Title
The Hierarchical Cortical Organization of Human Speech Processing

Journal
Journal of Neuroscience, 37(27)

ISSN
0270-6474

Authors
de Heer, Wendy A
Huth, Alexander G
Griffiths, Thomas L
et al.

Publication Date
2017-07-05

DOI
10.1523/jneurosci.3267-16.2017

Peer reviewed

Behavioral/Cognitive

The Hierarchical Cortical Organization of Human Speech Processing

Wendy A. de Heer,* Alexander G. Huth,* Thomas L. Griffiths, Jack L. Gallant, and Frédéric E. Theunissen
Department of Psychology and Helen Wills Neurosciences Institute, University of California, Berkeley, Berkeley, California 94720

Speech comprehension requires that the brain extract semantic meaning from the spectral features represented at the cochlea. To investigate this process, we performed an fMRI experiment in which five men and two women passively listened to several hours of natural narrative speech. We then used voxelwise modeling to predict BOLD responses based on three different feature spaces that represent the spectral, articulatory, and semantic properties of speech. The amount of variance explained by each feature space was then assessed using a separate validation dataset. Because some responses might be explained equally well by more than one feature space, we used a variance partitioning analysis to determine the fraction of the variance that was uniquely explained by each feature space. Consistent with previous studies, we found that speech comprehension involves hierarchical representations starting in primary auditory areas and moving laterally on the temporal lobe: spectral features are found in the core of A1, mixtures of spectral and articulatory in STG, mixtures of articulatory and semantic in STS, and semantic in STS and beyond. Our data also show that both hemispheres are equally and actively involved in speech perception and interpretation. Further, responses as early in the auditory hierarchy as in STS are more correlated with semantic than spectral representations. These results illustrate the importance of using natural speech in neurolinguistic research. Our methodology also provides an efficient way to simultaneously test multiple specific hypotheses about the representations of speech without using block designs and segmented or synthetic speech.

Key words: fMRI; natural stimuli; regression; speech

Significance Statement

To investigate the processing steps performed by the human brain to transform natural speech sound into meaningful language, we used models based on a hierarchical set of speech features to predict BOLD responses of individual voxels recorded in an fMRI experiment while subjects listened to natural speech. Both cerebral hemispheres were actively involved in speech processing in large and equal amounts. Also, the transformation from spectral features to semantic elements occurs early in the cortical speech-processing stream. Our experimental and analytical approaches are important alternatives and complements to standard approaches that use segmented speech and block designs, which report more laterality in speech processing and associated semantic processing to higher levels of cortex than reported here.

Introduction

The process of speech comprehension is often viewed as a series of computational steps that are carried out by a hierarchy of processing modules in the brain, each of which has a distinct functional role (Stowe et al., 2005; Price, 2010; Poeppel et al., 2012; Bornkessel-Schlesewsky et al., 2015). The classical view holds that the acoustic spectrum is analyzed in primary auditory cortex (AC; Hullett et al., 2016), then phonemes are extracted in secondary auditory areas, and words are extracted in lateral and ventral temporal cortex (for review, see DeWitt and Rauschecker, 2012). The meaning of the word sequence is then inferred based on syntactic and semantic properties by a network of temporal, parietal, and frontal areas (Rodd et al., 2005; Binder et al., 2009; Just et al., 2010; Visser et al., 2010; Fedorenko et al., 2012). Recent refinements to this serial view hold that speech comprehension

Received Oct. 21, 2016; revised May 22, 2017; accepted May 25, 2017.

Author contributions: W.A.d.H., A.G.H., T.L.G., J.L.G., and F.E.T. designed research; W.A.d.H. and A.G.H. performed research; W.A.d.H., A.G.H., and F.E.T. analyzed data; W.A.d.H., A.G.H., J.L.G., and F.E.T. wrote the paper.

*W.A.d.H. and A.G.H. contributed equally to this work.

The authors declare no competing financial interests.

The work was supported by grants from the National Science Foundation (IIS1208203); the National Eye Institute (EY019684); the National Institute on Deafness and Other Communication Disorders (NIDCD 007293); and the Center for Science of Information, a National Science Foundation Science and Technology Center, under grant agreement CCF-0939370. A.G.H. was also supported by the William Orr Dingwall Neurolinguistics Fellowship and a Career Award at the Scientific Interface from the Burroughs-Wellcome Fund.

Correspondence should be addressed to Frédéric E. Theunissen, 3210 Tolman Hall, University of California, Berkeley, Berkeley, CA 94720-1650. E-mail: theunissen@berkeley.edu.

DOI:10.1523/JNEUROSCI.3267-16.2017
 0270-6474/17/376539-19$15.00/0

separates into the following two streams (Turkeltaub and Coslett, 2010; Bornkessel-Schlesewsky et al., 2015): an anteroventral stream involved in auditory object recognition (DeWitt and Rauschecker, 2012) and a posterodorsal stream involved in processing temporal sequences (Belin and Zatorre, 2000); and sensorimotor transformations that represent speech sounds in terms of articulatory features (Scott and Johnsrude, 2003; Rauschecker and Scott, 2009). In principle, a hierarchically organized speech-processing system could contain multiple, mixed representations of various speech features. However, most neuroimaging studies assume that each brain area is dedicated to one computational step or level of representation and are designed to isolate a single computational step from the rest of the speech-processing hierarchy. In a typical experiment, responses are measured to stimuli that differ along a single dimension of interest (Binder et al., 2000; Leaver and Rauschecker, 2010); for review, see DeWitt and Rauschecker, 2012), and then subtraction is used to find brain regions that respond significantly more to one end of this dimension than the other. Although this approach provides substantial power for testing specific hypotheses about speech representation, investigating the entire speech-processing hierarchy this way is inefficient because it inevitably requires many separate experiments followed by meta-analyses. Furthermore, examining each computational step individually provides little information about relationships between representations at different levels in the speech-processing hierarchy.

Another limitation of most studies of speech processing is that they do not use natural stimuli. Neuroimaging experiments often use isolated sounds or segmented speech stimuli that are as different from natural language as sine-wave gratings are from natural images. It is well known that factors such as sentence intelligibility (Peelle et al., 2013) and attention (Mesgarani and Chang, 2012; Zion Golumbic et al., 2013) influence brain activity even in primary auditory areas, and these factors are likely to differ between non-natural stimuli and natural narrative speech. Because the speech-processing hierarchy likely contains important nonlinearities, it is unclear how well standard neuroimaging studies actually explain brain responses during natural speech comprehension.

Here we investigated speech processing at several different levels of representation—spectral, articulatory, and semantic—simultaneously using natural narrative speech. We conducted a functional MRI (fMRI) experiment in which subjects passively listened to natural stories and then used voxelwise modeling (VM; Nishimoto et al., 2011; Santoro et al., 2014; Huth et al., 2016) to determine which specific speech-related features are represented in each voxel. The stimulus was first nonlinearly transformed into three feature spaces—spectral, articulatory, and semantic—that span much of the sound-to-meaning pathway. Then, ridge regression was used to determine which specific features are represented in each voxel, in each individual subject. To examine the relationships between these representations, we used a variance-partitioning analysis. For each voxel, this analysis showed how much variance could be uniquely explained by each feature space versus that explained by multiple feature spaces (Lescroart et al., 2015). To the extent that different stages of speech comprehension involve these feature spaces, our results reveal how the computational steps in speech processing are represented across the cerebral cortex.

Materials and Methods

Participants

Functional data were collected from six male subjects (S1, 26 years of age; S2, 31 years of age; S3, 26 years of age; S4, 32 years of age; S7, 30 years of age) and two female subjects (S5, 31 years of age; S6, 25 years of age). Two of the subjects were authors of this article (S1, A.G.H.; S5, W.A.d.H.). All subjects were healthy and had no reported hearing problems. The use of human subjects in this study was approved by the University of California, Berkeley, Committee for Protection of Human Subjects.

Stimuli and feature spaces

The natural speech stimuli consisted of monologues taken from The Moth Radio Hour, produced by Public Radio International. In each story from The Moth Radio Hour, a single male or female speaker tells an autobiographical story in front of a live audience with no written notes or cues. The speakers are chosen for their story-telling abilities, and their stories are engaging, funny, and often emotional. For our experiments, the stimuli were split into separate model estimation and model validation sets. The model estimation dataset consisted of 10 stories that were 10 to 15 min in length played once each. The length of each scan was tailored to the story and also included 10 s of silence both before and after the story. Each subject heard the same 10 stories, of which 5 were told by male speakers and 5 by female speakers. The model validation dataset consisted of a single 10 min story told by a female speaker that was played twice for each subject to estimate response reliability.

Auditory stimuli were played over Sensimetrics S14 in-ear piezoelectric headphones. These headphones provide both high audio fidelity and some attenuation of scanner noise. A Berhinger Ultra-Curve Pro Parametric Equalizer was used to flatten the frequency response of the headphones. The sampling rate of the stimuli in their digitized form was 44.1 kHz, and the sounds were not filtered before presentation. Thus, the potential frequency bandwidth of the sound stimuli was limited by the frequency response of the headphones from 100 Hz to 10 kHz. The sounds were presented at comfortable hearing levels.

To study cortical speech processing these stories were represented using three distinct feature spaces chosen to capture different levels of processing: spectral-temporal power, articulatory features, and semantic features (Fig. 1). Time-varying spectral power was used to determine the areas of cortex that were sensitive to the spectral content of the sound as would be expected from primary auditory cortex but also other secondary auditory areas found on the temporal lobe (Hullett et al., 2016) as well as other cortical areas that have not been examined in previous studies. The articulatory features were used because they could represent both phonemes (each phoneme corresponds to a unique combination of articulations) and vocal gestures. Phonemic representation is thought to appear in intermediate auditory processing areas in the temporal lobe (DeWitt and Rauschecker, 2012; Mesgarani et al., 2014), while representations based on vocal gestures are thought to appear in premotor areas of the frontal lobe and in Broca's areas (Bouchard et al., 2013; Flinker et al., 2015). Finally, the semantic feature space is used to investigate areas that are sensitive to words and higher abstractions in the speech-processing stream (Huth et al., 2016). Stimuli that had been transformed into these feature spaces were then used to linearly predict the blood oxygenation level-dependent (BOLD) responses over the entire cortical sheet. In the equation used to describe our modeling of these cortical activities (described below), the sound stimulus is written as $s(t)$ and the feature space representation as $\vec{\Phi}(s(t)) = \vec{\Phi}(t)$, where $\vec{\Phi}(t)$ is a vector of features needed for a particular representation. At each time point [here discretized at the repetition time (TR) used for the fMRI data acquisition: TR = 2 s], the size of the vector is constant and corresponds to the number of parameters used in that representation. The three feature spaces used here had 80 spectral values, 22 articulatory values, and 985 semantic values. To construct the semantic and articulatory feature vectors, it was also necessary to determine the timing of each specific word and phoneme in the story. For this purpose, all of the stories were first transcribed manually. The transcriptions were then aligned with the sound and words were coded into phonemes using the Penn Phonetics Lab Forced Aligner software package (http://fave.ling.upenn.edu/). In this procedure, the beginning and end of each word and phoneme were estimated with millisecond accuracy. These temporal alignments were further verified and corrected by hand using the Praat (www.praat.org) visualization software.

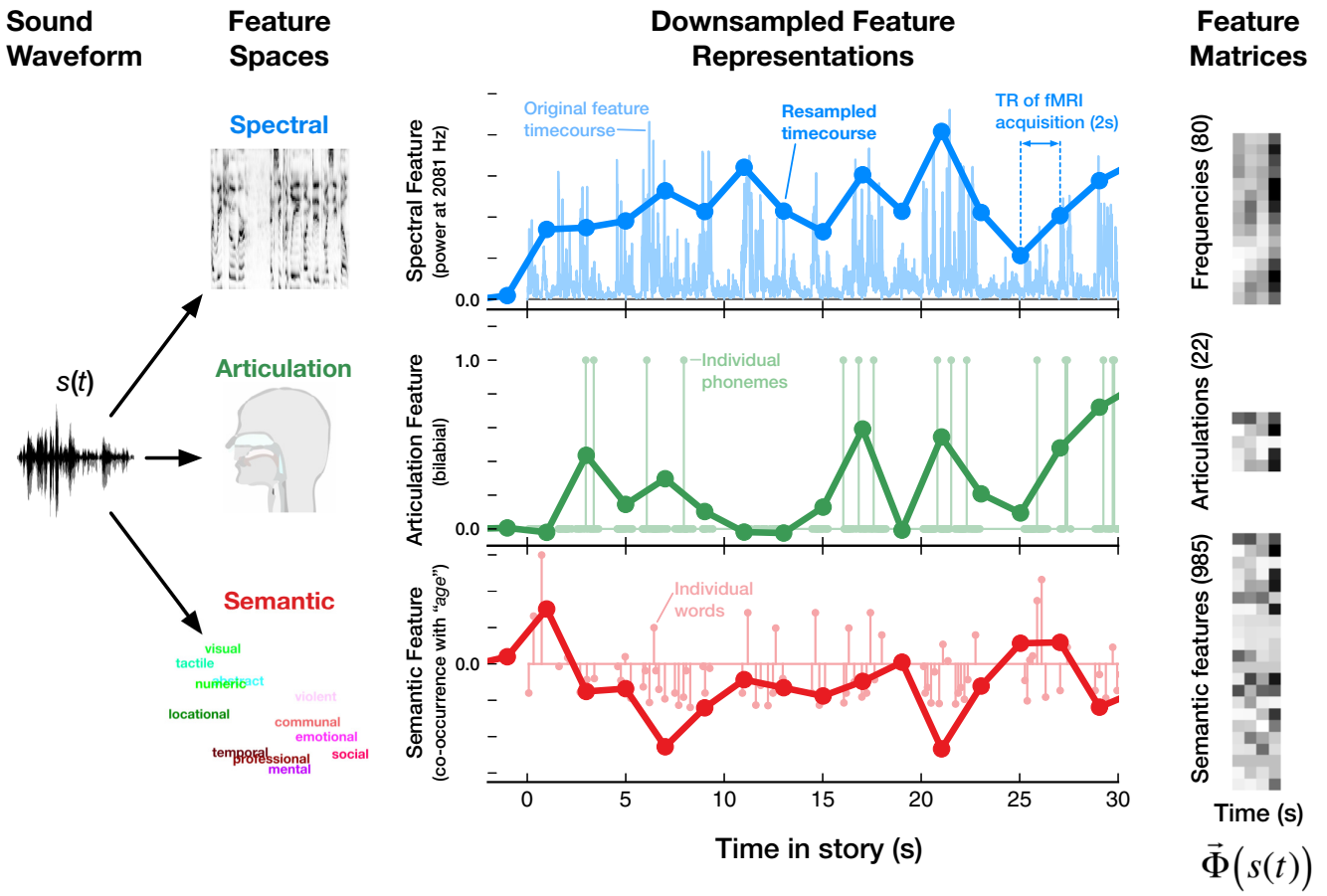

Figure 1. Feature spaces. Three feature spaces were used to predict the BOLD response of each voxel in each subject's brain: a spectral feature space, an articulatory feature space, and a semantic feature space. Each is realized by transforming the sound pressure waveform $s(t)$ into a vector of values at time t', the feature space corresponding to each model, $\vec{\Phi}(s(t)) = \vec{\Phi}(t')$. In this notation, the sound is sampled at 44,100 Hz and indexed with t, while the features are sampled at the TR (0.5 Hz) and are indexed with t'. Thus, this stimulus representation includes a transformation into features followed by low-pass filtering and resampling. The spectral features (blue) are the amplitudes of the 80 channels of a cochleogram, the articulatory features (green) are a 22-dimensional binary vector indicating the presence or absence of 22 articulatory and the semantic features are a 985-dimensional vector representing the statistical co-occurrences of each word in the story to 985 common words in the English language. The line plots in the figure show the time series for a single channel in the cochleogram or dimension in the articulatory and semantic feature vector before (light) and after (bold) the low-pass filtering and resampling step.

Prior single-unit studies (Gill et al., 2006) and fMRI studies (Santoro et al., 2014) have shown the importance of choosing the correct representation of sounds to predict responses in auditory cortex. Single-unit data in both mammals (Depireux et al., 2001; Miller et al., 2002) and birds (Sen et al., 2001) have clearly shown that neural responses in primary auditory cortex are well modeled by a modulation filter bank. The current working model holds that sounds are first decomposed into frequency bands by the auditory periphery, yielding a cochleogram, and then spectrotemporal receptive fields are applied to this time–frequency representation at the level of both the inferior colliculus (Escabi and Schreiner, 2002) and the auditory cortex to extract frequency-dependent spectral–temporal modulations. This cortical modulation filter bank is useful for extracting features that are important for percepts (Woolley et al., 2009) as well as to separate relevant signals from noise (Chi et al., 1999; Mesgarani et al., 2006; Moore et al., 2013). In a previous fMRI study, Santoro et al. (2014) showed that representing acoustic features using a full modulation filter bank that combines frequency information and joint spectrotemporal modulation yielded the highest accuracy for predicting responses to different natural sounds.

For this study, we investigated the predictive power of the following three different acoustic feature spaces: (1) the spectral power density expressed in logarithmic power units (in decibels); (2) a cochleogram that models the logarithmic filtering of the mammalian cochlea and the compression and adaption produced by the inner ear; and (3) a modulation power spectrum (MPS) that captures the power of a spectral–temporal modulation filter bank averaged across spectral frequencies (Singh and Theunissen, 2003). The power spectrum was obtained by estimating the power for 2 s segments of the sound signal (matching the rate of the fMRI acquisition) using the classic Welch method for spectral estimation density (Welch, 1967) with a Gaussian-shaped window that had an SD parameter of 5 ms (corresponding to a frequency resolution of 32 Hz), a length of 30 ms, and with successive windows spaced 1 ms apart. The power was expressed in dB units with a 50 dB ceiling threshold (to prevent large negative values). The final power spectrum consisted of a 449-dimensional vector that reflected the power of the signal between 0 Hz and ~15 kHz, in 33.5 Hz bands. The cochleogram model was estimated using a modified version of the Lyon (1982) Passive Ear model implemented by Slaney (1998) and modified by Gill et al. (2006; https://github.com/theunissenlab/tlab). This cochleogram uses approximately logarithmically spaced filters (more linear at low frequencies and log at higher frequencies) with a bandwidth given by the following:

$$\mathrm{BW} = \frac{\sqrt{\mathrm{cf}^2 + \mathrm{ebf}^2}}{Q},$$

where cf is the characteristic frequency, ebf is the earBreakFreq parameter of the model set at 1000 Hz, and Q is the quality factor (i.e., for log filters defined as the ratio of center frequency to bandwidth) set at 8. The output of this filter bank is rectified and compressed. In addition, the model includes adaptive gain control (for more details, see Lyon, 1982; Gill et al., 2006). The output of this biologically realistic cochlear filter bank consisted of 80 waveforms between 264 and 7630 Hz, spaced at 25%

of the bandwidth. Finally, the MPS features were generated from the spectrograms of 1 s segments of the story. The time–frequency scale of the spectrogram was set by the width of the Gaussian window used in the short time Fourier Transform: 32 Hz or 4.97 ms. The MPS was then obtained by calculating the 2D FFT of the log amplitude of the spectrogram with a ceiling threshold of 80 dB. We then limited the temporal and spectral modulations of the MPS that are the most relevant for the processing of speech (Elliott and Theunissen, 2009): −17 to 17 Hz for temporal amplitude modulations (dtm = 0.5 Hz) and 0 to 2.1 cycles/kHz for spectral modulations (dsm = 0.065 cycles/kHz). Thus, the modulation power spectrum feature space yielded 2272 (71 × 32) features at a 1 Hz sampling rate. Both the cochlear and modulation power spectrum features were low-pass filtered at a cutoff frequency of 0.25 Hz using a Lanczos antialiasing filter and downsampled to the fMRI acquisition rate of 0.5 Hz.

In preliminary testing, these three auditory feature spaces yielded similar predictions, with significant voxels found in similar brain areas but the model using the cochlear features systematically outperformed the models using the power spectrum features (in seven of seven subjects) and the modulation power spectrum features (in five of seven subjects). We also found that combining modulation power spectrum features with frequency features using a full modulation filter bank also yielded predictions similar to those the cochleogram (data not shown). This result is somewhat different from the findings in the study by Santoro et al. (2014), where the full modulation filter bank yielded the best prediction. However, it should be noted that stimuli are more restricted in our study (speech only vs multiple natural sound classes) and that the biggest difference between representations in the Santoro et al. (2014) study was found for high-resolution data obtained in a 7 T MRI scanner. Thus, for the goals of this analysis, the cochleogram was chosen as the representation for the low-level acoustic feature space. All further results presented here for the spectral feature space are based on the cochleogram representation.

For the articulatory feature space, phonemes obtained from the alignment procedure were represented by their associated articulations (Levelt, 1993). We created a 22-dimensional vector with a unique pattern of articulations per phoneme, measuring manner, place, and phonation for consonants and phonation, height, and front to back position for vowels (Table 1). These vectors (one per phoneme) were then low-pass filtered at a cutoff frequency of 0.25 Hz using a Lanczos antialiasing filter and downsampled to the fMRI acquisition rate of 0.5 Hz.

To represent the meaning of words, a 985-dimensional vector was constructed based on word co-occurrence statistics in a large corpus of text (Deerwester et al., 1990; Lund and Burgess, 1996; Mitchell et al., 2008; Turney and Pantel, 2010; Wehbe et al., 2014). First a 10,470 word lexicon was selected as the union of the set of all words appearing in the stimulus stories and the 10,000 most common words in the training corpus. Then 985 basis words were selected from Wikipedia's "List of 1000 Basic Words" (contrary to the title, this list contains only 985 unique words). This basis set was selected because it consists of common words that span a very broad range of topics. The training corpus used to construct this feature space includes the transcripts of 13 Moth stories (including the 10 used as stimuli), 604 popular books, 2,405,569 Wikipedia pages, and 36,333,459 user comments scraped from reddit.com. In total the 10,470 words in our lexicon appeared 1,548,774,960 times in this corpus. Next, a word co-occurrence matrix, C, was created with 985 rows and 10,470 columns. Iterating through the training corpus, 1 was added to C_{ij} each time word j appeared within 15 words of basis word i. The window size of 15 was selected to be large enough to suppress syntactic effects (i.e., word order) but no larger. Once the word co-occurrence matrix was complete, the counts were log-transformed, producing a new matrix, E, where $E_{ij} = \log(1 + C_{ij})$. Then each row of E was z scored to correct for differences in frequency among the 985 basis words, and each column of E was z scored to correct for frequency among the 10,470 words in our lexicon. Each column of E is now a 985-dimensional semantic vector representing one word in the lexicon. This representation tends to be semantically smooth, such that words with similar meanings (e.g., "dog" and "cat") have similar vectors, but words with very different meanings (such as "dog" and "book") have very different vectors. The semantic feature space for this experiment was then constructed from the stories: for each word–time pair (w, t) in each story, we selected the corresponding column of E, creating semantic feature vectors sampled at the word rate (~4 Hz). These vectors were then low-pass filtered and downsampled to the fMRI acquisition rate using a Lanczos filter.

Table 1. Phoneme to articulation conversion chart

Phoneme	Articulatory Features			
B	Bilabial	Plosive	Voiced	
CH	Postalveolar	Affricate	Unvoiced	
D	Alveolar	Plosive	Voiced	
DH	Dental	Fricative	Voiced	
F	Labiodental	Fricative	Unvoiced	
G	Velar	Plosive	Voiced	
HH	Glottal	Fricative	Unvoiced	
JH	Postalveolar	Affricate	Voiced	
K	Velar	Plosive	Unvoiced	
L	Alveolar	Lateral	Voiced	
M	Bilabial	Nasal	Voiced	
N	Alveolar	Nasal	Voiced	
NG	Velar	Nasal	Voiced	
P	Bilabial	Plosive	Unvoiced	
R	Alveolar	Approximant	Voiced	
S	Alveolar	Fricative	Unvoiced	
SH	Postalveolar	Fricative	Unvoiced	
T	Alveolar	Plosive	Unvoiced	
TH	Dental	Fricative	Unvoiced	
V	Labiodental	Fricative	Voiced	
W	Velar	Approximant	Voiced	
Y	Palatal	Approximant	Voiced	
Z	Alveolar	Fricative	Voiced	
ZH	Postalveolar	Fricative	Voiced	
AA	Low	Back		
AE	Low	Front		
AH	Mid	Central		
AO	Mid	Back		
AW	Low	Central	Mid	Back
AY	Low	Central	Mid	Front
EH	Mid	Front		
ER	Mid	Central		
EY	Mid	Front		
IH	Mid	Front		
IY	High	Front		
OW	Mid	Back		
OY	Mid	Back	High	Front
UH	High	Back		
UW	High	Back		

It is important to note that of the spectral, articulatory, and semantic models, only the spectral model was computed directly from the sound waveform, while the articulatory and semantic labeling were performed here by humans. We hope that improvements in the field of speech recognition will soon render this distinction obsolete.

Experimental design and statistical analysis

MRI data collection and preprocessing. Structural MRI data and BOLD fMRI responses from each subject were obtained while they listened to ~2 h and 20 min of natural stories. For five of the subjects, these data were collected during two separate scanning sessions that lasted no more than 2 h each. For two of the subjects (S1, author A.G.H., and S5, author W.A.d.H.) the validation data (two repetitions of a single story) were collected in a third separate session. MRI data were collected on a 3 T Siemens TIM Trio scanner at the University of California, Berkeley, Brain Imaging Center, using a 32-channel Siemens volume coil. Functional scans were collected using a gradient echo EPI sequence with TR = 2.0045 s, echo time = 31 ms, flip angle = 70°, voxel size = 2.24 × 2.24 × 4.1 mm, matrix size = 100 × 100, and field of view = 224 × 224 mm. Thirty-two axial slices were prescribed to cover the entire cortex. A

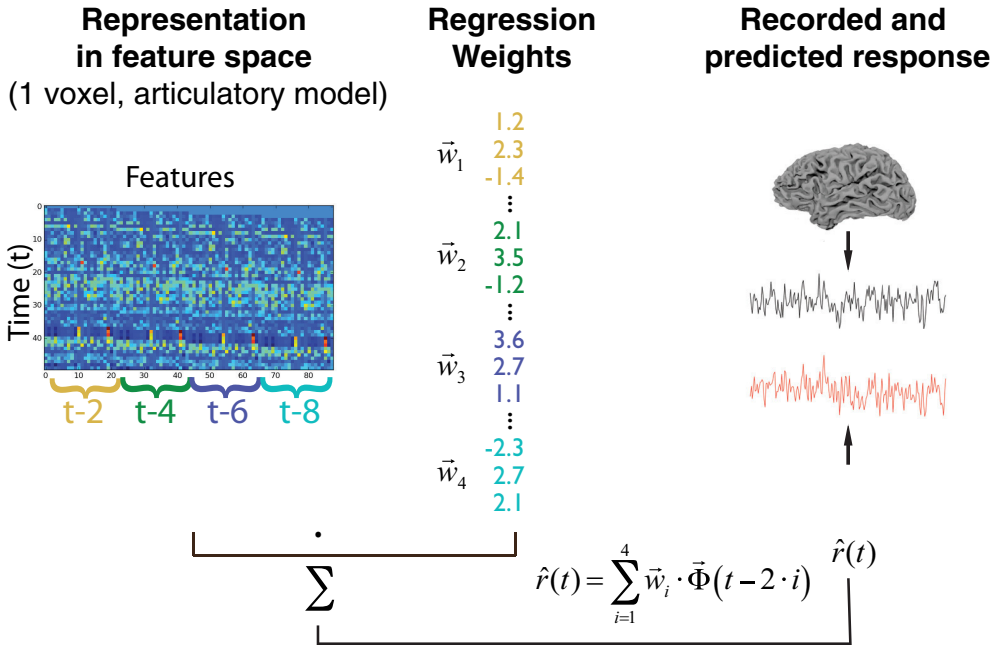

Figure 2. Linear regression. The stimulus is represented in a feature space (or combination of feature spaces) and is then used in linear regression to obtain a prediction $\hat{r}(t)$ (red curve) of the actual bold response, $r(t)$ (black curve) for each voxel in the brain. This linear regression is a linear filter with four point delays (t-2, t-4, t-6, and t-8): $\hat{r}(t) = \sum_{i=1}^{4} \vec{w}_i \cdot \vec{\Phi}(t - 2i)$. The diagram illustrates this operation: a row in the feature matrix shown on the left corresponds to a time t of a response and shows in a color code the features (here the articulation) at times t-2, t-4, t-6, and t-8 unfolded as a single vector (a single vector). The dot product of that vector with the regression weights (the $\vec{w}$) yields the predicted response at time t. These parameters were obtained using ridge regression, and model performance was assessed using cross-validation.

custom-modified bipolar water excitation radiofrequency pulse was used to avoid signals from fat tissue. Anatomical data were collected using a T1-weighted MP-RAGE (Brant-Zawadzki et al., 1992) sequence on the same 3 T scanner.

Each functional run was motion corrected using the fMRIB Linear Image Registration Tool (FLIRT) from FSL 4.2 (Jenkinson and Smith, 2001). All volumes in the run were then averaged to obtain a high-quality template volume. FLIRT was also used to automatically align the template volume for each run to the overall template, which was chosen to be the template for the first functional run for each subject. These automatic alignments were manually checked and adjusted for accuracy. The cross-run transformation matrix was then concatenated to the motion-correction transformation matrices obtained using MCFLIRT, and the concatenated transformation was used to resample the original data directly into the overall template space. Low-frequency voxel response drift was identified using a second-order Savitsky–Golay filter with a 120 s window, and this was subtracted from the signal. After removing this time-varying mean, the response was scaled to have unit variance (i.e., z scored).

Structural imaging and fMRI were combined to generate functional anatomical maps that included localizers for known regions of interests (ROIs). These maps were displayed either as 3D structures or as flatmaps using custom software (pycortex). For this purpose, cortical surface meshes were first generated from the T1-weighted anatomical scans using Freesurfer software (Dale et al., 1999). Five relaxation cuts were made into the surface of each hemisphere, and the surface crossing the corpus callosum was removed. The calcarine sulcus cut was made at the horizontal meridian in V1 using retinotopic mapping data as a guide. Known auditory and motor ROIs were then localized separately in each subject using standard techniques. To determine whether a voxel was responsive to auditory or motor stimuli, repeatability of the voxel response was calculated as an F statistic given by the ratio of the total variance in the responses over the residual variance. The residual variance was obtained by comparing responses in individual trials to the mean responses obtained from multiple repeats of the stimulus played back or of the motor action. AC localizer data were collected in one 10 min scan. The subject listened to 10 repeats of a 1 min auditory stimulus, which consisted of 20 s segments of music, speech, and natural sound. Motor localizer data were collected during one 10 min scan. The subject was cued to perform six different motor tasks in a random order in 20 s blocks (10 blocks per motor task for a total to 60 blocks). For the hand, mouth, foot, speech, and rest blocks, the stimulus was simply a word located at the center of the screen (e.g., "Hand"). For the Hand cue, the subject was instructed to make small finger-drumming movements with both hands for as long as the cue remained on the screen. Similarly, for the "Foot" cue the subject was instructed to make small toe movements for the duration of the cue. For the "Mouth" cue, the subject was instructed to make small mouth movements approximating the nonsense syllables *balabalabala* for the duration of the cue—this requires movement of the lips, tongue, and jaw. For the "Speak" cue, the subject was instructed to continuously subvocalize self-generated sentences for the duration of the cue. For the saccade condition, the written cue was replaced with a fixed pattern of 12 saccade targets and the subject was instructed to make frequent saccades between the targets. After preprocessing, a linear model was used to find the change in BOLD response of each voxel in each condition relative to the mean BOLD response. Weight maps for the foot, hand, and mouth responses were used to define primary motor area (M1) and somatosensory area (S1) for the feet (M1F, S1F), hands (M1H, S1H), and mouth (M1M, S1M); supplementary motor areas for the feet and hands; secondary somatosensory area for the feet (S2F) and, in some subjects, the hands (S2H); and, in some subjects, the ventral premotor hand area. The weight map for saccade responses was used to define the frontal eye field, frontal operculum eye movement area, intraparietal sulcus visual areas, and, in some subjects, the supplementary eye field. The weight map for speech production responses was used to define Broca's area and the superior ventral premotor speech area (sPMv).

Linear model fitting. The relationship between the speech stimulus represented in various feature spaces and the BOLD response was fitted with a linear filter estimated for every single voxel (i.e., VM) (Fig. 2). This filter is equivalent to a finite impulse response (FIR) model (Nishimoto et al., 2011; Huth et al., 2012) and to spatiotemporal receptive fields where the spatial dimensions correspond to the vector dimensions of particular feature spaces. The filters modeled here were estimated at the following

four delays: t-2 s, t-4 s, t-6 s, t-8 s. Therefore, the equation for the FIR can be written as follows:

$$\hat{r}(t) = \sum_{i=1}^{4} \vec{w}_i \cdot \vec{\Phi}(t - 2i).$$

Because auditory cortex has shorter hemodynamic delays than does visual cortex (Belin et al., 1999), this model incorporates a 2 s time delay that was not used in earlier vision publications from our group (Nishimoto et al., 2011; Huth et al., 2012). Models were fitted for each feature space independently, for all pairwise combinations of feature spaces, and for the combination of all three feature spaces (for a total of seven linearized models for each voxel).

The linear model weights $\vec{h}_i$ were estimated with regularized regression methods and cross-validation to avoid overfitting. Before fitting, both the BOLD response and the stimulus values in the feature space were z scored. Because each story had significantly different spectral, articulatory, and semantic content, and because the BOLD signals adapted to these average levels, a different mean and variance was used in the z-scoring operation for each story.

Regularization was performed using ridge regression. A separate ridge regularization parameter was estimated for each voxel, in each subject, and for each model (i.e., each combination of feature spaces). Ridge parameter estimation was performed by repeating a cross-validation regression procedure 50 times, in each subject, for each model. On each cross-validation iteration, 800 time points (consisting of 20 random blocks of 40 consecutive time points each) were selected at random and removed from the training dataset. Then the model parameters were estimated on the remaining 2937 time points, and each of 20 possible regularization hyperparameters were log-spaced between 10 and 10,000. These weights were used to predict responses for the 800 reserved time points, and R^2 was computed from these data (R^2 gives the fraction of variance that the predictions explain in the responses). After the 50 cross-validation iterations were complete, a regularization–performance curve was obtained by averaging the sample R^2 values across the 50 iterations. This curve was used to select the best regularization hyperparameter for the current model in each voxel and in each subject. Finally, the selected hyperparameter and all 3737 training time points were used to estimate the final model parameters.

For voxelwise models that combined multiple feature spaces, two different regularization approaches were investigated. First, the subspaces obtained from performing ridge regression on the individual features were combined, and a single regression analysis was performed in that joint subspace (see Joint ridge regression section below). Second, the features were concatenated without performing any dimensionality reduction and a new optimal ridge parameter was estimated for this joint space. Although results were similar in both cases, the second approach tended to yield a smaller error in the variance partitioning scheme. Only the results obtained with a single ridge parameter are shown in the Results section. However, we also describe the method for the joint ridge regression approach in the Materials and Methods section, because it is a principled approach for comparing nested models, and it can be used to verify the validity of other approaches for combining feature spaces in the context of regularized regression.

All model fitting and analysis was performed using custom software written in Python, which made heavy use of the NumPy (Oliphant, 2006) and SciPy (Jones et al., 2007) libraries.

Signal detection and model validation. Before voxelwise modeling was performed, the validation dataset was used to determine which voxels were significantly active in response to the stories (story-responsive voxels). The validation set consisted of BOLD responses to two repeats of the same story, as described above (Fig. 2). The correlation between these paired BOLD responses was calculated, and then the significance of this correlation was computed using an exact test. This test gives the probability of finding the observed correlation assuming that the two response vectors are bivariate Gaussians distributed with zero covariance. These p values were then corrected for multiple comparisons across voxels within each subject using the false discovery rate (FDR; Benjamini and Hochberg, 1995). The results of this analysis were used to show which voxels responded to speech but were not used to constrain or bias the voxelwise modeling analyses in any way.

After voxelwise model estimation was performed using the model estimation dataset, we estimated which voxels were significantly predicted by each model. First, the correlation between the actual BOLD response in the validation dataset (averaged across the two repetitions) and the model predictions were calculated, and then the significance of the observed correlation was computed as above. While the correlation between predicted response and actual mean response is an appropriate metric for assessing significance, it is biased downward due to noise in the validation data (Sahani and Linden, 2003; Hsu et al., 2004; David and Gallant, 2005). This is because the actual mean response is calculated using a finite number of repetitions (in this case, 2), and so it contains both signal and residual noise. This noise level is likely to vary across voxels due to vascularization and magnetic field inhomogeneity. The noise in the validation dataset was accounted using the method developed in the study by Hsu et al. (2004). In this method, the raw correlation is divided by the expected maximum possible model correlation (called the noise ceiling) for each voxel.

Joint ridge regression. The following section describes how to perform a joint ridge regression operation that preserves the individual shrinkages obtained in the single-feature space ridge regressions. This preservation of individual shrinkages in the joint model is important when comparing nested models.

Ridge regression as well as other forms of regularization shrink the parameter space of input features to prevent overfitting. These dimensionality reduction operations rely on feature spaces that have uniform physical dimensions and statistical properties in a Euclidian space. The joint models fitted here combined feature spaces with different units, different numbers of parameters, p, and different degrees of correlation across those parameters. When regularized linear models are fitted separately for each stimulus feature spaces, the solution that best generalizes to a novel dataset will be obtained with a different optimal value for the ridge parameter λ: stimulus feature spaces that have more dimensions and/or are less predictive of the response will require more shrinkage to prevent overfitting. When a model is based on a union of two feature spaces, such as one with a large number of parameters (e.g., semantic) and one with a low number of parameters (e.g., articulation), using a single large shrinkage value to prevent overfitting with this large combined feature space might significantly reduce the contribution to the prediction from the features of the smaller feature space. More importantly, using the same shrinkage in the joint model (i.e., the same projection to the same subspaces) is required to estimate the total variance (union in set theory), and from there both the shared (intersection in set theory) and unique [relative complement (RC) in set theory] variances explained (the variance partitioning in set theoretical terms is further developed below). The total variance explained of two or more regularized (shrunken) feature spaces is the variance explained by the union of these shrunken feature subspaces.

The voxelwise models predict responses, $\vec{r}$, from stimulus parameters $\mathbf{S}$. Here $\vec{r}$ is a column vector ($n \times 1$) corresponding to the BOLD response of a single voxel as a function of n discrete sampling points in time. $\mathbf{S}$ is a ($n \times p$) matrix where each row corresponds to the stimulus at the same n time points as $\vec{r}$, and the columns correspond to the values describing the stimulus in its feature space for a number of time slices: the number of columns (p) is equal to the dimension of the feature space (k) times the number of delays used in the FIR model (here $p = 4 \times k$ since four time slices are used). The maximum likelihood solution for the multiple linear regression is given by the following normal equation:

$$\vec{w} = \frac{\langle Sr \rangle}{\langle SS \rangle} = [\mathbf{S}^T\mathbf{S}]^{-1}[\mathbf{S}^T\vec{r}],$$

where $\vec{h}$ is the column vector of regression weights ($p \times 1$) also called the model parameters or the filter. The angle brackets ($\langle$ and $\rangle$) stand for averaging the cross-products across time (across rows). Note, more specifically, that the correct unbiased estimate of the stimulus–response cross-covariance (the numerator) and the stimulus auto-covariance (the denominator) are as follows:

$$\langle SS \rangle = \frac{\mathbf{S}^T\mathbf{S}}{n-1} \text{ and } \langle Sr \rangle = \frac{\mathbf{S}^T\vec{r}}{n-1}.$$

The prediction can then be obtained by:

$$\hat{\vec{r}} = \mathbf{S}\vec{w}.$$

The inverse of the symmetric and positive definite stimulus autocovariance matrix can easily be obtained from its eigenvalue decomposition or equivalently from the singular value decomposition (SVD) of **S**. The SVD of **S** can be written as follows:

$$\mathbf{S} = \mathbf{VWU}^T,$$

where **V** is a $(p \times p)$ matrix of orthonormal input vectors in columns (or left singular vectors), **W** is a diagonal $(p \times n)$ matrix of positive single values, and **U** is a $(n \times n)$ matrix of orthonormal output vectors in columns (or right singular vectors).

The eigenvalue decomposition of $\mathbf{S}^T\mathbf{S}$ is then given by the following:

$$\mathbf{S}^T\mathbf{S} = \mathbf{VW}^2\mathbf{V}^T,$$

where W^2 is the $(p \times p)$ diagonal matrix of eigenvalues.

To prevent overfitting (when p is large relative to n), a regularized solution for $\vec{w}$ can be obtained by ridge regression. The ridge regression is the maximum a posteriori solution (MAP) with a Gaussian prior on $\vec{w}$ with zero mean and covariance matrix given by $\mathbf{I}\lambda$.

Under these assumptions, the MAP solution is as follows:

$$\vec{w} = \mathbf{V}[\mathbf{W}^2 + \mathbf{I}\lambda]^{-1}\mathbf{V}^T\mathbf{S}^T\vec{r}.$$

If the normal equation can be interpreted as solving for $\vec{w}$ in the whitened-stimulus space (uncorrelated by the rotation given by **V** and normalized **W**), then the ridge regression decorrelates the stimulus space and provides a weighted normalization where the uncorrelated stimulus parameters with small variance (or small eigenvalues) are shrunken more than those with higher variance (or higher eigenvalues). The level of this relative shrinkage is controlled by the hyper-parameter λ, and its optimal value is found by cross-validation (see Linear model fitting above).

In the following equations, $\mathbf{S}_1$ is the $(n \times p_1)$ matrix of features for the first stimulus feature space, and $\mathbf{S}_2$ is the $(n \times p_2)$ matrix of features for the second stimulus feature space. $\mathbf{S}_{12}$ is the $(n \times (p_1 + p_2))$ matrix that combines the features from both spaces simply by column concatenation. In the joint ridge approach, the regression is performed in the rotated and scaled basis obtained for each of the models. The stimulus space in that new basis set is noted with a prime in the equations below. The decorrelation step is then performed in the joint stimuli but without performing any additional normalization

$$\mathbf{S}_1^{'T} = ([\mathbf{W}_1^2 + \mathbf{I}\lambda_1]^{-1/2}\mathbf{V}_1^T\mathbf{S}_1^T) \cdot \sqrt{n-1}$$

$$\mathbf{S}_2^{'T} = ([\mathbf{W}_2^2 + \mathbf{I}\lambda_2]^{-1/2}\mathbf{V}_2^T\mathbf{S}_2^T) \cdot \sqrt{n-1}.$$

After whitening the stimuli, a correlation coefficient matrix is created from the covariance matrix to decorrelate the stimuli without further normalization. The stimulus covariance matrix in this new stimulus space (denoted with the prime) can be obtained with $\mathbf{S}_{12}^{'T}\mathbf{S}_{12}^{'}$ divided by $n-1$, or the following:

$$\mathbf{S}_{12}^{'T}\mathbf{S}_{12}^{'} = (n-1)\begin{pmatrix} \sigma_{1,1}^2 & 0 & 0 & c_{1,1;2,1} & \cdots & c_{1,1;2,p_2} \\ 0 & \ddots & 0 & \vdots & \ddots & \vdots \\ 0 & 0 & \sigma_{1,p_1}^2 & c_{1,p_1;2,1} & \cdots & c_{1,p_1;2,p\_2} \\ c_{1,1;2,1} & \cdots & c_{1,p_1;2,1} & \sigma_{2,1}^2 & 0 & 0 \\ \vdots & \ddots & \vdots & 0 & \ddots & 0 \\ c_{1,1;2,p_2} & \cdots & c_{1,p_1;2,p_2} & 0 & 0 & \sigma_{1,p_2}^2 \end{pmatrix}$$

where σ^2 is variance and c is the covariance between individual parameters in each of the two feature spaces. The first index is for the feature space corresponding to model 1 or 2, and the second index runs over the parameters in that feature space. As one can notice from the form of this covariance matrix, the stimulus parameters are uncorrelated within each subset, but because of the relative shrinkage performed by the ridge they are not perfectly white. Therefore, the variance in the diagonals is not exactly equal to 1 but is slightly smaller and with decreasing values along each block diagonal. If at this stage the weights of the linear regression were to be obtained using a normal equation, the shrinkage performed in the ridge solution would be inverted. To prevent this unwanted normalization, the covariance matrix can be replaced with the correlation matrix. Dividing by the correlation matrix will decorrelate the stimulus features across the two component models while preserving the exact shrinkage from the separate ridge regressions. The correlation matrix obtained from the covariance matrix is given by the following:

$$\mathrm{Corr}(\mathbf{S}_{12}^{'}) = \begin{pmatrix} 1 & 0 & 0 & \frac{c_{1,1;2,1}}{\sigma_{1,1}\sigma_{2,1}} & \cdots & \frac{c_{1,1;2,p_2}}{\sigma_{1,p_1}\sigma_{2,1}} \\ 0 & \ddots & 0 & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & \frac{c_{1,p_1;2,1}}{\sigma_{1,p_1}\sigma_{2,1}} & \cdots & \frac{c_{1,p_1;2,p_2}}{\sigma_{1,p_1}\sigma_{2,p_2}} \\ \frac{c_{1,1;2,1}}{\sigma_{1,1}\sigma_{2,1}} & \cdots & \frac{c_{1,p_1;2,1}}{\sigma_{1,p_1}\sigma_{2,1}} & 1 & 0 & 0 \\ \vdots & \ddots & \vdots & 0 & \ddots & 0 \\ \frac{c_{1,1;2,p_2}}{\sigma_{1,1\sigma_{2,p_1}}} & \cdots & \frac{c_{1,p_1;2,p_2}}{\sigma_{1,p_1}\sigma_{2,p_2}} & 0 & 0 & 1 \end{pmatrix}$$

The combined ridge filter is then calculated as follows:

$$\vec{w}_{12}^{'} = \frac{\mathrm{Corr}(\mathbf{S}_{12}^{'})^{-1}\mathbf{S}_{12}^{'T}\vec{r}}{n-1}.$$

and is used to obtain predictions from the combined model with the following equation:

$$\hat{\vec{r}} = \mathbf{S}_{12}^{'}\vec{w}_{12}^{'},$$

For clarity, the derivation used here involved joint ridge regression on two models, but it can be extended to the joint ridge regression so that any number of feature spaces can be combined.

Partitioning of variance. To quantify the unique contribution of different stimulus features to the BOLD responses, we estimated the variance explained (R^2) uniquely by each individual feature space and the variance explained by the intersections of various combinations of these feature spaces (Lescroart et al., 2015). For this purpose, the results obtained from fitting models from individual feature spaces and combinations of two and three feature spaces were used to estimate R^2 for all of the nested models. Set theory was then used to calculate the common (as a set intersection) and unique (as a set difference) variances explained. (See Fig. 5 for a graphical representation of this process.) To be succinct, in the remainder of this section, the variance explained by the three feature spaces will be written as sets A–C. First, R^2 values for the following nested models were directly obtained using the following linear model fitting and cross-validation procedures described above:

$$A,\ B,\ C,\ A \cup B,\ A \cup C,\ B \cup C \text{ and } A \cup B \cup C.$$

The shared variances explained by the intersections of two sets was then obtained from the following:

$$A \cap B = A + B - A \cup B$$

$$A \cap C = A + C - A \cup C$$

$$B \cap C = B + C - B \cup C.$$

Similarly, the variance explained by the intersection of all three sets was obtained from the following:

$$A \cap B \cap C = A \cup B \cup C + A + B + C - A \cup B - A \cup C - B \cup C$$

The variance explained by the intersections of two models that did not include the variance explained by the intersection of all three models was then calculated from the following:

$$(A \cap B) \backslash C = A + B - A \cup B - A \cap B \cap C$$

$$(A \cap C) \backslash B = A + C - A \cup C - A \cap B \cap C$$

$$(B \cap C) \backslash A = B + C - B \cup C - A \cap B \cap C.$$

Finally, the variance solely explained by one model, with no overlap of variance explained by any of the other models, was calculated. This is known as the RC for each pair of models. The relative complement of BC, or BC^{RC}, is the portion of the variance explained exclusively by model A:

$$BC^{RC} = A \backslash (B \cup C) = A - A \cap B - A \cap C + A \cap B \cap C$$

$$AC^{RC} = B \backslash (A \cup C) = B - B \cap A - B \cap C + A \cap B \cap C$$

$$AB^{RC} = C \backslash (A \cup B) = C - C \cap A - C \cap B + A \cap B \cap C.$$

Set notation is used here because of its simplicity and its intuitive graphical representation of the results. However, one can easily rewrite these quantities in terms of R^2 and the sum of errors. For example, if SS_0 is used to represent the total sum of square errors (or the SS of a 0th order model predicting the mean response), then we have the following:

$$A = R_A^2 = \frac{SS_0 - SS_A}{SS_0}, \; B = R_B^2 = \frac{SS_0 - SS_B}{SS_0}, \text{ and}$$

$$C = R_C^2 = \frac{SS_0 - SS_C}{SS_0},$$

and:

$$A \cap B = R_{A \cap B}^2 = \frac{SS_0 - (SS_A + SS_B - SS_{A \cup B})}{SS_0} = \frac{SS_0 - SS_A}{SS_0} + \frac{SS_0 - SS_B}{SS_0} - \frac{SS_0 - SS_{A \cup B}}{SS_0}$$

$$A \cap B = R_{A \cap B}^2 = R_A^2 + R_B^2 - R_{A \cup B}^2$$

$$A \cap B = A + B - A \cup B.$$

Correction of variance partition estimates. Because empirical estimates of the variance explained by single and joint models contain sampling noise, the set theoretical approach detailed above sometimes produced results that were not theoretically possible. These sampling errors occurred both using the joint regression algorithm or when using a single ridge shrinkage parameter for each model. For example, the estimated variance explained by $A \cup B$ in the held-out validation dataset was sometimes smaller than the variance explained by A or B alone, due to overfitting of the larger $A \cup B$ model and sampling error. This happened most often by combining the semantic model, which has a large number of parameters and good predictive power, with the spectral or articulatory model, either of which has a small number of parameters and little additional predictive power in many regions of the brain. This situation produced nonsensical results, such as variance partitions with negative values. To mitigate this problem, a *post hoc* correction was applied to the estimated variance explained by each model in each voxel. This correction moved the estimates to the nearest values that produced no nonsensical results. Mathematically, this involved estimating a bias term for the variance explained by each model in each voxel. We began by assuming that the estimated variance explained by some model (R^2), $\hat{X}$, is a biased estimate of the true variance explained, X^*: $\hat{X} = X^* + b_X$.

Because there were seven models (each feature space alone, each pair, and all three together), this formulation yields seven bias parameters (b_x). Furthermore, because we know that the size of each variance partition must be at least equal to zero, the set theory equations that give the size of each partition can be used to define seven inequality constraints on the bias terms. Assuming that we want to find the smallest set of bias parameters (in an L2 sense) that produce no nonsensical results, this allowed us to set up a constrained function minimization problem, as follows:

$$\min\{\|b\|^2\} \text{ subject to } h_j(\vec{b}) \geq 0 \text{ for j } = 1..7,$$

where h are our seven inequality constraints.

This procedure was applied separately to the estimated values of the variance explained for each voxel. Applying this correction to simulated data verified that this scheme significantly decreases error, variance, and bias in the estimated variance partition sizes.

Auditory cortex axes and centers of mass. The center of mass was calculated along two axes of the auditory area for our seven variance partitions. First, all of the voxels within the auditory cortex (as defined by our localizer) were projected onto the following two different axes: an anterior–posterior axis and a medial–lateral axis. The medial–lateral axis was defined as the geodesic distance along the cortical surface from the crown of the superior temporal gyrus (STG), in millimeters. The anterior–posterior axis was defined as the geodesic anterior–posterior distance from the intersection of Heschl's gyrus and STG. The center of mass of each model and axis was then calculated for each subject, as follows:

$$cm = \frac{\sum_{i=0}^{n} a_i \cdot r_i}{\sum_{i=0}^{n} r_i}$$

where a_i is the location of the ith voxel projected on the chosen axis (in millimeters), and r is the voxel model performance, expressed as partial correlation, or the positive square root of the partial R^2.

Bootstrapping was used to calculate the SEs of these calculated centers of mass. For each model, subject, and axis, 1000 points were sampled (with replacement) along the chosen auditory axis, 1000 times. The center of mass was then calculated for each sample, and the SEM was computed from these data.

Linear mixed-effects modeling. Linear mixed-effects models (lme) were used to compare average responses in the left versus right hemispheres for all the cortical voxels as well as for specific ROIs. In these statistical tests, the subject ID was the random effect. The lme tests were run in R using the lme4 library. For *post hoc* tests, p values were corrected for multiple comparisons using the FDR (Benjamini and Hochberg, 1995).

Mapping semantic selectivity within auditory cortex. Variance partitioning shows that some of the response explained by the semantic features in AC cannot be explained away by spectral or articulatory features. However, the semantic model could still be picking up on other features that are correlated with semantics but were not included in the variance partitioning. One possibility is global modulatory effects such as attention or arousal, which one could easily imagine as being correlated with semantic content. If semantic models were capturing global modulatory signals, we would expect to find homogeneous semantic tuning across AC. To test for this possibility, we examined the variability of semantic selectivity within AC. The weights for the semantic model were projected onto the top 10 principal components (PCs) of the semantic models from all seven subjects (for details, see Huth et al., 2016; the PCs used here were taken from the analysis described in that article). This was done to reduce the semantic models to a dimensionality that could be easily examined while preserving as much structure as possible. We then selected the subset of the voxels where the unique contribution of the semantic features was the largest variance partition according to the variance partitioning analysis and computed the variance of the PC projections across that set of voxels for each of the 10 PCs. This analysis was performed for each subject and each hemisphere.

To obtain a baseline, we repeated this analysis for 200 random regions of the same size as AC in each subject and hemisphere. Random regions were formed by randomly selecting a point on the cortex and then selecting nearby vertices (according to geodesic distance) until the new region had the same number of vertices as AC for that subject and hemisphere.

Table 2. Fraction of cortex significantly active (q(FDR) < 0.05) and average response reliability in response to stories, mean of 7 subjects ± SE

	Fraction significant		Avg. reliability	
ROI	Left	Right	Left	Right
Auditory cortex	35.43% (±3.79)	37.19% (±4.25)	0.122 (±0.011)	0.134 (±0.016)
sPMv	42.86% (±6.46)	40.84% (±7.25)	0.134 (±0.019)	0.134 (±0.021)
Broca	42.86% (±6.51)	45.13% (±9.31)	0.138 (±0.018)	0.153 (±0.030)
Other cortex	18.60% (±3.17)	17.96% (±3.14)	0.065 (±0.009)	0.062 (±0.009)

Results

Story-related responses

The goal of this study was to investigate how different features of natural speech are represented across the cortex. Before examining specific features, however, we first aimed to discover which brain regions might be involved in any stage of speech comprehension. This was done by playing one 10 min naturally spoken narrative story twice for each subject and then computing, for each voxel in the cerebral cortex of each subject, the correlation between responses to the first and second presentations. Areas of cortex that are involved in almost any aspect of speech processing should respond reliably across the two presentations, while areas not involved in speech processing should not respond reliably.

We found significantly reliable responses (exact correlation test with $n = 290$, $q(\text{FDR}) < 0.05$) across presentations of the same story in a large fraction of the cortex (Table 2). In the AC, defined here as the region of temporal cortex that responds reliably to a selection of different sounds, and thus includes primary, secondary, and association areas of auditory cortex, we found that 35% of voxels in the left hemisphere and 37% of voxels in the right hemisphere were story responsive. In the speech-related sPMv (Wilson et al., 2004), we found that 43% of voxels in the left hemisphere and 41% of voxels in the right hemisphere were story responsive. In Broca's area, we found that 43% of voxels in the left hemisphere and 45% of voxels in the right hemisphere were story responsive. In the rest of cortex (i.e., excluding AC, sPMv, and Broca's area), 19% of voxels responded significantly in the left hemisphere and 18% of voxels responded significantly in the right hemisphere. Significantly story-responsive voxels span the entire putative speech-processing pathway (Stowe et al., 2005; Bornkessel-Schlesewsky et al., 2015) from early AC through premotor speech areas and high-level association cortex. This suggests that we should be able to use these data to build and test voxelwise models that capture computations at many different stages of speech processing.

Much of the neuropsychological literature on speech production has shown that cortical speech processing is strongly lateralized (Pujol et al., 1999; Knecht et al., 2002). Therefore, before assessing how different speech features are represented in the cortex or in particular ROIs, we investigated the extent of lateralization of the story-related responses in our dataset. To do this, we simply compared average response reliability (Table 2) across areas and hemispheres, using a linear mixed-effects model. This model had fixed effects of hemisphere (two levels: left and right) and cortical region (four levels: AC, sPMv, Broca's area, and other cortex) and included subject as a random effect. We found that average repeatability was significantly different across cortical regions (Wald χ^2 test, $p = 10^{-11}$) but was not significantly different between the two hemispheres ($p = 0.513$). There was also no significant interaction between hemisphere and cortical region ($p = 0.866$). These results are consistent with earlier reports that natural narrative speech evokes widespread, repeatable BOLD signals across the cerebral cortex (Lerner et al., 2011; Huth et al., 2016) and EEG (Di Liberto et al., 2015). This activity is not as strongly left lateralized as would be expected based on evaluations of the effects of lesions or stimulation on speech production (Pujol et al., 1999; Knecht et al., 2002), or BOLD responses obtained with segmented speech (DeWitt and Rauschecker, 2012).

Total variance explained by spectral, articulatory, and semantic features

The repeatability analysis showed that a relatively large fraction of cortical voxels responds reliably to natural speech and thus may represent information in natural speech. To investigate which types of information are represented in each brain area, we constructed voxelwise models that predict BOLD responses based on different features of the stimuli (Fig. 1). To span the transformation from sound to meaning, we selected the following three feature spaces: spectral features that describe amplitude in frequency channels obtained from a cochleogram; articulatory features that uniquely describe vocal gestures required to make English phonemes (Table 1); and semantic features that describe the meaning of words (see Materials and Methods for the exact definition and estimation of these three feature spaces). We first estimated separate voxelwise models for each feature space using ~2 h of BOLD responses collected while subjects listened to 10 different naturally spoken narrative stories (for details, see Materials and Methods). Model performance was quantified by examining the prediction accuracy using a 10 min story that had not been used for model estimation (Fig. 2). We initially computed model prediction performance as the fraction of variance in the responses explained by the predictions (R^2). However, to enable comparison with earlier studies from our group (Huth et al., 2016), we present results here as the positive square root of R^2, or R. These prediction performance values were then corrected for noise in the model validation dataset (Hsu et al., 2004).

Figure 3 shows the noise-corrected prediction performance values for each feature space, projected onto the cortical flatmap for one subject. The spectral model predicted responses of voxels located in early auditory areas (Fig. 3*A*) that lie in the more anterior and medial part of the AC. The articulatory model predicted responses in early auditory areas, lateral posterior temporal cortex, some sensory and motor mouth areas, and some prefrontal areas (Fig. 3*B*). The semantic model predicted responses broadly across cortex, including in relatively lateral auditory areas, in lateral and inferior regions of the temporal cortex, in many regions of the parietal cortex (specifically the temporoparietal junction and in the medial parietal cortex), and in many regions of the prefrontal cortex (Fig. 3*C*). These areas together have been previously defined as the semantic system (Binder et al., 2009; Huth et al., 2016). None of these feature spaces predicted responses of voxels located in visual cortex, most of somatomotor cortex, or most of insular cortex.

Jointly, these three feature spaces significantly predicted responses in a considerable portion of the story-responsive voxels. Although we did not find differences between left and right hemispheres in overall responses to speech, the analyses based on overall responses do not rule out lateralization effects for specific computations. Based on prior neurophysiological and fMRI studies, one might expect little lateralization for spectral–temporal processing but more for phonetic, articulatory, and semantic processing (DeWitt and Rauschecker, 2012). If this is true, the predictive power of the models using articulatory and semantic stimulus features should be higher in the left than right hemisphere, especially in higher-level areas that are putatively more specialized for language. To test this hypothesis, we used a linear

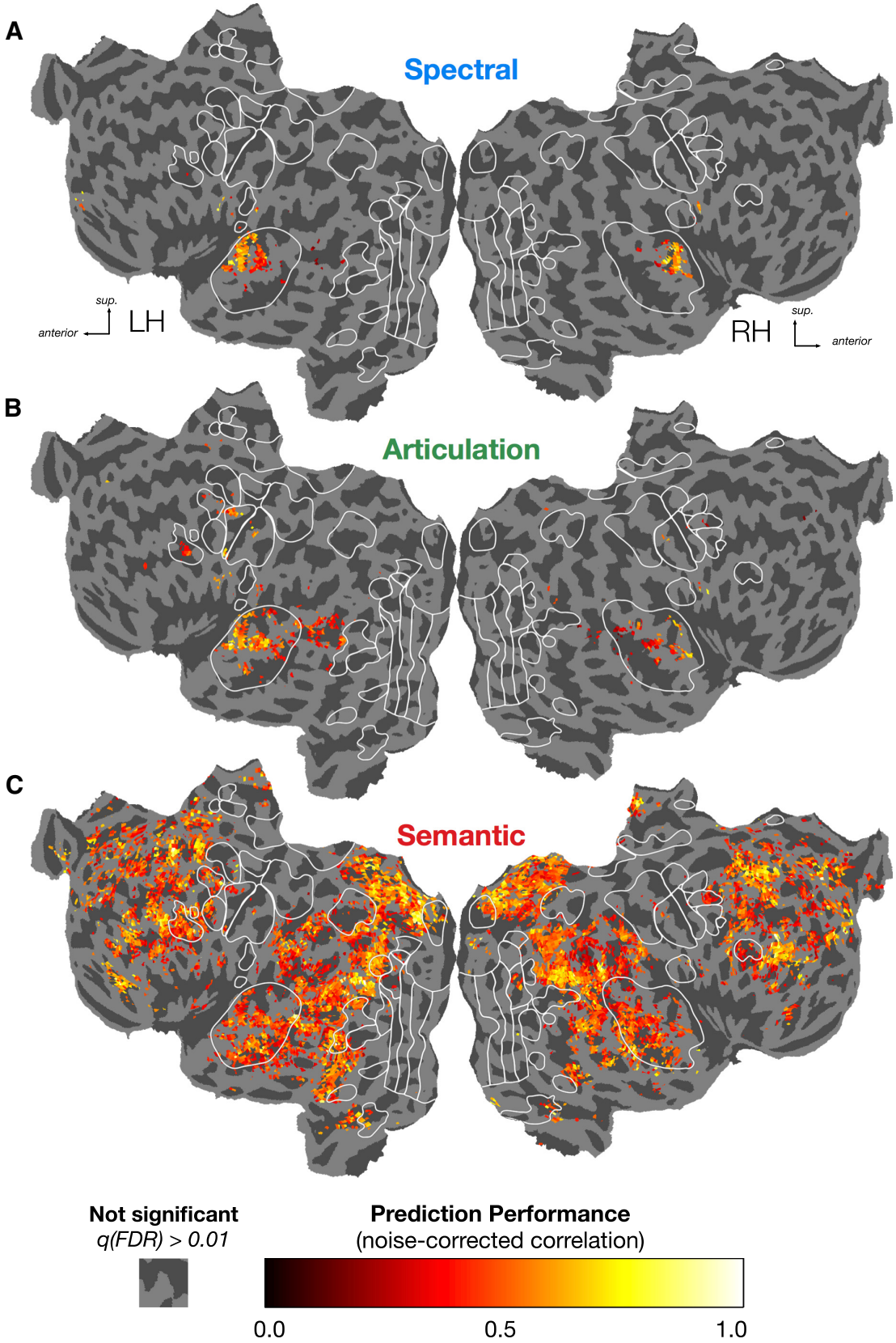

Figure 3. Prediction performance for each feature space. ***A***, Spectral model performance. Spectral model performance plotted on the flattened cortical surface of one subject (subject 2). Color shows the value of the noise-corrected correlation coefficient obtained by comparing the model prediction to actual BOLD responses for the story in the validation dataset. These correlations are normalized by the maximum correlation value that could be obtained given the noise in the signal (see Materials and Methods). Voxels for which the corrected correlation is not significantly different from zero are hidden, revealing the cortical curvature below. White lines encircle regions of interest obtained from separate localizer scans. In this subject, the spectral feature space only produces significant predictions in early auditory cortex around Heschl's gyrus. ***B***, Articulatory model performance. The articulatory feature space significantly predicts voxels in the auditory cortex, as well as in the posterior temporal cortex and frontal cortex. ***C***, Semantic model performance. The semantic feature space significantly predicts voxels in several large regions of cortex, including much of the temporal, parietal, and prefrontal cortex.

Table 3. Prediction performance in story-responsive voxels (noise-corrected correlation), mean of 7 subjects ± SE

	Left hemisphere	Right hemisphere
Spectral model		
Auditory cortex	0.244 (±0.030)	0.216 (±0.014)
sPMv	0.056 (±0.019)	0.077 (±0.017)
Broca	0.108 (±0.024)	0.096 (±0.019)
Other cortex	0.036 (±0.004)	0.023 (±0.007)
Articulatory model		
Auditory cortex	0.268 (±0.013)	0.271 (±0.017)
sPMv	0.220 (±0.029)	0.199 (±0.032)
Broca	0.164 (±0.035)	0.150 (±0.018)
Other cortex	0.071 (±0.010)	0.046 (±0.007)
Semantic model		
Auditory cortex	0.260 (±0.020)	0.303 (±0.012)
sPMv	0.276 (±0.032)	0.294 (±0.055)
Broca	0.259 (±0.031)	0.273 (±0.020)
Other cortex	0.228 (±0.021)	0.201 (±0.025)

mixed-effect model to compare average model performance in story-responsive voxels (Table 3) across feature spaces (three levels), hemispheres (two levels), and cortical regions (four levels: AC, sPMv, Broca's area, and other cortex), with subjects as a random effect. This showed that prediction performance varied significantly across regions (Wald χ^2 test, $p < 2.2 \times 10^{-16}$) and feature spaces ($p < 2.2 \times 10^{-16}$), but not between the cortical hemispheres ($p = 0.74$). There was also a significant interaction between region and feature space ($p = 8.9 \times 10^{-8}$), demonstrating that the feature spaces have different patterns of prediction performance across regions. However, we did not find significant differences between hemispheres either in main effects or interactions. Thus, we found clear evidence of specialized and hierarchical feature representation in different cortical reasons but little evidence for any lateralization. These results are well illustrated on the cortical maps shown for one subject on Figure 3.

Variance partitioning

The model prediction performance maps shown in Figure 3 suggest that responses of many voxels can be significantly predicted by more than one feature space. In more classical approaches, such as block designs using segmented speech, responses in two conditions that putatively differ by one level of language processing (e.g., words vs nonsense words) can be subtracted to find voxels where this difference is statistically significant. This type of result is often interpreted to mean that those voxels are located in brain regions that are specifically responsible for the cognitive function needed for that processing step function (e.g., word recognition; DeWitt and Rauschecker, 2012). Because our experiment used natural stimuli, we could not rely on traditional subtractive analyses to differentiate between neural representations of processing levels. Instead, we used a nested regression approach to distinguish between responses to different types of features. One clear advantage of our approach over traditional subtractive analyses is that we can identify single voxels whose responses can be described in multiple feature spaces. For example, a single voxel could be well modeled by both spectrotemporal features and by semantic features if the two models explained different parts of the response.

To facilitate model fitting and comparison, we designed a method for estimating both the fraction of variance explained by each feature space individually and the fraction that might be equally well explained by any combination of feature spaces (Lescroart et al., 2015). For this purpose, we fit models with all possible combinations of feature spaces, as follows: three models based on a single feature space (spectral, articulatory, and semantic); three models based on pairs of feature spaces (spectral–articulatory, spectral–semantic, and articulatory–semantic); and a single model that used all three feature spaces together. Then, using set theory, we divided the variance explained by these feature spaces into the following seven partitions: the variance explained uniquely by each feature space; the variance explained jointly by each pair of feature spaces but excluding the third; and the variance explained jointly by all three feature spaces (for details, see Materials and Methods). With this approach, we were able to quantify the extent to which any particular voxel represented features from each feature space, after taking into account its selectivity to other feature spaces.

Figure 4 shows the partial correlations (defined as the positive square root of the partial variance explained) for each of the seven partitions, projected on the cortical flatmap for one subject. The three feature spaces jointly explained variance in part of the AC, sPMv, and Broca's area. Unique contributions from the spectral feature space and the spectral–articulatory intersection (excluding contributions from the semantic feature space) were found mostly in medial AC. The articulatory feature space explained little unique variance outside of the AC, but the articulatory–semantic intersection explained some variance in prefrontal cortex and in lateral temporal cortex. The semantic feature space explained a large fraction of the unique variance everywhere outside of early AC and in lateral temporal cortex. The spectral–semantic intersection explained little variance anywhere.

The magnitude of these effects can be visualized by generating Venn diagrams showing the size of each variance partition for the entire cortex (Fig. 5, left), the AC (Fig. 5, middle left), sPMv (Fig. 5, middle right), and Broca's area (Fig. 5, right). Outside of the AC, semantic features best predicted BOLD activity, and the variance explained by the spectral and articulatory features largely overlaps with the variance explained by the semantic features. Each of the three feature spaces explained a similar amount of variance in AC, but there was little overlap in variance explained by spectral and semantic features. The fraction of the variance explained by the articulatory features is, to a large extent, also explained by either or both of the spectral and semantic features. This might be expected for a feature space that can be thought as an "intermediate" step between lower-level spectral representation and the higher-level semantic representation. This continuity in feature representation might be due to the existence of correlations between these features in natural speech. On the other hand, this result may be a consequence of the temporal low-pass nature of BOLD signals, which might limit sensitivity for identifying cortical representations of articulatory and phonemic features.

To visualize the fraction of shared variance in all brain regions, we also generated maps that showed the partition that captured the largest fraction of variance for each voxel (Fig. 6). The articulatory–semantic intersection (excluding any contribution from the spectral feature space) explained the most variance in some regions of lateral temporal cortex just posterior to AC and in prefrontal cortex near and within sPMv. Within AC, the medial portion was best explained by the unique contribution of the spectral feature space, while the intersection of all three feature spaces and the unique contribution of the articulatory feature space seemed to best explain voxels on the STG. Responses in the superior temporal sulcus (STS) were best explained by the unique contribution of the semantic feature space. These maps suggest that there might be a medial–lateral representational gradient in the AC, in which medial voxels represent low-level

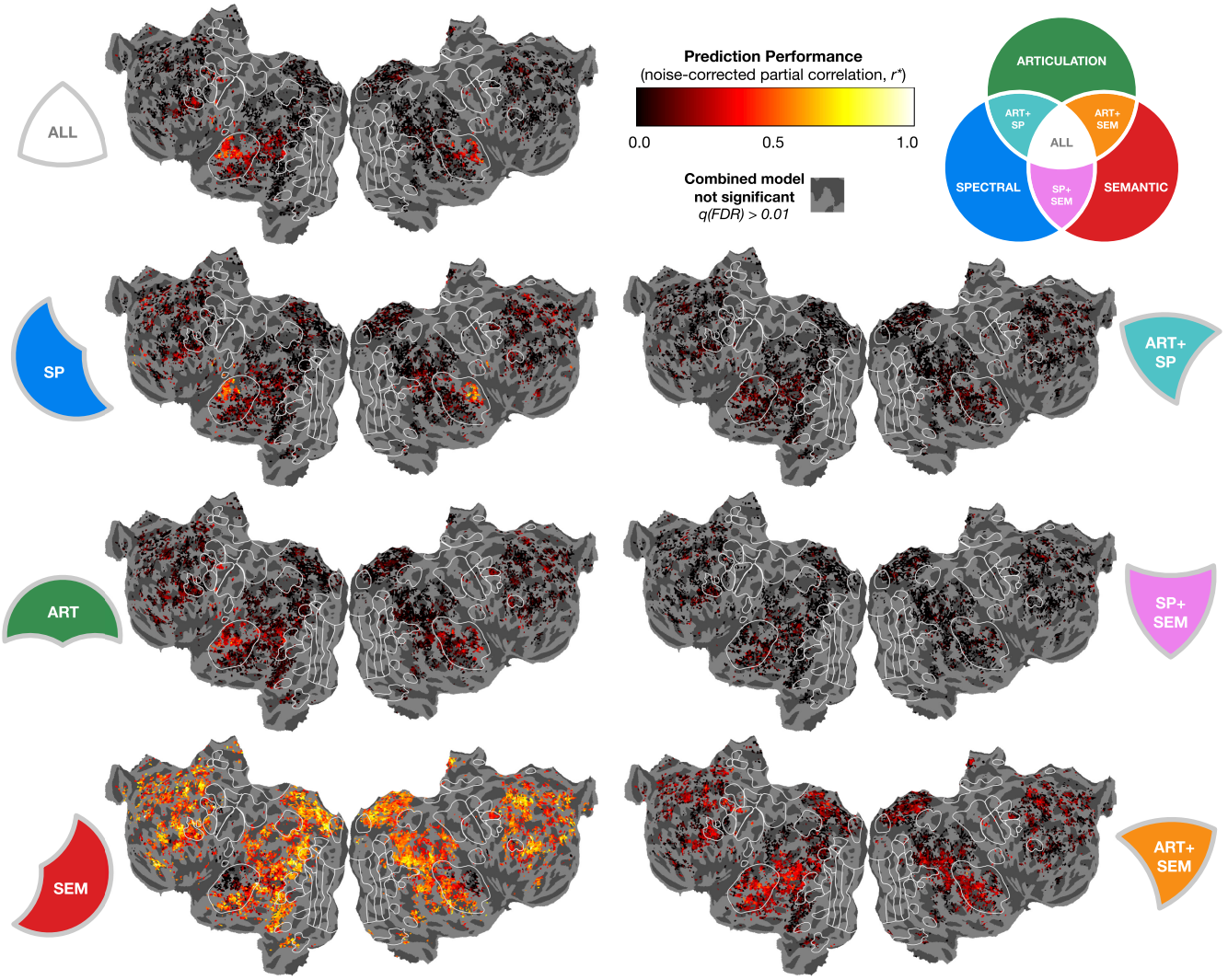

Figure 4. Variance partitions. A variance-partitioning analysis was used to separate the variance explained by the three feature spaces into the following seven partitions: the variance explained uniquely by each feature space; the variance explained jointly by each pair of feature spaces, excluding the third; and the variance explained by all three feature spaces. These flatmaps show noise-corrected partial correlations for each of the seven partitions on one subject. Unique contributions from the spectral feature space and the spectral–articulatory intersection (excluding contributions from the semantic feature space) are mostly limited to early auditory cortex. The articulatory feature space uniquely explains little variance outside of the auditory cortex, but the articulatory–semantic intersection explains some variance in prefrontal cortex and lateral temporal cortex. The semantic feature space uniquely explains a large fraction of the variance everywhere outside of early auditory cortex and lateral temporal cortex. The spectral–semantic intersection explains little variance anywhere.

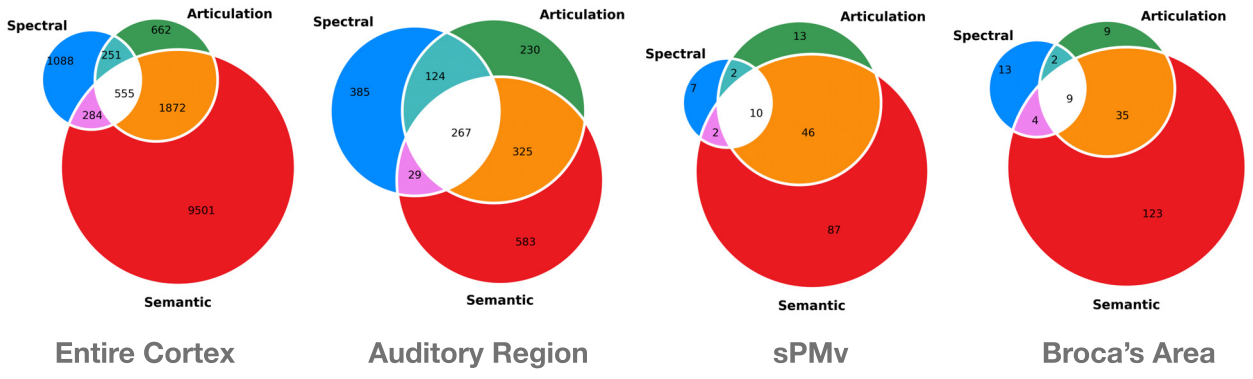

Figure 5. Venn diagrams of explained variance in selected ROIs. Venn diagrams of total explained variance across all subjects (calculated using only significantly predicted voxels) in the entire cortex and three speech-related ROIs. The proportion of variance explained in each partition differs across ROIs. The spectral feature space proportionally explains more variance in the auditory cortex (which includes early auditory cortex) than in other areas. The articulatory feature space and the articulatory–semantic intersection explain proportionally more variance in all speech-related ROIs than in the cortex taken as a whole. The semantic feature space explains proportionally more variance in the entire cortex than in speech-related ROIs.

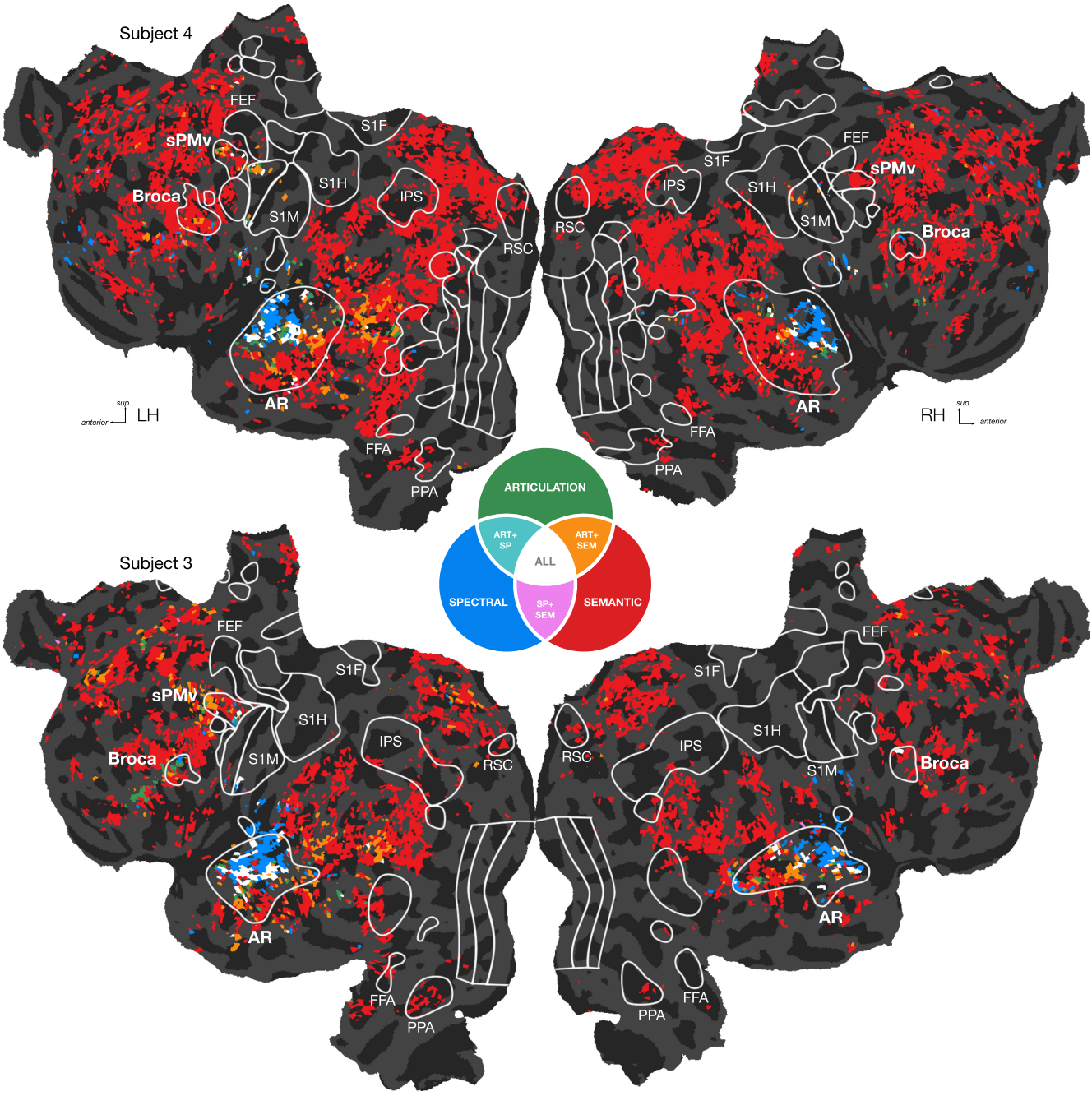

Figure 6. Largest variance partition for each voxel in cortex. Flatmaps show best variance partition for every significantly predicted voxel in two subjects. Each significantly predicted voxel is assigned a color corresponding to the partition that captured the most variance in that voxel. Colors are shown in the legend, center. Within the auditory cortex there is a diverse population of voxels that is best explained by the unique spectral, unique articulatory, unique semantic, spectral–articulatory, or articulatory–semantic features, as well as the combination of all three feature spaces. Some diversity is also seen in the prefrontal speech areas (sPMv and Broca's area). Outside of these areas, the vast majority of voxels is best explained by the semantic feature space alone.

spectral features, voxels on the STG represent mid-level articulatory features (as well as spectral and semantic features), and voxels in the STS represent high-level semantic features.

Our earlier analyses suggested that neither the story-related response nor the variance explained by each individual feature space was lateralized. To complete our analysis of potential lateralization, we also examined whether the variance partitions were different across hemispheres. For example, even though the variance explained by the semantic model was statistically indistinguishable between the left and right side of the brain, one could hypothesize a higher overlap between semantic and articulatory responses on the left hemisphere (based on specialization for word processing; Rodd et al., 2005) and a higher overlap between spectrotemporal features and semantic features on the right hemisphere (based on specialization for slow temporal features and prosody; Abrams et al., 2008). To test potential interactions such as these, we used a linear mixed-effect model to compare partial correlations for each partition (seven levels) across hemispheres (two levels) and cortical regions (four levels: AC, sPMv, Broca's area, and other cortex) with subject as a random effect. This shows that partial correlation varied significantly across cortical regions (Wald χ^2 test, $p = 8.2 \times 10^{-14}$) and partitions ($p <$

2.2×10^{-16}), but not hemispheres ($p = 0.70$). There were significant interactions between region and variance partition ($p < 2.2 \times 10^{-16}$), hemisphere, and variance partition ($p = 0.033$), and the three-way interaction of region, hemisphere, and variance partition ($p = 0.039$). The interaction between region and hemisphere was not significant ($p = 0.88$). A *post hoc* test comparing left and right hemispheres for each region and partition found a significant difference only in sPMv for an additional unique contribution of the semantic feature space (q(FDR) = 0.0022) on the right side. Thus, this interaction analysis reveals some degree of lateralization in feature space representations, but this effect is solely due to the relatively higher unique contribution of semantic features in right sPMV.

Feature space comparison within the auditory cortex

The cortical maps in Figures 4 and 6 suggest that representations of spectral, articulatory, and semantic information within localizer-defined AC (which includes both primary and secondary auditory cortex as well as auditory association cortex) are anatomically segregated. To quantify the organization in AC, we projected the partial correlations onto the medial–lateral axis of STG. We defined the medial–lateral axis by computing the distance from each point on the cortex to the nearest point along the crown of the STG. The partial correlation profiles are shown in Figure 7, averaged across subjects. Positive values are more medial (e.g., on the superior temporal plane), and negative values are more lateral (e.g., in the superior temporal sulcus). The results reveal a clear hierarchical map: the spectral feature space uniquely explains the most variance in medial areas (10–30 mm medial to STG), the articulatory feature space and its intersections explain the most variance around the crown of STG (0 mm) and the semantic feature space uniquely explains the most variance in lateral areas in and around the STS (5–40 mm lateral to STG).

Next, we estimated the center of mass for variance explained by each partition. We found that the centers of mass for the three unique feature space partitions are clearly ordered for both cortical hemispheres in all seven subjects: the spectral feature space is most medial, the semantic space is most lateral, and the articulatory space is represented in between. Among the other partitions, the spectral–articulatory intersection tends to lie medial to the spectral center of mass; the intersection of all models tends to lie between the articulatory and spectral centers of mass; and the articulatory–semantic intersection always lies between the semantic and articulatory centers of mass. (Because the spectral–semantic intersection explains very little variance, its center of mass is difficult to estimate.) This robust effect can clearly be observed across all subjects (Fig. 7, middle panels) and is statistically significant (Friedman rank-sum test: left hemisphere: Friedman $\chi^2 = 36.9$, df = 6, $p = 0.000002$; right hemisphere: Friedman $\chi^2 = 38.8$, df = 6, $p = 0.000001$). Overall, these results demonstrate that there is a strong relationship between the complexity of the speech features represented in BOLD responses and the medial–lateral location in the AC, with spectral features located medially, semantic features located laterally, and articulatory features located in between. This systematic organization is consistent with a hierarchical organization of auditory cortex, where regions that represent lower-level features are located more medially and regions that represent higher-level features are located more laterally.

One possible exception to this hierarchical organization is the spectral–articulatory partition, whose center of mass appears to fall medial to the spectral center of mass. Note, however, that the spectral–articulatory correlation profile mirrors the spectral profile but with lower overall correlation; both were highest at the most medial extent of AC and decreased laterally. We suspect that this biased our estimate of the center of mass for the spectral–articulatory partition to be more medial than the spectral partition.

Semantic selectivity within the auditory cortex

The feature space comparisons in Figure 7 and cortical maps in Figures 4 and 6 suggest that semantic features uniquely explain the largest proportion of variance as early in the auditory hierarchy as the lateral part of STG. This effect cannot be explained away by responses to spectral or articulatory features. However, it remains possible that the lateral STG is in fact responsive to some other features of the narrated stories that are correlated with our semantic representation but not with the spectral or articulatory representations; for example, the high predictions obtained from the semantic representation could reflect generic arousal, which in turn is correlated with particular semantic content. To gain further insights into these potential relationships, we examined the semantic selectivity of voxels within auditory cortex using methods from the study by Huth et al., 2016. In that study, we avoided interpreting semantic selectivity within AC due to confounds with lower-level features. However, using the variance partitioning methods in the current study we were able to select voxels in AC that were uniquely well explained by semantic features for further analysis.

If certain semantic features are correlated with increases in attention or arousal, we would expect to find that semantic feature tuning is homogeneous across the region such that every voxel is selective for the same specific semantic content. To test for this possibility, we examined the variability of voxel semantic selectivity along 10 orthogonal dimensions spanning the highest semantic variability in the regression weights (i.e., the principal components of the semantic weights across all subjects from Huth et al., 2016). For voxels in AC where the semantic features uniquely explained the most variance, the variability of semantic selectivity was no smaller than expected for random brain regions of the same size as AC and located elsewhere (one-sided permutation test, $p \geq 0.1$ for all PCs in left hemisphere; $p \geq 0.055$ for all PCs in the right hemisphere; Fig. 8). This suggests that semantic effects in AC are not driven by a single modulatory mechanism such as attention or arousal. While this analysis included all the semantic voxels in AC, cortical maps of semantic selectivity (Fig. 8) suggest that there is little difference between the semantic selectivity near the STG and the selectivity found deeper in the STS. Both of these areas seem to be primarily selective for social and emotional words. A detailed view of semantic selectivity for one subject can be seen at http://gallantlab.org/huth2016.

This result shows that a single confounding feature cannot explain the semantic effects in AC, but this does not exclude the possibility of confounds with other feature spaces that we did not study here. One possibility is intonation, which is known to drive responses in STG (Zhang et al., 2010). Future studies using natural language could explore additional feature spaces to test such specific hypothesis.

Discussion

This study examined how speech-related spectral, articulatory, and semantic features are represented across human cerebral cortex. To explore this issue, we recorded BOLD activity elicited by natural narrative stories. These stimuli elicit reliable BOLD responses across much of cerebral cortex, including primary and secondary auditory areas, intermediate association areas (Raus-

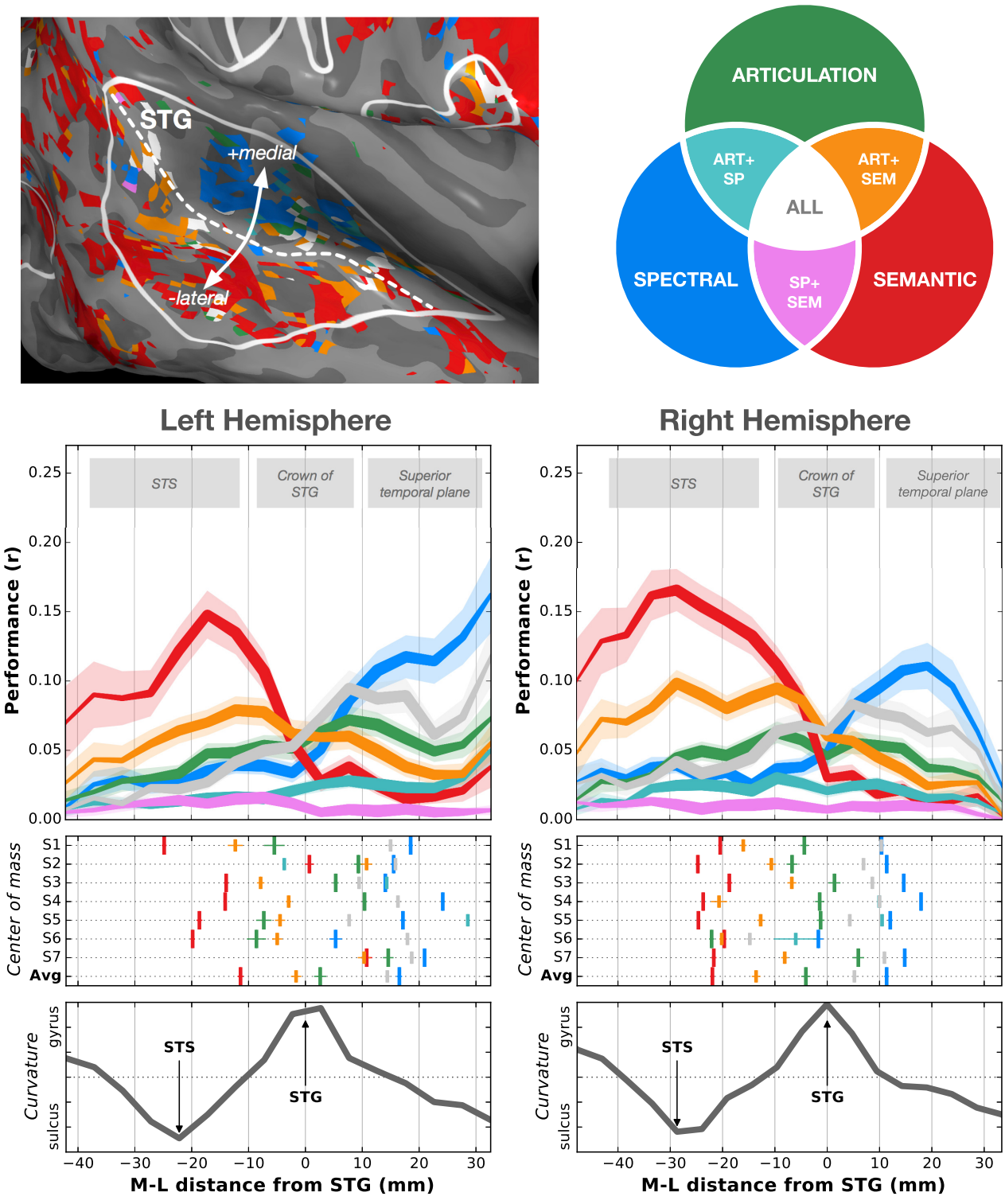

Figure 7. Organization within the auditory cortex. Each point in the auditory cortex was assigned a medial–lateral coordinate based on its distance from the crown of the STG, in millimeters. These were binned into 20 discrete bins, aggregated across subjects. Top panels, Mean noise-corrected partial correlation (among significantly predicted voxels) in each bin for each variance partition. Thickness of solid lines is proportional to the number of voxels falling into each bin, and the shaded areas show ±5 SEs. Middle panels, The center of mass was computed for each variance partition in each subject. Horizontal lines show ±1 SE. Spectral–semantic centers of mass (pink) are missing for some subjects because that partition explains too little variance. The bottom line shows the average center of mass locations across subjects. Bottom panels, Average cortical curvature in each bin, with positive curvature indicating convexity (as on a gyrus) and negative curvature indicating concavity (as in a sulcus).

checker and Scott, 2009; Bornkessel-Schlesewsky et al., 2015), and much of the semantic system (Binder et al., 2009). We then used a voxelwise modeling framework (Nishimoto and Gallant, 2011; Huth et al., 2012, 2016) to build predictive models based on three distinct feature spaces: spectral features, articulatory features, and semantic features. Because these features can be highly correlated in natural speech, we used a variance-partitioning analysis to determine how much of the variance was uniquely

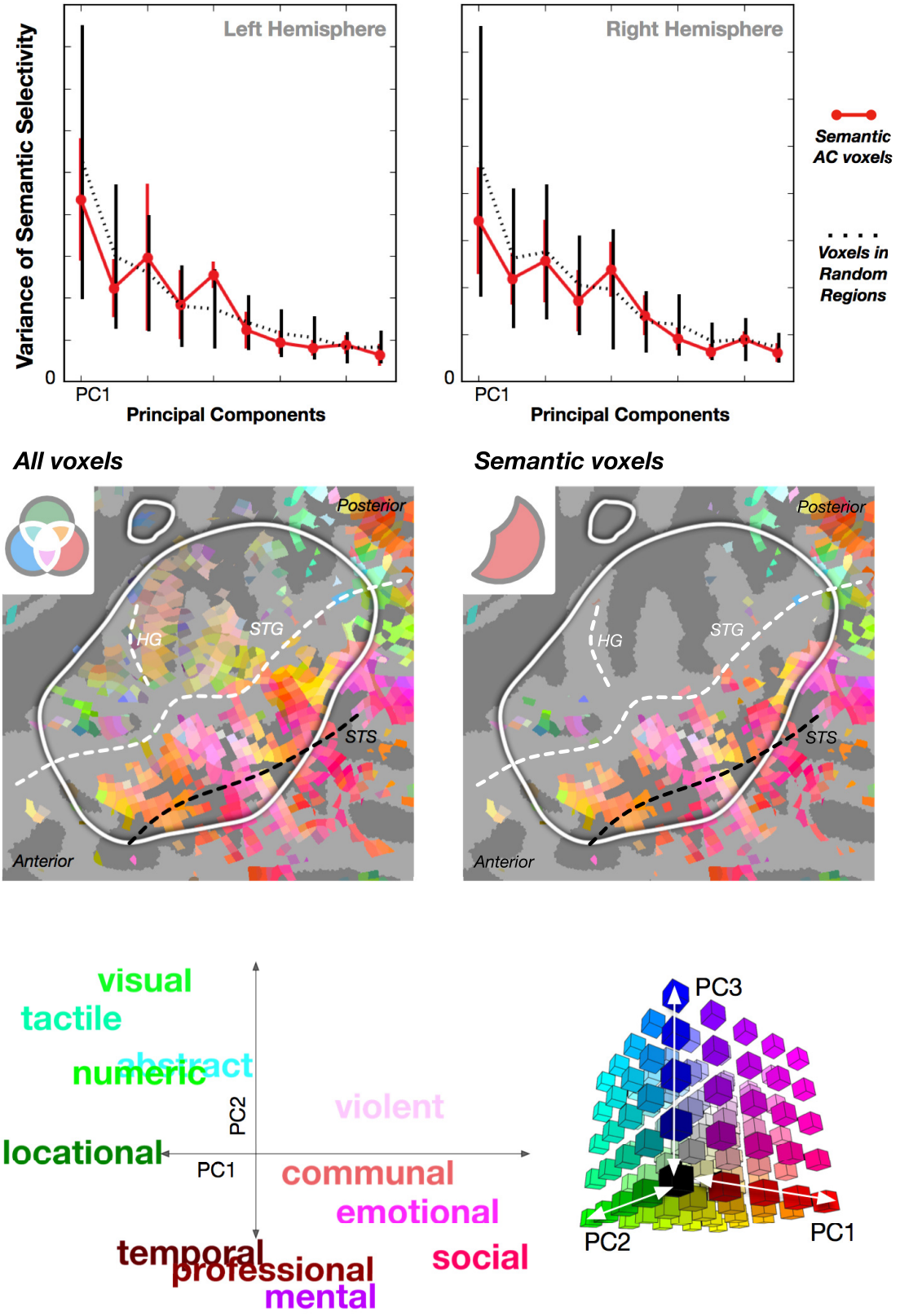

Figure 8. Semantic responses in auditory cortex. If global effects such as attention or arousal manifested were correlated with certain semantic features, then we might expect that semantic selectivity should be homogeneous across AC. Top panels, Variance of semantic selectivity across AC (red) vs 200 similarly sized random cortical regions (black dotted) among significantly predicted voxels where the unique contribution of the semantic feature space is the largest variance partition. Variance was calculated after projecting the regression weights for the semantic model (985 weights) onto the first 10 principal components obtained from the covariance matrix of all semantic dimensions across all voxels and all subjects (Huth et al., 2016). Error bars for the semantic models show SD across subjects. Error bars for the baseline show the SD over the 200 randomly sampled regions, averaged across subjects. Middle panels, Semantic map in the left AC of one subject obtained by color coding the regression weights projected into the first three PCs according to the color scheme shown in the bottom row. Left, The semantic map for all voxels for which any feature space yielded predictions above chance. Right, The semantic tuning for the responses in the voxels that were explained best by semantic features but not the spectral and articulatory features.

predicted by each feature space (Lescroart et al., 2015). We found that voxelwise models based on semantic features produced the most accurate predictions overall and significant predictions for the largest number of voxels. Models based on spectral features predicted activity uniquely in voxels located in primary AC. Models based on articulatory features predicted unique activity in a subset of voxels located in primary and secondary AC, and in a minority of voxels located in sPMv and Broca's area (Möttönen et al., 2013). In contrast, models based on semantic features predicted activity in large areas of temporal cortex (including much of auditory cortex), parietal cortex, and prefrontal cortex (including the speech-specific areas sPMv and Broca's area).

BOLD responses of many voxels in the AC were predicted by at least one of the three feature spaces. Analyzing the anatomical distribution of variance explained within AC revealed a clear hierarchical organization of representations along the medial–lateral axis, as follows: spectral features best predicted activity in the most medial voxels around Heschl's gyrus; articulatory features best predicted activity in the central region along the crown of the STG; and semantic features best predicted activity in the lateral region of auditory cortex along the STS and the medial temporal gyrus. These results are consistent with earlier studies showing that phoneme- and word-form information is represented on the STG, while longer timescale information is represented in the STS (DeWitt and Rauschecker, 2012). A recent fMRI study that also used a data driven approach to obtain functional maps of auditory cortical areas similarly found that the region in STG just lateral to primary AC to be principally responsive to speech features (Norman-Haignere et al., 2015).

One interesting avenue for future research will be to explore how articulatory or phonemic representations are constructed from spectral features. Neurophysiological studies suggest that neurons in auditory cortex selectively encode specific nonlinear combinations of spectral features (Suga et al., 1978; Rauschecker et al., 1995), and studies suggest that the middle STG contains many combination-selective (CS) neurons (Tian et al., 2001). Other neurons are likely to encode nonlinear combinations of features (or of CS neurons) that are invariant to certain spectral features. These invariant response neurons could respond to particular phonemes while ignoring acoustic differences. Although our analyses cannot directly address these underlying neural mechanisms, they may provide some insights. In regions along the crown of the STG, we found that the articulatory feature space explained variance not captured by the spectral feature space. By definition these regions respond invariantly to all sounds produced by that articulation. Future studies using voxelwise modeling might provide clues about these invariant responses. For example, one could extract all sounds that were associated with each specific articulation in the stories, describe their distribution, and then directly assess the degree of acoustic invariance for voxels in STG.

One important finding from this study is that semantic features predicted BOLD activity in a much larger fraction of cortex than articulatory or spectral features. One possible explanation is that semantic features change at a rate lower than the Nyquist limit of fMRI, but spectral and articulatory features change at a rate above the Nyquist frequency. We did, however, find that both spectral and articulatory features predicted responses in early auditory cortex, demonstrating that some information about the representations of these features can be recovered using fMRI. One factor that may have enabled us to recover these representations is that much of the high temporal frequency information in the acoustic signal is represented spatially in the articulatory and spectral features. The spectral feature space nonlinearly decomposes the extremely rapidly changing acoustic signal (up to 10 kHz) into frequency channels that each change much more slowly. Similarly, the articulatory feature space nonlinearly decomposes speech into 22 binary channels that also change much more slowly than the acoustic signal. Both the spectral and articulatory feature spaces serve to demodulate many slowly changing features from the extremely rapid acoustic signal. These feature spaces still discard information about the exact temporal sequencing of events, but otherwise retain much of the total information in the stimulus. In this light, it seems likely that the limited anatomical extent of predictive power that we find for spectral and articulatory features is veridical and due to the relatively small size of cortical regions that specialize in such processing.

One surprising finding is that semantic features predicted a large proportion of variance even within auditory cortex. The presence of semantic information as early as STG might reflect strong top–down modulation related to natural speech processing. Attention and behavioral task have been shown to produce strong top–down effects in auditory cortex in both animals (for review, see Fritz et al., 2007) and humans (Wild et al., 2012; Peelle et al., 2013). However, analyzing the regression weights for semantic features showed that semantic selectivity was highly variable across voxels in AC. Thus, if the semantic effect on BOLD activity in AC is due to an attentional mechanism, it also exhibits a certain degree of linguistic selectivity. The temporal resolution of fMRI does not allow one to disentangle top–down effects from bottom–up effects based on the relative timing of activations measured in different regions, but this question could be addressed in humans with natural speech using other neural signals such as an electrocortigram (Holdgraf et al., 2016) or with experimental designs that directly modulate attentional mechanisms (Çukur et al., 2013).

In the hypothesized dual-stream model of speech processing, the ventral stream is specialized for abstract concepts found in single words and word endings, while the dorsal stream is specialized for temporal sequences of phonemes and words (Turkeltaub and Coslett, 2010; Bornkessel-Schlesewsky et al., 2015). The semantic feature space used here likely reflects both types of information: abstracted concepts are captured directly by the word co-occurrence statistics; and word-order effects are captured at scales of >2 s are captured by the finite-impulse response model. This may account for the widespread effectiveness of the semantic model, which predicts responses well in temporal, frontal, and parietal cortex. In separate work using the same semantic feature space, we have shown that cortex can be parcellated into areas that represent specific semantic features (Huth et al., 2016). We expect that investigations using additional feature spaces could bridge our results and the dual-stream model by showing, for example, that similarly semantically selective brain areas can be distinguished by the roles they play in different computations (e.g., syntax) as postulated in the study by Bornkessel-Schlesewsky et al. (2015). Additional feature spaces that could have correlations with our semantic representation (e.g., intonation) should also be investigated to gain further insights or to rule out alternative explanations for the semantic contribution for speech processing that we found in AC.

We find that all three feature spaces predicted responses about equally well in both cortical hemispheres. This contradicts many previous studies suggesting that speech-related signals are lateralized at both earlier stages (Boemio et al., 2005; Abrams et al., 2008; Desai et al., 2008) and later stages (Pujol et al., 1999; Knecht et al., 2002) of speech processing. One potential explanation for

this discrepancy is that we used narrative speech rather than isolated sentences, segmented speech, or other sounds. At a high level of linguistic processing, narrative comprehension is both engaging and demanding, and it likely recruits brain circuits that are simply not required to perform simpler tasks. And indeed, earlier studies that used narrative speech also reported that both evoked BOLD activity and EEG were bilateral (Lerner et al., 2011; Di Liberto et al., 2015). The processing of low-level speech-like but nonlinguistic features has also shown to be bilateral: this is true for speech phonemes lacking linguistic information (Overath et al., 2015) as well as for tuning for spectral temporal modulations, which show greater within-hemisphere (rostral–caudal) specialization than across hemispheres (Santoro et al., 2014; Norman-Haignere et al., 2015). Second, the fact that all three feature spaces predicted activity in both hemispheres does not imply that the same specific features are represented in the two hemispheres. For example, a more detailed analysis of these data for the semantic features found that the representations in the left and right hemispheres were somewhat different (Huth et al., 2016). Further asymmetries might be revealed by testing feature spaces that capture specific hypotheses about lateralization.

In summary, on one hand, our results are consistent with the view that the brain extracts meaning from sound by a series of feature space transformations that begin in primary auditory cortex and extend laterally. On the other hand, our data suggest that this transformation begins at earlier stages of processing than was thought previously. Further, our results show that the processing of language-related features involves both hemispheres equally, although each hemisphere might be performing different computations. Future studies should examine how specific spectral and articulatory features are mapped across AC and additional feature spaces should be used to further explore specific hypotheses about the speech-processing system beyond AC. We believe that experiments using natural speech and analytical approaches based on nested model comparisons will play a critical role in furthering our understanding of language representation and will complement research based on more traditional approaches.

References

Abrams DA, Nicol T, Zecker S, Kraus N (2008) Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. J Neurosci 28:3958–3965. CrossRef Medline

Belin P, Zatorre RJ (2000) "What", "where" and "how" in auditory cortex. Nat Neurosci 3:965–966. CrossRef Medline

Belin P, Zatorre RJ, Hoge R, Evans AC, Pike B (1999) Event-related fMRI of the auditory cortex. Neuroimage 10:417–429. CrossRef Medline

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57:289–300.

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET (2000) Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex 10:512–528. CrossRef Medline

Binder JR, Desai RH, Graves WW, Conant LL (2009) Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex 19:2767–2796. CrossRef Medline

Boemio A, Fromm S, Braun A, Poeppel D (2005) Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nat Neurosci 8:389–395. CrossRef Medline

Bornkessel-Schlesewsky I, Schlesewsky M, Small SL, Rauschecker JP (2015) Neurobiological roots of language in primate audition: common computational principles. Trends Cogn Sci 19:142–150. CrossRef Medline

Bouchard KE, Mesgarani N, Johnson K, Chang EF (2013) Functional organization of human sensorimotor cortex for speech articulation. Nature 495:327–332. CrossRef Medline

Brant-Zawadzki M, Gillan GD, Nitz WR (1992) MP RAGE - A 3-dimensional, T1-Weighted, Gradient-Echo Sequence—initial experience in the brain. Radiology 182:769–775. CrossRef Medline

Chi T, Gao Y, Guyton MC, Ru P, Shamma S (1999) Spectro-temporal modulation transfer functions and speech intelligibility. J Acoust Soc Am 106:2719–2732. CrossRef Medline

Çukur T, Nishimoto S, Huth AG, Gallant JL (2013) Attention during natural vision warps semantic representation across the human brain. Nat Neurosci 16:763–770. CrossRef Medline

Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis - I. Segmentation and surface reconstruction. Neuroimage 9:179–194. CrossRef Medline

David SV, Gallant JL (2005) Predicting neuronal responses during natural vision. Network 16:239–260. CrossRef Medline

Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41:391–407. CrossRef

Depireux DA, Simon JZ, Klein DJ, Shamma SA (2001) Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophysiol 85:1220–1234. Medline

Desai R, Liebenthal E, Waldron E, Binder JR (2008) Left posterior temporal regions are sensitive to auditory categorization. J Cogn Neurosci 20:1174–1188. CrossRef Medline

DeWitt I, Rauschecker JP (2012) Phoneme and word recognition in the auditory ventral stream. Proc Natl Acad Sci U S A 109:E505–E514. CrossRef Medline

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol 25:2457–2465. CrossRef Medline

Elliott TM, Theunissen FE (2009) The modulation transfer function for speech intelligibility. PLoS Comput Biol 5:e1000302. CrossRef Medline

Escabi MA, Schreiner CE (2002) Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. J Neurosci 22:4114–4131. Medline

Fedorenko E, Nieto-Castañón A, Kanwisher N (2012) Syntactic processing in the human brain: what we know, what we don't know, and a suggestion for how to proceed. Brain Lang 120:187–207. CrossRef Medline

Flinker A, Korzeniewska A, Shestyuk AY, Franaszczuk PJ, Dronkers NF, Knight RT, Crone NE (2015) Redefining the role of Broca's area in speech. Proc Natl Acad Sci U S A 112:2871–2875. CrossRef Medline

Fritz JB, Elhilali M, David SV, Shamma SA (2007) Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in Al? Hear Res 229:186–203. CrossRef Medline

Gill P, Zhang J, Woolley SM, Fremouw T, Theunissen FE (2006) Sound representation methods for spectro-temporal receptive field estimation. J Comput Neurosci 21:5–20. CrossRef Medline

Holdgraf CR, de Heer W, Pasley B, Rieger J, Crone N, Lin JJ, Knight RT, Theunissen FE (2016) Rapid tuning shifts in human auditory cortex enhance speech intelligibility. Nat Commun 7:13654. CrossRef Medline

Hsu A, Borst A, Theunissen FE (2004) Quantifying variability in neural responses and its application for the validation of model predictions. Network 15:91–109. CrossRef Medline

Hullett PW, Hamilton LS, Mesgarani N, Schreiner CE, Chang EF (2016) Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. J Neurosci 36:2014–2026. CrossRef Medline

Huth AG, Nishimoto S, Vu AT, Gallant JL (2012) A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76:1210–1224. CrossRef Medline

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532:453–458. CrossRef Medline

Jenkinson M, Smith S (2001) A global optimisation method for robust affine registration of brain images. Med Image Anal 5:143–156. CrossRef Medline

Jones E, Oliphant T, Peterson P (2007) SciPy: open source scientific tools for Python. Beaverton, OR: Python Software Foundation. Available at: http://www.scipy.org/.

Just MA, Cherkassky VL, Aryal S, Mitchell TM (2010) A neurosemantic theory of concrete noun representation based on the underlying brain codes. PLoS One 5:e8622. CrossRef Medline

Knecht S, Flöel A, Dräger B, Breitenstein C, Sommer J, Henningsen H, Ringelstein EB, Pascual-Leone A (2002) Degree of language lateralization determines susceptibility to unilateral brain lesions. Nat Neurosci 5:695–699. CrossRef Medline

Leaver AM, Rauschecker JP (2010) Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J Neurosci 30:7604–7612. CrossRef Medline

Lerner Y, Honey CJ, Silbert LJ, Hasson U (2011) Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J Neurosci 31:2906–2915. CrossRef Medline

Lescroart MD, Stansbury DE, Gallant JL (2015) Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. Front Comput Neurosci 9:135. CrossRef Medline

Levelt WJ (1993) Speaking: from intention to articulation. Cambridge, MA: MIT.

Lund K, Burgess C (1996) Hyperspace analogue to language (HAL): a general model semantic representation. Brain Cogn 30:5.

Lyon RF (1982) A computational model of filtering, detection and compression in the cochlea. Paper presented at IEEE International Conference on Acoustics, Speech, and Signal Processing. Paris, France, May.

Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485:233–236. CrossRef Medline

Mesgarani N, Slaney M, Shamma SA (2006) Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. IEEE Trans Audio Speech Lang Process 14:920–930. CrossRef

Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. Science 343:1006–1010. CrossRef Medline

Miller LM, Escabí MA, Read HL, Schreiner CE (2002) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol 87:516–527. CrossRef Medline

Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, Just MA (2008) Predicting human brain activity associated with the meanings of nouns. Science 320:1191–1195. CrossRef Medline

Möttönen R, Dutton R, Watkins KE (2013) Auditory-Motor Processing of Speech Sounds. Cereb Cortex 23:1190–1197. CrossRef Medline

Moore RC, Lee T, Theunissen FE (2013) Noise-invariant neurons in the avian auditory cortex: hearing the song in noise. PLoS Comput Biol 9:e1002942. CrossRef Medline

Nishimoto S, Gallant JL (2011) A three-dimensional spatiotemporal receptive field model explains responses of area MT neurons to naturalistic movies. J Neurosci 31:14551–14564. CrossRef Medline

Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL (2011) Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol 21:1641–1646. CrossRef Medline

Norman-Haignere S, Kanwisher NG, McDermott JH (2015) Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88:1281–1296. CrossRef Medline

Oliphant TE (2006) A guide to NumPy. Austin, TX: Continuum Publishing.

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

Peelle JE, Gross J, Davis MH (2013) Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb Cortex 23:1378–1387. CrossRef Medline

Poeppel D, Emmorey K, Hickok G, Pylkkänen L (2012) Towards a new neurobiology of language. J Neurosci 32:14125–14131. CrossRef Medline

Price CJ (2010) The anatomy of language: a review of 100 fMRI studies published in 2009. Ann N Y Acad Sci 1191:62–88. CrossRef Medline

Pujol J, Deus J, Losilla JM, Capdevila A (1999) Cerebral lateralization of language in normal left-handed people studied by functional MRI. Neurology 52:1038–1043. CrossRef Medline

Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci 12:718–724. CrossRef Medline

Rauschecker JP, Tian B, Hauser M (1995) Processing of complex sounds in the macaque nonprimary auditory cortex. Science 268:111–114. CrossRef Medline

Rodd JM, Davis MH, Johnsrude IS (2005) The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. Cereb Cortex 15:1261–1269. Medline

Sahani M, Linden J (2003) Evidence optimization techniques for estimating stimulus-response functions. In: Advances in neural infomation processing systems (Becker S, Thrun S, Obermeyer K, eds), pp 301–308. Cambridge, MA: MIT.

Santoro R, Moerel M, De Martino F, Goebel R, Ugurbil K, Yacoub E, Formisano E (2014) Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput Biol 10:e1003412. CrossRef Medline

Scott SK, Johnsrude IS (2003) The neuroanatomical and functional organization of speech perception. Trends Neurosci 26:100–107. CrossRef Medline

Sen K, Theunissen FE, Doupe AJ (2001) Feature analysis of natural sounds in the songbird auditory forebrain. J Neurophysiol 86:1445–1458. Medline

Singh NC, Theunissen FE (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. J Acoust Soc Am 114:3394–3411. CrossRef Medline

Slaney M (1998) Lyon's cochlear model. Interval Research Corproation Technical Report #1998-010.

Stowe LA, Haverkort M, Zwarts F (2005) Rethinking the neurological basis of language. Lingua 115:997–1042. CrossRef

Suga N, O'Neill WE, Manabe T (1978) Cortical neurons sensitive to combinations of information-bearing elements of biosonar signals in the moustache bat. Science 200:778–781. CrossRef Medline

Tian B, Reser D, Durham A, Kustov A, Rauschecker JP (2001) Functional specialization in rhesus monkey auditory cortex. Science 292:290–293. CrossRef Medline

Turkeltaub PE, Coslett HB (2010) Localization of sublexical speech perception components. Brain Lang 114:1–15. CrossRef Medline

Turney PD, Pantel P (2010) From frequency to meaning: vector space models of semantics. J Artif Intell Res 37:141–188. CrossRef

Visser M, Jefferies E, Lambon Ralph MA (2010) Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature. J Cogn Neurosci 22:1083–1094. CrossRef Medline

Wehbe L, Murphy B, Talukdar P, Fyshe A, Ramdas A, Mitchell T (2014) Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. PLoS One 9:e112575. CrossRef Medline

Welch P (1967) The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans Audio Electroacoust 15:70–73. CrossRef

Wild CJ, Davis MH, Johnsrude IS (2012) Human auditory cortex is sensitive to the perceived clarity of speech. Neuroimage 60:1490–1502. CrossRef Medline

Wilson SM, Saygin AP, Sereno MI, Iacoboni M (2004) Listening to speech activates motor areas involved in speech production. Nat Neurosci 7:701–702. CrossRef Medline

Woolley SM, Gill PR, Fremouw T, Theunissen FE (2009) Functional groups in the avian auditory system. J Neurosci 29:2780–2793. CrossRef Medline

Zhang L, Shu H, Zhou F, Wang X, Li P (2010) Common and distinct neural substrates for the perception of speech rhythm and intonation. Hum Brain Mapp 31:1106–1116. CrossRef Medline

Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013) Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron 77:980–991. CrossRef Medline