MAPPING WHISPER REPRESENTATIONS TO HUMAN ECOG RESPONSES WITH INTERPRETABLE TIME-RESOLVED NEURAL ENCODING

Matteo Ciferri[1,*], **Tommaso Boccato**[2], **Michal Olak**[2], **Matteo Ferrante**[1,2,†], **Nicola Toschi**[1,3,†]

[1]University of Rome Tor Vergata, Department of Biomedicine and Prevention
[2]Tether Evo
[3]A.A. Martinos Center for Biomedical Imaging, Harvard Medical School, Boston, USA

ABSTRACT

Understanding how hierarchical speech representations map onto human cortical activity is a central challenge in computational neuroscience. In this work, we study how internal representations from Whisper, a large-scale speech recognition model, predict intracranial electrophysiological (ECoG) responses during naturalistic speech perception. We introduce a time-resolved neural encoding model that aligns Whisper embeddings to word-locked cortical responses using a recurrent architecture with soft temporal attention. Our results show that intermediate Whisper layers consistently provide the best predictions of neural activity, revealing a correspondence between model hierarchy and cortical speech processing. In addition, a phonemic interpretability analysis uncovers anatomically coherent, phoneme-selective clusters in superior temporal cortex, providing converging evidence that intermediate speech model representations capture neural computations underlying human speech perception.

1 INTRODUCTION

Understanding how the human brain encodes the acoustic, phonetic, and linguistic structure of speech is a central challenge in computational neuroscience and cognitive science. Intracranial electrophysiology (ECoG) provides a unique opportunity to study these mechanisms with high temporal and spatial precision, revealing how auditory and language-responsive cortical regions transform continuous acoustic input into increasingly abstract linguistic representations. Recent advances in large-scale language models have shown that their hierarchical internal representations mirror aspects of neural processing, suggesting a promising bridge between machine-learned features and cortical dynamics (Raugel et al., 2025; Antonello et al., 2023; Oota et al., 2023; Ciferri et al., 2025).

At the same time, modeling the fine-grained mapping between continuous speech and neural activity remains difficult. Naturalistic listening conditions introduce substantial variability, and neural responses reflect a mixture of low-level acoustic cues, phonetic structure, and higher-order linguistic information unfolding over time. Capturing these dependencies requires models that are both temporally aware and capable of leveraging the hierarchical organization of modern speech encoders.

In this work, we investigate how representations from Whisper (Radford et al., 2023), a large speech recognition model trained at scale, predict ECoG activity during naturalistic speech perception. Using the Podcast ECoG dataset (Zada et al., 2025), we construct a time-resolved neural encoding model that aligns Whisper embeddings to cortical responses. This architecture allows us to examine which layers of the speech model best correspond to neural dynamics and how temporal information is integrated during word processing.

Beyond encoding performance, we introduce a phonemic interpretability framework that identifies electrodes with selective responses to articulatory phoneme categories. This analysis reveals

[*]Corresponding author: matteo.ciferri@uniroma2.it
[†]These authors contributed equally

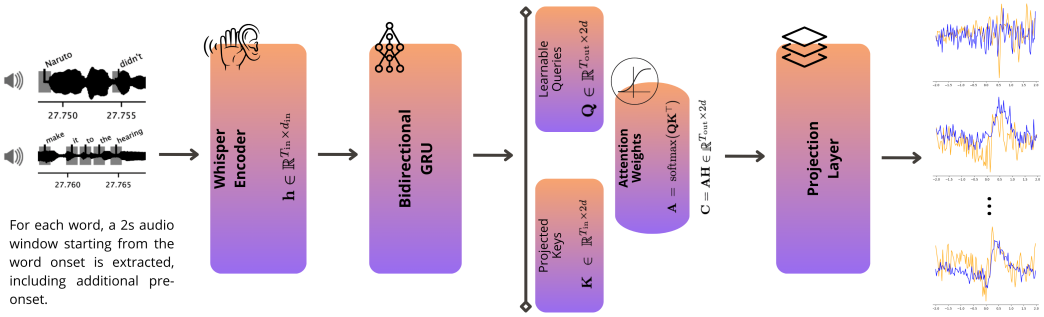

Figure 1: Overview of the neural encoding architecture. Word-aligned speech segments are processed by the Whisper encoder and a bidirectional GRU, followed by a temporal attention mechanism that maps speech representations to neural timepoints and a linear projection that predicts ECoG activity.

anatomically structured clusters across the superior temporal cortex, providing a bridge between model-derived features and known phonetic organization in auditory cortex.

Overall, our results demonstrate that (i) intermediate layers of Whisper best predict neural activity, (ii) the learned temporal alignment reflects both anticipatory and feedforward auditory processing, and (iii) phoneme-selective responses follow interpretable spatial gradients in the human cortex. Together, these findings highlight how modern speech models provide a powerful lens for understanding the neural basis of language perception.

2 Related Work

Early work on the neural basis of speech perception used hypothesis-driven encoding models based on spectro-temporal receptive fields or hand-crafted phonetic features to explain responses in auditory cortex (Mesgarani et al., 2009; Chang et al., 2010).

The advent of self-supervised and large-scale speech models has enabled more powerful, data-driven encoding approaches. Representations from models such as Wav2Vec2 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) have been used to predict fMRI, MEG, and ECoG responses, revealing layer-wise correspondences between model hierarchies and auditory pathways, from subcortical nuclei to higher-order language areas (Li et al., 2023; Anderson et al., 2024), such as the superior temporal gyrus (STG). In parallel, invasive speech brain–computer interface (BCI) work has focused primarily on decoding, using high-resolution intracortical or ECoG recordings together with recurrent or transformer architectures to reconstruct text or audio from attempted or imagined speech (Willett et al., 2023; Metzger et al., 2023; Chen et al., 2024). These studies demonstrate the feasibility of rich speech decoding, but typically optimize task performance rather than building explicit encoding models of continuous natural language comprehension.

Closer to our setting, several recent works have begun to exploit large speech–language models as joint models of brain and behavior. Goldstein et al. (2025) show that a unified acoustic-to-speech-to-language embedding space derived from Whisper captures neural activity during everyday conversations across widespread cortical networks. Their analysis, however, focuses on large-scale alignment between model stages and brain regions rather than modeling fine-grained, time-resolved mappings around individual word events.

In contrast, our work explicitly targets word-locked, time-resolved neural encoding, introducing a temporally aware alignment mechanism that enables phoneme level interpretability of the learned speech–brain mappings.

3 MATERIALS AND METHODS

In the following sections we describe the dataset used and propose the brain encoding framework to estimate neural activity from embeddings of a natural speech stimulus (Figure 1). The goal is to extract meaningful auditory information from the most responsive neural channels.

The dataset is freely accessible on the OpenNeuro platform (dataset number "*ds005574*"). Implementation code for reproducibility is available at the anonymised repository: `https://anonymous.4open.science/r/ecog_encoding-9A21/`. All experiments were carried out on a high performance server equipped with eight NVIDIA A100 GPUs (80 GB each, interconnected via NVLINK), 256 CPU threads and 2 TB of system memory.

3.1 DATASET

We analyzed the publicly available Podcast ECoG Dataset (Zada et al., 2025), which contains intracranial electrophysiological recordings (ECoG) acquired from 9 subjects while listening to a naturalistic 30-minute spoken narrative comprising 5,137 words. The dataset includes a total of 1,330 electrodes covering auditory, premotor, and higher-order language regions across participants. Neural signals were sampled at 512 Hz and stored in BIDS-compatible structures (Gorgolewski et al., 2016). During listening, participants heard the continuous podcast "Monkey in the Middle", and no behavioral task was required. This naturalistic setting enables the study of spontaneous cortical responses to ecologically valid speech.

We used the high-gamma ECoG derivatives, following the preprocessing pipeline released with the Podcast ECoG dataset (including automated artifact handling and high-gamma range 70–200 Hz filtering). Neural data were subsequently downsampled from 512 Hz to 32 Hz using MNE's FFT-based resampling procedure, which applies an implicit low-pass filter prior to resampling to prevent aliasing, as in the official tutorial (Zada et al., 2025). Finally, neural signals were z-scored channel-wise using normalization statistics computed on the training data and applied to the held-out test data. This preprocessing substantially reduces computational complexity while preserving the temporal resolution necessary to capture word-level cortical dynamics unfolding over hundreds of milliseconds.

Since both neural and audio data are continuous signals, we defined each data sample as a fixed temporal segment centered on a spoken word. Specifically, for each word onset provided in the time-aligned transcript, we extracted a neural segment spanning from -2 s to +2 s relative to onset. The corresponding audio segment was aligned to the same onset but included a shorter pre-onset context of 200 ms, avoiding the inclusion of temporally distant speech segments that are unlikely to contribute to the word-locked neural response, while retaining the post-onset portion corresponding to the same temporal window. The resulting word-aligned neural and audio segments were then paired to form a collection of (audio, neural) samples used for model training and evaluation.

3.2 AUDIO PROCESSING

Audio segments were resampled to 16 kHz and passed to the Whisper-base encoder (Radford et al., 2023) as a frozen feature extractor. Encoder hidden states were extracted at the model's intrinsic temporal resolution of approximately 50 Hz (one frame every $\sim$20 ms). Whisper is designed to process audio segments of 30 s in duration; in our setting we provide shorter, word-aligned segments. Accordingly, for each input we retained the first $50\times(T_{\text{pre}}+T_{\text{post}})$ encoder timesteps, corresponding to the temporal extent of the segmented audio window (e.g., 110 timesteps for a 2.2 s segment).

For each segment $x(t)$, the n-th hidden layer of Whisper provided a sequence of latent speech embeddings: $\mathbf{h}_t = f^{(n)}_{\text{Whisper}}(x_t)$. Prior analyses suggest that Whisper representations may capture information at multiple levels of abstraction across encoder layers (Goldstein et al., 2025): early layers predominantly encode the acoustic envelope and fine-grained spectral structure, whereas deeper layers progressively encode phonetic and articulatory information. Because Whisper retains the temporal structure of the input, its hidden states constitute an expressive basis for modeling time-resolved brain activity.

3.3 NEURAL ENCODING ARCHITECTURE

In order to model the mapping from hierarchical Whisper representations to neural activity, we adopted a time-aware neural encoder composed of a bidirectional Gated Recurrent Unit (GRU) followed by a temporal attention mechanism and a linear projection stage.

Let $\mathbf{h} \in \mathbb{R}^{T_{\text{in}} \times d_{\text{in}}}$ denote the sequence of Whisper embeddings for a speech segment, where T_{in} is the number of input timepoints and d_{in} is the input dimensionality. The embeddings are first processed by a bidirectional GRU with hidden size d, yielding the contextualized representation $\mathbf{H} = \text{BiGRU}(\mathbf{h}) \in \mathbb{R}^{T_{\text{in}} \times 2d}$, where each row $\mathbf{H}_t$ corresponds to the concatenated forward and backward hidden states at time t.

To align these contextualized speech embeddings with the neural response over time, we introduced a learnable soft alignment mechanism. The model learns a set of T_{out} temporal queries, $\mathbf{Q} \in \mathbb{R}^{T_{\text{out}} \times 2d}$, where T_{out} denotes the number of neural timepoints to be predicted. The contextualized representations are projected into a key space, $\mathbf{K} \in \mathbb{R}^{T_{\text{in}} \times 2d}$, and attention weights are obtained by $\mathbf{A} = \text{softmax}\big(\mathbf{Q}\mathbf{K}^\top\big) \in \mathbb{R}^{T_{\text{out}} \times T_{\text{in}}}$, where the softmax is applied along the input-time dimension T_{in}. Each row of $\mathbf{A}$ distributes attention over the Whisper time axis for a given predicted neural time index.

The resulting context vectors are computed as $\mathbf{C} = \mathbf{A}\mathbf{H} \in \mathbb{R}^{T_{\text{out}} \times 2d}$, providing a soft temporal alignment of speech features to neural timepoints. This mechanism allows the model to integrate temporally distant acoustic or phonetic cues when predicting the neural signal. Finally, the context vectors are mapped onto predicted neural activity by a linear projection $\hat{\mathbf{Y}} = \mathbf{C}\mathbf{W}_{\text{out}}^\top \in \mathbb{R}^{T_{\text{out}} \times S}$, where S is the number of electrodes.

The model was trained using a contrastive objective that aligns predicted and ground-truth neural responses. Predicted and real ECoG responses were compared using a temperature-scaled cosine similarity loss, encouraging each prediction to be most similar to its corresponding target within a batch. The temperature parameter was learned jointly with the model.

We chose a recurrent backbone to explicitly model short-range temporal dependencies while keeping the architecture lightweight. Interpretability is introduced through the temporal attention mechanism, which provides a transparent weighting of speech embeddings when predicting neural responses. More expressive architectures were avoided to reduce overfitting, given the limited number of word-aligned training samples.

3.4 EVALUATION

Model evaluation was performed using 5-fold cross-validation with temporally contiguous splits. Specifically, the continuous audio stream was partitioned into five non-overlapping segments, and in each fold one segment was held out for testing while the remaining segments were used for training. This strategy was adopted to prevent potential data leakage arising from temporal autocorrelations in naturalistic speech, and to avoid inflated performance that could result from the model interpolating between temporally adjacent and highly similar samples.

Model performance was quantified using Pearson's correlation between real and predicted neural activity: $r = \frac{\text{Cov}(\hat{\mathbf{y}}, \mathbf{y})}{\sigma_{\hat{\mathbf{y}}}\, \sigma_{\mathbf{y}}}$. For each subject and Whisper layer, we computed r for each electrode and neural timepoint across held-out test samples, and then averaged scores across cross-validation folds. Neural timepoints were defined relative to word onset, with negative lags corresponding to pre-onset activity and positive lags corresponding to post-onset responses.

3.5 PHONEMIC INTERPRETABILITY

In order to better understand how linguistic elements are represented in neural activity, a multi-stage interpretability framework was implemented. This module identifies significant neural activations time-locked to words, relates them to underlying phonemic categories, and organizes the results into spatial clusters on the cortical surface. The clusters are extracted using real neural data with only significant channels from the encoding results.

Table 1: Phoneme categories used in the interpretability analysis. Phonemes are reported in ARPA-bet notation.

Category	**Phonemes**
Vowels (Front)	IY, IH, EY, EH
Vowels (Central)	AH, ER
Vowels (Back)	UW, UH, AO, AA
Diphthongs	AY, OY, AW
Bilabial	P, B
Alveolar	T, D, N, S, Z
Velar–Palatal	K, G, NG
Post-Alveolar	SH, ZH
Labio-Dental	F, V, M
Dental / Glottal	TH, DH, HH
Approximants	L, R, W, Y

Neural activation significance was assessed using a sliding-window analysis applied to z-scored ECoG time series aligned with each spoken word. For each word-channel pair (w, s), the mean z-score within each window was extracted. A window was considered significant if the summary statistic (the mean) exceeded an empirical threshold z_{thr}. For each significant window, the peak index $t^* = \arg\max_t |\bar{z}_{w,s}(t)|$ was stored, representing the most responsive neural moment relative to the linguistic event and channel. This approach is conceptually related to recent analyses of phoneme-selective activation patterns in Whisper encoder representations, where a similar strategy is used to characterize category-specific temporal peaks in model layer activations.

Each spoken word was decomposed into phonemic sequences using a grapheme-to-phoneme (G2P) model (Bisani & Ney, 2008). Each phoneme was assigned to a broader articulatory category (following the Willett et al. (2023) results) such as front vowels, alveolar consonants, bilabial consonants, diphthongs, and others (Table 1).

We quantified category-specific phoneme selectivity at each channel using a multinomial likelihood ratio test (G-test). For each channel, we compared the observed distribution of phoneme categories to the global distribution computed across the entire dataset. Let O_{sc} be the observed count of category c in channel s, and $E_{sc} = n_s p_c$ the expected count under the global proportions p_c. The goodness-of-fit statistic is: $G_s = 2 \sum_c O_{sc} \log \left(\frac{O_{sc}}{E_{sc}} \right)$. Channels with large G_s deviate significantly from the global phoneme distribution. To assess which categories drive this deviation, we used the per-category contribution G_{sc} and computed one-sided chi-square p-values (denoted by $p_{s,c}$) testing for over-representation.

For each channel, we defined the set of significantly over-represented categories as $C_s = \{c \mid p_{s,c} < \alpha,\ O_{sc} > E_{sc}\}$, and assigned an initial label $c_s^* = \arg\min_{c \in C_s} p_{s,c}$. To ensure anatomical coherence, we applied a spatial refinement step based on majority voting among neighbors in MNI space. For each channel s, we identified all electrodes $N_r(s)$ within a fixed Euclidean radius and reassigned its label according to $c_s^{\text{refined}} = \arg\max_{c \in C_s} \left| \{ j \in N_r(s) \mid c_j^* = c \} \right|$, while restricting the vote to categories in C_s to ensure that no label lacking local statistical support could be assigned.

In order to quantify this organization, we compared the spatial coherence of phoneme-category clusters obtained from the subset of encoding-selected electrodes (most responsive channels) against clusters obtained by running the same pipeline on all electrodes. We used complementary clustering metrics that capture separation, compactness, and local consistency. The Silhouette score (Rousseeuw, 1987) quantifies how well electrodes assigned to the same phoneme category are separated from electrodes assigned to other categories, with higher values indicating better-defined clusters. The Davies–Bouldin index (Davies & Bouldin, 2009) measures the ratio between within-cluster dispersion and between-cluster separation, where lower values reflect more compact and well-separated clusters. Finally, local label entropy (Shannon entropy) quantifies the diversity of phoneme labels within a spatial neighborhood of each electrode, with lower entropy indicating greater local homogeneity and cleaner anatomical organization.

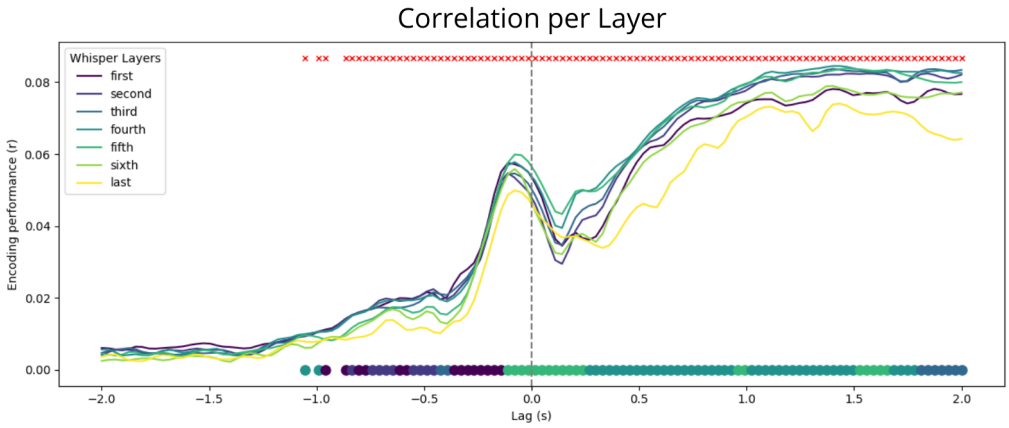

Figure 2: Time-resolved encoding performance across Whisper layers. The curves show the mean Pearson correlation (averaged across channels and subjects) between predicted and observed neural activity. Colored dots along the bottom indicate, for each timepoint, the layer achieving the highest performance. Red crosses along the top mark timepoints that reached statistical significance under a permutation test against a null distribution obtained by shuffling word–neural pairings. Intermediate layers consistently outperform both lower and higher layers, with layer 4 emerging as the statistically best overall: a paired bootstrap analysis across subjects (using one averaged value per subject) revealed that layer 4 significantly exceeds all other layers after FDR correction (the least significant comparison still yielded $q < 0.04$, and all other comparisons reached $q < 1 * 10^{-15}$).

4 Results

We first evaluate how well hierarchical speech embeddings predict time-resolved ECoG responses during naturalistic speech perception, revealing a structured correspondence between Whisper layers and neural dynamics. Finally, we assess the interpretability of the learned mappings, showing that the model captures meaningful temporal alignment and recovers anatomically coherent, phoneme-selective cortical organization.

4.1 Encoding Performance

Hierarchical representations from Whisper reliably predicted ECoG responses during naturalistic speech perception (Fig. 2). All layers exhibited a rise in encoding performance shortly before word onset with a sharp peak followed by a brief dip and a subsequent increase during the post-onset interval. While layer depth is often associated with increasing representational abstraction in speech models, we do not directly manipulate or isolate specific linguistic levels. Accordingly, our results are interpreted as evidence for layer-dependent differences in representational alignment with neural activity (Goldstein et al., 2025).

Layer-specific analyses showed that intermediate Whisper layers provided the best overall predictions. Early layers aligned with pre-onset and onset-related responses, while middle layers dominated after onset, consistent with the timing of acoustic processing (Figure 3).

4.2 Baseline Comparisons

To contextualize the performance of our proposed time-aware encoding model, we compared it against two linear baseline models commonly used in prior works. The first baseline uses Whisper representations and mirrors our feature extraction pipeline, but replaces the recurrent and attention-based temporal alignment with a linear regression applied to a single word-level window, obtained by averaging Whisper embeddings across time. The second baseline follows the approach adopted in the original Podcast ECoG paper (Zada et al., 2025), using middle-layer GPT-2 representations (lacks access to acoustic and phonetic information) with a linear encoding model.

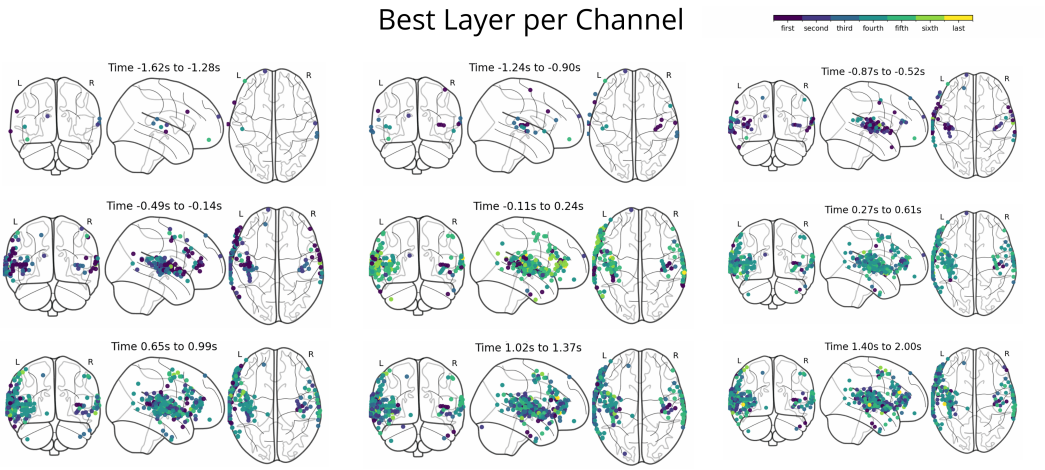

Figure 3: For each electrode, the Whisper layer yielding the highest encoding performance is shown across successive temporal windows relative to word onset. Early windows are dominated by lower-level layers, reflecting acoustic tracking, while later windows show a shift toward middle layers, consistent with phonetic and articulatory processing. Spatial patterns reveal a progression from posterior auditory cortex to more anterior superior temporal and frontal regions, suggesting a temporal-to-linguistic representational gradient.

Figure 4 summarizes the comparison across models in terms of temporal encoding performance and spatial distribution of predictive electrodes. While both linear baselines capture coarse stimulus–response relationships, they exhibit substantially weaker performance compared to our model.

The comparison between our model (fourth layer as the best) and the linear Whisper baseline demonstrates that the performance gains are not merely due to the choice of representation, but arise from the explicit modeling of temporal dependencies. By learning a soft-alignment between speech embeddings and neural timepoints, our model captures both anticipatory and post-onset dynamics that are largely inaccessible to linear, word-averaged approaches.

4.3 Interpretability Results

The learned attention map in Figure 5 (left) shows a smooth diagonal structure between Whisper embedding time steps and neural timepoints, indicating that the model tends to place higher weight on input frames that are temporally close to (and often slightly preceding) the neural timepoint being predicted. The observed structure suggests that the model leverages local temporal context in the stimulus representations when mapping speech features to time-resolved neural activity.

The phoneme-selective electrodes identified by our interpretability analysis form anatomically coherent spatial clusters across the temporal cortex (Figure 5, right). Although individual phoneme classes partially overlap, their overall organization reveals a systematic structure consistent with known articulatory and acoustic gradients in the human auditory system. Vowel-related categories (front vowels, central vowels, diphthongs) tend to occupy more medial portions of the superior temporal gyrus (STG) and the internal part of the planum temporale. These regions are known to encode spectral and harmonic structure, consistent with the acoustic properties of vowel production (Steinschneider, 2023; Chang et al., 2010). Consonant-related categories, by contrast, form denser clusters distributed along the STG, with alveolar, labio-dental, and dental categories appearing more anteriorly, and post-alveolar categories located more posteriorly. Bilabial consonants represent a partial exception, showing a more medial distribution.

Across complementary metrics, the encoding-selected configuration yielded substantially more separable and spatially consistent clusters (Table 2): silhouette scores increased markedly, Davies–Bouldin indices decreased, and local neighborhood label entropy was reduced. Bootstrap resampling over electrodes confirmed that these improvements were statistically reliable, indicating that restricting the analysis to encoding-informative channels enhances the anatomical specificity of phoneme-

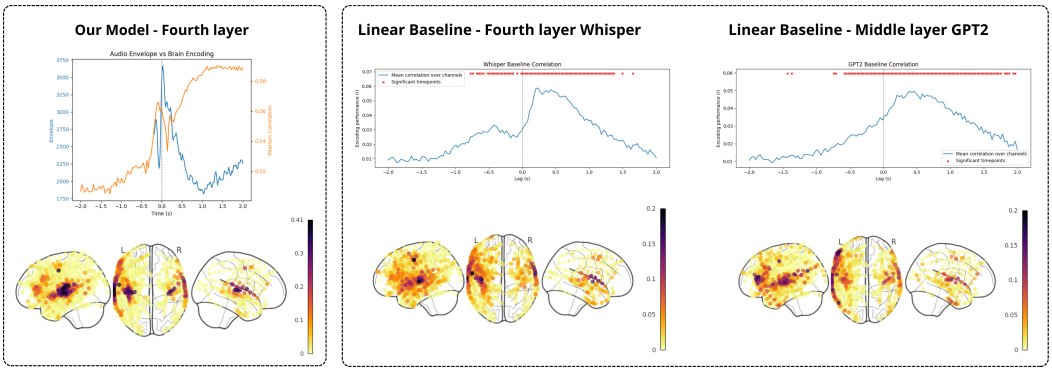

Figure 4: Comparison between our time-aware encoding model and linear baseline models. **Left:** Results obtained with our model using the most predictive Whisper layer (layer 4). This panel illustrates the temporal relationship between encoding performance and the acoustic envelope, plotted with dual y-axes for visualization purposes. In addition to achieving the highest overall encoding performance, our model exhibits a pronounced and structured temporal correlation profile. Encoding performance rises prior to word onset, shows a transient dip around the time of maximum acoustic envelope energy, and subsequently increases again after onset. This non-monotonic pattern is consistent with a transition from early acoustic feature tracking to the integration of higher-level linguistic information. **Middle:** Linear baseline using the same Whisper layer but with time-averaged embeddings and no explicit temporal modeling. While correlations are significant around word onset, overall performance is substantially reduced. **Right:** Linear baseline using middle-layer GPT-2 representations, following the approach adopted in the original dataset paper.

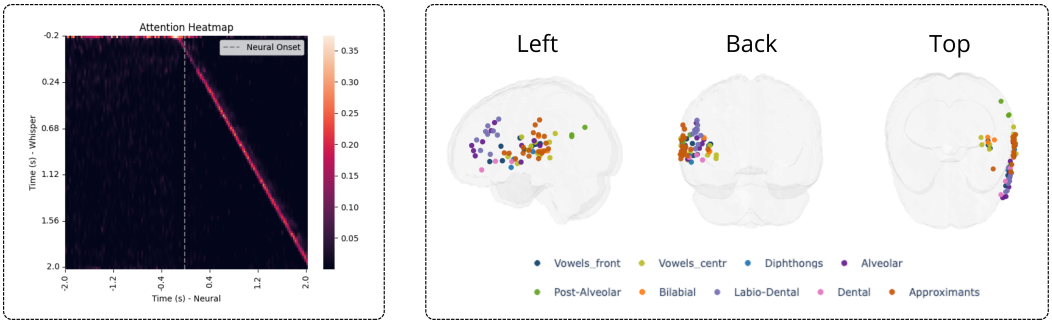

Figure 5: **Left**: Attention map showing how neural timepoints (horizontal axis, aligned to word onset) weight Whisper encoder representations at different relative temporal offsets (vertical axis). The map is obtained by averaging attention weights across all test samples (i.e., words, electrodes, and subjects). A diagonal structure is observed, indicating a systematic temporal alignment between neural activity and stimulus representations computed over nearby acoustic frames. **Right**: Spatial distribution of phoneme-selective electrodes across the temporal lobe. Colors denote broad phoneme categories defined based on articulatory properties (e.g., front/central/back vowels, bilabial, alveolar, dental, approximants).

Table 2: Quantitative comparison of phoneme-cluster spatial coherence using encoding-selected electrodes (most responsive channels) vs. all electrodes. Bootstrap statistics are computed by resampling electrodes ($n = 1000$) and performing a t-test between distributions.

Metric	**Selected (obs)**	**All (obs)**	**Bootstrap mean±std**	p **(one-tailed)**
Silhouette ↑	0.255	0.020	0.303±0.073 vs. 0.031±0.030	4.12×10^{-8}
DBI ↓	2.356	6.901	2.463±0.586 vs. 5.892±1.477	5.67×10^{-11}
Local entropy ↓	1.733	1.911	1.531±0.128 vs. 1.736±0.063	6.11×10^{-5}

selective organization rather than simply reducing sample size. Consequently, encoding-based electrode selection does not trivially induce spatial clustering, as shown by the markedly lower coherence observed when applying the same pipeline to all electrodes.

5 Discussion

The primary finding of this work is that intermediate layers of Whisper yield the highest encoding performance when predicting ECoG activity during naturalistic speech perception. This result is consistent with a robust pattern observed across brain–model alignment studies: representations from middle layers of large language and speech models tend to best match neural responses, whereas early layers are dominated by low-level sensory features and deeper layers become increasingly semantic-specific. Similar effects have been reported for text-based foundation models such as GPT-2 and BERT, where intermediate layers most strongly predict fMRI, MEG, and ECoG responses during reading or listening (Jain & Huth, 2018; Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022). Our results extend this principle to a large-scale speech foundation model and to intracranial, time-resolved recordings, supporting the idea that intermediate representational stages capture the level of abstraction most relevant for cortical speech processing.

The learned attention alignment reveals a structured temporal relationship between Whisper embeddings and neural responses, with cortical activity relying primarily on slightly earlier acoustic frames. This pattern is consistent with known auditory cortical latencies and with predictive mechanisms during speech perception, whereby upcoming linguistic information is partially anticipated based on context. Similar anticipatory dynamics have been reported in electrophysiological studies comparing neural activity to hierarchical representations from deep speech and language models, suggesting that temporal structure in these models mirrors neural processing timescales rather than reflecting a simple stimulus–response lag.

Beyond encoding accuracy, our phonemic interpretability analysis identifies electrodes selectively responsive to articulatory phoneme categories, forming anatomically coherent clusters across the temporal lobe. This organization is strongly consistent with prior intracranial and electrophysiological evidence showing that STG encodes phonetic and articulatory features of speech. Classic ECoG studies have demonstrated systematic representation of phoneme categories and phonetic features in STG during continuous speech perception (Chang et al., 2010; Mesgarani et al., 2014), as well as separable neural responses to vowels and consonants. Our results do not introduce new phoneme categories per se, but demonstrate that such cortical structure can be recovered through an encoding model grounded in Whisper representations, linking phoneme-level neural selectivity to representations learned by a modern speech foundation model.

6 Conclusion

Together, these findings support a convergent picture in which intermediate representations of large speech models provide the best match to cortical speech processing, both in terms of encoding performance and interpretability. The results strengthen the view that foundation models trained on large-scale speech data capture representational stages that are not only useful for recognition, but also closely aligned with the neural computations underlying human speech perception.

AI USE DISCLOSURE

The authors used large language models as an assistive tool for editing and improving the clarity of the manuscript. All scientific content, experimental design, data analysis, and interpretation of results were performed and verified by the authors. The authors take full responsibility for the content of this work.

REFERENCES

Andrew J Anderson, Chris Davis, and Edmund C Lalor. Deep-learning models reveal how context and listener attention shape electrophysiological correlates of speech-to-language transformation. *PLOS Computational Biology*, 20(11):e1012537, 2024.

Richard Antonello, Aditya Vaidya, and Alexander G. Huth. Scaling laws for language encoding models in fmri, 2023.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451, 2008.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, December 2022. ISSN 2399-3642. doi: 10.1038/s42003-022-03036-1. URL https://www.nature.com/articles/s42003-022-03036-1.

Edward F Chang, Jochem W Rieger, Keith Johnson, Mitchel S Berger, Nicholas M Barbaro, and Robert T Knight. Categorical speech representation in human superior temporal gyrus. *Nature neuroscience*, 13(11):1428–1432, 2010.

Xupeng Chen, Ran Wang, Amirhossein Khalilian-Gourtani, Leyao Yu, Patricia Dugan, Daniel Friedman, Werner Doyle, Orrin Devinsky, Yao Wang, and Adeen Flinker. A neural speech decoding framework leveraging deep learning and speech synthesis. *Nature Machine Intelligence*, 6(4): 467–480, 2024.

Matteo Ciferri, Matteo Ferrante, and Nicola Toschi. Reconstructing music perception from brain activity using a prior guided diffusion model. *Scientific Reports*, 15(1):42108, 2025.

David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 2009.

Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, March 2022. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-022-01026-4. URL https://www.nature.com/articles/s41593-022-01026-4.

Ariel Goldstein, Haocheng Wang, Leonard Niekerken, Mariano Schain, Zaid Zada, Bobbi Aubrey, Tom Sheffer, Samuel A Nastase, Harshvardhan Gazula, Aditi Singh, et al. A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature human behaviour*, pp. 1–15, 2025.

Krzysztof J Gorgolewski, Tibor Auer, Vince D Calhoun, R Cameron Craddock, Samir Das, Eugene P Duff, Guillaume Flandin, Satrajit S Ghosh, Tristan Glatard, Yaroslav O Halchenko, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific data*, 3(1):1–9, 2016.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. *Advances in neural information processing systems*, 31, 2018.

Yuanning Li, Gopala K Anumanchipalli, Abdelrahman Mohamed, Peili Chen, Laurel H Carney, Junfeng Lu, Jinsong Wu, and Edward F Chang. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, 26(12):2213–2225, 2023.

Nima Mesgarani, Stephen V David, Jonathan B Fritz, and Shihab A Shamma. Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *Journal of neurophysiology*, 102(6):3329–3339, 2009.

Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.

Sean L Metzger, Kaylo T Littlejohn, Alexander B Silva, David A Moses, Margaret P Seaton, Ran Wang, Maximilian E Dougherty, Jessie R Liu, Peter Wu, Michael A Berger, et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046, 2023.

Subba Reddy Oota, Manish Gupta, Raju S. Bapi, Gael Jobard, Frederic Alexandre, and Xavier Hinaut. Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey), July 2023. URL http://arxiv.org/abs/2307.10246. arXiv:2307.10246 [cs, q-bio].

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Joséphine Raugel, Stéphane d'Ascoli, Jérémy Rapin, Valentin Wyart, and Jean-Rémi King. Scaling and context steer llms along the same computational path as the human brain. *arXiv preprint arXiv:2512.01591*, 2025.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Mitchell Steinschneider. Toward an understanding of vowel encoding in the human auditory cortex. *Neuron*, 111(13):1995–1997, 2023.

Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

Zaid Zada, Samuel A Nastase, Bobbi Aubrey, Itamar Jalon, Sebastian Michelmann, Haocheng Wang, Liat Hasenfratz, Werner Doyle, Daniel Friedman, Patricia Dugan, et al. The "podcast" ecog dataset for modeling neural activity during natural language comprehension. *Scientific Data*, 12(1):1135, 2025.

A Appendix

A.1 Projection to Glasser-Atlas

To assess whether the encoding effects are anatomically coherent at the ROI level, we projected channel-wise encoding correlations onto the Glasser cortical atlas (Glasser et al., 2016). Specifically, for each ROI we aggregated the correlations of electrodes falling within that parcel and reported the mean correlation and number of contributing channels (Figure A1).

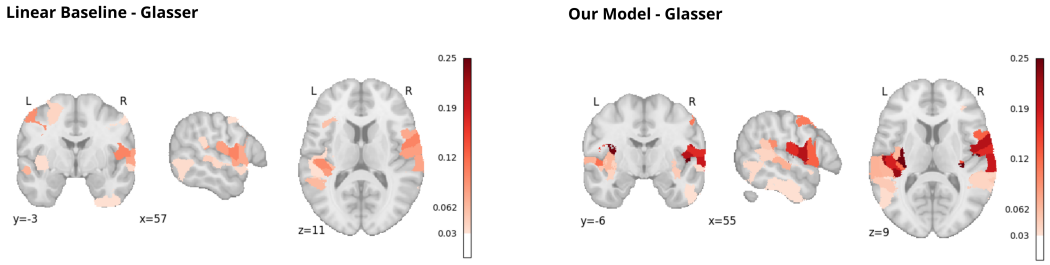

Figure A1: ROI-level projection of encoding correlations onto the Glasser atlas. **Left:** linear baseline, Whisper embeddings with time-averaging and linear regression. **Right:** our time-aware model. Warm colors indicate higher mean Pearson correlation within each ROI (aggregated over electrodes assigned to that parcel; colorbar shared across panels). Both approaches highlight auditory and peri-auditory cortex, but our model yields stronger and more spatially concentrated ROI responses, with prominent effects in core/belt auditory parcels and adjacent opercular/insular regions.

Table A1: Top Glasser ROIs by mean encoding correlation. For each ROI we report the model, hemisphere (hemi), mean Pearson correlation across electrodes within the ROI, and the number of contributing channels.

Model	**ROI**	**Hemi**	**Mean r**	**#Ch**
Linear	Area 43	L	0.1439	2
	Primary Auditory Cortex	R	0.1092	4
	Area 55b	R	0.1063	1
	Lateral Belt Complex	R	0.0956	5
	Area OP4/PV	L	0.0836	2
	Auditory 4 Complex	L	0.0795	5
	Rostral Area 6	L	0.0760	2
	Area STSd Anterior	R	0.0682	1
	Medial Belt Complex	R	0.0644	9
	Area Temporoparietotooccipital Junction 1	R	0.0626	1
Our Model	Area 52	R	0.3160	1
	Medial Belt Complex	L	0.3116	4
	Insular Granular Complex	R	0.2438	2
	Lateral Belt Complex	R	0.2237	5
	Area OP1/SII	R	0.2101	2
	Area 43	L	0.2044	2
	Auditory 4 Complex	L	0.1896	5
	Primary Auditory Cortex	R	0.1801	4
	Area OP4/PV	L	0.1706	2
	Area OP2–3/VS	R	0.1578	1

Both the linear Whisper baseline and our model highlight auditory and peri-auditory regions, indicating that the strongest encoding effects are not spatially random. However, our model produces a markedly stronger and more focal ROI pattern: peak correlations are higher and concentrate in

core and belt auditory areas (e.g., Area 52, Medial Belt Complex, Lateral Belt Complex), as well as adjacent opercular/insular regions (e.g., OP1/SII, Insular Granular Complex). In contrast, the linear baseline yields lower ROI means and a weaker overall anatomical contrast (Table A1), consistent with the reduced time-locked encoding performance observed in the baseline comparison.

A.2 Temporal Ablation

In this chapter we address the presence of encoding performance at timepoints distant from word onset, potentially reflecting temporal leakage induced by the bidirectional architecture or by the choice of temporal context. We conducted two complementary ablation analyses targeting (i) architectural temporal information and (ii) input temporal windowing.

To assess whether future information contributes to the observed encoding patterns, we replaced the bidirectional GRU with a strictly causal (unidirectional) GRU, preventing the model from accessing future input frames when predicting neural responses. Results show that the overall temporal profile of encoding performance remains qualitatively unchanged (Figure A2 first raw). In particular, the post-onset rise in correlation and the sustained encoding performance are preserved under the causal architecture. This indicates that the model does not rely on future information to achieve its performance, and suggests that the observed temporal structure reflects genuine alignment between stimulus representations and neural activity.

A second potential source of anticipatory effects is the inclusion of pre-onset acoustic context (200 ms before word onset) in the input. To evaluate its impact, we repeated the analysis removing any pre-onset audio, such that the model only receives information starting from word onset.

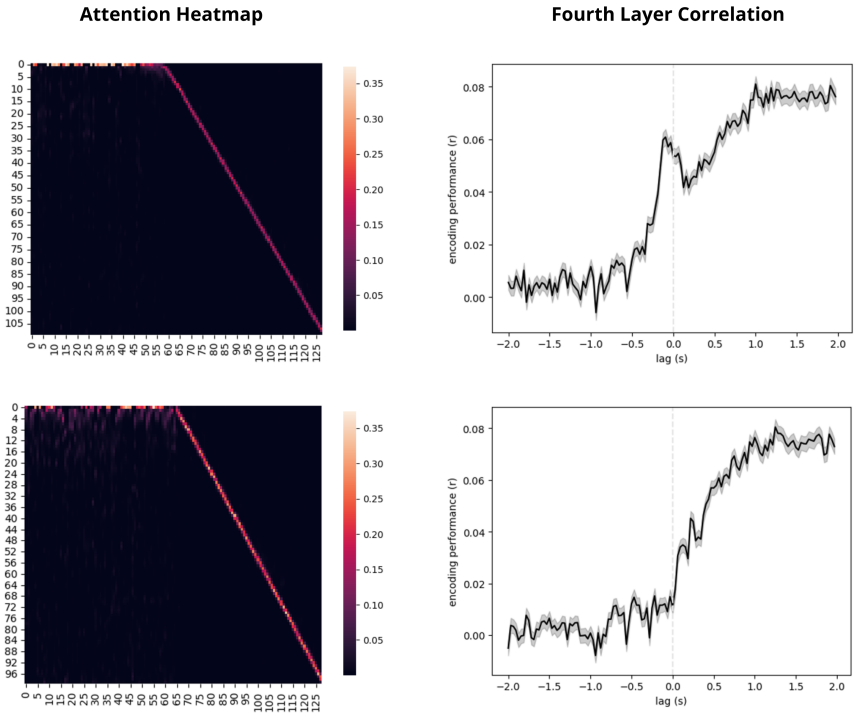

Figure A2: Temporal ablation analysis. Left: average temporal attention maps showing the alignment between input speech representations and predicted neural timepoints. Right: encoding performance. Top: unidirectional GRU model with pre-onset context (200 ms). Bottom: same GRU model without pre-onset context. The removal of pre-onset input eliminates the pre-onset peak in encoding performance, while preserving the post-onset dynamics.

As shown in Figure A2 line (b), removing pre-onset context eliminates the peak in encoding performance observed before word onset. The resulting temporal profile becomes strictly post-onset,

with encoding performance rising only after stimulus presentation. This result demonstrates that the pre-onset peak observed in the main model is not an artifact, but rather reflects the availability of preceding acoustic context. Importantly, this context corresponds to natural speech continuity, where upcoming words are partially predictable from prior linguistic input.