Self-Supervised Models of Audio Effectively Explain Human Cortical Responses to Speech

Aditya R. Vaidya [1] **Shailee Jain** [1] **Alexander G. Huth** [1 2]

Abstract

Self-supervised language models are very effective at predicting high-level cortical responses during language comprehension. However, the best current models of lower-level auditory processing in the human brain rely on either hand-constructed acoustic filters or representations from supervised audio neural networks. In this work, we capitalize on the progress of self-supervised speech representation learning (SSL) to create new state-of-the-art models of the human auditory system. Compared against acoustic baselines, phonemic features, and supervised models, representations from the middle layers of self-supervised models (APC, wav2vec, wav2vec 2.0, and HuBERT) consistently yield the best prediction performance for fMRI recordings within the auditory cortex (AC). Brain areas involved in low-level auditory processing exhibit a preference for earlier SSL model layers, whereas higher-level semantic areas prefer later layers. We show that these trends are due to the models' ability to encode information at multiple linguistic levels (acoustic, phonetic, and lexical) along their representation depth. Overall, these results show that self-supervised models effectively capture the hierarchy of information relevant to different stages of speech processing in human cortex.

1. Introduction

Self-supervised learning (SSL) has emerged as a popular and successful pre-training objective. In natural language processing, self-supervised language models (LM) encode diverse linguistic information and achieve excellent zero-shot performance on many language tasks (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019). Capitalizing on these findings, neuroimaging studies have shown that representations extracted from LMs are highly effective at predicting brain activity elicited by natural language (Wehbe et al., 2014; Jain & Huth, 2018; Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux et al., 2021; Goldstein et al., 2020) and can help reveal how linguistic representations are organized across human cortex (Jain et al., 2020; Antonello et al., 2021). Even outside the domain of language, self-supervised models have proven fruitful in modeling the mammalian visual system (Zhuang et al., 2021).

In automatic speech recognition there has been a similar trend towards SSL-based approaches. Using autoregressive, contrastive, and masked prediction losses, SSL models learn powerful acoustic representations that are highly transferable across applications (Hsu et al., 2021; Schneider et al., 2019; Baevski et al., 2020). Much like language models, SSL models of speech have also been shown to capture higher-order linguistic structure without explicit supervision (Pasad et al., 2021; Yang et al., 2021). This can be attributed to their ability to integrate information over a broader context than traditional acoustic filters. However, SSL models have not yet been combined with neuroimaging to study auditory processing in the human brain, where the best models are currently hand-constructed features (Norman-Haignere & McDermott, 2018; Venezia et al., 2019; Chi et al., 2005; Mesgarani et al., 2014) or supervised neural networks (Millet & King, 2021). In this paper, we investigate the potential of SSL speech representations for predicting human cortical responses to natural speech.

One approach to studying sensory representations in the brain is through "encoding models" (Wu et al., 2006). These predictive models of brain activity learn a mapping from stimulus features to responses measured using a neuroimaging technique like fMRI. In this work, we use SSL models to extract acoustic features of natural speech and in turn, build voxel-wise speech encoding models using data from an fMRI experiment (Figure 1). Our experiment comprised human participants passively listening to English-language narrative stories while their whole-brain fMRI BOLD activ-

[1]Department of Computer Science, [2]Department of Neuroscience, The University of Texas at Austin. Correspondence to: Aditya Vaidya <avaidya@utexas.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

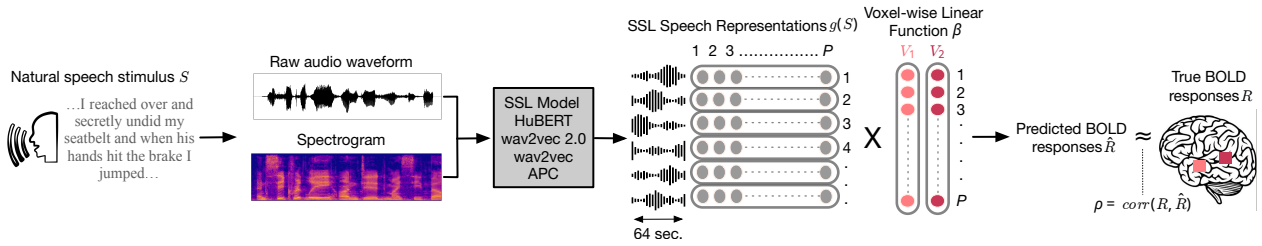

Figure 1. Voxel-wise encoding models from SSL models of speech. 64 second spans of narrative stimuli (represented as waveforms or spectrograms) are fed into an SSL model trained to learn the statistics of speech. Hidden states from a layer in the neural network are extracted to form representation $g(S)$ after downsampling to the rate of fMRI acquisition. $g(S)$ is then used to fit encoding models that predict fMRI BOLD responses to natural speech. Encoding model weights $\beta \in \mathbb{R}^{P \times V}$ are estimated with ridge regression. Models are then used to predict $\hat{R}$, the BOLD response to an unseen stimulus, and are evaluated by the correlation between R and $\hat{R}$.

ity was being recorded. Prior work on LM-based language encoding models has found performance differences across LM layers (Jain & Huth, 2018; Toneva & Wehbe, 2019), and layer-wise analyses of SSL models have shown that they capture different types of acoustic information (Pasad et al., 2021). Motivated by these findings, we built separate encoding models for each layer in four SSL models — APC, wav2vec, wav2vec 2.0 and HuBERT (Table 1). Their performance was compared to low-level acoustic baselines like Mel spectrograms and spectrotemporal features, mid-level phoneme articulations, and high-level semantic features.

Overall, we found that SSL models are highly effective at predicting cortical responses to natural speech, beating supervised and hand-engineered acoustic representations. Despite being trained on audiobooks, the SSL representations transferred well to encoding models of natural speech. We further show that encoding performance greatly varied across SSL model layers and brain areas. While the upper-middle layers of some models had the best performance broadly across cortex, lower layers were comparable only in low-level primary auditory cortex. To better understand why these models worked so well, we did variance partitioning across pairs of feature spaces. This revealed that the best SSL layers capture spectral, phonemic, and semantic information. Finally, we directly assessed linguistic representations in each SSL layer using linear probes. This showed that the evolution of representations across SSL layers mirrors the presumed stages of speech processing. Together with the encoding model results, this shows that SSL models capture diverse acoustic information across their representational depth which enables them to effectively model different stages of speech processing in cortex.

2. Natural Language fMRI Experiment

To build SSL encoding models, we used data from an fMRI experiment comprising 7 participants (3 female) listening to over five hours of spoken narrative stories (27 stories; ∼57,900 total words) from *The Moth Radio Hour*. These rich, diverse naturalistic stimuli are highly representative of speech humans encounter on a daily basis. To find the exact timing of each word and phoneme, the stories were transcribed and a forced aligner was used to align each transcript to its audio. All subjects were healthy with normal hearing, had fluent English language comprehension, and gave written informed consent. Whole-brain MRI data was collected every 2 seconds (TR). The experimental procedure was approved by the local Institutional Review Board. More MRI acquisition details can be found in Appendix D.

3. Voxel-wise Encoding Models

Encoding models aim to approximate $f(S) = R$, the mapping between a stimulus S and the elicited BOLD response R measured in some brain area (Figure 1). Here, the stimulus S can take any form, like images, audio waveforms, or words. To make model fitting tractable with limited fMRI data, we restricted encoding models to the *linearized* form $f(S) = g(S)\beta$, where g is a pre-specified non-linear transformation on the stimulus and β is a vector of learned linear weights (Wu et al., 2006). If g transforms S into a P-dimensional feature space and V is the number of cortical voxels (regression targets), then β is a $P \times V$ linear transformation between the feature space and each voxel's predicted response. We estimated a separate encoding model $\hat{f}_v$ for each voxel v on the training dataset $(S_{\text{train}}, R_{\text{train}})$ using ridge regression. To select the ridge parameter independently for each voxel, we used 50 iterations of cross-validation. Since fMRI data is auto-correlated, for each cross-validation run we randomly sampled 40 different chunks of the training data, each totaling over 4 minutes. The training set comprised 26 stories, totaling 5.4 hours.

For evaluation, we used the learned encoding models to predict the response timecourse of each voxel v on a held-out test set, $\hat{R}_{\text{test},v} = \hat{f}_v(S_{\text{test}})$. We then calculated the linear correlation between true and predicted responses to

Table 1. Model architectures and training objectives. For each model, its input representation is fed into an encoder module to produce latent speech representations. A second module then produces contextualized representations that capture information across the input. Both modules are trained end-to-end to fulfill their supervised or self-supervised objective.

Model	Input	Encoder	Contextualizer	Loss type
APC (Chung et al., 2019; Chung & Glass, 2020)	log Mel spectrogram	N/A	3-layer GRU	Autoregressive
wav2vec (Schneider et al., 2019)	Waveform	7-layer CNN	12-layer CNN	Contrastive
wav2vec 2.0 BASE (Baevski et al., 2020)	Waveform	7-layer CNN	12-layer Transformer	Masked contrastive
HuBERT BASE (Hsu et al., 2021)	Waveform	7-layer CNN	12-layer Transformer	Masked predictive
Deep Speech 2 (Amodei et al., 2016)	log spectrogram	2-layer CNN	5-layer LSTM	CTC (supervised)

determine encoding performance $\rho_v = \text{corr}(R_{\text{test},v}, \hat{R}_{\text{test},v})$. The test set comprised 1 held-out story (10 minutes) that did not participate in model estimation.

Using the encoding model framework, we can compare how well different feature spaces model speech processing in the brain. Consider a feature space g_i on which we estimate and test encoding models, yielding prediction performance ρ_v^i for a voxel v. If $\rho_v^1 > \rho_v^2$ for some feature spaces g_1 and g_2, we can conclude that g_1 is a closer match to the information encoded by voxel v while processing speech. By examining the type of information encoded in g_1, we can consequently estimate voxel function. Here we use this approach to compare representations extracted from SSL speech models with several well-known baselines and supervised alternatives.

3.1. Extracting Speech Features from SSL Models

We extracted hidden state representations from each layer of four self-supervised models: APC (Chung et al., 2019; Chung & Glass, 2020), wav2vec (Schneider et al., 2019), wav2vec 2.0 (Baevski et al., 2020), and HuBERT (Hsu et al., 2021). These models are pre-trained on 960 hours of speech from LibriSpeech (Panayotov et al., 2015), a corpus of English audiobooks from the LibriVox project. The SSL models are not finetuned on any annotated samples. See Table 1 for an overview of model architectures and training methods. In the main analyses, we evaluate the output of CNN encoders and all layers of the contextualizer.

The self-supervised speech models effectively capture statistical regularities in speech, either by learning to predict the content of future or unknown time blocks (autoregressive & masked prediction objectives), or by learning features that are effective at discriminating future samples from other samples (contrastive objective). While APC operates on Mel spectrograms of the input, the other three models operate directly on the waveform.

3.2. Supervised Speech Model Baseline

To compare SSL-based encoding models with related work (Millet & King, 2021), we additionally extracted representations from Deep Speech 2 (Amodei et al., 2016). This automatic speech recognition model was trained with full supervision on the same amount of data as the self-supervised models (960 hours of LibriSpeech).

3.3. Baselines

We also compared encoding models trained with SSL representations against those trained with traditional acoustic and hand-labeled features.

We first used spectrotemporal modulations, which are a standard model of primary auditory cortex responses to sound (Norman-Haignere & McDermott, 2018; Chi et al., 1999; 2005; Venezia et al., 2019). These features are computed by convolving spectrograms with filters that are selective to different rates of spectral and/or temporal modulation. We also computed spectral features ("FBANK") by applying Mel-scale triangular filters to the power spectrum. While FBANK captures frequencies, spectrotemporally-modulated spectrograms may additionally capture changing tones or harmonics across time.

Since parts of auditory cortex are known to selectively respond to phonemes (Mesgarani et al., 2014), we also used phoneme articulations as a mid-level speech feature. These were derived by mapping hand-labeled phonemes onto 14 articulatory features.

For word-level features, we used a 985-dimensional word embedding to capture lexical and semantic information (Huth et al., 2016). Finally, we extracted deep contextual features from the 9th layer of GPT (Radford et al., 2018), since similar LMs are the most effective representations for modeling the brain's response to language (Toneva & Wehbe, 2019; Caucheteux et al., 2021).

Since some of the artificial neural networks discussed here are bi-directional, we enforced causality by extracting representations at the end of a sliding window of size 64 s with

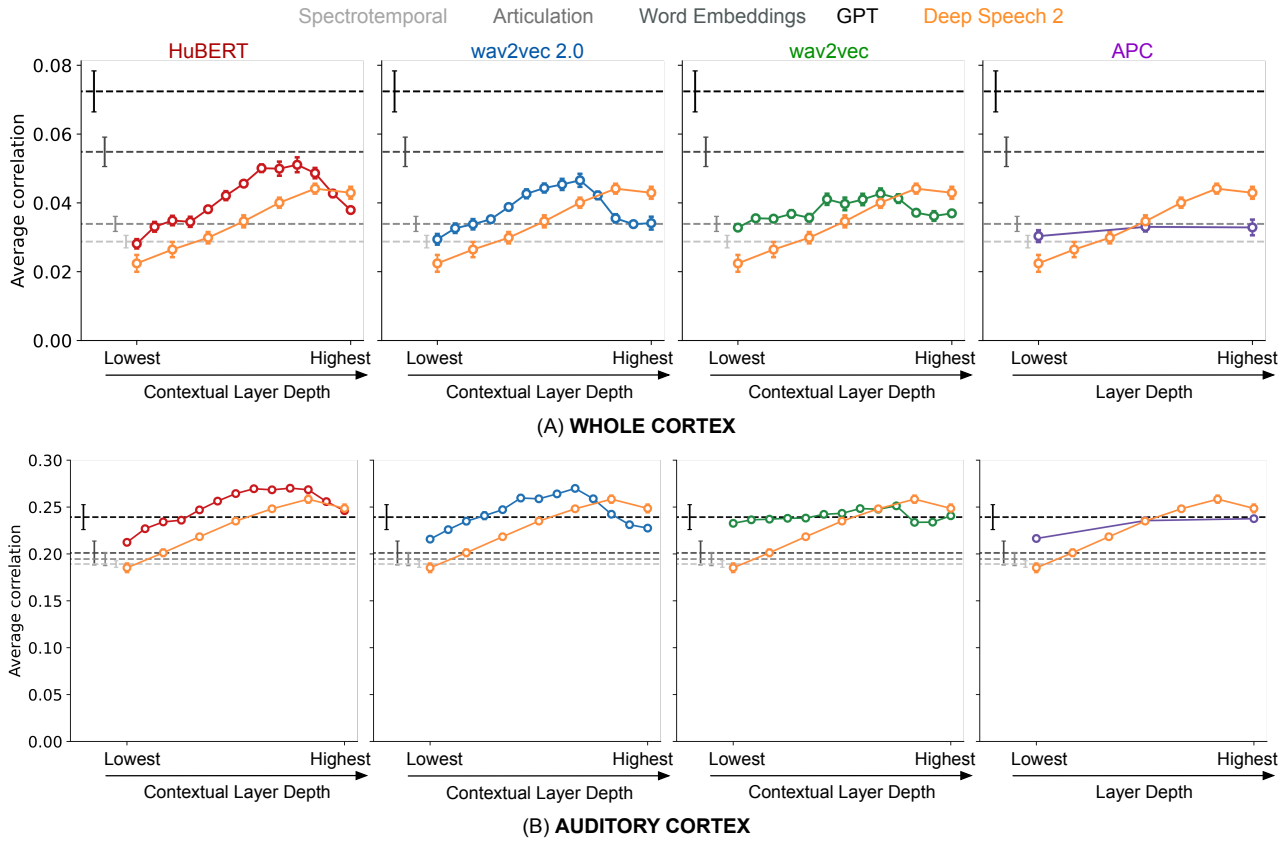

Figure 2. Encoding performance for each feature space, averaged across voxels and then subjects ($N = 7$). Error bars show standard error of the mean (SEM) across subjects after correcting for per-subject overall performance (i.e., subtracting the mean performance across all models from each subject's values). Each column compares one SSL model against baseline representations. (A) Encoding performance is averaged across all voxels in the cortex. HuBERT layer 9 achieves the highest performance, outperforming both auditory baselines. (B) To focus on encoding performance for lower-level features, we averaged only within auditory cortex (broadly defined). All SSL models and Deep Speech 2 outperform the hand-engineered baselines, with HuBERT still outperforming all other models. More speech ROIs can be seen in Appendix A.1. The encoding model built on FBANK features had the worst performance and is not visualized here.

a stride of 10 ms. Before their use in encoding models, all features were down-sampled to the rate of fMRI acquisition (0.5 Hz) with Lanczos resampling. The haemodynamic response function for each encoding model was estimated using a finite impulse response model with 4 delays.

4. Experiments

4.1. Encoding Performance Comparisons

Earlier studies have found that encoding model performance can vary widely across layers of LMs (Jain & Huth, 2018; Toneva & Wehbe, 2019) and supervised speech models (Millet & King, 2021; Kell et al., 2018). To test whether the same is true for SSL speech models, we measured prediction performance for encoding models built using each layer of each network shown in Table 1. For comparison, we also tested encoding models using each layer of the supervised speech baseline and hand-constructed feature baselines.

Figure 2A shows the average voxel prediction performance across the whole cortex for each encoding model. Across layers, this analysis reveals a consistent trend — upper-middle layers of every model have the best encoding performance, with lower performance for the shallowest and deepest layers. This trend is most evident in wav2vec 2.0, HuBERT, and Deep Speech 2, and is consistent with previous literature in higher-level language encoding models that also found best performance in the upper-middle layers (Jain & Huth, 2018; Toneva & Wehbe, 2019; Caucheteux et al., 2021; Goldstein et al., 2020).

Across models, we see that the best layer from each SSL and Deep Speech 2 layer outperform the auditory baselines (articulation and spectrotemporal modulation). Despite being trained on the same amount of data, the best layers of SSL models wav2vec 2.0 and HuBERT outperform the fully supervised Deep Speech 2 model that was used in prior work investigating speech processing across cortex

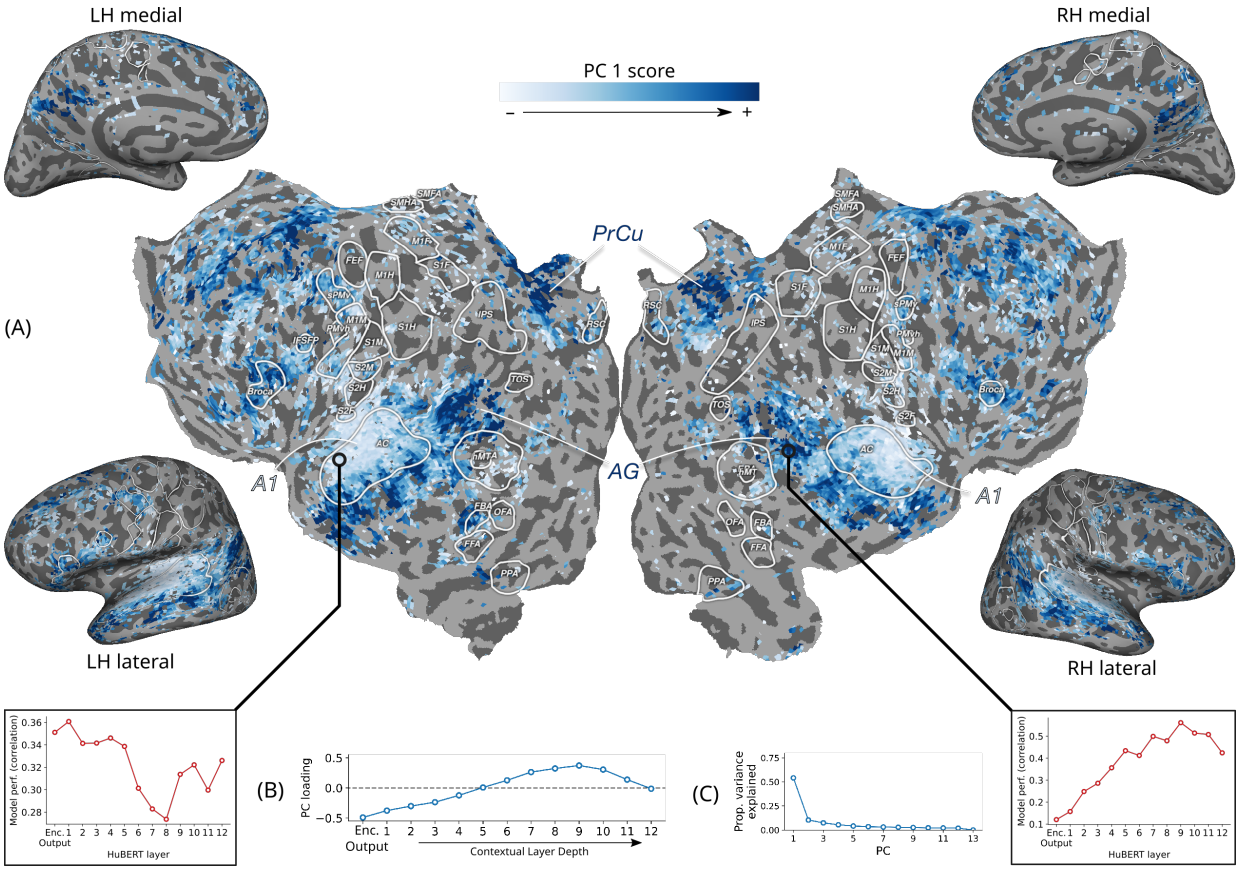

Figure 3. Cortical map of HuBERT layer selectivity. For each individual subject, PCA was applied to an $n_{\text{voxels}} \times n_{\text{layers}}$ matrix C containing the subject's per-voxel encoding performance for each layer in HuBERT. (A) Voxel scores for the first principal component (PC 1) across well-predicted voxels (mean $(\rho_v) > 0.15$) for a single subject. The cortical surface is first inflated (medial and lateral views), and then cut and flattened (flatmap, center). (B) Loadings of PC 1 (averaged across subjects) indicate that the primary dimension of variance is between lower layers and upper-middle layers. Voxels with low scores (white) prefer earlier layers, while those with high scores (blue) prefer later layers. For all loadings, the SEM across subjects was less than 0.02. Insets show the layer-wise encoding performance for a voxel that prefers earlier layers (left) and a voxel that prefers later layers (right). (C) Proportion of variance explained by each PC (averaged across subjects). We focus our analysis on the first PC because it explains much more variance than any other. SEM across subjects was less than 2% for PC 1, and less than 1% for all other PCs. Appendix B shows cortical selectivity maps for additional subjects.

(Millet & King, 2021). Further, the best layer of HuBERT approaches the performance of the word embedding model, despite not receiving any explicit supervision on lexical or semantic information. Nonetheless, there is a substantial gap between even the best SSL audio models and GPT, indicating that the audio models do not make word-level models obsolete.

Figure 2B shows average voxel prediction performance only for voxels within auditory cortex (AC), which is thought to be responsible for extracting higher-level features such as words from incoming acoustic information (Hickok & Poeppel, 2007). Here, most SSL layers and supervised Deep Speech 2 layers outperform all of the hand-constructed feature spaces, including word embeddings. The best encoding model overall is the same, however: layer 9 of HuBERT.

These results could indicate that the best SSL layers are able to capture some of the same semantic information that is in word embeddings, or that the SSL models are simply better acoustic representations than the other models, or a combination of both. Our next analyses are aimed at disentangling what types of information are captured by the different SSL layers.

4.2. Voxel-wise Layer Selectivity

One way to disentangle what is represented in the SSL layers is to compare which brain areas are predicted by each layer. Earlier results provide relatively strong priors about the function of each brain area; for example, primary auditory cortex (A1) on Heschl's gyrus represents the acoustic

properties of incoming sound (Hickok & Poeppel, 2007), while angular gyrus (AG) and precuneus (PrCu) are part of the brain's 'semantic system' (Binder et al., 2009). If a layer is highly predictive of A1 we may infer it contains acoustic information, and if a layer is highly predictive of AG or PrCu we may infer it contains semantic information.

Rather than examining brain maps for each layer separately, we found the major patterns of variation in prediction performance across model layers using principal components analysis (PCA). For an SSL model, we first construct a data matrix matrix C with dimensions $V \times L$, where V is the number of voxels for a subject and L is the number of layers in the model. C_{vl} is then the encoding performance of layer l in the SSL model for voxel v. To account for overall performance differences between layers and voxels, we centered each row and column to have zero mean. Finally, we applied principal components analysis (PCA) to C.

For simplicity, here we focus on the overall best-performing SSL model, HuBERT ($L = 13$). The first principal component (PC) explained 54% of the variance in C, while each subsequent PC explained less than 10% of the variance (Figure 3C). Inspecting the first PC loadings shows that it separates voxels with high performance in the upper-middle layers from voxels with high performance in the lower layers (Figure 3B). We visualized layer selectivity across the cortical surface of one subject by projecting C onto its first PC (Figure 3A).

Overall, layer selectivity roughly follows the hierarchy of speech processing across cortex (Hickok & Poeppel, 2007; Binder et al., 2009). While low-level regions like primary auditory cortex (A1) preferred lower HuBERT layers (negative PC 1 score), high-level regions like the angular gyrus (AG) and precuneus (PrCu) preferred the upper-middle layers (positive PC 1 score). We also compared the first PC scores to voxel encoding performance for low-level acoustic features and word embeddings. As expected from the cortical organization of PC 1 scores, low-level encoding performance was negatively correlated with PC 1 scores ($\rho = -0.330$), while word embedding encoding performance was positively correlated ($\rho = 0.449$) (Appendix B). This suggests that the type of information varies along the depth of SSL models, with the most semantic information appearing at the upper-middle layers and most acoustic information appearing at the lowest layers.

4.3. Partitioning Explained Variance between SSL Models and Hand-Engineered Baselines

Another way to disentangle SSL representations is to measure the overlap in brain variance explained by SSL model layers and the hand-constructed feature spaces. To accomplish this we used variance partitioning, which separates the brain response variance that can be explained by two models into their unique and overlapping contributions (de Heer et al., 2017; LeBel et al., 2021).

For two feature spaces, this is done by fitting separate encoding models for each space as well as a joint encoding model, obtained by concatenating the features. Consider the resultant encoding performances to be ρ_v^1 for feature space 1, ρ_v^2 for feature space 2 and $\rho_v^{1\cup 2}$ for the joint model. The amount of variance in the BOLD responses explained by any encoding model can be approximated as the signed squared correlation coefficient $(\rho)^2 \cdot \text{sgn}(\rho)$, which we will denote $(\rho)^2$ for simplicity. Using set arithmetic, we can then derive the size of the intersection $(\rho^2)_v^{1\cap 2} = (\rho^2)_v^1 + (\rho^2)_v^2 - (\rho^2)_v^{1\cup 2}$ and $\rho_v^{1\cap 2} = \sqrt{(\rho^2)_v^{1\cap 2}}$. This measures the amount of BOLD response in voxel v that can be equally well explained by either of the two feature spaces. Similarly, the unique contribution of feature space 1 can be computed as $(\rho^2)_v^{1\setminus 2} = (\rho^2)_v^1 - (\rho^2)_v^{1\cap 2}$ with $\rho_v^{1\setminus 2} = \sqrt{(\rho^2)_v^{1\setminus 2}}$ and that of feature space 2 as $(\rho^2)_v^{2\setminus 1} = (\rho^2)_v^2 - (\rho^2)_v^{1\cap 2}$ with $\rho_v^{2\setminus 1} = \sqrt{(\rho^2)_v^{2\setminus 1}}$. These values indicate variance uniquely explained by one feature space that is absent from the other.

Variance partitioning quickly becomes intractable as the number of feature spaces increases. Thus we restrict our analyses to pairwise comparisons that involve an upper-middle layer in HuBERT (layer 9) or a lower layer (layer 1). We compared each of these to the hand-constructed spectrotemporal, articulation, and word embedding feature spaces, totaling $2 \times 3 = 6$ comparisons. For robustness, we used banded ridge regression, which allows different regularization parameters for each feature space in the joint encoding model (Nunez-Elizalde et al., 2019). By examining the amount of unique and shared variance between each baseline and HuBERT layer, we can quantify the amount of brain-relevant acoustic, phonetic, or lexical information in each of the HuBERT layers.

Figure 4 shows the partial correlations $\rho_v^{1\cap 2}$, $\rho_v^{1\setminus 2}$ and $\rho_v^{2\setminus 1}$ for each feature space pair. Brain images (above) show the largest component of the three for each voxel across the cortical surface of one participant, and bar plots (below) show the average across voxels and subjects.

We first compared the two HuBERT layers against the lowest-level baseline model, spectrotemporal modulations (Figure 4, column 1). For both layers, there is a relatively large amount of shared variance and very little variance that is uniquely explained by the acoustic baseline. Layer 1 also uniquely explains very little variance, suggesting that it contains very similar information to the spectrotemporal feature space. Layer 9, however, uniquely explains a great deal of variance over the spectrotemporal features.

Next, we compared the HuBERT layers against phoneme articulation features (Figure 4, column 2). Phonemes con-

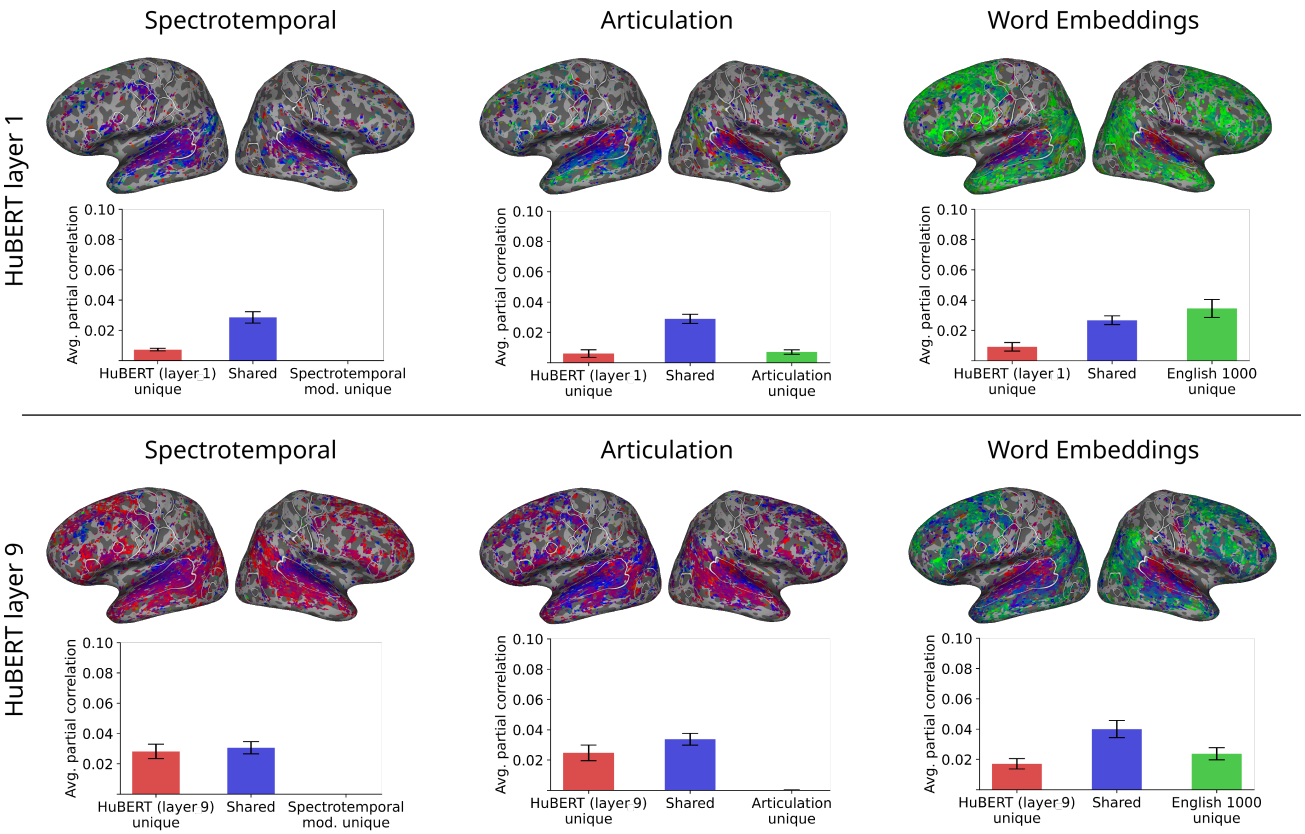

Figure 4. Partitioning variance explained by baselines and HuBERT layers 1 and 9. For each pair of features 1 and 2, bar plots (below) show $\rho_v^{1\setminus 2}$ (red), $\rho_v^{1\cap 2}$ (blue) and $\rho_v^{2\setminus 1}$ (green) averaged across cortex and subjects. Error bars indicate SEM across subjects. Cortical maps (above) show the largest partition per voxel for one subject. Voxels are only shown if $\rho_v^{1\cup 2} > 0.15$. Baselines are ordered by increasing level of abstraction. Every baseline has unique variance when partitioned against layer 1, whereas spectrotemporal (left) and articulatory (middle) features explain no unique variance when partitioned against layer 9. Word embeddings (right) share more variance explained with layer 9 than with layer 1.

stitute more abstract information than acoustic filters and are generally more predictive of cortical activity (Figure 2). Here there is a larger difference between the two HuBERT layers. Layer 1 uniquely explains early auditory cortex, while articulations uniquely explain more lateral and posterior temporal cortex. In contrast, layer 9 appears to be a strict superset of the articulations: there is substantial shared variance, but little or no variance is uniquely explained by the articulatory model. This layer not only captures articulatory features but also contains additional information that is not captured by those features but is relevant to the brain.

Finally, we compared the two HuBERT layers against word embeddings (Figure 4, column 3), which are more abstract than articulatory features. Here, layer 9 has substantially more unique *and* shared variance with word embeddings than layer 1, suggesting that the it contains more semantic information than the lower layer. Across cortex, the word embedding model uniquely explains more variance in high-level regions while layer 1 better explains AC. For layer 9, the pattern is similar but less pronounced, as most variance is shared between the two models.

Overall, these results show that the variance captured by HuBERT layer 9 entirely encompasses the low-level spectrotemporal and mid-level articulatory features, and overlaps to a great extent with the high-level word embedding features. In contrast, HuBERT layer 1 is very similar to the spectrotemporal features but fails to capture some articulatory and most semantic information.

4.4. Probing SSL Models for Linguistic Structure

The previous analyses explained much of why the SSL features are so capable at predicting brain data: the single best layer encompasses spectrotemporal, articulatory, and some semantic information. To get a finer-grained understanding of how linguistic representations evolve across these models, we turned away from the fMRI data and instead compared each layer directly to known linguistic features. Inspired

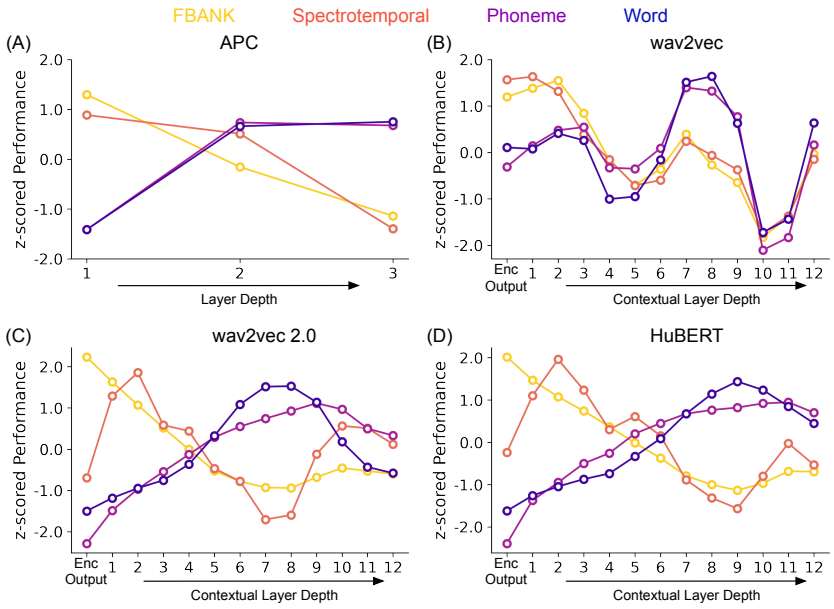

Figure 5. Probing the information represented across layers in SSL models. Each panel shows normalized performance of different layers in an SSL model predicting FBANK features, spectrotemporal features, phoneme identity, and word identity. Upper-middle layers in wav2vec, wav2vec 2.0 and HuBERT best encode phoneme- and word-level information, while the lower layers are better at predicting the acoustic features. This suggests that SSL models recapitulate the putative stages of speech processing.

by prior work in computer vision (Alain & Bengio, 2017), natural language processing (Ettinger et al., 2016; Shi et al., 2016) and speech (Pasad et al., 2021; Yang et al., 2021), we did this by linearly probing each layer's representations for spectral features (FBANK), spectrotemporal features, phoneme identity, and word identity.

To build linear probes, we first obtained parallel SSL model and baseline features using the 27 fMRI stimulus stories. For baseline features with a lower sampling rate (spectrotemporal, phonemes, words), SSL layer representations were mean-pooled across time. To predict FBANK and spectrotemporal features, we used ridge regression to map each SSL model layer to each feature. Probing performance was measured by the correlation between the predicted and true feature value across the test set, averaged across all features in each space. For phoneme and word identity we trained linear classifiers and measured performance with classification accuracy and perplexity, respectively. We used 100-D GloVe embeddings (Pennington et al., 2014) and the associated vocabulary for the word-level tasks. To ensure that these results were not biased by the limited capabilities of linear probes, we repeated some analyses using linear MLPs with a single bottleneck layer, and found no difference (Appendix C). Each linear probe was trained and evaluated on 3 different seeds. For each run, we randomly divided the 27 stories into an 80-10-10 train-validation-test split.

Figure 5 shows the normalized probing performance of different tasks as a function of SSL model depth. These results suggest that the SSL layers recapitulate the putative stages of speech processing (Hickok & Poeppel, 2007; Lieberman, 1970; Pisoni & Sawusch, 1975). For most models the simplest features (FBANK) are best captured by the lowest layers, slightly more complex features (spectrotemporal) by the lower-middle layers, and high-level features (words) by the upper-middle layers. Yet phonemes, which are generally thought to fall between spectrotemporal and word-level features in speech analysis (Hickok & Poeppel, 2007), are best captured by the uppermost layers in wav2vec 2.0 and HuBERT. This may reflect how phoneme perception is influenced by word context (Elman & McClelland, 1988) rather than being purely based on acoustics. This effect appears to mirror human behavior despite the fact that these models have no explicit representation of words or phonemes.

5. Conclusion

We developed computational models of speech processing in the human cortex using representations from four different self-supervised learning (SSL) models – APC, wav2vec, wav2vec 2.0 and HuBERT. Voxel-wise encoding models using SSL-derived features were better models of cortical activity than either phonemes or traditional acoustic features like spectrotemporal modulations and spectrograms. The best layers of wav2vec 2.0 and HuBERT also outperformed features from supervised speech models used in prior work

and approached the performance of word embedding models that capture semantic information.

To better understand what information is captured by these models, we first analyzed which model layers best predicted different areas of cortex. This showed that lower layers effectively modeled areas involved in low-level acoustic processing, while the upper layers were better predictors of areas capturing phonetic and semantic information. Second, we measured the degree to which SSL layers and hand-engineered feature spaces predicted the same or different variance in brain responses. This suggested that the best SSL layer achieved its encoding performance by capturing features across several scales, including spectrotemporal, phonemic, and some semantic information. Finally, we directly probed the SSL layers for acoustic and linguistic features, demonstrating that SSL models seem to recapitulate the speech processing hierarchy surprisingly well.

Overall, our results suggest that deeper SSL model layers capture increasingly high-level features while maintaining low-level information. Coupled with our finding that SSL models are currently the best sound-based models of cortical speech processing, this suggests that further development and analysis of SSL models may reveal much about how speech is processed by the human brain.

Acknowledgments

We would like to thank David Harwath, as well as the anonymous reviewers, for their insights and suggestions. Funding for this work was provided by the Burroughs-Wellcome Fund Career Award at the Scientific Interface (CASI), the Whitehall Foundation, and the Alfred P. Sloan Foundation.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. February 2017.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L. V., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., and Zhu, Z. Deep speech 2: End-to-end speech recognition in English and mandarin. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 173–182, New York, NY, USA, June 2016. JMLR.org.

Antonello, R., Turek, J. S., Vo, V. A., and Huth, A. Low-dimensional Structure in the Space of Language Representations is Reflected in Brain Responses. In *Advances in Neural Information Processing Systems*, May 2021.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv:2006.11477 [cs, eess]*, September 2020.

Binder, J. R., Desai, R. H., Graves, W. W., and Conant, L. L. Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex*, 19(12):2767–2796, December 2009. ISSN 1047-3211. doi: 10.1093/cercor/bhp055.

Caucheteux, C., Gramfort, A., and King, J.-R. GPT-2's activations predict the degree of semantic comprehension in the human brain, September 2021.

Chi, T., Gao, Y., Guyton, M. C., Ru, P., and Shamma, S. Spectro-temporal modulation transfer functions and speech intelligibility. *The Journal of the Acoustical Society of America*, 106(5):2719, October 1999. ISSN 0001-4966. doi: 10.1121/1.428100.

Chi, T., Ru, P., and Shamma, S. A. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906, August 2005. ISSN 0001-4966. doi: 10.1121/1.1945807.

Chung, Y.-A. and Glass, J. Generative Pre-Training for Speech with Autoregressive Predictive Coding. *arXiv:1910.12607 [cs, eess]*, January 2020.

Chung, Y.-A., Hsu, W.-N., Tang, H., and Glass, J. An Unsupervised Autoregressive Model for Speech Representation Learning. *arXiv:1904.03240 [cs, eess]*, June 2019.

Dale, A. M., Fischl, B., and Sereno, M. I. Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage*, 9(2):179–194, February 1999. ISSN 1053-8119. doi: 10.1006/nimg.1998.0395.

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., and Theunissen, F. E. The Hierarchical Cortical Organization of Human Speech Processing. *The Journal of Neuroscience*, 37(27):6539–6557, July 2017. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3267-16.2017.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Elman, J. L. and McClelland, J. L. Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, 27(2):143–165, April 1988. ISSN 0749-596X. doi: 10.1016/0749-596X(88)90071-X.

Ettinger, A., Elgohary, A., and Resnik, P. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 134–139, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2524.

Gao, J. S., Huth, A. G., Lescroart, M. D., and Gallant, J. L. Pycortex: An interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, 9, 2015. ISSN 1662-5196.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, S. C., Casto, C., Fanda, L., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Reichart, R., Devore, S., Flinker, A., Hasenfratz, L., Levy, O., Hassidim, A., Brenner, M., Matias, Y., Norman, K. A., Devinsky, O., and Hasson, U. Thinking ahead: Spontaneous prediction in context as a keystone of language in humans and machines. Preprint, Neuroscience, December 2020.

Hewitt, J. and Liang, P. Designing and Interpreting Probes with Control Tasks. *arXiv:1909.03368 [cs]*, September 2019.

Hickok, G. and Poeppel, D. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5): 393–402, May 2007. ISSN 1471-0048. doi: 10.1038/nrn2113.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv:2106.07447 [cs, eess]*, June 2021.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, April 2016. ISSN 1476-4687. doi: 10.1038/nature17637.

Jain, S. and Huth, A. Incorporating Context into Language Encoding Models for fMRI. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 6628–6637. Curran Associates, Inc., 2018.

Jain, S., Vo, V., Mahto, S., LeBel, A., Turek, J. S., and Huth, A. Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. *Advances in Neural Information Processing Systems*, 33:13738–13749, 2020.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., and McDermott, J. H. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644.e16, May 2018. ISSN 0896-6273. doi: 10.1016/j.neuron.2018.03.044.

LeBel, A., Jain, S., and Huth, A. G. Voxelwise Encoding Models Show That Cerebellar Language Representations Are Highly Conceptual. *Journal of Neuroscience*, 41 (50):10341–10355, December 2021. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0118-21.2021.

Lieberman, P. Towards a Unified Phonetic Theory. *Linguistic Inquiry*, 1(3):307–322, 1970. ISSN 0024-3892.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, 343(6174):1006–1010, February 2014. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1245994.

Millet, J. and King, J.-R. Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech. *arXiv:2103.01032 [cs, eess, q-bio]*, February 2021.

Norman-Haignere, S. V. and McDermott, J. H. Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLOS Biology*, 16(12):e2005127, December 2018. ISSN 1545-7885. doi: 10.1371/journal.pbio.2005127.

Nunez-Elizalde, A. O., Huth, A. G., and Gallant, J. L. Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage*, 197:482–492, August 2019. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2019.04.012.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, South Brisbane, Queensland, Australia, April 2015. IEEE. ISBN 978-1-4673-6997-8. doi: 10.1109/ICASSP.2015.7178964.

Pasad, A., Chou, J.-C., and Livescu, K. Layer-wise Analysis of a Self-supervised Speech Representation Model. *arXiv:2107.04734 [cs, eess]*, October 2021.

Pennington, J., Socher, R., and Manning, C. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202.

Pisoni, D. B. and Sawusch, J. R. Some Stages of Processing in Speech Perception. In Cohen, A. and Nooteboom, S. G. (eds.), *Structure and Process in Speech Perception*, Communication and Cybernetics, pp. 16–35, Berlin, Heidelberg, 1975. Springer. ISBN 978-3-642-81000-8. doi: 10.1007/978-3-642-81000-8_2.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving Language Understanding by Generative Pre-Training. pp. 12, 2018.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. Wav2vec: Unsupervised Pre-training for Speech Recognition. *arXiv:1904.05862 [cs]*, September 2019.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118 (45), November 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2105646118.

Shi, X., Padhi, I., and Knight, K. Does String-Based Neural MT Learn Source Syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1526–1534, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1159.

Toneva, M. and Wehbe, L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, pp. 14954–14964, 2019.

Venezia, J. H., Thurman, S. M., Richards, V. M., and Hickok, G. Hierarchy of speech-driven spectrotemporal receptive fields in human auditory cortex. *NeuroImage*, 186:647–666, February 2019. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2018.11.049.

Wehbe, L., Vaswani, A., Knight, K., and Mitchell, T. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 233–243, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1030.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6.

Wu, M. C.-K., David, S. V., and Gallant, J. L. Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, 29(1): 477–505, July 2006. ISSN 0147-006X, 1545-4126. doi: 10.1146/annurev.neuro.29.051605.113024.

Yang, S.-w., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G.-T., Huang, T.-H., Tseng, W.-C., Lee, K.-t., Liu, D.-R., Huang, Z., Dong, S., Li, S.-W., Watanabe, S., Mohamed, A., and Lee, H.-y. SUPERB: Speech processing Universal PERformance Benchmark. *arXiv:2105.01051 [cs, eess]*, October 2021.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., and Yamins, D. L. K. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), January 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2014196118.

A. Voxel-wise encoding models of speech

We determine the significance of performance differences between layers via significance tests. Figure 6 shows the results for all layers of HuBERT.

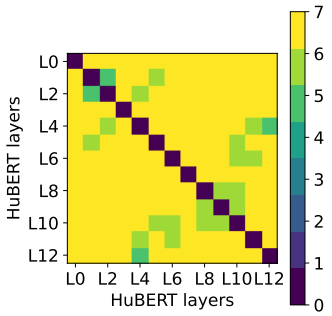

Figure 6. Pairwise significance tests between each layer of HuBERT. The color of cell i, j indicates the number of subjects for which the voxel-wise encoding performance of layer i was significantly different from that of layer j ($p < 0.01$ for a two-sided t-test). Layer 0 ("L0") is the encoder layer.

A.1. Performance in speech ROIs

In Figure 7, we show the encoding performance of each representation in two additional ROIs: Broca's area and sPMv.

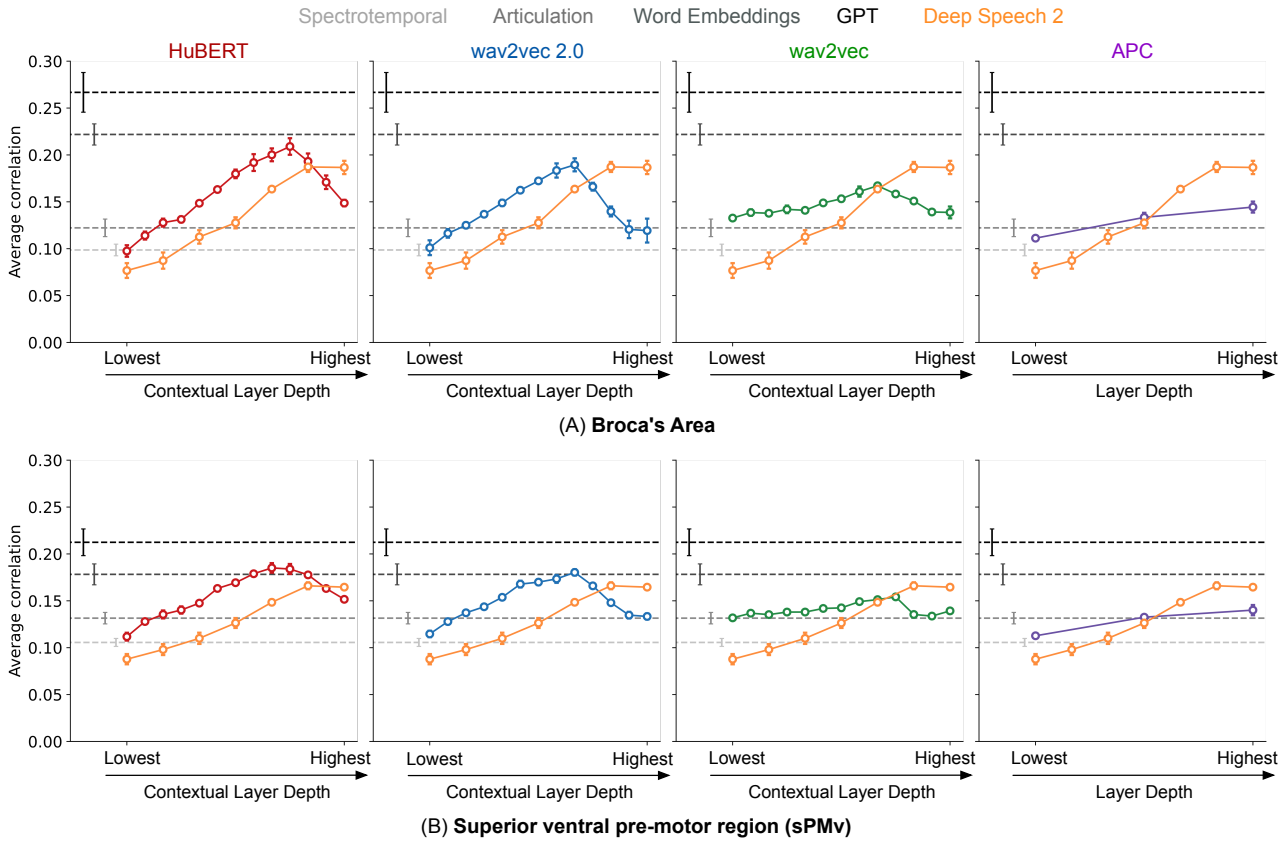

Figure 7. Average encoding performance across subjects ($N = 7$), for every baseline and SSL model representation in two speech ROIs. As with Figure 2, error bars show subject-adjusted SEM.

B. Voxel-wise layer selectivity

We show the relationship between HuBERT layer selectivity and baseline encoding performance in Figure 8.

(a) (b)

Figure 8. 2D histograms of voxels showing correlation between voxel-wise HuBERT layer selectivity (PC 1 in Figure 3) and two baselines. (a) Correlation of PC 1 and word embedding encoding performance across all of cortex are correlated ($\rho = 0.449$), suggesting that PC 1 is positively correlated with semantics. (b) Correlation of PC 1 and spectrotemporal mod. encoding performance within voxels in AC. That these are negatively correlated ($\rho = -0.330$) gives further evidence that PC 1 is capturing higher-level processing.

Layer selectivity maps for six subjects are shown in Figure 9.

C. Probing SSL model representations for linguistic structure

We visualize the full set of probing results for the four tasks (FBANK features, spectrotemporal features, phoneme identity, word identity) for each layer of the four SSL models (Figure 10-Figure 13). Overall, the trends are consistent with the summaries reported in Figure 5 — lower layers best predict low-level acoustic features, while the upper-middle layers are better at the phoneme and word tasks. In addition to the four probes, for HuBERT we also built MLP probes that predict the entire FBANK feature/GloVe embedding as opposed to regression probes that treat each feature in the representation independently. Such probes allow us to test the geometry of HuBERT's representations. To do this, we trained a linear MLP with a single bottleneck layer (50-D). The model was trained on MSE loss and we applied L2 regularization to the output. Performance was measured as the linear correlation between true and predicted representation, averaged across the test set.

For the phoneme and word classification tasks, we developed a baseline that predicts the most frequent output category in the training set for all test set examples. All layers in the four SSL models beat the baseline. We do not visualize the word classification baseline perplexity since it is significantly worse than the highest layer PPX (perplexity). For the word embedding MLP probe, we constructed two different baselines – randomly sampling embedding vectors from a normal distribution ("random") and randomly shuffling the GloVe embeddings between words ("shuffle"). All layers in HuBERT beat the baseline with a considerable gap, indicating that the model captures semantic similarity. (The baselines remove similarity from their representations.) Overall, the substantial differences in the performance of probing baselines and the SSL model layers suggest that the probes have high *selectivity* (Hewitt & Liang, 2019).

D. MRI acquistion, preprocessing, and experiment details

D.1. Participants

All participants were healthy and had normal hearing, and normal or corrected-to-normal vision. To stabilize head motion during scanning sessions participants wore a personalized head case that precisely fit the shape of each participant's head. Anatomical data for subject S-02 were collected on a 3T Siemens TIM Trio scanner using a 32-channel Siemens volume coil at a different site. The same MP-RAGE sequence was used.

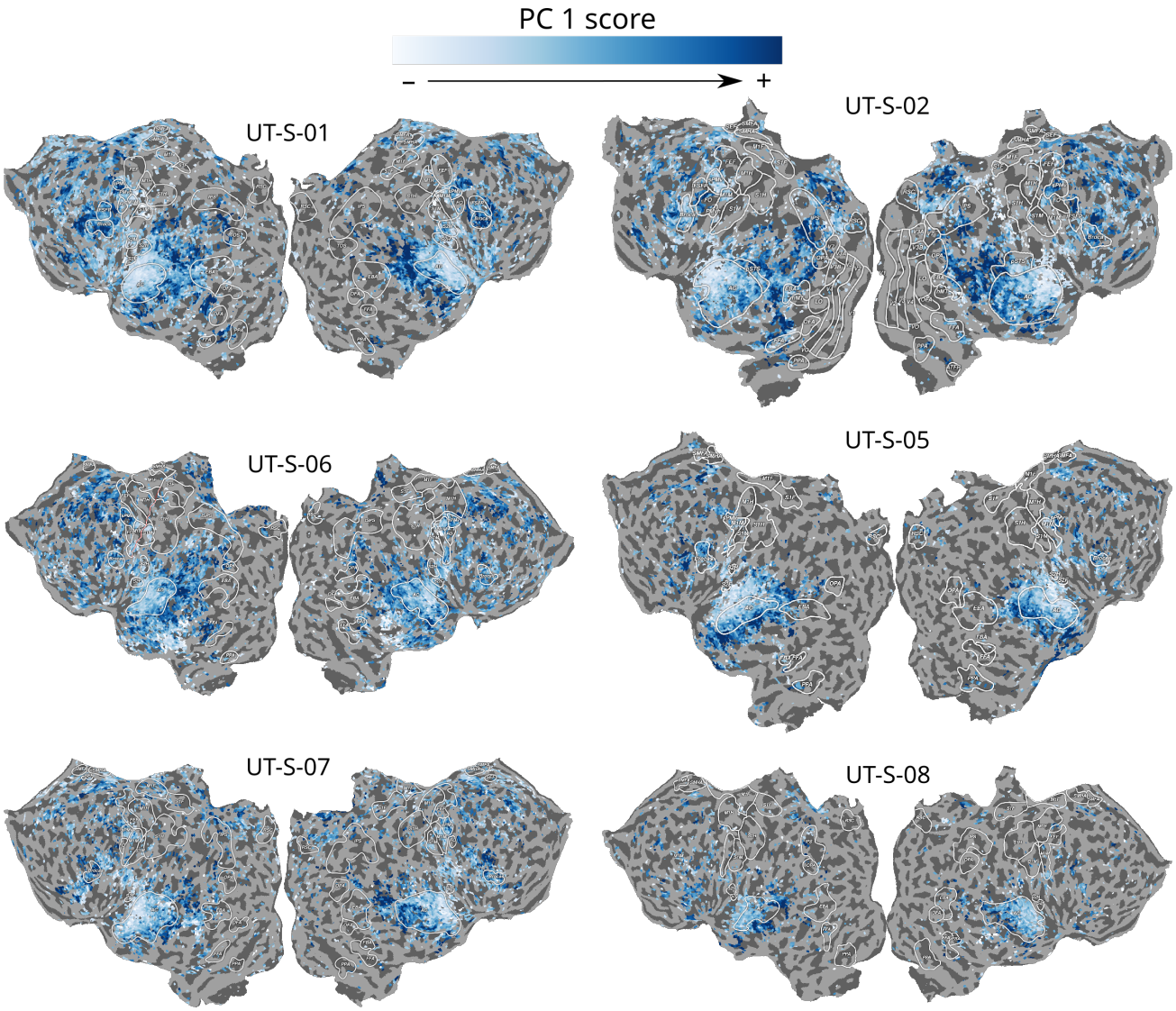

Figure 9. Layer selectivity maps for HuBERT. Shown are the flatmaps for six more subjects in the fMRI dataset. (The subject presented in the main text is UT-S-03.) PCA was computed individually for each subject using the procedure described in Section 4.2, and visualized above is each voxel's score for PC 1. Per the loadings averaged across all subjects in Figure 3B, lighter voxels are better predicted by earlier layers, while bluer voxels are better predicted by later layers.

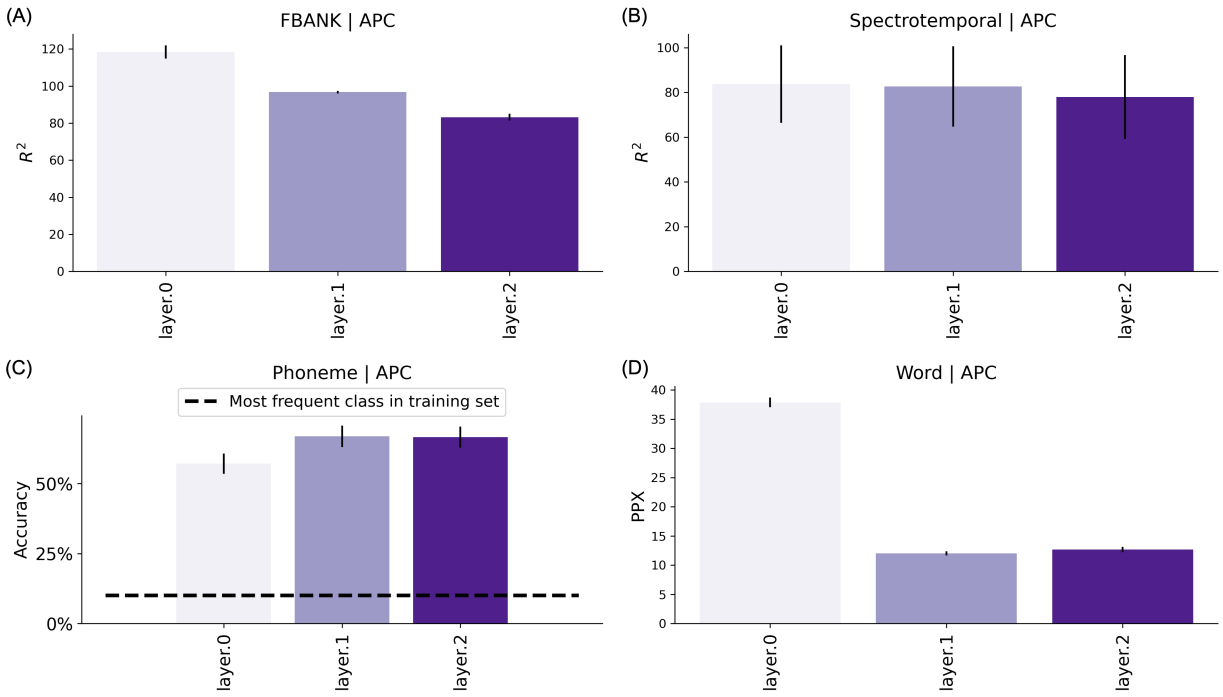

Figure 10. Probing results for APC.

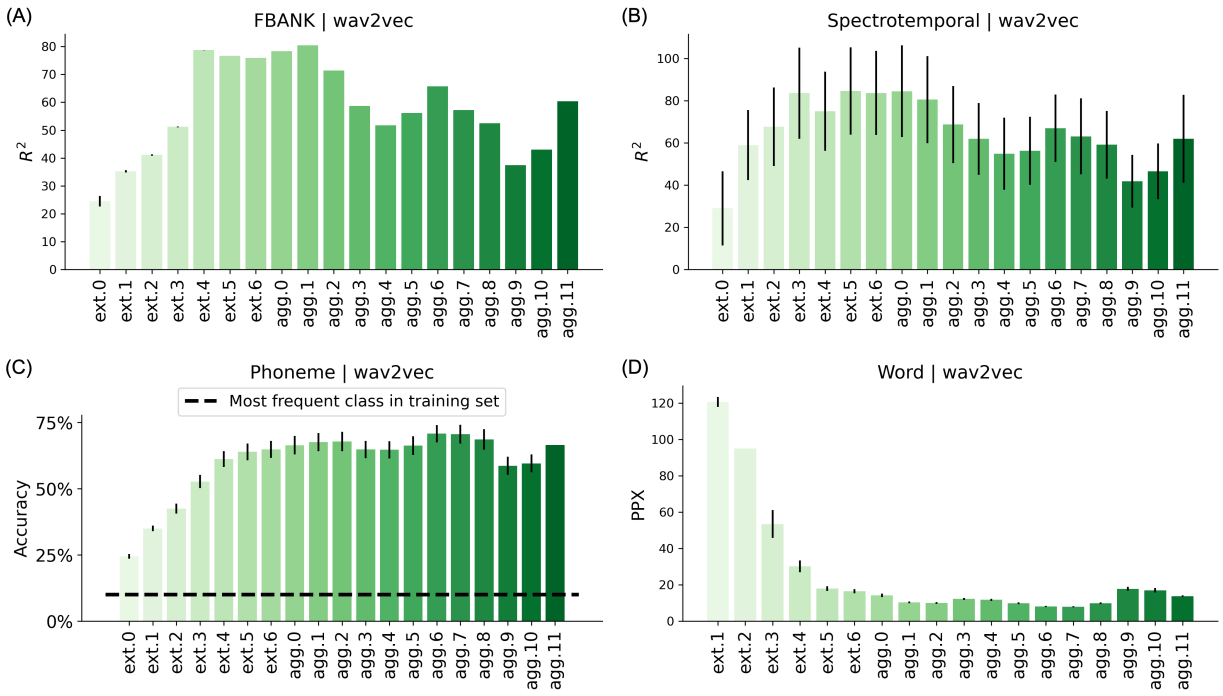

Figure 11. Probing results for wav2vec.

D.2. Stimulus preparation and presentation

Story stimuli were played over Sensimetrics S14 in-ear piezoelectric headphones. The audio for each story was filtered to correct for frequency response and phase errors induced by the headphones using calibration data provided

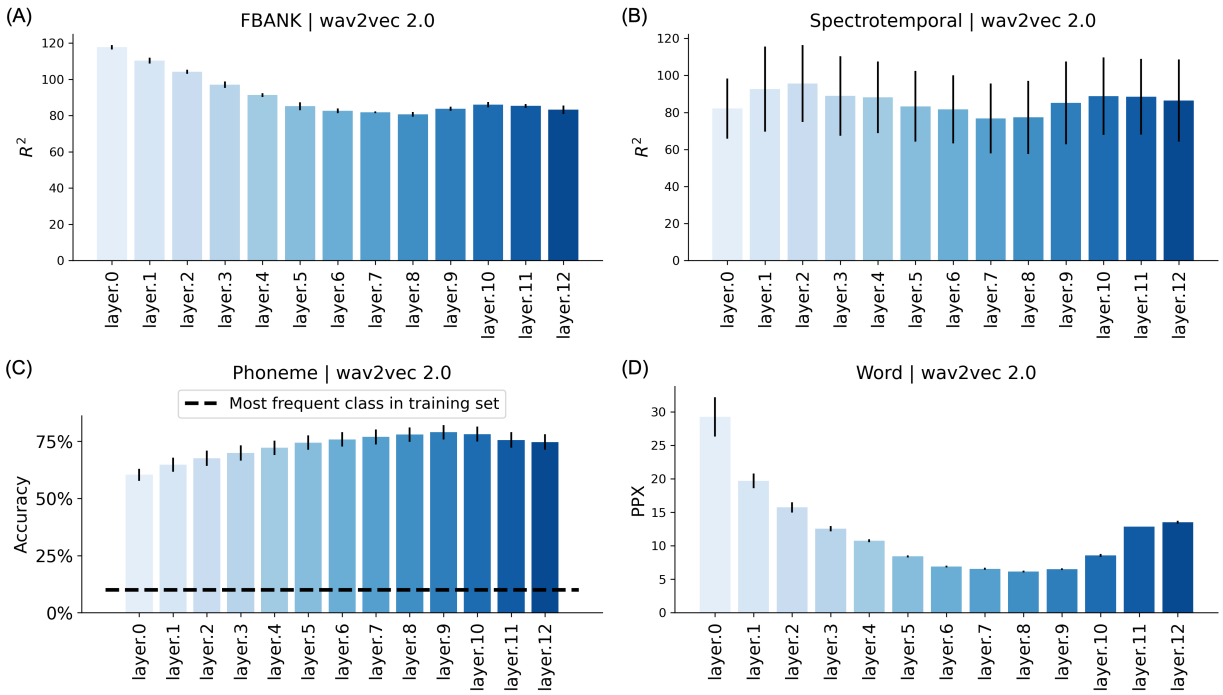

Figure 12. Probing results for wav2vec 2.0.

by sensimetrics and custom Python code[1] All stimuli were played at 44.1 kHz using the pygame library in Python (https://www.pygame.org/news).

D.3. Acquisition parameters

Whole-brain MRI data was collected on a 3T Siemens Skyra scanner using a 64-channel Siemens volume coil. Functional MRI (fMRI) data were collected using a gradient echo EPI sequence, multi-band factor of 2. Scan parameters included repetition time (TR)=2.00s, echo time (TE)=30.8 ms, flip angle=71°, voxel size=2.6mm^3, matrix size=84x84, field of view=220 mm. Anatomical MRI data were collected with a T1-weighted multi-echo MP-RAGE sequence with voxel size=1mm^3.

D.4. Processing MRI data

Functional data. All functional data were motion corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL 5.0. FLIRT was used to align all data to a template that was made from the average across the first functional run in the first story session for each subject. These automatic alignments were manually checked for accuracy.

Low frequency voxel response drift was identified using a 2nd order Savitzky-Golay filter with a 120 second window and then subtracted from the signal. To avoid onset artifacts and poor detrending performance near each end of the scan, responses were trimmed by removing 20 seconds (10 volumes) at the beginning and end of each scan, which removed the 10-second silent period and the first and last 10 seconds of each story. The mean response for each voxel was subtracted and the remaining response was scaled to have unit variance.

To account for physiological and behavioral noise, encoding models for all analyses were fit with responses that had been corrected for nuisance regressors that accounted for closed-eye eye movements and head motion. This has little effect on the overall model performance, but accounts for spurious eye movement and stimulus-related responses in visual cortex.

[1] https://github.com/alexhuth/sensimetrics_filter

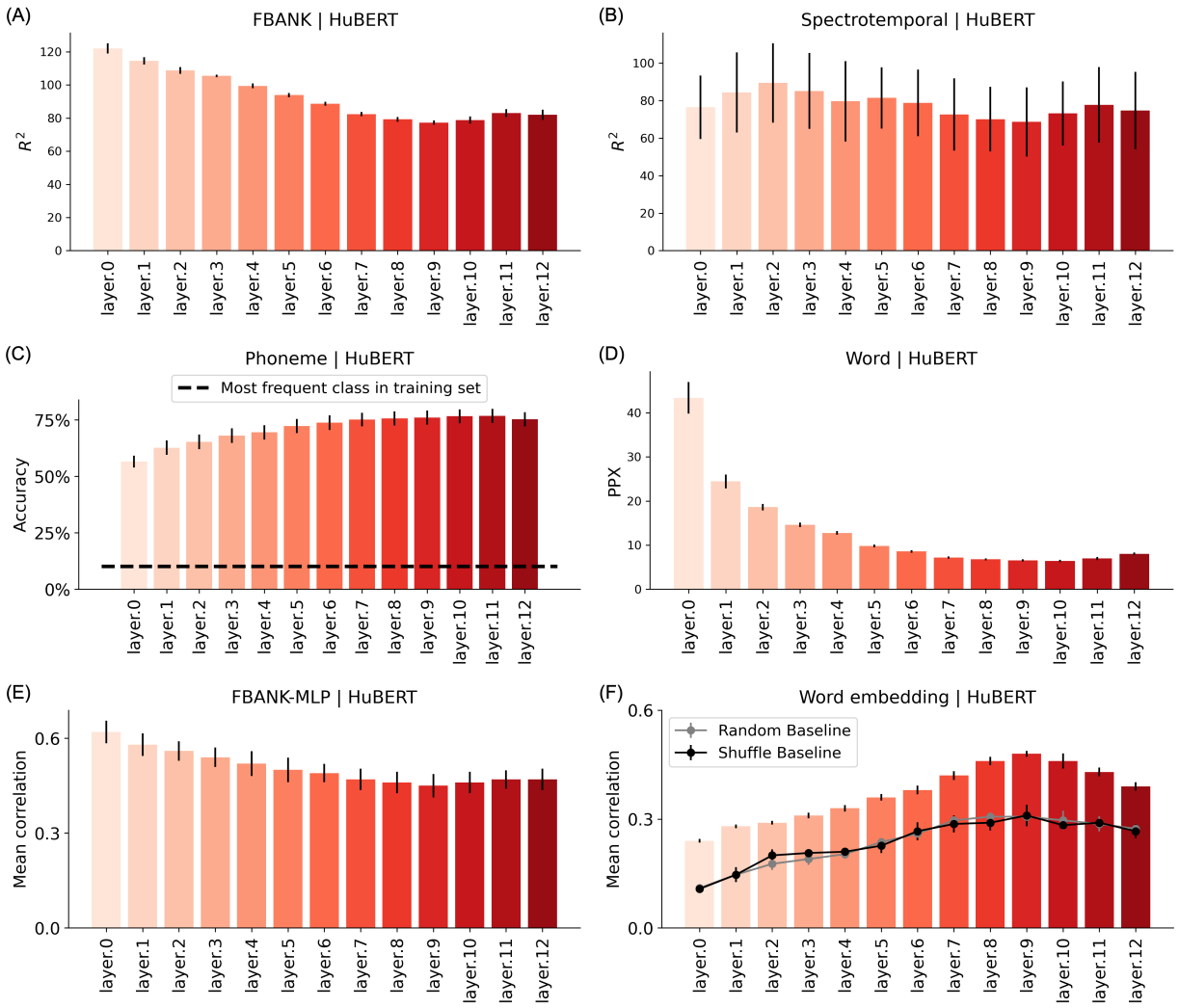

Figure 13. Probing results for HuBERT.

Anatomical data. Cortical surface meshes were generated from the T1-weighted anatomical scans using FreeSurfer (Dale et al., 1999). Before surface reconstruction, anatomical surface segmentations were hand-checked and corrected. Blender was used to remove the corpus callosum and make relaxation cuts for flattening. Functional images were aligned to the cortical surface using boundary based registration (BBR) in FSL. These alignments were manually checked for accuracy and adjustments were made as necessary. Flat maps were created by projecting the values for each voxel onto the cortical surface using the "nearest" scheme in pycortex (Gao et al., 2015).

D.5. Defining regions of interest (ROIs)

Known regions of interest (ROIs) were localized separately in each participant. Three different tasks were used to define ROIs; a visual category localizer, an auditory cortex localizer, and a motor localizer.

For the visual category localizer, data were collected in six 4.5 minute scans consisting of 16 blocks of 16 seconds each. During each block 20 images of either places, faces, bodies, household objects, or spatially scrambled objects were displayed. participants were asked to pay attention to the same image being presented twice in a row. The corresponding ROIs defined in the cerebral cortex with this localizer were the fusiform face area (FFA), occipital face area (OFA), extrastriate body area (EBA), parahippocampal place area (PPA), and occipital place area (OPA).

The motor localizer data were was collected during 2 identical 10-minute scans. The participant was cued to perform six

different tasks in a random order in 20 second blocks. The cues were “hand”, “foot”, “mouth”, “speak”, saccade, and “rest” presented as a word at the center of the screen, except for the saccade cue which was presented as an array of dots. For the “hand” cue, participants were instructed to make small finger-drumming movements for the entirety of the cue display. For the “foot” cue, participants were instructed to make small foot and toe movements. For the “mouth” cue, participants were instructed to make small vocalizations that were nonsense syllables such as balabalabala. For the “speak” cue, participants were instructed to self-generate a narrative without vocalization. For the saccade cue, participants were instructed to make frequent saccades across the display screen for the duration of the task.

Weight maps for the motor areas were used to define primary motor and somatosensory areas for the hands, feet, and mouth; supplemental motor areas for the hands and feet, secondary somatosensory areas for the hands, feet, and mouth, and the ventral premotor hand area. The weight map for the saccade responses was used to define the frontal eye fields and intraparietal sulcus visual areas. The weight map for speech was used to define Broca’s area and the superior ventral premotor (sPMv) speech area.

Auditory cortex localizer data were collected in one 10 minute scan. The participant listened to 10 repeats of a 1-minute auditory stimulus containing 20 seconds of music (Arcade Fire), speech (Ira Glass, This American Life), and natural sound (a babbling brook). To determine whether a voxel was responsive to auditory stimulus, the repeatability of the voxel response across the 10 repeats was calculated using an F-statistic. This map was used to define the auditory cortex (AC).

Data availability. The fMRI data used in this study are publicly available at https://openneuro.org/datasets/ds003020/versions/1.0.2.