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-
I reached over and secretly undid my seatbelt and when his hands hit the brake I jumped...
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.

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

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
Self-supervised models of audio effectively explain human cortical responses to speech

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.

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.

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.

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.
Self-supervised models of audio effectively explain human cortical responses to speech

(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

Figure 3. Cortical map of HuBERT layer selectivity. For each individual subject, PCA was applied to an n<sub>voxels</sub> × n<sub>layers</sub> 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 (ρ<sub>v</sub>) > 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.
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 with dimensions V x 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 and most 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 (ρ = −0.330), while word embedding encoding performance was positively correlated (ρ = 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 ρ_v^{1/2} 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 (ρ)^2 · sgn(ρ), which we will denote (ρ)^2 for simplicity. Using set arithmetic, we can then derive the size of the intersection (ρ_v^{1/2})_v = (ρ_v^1)_v + (ρ_v^2)_v − (ρ_v^{1/2})_v and (ρ_v^{1/2})_v = √((ρ_v^1)_v · (ρ_v^2)_v). 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 (ρ_v^{1/2})_v = (ρ_v^1)_v − (ρ_v^{1/2})_v with ρ_v^{1/2} = √((ρ_v^1)_v · (ρ_v^2)_v) and that of feature space 2 as (ρ_v^{1/2})_v = (ρ_v^2)_v − (ρ_v^{1/2})_v with ρ_v^{1/2} = √((ρ_v^1)_v · (ρ_v^2)_v). 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 x 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 ρ_v^{1/2}, ρ_v^{1/2} and ρ_v^{1/2} 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-
Self-supervised models of audio effectively explain human cortical responses to speech

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 $\rho_{1 \setminus 2}$ (red), $\rho_{1 \setminus 2}$ (blue) and $\rho_{1 \setminus 2}$ (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 $\rho_{1 \setminus 2} > 0.15$. 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.

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
Self-supervised models of audio effectively explain human cortical responses to speech

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.

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


Self-supervised models of audio effectively explain human cortical responses to speech
Self-supervised models of audio effectively explain human cortical responses to speech


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.

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 acquisition, 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.
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.
Self-supervised models of audio effectively explain human cortical responses to speech

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
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 collected during 2 identical 10-minute scans. The participant was cued to perform six
Self-supervised models of audio effectively explain human cortical responses to speech

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.