Parallel hierarchical encoding of linguistic representations in the human auditory cortex and recurrent automatic speech recognition systems

Menoua Keshishian [1,2], Gavin Mischler [1,2], Samuel Thomas [3], Brian Kingsbury [3], Stephan Bickel [4,5], Ashesh D. Mehta [4,5], Nima Mesgarani [1,2,*]

Abstract

The human brain's ability to transform acoustic speech signals into rich linguistic representations has inspired advancements in automatic speech recognition (ASR) systems. While ASR systems now achieve human-level performance under controlled conditions, prior research on their parallels with the brain has been limited by the use of biologically implausible models, narrow feature sets, and comparisons that primarily emphasize predictability of brain activity without fully exploring shared underlying representations. Additionally, studies comparing the brain to text-based language models overlook the acoustic stages of speech processing, an essential part in transforming sound to meaning. Leveraging high-resolution intracranial recordings and a recurrent ASR model, this study bridges these gaps by uncovering a striking correspondence in the hierarchical encoding of linguistic features, from low-level acoustic signals to high-level semantic processing. Specifically, we demonstrate that neural activity in distinct regions of the auditory cortex aligns with representations in corresponding layers of the ASR model and, crucially, that both systems encode similar features at each stage of processing—from acoustic to phonetic, lexical, and semantic information. These findings suggest that both systems, despite their distinct architectures, converge on similar strategies for language processing, providing insight in the optimal computational principles underlying linguistic representation and the shared constraints shaping human and artificial speech processing.

[1] Department of Electrical Engineering, Columbia University, New York, NY, USA
[2] Zuckerman Institute, Columbia University, New York, NY, USA
[3] IBM Research, Yorktown Heights, NY, USA
[4] The Feinstein Institutes for Medical Research, Northwell Health, Manhasset, NY, USA
[5] Department of Neurosurgery, Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, USA
[*] Correspondence: nima@ee.columbia.edu

Introduction

Understanding how the human brain processes and encodes linguistic information is a fundamental challenge in neuroscience and artificial intelligence. The human auditory cortex is capable of extracting meaning and structure from spoken language with remarkable efficiency. In parallel, automatic speech recognition (ASR) systems have achieved near-human accuracy in recognizing and transcribing speech (*1*). However, it remains unclear how closely the internal computation and representation of these ASR systems mirror those of the human brain, leaving a significant gap in our understanding of speech processing in humans and machines. Are both systems converging toward similar strategies, or are they reaching different solutions independently as they optimize for performance?

Neuroimaging studies of speech processing in the brain have revealed an emergent encoding of linguistic hierarchies, progressing from primary to nonprimary areas of the auditory cortex (*2*–*4*). These studies show that the brain distributes, yet jointly encodes, various linguistic features, including phonemes, phonotactics, lexical-phonological, and lexical-semantic information. However, parallels between these hierarchical patterns in the brain and those found in ASR algorithms have not been directly established. Several studies have examined how end-to-end ASR systems represent linguistic information (*5*) and explored similarities between these systems and the brain (*6*). Additional research has also investigated the representational similarities between large language models (LLMs) and the brain (*7*–*11*). While these studies have provided valuable insights, they have some limitations. Many lacked the temporal precision necessary for speech processing, as they relied on functional MRI (fMRI) (*7*, *10*) which is slow in capturing the neural dynamics. Moreover, by comparing the brain to text-based LLMs, these studies inherently overlooked the acoustic stages of speech processing even though the subjects listened to the stimuli (*7*–*11*). Other studies used models that were biologically implausible, such as transformers (*6*–*11*) or non-causal architectures (*6*). In addition, some studies did not explicitly analyze linguistic representation in the models (*7*, *9*–*11*), or used only a narrow set of features (*6*, *8*). Therefore, even though it has been shown that speech processing models and the brain use increasingly similar representations as revealed by the predictability of neural responses

from these models, the precise nature of this convergent similarity remains unclear, replacing one black box with another one without explaining the full picture of speech understanding.

This study addresses several gaps in the current understanding of speech processing in both biological and artificial systems. First, we expand on previous studies by investigating a broad range of linguistic features, providing a comprehensive analysis that includes phonetic, lexical, and semantic representations. Second, we use a biologically plausible recurrent neural network transducer (RNN-T) (*12*) model that processes speech in a causal and incremental manner, aligning more closely with how the human brain processes speech in real time. Finally, by incorporating high-resolution intracranial electroencephalography (iEEG) data from participants listening to continuous speech, we establish a direct comparison between neural activations and the internal states of the ASR model.

Through a detailed node-level analysis, we compare specific neural sites in the brain with individual nodes in the ASR model to directly assess representational alignment. Additionally, a layer-level analysis provides a more comprehensive view of the ASR model's internal hierarchical representations, allowing us to examine how the structure of these representations parallels the cortical encoding of speech in humans. This dual approach enables us to uncover both fine-grained and hierarchical similarities in how speech is processed by biological and artificial systems, shedding light on shared mechanisms of linguistic encoding while also revealing key divergences.

Results

We recorded intracranial electroencephalography (iEEG) data from fifteen human participants implanted with subdural (electrocorticography; ECoG) and depth (stereotactic EEG) electrodes. The participants listened to 30 minutes of continuous speech spoken by four speakers. To ensure that the subjects were engaged in the task, we paused the audio at random intervals and asked the subjects to report the last sentence before the pause. All subjects were attentive and could correctly repeat the speech utterances. We extracted the envelope of the high-gamma frequency band (70-150 Hz), which has been shown to correlate with neural firing in the proximity of the recording electrode (*13*, *14*), as the neural response measure of the recorded signals.

We restricted our analyses to electrode sites in the auditory cortex (AC; $N = 335$) and the inferior frontal gyrus (IFG; $N = 191$). Figure 1A shows the general location of the IFG and the subregions of the auditory cortex on the FreeSurfer average brain (*15*). We further limited our analysis to sites that were determined to be speech-responsive ($N = 291/526$, determined by a t-test between responses during speech and in silence). We labeled the neural sites in both hemispheres based on their anatomical region to enable population tests: posteromedial Heschl's gyrus (pmHG; $N_L = 12/15$, $N_R = 16/19$), anterolateral Heschl's gyrus (alHG; $N_L = 34/36$, $N_R = 32/34$), planum temporale (PT; $N_L = 9/12$, $N_R = 18/20$), middle superior temporal gyrus (mSTG; $N_L = 44/53$, $N_R = 17/18$), posterior superior temporal gyrus (pSTG; $N_L = 27/32$, $N_R = 21/23$), anterior superior temporal gyrus (aSTG; $N_L = 19/60$, $N_R = 0/13$), and inferior frontal gyrus (IFG; $N_L = 32/119$, $N_R = 10/72$). The electrode locations and their responsiveness are plotted in Figure 1B on the average FreeSurfer brain, where the color indicates whether an electrode is speech-responsive.

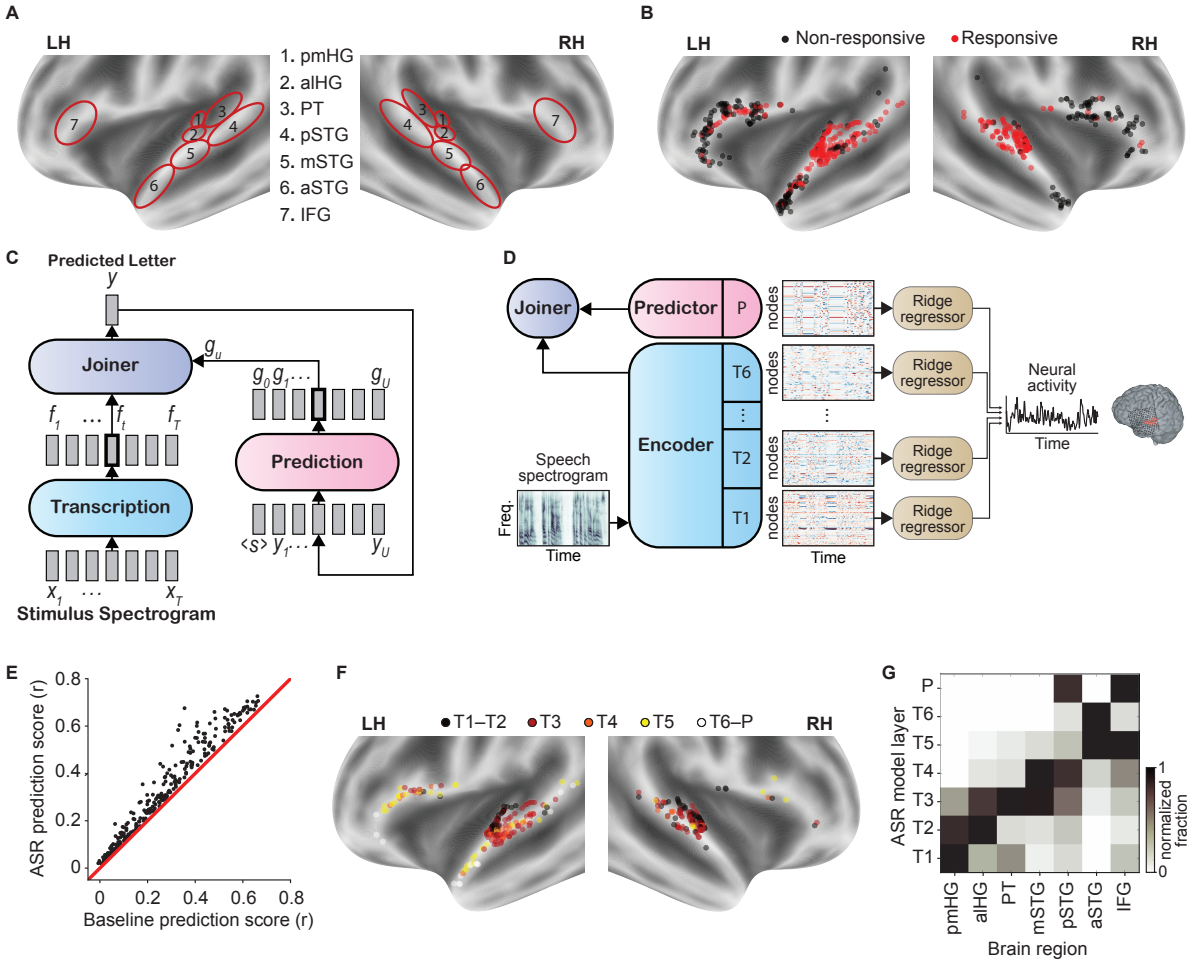

Figure 1. Stages of speech processing in the brain based on ASR modelling.

(A) General location of the inferior frontal gyrus (IFG) and subregions of the auditory cortex (AC). **(B)** Electrode locations within the region of interest. Colors represent whether the electrode showed higher activation in the presence of speech stimuli compared to moments of pre-trial silence. **(C)** RNN-Transducer architecture. $\boldsymbol{x_t}$ are spectrogram frames, $\boldsymbol{f_t}$ are the encoder's outputs, $\boldsymbol{h_{t,u}}$ are model outputs used to generate output labels, $\boldsymbol{y_u}$ are output labels predicted by the model (feedback), and $\boldsymbol{g_u}$ are the prediction network's outputs. **(D)** Predicting brain responses from the ASR model. To determine the mapping between network layers and electrode sites, we fit a Ridge regression model to predict the neural response at the electrode (r_S) while the participant was listening to some stimuli S, from the activation matrix $\Gamma_{i,S}$ corresponding to the response of the i-th layer of the ASR model to the same stimuli S. **(E)** Improvement in prediction score of electrode sites when using the best ASR layer compared to a

spectrogram baseline predictor. The best ASR layer for an electrode was determined by 10-fold cross-validation. **(F)** We color each electrode based on the layer of the ASR model that predicts it best. **(G)** Within each subregion of interest, we compute the fraction of electrodes that are best predicted by each layer of the ASR. We then normalize these fractions by dividing each row by its maximum value. This metric shows in which subregion of interest a given layer of the model is represented relatively stronger.

Stages of speech processing in the brain based on ASR modeling

We use an RNN-Transducer (RNN-T) (*12*), a recurrent neural network trained to predict letters (graphemes) from the speech spectrogram, to model the speech recognition process in the brain. RNN-T models take spectro-temporal speech signals as input through a causal mechanism, and have been widely used for speech recognition (*16–20*), making them an ideal candidate model to compare with the full hierarchy of speech processing in the brain. The model consists of two branches—transcription and prediction (Figure 1C). The transcription branch, consisting of 6 uni-directional LSTM layers (T1–T6), acts as a bottom-up encoder by producing acoustic embeddings from the input speech spectrogram. The prediction branch, consisting of 1 uni-directional LSTM layer (P), acts as an internal language model in that it is conditioned on previous non-blank symbols produced by the model. It stores an internal representation of the prior predictions of the model that together with the encoder's output are combined by a shallow joint network and used to make the next prediction. We computed layer activations for all layers of these two branches in response to the same 30-minute stimulus set that the human participants listened to.

To examine how the speech processing pathways of the brain and the model map to each other, we find the best matching model layer for each electrode site (Figure 1D). First, for each electrode-layer pair, we fit a single-lag regression model that predicts the electrode activity from the layer activations in response to the same stimuli (reduced to top 256 principal components):

$$y(t) = X(t - \delta)\beta + c + \varepsilon$$

where y is the $T \times 1$ response vector for an electrode site, X is the $T \times H$ matrix of layer activations, δ is a constant time lag, β is the $H \times 1$ vector of regression coefficients, c is a constant bias term and ε is the regression error. We also define a baseline 256-dimensional predictor which is obtained from a 128-dimensional log-Mel-spectrogram stacked on top of a 1-sample delayed version of itself, such that $X(t, 1{:}128) = S(t, :)$ and $X(t, 129{:}256) = S(t-1, :)$, where X is the $T \times 256$ baseline predictor and S is the $T \times 128$ spectrogram. For each electrode, we also fit a single-lag linear regression model that predicts the neural activity from the baseline predictor, allowing for a fixed time lag δ. Figure 1E shows the improvement of cross-validated prediction scores for each electrode when predicting from the ASR model compared to the baseline spectrogram predictor.

We associate each electrode with the layer of the model that predicts it with the highest accuracy. The colors in Figure 1F indicate the corresponding layer of the model for each neural site in the AC and IFG; and Figure 1G shows the normalized fraction of electrodes in each brain region best predicted by each layer of the model (values indicate fraction of neural sites in brain region, divided by sum of fractions within layer/row). The model layers map to the cortex such that as we move deeper in the model, in the cortex we move from the primary auditory cortex (pmHG) laterally to PT and mSTG, and from there to pSTG, aSTG, and IFG. The corresponding model depth (layer number) of an electrode is also correlated with various other metrics associated with being downstream in speech processing – relative prediction improvement from ASR over baseline (Spearman's $r = 0.582$, $p \ll 1e-4$), neural response latency ($r = 0.373$, $p \ll 1e-4$), neural site's distance from the primary auditory cortex (center of pmHG chosen as reference for primary AC; $r = 0.613$, $p \ll 1e-4$). These results together suggest a clear association between the biological and artificial processing pathways.

The brain results, however, show a difference in encoding between hemispheres, where neural sites corresponding to deeper layers of the ASR model are predominantly found in the left hemisphere. We can see this by a statistical comparison between the distribution of associated model depth of electrodes (one-sided two-sample t-test, $t = 6.31$, $p \ll 1e-4$, $df = 289$). More interesting is the difference by anatomical region, where as we move further away from the primary AC, lateralization increases – pmHG ($t = -0.183$, $p = 0.57$, $df = 26$), alHG ($t =$

-1.25, $p = 0.89$, $df = 64$), PT ($t = 2.57$, $p = 0.008$, $df = 25$), mSTG ($t = 2.27$, $p = 0.0134$, $df = 59$), pSTG ($t = 2.78$, $p = 0.0039$, $df = 46$), aSTG (no speech-responsive sites in the right hemisphere), IFG ($t = 2.98$, $p = 0.0024$, $df = 40$). These results are in line with previous studies showing left-lateralization of higher-order regions in linguistic processing (*21*, *22*): lower-level acoustic processing (according to ASR) is similar between hemispheres, while higher-level processing is biased towards the left hemisphere.

Similar node-level linguistic encoding across the brain and ASR

We used a regression-based method with temporal receptive fields (TRFs) (*23*) to measure the degree of encoding of different levels of linguistic information in individual nodes of the ASR model and individual neural sites (Figure 2A). In this method, we first predicted the neural activity of a site from a list of acoustic and linguistic features. Next, we determined the contribution of each feature in predicting the response by replacing it with a control variable and interpret the drop in predictive power as the contribution of that feature to the prediction. We repeated the replacement process for each linguistic feature 100 times and measured the t-statistic between actual and control features to determine the significance of the feature.

Figure 2B shows the normalized fraction of nodes in each ASR layer that significantly encode each linguistic feature group (fractions divided by sum across feature/row). Figure 2C shows the same, but for normalized fraction of neural sites in each brain region. Figure S1 show the average t-statistic without thresholding, normalized by the sum of each row (feature). We observed that the brain results show broader encoding of higher-level linguistic features in the nonprimary auditory cortex (PT and STG) compared to the primary auditory cortex (HG). This trend is also mirrored in the ASR model. These trends do not depend on the specific choice of the threshold value, as we can see a similar result using average t-values of groups instead of fractions.

Figure 2D shows the temporal order of the encoding of these linguistic features in the ASR model and the brain. To measure the encoding latency for each feature, we first select the group of nodes/electrodes that significantly encoded that feature. For each node/electrode, we compute the latency of the peak absolute value of the TRF weight associated with that feature. We average the latencies to obtain a single latency value for each feature. Finally, using bootstrapping, we

obtain 95% confidence intervals for the average latency of each feature. In line with prior findings in the human brain (*24*), the results show a particular temporal order of emergence for the different levels of linguistic information in both systems.

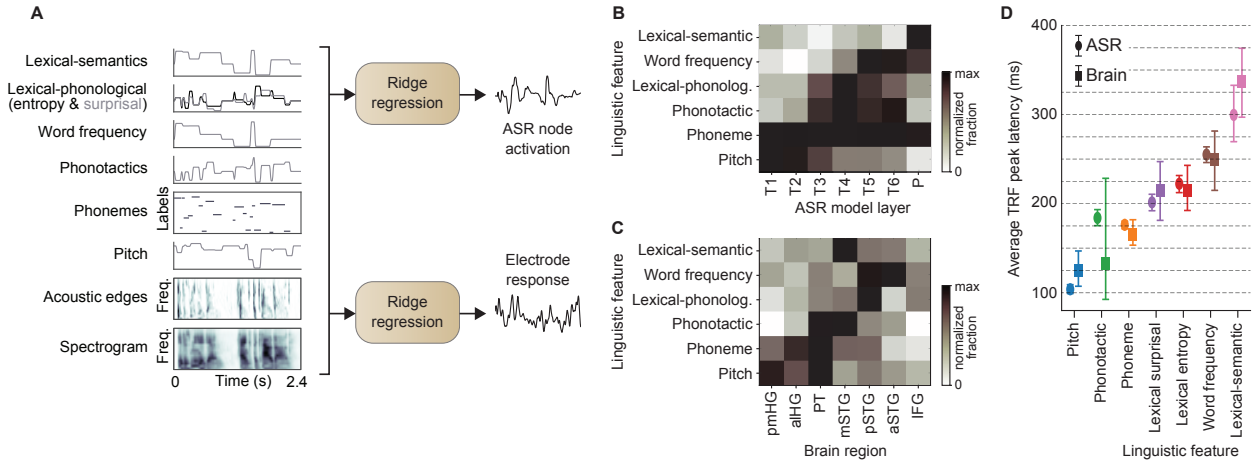

Figure 2. Similar node-level linguistic encoding across the brain and ASR.

(A) Determining node-level linguistic encoding. To measure which linguistic features are encoded in a given electrode site or ASR node, we use a similar approach to [1]. We first predict the neural response r_S to the stimuli S from a set of time-aligned acoustic-linguistic features $\mathcal{F}_S = \{f_{j,S}\}$, where $f_{j,S}$ is the j-th linguistic feature of the stimulus S. Then to find if feature f_j is encoded in the neural response, we replace it with $\widehat{f_j}$ and measure the difference between prediction score between the true and control cases. We repeat this process 100 times and compute a statistic of significance of encoding. **(B and C)** Node-level linguistic encoding in ASR layers and brain regions. The values denote fraction of nodes/electrode in a layer/region significantly encoding a given feature. Fractions are normalized by dividing each row by the maximum. **(D)** Temporal order of node-level linguistic encoding in the brain and ASR. Error bars indicate 95% confidence intervals obtained using bias-corrected and accelerated bootstrap (BCa; $N = 1000$) on the mean latency of encoding across nodes/electrodes.

Population-level linguistic encoding in the ASR

While the regression approach enabled us to compare the results between the two systems, it cannot capture population-level representations. This is especially relevant in the case of the ASR model, since we have full access to the entire neural population. To determine the patterns of linguistic encoding in the model, we decoded each linguistic feature from the activations of the model (Figure 1A). Because linguistic features are defined at different resolutions and linguistic units have different lengths, we performed a unit-aligned analysis. For example, to predict linguistic features that are defined at phoneme resolution—pitch, phoneme, biphone probability, lexical entropy and surprisal—we extracted a layer's activations at time points $\{p_i + \delta\}$, where $\{p_i\}$ were the time points corresponding to phoneme centers and δ was a constant time lag. Similarly, for word resolution features—word frequency, semantic neighborhood density, contextual embedding—we extracted activations at time points $\{w_i + \delta\}$, where $\{w_i\}$ were the time points corresponding to word centers and δ was a constant time lag.

For predicting the categorical phoneme labels we fit linear Ridge classifiers, and for the rest of the features, Ridge regressors (*24*). The best time lag (δ) and regularization parameter (λ) were determined independently for each layer-feature pair by maximizing the 10-fold cross-validated prediction scores.

We also decoded each linguistic feature from the 256-dimensional baseline predictor described earlier. Figure 3B shows the prediction scores of predicting each feature from each layer of the model, normalized per feature such that zero corresponds to the baseline prediction score and 1 corresponds to the best score across all layers of the model. We can see that as we move deeper into the model, representations of higher-order linguistic information emerge. Based on these scores, we can associate each linguistic feature with a layer of the model and find the best lag for that layer-feature pair. As a result, we can describe the place of a linguistic feature in the speech recognition process by its time lag and layer depth (Figure 3C). Together, these results show a

pronounced emergent linguistic representation in the model through time and space and with a specific order, enabling a direct comparison with the results observed in the brain.

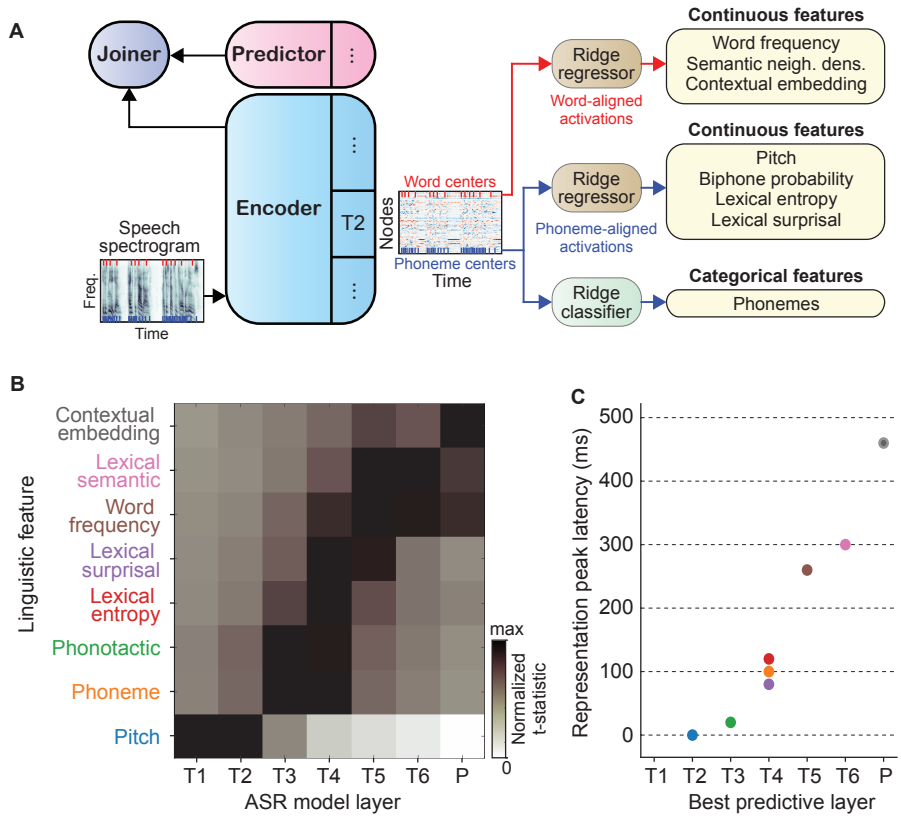

Figure 3. Population-level linguistic encoding in the ASR.

(A) Decoding linguistic information in the ASR model. To measure how strongly linguistic features are encoded at the population-level (layer embedding), we decode the features from the activations of the different layers of the ASR. We use a single-lag Ridge regressor or classifier. **(B)** Population-level linguistic encoding in different layers of the model. Shown values are min-max normalized between baseline (spectrogram) prediction score and maximum prediction score across layers. **(C)** Temporal order of population-level linguistic encoding in the ASR. For each linguistic feature, we find the layer of the model that best predicts it, then

similarly determine the best time lag where the feature was maximally predictable (see Methods).

Effect of model training on linguistic representation

Finally, we explored the conditions in which such a linguistic representation can emerge in the model. We tested two hypotheses: the emergent representation is training-dependent, and the emergent representation is language-dependent. To do this, we decoded the same list of English linguistic features from three different bi-directional RNN-T models: one trained on English, one on French, and a randomly initialized model which underwent no training (Figures 4A–C). We used bi-directional models for this analysis because we do not directly compare the models with the brain, and bi-directional models perform better and are more commonly used in ASR tasks. Additionally, the layers of the bi-directional models have more consistent latency which makes direct comparison between them more accurate, as we do next. The results show a stark difference between the English and random models, such that the random model does not show a strong linguistic representation for any feature compared to the baseline, especially for deeper layers. The French model shows an identical representation of pitch compared to the English model and a relatively strong representation of only phonemes, phonotactics, and word frequency compared to the random model.

We can also compare these models more directly by computing the similarities of their representations to the same stimulus using centered kernel alignment (CKA) (*25*). We compared the representations of each layer of the English model to all layers of all three models (Figures 4D–F). We observed that the English and French models share similar representations of their first 3 layers to a large degree, but steadily decrease in similarity over layers overall. In contrast, the random model has weak similarity to the English model, even from early layers. Put together, these results show that the emergent linguistic representation observed in the RNN-T is both training-dependent and language-dependent.

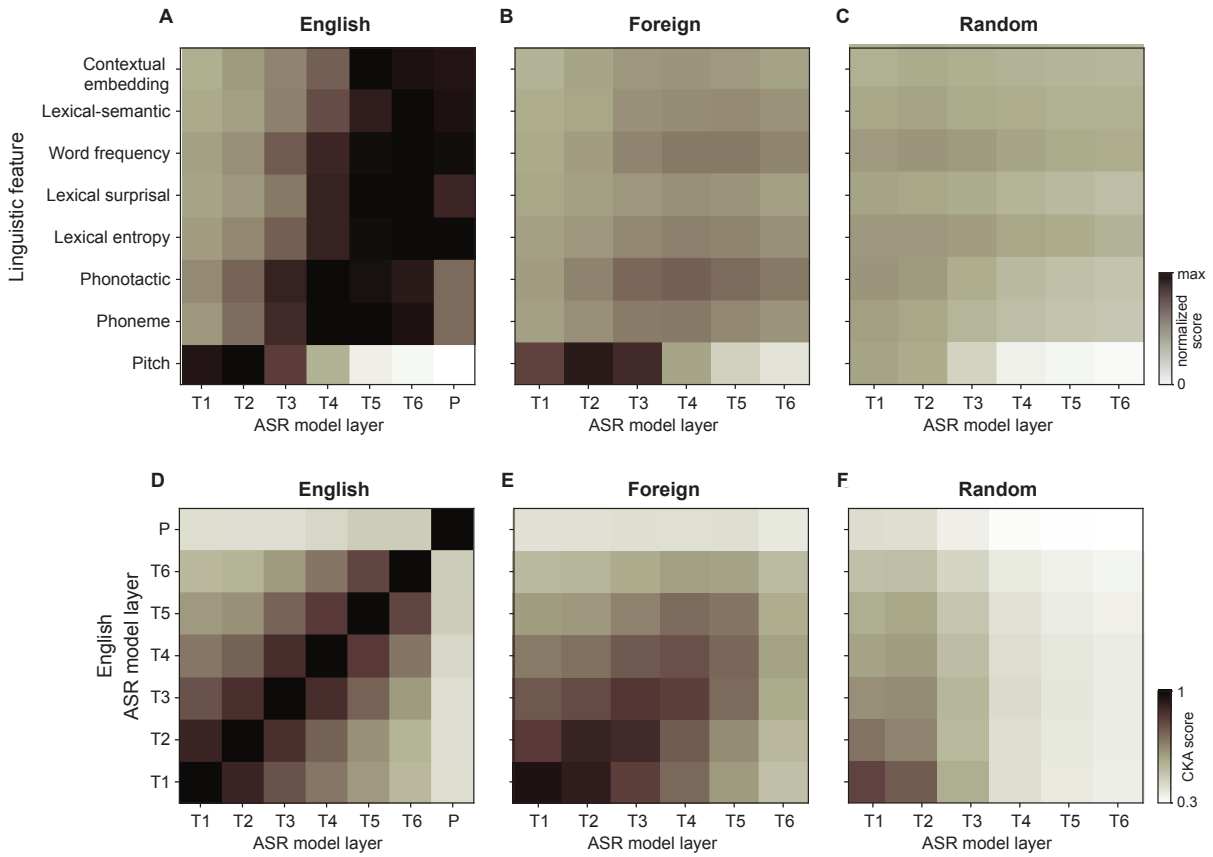

Figure 4. Effect of model training on linguistic representation.

(A–C) Linguistic encoding in control ASR models. We perform the same analysis as in Figure 3 on three bi-directional RNN-Transducers–a model trained to recognize English speech, a model trained to recognize French ("foreign"), and a randomly initialized model ("random"). For the foreign and random models we discard the prediction branch (P) activations, as they fail at transcribing speech so their top-down branch cannot be aligned to the bottom-up branch. Values are min-max normalized between baselined (spectrogram) predictions score and maximum prediction score across all layers of all models. **(D–F)** Pairwise representation similarity analysis between ASR model layers. Each of the three panels shows the centered

kernel alignment (CKA) similarity between the activations of each of its layers and each of the layers of the English model.

Discussion

This study presents a detailed comparison of the linguistic representations in the human brain and an ASR model, specifically focusing on the hierarchical encoding of speech. By using a biologically plausible RNN-Transducer model, we demonstrate similarities and differences in how linguistic features are processed across systems. By selecting a model architecture that mirrors the causal and incremental nature of human speech processing, and incorporating a biologically plausible RNN architecture which processes the input at the spectrogram-level rather than the word-level, we provide a fair ground for comparing how each system encodes various levels of linguistic information, separating this study from prior works using transformer-based large language models or non-causal architectures (*6–11*). Our findings show a striking alignment between specific brain regions and ASR model layers. Both systems exhibit a hierarchical progression in the encoding of linguistic information, moving from low-level acoustic analysis to higher-level semantic understanding. The correlations between layers and neural regions reconstructs the dorsal and ventral streams of speech processing (*26–28*), with increasing alignment with deeper layers. Specifically, we find increasing correlation with deeper layers of the model along the dorsal pathway from mSTG to pSTG and finally to IFG, as well as the ventral pathway from mSTG to aSTG and IFG. The fact that IFG most strongly matches the later layers and particularly the prediction branch of the ASR model, which receives feedback from prior unified representations, suggesting its role in unifying lexical information from previous stages into a single high-level representation, as has been suggested by prior research (*29*). Our findings of the brain's hierarchical representations support prior studies of the neural hierarchy of linguistic representations (*26*, *30–32*), and we advance this knowledge by highlighting the specific regions and representational latencies at each stage of this hierarchy which are common across both the human brain and the ASR model. This functional congruence underscores the shared

mechanisms of linguistic processing in both the brain and ASR systems, revealing that artificial systems trained to emulate human behavior can mirror aspects of biological speech recognition across the full hierarchy of linguistic feature extraction.

Our analysis of ASR models with foreign language training or no training also yielded new insights into how the linguistic feature hierarchy forms. We showed that in the French-trained model, only acoustic features like pitch were strongly represented, with higher-level linguistic representations of the English stimuli being weakly encoded, and features at all levels being weakly encoded in the random model. When comparing the representations of the models to each other, we found that the layer representations of the French model quickly diverged from those of the English model after the first two layers, demonstrating where the effects of training begin to compound. One recent study compared the processing of Chinese and English speech by unsupervised transformer speech representation models and the brain (*6*). They found that models trained on one language better predicted neural responses recorded in STG from native speakers of that same language, whether English or Chinese, and that this performance difference increased over the model's layers. Our results support this and offer a more granular explanation by analyzing the encoding of higher-level linguistic features in biologically-plausible models and over deeper regions in the brain's language processing pathway, showing that models trained on different languages (e.g. French) or untrained models do not exhibit the same level of linguistic representation for English speech as a native English model. These findings suggest that high-level linguistic representations are not simply a byproduct of deeper layers processing longer segments of speech (*33*) but are instead tightly linked to the model's training. A closer look at the training process, which is akin to language acquisition in humans, could be utilized to study the development of spoken language processing in infants or second language learners.

Despite these similarities, the RNN-T ASR model lacks the brain's top-down feedback connections and flexible, multi-directional communication between regions (*34*, *35*). While the ASR model processes speech in a non-lateralized fashion, our results show that the brain shows a strong left-hemisphere bias for higher-level linguistic tasks, in accordance with prior studies (*21*, *22*). This lateralization highlights a key divergence between the two systems, which the current ASR

339 architecture cannot fully account for. These architectural differences suggest that while ASR
340 models can emulate aspects of human speech processing, more biologically accurate models are
341 needed to capture the full complexity of the brain's hierarchical and interactive mechanisms.
342 Future models incorporating binaural input, top-down feedback, and multitask objectives could
343 provide deeper insights into the brain's speech processing pathways. Additionally, ASR models
344 are trained with only a speech recognition loss, forcing them to develop speaker-independent
345 features which likely remove the speaker-specific information such as pitch, matching our finding
346 that pitch is only strongly encoded in the early layers. This may influence the hierarchy of
347 representations formed over layers and represents a point of divergence between ASR models
348 and the brain (*36*, *37*). Nonetheless, the insight that a high-performing ASR system represents
349 linguistic features, from low-level acoustic to high-level contextual features, in a similar
350 hierarchical manner to the brain suggests that similar recent parallels shown between language
351 models and the brain (*7–11*) can arise in a system that is performing the same overall task as a
352 human listener, from input sound waveform to output words.

353 This study demonstrates a detailed correspondence between linguistic representations in the
354 human brain and an RNN-Transducer model. The RNN-T's causal and incremental processing
355 mirrors key aspects of human speech processing, making it a reasonable proxy for investigating
356 brain-like hierarchical encoding. While our findings are specific to this model, they likely reflect
357 broader principles of linguistic processing shared across systems. The hierarchical progression of
358 encoding observed here, from low-level acoustic to high-level semantic features, is consistent
359 with patterns identified in prior studies using other architectures (*38–40*), suggesting robustness
360 in the hierarchical nature of linguistic representation even if layer-to-region mappings vary.
361 Crucially, our findings are grounded in empirical data from the human brain, providing a
362 benchmark for comparison and aligning with known principles of neural processing, such as the
363 ventral and dorsal streams of speech (*27–29*) and left-hemisphere lateralization (*21*, *22*). Future
364 work should explore whether these similarities extend to other biologically plausible ASR
365 architectures to disentangle architecture-specific phenomena from universal principles. While
366 model-specific details may vary, this study highlights that biologically inspired architectures can

reveal shared computational strategies between the brain and artificial systems, offering insights into the constraints and mechanisms underlying human and machine language processing.

In conclusion, this study bridges neuroscience and artificial intelligence by revealing how both systems process speech in hierarchical layers. Our findings not only advance our understanding of the brain's mechanisms for linguistic encoding but also offer valuable guidance for the development of more sophisticated and biologically informed ASR models. Moving forward, integrating more brain-like features into artificial systems may unlock new possibilities for understanding and enhancing human and machine communication.

Methods

Data

Participants, neural data, task, and stimuli

Fifteen human patients (7 female, mean age: 36, range: 19-58) with drug-resistant epilepsy were studied. All patients were implanted with intracranial electroencephalography (iEEG) electrodes for epileptogenic foci localization. Twelve of the patients has stereoelectroencephalographic (sEEG) depth electrodes, while the other three had both depth electrodes and subdural grids and/or strips. All recordings were inspected by an epileptologist to ensure they were free of interictal spikes. The patients provided written, informed consent to participate in the research study prior to implantation, and the protocol was approved by the institutional review board at the Feinstein Institutes for Medical Research.

The subjects listened to approximately 30 minutes of stories spoken by voice actors. Occasional pauses in the story were added, resulting in 53 trial segments, and the subjects were asked to repeat the most recent sentence before the pause to ensure they were paying attention. All subjects were able to repeat the sentences without issue. iEEG signals were acquired at 3 kHz sampling rate, and the envelope of the high-gamma band (70-150 Hz) was extracted with the Hilbert transform (*41*). This signal was then z-scored and resampled to 100 Hz.

Electrode selection

Electrodes were projected onto the nearest cortical surface. We selected all electrode sites within the auditory cortex (AC; $N = 335$) and the inferior frontal gyrus (IFG; $N = 191$). Figure 1A shows the general location of IFG and the subregions of the auditory cortex on the FreeSurfer average brain (*15*).

To determine whether an electrode site was speech-responsive, we first selected 53 pre-trial silence segments ([-1 s, 0 s] relative to segment onset) and using a 200 ms wide non-overlapping moving average window reduced it into 53×5 data points. We then selected 53 post-onset speech segments ([0.4 s, 1.2 s] relative to segment onset) and using a similar moving average window reduced it into 53×4 data points. We performed a two-sample t-test between the two

403 distributions for each electrode to obtain a p-value, and then performed a Benjamini-Hochberg
404 (false discovery rate) correction with an alpha of 0.05 to determine the speech-responsive
405 electrodes. This method was observed to be more robust compared to some alternatives when
406 viewing across the entire regions of interest which included electrodes inside and outside the
407 auditory cortex with different response latencies.

408 We limited our analysis to sites that were determined to be speech-responsive ($N = 291/526$).
409 We labeled the neural sites in both hemispheres based on their anatomical region to enable
410 population tests: posteromedial Heschl's gyrus (pmHG; $N_L = 12/15$, $N_R = 16/19$), anterolateral
411 Heschl's gyrus (alHG; $N_L = 34/36$, $N_R = 32/34$), planum temporale (PT; $N_L = 9/12$, $N_R =$
412 $18/20$), middle superior temporal gyrus (mSTG; $N_L = 44/53$, $N_R = 17/18$), posterior superior
413 temporal gyrus (pSTG; $N_L = 27/32$, $N_R = 21/23$), anterior superior temporal gyrus (aSTG; $N_L =$
414 $19/60$, $N_R = 0/13$), and inferior frontal gyrus (IFG; $N_L = 32/119$, $N_R = 10/72$). The electrode
415 locations and their responsiveness are plotted in Figure 1 on the average FreeSurfer brain, where
416 the color indicates whether an electrode is speech-responsive.

417

418 **RNN-Transducer**

419 To model the speech processing mechanism, we use an RNN-Transducer (RNN-T) (*12*), a
420 recurrent neural network trained to perform automatic speech recognition, i.e., predicting
421 graphemes from the speech spectrogram. The model consists of two branches—transcription
422 and prediction (Figure 1C). The transcription branch, consisting of six uni-directional LSTM layers
423 of 640 nodes each (T1–T6), acts as a bottom-up encoder. It transforms the input speech
424 spectrogram into a representation used to predict the output labels. The prediction branch,
425 consisting of a single uni-directional LSTM layer of 1024 nodes (P), acts as an internal language
426 model. It stores an internal representation of the prior predictions of the model that, together
427 with the encoder's output, are used to make the next prediction. A joint network merges the final
428 transcription branch layer with the prediction branch's embedding and predicts text characters
429 to generate the speech transcription. Bi-directional versions of the model were used for the
430 analysis of training language dependencies since uni-directional models trained on French were

not available. The English network outputs are projected to 42 logits, corresponding to 41 characters plus BLANK. Similarly, for French, we use 72 output units. The model is trained with the RNN-T loss on the Switchboard and Fisher datasets, which, put together, consist of about 2,000 hours of English phone conversations (see (*42*) for training details). An additional 4-fold data augmentation was applied to the input spectrograms to allow better generalization of the model, bringing the duration of the unique training data to 10,000 hours.

We computed layer activations for all layers of the model in response to the same 30-minute stimulus set that the human participants listened to.

Time-aligning RNN-Transducer activations to neural data

The processed neural data is sampled at 100 Hz and has shape $T \times C$, where T is the number of time steps and C is the number of electrode channels. The activations of the layers in the transcription branch of the RNN-T models are sampled at 50 Hz and have the shape $\left\lfloor \frac{T}{2} \right\rfloor \times H_T$, where H_T is the number of nodes in each hidden transcription layer. These activation matrices can be trivially aligned with the neural data sampled through a resample with the integer factor 2. The activations of the prediction branch, however, have shape $U \times H_P$, where U is the number of output graphemes and H_P is the number of nodes in the hidden prediction layer. These activation matrices cannot be directly aligned with the neural data since they are based on graphemes instead of time. To make this alignment possible, we need to find the most likely "time warping" between output grapheme indices ($1 \leq u \leq U$) and time ($1 \leq t \leq \left\lfloor \frac{T}{2} \right\rfloor$). We run the forward-backward algorithm defined in section 2.4 of Graves (*12*) on the grapheme output probability lattice produced by the model to obtain this "time warping" between the two sequences. We then use the $\alpha: t \rightarrow u$ alignment to stretch the activation matrix Z_P with shape $U \times H_P$ into a matrix Z'_P with shape $\left\lfloor \frac{T}{2} \right\rfloor \times H_P$ where $Z'_P[t, :] = Z_P[\alpha(t), :]$. Since Z'_P has the same shape as transcription layer activations, we can trivially align it to the neural data.

Linguistic features of speech

Pitch

As a measure of pitch, we computed the pitch contour (F0) of the speech signal using the PyWORLD python package, which is a python wrapper for the WORLD vocoder (*43*). We then averaged the value across the duration of each phoneme to obtain a phoneme-average pitch.

Phonemes

For phonetic features, we used the categorical (one-hot encoded) representation of ARPAbet phonemes. We chose this because it allows classification in the layer-level analysis.

Phonotactic features

Phonotactics represent phoneme transition probabilities, so we used the biphone probability P_{ab} for phoneme bigram ab:

$$\boldsymbol{P_{ab}} = \frac{\mathbf{freq_{ab}}}{\sum_{p \in S} \mathbf{freq_{ap}}}$$

where S is the set of all ARPAbet phonemes. To compute the frequencies of each biphone, we used the CMU dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict) to convert words to phoneme sequences and counted the occurrence of each biphone using the SUBTLEX-US corpus, which is an English word frequency dataset calculated from movie subtitles (*44*). Since 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/. The frequency of a phoneme bigram represents the degree of exposure of an average native listener to the 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 (*45*)) to maximally dissociate this effect from lexical processes. This level represents the expectation and surprisal of the listener when hearing a new

482 phoneme, based on the immediate past. This prelexical phonotactics feature could indicate
483 predictive coding mechanisms that operate on the phonemic level (*46–49*).

484 **Lexical-phonological features**

485 To measure the lexical-phonological effect, we used lexical entropy and surprisal. These values
486 were calculated for each phoneme within a word from the previous phonemes in that word. The
487 surprisal caused by phoneme φ_i, $S(i)$, in word $w=\varphi_1...\varphi_K$ indicates the improbability of hearing
488 phoneme φ_i based on the previous i−1 phonemes that came before it in the word and is
489 calculated as follows:

490 $$S(i) = -\log_2 \frac{\mathrm{freq}(\mathrm{cohort}_i)}{\mathrm{freq}(\mathrm{cohort}_{i-1})}$$

491 where $\mathrm{freq}(\mathrm{cohort}_i)$ is the summed frequency of all words that start with the phoneme
492 sequence $\varphi_1...\varphi_i$. On the other hand, the lexical entropy, E(i), for phoneme φ_i, is the entropy
493 within all words that start with the phoneme sequence $\varphi_1...\varphi_i$ (the cohort) (*50*):

494 $$\mathrm{E(i)} = -\sum_{\mathrm{word} \in \{\mathrm{cohort}_i\}} \mathrm{p(word)} \log_2 \mathrm{p(word)}$$

495 where $\mathrm{p(word)}$ indicates the relative frequency of the word within the cohort. These two
496 parameters together encode the incremental lexical competition among all phonologically
497 consistent candidates as a word is being heard, weighted by their frequency. To compute lexical
498 surprisal for the word-initial phoneme, we assumed a transition from the entire lexicon, i.e., how
499 surprising it is to hear a word starting with phoneme p given the frequencies of all the words in
500 the lexicon.

501 **Lexical-semantic features**

502 To study the encoding of semantic information, we represented each word with its semantic
503 neighborhood density (SND) obtained from the English Lexicon Project (*51*, *52*), which refers to
504 the relative distance between a word and its closest neighbors based on a global co-occurrence
505 model (*51*, *52*). The neighborhood density can encode the degree of activation of semantically
506 related words in the lexicon upon hearing the target word.

507

Contextual embedding

We used the embedding obtained from the last hidden layer of a pre-trained GPT-2 XL model obtained from Hugging Face (*53*). This 1.5B parameter version of GPT-2, a transformer-based language model was pretrained using a causal language modeling (CLM) objective on English language data. CLM is a training goal where the model predicts the next token in a sequence given its preceding tokens, ensuring that the prediction for a position can only depend on known outputs at previous positions.

To associate contextual embedding to words in the analysis data, we concatenated the transcripts for all 53 trials, in the same order the participants heard them. Then, for each word, we gave the model the last token of the target word and the 511 tokens preceding it (if available) to obtain a contextual embedding for that word. The total 512 tokens are half the maximum context window size of GPT-2 XL and correspond to about 3 minutes of context in the experiment data.

Predicting neural responses from ASR activations

We predict the neural activity recorded at each electrode from each of the layers of the ASR model (Figure 1). We map ASR layers instead of ASR nodes to electrodes, because the high-gamma envelope of an iEEG electrode represents a readout of the activity of neighboring neural populations consisting of thousands of neurons rather than individual neurons.

To predict a neural response from a layer of the model, we fit a linear single-lag Ridge regressor that predicts the electrode activity from the layer activations in response to the same stimuli. Discounting a constant lag term δ, the output at time t is predicted only from the input at time $t + \delta$. In other words, $y = A_i^{+\delta}\beta + \varepsilon$, where y is the $T \times 1$ neural response vector, $A_i^{+\delta}$ is the $T \times H$ activation matrix for layer i of the ASR model with H nodes shifted by δ time steps across its first dimension, β is the $H \times 1$ regression model that maps $A_i^{+\delta}$ to y, and ε is the regression error.

We use the Ridge regressor from the python scikit-learn package to fit these models. The optimal lag value (δ) and ridge regularization parameter (λ) are chosen independently for each electrode-layer pair by maximizing the prediction score through 10-fold cross-validation.

Predicting neural responses and ASR activations from linguistic features

We used a temporal receptive field (TRF) (*23*) framework to measure the extent to which different linguistic features are encoded in the neural responses recorded from the brain and activations extracted from the ASR nodes (Figure 2). We used a broad spectrum of acoustic and linguistic features as predictors: spectrogram, acoustic edges, pitch (phoneme-average F0), phonemes, phonotactics ($-\log P_{ab}$), word frequency, lexical-phonological (lexical entropy and lexical surprisal), and lexical-semantics (semantic neighborhood density). The enormous dimensionality of the contextual embedding feature prevented its use in this framework.

To produce a distribution of shuffled models for statistics, we replaced each linguistic feature with a shuffled version—one at a time—100 times and measured a t-statistics of encoding, using $t = 19$ as the threshold of significant encoding. Special care was taken for each type of feature to ensure that the shuffling did not alter the influence of lower-level features in the hierarchy, as described below.

Pitch features

We grouped all the words within our 30-minute data based on their number of phonemes. Then we shuffled the word-to-pitch sequence association map within each group.

Phonetic features

We took the CMU 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 will have a consistent pronunciation at ever occurrence, but words that share phonemes will have independent pronunciations, e.g. /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 the same length are more similar in frequency of occurrence (i.e. 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

To generate controls for phonotactic features, we shuffled the bigram-to-frequency associations (i.e. 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 vs. non-first phoneme grouping, we perform the above shuffling separately for first phones (ones starting with #) and non-first biphones, so that any first vs. non-first effect will be maintained in the control, and thus discounted.

Word frequency

We grouped words based on their phoneme length and shuffled the word-to-frequency associations within each group.

Lexical-phonological features

We grouped all cohorts based on 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, i.e. if two words share their first k phonemes, the cohort information for their first k positions would be the same because the same cohorts are mapped to the same information.

Lexical-semantic features

The control for the semantic condition was constructed by grouping words based on their phoneme length and shuffling the word-to-SND associations within each group.

Predicting linguistic features from ASR activations

We predicted different linguistic features from each of the layers of the ASR model. Because linguistic features are defined at different resolutions and linguistic units have different lengths, we performed a unit-aligned analysis. For example, to predict linguistic features that are defined at phoneme resolution—pitch, phoneme, biphone probability, lexical entropy and surprisal—we extracted a layer's activations at time points $\{p_i + \delta\}$, where $\{p_i\}$ were the time points corresponding to phoneme centers and δ was a constant time lag. This $N \times H$ activation matrix, where N is the number of phonemes and H is the number of nodes, was multiplied by an $H \times D$ linear decoder to predict the D-dimensional linguistic feature. Similarly, for word resolution features—word frequency, semantic neighborhood density, contextual embedding—we extracted activations at time points $\{w_i + \delta\}$, where $\{w_i\}$ were the time times corresponding to word centers and δ was a constant time lag.

For predicting the categorical phoneme labels we fit linear Ridge classifiers, and for the rest of the features Ridge regressors. The best time lag (δ) and regularization parameter (λ) was determined independently for each layer-feature pair by maximizing the 10-fold cross-validated prediction scores.

Representation similarity of ASR model layers

We use two methods to compare representations of two ASR layers. The first is to compare their linguistic decoding results. The second is to compare the representations directly with the centered kernel alignment (CKA) method, which is used to measure the similarity between two sets of high-dimensional vectors, especially representations learned by neural networks (*25*). For this comparative analysis, we used three bi-directional LSTM models with a similar architecture to the causal one described earlier: a model trained on the same English data; a "random" model that has the same architecture as the English model, but randomly initialized and untrained; a "foreign" model that was trained on a French ASR task and was not exposed to English. We chose bi-directional models for the between-model analysis because the layers of the bi-directional models have more consistent latency which makes the CKA layer comparison more accurate.

Code availability

Code for preprocessing neural data, selecting responsive electrodes, and creating brain plots is available in the naplib-python package (*54*).

Acknowledgements

This study was funded by the National Institute on Deafness and Other Communication Disorders. S.B. was also supported by the National Institute on Deafness and Other Communication Disorders, R01DC019979. The funders had no role in the study design, data collection and analysis, decision to publish and manuscript preparation.

Competing interests

The authors declare no competing interests.

References

1. A. Stolcke, J. Droppo, Comparing Human and Machine Errors in Conversational Speech Transcription. arXiv arXiv:1708.08615 [Preprint] (2017). https://doi.org/10.48550/arXiv.1708.08615.
2. I. DeWitt, J. P. Rauschecker, Phoneme and word recognition in the auditory ventral stream. *Proc. Natl. Acad. Sci.* **109**, E505–E514 (2012).
3. C. J. Price, The anatomy of language: a review of 100 fMRI studies published in 2009. *Ann. N. Y. Acad. Sci.* **1191**, 62–88 (2010).
4. D. Poeppel, The neuroanatomic and neurophysiological infrastructure for speech and language. *Curr. Opin. Neurobiol.* **28**, 142–149 (2014).
5. Y. Belinkov, J. Glass, "Analyzing Hidden Representations in End-to-End Automatic Speech Recognition Systems" in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2017; https://proceedings.neurips.cc/paper_files/paper/2017/hash/b069b3415151fa7217e870017374de7c-Abstract.html)vol. 30.
6. Y. Li, G. K. Anumanchipalli, A. Mohamed, P. Chen, L. H. Carney, J. Lu, J. Wu, E. F. Chang, Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nat. Neurosci.* **26**, 2213–2225 (2023).
7. C. Caucheteux, A. Gramfort, J.-R. King, Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat. Hum. Behav.* **7**, 430–441 (2023).
8. C. Caucheteux, J.-R. King, Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5**, 134 (2022).
9. A. Goldstein, Z. Zada, E. Buchnik, M. Schain, A. Price, B. Aubrey, S. A. Nastase, A. Feder, D. Emanuel, A. Cohen, others, Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).
10. E. A. Hosseini, M. Schrimpf, Y. Zhang, S. Bowman, N. Zaslavsky, E. Fedorenko, Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *BioRxiv*, 2022–10 (2022).
11. G. Mischler, Y. A. Li, S. Bickel, A. D. Mehta, N. Mesgarani, Contextual feature extraction hierarchies converge in large language models and the brain. *Nat. Mach. Intell.*, 1–11 (2024).
12. A. Graves, Sequence Transduction with Recurrent Neural Networks. arXiv arXiv:1211.3711 [Preprint] (2012). https://doi.org/10.48550/arXiv.1211.3711.
13. S. Ray, J. H. Maunsell, Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biol.* **9**, e1000610 (2011).
14. M. Steinschneider, Y. I. Fishman, J. C. Arezzo, Spectrotemporal analysis of evoked and induced electroencephalographic responses in primary auditory cortex (A1) of the awake monkey. *Cereb. Cortex* **18**, 610–625 (2008).
15. B. Fischl, A. Van Der Kouwe, C. Destrieux, E. Halgren, F. Ségonne, D. H. Salat, E. Busa, L. J. Seidman, J. Goldstein, D. Kennedy, others, Automatically parcellating the human cerebral cortex. *Cereb. Cortex* **14**, 11–22 (2004).

16. Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S. Chang, K. Rao, A. Gruenstein, "Streaming End-to-end Speech Recognition for Mobile Devices" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019; https://ieeexplore.ieee.org/abstract/document/8682336), pp. 6381–6385.
17. K. Rao, H. Sak, R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer" in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (2017; https://ieeexplore.ieee.org/abstract/document/8268935), pp. 193–199.
18. J. Li, R. Zhao, H. Hu, Y. Gong, "Improving RNN Transducer Modeling for End-to-End Speech Recognition" in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (2019; https://ieeexplore.ieee.org/abstract/document/9003906), pp. 114–121.
19. L. E. Shafey, H. Soltau, I. Shafran, Joint Speech Recognition and Speaker Diarization via Sequence Transduction. arXiv arXiv:1907.05337 [Preprint] (2019). https://doi.org/10.48550/arXiv.1907.05337.
20. M. Ghodsi, X. Liu, J. Apfel, R. Cabrera, E. Weinstein, "Rnn-Transducer with Stateless Prediction Network" in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020; https://ieeexplore.ieee.org/abstract/document/9054419), pp. 7049–7053.
21. C. McGettigan, S. Evans, S. Rosen, Z. K. Agnew, P. Shah, S. K. Scott, An Application of Univariate and Multivariate Approaches in fMRI to Quantifying the Hemispheric Lateralization of Acoustic and Linguistic Processes. *J. Cogn. Neurosci.* **24**, 636–652 (2012).
22. R. J. Zatorre, A. C. Evans, E. Meyer, A. Gjedde, Lateralization of Phonetic and Pitch Discrimination in Speech Processing. *Science* **256**, 846–849 (1992).
23. M. J. Crosse, G. M. Di Liberto, A. Bednar, E. C. Lalor, The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* **10**, 604 (2016).
24. M. Keshishian, S. Akkol, J. Herrero, S. Bickel, A. D. Mehta, N. Mesgarani, Joint, distributed and hierarchically organized encoding of linguistic features in the human auditory cortex. *Nat. Hum. Behav.* **7**, 740–753 (2023).
25. S. Kornblith, M. Norouzi, H. Lee, G. Hinton, "Similarity of neural network representations revisited" in *International Conference on Machine Learning* (PMLR, 2019), pp. 3519–3529.
26. G. Hickok, D. Poeppel, The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
27. G. Hickok, D. Poeppel, Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* **92**, 67–99 (2004).
28. A. D. Friederici, The cortical language circuit: from auditory perception to sentence comprehension. *Trends Cogn. Sci.* **16**, 262–268 (2012).
29. P. Hagoort, On Broca, brain, and binding: a new framework. *Trends Cogn. Sci.* **9**, 416–423 (2005).

30. U. Hasson, E. Yang, I. Vallines, D. J. Heeger, N. Rubin, A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* **28**, 2539–2550 (2008).
31. Y. Lerner, C. J. Honey, L. J. Silbert, U. Hasson, Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* **31**, 2906–2915 (2011).
32. N. Ding, L. Melloni, A. Yang, Y. Wang, W. Zhang, D. Poeppel, Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). *Front. Hum. Neurosci.* **11**, 481 (2017).
33. M. Keshishian, S. Norman-Haignere, N. Mesgarani, "Understanding Adaptive, Multiscale Temporal Integration In Deep Speech Recognition Systems" in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2021; https://proceedings.neurips.cc/paper/2021/hash/ccce2fab7336b8bc8362d115dec2d5a2-Abstract.html)vol. 34, pp. 24455–24467.
34. C. D. Gilbert, M. Sigman, Brain States: Top-Down Influences in Sensory Processing. *Neuron* **54**, 677–696 (2007).
35. K. Kveraga, A. S. Ghuman, M. Bar, Top-down predictions in the cognitive brain. *Brain Cogn.* **65**, 145–168 (2007).
36. C. Tang, L. S. Hamilton, E. F. Chang, Intonational speech prosody encoding in the human auditory cortex. *Science* **357**, 797–801 (2017).
37. Y. Li, C. Tang, J. Lu, J. Wu, E. F. Chang, Human cortical encoding of pitch in tonal and non-tonal languages. *Nat. Commun.* **12**, 1161 (2021).
38. T. Nagamine, N. Mesgarani, "Understanding the Representation and Computation of Multilayer Perceptrons: A Case Study in Speech Recognition" in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup, Y. W. Teh, Eds. (PMLR, 2017; https://proceedings.mlr.press/v70/nagamine17a.html)vol. 70 of *Proceedings of Machine Learning Research*, pp. 2564–2573.
39. T. Nagamine, M. L. Seltzer, N. Mesgarani, "Exploring how deep neural networks form phonemic categories." in *Interspeech* (2015), pp. 1912–1916.
40. Q. Raymondaud, M. Rouvier, R. Dufour, Probing the Information Encoded in Neural-based Acoustic Models of Automatic Speech Recognition Systems. arXiv arXiv:2402.19443 [Preprint] (2024). https://doi.org/10.48550/arXiv.2402.19443.
41. E. Edwards, M. Soltani, W. Kim, S. S. Dalal, S. S. Nagarajan, M. S. Berger, R. T. Knight, Comparison of time–frequency responses and the event-related potential to auditory speech stimuli in human cortex. *J. Neurophysiol.* **102**, 377–386 (2009).
42. G. Saon, Z. Tüske, D. Bolanos, B. Kingsbury, "Advancing RNN Transducer Technology for Speech Recognition" in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021; https://ieeexplore.ieee.org/abstract/document/9414716), pp. 5654–5658.
43. M. Masanori, K. Hideki, K. Haruhiro, Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech. *J. Audio Eng. Soc.* (2009).
44. M. Brysbaert, B. New, 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).

45. M. S. Vitevitch, P. A. Luce, Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition. *J. Mem. Lang.* **40**, 374–408 (1999).
46. C. Brodbeck, L. E. Hong, J. Z. Simon, Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Curr. Biol.* **28**, 3976-3983.e5 (2018).
47. S. Ylinen, M. Huuskonen, K. Mikkola, E. Saure, T. Sinkkonen, P. Paavilainen, Predictive coding of phonological rules in auditory cortex: A mismatch negativity study. *Brain Lang.* **162**, 72–80 (2016).
48. K. Friston, The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* **13**, 293–301 (2009).
49. P. Gagnepain, R. N. Henson, M. H. Davis, Temporal Predictive Codes for Spoken Words in Auditory Cortex. *Curr. Biol.* **22**, 615–621 (2012).
50. M. K. Leonard, K. E. Bouchard, C. Tang, E. F. Chang, Dynamic Encoding of Speech Sequence Probability in Human Temporal Cortex. *J. Neurosci.* **35**, 7203–7214 (2015).
51. D. A. Balota, M. J. Yap, K. A. Hutchison, M. J. Cortese, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson, R. Treiman, The English Lexicon Project. *Behav. Res. Methods* **39**, 445–459 (2007).
52. C. Shaoul, C. Westbury, Exploring lexical co-occurrence space using HiDEx. *Behav. Res. Methods* **42**, 393–413 (2010).
53. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, others, Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).
54. G. Mischler, V. Raghavan, M. Keshishian, N. Mesgarani, naplib-python: Neural acoustic data processing and analysis tools in python. *Softw. Impacts* **17**, 100541 (2023).