Behavioral/Cognitive

Neural Tuning to Low-Level Features of Speech throughout the Perisylvian Cortex

Julia Berezutskaya,[1,2] Zachary V. Freudenburg,[1] Umut Güçlü,[2] Marcel A.J. van Gerven,[2] and Nick F. Ramsey[1]
[1]Brain Center Rudolf Magnus, Department of Neurology and Neurosurgery, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands, and
[2]Radboud University, Donders Institute for Brain, Cognition and Behaviour, 6525 HR Nijmegen, The Netherlands

Despite a large body of research, we continue to lack a detailed account of how auditory processing of continuous speech unfolds in the human brain. Previous research showed the propagation of low-level acoustic features of speech from posterior superior temporal gyrus toward anterior superior temporal gyrus in the human brain (Hullett et al., 2016). In this study, we investigate what happens to these neural representations past the superior temporal gyrus and how they engage higher-level language processing areas such as inferior frontal gyrus. We used low-level sound features to model neural responses to speech outside of the primary auditory cortex. Two complementary imaging techniques were used with human participants (both males and females): electrocorticography (ECoG) and fMRI. Both imaging techniques showed tuning of the perisylvian cortex to low-level speech features. With ECoG, we found evidence of propagation of the temporal features of speech sounds along the ventral pathway of language processing in the brain toward inferior frontal gyrus. Increasingly coarse temporal features of speech spreading from posterior superior temporal cortex toward inferior frontal gyrus were associated with linguistic features such as voice onset time, duration of the formant transitions, and phoneme, syllable, and word boundaries. The present findings provide the groundwork for a comprehensive bottom-up account of speech comprehension in the human brain.

Key words: inferior frontal gyrus; language; modeling; neural encoding; speech comprehension

Significance Statement

We know that, during natural speech comprehension, a broad network of perisylvian cortical regions is involved in sound and language processing. Here, we investigated the tuning to low-level sound features within these regions using neural responses to a short feature film. We also looked at whether the tuning organization along these brain regions showed any parallel to the hierarchy of language structures in continuous speech. Our results show that low-level speech features propagate throughout the perisylvian cortex and potentially contribute to the emergence of "coarse" speech representations in inferior frontal gyrus typically associated with high-level language processing. These findings add to the previous work on auditory processing and underline a distinctive role of inferior frontal gyrus in natural speech comprehension.

Introduction

Speech is a specific kind of complex sound and, similarly to complex object processing in vision (e.g., face processing), a substantial amount of research has focused on the neural tuning to low-level properties of this type of sensory input. Comparing tones revealed that primary auditory cortex and immediately adjacent superior temporal gyrus (STG) are organized tonotopically (Wessinger et al., 1997; Bilecen et al., 1998; Formisano et al., 2003) and modulated by various spectral and temporal properties of sound (Giraud et al., 2000; Joris et al., 2004; Altmann et al., 2010; Leaver and Rauschecker, 2010).

There is also a large body of literature that describes speech in the context of language theory. These studies relate different higher-level language features from phonemes to phrase units to the neural responses to speech throughout the cortex (Giraud and Poeppel, 2012; Hagoort and Indefrey, 2014; Ding et al., 2016). Altogether, these studies outline a network of key language processing regions: STG, anterior temporal cortex, inferior frontal gyrus (IFG), and middle temporal, precentral, and angular gyri (Hickok and Poeppel, 2007; Friederici, 2012; Hagoort, 2013). However, establishing a connection between low-level auditory

Received Jan. 25, 2017; revised July 4, 2017; accepted July 9, 2017. Author contributions: J.B., Z.V.F., and N.F.R. designed research; J.B., Z.V.F., U.G., M.A.J.v.G., and N.F.R. performed research; J.B. analyzed data; J.B., Z.V.F., U.G., M.A.J.v.G., and N.F.R. wrote the paper.

This work was supported by the European Research Council (Advanced iConnect Project Grant ADV 320708) and the Netherlands Organisation for Scientific Research (Language in Interaction Project Gravitation Grant 024.001.006). We thank Frans Leijten, Cyrille Ferrier, Geertjan Huiskamp, and Tineke Gebbink for help in collecting data; Peter Gosselaar and Peter van Rijen for implanting the electrodes; the technicians and staff of the clinical neurophysiology department and the patients for their time and effort; and the members of the UMC Utrecht ECoG research team for data collection.

The authors declare no competing financial interests.

Correspondence should be addressed to Julia Berezutskaya, Oranjesingel 72, 6511 NZ Nijmegen, The Netherlands. E-mail: y.berezutskaya@umcutrecht.nl.

DOI:10.1523/JNEUROSCI.0238-17.2017

 0270-6474/17/377906-15$15.00/0

features and higher-level language structures in continuous speech remains a challenging task.

In the present study, we focus on the neural tuning to acoustic features of speech along the perisylvian cortex. We believe that understanding the anatomical and temporal characteristics of the neural tuning to the low-level features of naturalistic speech will provide the groundwork for a bottom-up account of speech perception. Hopefully, a comprehensive bottom-up account of speech perception can be further integrated into existing theoretical frameworks of language processing in the brain (Hickok and Poeppel, 2007; Friederici, 2012; Hagoort, 2013), altogether providing an extensive model of neural processing of speech.

Here, we used continuous naturalistic stimuli, which were segments from a feature film ("Pippi Longstocking," 1969), to investigate the neural tuning to speech and compare it with the neural tuning to another type of complex sound: music. We used linear regression to model how neural responses are elicited by low-level sound features. The neural responses to the movie were obtained using electrocorticography (ECoG). Subsequently, we obtained and processed analogous fMRI data to assess the reproducibility of ECoG findings with a more commonly accessible technique and at higher spatial sampling.

Our results suggest a propagation of temporal features of the speech audio signal extending beyond the auditory network into the IFG during speech comprehension. The temporal features passed along the pathway were associated with distinct linguistic markers (from subphonemic features to syllable and word boundaries). The present evidence suggests that, as in the visual domain, increasingly complex stimulus features are processed in downstream areas, pinpointing the similarity in neural coding across perceptual modalities.

Materials and Methods

Movie stimulus and audio processing

A series of 30 s video fragments from a feature film ("Pippi Longstocking," 1969) were concatenated into a coherent storyline. Fragments containing conversations between people were alternated with fragments containing music (no speech) while the visual story proceeded. The resulting movie lasted 6.5 min. Speech was dubbed because the original language was Swedish. There were two repeating music fragments; therefore, one of them was omitted from all the analyses. Therefore, six music and six speech 30 s fragments were analyzed.

The NSL toolbox (Chi et al., 2005) was used to extract low-level spectrotemporal features from the audio signal. The computational model by Chi et al. (2005) is biologically inspired and implements two stages of sound processing: the cochlea stage, during which the audio spectrogram is computed at logarithmically spaced central frequencies in the range 180–7000 Hz, and the early cortical stage, during which a nonlinear spectrotemporal representation of sound is obtained by filtering the spectrogram with various modulation filters.

The soundtrack of the short movie was down-sampled to 16 kHz and a sound spectrogram was obtained at 8 ms frames for 128 logarithmically spaced frequency bins in the range of 180–7000 Hz (Fig. 1*A*). Next, a nonlinear modulation-based feature representation of the sound was obtained by filtering the spectrogram at every audio frequency bin and time point with a bank of 2D Gabor filters varying in spatial frequency and orientation. The output was a 4D set of low-level audio features: spectral modulations (SMs) × temporal modulations (TMs) × frequency bin × time point. The nonlinear features were obtained along the dimensions of SMs over the range of 0.2–16 cycles per octave (cyc/oct) in linear steps of 0.2 cyc/oct, and TMs over the range of 0.2–16 Hz in linear steps of 0.2 Hz. The resulting feature set was of size 81 × 162 × 128 × 3750 per each 30 s fragment. The features along the dimension of TMs accounted for upward and downward sweeps in the spectrogram and therefore contained twice as many features.

To quantify the difference in the spectrotemporal features between speech and music fragments, within each 30 s fragment, the feature set was averaged along the frequency bins. A 1-s-long vector per each combination of SM and TM features was obtained for speech and music separately by averaging the SM–TM features over all seconds of all sound-specific fragments. For each combination of SM and TM features, the music and speech 1-s-long vectors were compared with a two-tailed *t* test. The *p*-values were set at 0.001, Bonferroni corrected for the number of SM–TM features (81 × 162 in total). Throughout the manuscript, we report the *p*-value threshold before its correction for multiple comparisons. Therefore, the actual *p*-value threshold is smaller than the reported one. In case of the Bonferroni correction, the actual *p*-value threshold is the reported threshold over the number of comparisons.

For the visualization (Fig. 1*B*,*C*), within each fragment, the original size feature set (81 × 162 × 128 × 3750) was averaged along the frequency bins and time points. Then, the SM–TM features were averaged across sound-specific fragments to obtain a single averaged SM–TM feature set for speech and music separately.

Experimental design and statistical analysis

ECoG experiment

ECoG setup and movie-watching experiment. Fifteen patients (age 28 ± 12, nine females) with medication-resistant epilepsy underwent subdural electrode implantation to determine the source of seizures and test the possibility of surgical removal of the corresponding brain tissue. All patients gave written informed consent to participate in accompanying ECoG and fMRI recordings and gave permission to use their data for scientific research. The study was approved by the Medical Ethical Committee of the Utrecht University Medical Center in accordance with the Declaration of Helsinki (2013).

All patients were implanted with clinical electrode grids (2.3 mm exposed diameter, interelectrode distance 10 mm, between 64 and 120 contact points), one patient had a high-density grid (1.3 mm exposed diameter, interelectrode distance 3 mm). Thirteen patients were implanted with left hemispheric grids. Most had left hemisphere as language dominant (based on fMRI or Wada test). All patients had perisylvian grid coverage and most had electrodes in frontal and motor cortices. The total brain coverage is shown in Figure 2. Patient-specific information about the grid hemisphere, number of electrodes, and cortices covered is summarized in Table 1.

In the movie-watching experiment, each patient was asked to attend to the short movie made of fragments from "Pippi Longstocking". The video was delivered on a computer screen (21 inches in diagonal). The stereo sound was delivered through speakers with the volume level adjusted for each patient.

During the experiment, ECoG data were acquired with a 128 channel recording system (Micromed) at a sampling rate of 512 Hz filtered at 0.15–134.4 Hz. The movie was presented using Presentation software (Version 18.0, Neurobehavioral Systems) and sound was synchronized with the ECoG recordings.

ECoG data processing. All electrodes with noisy or flat signal (visual inspection) were excluded from further analyses. After applying a notch filter for line noise (50 and 100 Hz), common average rereferencing was applied separately for clinical and high-density grids. Data were transformed to the frequency domain using Gabor wavelet decomposition at 1–120 Hz in 1 Hz bins with decreasing window length (4 wavelength full-width at half maximum, FHWH). Finally, high-frequency band (HFB) amplitude was obtained by averaging amplitudes for the 60–120 Hz bins and the resulting time series per electrode were down-sampled to 125 Hz.

Electrode locations were coregistered to the anatomical MRI in native space using computer tomography scans (Hermes et al., 2010) and Freesurfer (http://surfer.nmr.mgh.harvard.edu/). The Desikan–Killiany atlas (Desikan et al., 2006) was used for anatomical labeling of electrodes (closest cortical structure in the radius of 5 mm). All electrode positions were projected to Montreal Neurological Institute (MNI) space using SPM8 (Wellcome Trust Centre for Neuroimaging, University College London).

Neural tuning to low-level sound features in ECoG

Feature space. To model ECoG HFB responses, features describing SMs, TMs, and time lag were used. To reduce the computational load, reduced spectrotemporal feature sets were obtained. The reduced feature set in-

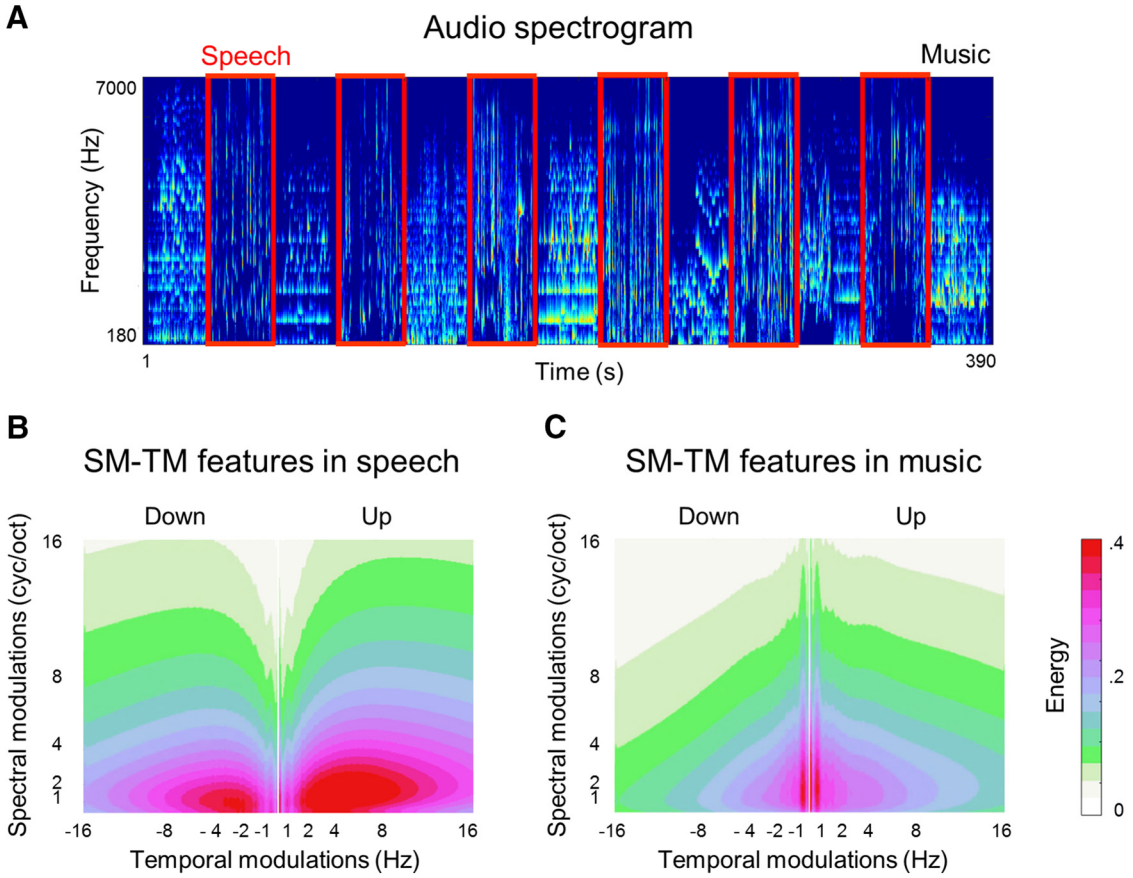

Figure 1. Low-level properties of sound. ***A***, Audio spectrogram of the soundtrack of the 6.5 min video stimulus. Horizontal axis represents time and vertical axis represents frequency scale. The spectrogram was obtained using NSL toolbox at 8 ms windows (sampling rate = 125 Hz). The frequency bins are spaced logarithmically (128 bins) in the range of 180–7000 Hz. Speech blocks are framed in red. Music blocks are the rest. ***B***, Low-level spectrotemporal (SM–TM) features of speech. Horizontal axis represents TMs measured in Hertz; vertical axis represents SMs measured in cycles per octave (cyc/oct). The features were extracted using the NSL toolbox by convolving the spectrogram with a bank of 2D Gabor filters at different spatial frequencies and orientations. The SM–TM features were extracted for every time point and every frequency bin. TMs were obtained at 0.25–16 Hz, in linear steps of 0.2 Hz. SMs were obtained at 0.25–16 cyc/oct in linear steps of 0.2 cyc/oct. Positive and negative sign representations along the TM axis capture the direction of energy sweeps in the spectrogram (up and down). The extracted SM–TM features were averaged over time points within block and across blocks. ***C***, Low-level SM–TM features of music. The feature extraction and visual representation are identical to the SM–TM features of speech in ***B***.

cluded 17 features spaced logarithmically along the spectral dimension (0.03–8 cyc/oct) and 17 features spaced logarithmically along the temporal dimension (0.25–64 Hz). The resulting feature set was of size $17 \times 34 \times 128 \times 3750$ per sound-specific fragment. The features along the dimension of TMs accounted for upward and downward sweeps in the spectrogram and therefore contained twice as many features. The features were averaged along 128 frequency bins, as well as along upward and downward sweeps, resulting in a $17 \times 17 \times 3750$ feature set per fragment. Due to a possible delay of the neural responses with respect to the audio onset, all negative time lags up to −500 ms relative to the audio onset were added to the feature set as the third feature dimension. Therefore, to model a neural response at time t, all SMs and TMs within $[t - 500, t]$ were used. At the sampling rate of 125 Hz, this resulted in 63 time lags. The resulting feature matrix had $17 \times 17 \times 63$ features per time point.

Encoding model. Kernel ridge regression (Murphy, 2012) was used to model responses of an electrode (y_e) as a linear combination of nonlinear stimulus features (X) as follows:

$$\mathbf{y}_e = \boldsymbol{\beta}_e{}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\varepsilon}_e \quad (1)$$

where $\boldsymbol{\varepsilon}_e \sim \mathcal{N}(0, \sigma^2)$.

An L^2 penalized least-squares loss function was analytically minimized to estimate the regression coefficients (β_e) and the kernel trick was used to avoid large matrix inversions in the feature space as follows:

$$\boldsymbol{\beta}_e = \mathbf{X}^{\mathrm{T}}(\mathbf{K} + \lambda_e \mathbf{I}_n)^{-1}\mathbf{y}_i \quad (2)$$

where X and y_e is an estimation set with n data cases and matrix K is the Gram matrix as follows:

$$\mathbf{K} = \mathbf{X}\mathbf{X}^{\mathrm{T}} \quad (3)$$

A nested cross-validation was used to estimate the complexity parameter that controls the amount of regularization (λ_e) (Güçlü and van Gerven, 2014). First, a grid of the effective degrees of freedom of the model fit was specified. Then, Newton's method was used to solve the effective degrees of freedom for λ_e. Finally, the λ_e that resulted in the lowest nested cross-validation error was taken as the final estimate.

Model evaluation. A fivefold cross-validation was used to validate the model performance. The validation procedure was used to tackle overfitting; that is, to obtain results generalizable to novel audio data. The model was trained using data from music and speech fragments combined. We tested the model on predicting neural responses to music and speech separately. Each of the sound-specific 30 s fragments was partitioned into five 6 s chunks. When testing the model on speech data, one chunk was selected per speech fragment to make up a test set. In each cross-validation fold, different chunks were selected per fragment so that no data points were shared in test datasets across five folds. The first second of each chunk used as a test set was discarded because the preceding lag information during the chunk's first 500 ms was not part of the test set. The five truncated chunks were concatenated and constituted the test set. The remaining data from both speech and music fragments were used for training. The data immediately following the chunk used in the test set (up to 1 s) were also excluded from the training set because the lag information for these data was used in the test set. Therefore, no data points were shared between train and test datasets in one cross-validation fold. The procedure was identical when selecting test data from music fragments to test the model on music data.

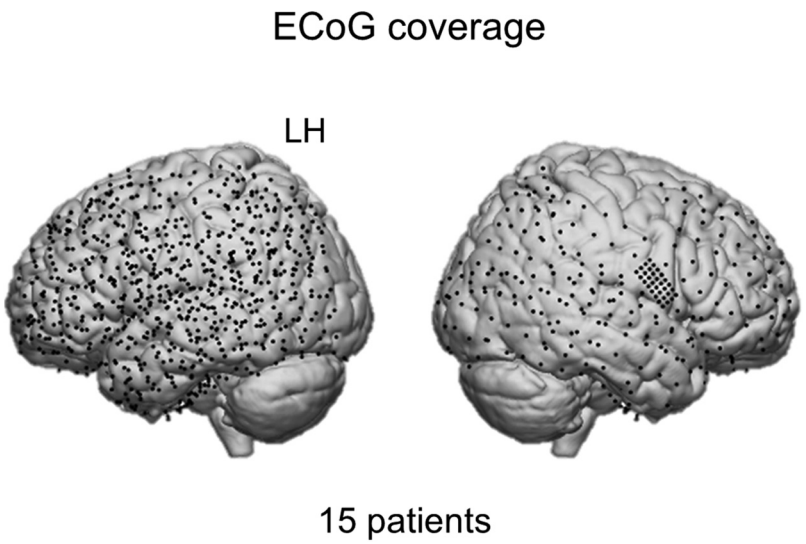

Figure 2. ECoG coverage. Per patient, the electrode locations were normalized using individual affine transformation matrices obtained with SPM8. The normalized electrode locations were pooled across patients and displayed on the standard MNI brain to show overall brain coverage.

Table 1. Electrode grid information

Patient	No. of electrodes	Hemisphere	Cortices covered	Handedness	Language dominance
1	96	LH	F, T, P	R	L (Wada)
2	112	LH	F, M, T, P, O	R	L (fMRI)
3	96	LH	F, M, T, P, O	R	L (Wada)
4	104	LH	F, M, T, P	R	L (Wada)
5	48	LH	F, M, T, P	L	L (fMRI)
6	120	LH	F, M, T	R	L (Wada)
7	112	LH	F, M, T, P	R	L (fMRI)
8	64	LH	F, M, T	R	L (fMRI)
9	112	LH	M, T, P, O	R	L (fMRI)
10	64	LH	M, T, P	R	L (Wada)
11	96	RH	M, T, P, O	L	R (Wada)
12	88	LH	T, P, O	R	L (fMRI)
13	112	LH	F, T, P	R	R (Wada)
14	120	RH	F, M, T	R	L (Wada)
15	96	RH	F, T, P, O	L	L (Wada)

Shown is information about the number of electrodes, grid hemisphere, covered cortices, handedness, and language-dominant hemisphere per patient.

LH, Left hemisphere; RH, right hemisphere; F, frontal cortex; M, motor cortex; T, temporal cortex; P, parietal cortex; O, occipital cortex.

Model performance was measured as the Spearman correlation between predicted and observed neural responses in the held-out test set. The correlation values were averaged across five cross-validation folds and were transformed to t values for determining significance (Kendall and Stuart, 1973). The correlation values reported here were significant at $p < 0.001$, Bonferroni corrected for the number of electrodes.

Retraining the encoding model with varying amounts of training data. We retrained the encoding model using smaller training datasets: 10%, 20%, 40%, 60%, and 80% of all available training dataset to investigate the sufficiency of the amount of training data in the encoding analysis. The model was retrained to predict the neural responses to speech. The prediction accuracy obtained using different amounts of training data was compared. The results for all 1283 electrodes are shown in Figure 3*A*. Next, the 130 electrodes with significant model performance in speech were selected. Their prediction accuracy obtained using different amounts of training data was compared using one-way ANOVA test ($F = 65.57$, $p = 4.1 \times 10^{-57}$). Tukey's honest significance difference criterion was used to correct for multiple comparisons. Mean prediction accuracy was significantly lower when using 10% training data compared with all other models ($p = 2.07 \times 10^{-8}$ for all comparisons). Mean prediction accuracy for models using 60%, 80%, and 100% of training data did not differ significantly: $p = 0.63$ (60% vs 80%), $p = 0.21$ (60% vs 80%), and $p = 0.98$ (80% vs 100%). The details are summarized in Figure 3*B*. The results of the model comparison suggested leveling out of the mean prediction accuracy over electrodes tuned to speech starting at 60% of training data toward using 100% of training data. Altogether, this indicates that the amount of training data in this study was sufficient for robust assessment of the speech encoding model.

Anterior–posterior gradient of model prediction accuracy. A one-way ANOVA test was used to determine the effect of the anatomical label of the electrode on the model prediction accuracy in speech. Five cortical regions along the IFG–STG axis were chosen: IFG pars triangularis and pars opercularis, precentral gyrus, postcentral gyrus, and STG. Temporal pole and pars orbitalis of IFG were not included because each only comprised one electrode with significant model performance. Each of the five chosen cortical regions comprised five or more electrodes. Across patients, 102 electrodes were included in the analysis. Given that the F statistic was significant at $p = 0.02$, the least significant difference test was applied to establish which cortical regions had significantly different mean prediction accuracy.

In addition, a linear regression was fit to model the prediction accuracy in speech per electrode (r_e) based on the electrode's anatomical label (Eq. 4). The five previously considered anatomical labels were replaced by integer values (from 1 to 5; l_e) based on the distance of the label from STG; therefore, for the electrodes in STG, $l_e = 1$, whereas for the electrodes in pars opercularis, $l_e = 5$, as follows:

$$r_e = \boldsymbol{\beta}_e{}^{\mathrm{T}}\mathbf{l}_e + \varepsilon_e \tag{4}$$

where $\mathbf{l}_e$ is a vector: $[1, l_e]^{\mathrm{T}}$.

An F statistic comparing the model fit using the five anatomical labels against the constant model was computed.

Distinct tuning profiles across cortical sites in ECoG

Clustering of regression coefficients. The learned regression coefficients were selected for the electrodes with significant model performance (130 electrodes for speech and 47 electrodes for music). Per electrode, the regression coefficients were z-scored over the features. To improve display of results the regression coefficients were also z-scored over the selected electrodes; however, comparable clustering results were obtained without z-scoring over the electrodes.

Affinity propagation clustering (Frey and Dueck, 2007) was applied to the regression coefficients across the electrodes to obtain clusters of electrodes with similar feature tuning profiles. The 3D regression coefficient matrices (SMs × TMs × time lags) were vectorized. Euclidian distance was used as a measure of similarity between resulting regression coefficient vectors. The grouping criterion, or similarity preference parameter (p), was set to one of the default estimates (Frey and Dueck, 2007) as follows:

$$p = \frac{5\left(\frac{\sum_{i=1}^{m}\sum_{j=1}^{m} D_{ij}}{m}\right)^2 + \min_{i,j} D_{ij}}{6} \tag{5}$$

where D is the Euclidian distance matrix for the regression coefficient vectors and m is the number of electrodes. These parameters produced better visualizations; however, various parameter settings were also tried. Affinity propagation clustering determined the optimal number of clusters automatically based on the similarity preference parameter (p).

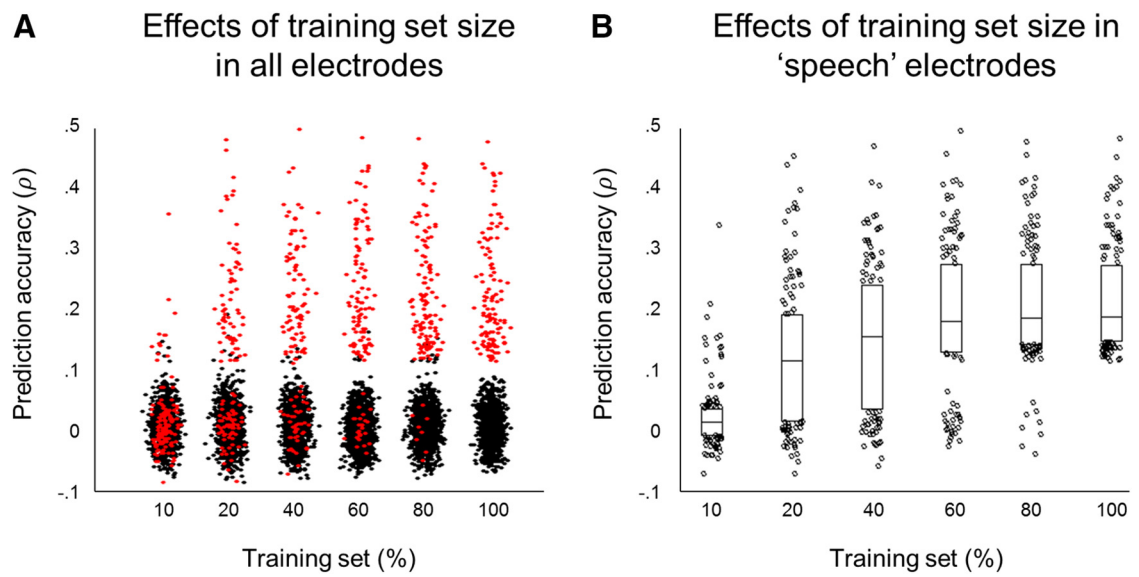

Figure 3. Effects of training size on model performance when tested on speech in ECoG. ***A***, Dot plot of prediction accuracy in all electrodes ($N = 1283$). Red dots indicate electrodes with significant model performance when trained on 100% training data and tested on speech (130 electrodes). Black plots are remaining electrodes. ***B***, Box plots of prediction accuracy in 130 electrodes with significant model performance when trained on 100% training data and tested on speech. The lower and upper boundaries of the box plots show 25 th and 75 th quantile, respectively. The horizontal bars show the median. The dots are the remaining accuracy values. One-way ANOVA results showed significant difference between results obtained with 10% training data and all other groups. No significant difference in prediction accuracy was revealed between results obtained with 60%, 80%, or 100% of all training data.

For subsequent analyses the three feature dimensions (SM, TM, and time lag) were separated and averaged across electrodes within each cluster. To obtain tuning profiles for each dimension, regression coefficients along the other two dimensions were averaged. Significance of the resulting tuning profiles was assessed by permutation testing. The procedure was performed on the regression coefficients averaged over all cluster members. The electrode-to-cluster assignments were permuted 10,000 times, resulting in a null distribution of tuning profiles for each feature along each feature dimension: 17 for SMs, 17 for TMs, and 63 for time lags. The original cluster tuning profiles were then evaluated in each feature relative to those null distributions, yielding exact statistical *p*-values. The cluster mean regression coefficients reported here were significant at $p < 0.01$, Bonferroni corrected for the number of features along each feature dimension.

Stepwise linear regression to relate tuning profiles to prediction accuracy

A stepwise linear regression as implemented in MATLAB 2016a (The MathWorks) was used to model prediction accuracy as a function of time lag, TM, and SM tuning. The regression was fit on 130 electrodes with significant model performance in speech. For each electrode maximal regression coefficients for time lag, TMs and SMs were calculated and used as predictors of the electrode's model prediction accuracy (r_e). The full interaction model was specified where apart from the individual predictors (lag, TMs, and SMs), products of all their combinations were added to the regression as follows:

$$r_e = \beta_e^{(0)} + \sum_{i=1}^{3} \beta_e^{(i)} w_e^{(i)} + \sum_{i=1}^{2} \sum_{\substack{j=2, \\ j \neq i}}^{3} \beta_e^{(ij)} w_e^{(i)} w_e^{(j)} + \beta_e^{(123)} \prod_{i=1}^{3} w_e^{(i)} + \varepsilon_e \quad (6)$$

where $\beta_e^{(0)}$ is the bias term and $w_e^{(i)}$ is the electrode's maximal regression coefficient for each of the features (i): time lag, TMs, and SMs.

Based on the analysis of the residuals, the algorithm determined which predictors contributed significantly to the model performance. An *F* statistic comparing the model fit using significant predictors against the constant model was computed. The estimated regression coefficients for the significant predictors are reported.

Retraining the encoding model with reduced lag representations

To further quantify the effects of time lag on the model performance and electrode-to-cluster assignment (six clusters in speech), the model was retrained using reduced lag representations and tested to predict neural responses to speech. In the first case, the model was retrained without using any lag information; in the second case, the model was retrained to predict the neural responses to speech only using sound features at a lag of −500 ms. The resulting prediction accuracy produced by varying input lag information (Lag0–500, Lag0, and Lag500) were compared using a two-way ANOVA test with unbalanced design (due to varying number of electrodes per cluster). The two main factors in the test were the time lag (Lag0–500, 0, or Lag500) and the cluster assignment (six speech clusters). The interaction between the main factors was also estimated. Given that the *F* statistic was significant at $p < 0.01$, the least significant difference test was applied to establish which models had significantly different mean prediction accuracy per cluster.

In addition, for each cluster, a one-way ANOVA test was performed to quantify the difference in the SM and TM tuning learned in all three cases. For Lag0–500, the regression coefficients were averaged over time lags. For each test, cluster average tuning profiles were used. Only electrodes with significant model performance in speech (130 electrodes) were considered.

Relation to language features

Annotation of the movie sound track with linguistic markers was done manually using Praat (Boersma and Weenink, 2016) (http://www.praat.org/). Five linguistic features were considered: voice onset time, formant transitions, and phoneme, syllable, and word boundaries. For each linguistic feature, we calculated the duration of each instance (d_i) measured in milliseconds and converted it to the rate value (r_i) in Hertz. The rate (r_i) was obtained by dividing 1000 over the duration (d_i).

FMRI experiment

FMRI subjects and setup. Eleven patients watched the same short movie during the presurgical fMRI recordings. The video was delivered on a screen through a scanner mirror and the audio was delivered through earphones.

Functional images were acquired on a Philips Achieva 3T MRI scanner using 3D-PRESTO (Neggers et al., 2008) (TR/TE = 22.5/33.2, time per

volume 608 ms, FA = 10°, 40 slices, FOV = 224 × 256 × 160 mm, voxel size 4 mm). Anatomical T1 images were acquired using TR/TE 8.4/3.2 ms, FA = 8°, 175 = slices, FOV = 228 × 228 × 175, voxel size of 1 × 1 × 2 mm.

FMRI data preprocessing. The functional data were corrected for motion artifacts and coregistered with the anatomical scan using SPM8 (Welcome Trust Centre for Neuroimaging, University College London). Four patients were discarded because they moved >4 mm in at least one of the directions. The data were despiked using ArtRepair (Mazaika et al., 2009) at 10% variance, detrended, and high-pass filtered at 0.009 Hz using mrTools (http://gru.stanford.edu/doku.php/mrTools/download). Only voxels in gray matter voxels were analyzed. No smoothing was applied before encoding or clustering analyses.

FMRI encoding model and clustering of regression coefficients. The 4D SM–TM feature set used in ECoG was adjusted to accommodate analysis with the lower sampling rate of fMRI (to 4 samples/s). Time lag was not included. Based on the previous findings (Santoro et al., 2014), the reduced frequency dimension was added to the feature set, resulting in 17 × 17 × 4 (SMs × TMs × frequency) features per time point. Each feature was convolved with the classical hemodynamic response function (spm_hrf) using default settings. For each patient, six motion regressors were added to the feature set. The model-fitting procedures were identical to the model fit on the ECoG neural responses. The split into training and test set was done in the same way except that no time lag information was used and therefore no truncation of chunks in sound-specific fragments was applied. No data points were shared between the train and test datasets in one cross-validation fold. No data were shared in the test datasets across the five cross-validation folds. Individual prediction accuracy was thresholded at $p < 0.05$, false discovery rate (FDR) corrected. For the display purposes, the individual prediction accuracy maps were smoothed at FHWH = 10 mm.

Voxels with similar tuning profiles were clustered using affinity propagation (Frey and Dueck, 2007) separately for speech and music. Various clustering parameters were used. Significance of the cluster-specific tuning profiles was assessed by permutation testing. The voxel-to-cluster assignments were permuted 10,000 times, resulting in a null distribution of tuning profiles for each feature along each feature dimension. The original cluster-tuning profiles were then evaluated in each feature relative to those null distributions, yielding exact statistical *p*-values. The cluster tuning profiles reported here were significant at $p < 0.01$, Bonferroni corrected. The brain maps with clustering results were created only for the purpose of visualization and no explicit group-based statistical analysis was performed. The histograms in Figure 7 show that the cluster map was not biased by one patient and included voxels from at least a half of all patients. To project the individual results on the standard MNI template, the individual maps (e.g., maps of clustering of regression coefficients) were normalized using the DARTEL toolbox (Ashburner, 2007) and an aggregated mask over patients was created in MNI space. For display purposes, the cluster maps were smoothed at FHWH = 10 mm.

Agreement between ECoG and fMRI speech encoding results. To estimate the voxels corresponding to each ECoG electrode, a sphere was defined in individual fMRI prediction accuracy maps using a 10 mm radius around each electrode's center coordinate. Voxels within that sphere (only gray matter) were associated with the corresponding electrode. No voxel was associated with more than one electrode. Pearson correlation coefficients were calculated between prediction accuracy in ECoG electrodes and corresponding fMRI voxels. The voxel with maximal correlation was determined per sphere. We used prediction accuracy values, uncorrected for significance, because we assumed that low prediction accuracy in both fMRI voxels and ECoG electrodes was indicative of a consistency of the results. Significance was assessed by permuting the prediction accuracy in ECoG and therefore disrupting the spatial association between ECoG electrodes and fMRI voxels. The permutations were conducted 10,000 times per each electrode–voxel pair and a null distribution of the Pearson correlation coefficients was obtained. The statistical *p*-values of the original Pearson correlation coefficients were determined and Bonferroni corrected for the number of electrodes per patient.

In addition, the correspondence between ECoG-based and fMRI-based feature tuning profiles, or regression coefficients, was quantified. Per patient, we correlated regression coefficients between electrodes with significant model performance and corresponding fMRI spheres using the Pearson correlation coefficient. The voxel with maximal correlation was determined per sphere. Significance was assessed by permuting the regression coefficients in ECoG 10,000 times and obtaining a null distribution of the Pearson correlation coefficients. The statistical *p*-values of the original Pearson correlation coefficients were determined and Bonferroni corrected for the number of electrodes per patient. Per patient, the Pearson correlation coefficients, significant at $p < 0.001$, Bonferroni corrected, were averaged over all the electrodes of the corresponding patient that were used in the analysis. The averaged Pearson correlation coefficient was weighted by the percentage of electrodes, shown to be significant in this analysis per patient.

The agreement between ECoG and fMRI encoding results was quantified for both speech and music; however, only speech results showed significant agreement.

Results

Segments from the feature film "Pippi Longstocking" were edited together to make a 6.5 min video with a coherent plot. The video was accompanied by a soundtrack that contained contrasting fragments of speech and music of 30 s each. This naturalistic stimulus was used in both ECoG and fMRI experiments of the present study. During both experiments, the participants were asked to attend to the short movie. The linear model using low-level sound features was trained on ECoG and fMRI data separately. Due to its high spatial sampling, fMRI results were used as complementary evidence of the cortical topography of model performance and feature tuning observed in ECoG.

We first explored the low-level acoustic properties of the soundtrack of the movie. The NSL toolbox (Chi et al., 2005) was used to obtain spectrotemporal modulations (SMs and TMs) of the sound based on the audio spectrogram (Fig. 1*A*). These modulations can intuitively be interpreted as representations of sound at different resolutions along the spectral and temporal dimensions. Finer modulations capture the sound at a higher resolution and thus reflect finer changes along the spectral and temporal dimensions. Coarser modulations capture the sound at a lower resolution and reflect smoothed energy distributions along both dimensions.

In the soundtrack of the movie used in the present study, low-level spectrotemporal features differed between speech and music fragments, as assessed with two-tail *t* tests run for each SM–TM feature ($p < 0.001$, Bonferroni corrected). Speech exhibited a larger spread along the TM dimension (2–8 Hz) than music (<2 Hz). Conversely, music exhibited a larger spread over the SM dimension (1–16 cyc/oct at TMs <2 Hz) compared with speech (<4 cyc/oct) (Fig. 1*B*,*C*).

Neural tuning to low-level features of speech and music in ECoG: encoding model performance

Fifteen patients watched the movie in an ECoG experiment. Previous studies have shown that the HFB amplitude closely correlates with neuronal firing rates and with fMRI blood-oxygenation-level-dependent response (Lachaux et al., 2007; Hermes et al., 2012). Here, we extracted the 60–120 Hz amplitude from the signal. A linear kernel ridge regression model was applied to predict neural responses from the sets of SMs (0.03–8 cyc/oct) and TMs (0.25–64 Hz). To account for time differences between neural responses and sound onset, the analysis included time lags between 0 and −500 ms, which constituted a third feature dimension. Per electrode, the model was trained on

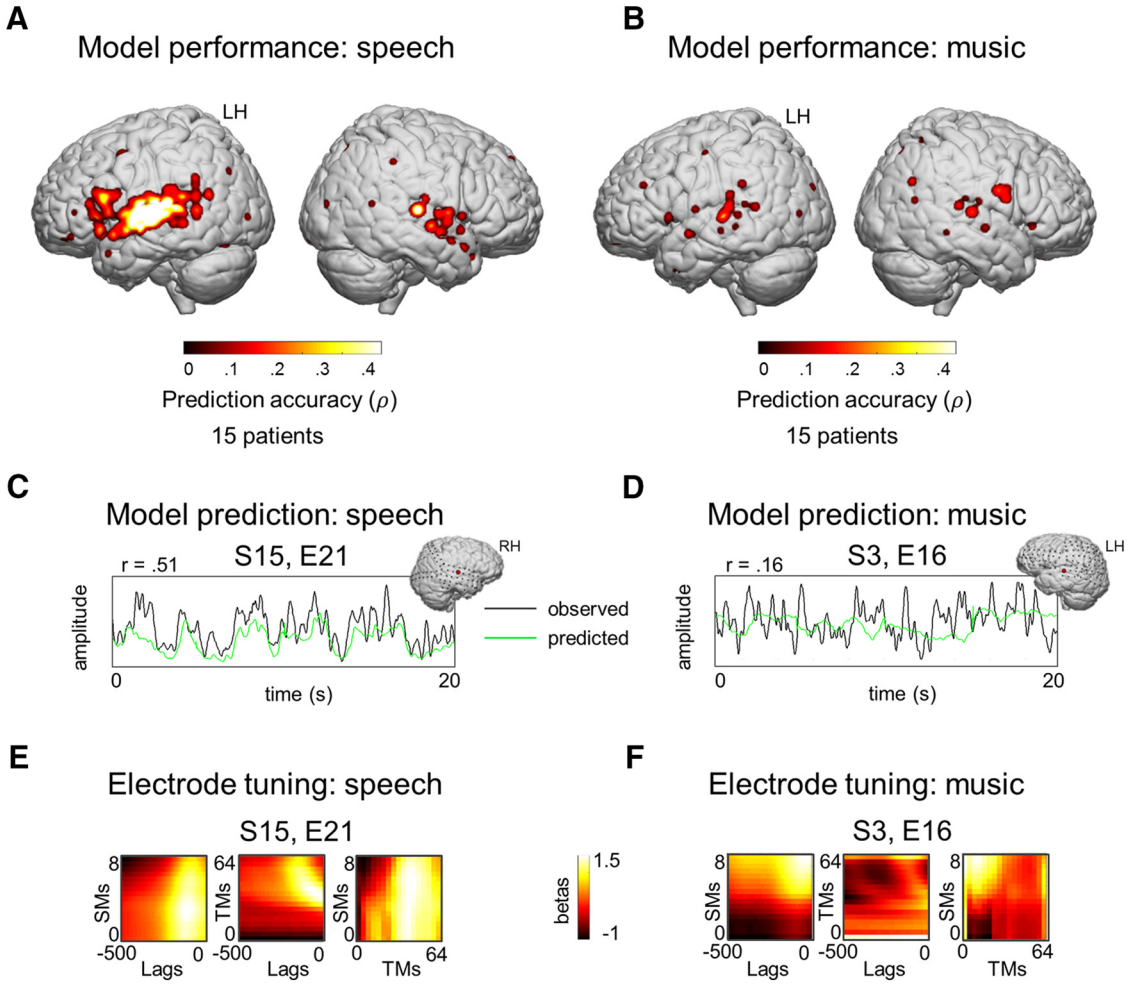

Figure 4. Encoding model performance in ECoG. A linear encoding model was trained on low-level SM–TM features to model neural responses in ECoG evoked by speech and music. The model was trained per individual electrode. The model performance was quantified in terms of prediction accuracy: the held-out neural responses (test set) were correlated to the corresponding time courses predicted by the model trained on low-level SM–TM features of the audio. The model performance assessment was performed separately for speech and music. Spearman correlation scores were used. The presented maps were thresholded at $p < 0.001$, Bonferroni corrected for the number of electrodes (1283 electrodes over 15 patients). For the visualization purposes, a 2D Gaussian kernel (FWHM = 10 mm) was applied to the coordinate on the brain surface corresponding to the center of the electrode, so that the prediction accuracy value obtained by the model faded out from the center of the electrode toward its borders. The results are shown on a standard MNI brain. Individual electrode locations were normalized to the MNI space using patient-specific affine transformation matrices obtained with SPM8. ***A***, Prediction accuracy maps for modeling neural responses to speech (130 electrodes). ***B***, Prediction accuracy maps for modeling neural responses to music (47 electrodes). ***C***, Example of a neural response predicted by the model (green) plotted against the observed response (black) in speech. ***D***, Example of predicted and observed neural responses in music. ***E*** and ***F*** are examples of electrode feature tuning profiles (***E*** shows electrode from ***C***, ***F*** shows electrode from ***D***). Non *z*-scored regression coefficients were used.

speech and music data combined and was tested to predict HFB responses to speech and music separately. All individual accuracy scores were thresholded at $p < 0.001$, Bonferroni corrected. The model predicted the neural responses evoked by speech in 10% of all electrodes (Fig. 4*A*). The neural responses evoked by music could be predicted in 3% of all electrodes (Fig. 4*B*). The accuracy scores varied considerably depending on the cortical region but, on average, across patients, the maximal accuracy score (r_{max}) reached 0.39 ± 0.08 when predicting neural responses to speech and 0.18 ± 0.05 when predicting neural responses to music. In general, modeling neural responses to speech and music produced high prediction accuracy for nonoverlapping sets of electrodes. Examples of observed and predicted neural responses are shown in Figure 4*C* for speech and in Figure 4*D* for music.

Low-level speech encoding: a posterior–anterior prediction accuracy gradient

In the case of speech, we observed a posterior–anterior prediction accuracy gradient along the perisylvian cortex: prediction accuracy values were highest in posterior STG and declined as the location of the electrode moved toward IFG (Fig. 4*A*). The effect was significant as tested with a one-way ANOVA comparing anatomical electrode location to prediction accuracy ($F_{(4)} = 3.25$, $p = 0.02$). Locations were confined to five cortical regions covering the anterior–posterior axis along the perisylvian cortex: IFG pars triangularis and pars opercularis, precentral gyrus, postcentral gyrus, and STG. Electrodes in STG exhibited higher accuracy compared with electrodes in pars opercularis and pars triangularis: $p = 5 \times 10^{-3}$ and $p = 0.01$, respectively, based on least significant difference test. The ANOVA test showed a difference in model performance between the cortical regions. To quantify the effect of the poster–anterior direction of the difference in model performance, we assigned numerical labels to the 5 cortical regions based on the ranked distance from STG (1–5) and fit a linear trend on the prediction accuracy along the five regions. The model fit was significant compared with the constant model: $\beta_{label} = 0.02$, $F_{(100)} = 12.9$, $p = 5.1 \times 10^{-4}$ and $\beta_{intercept} = 0.16$ ($p = 3.5 \times 10^{-8}$).

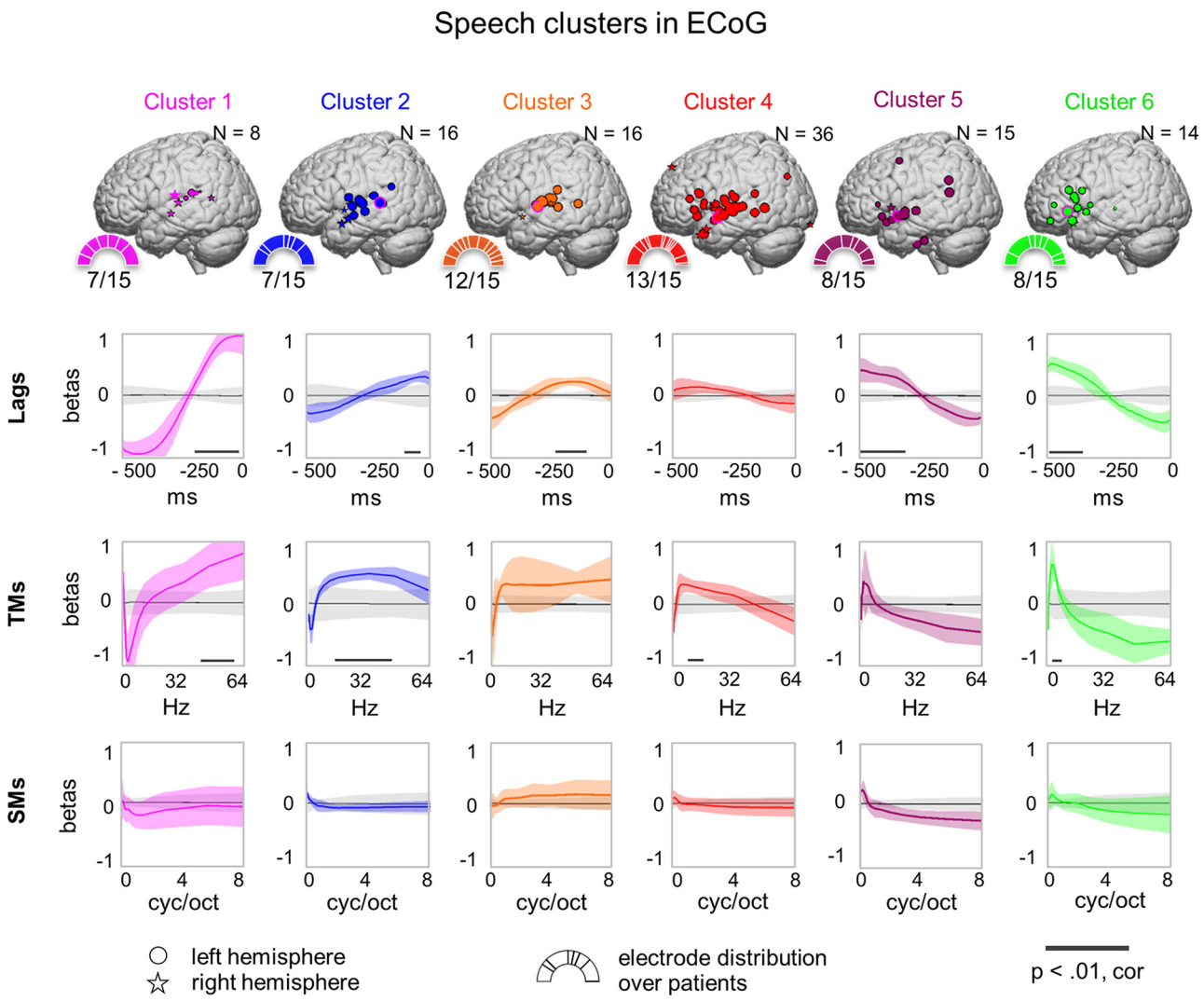

Figure 5. Affinity propagation clustering of regression coefficients in ECoG electrodes tuned to speech. Shown are six clusters produced by clustering on regression coefficients of electrodes involved in speech tuning. For each cluster, a number of properties are shown: the number of electrodes belonging to the cluster (*N*), the normalized electrode locations on the MNI brain, the distribution of electrodes over patients (half pie charts with number of patients out of 15), and the feature tuning profiles over lags, TMs, and SMs. Right hemisphere electrodes were projected on the left hemisphere and marked with star symbols; circles show left hemisphere electrodes. The size of electrodes on the MNI brain reflects the similarity of the electrode's feature tuning profile to the exemplar electrode of the cluster (framed in pink). The sectors of the half pie charts show percentage of each patient's electrodes relative to the overall number of electrodes in the cluster (*N*). The lags were used in the range of 0 to −500 ms with respect to audio onset; TMs were in the range of 0.25–64 Hz; SMs were in the range of 0.03–8 cyc/oct. Per cluster, mean (colored bold curve) and SD (colored shading) over regression coefficients are plotted for each dimension. The mean (gray bold curve) and SD (gray shading) of the null distribution based on permutations of cluster assignments are shown. The black thick line at the bottom of several plots represents significant segments of the tuning profile compared with the null distribution ($p < 0.01$, Bonferroni corrected).

Distinct tuning profiles across cortical sites in ECoG

Next, we explored whether different cortical sites, where responses were well predicted by the model, exhibited distinct feature tuning profiles. Examples of feature tuning profiles are shown in Figure 4*E* for speech and in Figure 4*F* for music. We aimed at uncovering brain sites with similar feature tuning in an unsupervised fashion to avoid superimposing any assumptions about the topography of the tuning profiles. Therefore, an affinity propagation clustering (Frey and Dueck, 2007) analysis was performed on the regression coefficients and *z*-scored over the electrodes. The electrodes with similar regression coefficients were clustered together as having a similar feature tuning profile. The clustering analyses were performed separately for speech and music because modeling neural responses to speech and music resulted in different sets of electrodes. Permutation testing was used to assess the ranges over the three feature dimensions (TMs, SMs, and time lags) to which the clusters were significantly tuned.

Neural tuning to temporal and spectral characteristics of speech

In the case of speech, the affinity propagation clustering returned six clusters (clusters 1–6), each of which contained electrodes from about half of all patients. Altogether, these clusters contained 81% (105 electrodes) of all electrodes used in the clustering analysis (130 electrodes). Of the remaining clusters, seven clusters only comprised electrodes from one patient, and two clusters comprised electrodes from three patients (altogether, 25 electrodes).

Across six speech clusters, most variance was concentrated along the dimensions of TMs and time lags (Fig. 5). Electrodes in cluster 2 exhibited tuning to fast TMs (11–45 Hz) within a 150 ms window after the sound onset. Conversely, electrodes in cluster 6 were tuned to slow TMs (0.7–4 Hz) within a 315–500 ms window after the sound onset. Clusters 1–6 did not exhibit distinct SM tuning, but showed some preference to coarse SMs (<1 cyc/oct).

From the anatomical point of view, clusters 1–3 comprised electrodes primarily in posterior STG, whereas clusters 4–6 comprised electrodes in IFG and anterior STG. Clusters 2–5 comprised a small number of electrodes in supramarginal gyrus.

With respect to the negative regression coefficients (Fig. 5), some electrodes showed negative tuning to certain features before any *z*-scoring procedures. We believe that the negative tuning indicates a decrease in the HFB amplitude of an electrode relative to its mean HFB amplitude as a response to the stimulus (estimated by the bias term).

In addition, we looked at the average tuning profiles per anatomical region along the perisylvian cortex rather than the clusters. The tuning plots per anatomical region corresponded to the ones recovered with the clustering (Fig. 5), but exhibited more variance in tuning along each feature dimension. Moreover, averaging the regression coefficients over the whole STG did not result in any specific tuning, whereas using the clustering approach allowed us to uncover a number of functional subdivisions along STG (clusters 1 - 5).

No hemisphere-specific tuning profile was found across the patients. In general, the electrodes located similarly across the hemispheres were assigned to the same cluster.

Noticeably, clusters 2–3 and clusters 4–5 showed similar cortical topography with varying tuning profiles. These local variations could be due to the anatomical variability among the patients. Overall, the change in tuning to TMs and time lags followed a global trend: from tuning to shorter lags and faster TMs in posterior STG toward tuning to larger lags and slower TMs in IFG and anterior STG. The change in tuning to TMs and time lags mapped on the posterior-anterior prediction accuracy gradient as estimated with a stepwise linear regression: $\beta_{\text{lag}} = 1.33 \times 10^{-4}$, $\beta_{\text{TM}} = 9.67 \times 10^{-4}$, $F_{(127)} = 13.3$, $p = 5.92 \times 10^{-6}$.

We further explored the effect of time lag on the model performance across different clusters by retraining the model with reduced lag feature sets and comparing the model performances. In addition to using audio features at all lags from 0 to −500 ms (Lag0–500), we retrained the model using no input lag information (Lag0) or using only audio features at a lag of −500 ms (Lag500). Interaction between lag information and assignment to clusters 1–6 was tested using a two-way ANOVA with unbalanced design: $F_{(10)} = 10.32$, $p = 6.06 \times 10^{-15}$ (Fig. 6). Prediction accuracy in cluster 1 was higher for Lag0 and Lag0–500 compared with Lag500: $p = 4.64 \times 10^{-6}$ and $p = 5.32 \times 10^{-6}$, respectively. Prediction accuracy in cluster 6 was higher for Lag500 and Lag0–500 compared with Lag0: $p = 1.7 \times 10^{-3}$ and $p = 3.7 \times 10^{-3}$, respectively. In the case of Lag0, more electrodes were modeled in clusters 1–3 compared with Lag500; the opposite was observed for clusters 4–6. Despite the lag difference, in all cases, the encoding model learned the same tuning to SMs and TMs in all clusters as quantified with multiple one-way ANOVA tests on cluster mean regression coefficients, separately for TMs and SMs (TMs: $0.1 \leq F_{(2)} \leq 0.52$, SMs: $0.77 \leq F_{(2)} \leq 4.27$, both at $p > 0.01$).

Link between low-level TMs of speech and rates of language features

To obtain a more intuitive interpretation of the TM tuning results, we compared the TM tuning profiles of the different clusters with the temporal rates of well known language features. We estimated the temporal rates (see Materials Methods) for a number of critical language features from the acoustic data (Fig. 7*A*). We considered the rates of three features that represent important building blocks of language: phonemes, syllables, and words.

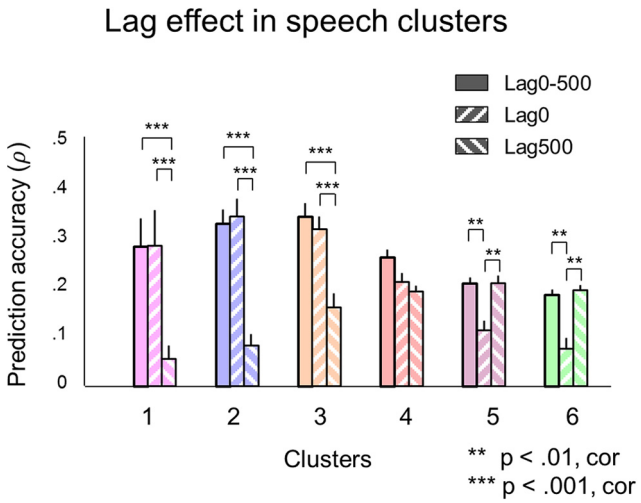

Figure 6. Lag effects in encoding of speech in ECoG responses. In addition to the full model using all lags from 0 to −500 ms of the audio features to predict ECoG responses (Lag0–500), two more models with reduced lag representation were trained. The first additional model included no lags at all (Lag0) and the second additional model only included one lag at −500 ms (Lag500). The bars show mean prediction accuracy over all electrodes in the cluster. The error bars indicate SEM. A two-way ANOVA test with unbalanced design was performed to estimate the main effect of the lag factor and the interaction between the lag factor and cluster assignment (on 105 electrodes). Groups with significantly different means are shown at $p < 0.01$ and $p < 0.001$. The *p*-values were corrected for multiple comparisons using Fisher's least significance difference under the condition that the null hypothesis had been rejected: $F_{(10)} = 10.32$, $p = 6.06 \times 10^{-15}$.

We also considered two features at the scale smaller than a phoneme: voice onset time and formant transitions. These two features are crucial for the categorization of a large number of phonemes and are therefore central to speech comprehension (Liberman et al., 1967; Ganong, 1980; Mesgarani et al., 2014). Voice onset time is a characteristic of plosive consonants and represents the amount of time between the release and the voice onset of the subsequent vowel. Based on this feature, voiced plosives, which are associated with shorter voice onset time, can be distinguished from voiceless plosives, which are associated with longer voice onset time. Formant transitions are characteristics of vowels and are indicative of the place of articulation, and thus the category, of the neighboring phonemes. For all of these features, we calculated their temporal rates to determine whether there might be a relationship between the language features and low-level TM tuning in the observed ECoG responses.

On average, the language markers of phoneme and syllable boundaries showed moderate correspondence to TM tuning of cluster 2 and cluster 4, respectively (Fig. 7*B*). Subphonemic features such as voice onset time and formant transitions changed at a temporal rate similar to the TM tuning of clusters 1 and 2 (Fig. 7*B*). The TM tuning profile of cluster 6 seemed to overlap to some extent with the temporal rates of word boundaries. However, it appeared that our naturalistic speech stimuli contained a lot of monosyllabic words and the temporal rate of word boundaries was similar to the temporal rate of syllable boundaries. Cluster 6 showed tuning to potentially slower TMs than the rate of individual word transitions.

Neural tuning to temporal and spectral characteristics of music

Analogous affinity propagation clustering of regression coefficients was performed for music encoding. Two clusters with opposite tuning profiles were revealed (Fig. 8). Distinct SM and TM temporal profiles were uncovered; however, no specific time lag

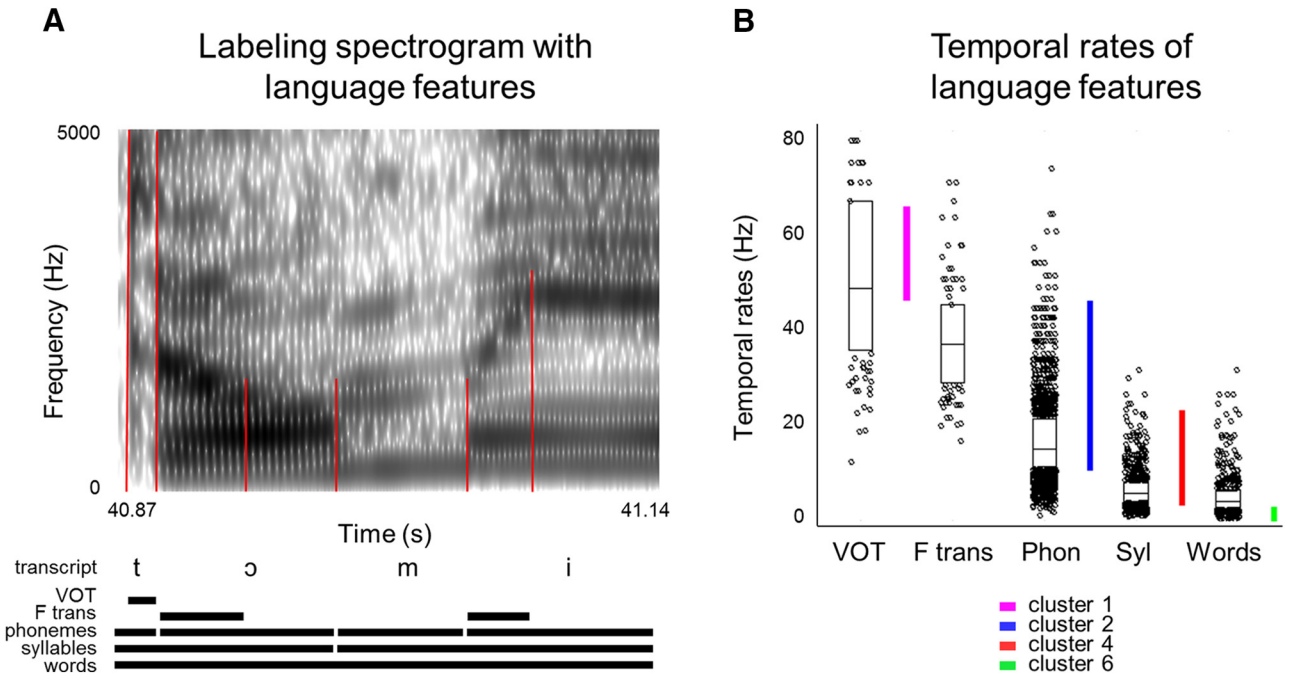

Figure 7. Rates of various language features. ***A***, Details of annotating the spectrogram with the language features. Based on the spectrogram, we identified the boundaries for formant transitions, voice onset time, phonemes, syllables, and words. Red lines outline the acoustic features based on which the annotations were made; for example, here: noise before the voicing of the vowel, changes in the second formant of the vowel, and boundaries between different formant structures of the vowels and a sonar consonant. ***B***, Distributions of rates for each language unit in Hertz: VOT (voice onset time), F trans (formant transitions), Phon (phonemes), Syl (syllables), and Words. The rates were calculated based on the duration of each instance of each language feature. The durations were then expressed in Hertz. The lower and upper boundaries of the box plots show 25th and 75th quantile, respectively. The horizontal bars show the median. The dots are the remaining rate values. Colored vertical bars show TM tuning profiles for clusters 2, 4, and 6.

tuning was found. Cluster 3 included electrodes without specific tuning to any of the low-level features (Fig. 8). No hemisphere-specific tuning profile was found across the patients. The reported three clusters comprised 100% of all electrodes used in the clustering analysis in music (47 electrodes).

We tested to what extent the apparent difference in model performance between speech and music could be attributed to our choice in the range of low-level spectrotemporal features. Given that music exhibited a larger spread over the dimension of SMs, the model was retrained to predict neural responses to music on another set of SM features spanning from 0.5 to 32 cyc/oct keeping the TMs fixed (0.25–64 Hz). The resulting prediction accuracy scores did not differ significantly from the prediction accuracy scores obtained with the previously used set of features based on one-tailed t test: $t = 0.74$, $p = 0.46$. Therefore, the limited range of SMs (0.03–8 cyc/oct) chosen for the encoding analyses did not explain why neural encoding of music was less prominent than that of speech.

Complementary findings with fMRI

Seven of the 15 patients whose ECoG results are described above watched the same movie in the MRI scanner. Analogous analyses were performed on the obtained fMRI data.

A linear kernel ridge regression model was trained to predict voxel responses from SMs, TMs, and audio frequency. Time lag was not included due to the slow temporal resolution of fMRI. Across patients, responses in STG were modeled best. On average across patients, the maximal accuracy (r_{max}) reached 0.54 ± 0.1 for speech encoding, and $r_{max} = 0.48 \pm 0.08$ for music encoding (all at $p < 0.05$, FDR corrected). Distinct feature tuning profiles were estimated using affinity propagation clustering. The clustering was performed separately for speech and music, but in both cases, two clusters were revealed. Cluster 1 comprised voxels throughout STG tuned to a broad range of TMs (3–45 Hz), coarse SMs (<0.25 cyc/oct), and high frequencies (>3 kHz) (Fig. 9*B*). Cluster 2 comprised voxels in the prefrontal, parietal, and latero-occipital cortices and did not exhibit distinct tuning profiles along any of the dimensions.

Finally, the anatomic similarity between ECoG and fMRI speech encoding results was quantified. For each individual ECoG electrode, a corresponding sphere of voxels was defined within a radius of 10 mm around the electrode center coordinate. Five of seven subjects showed significant correlation between prediction accuracy in ECoG electrodes and corresponding fMRI spheres: $r_{(1)} = 0.35$ ($p = 4 \times 10^{-3}$), $r_{(3)} = 0.44$ ($p = 5 \times 10^{-4}$), $r_{(5)} = 0.35$ ($p = 0.02$), $r_{(6)} = 0.53$ ($p = 3 \times 10^{-4}$), and $r_{(7)} = 0.36$ ($p = 5 \times 10^{-3}$). Figure 9*A* shows the overlap in model performance when using ECoG and fMRI data for one representative patient. The most similarity in prediction accuracy between ECoG and fMRI results was observed along STG.

In addition, we assessed the similarity of feature tuning profiles in ECoG electrodes and corresponding fMRI spheres. We first considered speech encoding results. The SM–TM tuning profiles uncovered with ECoG and fMRI were correlated (Fig. 9*C*,*D*). In six of seven patients ECoG-based and fMRI-based feature tuning profiles were significantly correlated (Fig. 9*D*): mean $r_{(1)} = 0.57$ (100% of electrodes), mean $r_{(3)} = 0.56$ (80% of electrodes), mean $r_{(4)} = 0.16$ (40% of electrodes), mean $r_{(5)} = 0.23$ (50% of electrodes), mean $r_{(6)} = 0.41$ (60% of electrodes), and mean $r_{(7)} = 0.38$ (80% of electrodes). Figure 9*C* shows some examples of similarity of tuning profiles in ECoG electrodes and corresponding fMRI spheres. The overall contour of the feature tuning profile was reproduced in the fMRI results; however, the gradient of the TM tuning across the perisylvian cortex was not

observed in the fMRI data. The latter is likely due to low temporal resolution of fMRI combined with the use of continuous, naturalistic speech stimuli. The most similarity in feature tuning profiles between ECoG and fMRI results was observed along STG. Analogous analyses were performed on the results of music encoding; however, for music, less than half of patients showed agreement between ECoG and fMRI.

Discussion

Our results show that a linear model using low-level sound features can predict the neural responses to naturalistic speech in the perisylvian regions. This result was observed with both ECoG and fMRI. Exploiting the fine-grained temporal resolution of ECoG, we also found a gradient of tuning to the temporal characteristics of speech spreading from posterior STG to IFG. Moreover, IFG was tuned to coarse acoustic features of speech sounds at a temporal rate close to or slower than the rate of word transitions.

Information propagation in speech perception: ventral pathway from posterior STG to IFG

Low-level feature tuning along the perisylvian cortex was observed in ECoG and reproduced using fMRI. Therefore, we confirmed that sparse spatial coverage of ECoG did not bias the encoding results. The present fMRI results were obtained using continuous naturalistic stimuli and resembled the tuning patterns reported previously with more controlled stimuli (Norman-Haignere et al., 2015). In ECoG, the neural responses in posterior STG were modeled better compared with the more anterior associative cortices such as IFG. The prediction accuracy dropped gradually as the electrode location moved anteriorly and away from the posterior STG. Regional differences in the prediction accuracy were associated with tuning to different time lags, with STG being tuned to features at short time lags (<150 ms) after the sound onset and IFG being tuned to longer time lags (up to 500 ms).

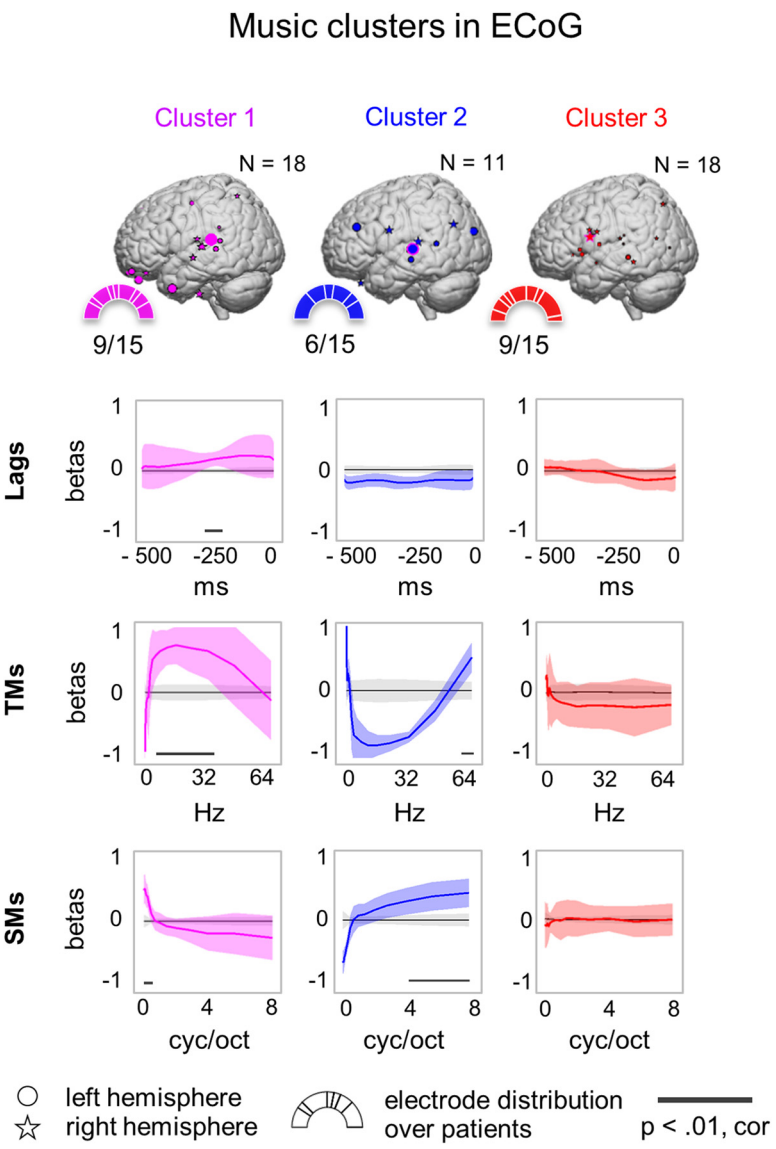

Figure 8. Affinity propagation clustering of regression coefficients in ECoG electrodes tuned to music. Clustering setup, significance testing, and the display format are identical to Figure 5. However, in the case of music encoding, a different set of electrodes was significantly well modeled based on low-level sound features (47 electrodes). The choice of cluster colors is based on convenience; there is no relationship between speech and music clusters represented by the same color.

Multiple electrophysiological studies have reported a temporal evolution of the activation from posterior STG to IFG during sentence comprehension (Halgren et al., 2002; Brennan and Pylkkänen, 2012; Kubanek et al., 2013). A recent study reported propagation of acoustic features of speech along the STG (Hullett et al., 2016). The evidence presented here is consistent with these findings. However, it also shows that speech-specific low-level auditory features propagate from early sensory areas to distal associative cortex as far as the IFG. Evidence for propagation of low-level sensory features in the brain has been reported previously for both visual and auditory modalities (Fontolan et al., 2014; Bastos et al., 2015). In visual perception, information is transferred from early visual processing areas toward the inferior temporal cortex, where high-level object representations (such as object categories) arise (Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and van Gerven, 2015). An analogous representational gradient from early processing areas (posterior STG) toward higher-level cortex (IFG) has been reported for auditory perception using neural responses to musical pieces containing lyrics (Güçlü et al., 2016). Our results suggest that a similar form of signal transfer might take place in speech perception.

The present findings are connected to the research on language comprehension along the perisylvian cortex through a dual-stream theory of language processing (Rauschecker, 1998; Hickok and Poeppel, 2004). In particular, these findings are consistent with the view that there is a ventral pathway of language processing in the brain (Rauschecker and Tian, 2000; Hickok and Poeppel, 2004; Saur et al., 2008). Along this pathway, the information is passed from posterior STG to anterior STG and IFG. The ventral stream is thought to accommodate transition of sound to meaning.

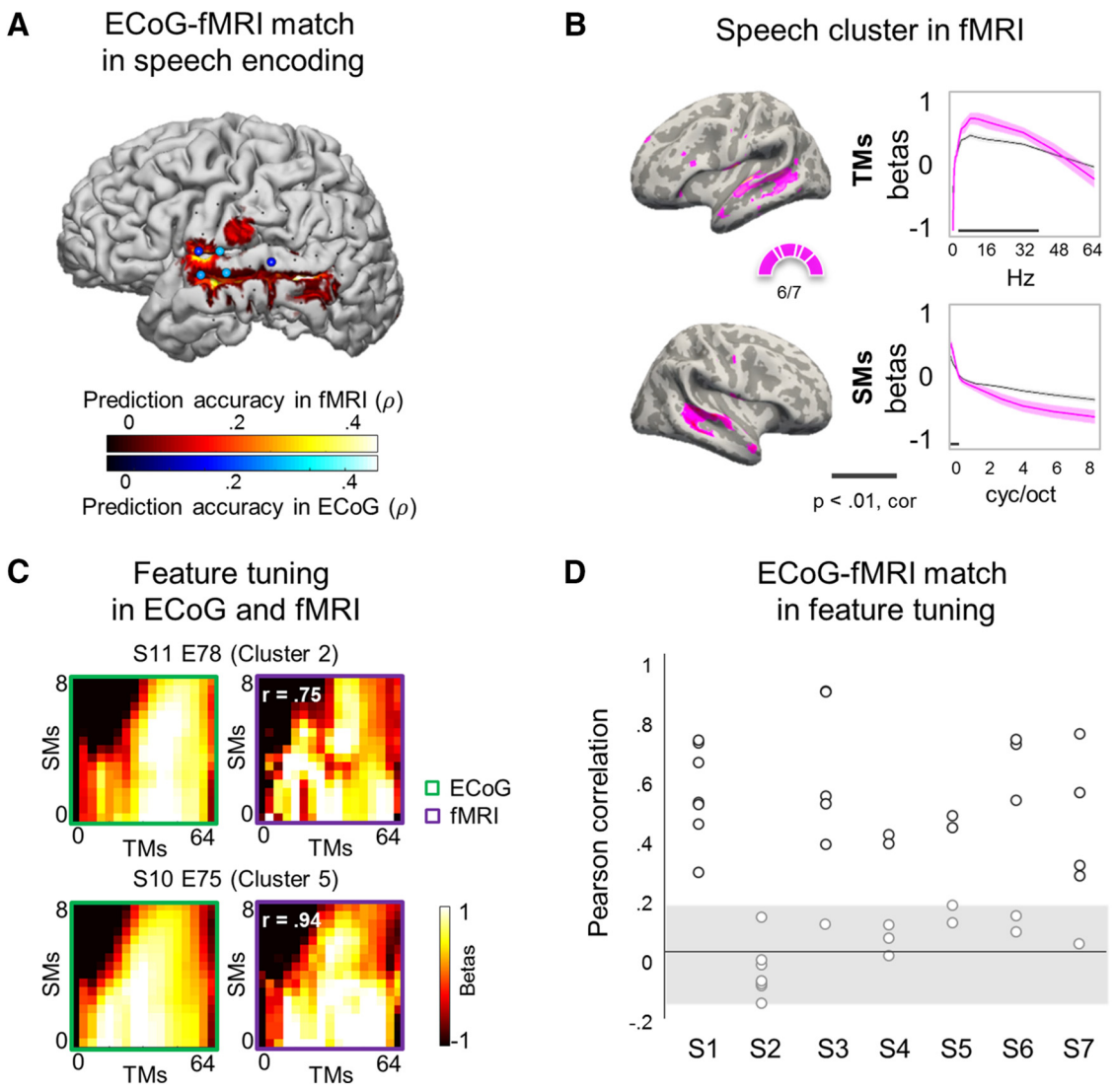

Figure 9. Agreement between ECoG and fMRI in tuning to low-level features of speech. ***A***, Prediction accuracy for modeling neural responses to speech with ECoG and fMRI for one representative patient. The map of electrodes tuned to low-level speech features ($p < 0.001$, Bonferroni corrected) is overlaid on top of the map of the voxels tuned to low-level speech features ($p < 0.05$, FDR corrected). For display purposes, the fMRI map was smoothed at FWHM = 10 mm. The results are shown in the native space of one patient's brain. ***B***, Speech cluster in fMRI matching ECoG speech clusters 1–6. A similar affinity propagation clustering analysis was performed on the regression coefficients learned from the fMRI data. The black line at the bottom of the plots shows significant segments of the tuning profile compared with the null distribution ($p < 0.01$, Bonferroni corrected). The displayed cluster showed a feature tuning profile along the perisylvian cortex similar to speech clusters 1–6 in ECoG. The fMRI speech cluster was tuned to fast TMs and coarse SMs. ***C***, Examples of SM–TM feature tuning profiles observed in ECoG electrodes and corresponding fMRI spheres. In each patient, for every ECoG electrode, a corresponding fMRI sphere was calculated within a radius of 10 mm. The SM–TM feature tuning profile of the best modeled voxel within the sphere was correlated with the feature tuning profile of the corresponding electrode. Pearson correlation between the ECoG and fMRI-based SM–TM profiles is reported per example. ***D***, Pearson correlation between SM–TM feature tuning profiles in ECoG electrodes and corresponding fMRI spheres per patient. Only electrodes and voxels with significant model performance in speech were used. The null distribution based on 10,000 permutation tests is shown in gray (0.1 and 99.9 percentiles). Significant correlation ($p < 0.001$, Bonferroni corrected) have black outlines; insignificant correlations are greyed out. Per patient, we report the number of electrodes with significant correlations compared with the number of electrodes used in the analysis: $N_1 = 7$ (of 7), $N_3 = 5$ (of 6), $N_4 = 2$ (of 5), $N_5 = 2$ (of 4), $N_6 = 3$ (of 5), $N_7 = 4$ (of 5).

Tuning to low-level features of speech along the perisylvian cortex: temporal encoding gradient

It has so far been challenging to uncover how exactly the sound-to-meaning mapping is implemented in the brain. Here, we investigated which specific features in the auditory input were most relevant for modeling the neural responses in different brain areas along the perisylvian cortex. In an effort to interpret the tuning preferences in different cortical regions, we explored how the low-level TMs mapped onto the temporal rates of various language units.

Multiple previous studies have shown a relationship between the neural responses in posterior STG and the spectrotemporal features of sound (Schönwiesner and Zatorre, 2009; Pasley et al., 2012; Martin et al., 2014; Santoro et al., 2014; Hullett et al., 2016). Here, with both ECoG and fMRI, we found that the cortical sites along the pathway from posterior STG to IFG showed tuning to coarse SMs (<1 cyc/oct). Coarse SMs represent a spread of energy over a broad range of frequencies (Chi et al., 2005). Such modulations are characteristic of formants and their transitions (Elliott and Theunissen, 2009). Using behavioral experiments (Drullman, 1995; Fu and Shannon, 2000; Elliott and Theunissen, 2009), it was shown that coarse SMs (<1 cyc/kHz) are crucial for speech intelligibility. In our audio data, speech also exhibited more energy along the dimension of TMs compared with SMs,

where the energy concentrated in the range of 0.25–4 cyc/oct (Fig. 1*B*). Notably, in the case of modeling neural responses to music, we observed tuning to both fine and coarse SMs.

TMs capture the dynamics in the audio signal. The energy along a specific frequency range can decline rapidly over time, and thus occur at a fast temporal rate, or last for longer periods of time, and thus occur at a slower temporal rate (Chi et al., 2005). The encoding of the temporal organization of speech in the ECoG responses showed distinct profiles along the ventral pathway. The TM tuning profiles in STG were consistent with the previous findings (Hullett et al., 2016). At the same time, the present results extend more posteriorly (posterior STG: TMs > 10 Hz) and more anteriorly toward IFG. We show that IFG is tuned to even slower TMs compared with anterior STG (0.5–4 Hz). Based on perceptual experiments, a number of previous studies have linked speech intelligibility to TMs in the range of 0.25–30 Hz (Drullman et al., 1994; Arai et al., 1999; Elliott and Theunissen, 2009). Specifically, various subphonemic features such as formant transitions and voice onset time have been linked to TMs in the 30–50 Hz range (Giraud and Poeppel, 2012), phoneme identification has been linked to TMs of >16 Hz (Drullman et al., 1994), and syllable identification has been linked to TMs at 4 Hz (Houtgast and Steeneken, 1985; Arai et al., 1999). Neural tracking of sentence-specific (1 Hz) and syllable-specific (4 Hz) TMs have been shown using highly controlled synthesized speech stimuli (Ding and Simon, 2012). Here, we show that a correspondence between the stimulus TMs and neuronal response preferences can also be observed during naturalistic speech comprehension. Tuning to the TMs of voice onset time, formant transitions, and phoneme, syllable, and word boundaries varied in an orderly fashion along the pathway from posterior STG to IFG.

Role of IFG in continuous speech comprehension

The present results conform to the previously denoted roles of STG and IFG in language processing. Posterior STG is tuned to fast TMs (>10 Hz) and short time lags (<150 ms), which allows efficient processing of fine changes in the auditory input and accommodates phonological processing and phoneme identification (Dehaene-Lambertz et al., 2000; Formisano et al., 2008; Chang et al., 2010). As the information about the input signal passes through the ventral stream of language processing, it becomes represented at coarser temporal rates and the sound representation exhibits longer time lags. Eventually, IFG processes the incoming speech signal at a 300–500 ms delay compared with the sound onset. Here, the signal is spread spectrally over >1 octave and temporally over 250–1000 ms. A comparison of the TM tuning of IFG to the temporal rate of individual word transitions (Fig. 7*B*) suggests that this region may be capturing speech-related information integrated over several words.

A subregion of IFG that exhibited tuning to coarse features of speech was also engaged in a different language task (verb generation, same patients). Given that the same neuronal populations appear to be involved in both maintaining coarse representations of the incoming speech and a verb generation task, it becomes more challenging to agree on one specific function of IFG. One influential theory, which attempts to integrate over multiple findings regarding the function of IFG, is the memory, unification, and control theory (Hagoort, 2005, 2013). According to this theory, IFG accommodates the general process of unification or the combining of the building blocks of language information into larger structures (Hagoort, 2013). With respect to the present results, during speech perception, IFG appears to be involved in the integration of incoming speech information over windows of 500 ms and the maintenance of a more "abstract," integrated representation of the input that arises in neuronal populations of IFG 300–500 ms after the sound onset.

Limitations and future directions

The present work has a number of limitations. For instance, the results obtained using ECoG data should be interpreted with care because ECoG is an invasive method that only registers neural responses from the cortical surface and not from deep and folded brain regions.

We studied the neural responses to speech using fragments of a feature film. This experimental setup allowed us to obtain neural responses to speech in naturalistic circumstances without introducing experimental artifacts and explore the data using bottom-up approaches. In the case of fMRI, we recovered the general contour of the low-level feature tuning shown in related work (Schönwiesner and Zatorre, 2009; Santoro et al., 2014; Norman-Haignere et al., 2015). However, using continuous naturalistic stimuli and a task-free experimental paradigm did not allow us to uncover the fine-grained tuning profiles in STG reported in these studies. In addition, the adopted paradigm made it unclear to what extent the observed pattern of neural responses was explained by attention shifts, multisensory integration, and other top-down processes such as the McGurk effect, which represents interaction between visual and auditory speech modalities (McGurk and MacDonald, 1976). It is conceivable that our use of a dubbed soundtrack affected the latency of the ECoG responses to speech. However, we believe that it is unlikely to have influenced the specific low-level tuning of different cortical regions or their lag difference relative to each other because the observed tuning profiles are consistent with previous reports on auditory processing (Hullett et al., 2016).

One of the possible future directions of the present work could be to investigate the feedback connections during continuous speech comprehension. Specifically, it will be interesting to investigate how our results relate to the predictive coding account of speech perception (Arnal and Giraud, 2012; Hickok, 2012). To address this issue, we would investigate whether the neural responses contain an indication about the features of upcoming speech and take into account oscillatory patterns in lower frequency bands.

Conclusions

In the present study, we investigated neural tuning to low-level acoustic features of speech sounds while attending to a short movie. The present findings support the notion of low-level sound information propagating along the speech perception pathway. In particular, increasingly coarse temporal characteristics of speech are encoded along the pathway from sensory (STG) areas toward anterior associative cortex (IFG). Finally, we observed that the sound features reflecting markers of words and syllable transitions travel as far as the IFG, which is associated with language processing.

References

Altmann CF, Gomes de Oliveira Júnior C, Heinemann L, Kaiser J (2010) Processing of spectral and amplitude envelope of animal vocalizations in the human auditory cortex. Neuropsychologia 48:2824–2832. CrossRef Medline

Arai T, Pavel M, Hermansky H, Avendano C (1999) Syllable intelligibility for temporally filtered LPC cepstral trajectories. J Acoust Soc Am 105:2783–2791. CrossRef Medline

Arnal LH, Giraud AL (2012) Cortical oscillations and sensory predictions. Trends Cogn Sci 16:390–398. CrossRef Medline

Ashburner J (2007) A fast diffeomorphic image registration algorithm. Neuroimage 38:95–113. CrossRef Medline

Bastos AM, Vezoli J, Bosman CA, Schoffelen JM, Oostenveld R, Dowdall JR, De Weerd P, Kennedy H, Fries P (2015) Visual areas exert feedforward and feedback influences through distinct frequency channels. Neuron 85:390–401. CrossRef Medline

Bilecen D, Scheffler K, Schmid N, Tschopp K, Seelig J (1998) Tonotopic organization of the human auditory cortex as detected by BOLD-FMRI. Hear Res 126:19–27. CrossRef Medline

Boersma P, Weenink D (2016) Praat: doing phonetics by computer [computer program], Version 6.0. 14.

Brennan J, Pylkkänen L (2012) The time-course and spatial distribution of brain activity associated with sentence processing. Neuroimage 60:1139–1148. CrossRef Medline

Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT (2010) Categorical speech representation in human superior temporal gyrus. Nat Neurosci 13:1428–1432. CrossRef Medline

Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Am 118:887–906. CrossRef Medline

Dehaene-Lambertz G, Dupoux E, Gout A (2000) Electrophysiological correlates of phonological processing: a cross-linguistic study. J Cogn Neurosci 12:635–647. CrossRef Medline

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31:968–980. CrossRef Medline

Ding N, Simon JZ (2012) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 107:78–89. CrossRef Medline

Ding N, Melloni L, Zhang H, Tian X, Poeppel D (2016) Cortical tracking of hierarchical linguistic structures in connected speech. Nat Neurosci 19:158–164. CrossRef Medline

Drullman R (1995) Temporal envelope and fine structure cues for speech intelligibility. J Acoust Soc Am 97:585–592. CrossRef Medline

Drullman R, Festen JM, Plomp R (1994) Effect of reducing slow temporal modulations on speech reception. J Acoust Soc Am 95:2670–2680. CrossRef Medline

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

Fontolan L, Morillon B, Liegeois-Chauvel C, Giraud AL (2014) The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. Nat Commun 5:4694. CrossRef Medline

Formisano E, Kim DS, Di Salle F, van de Moortele PF, Ugurbil K, Goebel R (2003) Mirror-symmetric tonotopic maps in human primary auditory cortex. Neuron 40:859–869. CrossRef Medline

Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322:970–973. CrossRef Medline

Frey BJ, Dueck D (2007) clustering by passing messages between data points. science 315:972–976. CrossRef Medline

Friederici AD (2012) The cortical language circuit: from auditory perception to sentence comprehension. Trends Cogn Sci 16:262–268. CrossRef Medline

Fu QJ, Shannon RV (2000) Effect of stimulation rate on phoneme recognition by Nucleus-22 cochlear implant listeners. J Acoust Soc Am 107:589–597. CrossRef Medline

Ganong WF 3rd (1980) Phonetic categorization in auditory word perception. J Exp Psychol Hum Percept Perform 6:110–125. CrossRef Medline

Giraud AL, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. Nat Neurosci 15:511–517. CrossRef Medline

Giraud AL, Lorenzi C, Ashburner J, Wable J, Johnsrude I, Frackowiak R, Kleinschmidt A (2000) Representation of the temporal envelope of sounds in the human brain. J Neurophysiol 84:1588–1598. Medline

Güçlü U, van Gerven MA (2015) Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J Neurosci 35:10005–10014. CrossRef Medline

Güçlü U, van Gerven MA (2014) Unsupervised feature learning improves prediction of human brain activity in response to natural images. PLoS Comput Biol 10:e1003724. CrossRef Medline

Güçlü U, Thielen J, Hanke M, van Gerven MAJ (2016) Brains on beats. Adv Neural Inf Process Syst 29:2101–2109.

Hagoort P (2005) On Broca, brain, and binding: a new framework. Trends Cogn Sci 9:416–423. CrossRef Medline

Hagoort P (2013) MUC (memory, unification, control) and beyond. Front Psychol 4:416. CrossRef Medline

Hagoort P, Indefrey P (2014) The neurobiology of language beyond single words. Annu Rev Neurosci 37:347–362. CrossRef Medline

Halgren E, Dhond RP, Christensen N, Van Petten C, Marinkovic K, Lewine JD, Dale AM (2002) N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. Neuroimage 17:1101–1116. CrossRef Medline

Hermes D, Miller KJ, Noordmans HJ, Vansteensel MJ, Ramsey NF (2010) Automated electrocorticographic electrode localization on individually rendered brain surfaces. J Neurosci Methods 185:293–298. CrossRef Medline

Hermes D, Miller KJ, Vansteensel MJ, Aarnoutse EJ, Leijten FS, Ramsey NF (2012) Neurophysiologic correlates of fMRI in human motor cortex. Hum Brain Mapp 33:1689–1699. CrossRef Medline

Hickok G (2012) Computational neuroanatomy of speech production. Nat Rev Neurosci 13:135–145. CrossRef Medline

Hickok G, Poeppel D (2004) Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92:67–99. CrossRef Medline

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402. CrossRef Medline

Houtgast T, Steeneken HJM (1985) A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J Acoust Soc Am 77:1069–1077. CrossRef

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

Joris PX, Schreiner CE, Rees A (2004) Neural processing of amplitude-modulated sounds. Physiol Rev 84:541–577. CrossRef Medline

Kendall MG, Stuart A (1973) The advanced theory of statistics. Vol. 2: inference and relationship. London: Griffin.

Khaligh-Razavi SM, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain it cortical representation. PLoS Comput Biol 10:e1003915. CrossRef Medline

Kubanek J, Brunner P, Gunduz A, Poeppel D, Schalk G (2013) The tracking of speech envelope in the human cortex. PLoS One 8:e53398. CrossRef Medline

Lachaux JP, Fonlupt P, Kahane P, Minotti L, Hoffmann D, Bertrand O, Baciu M (2007) Relationship between task-related gamma oscillations and BOLD signal: new insights from combined fMRI and intracranial EEG. Hum Brain Mapp 28:1368–1375. 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

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. Psychol Rev 74:431–461. CrossRef Medline

Martin S, Brunner P, Holdgraf C, Heinze H-J, Crone NE, Rieger J, Schalk G, Knight RT, Pasley BN (2014) Decoding spectrotemporal features of overt and covert speech from the human cortex. Front Neuroengineering 7.

Mazaika PK, Hoeft F, Glover GH, Reiss AL (2009) Methods and software for fMRI analysis of clinical subjects. Neuroimage 47:S58.

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748. CrossRef Medline

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

Murphy KP (2012) Machine learning: a probabilistic perspective. Cambridge, MA: MIT.

Neggers SF, Hermans EJ, Ramsey NF (2008) Enhanced sensitivity with fast three-dimensional blood-oxygen-level-dependent functional MRI: comparison of SENSE-PRESTO and 2D-EPI at 3 T. NMR Biomed 21:663–676. 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

Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF (2012) Reconstructing speech from human auditory cortex. PLoS Biol 10:e1001251. CrossRef Medline

Rauschecker JP (1998) Cortical processing of complex sounds. Curr Opin Neurobiol 8:516–521. CrossRef Medline

Rauschecker JP, Tian B (2000) Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proc Natl Acad Sci U S A 97: 11800–11806. CrossRef Medline

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

Saur D, Kreher BW, Schnell S, Kummerer D, Kellmeyer P, Vry MS, Umarova R, Musso M, Glauche V, Abel S, Huber W, Rijntjes M, Hennig J, Weiller C (2008) Ventral and dorsal pathways for language. Proc Natl Acad Sci U S A 105:18035–18040. CrossRef Medline

Schönwiesner M, Zatorre RJ (2009) Spectrotemporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc Natl Acad Sci U S A 106:14611–14616. CrossRef Medline

Wessinger CM, Buonocore MH, Kussmaul CL, Mangun GR (1997) Tonotopy in human auditory cortex examined with functional magnetic resonance imaging. Hum Brain Mapp 5:18–25. CrossRef Medline