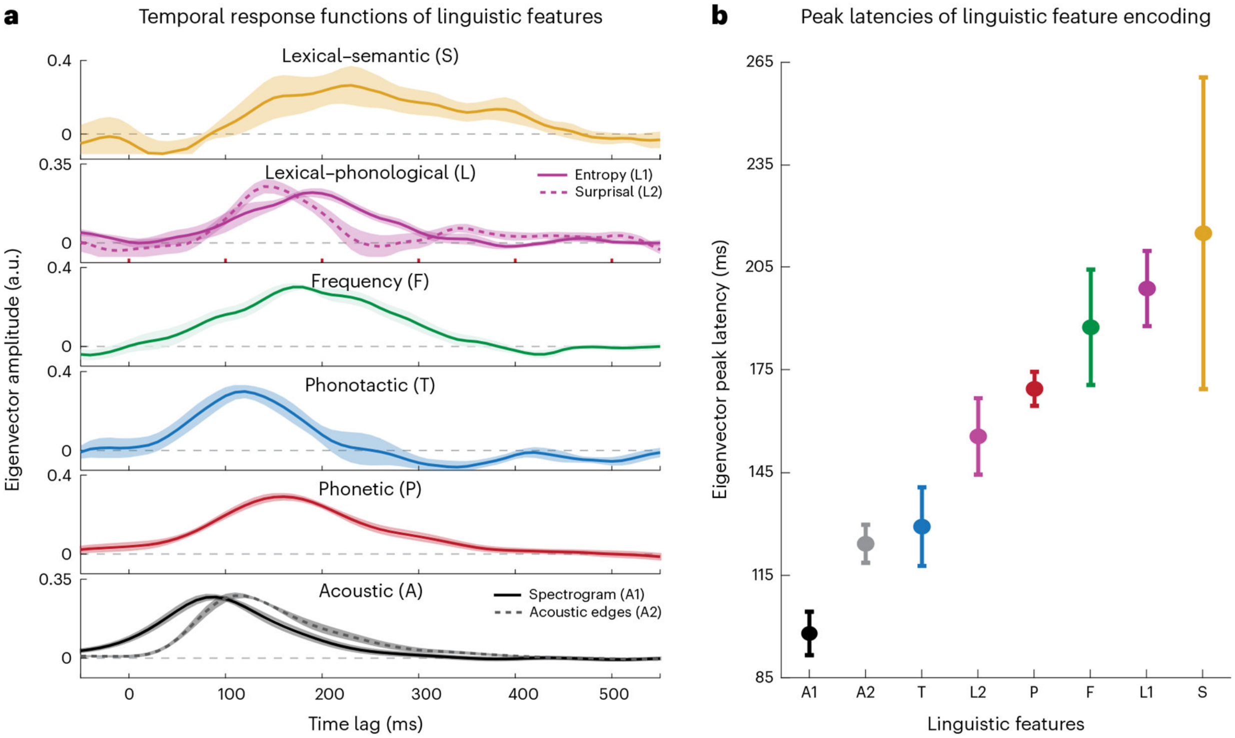


Fig. 3 |. Temporal profile of linguistic feature encoding.

a, Mean (lines) and standard deviation (shading) of the first eigenvector of the regression weights assigned to each linguistic feature computed over all neural sites that showed significant encoding for that feature ($t > 19$; sample sizes: A1/2, 242; P, 174; T, 51; F, 73; L1/2, 55; S, 32 electrodes). The statistics were calculated using a 1,000-sample bootstrap. **b**, Mean and standard deviation of the peak latencies of the first PCs computed in **a** ($n = 1,000$ bootstraps).

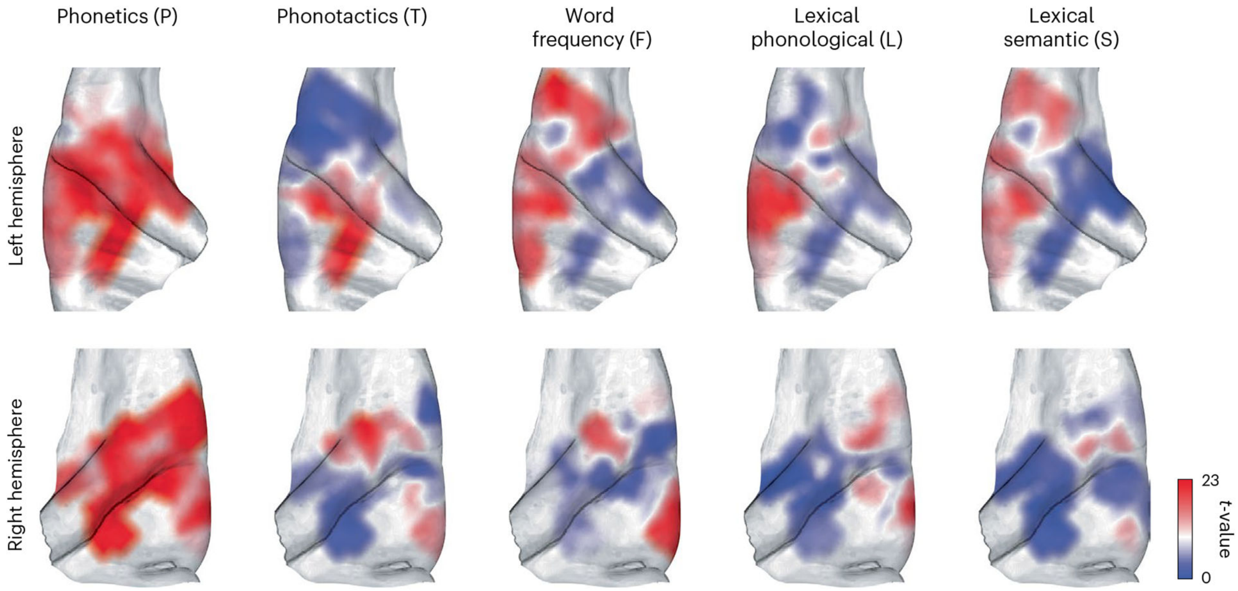


Fig. 4 |. Spatial profile of linguistic feature encoding.

Interpolated two-dimensional maps of *t* statistics representing linguistic feature encoding across the auditory cortex. The interpolation was performed using *k*-nearest neighbours ($k = 5$). Interpolated *t* values greater than 23 are shown with the same dark red and those less than 0 are shown with the same dark blue. Darker shades of red indicate stronger encoding of the corresponding feature by neighbouring electrodes.