Hierarchy of Speech-Driven Spectrotemporal Receptive Fields in Human Auditory Cortex

Jonathan H. Venezia[1,2], Steven M. Thurman[3], Virginia M. Richards[4], and Gregory Hickok[4]

[1]VA Loma Linda Healthcare System, Loma Linda, CA

[2]Dept. of Otolaryngology, School of Medicine, Loma Linda University, Loma Linda, CA

[3]U.S. Army Research Laboratory, Aberdeen Proving Ground, MD

[4]Depts. of Cognitive Sciences and Language Science, University of California, Irvine, Irvine, CA

Running title: Speech-driven STRFs in human cortex

Corresponding Author:

Jonathan H. Venezia

11201 Benton Street

Loma Linda, CA 92357

Email: jonathan.venezia@va.gov

Phone: 909-825-7084 x7292

Fax: 909-777-3244

Abstract

Existing data indicate that cortical speech processing is hierarchically organized. Numerous studies have shown that early auditory areas encode fine acoustic details while later areas encode abstracted speech patterns. However, it remains unclear precisely what speech information is encoded across these hierarchical levels. Estimation of speech-driven spectrotemporal receptive fields (STRFs) provides a means to explore cortical speech processing in terms of acoustic or linguistic information associated with characteristic spectrotemporal patterns. Here, we estimate STRFs from cortical responses to continuous speech in fMRI. Using a novel approach based on filtering randomly-selected spectrotemporal modulations (STMs) from aurally-presented sentences, STRFs were estimated for a group of listeners and categorized using a data-driven clustering algorithm. 'Behavioral STRFs' highlighting STMs crucial for speech recognition were derived from intelligibility judgments. Clustering revealed that STRFs in the supratemporal plane represented a broad range of STMs, while STRFs in the lateral temporal lobe represented circumscribed STM patterns important to intelligibility. Detailed analysis recovered a bilateral organization with posterior-lateral regions preferentially processing STMs associated with phonological information and anterior-lateral regions preferentially processing STMs associated with word- and phrase-level information. Regions in lateral Heschl's gyrus preferentially processed STMs associated with vocalic information (pitch).

Keywords: speech perception, spectrotemporal modulations, fMRI, bubbles, classification images

Highlights

- A new method, “auditory bubbles”, is developed to estimate speech-driven spectrotemporal receptive fields (STRFs) using fMRI
- STRFs are estimated at locations throughout the auditory cortex
- Groups of STRFs with similar functional properties are identified using an unsupervised clustering algorithm
- Results support an interpretation in which STRFs are hierarchically organized and specialized within hierarchical levels
- Early auditory areas encode vocalic information (pitch)
- Posterior and anterior superior-temporal regions encode phonetic information on temporal scales associated with phonemes/syllables and words/phrases, respectively

1. Introduction

Current functional neuroanatomical models (de la Mothe et al., 2006; Hackett, 2011; Hackett et al., 2014; Hackett et al., 1998; Kaas and Hackett, 1998, 2000) suggest that primate auditory cortex is organized as a regional hierarchy in which information flows along two major anatomical axes: (1) from core to belt to parabelt regions; and (2) from caudal to rostral regions. The hierarchical nature of this organization has been confirmed by physiological data, which show that temporal integration windows, frequency tuning bandwidth, response latency, and stimulus selectivity (i.e., receptive field complexity) tend to increase along these axes (Bendor and Wang, 2008a; Brugge and Merzenich, 1973; Camalier et al., 2012; Kikuchi et al., 2010; Kuśmierek and Rauschecker, 2009; Lakatos et al., 2005; Rauschecker, 1998; Rauschecker and Tian, 2004; Rauschecker et al., 1995; Rauschecker et al., 1997; Recanzone et al., 2000; Scott et al., 2011). Response patterns in human auditory cortex measured using electrocorticography (ECoG) and functional magnetic resonance imaging (fMRI) largely mirror this pattern (Bitterman et al., 2008; Brugge et al., 2009; Brugge et al., 2008; Chevillet et al., 2011; Howard et al., 2000; Leaver and Rauschecker, 2010; Liegeois-Chauvel et al., 1994; Liegeois-Chauvel et al., 1991; Nourski et al., 2013; Nourski et al., 2014; Nourski et al., 2012; Wessinger et al., 2001; Woods et al., 2010). Some exceptions have been noted including short latency responses in human posterolateral superior temporal gyrus (STG; Nourski et al., 2014) and sensitivity to complex features of synthetic speech sounds in the primary auditory cortex of ferrets (Bizley et al., 2009; Town and Bizley, 2013), although these findings remain interpretable within a hierarchical framework (Bizley and Cohen, 2013; Nourski et al., 2014).

Taking note of these data and, indeed, of the general trend for sensory cortices to analyze and represent complex inputs via hierarchical, feedforward processing (Felleman and Van Essen, 1991; Foxe and Schroeder, 2005; Griffiths and Warren, 2004; Hilgetag et al., 2000; Riesenhuber and Poggio, 2002; Serre et al., 2007), many speech researchers have embraced the notion that cortical analysis of speech sounds proceeds in a hierarchical fashion (Bornkessel-Schlesewsky et al., 2015; Peelle et al., 2010; Poeppel et al., 2012; Okada et al., 2010; Rauschecker and Scott, 2009). Early work in human auditory

neuroimaging demonstrated that lower-level (core-like) regions of the auditory cortex respond well to simple stimuli such as tones or unmodulated broadband noise, while belt-like regions in the supratemporal plane anterior and posterior to the auditory core respond more strongly to temporally-modulated signals, and parabelt-like regions in the lateral STG and superior temporal sulcus (STS) respond best to spectrotemporally-complex stimuli such as speech (Binder et al., 2000; Hickok and Poeppel, 2004; Scott and Johnsrude, 2003; Scott and Wise, 2003; Zatorre et al., 2002). Moreover, a subset of these later auditory regions respond preferentially to intelligible speech compared to unintelligible sounds with similar spectrotemporal complexity, e.g., noise-vocoded speech or spectrally rotated speech (Davis and Johnsrude, 2003; Narain et al., 2003; Scott et al., 2000).

While there is broad agreement that the human auditory cortex is hierarchically organized for speech, it remains unclear exactly what speech information is being encoded within different levels of the hierarchy. Several recent imaging studies using multivariate analysis methods have shown that early auditory regions in and around Heschl's gyrus are able to distinguish intelligible speech from acoustically complex control stimuli including spectrally rotated speech (Evans et al., 2014; McGettigan et al., 2012; Okada et al., 2010). In line with hierarchical interpretations, these and other studies demonstrated that such discriminative capacity is likely driven by the exquisite sensitivity of early auditory areas to slight variation in acoustic form, while higher-level speech-selective regions are relatively invariant to superficial acoustic variation (Evans, 2017; Evans and Davis, 2015; Okada et al., 2010). However, work by Poeppel and others (Boemio et al., 2005; Overath et al., 2015) suggests that both early and late regions are sensitive to acoustic form, where the distinction between hierarchical levels concerns their tuning to temporal patterns on different time scales, with later regions such as the STS generally preferring longer time scales. A lack of data regarding the details of acoustic vs. abstract speech encoding at different cortical levels leads to at least two outstanding questions: (a) whether processing at higher levels of the hierarchy is bilaterally organized or left-lateralized; and (b) whether the processing hierarchy proceeds along posterior or anterior pathways (or both).

A promising technique to probe the detailed information encoded in cortical responses to natural sounds is the estimation of spectrotemporal receptive fields (STRFs). A STRF is a linear filter in the time-frequency domain showing the spectrotemporal patterns that best drive an individual neuron or neuronal population. Though STRF analysis was initially developed to characterize single-unit responses in animal models (cf., Theunissen and Elie, 2014), speech-driven STRFs have recently been derived from human electroencephalography (EEG), magnetoencephalography (MEG), and ECoG data (Ding and Simon, 2012; Lalor and Foxe, 2010; Mesgarani and Chang, 2012). At the cortical level, it has proven useful to characterize STRFs in the spectrotemporal modulation (STM) domain (Hullett et al., 2016; Kowalski et al., 1996; Shamma, 2001). An STM is a fluctuation in acoustic energy at a given rate (i.e., over time) and scale (i.e., over frequency). In the context of speech, different STM patterns are associated with different levels of acoustic or linguistic analysis, e.g., formant vs. harmonic structure (Elliott and Theunissen, 2009) or phonemes vs. syllables/words (Hullett et al., 2016). Recent ECoG studies demonstrate that speech-selective regions of the STG exhibit spatially organized tuning to a range of STMs (Hullett et al., 2016), that speech signals can be reconstructed from patterns of activity in the STG using STM-based STRF models (Pasley et al., 2012), and that shifts in STM tuning within the auditory cortex underlie cortical plasticity associated with priming-induced changes in speech intelligibility (Holdgraf et al., 2016). However, ECoG is inherently limited due to its invasive nature and relatively limited coverage of auditory cortical areas in the supratemporal plane (Reddy et al., 2010), and related non-invasive techniques (EEG/MEG) are limited by relatively poor spatial resolution. As a result, following pioneering work by Schönwiesner and Zatorre (2009) who used (synthetic) dynamic spectral ripple stimuli to derive voxel-wise STRFs in fMRI, several recent studies have developed fMRI encoding (Naselaris et al., 2011) paradigms for measuring STRFs throughout the auditory cortex using speech and other natural sounds as driving stimuli (de Heer et al., 2017; Santoro et al., 2014; Santoro et al., 2017). A limitation of these encoding methods for studying speech is that there may not be sufficient long-term acoustic variability in natural speech signals to allow derivation of good-fitting STRFs with fMRI, which has relatively poor temporal resolution. This may be particularly true in the STM domain (de Heer et al.,

2017), where long-term patterns of acoustic energy in the speech modulation power spectrum (MPS) are quite stable across utterances (Elliott and Theunissen, 2009; Fig. 1A).

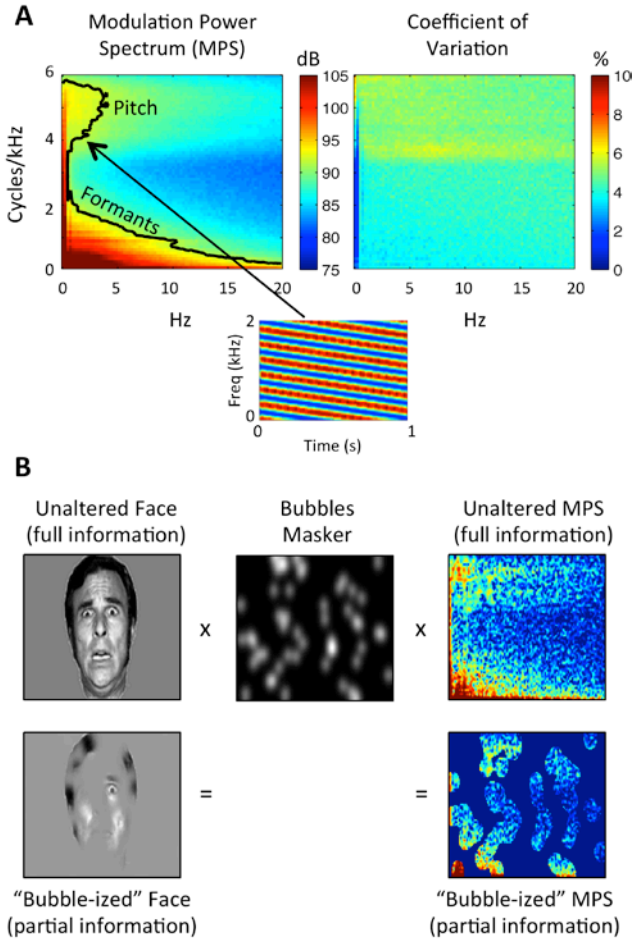

Figure 1. (A) Speech Modulation Power Spectrum*. Left: Average MPS of 452 sentences spoken by a single female talker. The MPS describes speech as a weighted sum of spectrotemporal ripples containing energy at a unique combination of temporal (Hz; abscissa) and spectral (cycles/kHz; ordinate) modulation rate. Modulation energy (dB, arb. ref; color scale) clusters into two discrete regions: a high-spectral-modulation-rate region corresponding to finely spaced harmonics of the fundamental (a "pitch region") and a low-spectral-modulation-rate region corresponding to coarsely spaced resonant frequencies of the vocal tract (a "formant region"). The black contour line indicates the modulations accounting for 80% of the total modulation power. A spectrogram of an example spectrotemporal ripple (2 Hz, 4 cyc/kHz) is shown beneath. Right: Coefficient of variation across the 452 sentences (sd/mean), expressed as a percentage (color scale). Plotted on the same axes as the MPS. There is relatively little variation across utterances (maximum CV ~7%).* ***(B) Bubbles Procedure****. Bubbles (middle) are applied to an image of a face (left) and the MPS of an individual sentence (right). In either case, bubbles reduce the information in the stimulus. Different random bubble patterns are applied across trials of an experiment. For auditory bubbles, we in practice use a binary masker with bubbles that are larger than those shown in the example.*

In the present fMRI study, we *induce* random variation in the speech MPS to derive speech-driven STRFs using a classification image technique known as "bubbles" (Gosselin and Schyns, 2001).

Used frequently in vision research, the bubbles procedure involves masking randomly-selected regions of an image (e.g., a face; Fig. 1B, left) and relating the masker patterns to behavior (e.g., emotion identification) using reverse correlation to identify task-relevant features of the stimulus (i.e., a 'perceptual receptive field'). We recently extended the bubbles procedure to the auditory domain by applying bubbles-like filters to the MPS of auditory sentences (Fig. 1B, right) and deriving behavioral classification images describing the relative contributions of different STMs to intelligibility (Venezia et al., 2016). Here, we apply the auditory bubbles procedure to fMRI by using single-trial blood-oxygen-level dependent (BOLD) response magnitudes (cf., Smith et al., 2008) to derive speech-driven STRFs in the STM domain (Fig. 2). We then apply an unsupervised clustering algorithm to reveal the large-scale organization of STRF patterns in the human auditory cortex. The broad goal of this data-driven approach was to reveal the precise speech information encoded in different regions of the auditory cortex. The specific goals were three-fold: (1) to estimate speech-driven STRFs throughout the auditory cortex; (2) to probe for characteristic patterns of STRF organization at different levels of cortical processing as they relate to different levels of acoustic or linguistic analysis of speech; and (3) to compare behavioral classification images for intelligibility (perceptual receptive fields) to BOLD-based measures (STRFs). Here and throughout, it is assumed that cortical analysis of speech sounds proceeds hierarchically in terms of increasing functional complexity (e.g., spectrotemporal features→phonemes→syllables→words), and that STRF patterns correlated with these different levels of analysis can be used to capture the computational roles of different stages in the feedforward cortical speech network. Therefore, we did not aim to explicitly test for a hierarchical organization, but rather to probe the detailed organization of different levels of the presumed cortical speech hierarchy.

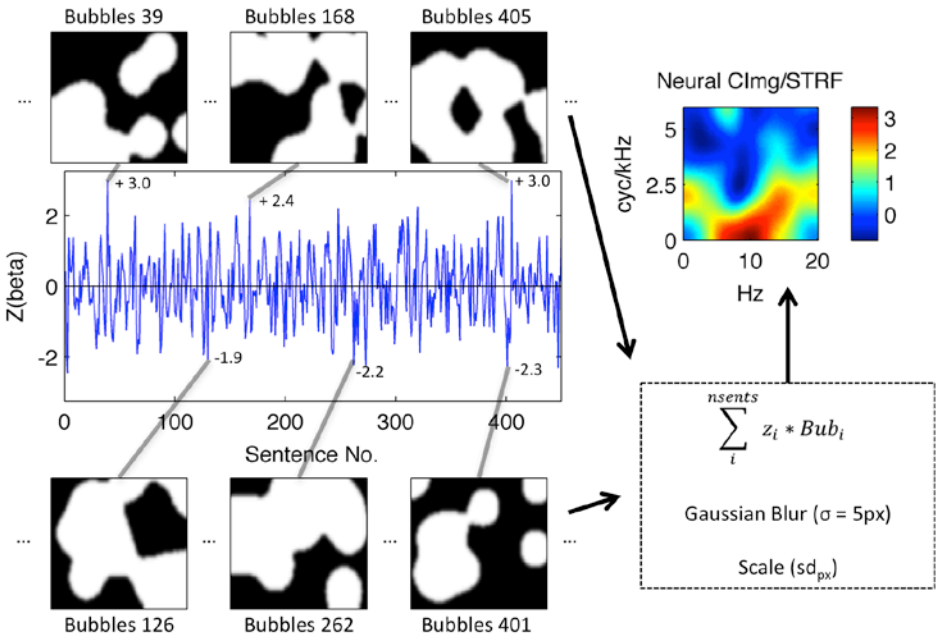

Figure 2. Bubbles Analysis Schematic. *A BOLD activation time-course from a single voxel in left Heschl's gyrus of a representative subject is shown (blue line). The time-course plots the z-scored time-series of single-trial activation magnitudes (beta; ordinate) evoked by "bubble-ized" sentences (Sentence No., abscissa). Example bubble patterns (black-and-white panels) associated with sentences that evoked relatively large (top) and small (bottom) activations are plotted and identified by their sentence number. Z-scored activation magnitudes associated with these examples are shown next to the corresponding point in the activation time-course. Bubbles are applied to the MPS of each sentence as shown in Fig. 1. White pixels show regions of the MPS that are transmitted to the listener, while black pixels show regions of the MPS that are removed. Each bubble pattern is multiplied by its associated z-score, and the series of bubble patterns is summed pixel-by-pixel. The resulting summed image is then blurred (Gaussian filter with sigma = 5 pixels) and scaled by the across-pixel standard deviation (sd_{px}). The result is a STRF (top right) showing which regions of the MPS best activated this voxel. The STRF color scale is in across-pixel standard deviation units, where large positive values (yellow-red) correspond to regions of the MPS that evoked relatively large activations.*

The results showed that all reliably-tuned STRFs were located in the auditory cortex and lateral superior temporal lobe. Data-driven clustering identified four groups of STRF patterns: (1) broad responses covering most of the speech MPS; (2) responses preferring high temporal modulation rates; and (3)/(4) responses closely matched to behavioral classification images for intelligibility. Group 1 was primarily located in the supratemporal plane including Heschl's gyrus/sulcus and the posterior STG, Group 2 was located primarily in medial supratemporal regions, and Groups 3/4 were located in the lateral STG and STS. The latter groups split anatomically and functionally along an STG-STS division with Group 3, which preferred slightly higher temporal modulation rates, located in more prominently in the STG, while Group 4 was located more prominently in the STS. Within each cluster group, there were

specialized STRF patterns including a high-spectral-modulation-rate response on part of Heschl's gyrus, and a low-temporal-modulation-rate response in the anterior temporal lobe. Together, these results characterize the cortical speech hierarchy in acoustic-informational terms and reveal processing specializations within and across levels of the hierarchy that relate directly to perception of intelligible speech.

2. Materials and Methods

2.1 Participants

Ten participants (mean age = 26, range = 20-33, 2 females) took part in the experiment. All participants were right-handed, native speakers of American English with self-reported normal hearing and normal or corrected-to-normal vision. All participants provided informed consent in accordance with the University of California, Irvine Institutional Review Board guidelines. No statistical methods were used to predetermine sample size; rather, our sample size was set to match that used in our previous psychophysical study using the same experimental paradigm (Venezia et al., 2016).

2.2 Bubbles Stimuli

The stimuli used in this study have been described in detail in our previous paper on auditory bubbles (refer to the "uncompressed" or UC stimuli; Venezia et al., 2016). Briefly, the speech source signals were recordings of 452 sentences from the Institute of Electrical and Electronics Engineers (IEEE) sentence corpus (IEEE, 1969) spoken by a single female talker. Each sentence was stored as a separate .wav file digitized at 22050 Hz with 16-bit quantization. The sound files were zero-padded to an equal duration of 3.29 s. To create the bubbles stimuli, the sentence audio files were filtered to remove randomly-selected patterns of energy in the spectrotemporal modulation domain. For each sentence, a log-power (dB) spectrogram was obtained using Gaussian windows with a 4.75 ms-33.5 Hz time-frequency scale. The 2D modulation spectrum was then obtained as the modulus of the 2D Fourier

transform of the spectrogram. The modulation spectrum was restricted to temporal modulation rates less than 20 Hz and spectral modulation rates less than 6 cyc/kHz, a region containing approximately 90-95% of speech energy (Elliott and Theunissen, 2009). A 2D filter of the same dimensions as the modulation spectrum was created by generating an identically-sized image with a set number of randomly-chosen pixel locations assigned the value 1 and the remainder of pixels assigned the value 0. A symmetric Gaussian blur (sigma = 7 pixels) was applied to the image and all resultant values above 0.1 were set to 1 while the remaining values were set to 0. This produced a binary image with a number of randomly-located contiguous regions with value 1. A second Gaussian blur (sigma = 1 pixel) was applied to smooth the edges between 0- and 1-valued regions, producing the final 2D filter. The number of pixels originally assigned a value of 1 (i.e., prior to any blurring) corresponds to the number of "bubbles" in the filter. The modulation spectrum was then multiplied by the filter, thus removing randomly-selected sections of modulation energy from the signal. The multiplication procedure was performed identically for upward- and downward-sweeping spectrotemporal modulations, effectively collapsing over these representations in subsequent analyses (see Venezia et al., 2016 for a discussion). A filtered speech waveform was obtained from the degraded modulation spectrum by performing an inverse 2D Fourier transform followed by iterative spectrogram inversion (Griffin and Lim, 1984). For each of the 452 sentences, filtered versions were created using independent, randomly-generated filter patterns. This renders some filtered items unintelligible while others remain intelligible depending on the filter pattern. Separate sets of filtered stimuli were created using different numbers of bubbles (20-100 in steps of five). This produced a total of 7684 filtered sentences. All stimuli were generated offline and stored prior to the experiment. For reference, the average proportion of the STM spectrum revealed to the listener is ~ 0.25 for 20 bubbles and ~ 0.7 for 100 bubbles, and the relation between number of bubbles and proportion of the STM spectrum revealed is exponential.

2.3 Procedure

Participants listened to filtered sentences during sparse acquisition fMRI scanning. On each trial of the experiment, a single filtered sentence was presented in the silent period (4 s) between image acquisitions (2 s). Stimulus presentation was triggered 400 ms into the silent period and sentence duration ranged from 1.57-3.29 s (mean = 3.02 s). At the end of sentence presentation, participants were visually cued to make a subjective yes-no judgment indicating whether the sentence was intelligible or not. The number of bubbles was adjusted trial-by-trial using an up-down staircase procedure such that participants rated sentences as intelligible on ~ 50% of trials. Equilibrium for the staircase procedure was reached after ~ 10-20 trials, and the adaptive track was maintained continuously across scan runs. It should be noted that performance increases systematically with number of bubbles only in the long run, since performance depends not only on the number of bubbles (total information transmitted) but on the pattern of bubbles (particular information transmitted). Therefore, in practice the equilibrium number of bubbles varies over a range. Our previous behavioral work (Venezia et al., 2016) shows that performance nonetheless converges to the expected average response rate of 50% over a large enough number of trials (consistent with the number of trials performed here).

A total of 45 experimental trials were performed per scan run. An additional 5 baseline trials were also presented on quasi-randomly chosen trials during each scan run. On these trials, participants viewed short clips (3.33 s duration, 15 fps) from the video game "Snake" in which a moving object navigates though a 2D bordered grid. Video onsets occurred 400 ms into the silent period between image acquisitions. At the end of each video clip participants were visually cued to make a yes-no judgment indicating whether the moving object made at least one left-hand turn. All participants performed well on the baseline task (mean = 91% correct, sd = 6%, min = 82%). At the beginning of each trial, participants were visually cued to "listen" (experimental trials) or "watch" (baseline trials).

Behavioral responses were generated by button-press on an MR-compatible response box. Button-presses in the experimental and baseline tasks were generated during the subsequent MR image acquisition. Prior to scanning, participants completed a short behavioral session outside the scanner to

familiarize themselves with the stimuli and tasks. Two participants completed a total of 9 scan runs (405 experimental trials) and the remaining eight participants completed a total of 10 scan runs (450 experimental trials). For each participant, sentences were selected from the list of 452 source sentences without replacement in randomly permuted order. On any trial, the stimulus was drawn from the pool of 7684 filtered sentences based on the selected sentence number and the number of bubbles (adjusted adaptively). Thus, it was possible for particular filtered sentences to be repeated across participants. In practice, filtered sentences were repeated on average only 1.39 times across participants (i.e., most items were not repeated). Auditory stimuli were amplified using a Dayton DTA-1 model portable amplifier and presented diotically over Sensimetrics S14 piezoelectric earphones. Participants were asked to adjust the volume to a comfortable level slightly above that of conversational speech (~75-80 dB SPL). Visual stimuli were back-projected via a Christie DLV1400-DX DLP projector onto a screen at the head end of the magnet bore (spatial resolution: 1024x768 pixels; refresh rate: 60 Hz). Participants viewed the display on an angled front surface mirror mounted on the head coil with a viewing distance of ~ 70 cm. Stimulus presentation was controlled using the Psychophysics Toolbox Version 3 (Kleiner et al., 2007). A single high-resolution T1 anatomical image was acquired for each participant at the end of fMRI scanning.

2.4 Image acquisition

Images were acquired on a Philips Achieva 3T MRI scanner with a 32-channel sensitivity encoding (SENSE) head coil located at the University of California, Irvine Neuroscience Imaging Center. T2*-weighted images (gradient-echo EPI) were acquired using a sparse acquisition sequence (35 axial slices, interleaved slice order, TR = 6 s, TA = 2 s, TE = 30 ms, flip = 90°, SENSE factor = 1.7, reconstructed voxel size = 1.875 x 1.875 x 3 mm, matrix = 128 x 128, no gap). Fifty-two EPI volumes were collected per scan run. A single high-resolution, T1-weighted anatomical image was collected for each participant using a magnetization prepared rapid gradient echo (MPRAGE) sequence (160 axial

slices, TR = 8.4 ms, TE = 3.7 ms, flip = 8°, SENSE factor = 2.4, 1 mm isotropic voxels, matrix = 256 x 256).

2.5 Behavioral Analysis

For the experimental task, the goal was to calculate a behavioral classification image based on participants' yes-no responses indicating whether each filtered sentence was judged to be intelligible or not. To accomplish this, the 2D bubbles filter patterns associated with each sentence were treated as predictors of yes-no intelligibility judgments. Specifically, for each participant, a weighted sum of the 2D bubbles filters across trials was performed in which "no" trials received a negative weight equal to the proportion of "yes" trials, p_{yes}, and "yes" trials received a positive weight equal to the complement of p_{yes}:

$$CImg_B = \sum_{i=1}^{ntrials} w_i * Bub_i$$

where i is the trial index, w_i is the weight associated with the response on a given trial (p_{yes} or its complement), Bub_i is the 2D bubbles filter applied on a given trial, and $CImg_B$ is the resulting behavioral classification image showing which regions of the speech modulation spectrum predict a "yes" judgment (i.e., support intelligibility). Trials in which no button press was recorded were excluded from analysis. To form a group-level classification image, the behavioral classification images from each participant were summed, smoothed with a symmetric Gaussian filter (sigma = 5 pixels), and z-scored (Venezia et al., 2016).

2.6 MR Image Preprocessing.

Automated cortical surface reconstruction based on the T1-weighted anatomical images was performed in Freesurfer v5.3 (Fischl, 2012). For each participant, the inflated surface mesh and white matter segmentation volume were manually checked to ensure no large-scale errors occurred during automated tissue segmentation. Right and left hemisphere cortical surface meshes were then converted to

AFNI/SUMA format, co-registered to the participant's native-space anatomical volume, resampled to a standard topology via linear icosahedral tessellation with 128 edge divides, and merged into a single surface containing 327684 nodes using the prep_afni_surf.py function of Oosterhof's "surfing" toolbox v0.6 (https://github.com/nno/surfing; Oosterhof et al., 2011). The standard-topology mesh is nearly identical in geometry (i.e., cortical folding patterns) to the original surface but has been re-aligned to a template such that each surface node represents the same cortical location across participants (Saad and Reynolds, 2012). Group-level results are plotted on a surface mesh generated from the Colin 27 template brain after resampling to the same standard topology.

Preprocessing of the functional data was performed using AFNI v17.0.05 (Cox, 2012). Functional images were slice-timing corrected based on slice time offsets extracted from the Philips PAR files, followed by realignment (motion correction) and coregistration to the T1-weighted anatomical image. The functional data were then mapped to the merged, standard-topology surface mesh and smoothed to a target level of 4 mm full width at half maximum. An iterative procedure (AFNI SurfSmooth) was implemented in which the level of smoothness in the data (~ 2.5 mm intrinsic smoothness at the outset) was estimated from the residual time series after high-order detrending, and additional smoothing was applied in small increments until the target level was reached. Finally, the data from each scan run were scaled to have a mean of 100 across time points subject to a range of 0-200.

2.7 fMRI Beta Time Series Estimation

The onset and offset of sound energy for each experimental-stimulus sound file were identified based on the windowed root-mean-square amplitude of the signal (silence threshold = 0.0035). These measurements were used to generate a series of stimulus onsets and durations that defined the event timing of the experiment for each participant. The resulting event timing was entered as an input to the 3dDeconvolve function in AFNI using the stim_times_IM option with a duration-modulated BLOCK hemodynamic response function. This call to 3dDeconvolve produced a predicted activation time-course separately for each experimental trial. These predicted time-courses were at first sampled with a temporal

resolution of 0.1 s, but were subsequently down-sampled by averaging together the values occurring during periods of image acquisition (i.e., the 2 s TA within each 6 s TR). This produced a final set of predicted activation time-courses with a temporal resolution of 2 s, accounting for temporal discontinuities introduced by sparse sampling (Perrachione and Ghosh, 2013). This set of predictors along with additional baseline and third-order polynomial drift terms, and six regressors of no interest corresponding to motion parameters estimated during the realignment stage of preprocessing, all appropriately down-sampled, were combined to create an experimental design matrix. Baseline-task events were not modeled explicitly and were thus captured by the baseline term of the design matrix. The 3dLSS function in AFNI was then used to perform "least squares-separate" (Mumford et al., 2012) regression on the preprocessed fMRI data using the aforementioned design matrix. The output from 3dLSS was a beta time-series at each voxel representing the overall magnitude of activation for each experimental trial over the duration of the experiment. Extreme (outlier) beta values were excluded based on the following formula for outlier detection:

$$C = \alpha * \sqrt{\pi/2} * MAD;\ \alpha = \Phi^{-1}\left(1 - 0.001/N\right)$$

where C is the outlier cutoff, MAD is the median absolute deviation, Φ is the cumulative normal distribution function, and N is the number of time points. Beta values for which the absolute deviation from the median exceeded C were excluded.

2.8 Spectrotemporal Receptive Field Estimation

Neural classification images (STRFs) were calculated just as behavioral classification images except that bubbles filter patterns were used to predict the estimated fMRI beta time series at each cortical surface node rather than participant behavior (see Fig. 2). That is, for each surface node (voxel-like unit) in each participant, a weighted sum of the 2D bubbles filters across trials was performed such that each trial received a weight equal to the z-scored activation magnitude on that trial:

$$STRF = \sum_{i=1}^{ntrials} z_i * Bub_i$$

where *i* is the trial index, z_i is the magnitude of neural activation on a given trial (taken from the z-scored beta time series), Bub_i is the 2D bubbles filter applied on a given trial, and *STRF* is the resulting neural classification image showing which regions of the speech modulation spectrum best activated a given cortical surface node[1]. Trials in which no behavioral response was given were excluded from analysis. To create a group-level STRF, individual-participant STRFs were smoothed with a symmetric Gaussian filter (sigma = 5 pixels), scaled by their across-pixel standard deviation, averaged across subjects, and scaled pixel-wise by the between-subject standard error to produce a t-score image. This was performed separately for each node in the standard-topology cortical surface mesh.

The STRFs estimated in this manner may contain an intelligibility bias – that is, if the magnitude or variance of the neural signal at a given surface node is influenced by the intelligibility of the speech signal, then relatively more weight could be placed on intelligible compared to unintelligible trials or vice versa. In short, the STRF could reflect a global effect of intelligibility. Therefore, in addition to the primary STRF analysis, we estimated STRFs separately for intelligible and unintelligible trials as follows:

$$STRF_{Intel} = \sum_{i=1}^{n\_intel} z_i * Bub_i$$

where *n_intel* is the number of trials judged as intelligible by the listener, *i* is the trial index, z_i is the magnitude of neural activation on a given intelligible trial, Bub_i is the 2D bubbles filter applied on a given intelligible trial, and $STRF_{Intel}$ is the resulting neural classification image for intelligible trials; and

[1] Note that the average number bubbles at behavioral threshold (50% "yes" responses) can vary across participants, but this is not expected to bias STRF estimates given the number of trials performed in this study. Bias of this sort could only occur if the number of trials was too few to allow a comprehensive random sampling of the stimulus domain at a given average number of bubbles. The typical average number of bubbles in this study was ~ 50, at which point ~ 50% of the stimulus space was revealed on each trial (i.e., optimum sampling of the stimulus space).

$$STRF_{Unintel} = \sum_{i=1}^{n\_unintel} z_i * Bub_i$$

where *n_unintel* is the number of trials judged as unintelligible by the listener, *i* is the trial index, z_i is the magnitude of neural activation on a given unintelligible trial, Bub_i is the 2D bubbles filter applied on a given unintelligible trial, and $STRF_{Unintel}$ is the resulting neural classification image for unintelligible trials. Crucially, neural beta time series were z-scored *separately* for intelligible and unintelligible trials, thus removing any differences in the mean and variance of the neural signal due to intelligibility alone. An unbiased STRF estimate was then generated as follows:

$$STRF_{Unbiased} = \frac{STRF_{Intel} + STRF_{Unintel}}{2}$$

where $STRF_{Unbiased}$ is a neural classification image with the global effect of intelligibility removed. An unweighted average is used because the number of intelligible trials is kept equal to the number of unintelligible trials by the up-down staircase implemented on yes-no intelligibility judgments (2.3). The procedure for calculating $STRF_{Unbiased}$ is thus equivalent to separately z-scoring the beta time series for intelligible and unintelligible trials, recombining those time series in the original order, and performing a weighted sum of bubbles filters across all trials as described above for uncorrected STRFs. A similar approach is often taken in the context of multivariate pattern analysis (MVPA), where the neural signal is normalized across voxels at each time point – and thus implicitly normalized across conditions or classes – in order to prevent decoding algorithms from predicting trial types based on condition-mean differences in signal amplitude as opposed to differences in multi-voxel patterns (Coutanche, 2013). Note that $STRF_{Unbiased}$ is still biased in the more straightforward sense that only (intelligible and unintelligible) speech, and no other class of sounds, was used as a driving stimulus. Group-level (t-scored) versions of $STRF_{Unbiased}$ were calculated as described above for uncorrected STRFs. The subsequently-described

analyses were performed only on the uncorrected STRFs, as $STRF_{Unbiased}$ was calculated primarily as a basis for comparison to $STRF$.

2.9 Quantification of Modulation Tuning

To determine whether a given cortical surface node displayed significant modulation tuning across participants, we tested whether the group-level STRF at that surface node demonstrated a statistically significant non-zero response. Specifically, a p-value was calculated for each pixel in the STRF based on the group-level t-score at each pixel location (i.e., a p-value image was generated), and the p-values were then adjusted for multiple comparisons across pixels using the false discovery rate (FDR) procedure (Benjamini and Hochberg, 1995). Multiple comparisons across cortical surface nodes were then adjusted for multiple comparisons via the 'wild bootstrap,' a nonparametric technique that has been shown to provide appropriate control of false positives in most situations (Eklund et al., 2016). Specifically, each participant's first-level STRF was multiplied by 1 or -1, first-level STRFs were combined across participants (t-scoring), and this was repeated for different permuted sign-flip vectors. For each permutation order, the minimum FDR-corrected p-value was computed for each second-level (t-scored) STRF, and the size of the maximum cluster of nodes (surface area in mm^2) with a minimum p-value satisfying FDR < 0.05 was calculated and placed in a null distribution of maximum cluster sizes. Since we had 10 subjects, we were able to perform all possible sign-flip permutations (2^{10}) to form a null distribution of 1024 maximum cluster sizes. The 95th percentile of this distribution (node-level corrected $p < 0.05$) was 116 mm^2. This surface area threshold was implemented separately for positively tuned (peak t-score is positive) and negatively tuned (peak t-score is negative) surface nodes. Note that the null distribution of maximum cluster sizes is identical for positive and negative t-scores due to the symmetry of the sign-flip orders across all possible permutations.

For each significantly tuned surface node surviving the area threshold, the best temporal modulation frequency (tBMF) and best spectral modulation frequency (sBMF) were calculated for each of the individual-participant STRFs at that surface node. That is, the pixel with the largest response

magnitude was identified and the temporal modulation rate (Hz) and spectral modulation rate (cyc/kHz) represented by that pixel location were recorded. Similarly, the temporal peak modulation frequency (tPMF) and spectral peak modulation frequency (sPMF) were calculated from the group-level (t-score) neural STRF. We distinguish between BMF at the individual participant level and PMF at the group level because the BMF reflects the highest magnitude response per participant while the PMF reflects the most reliable response across participants.

2.10 Unsupervised Clustering of STRFs

Two large sets of positively tuned cortical surface nodes were identified in the left and right auditory cortices. To explore whether different groups of auditory-cortical surface nodes systematically represented different patterns of speech modulation energy, we performed unsupervised clustering using a Gaussian mixture model (GMM) analysis. The group-level STRFs at each significantly tuned auditory-cortical surface node were first down-sampled to 4 x 8 pixels and re-scaled to the range [0 1]. This re-scaling was performed because we were interested in identifying differences related to the pattern of tuning rather than the overall magnitude of the tuned response. The re-scaled neural STRFs were then vectorized to produce a length-32 feature vector at each cortical surface node. The feature vectors across all significantly tuned surface nodes served as the input to GMM analysis. Briefly, in the GMM analysis the distribution of each observation is specified by a probability density function through a finite mixture of *K* 32-dimensional Gaussian distributions:

$$f(x_i;\ \Psi) = \sum_{k=1}^{K} \pi_k f_k(x_i;\ \Theta_k),$$

where $\Psi = \pi_1, \dots, \pi_{K-1}, \theta_1, \dots, \theta_K$ are the model parameters; $f_k(x_i;\ \Theta_k)$ is the *k*th Gaussian distribution for observation x_i with parameter vector Θ_k; $\pi_1, \dots, \pi_{K-1}$ are the mixing weights such that $\sum_{k=1}^{K} \pi_k = 1$; and K is the number of mixture components. For each component, $f_k(x_i;\ \Theta_k) \sim N(\mu_k, \Sigma_k)$ where μ_k is the vector of means (i.e., representing that component's average feature vector) and Σ_k is the covariance

matrix. The GMM was fit in R v3.3.2 using the package mclust v5.3 (Scrucca et al., 2016). In mclust, the GMM parameters are estimated by maximization of the log-likelihood function using the expectation-maximization (EM) algorithm (McLachlan and Peel, 2004). The EM algorithm is initialized by a model-based agglomerative clustering routine. The mclust package allows for specification of covariance structures of varying complexity such that the volume, shape, and orientation of the covariance matrix, Σ_k, can be equal (E) or variable (V) across the *K* components (e.g., a model with equal volume, equal shape, and variable orientation would be coded EEV). The volume, shape, and orientation can also be set to identity (I) to generate simpler (non-ellipsoidal) GMMs. We implemented a model selection procedure in mclust where the best fitting GMM was selected from a set of models in which *K* varied from 2 to 40 and ten possible covariance structures were explored for each *K*.

The final number of components, *K*, and the best-fitting covariance parameterization were selected by choosing the model with the largest Bayesian Information Criterion (BIC), defined in mclust as:

$$BIC = 2 * \ell(\Psi; x_1, \dots, x_n) - df * \log(n),$$

where $\ell(\Psi; x_1, \dots, x_n)$ is the log-likelihood, *df* is the number of model parameters, and *n* is the number of observations. The initial agglomerative clustering was carried out on a subset of observations consisting of every other cortical surface node. A conjugate prior on the means and covariance matrices of the components was specified using the default options in mclust. The model selection procedure ultimately identified a GMM with $K = 18$ and a VVV covariance structure. Each cortical surface node was assigned to the component (cluster) with the maximum mixing weight, π_K. Cluster labels were then reassigned such that the correlation between μ_k at adjacent cluster labels was maximized. This was done by reordering the correlation matrix of μ_k's such that large values were shifted toward the diagonal; reordering was performed using hierarchical clustering based on Ward's distance. After label reassignment, it was clear from visual inspection that STRFs could be assigned to one of four "cluster groups" with similar response properties.

2.11 Intelligibility Maps

To facilitate comparison with previous work on cortical speech processing, we obtained surface-node-wise maps of the brain regions that responded significantly to intelligible speech. This was performed in two ways. First, following the classic cognitive subtraction approach (Petersen et al., 1989), we performed a traditional whole-brain general linear model analysis using a design matrix identical to that described above for beta time series estimation with the following exception: only two predictors of interest were included, one coding the predicted activation time-course for experimental trials in which the participant indicated that the sentence presented on that trial was intelligible, and a second coding experimental trials for which the participant responded unintelligible. For each participant, a contrast coefficient was calculated by comparing the response on intelligible vs. unintelligible trials at each cortical surface node. Second-level maps were computed by performing one-sample t-tests on the contrast coefficients at each cortical surface node. Multiple comparisons were corrected for via the 'wild bootstrap': second-level t-tests were repeated after flipping the sign of first-level contrast coefficients in different permuted orders across participants. For each permutation order, the maximum cluster of nodes with one-tailed (intelligible > unintelligible) $p < 0.005$ was calculated and placed in a null distribution of maximum cluster sizes. We again performed all possible sign-flip permutations (2^{10}) to form a null distribution of 1024 maximum cluster sizes. The 95th percentile of this distribution was 117 mm^2.

For the second approach, intelligibility maps were obtained by examining the correlation between behavioral classification images and STRFs. Specifically, for each participant at each surface node, the pixel-by-pixel Pearson correlation of the behavioral classification image and the STRF at that node was calculated. The correlation values were then subjected to Fisher's z transformation and the resulting z-maps were entered into a second-level analysis (one sample t-test). A second-level z-score was considered significant if the node-wise $p < 0.005$ (one-tailed, positive z) and the cluster size exceeded 130 mm^2 as determined by the 'wild bootstrap' implemented as described above for the intelligibility contrast coefficients.

2.12 Linear Mixed Effects Modeling

There are two significant concerns regarding the application of GMM clustering (2.10) to group-level STRFs: (1) cluster-level (i.e., aggregate) STRFs may not strongly reflect the patterns of individual STRFs at the constituent cortical surface nodes within a given cluster; and (2) group-level STRFs may not strongly reflect the patterns of the individual-participant STRFs from which the group-level data were derived. That is, aggregate STRF patterns are not guaranteed to be representative of the individual STRFs entered into the aggregate (Joosten & Neri, 2012). Therefore, individual-participant estimates of STRF-summary scalar metrics – tBMF, sBMF, and behavioral-neural intelligibility correlation (z) – were entered as the dependent variables in separate linear mixed effects (LME) models to evaluate their distribution across the cortical regions defined by data-driven STRF clustering (i.e., regions defined by cluster group membership). For comparison, we also examined the distribution of these scalar metrics across anatomically defined regions. The LME models included two fixed effects (*hemisphere* and cluster/anatomical *region*) and their interaction. The models also included random effects parameterized in such a way as to approximate traditional repeated measures ANOVA – namely, random intercepts were included for each within-participant error stratum implied by the fixed effects design. Models were fit in the R statistical computing environment using the 'mixed' function included in the afex package (Singmann and Kellen, 2017) version 0.18-0. The R model formulae were specified as follows:

$$DV \sim hemi + region + hemi{:}region + (1 \mid sub) + (1 \mid hemi{:}sub) + (1 \mid region{:}sub) + (1 \mid hemi{:}region{:}sub),$$

where *DV* is the dependent variable (tBMF, sBMF, or z), *sub* is a factor variable representing the identity of each participant, ':' represents an interaction, and (1 | …) represents the random intercept for a given error stratum as specified by the variables to the right of the vertical bar. Crucially, models were fit to *un-aggregated data* such that each participant contributed 6403 observations, one for each significantly tuned auditory-cortical surface node. A significant difference in the distribution of tBMF, sBMF, or behavioral-

neural intelligibility correlation (z) across cluster groups would provide evidence for the reliability of STRF patterns across the cortical surface nodes within a cluster group and across individual participants.

To test for a tradeoff in spectral and temporal resolution across significantly tuned STRFs, an additional model examining the node-by-node relationship between tBMF and sBMF was estimated using the following formula:

$$tBMF \sim hemi + sBMF + hemi{:}\,sBMF + (sBMF \mid sub) + (sBMF \mid hemi{:}\,sub),$$

where *tBMF* is the dependent variable, *sBMF* is a continuous covariate, and (*sBMF* | …) represents a random intercept plus random slope of sBMF within the error stratum specified by the variables to the right of the vertical bar. An analogous model with sBMF as the dependent variable and tBMF as the continuous covariate was also estimated. For all models, categorical independent variables were coded using a weighted-sum-to-zero approach such that the mean contrast weight was equal to zero after accounting for imbalances in the number of observations coming from each hemisphere and/or cortical region. Continuous covariates were centered on zero. Statistical significance of the fixed effects was evaluated by F-test with Satterthwaite approximation of the denominator degrees of freedom (Luke, 2017) and type III sums of squares. Observations for which the z-scored STRF peak magnitude was less than 2 were excluded from model fitting (4.2% of the total observations across participants).

3. Results

3.1 Cluster analysis reveals four groups of distinct STRF patterns in human auditory cortex

During fMRI scanning, participants listened to 400-450 “bubble-ized” sentences and made yes-no intelligibility judgments for each sentence by button press. A behavioral classification image (Fig. 3C) showing which STMs were important for intelligibility was calculated from the button-press data and neural classification images (STRFs) in the STM domain were calculated for each node (roughly a voxel-

like unit) in a standard-topology cortical surface model (essentially a group-level template brain that respects the gray matter folding patterns of each individual participant; see 2.6). Positively tuned STRFs (particular STMs produce an increased BOLD response, n = 6403) were considered separately from negatively tuned STRFs (particular STMs produce a decreased BOLD response, no significant nodes detected). Using these criteria, 100% of positively tuned surface nodes were located in the auditory cortex including the supratemporal plane and lateral superior temporal lobe. Of these, 58.8% were located in the left hemisphere; 97.3% were located in either Heschl's gyrus/sulcus, STG, STS, or posterior Sylvian cortex, based on the Destrieux (Destrieux et al., 2010) anatomical atlas in Freesurfer v5.3 (Fischl, 2012). For subsequent reporting of anatomical locations, the STG and STS were split into posterior and anterior segments by marking the midpoint of Heschl's gyrus and drawing a plane perpendicular to the Sylvian fissure.

To determine whether STRF patterns were organized within the auditory cortex, we applied an unsupervised Gaussian mixture model (GMM) clustering algorithm. The GMM essentially grouped STRFs according to their *functional* patterns within the STM domain. Crucially, clusters were not constrained to include STRFs from neighboring anatomical locations. The best-fitting GMM identified 18 STRF clusters within the auditory cortex (Fig. 3B). These clusters were then sorted to maximize the correlation between neighboring cluster-level STRFs and merged into four distinct "cluster groups" (Fig. 3A) by visual inspection. The cluster groups can be described as follows: Cluster Group 1 was located primarily within the supratemporal plane and posterior STG and responded to STMs throughout the entire range of the speech MPS including "pitch" and "formant" regions (see Fig. 1); Cluster Group 2 was located primarily in the posterior Sylvian region and medial supratemporal regions and responded to STMs with low spectral modulation rates and relatively high temporal modulation rates; Cluster Groups 3 and 4 were located primarily in the lateral STG and STS and responded to STMs important for intelligibility, based on comparison to the group-level behavioral classification image for intelligibility (Fig. 3C). Cluster Group 3 responded most reliably to temporal modulations about one-half octave higher than Cluster Group 4. To summarize, an STRF was derived empirically for each node. We then defined

18 clusters of nodes based on STRF similarity, each with its own cluster-level STRF. Finally, groups of clusters were formed by sorting the cluster-level STRFs to maximize the similarity between neighboring STRFs, and then manually identifying groups based on visually apparent, characteristic STRF profiles.

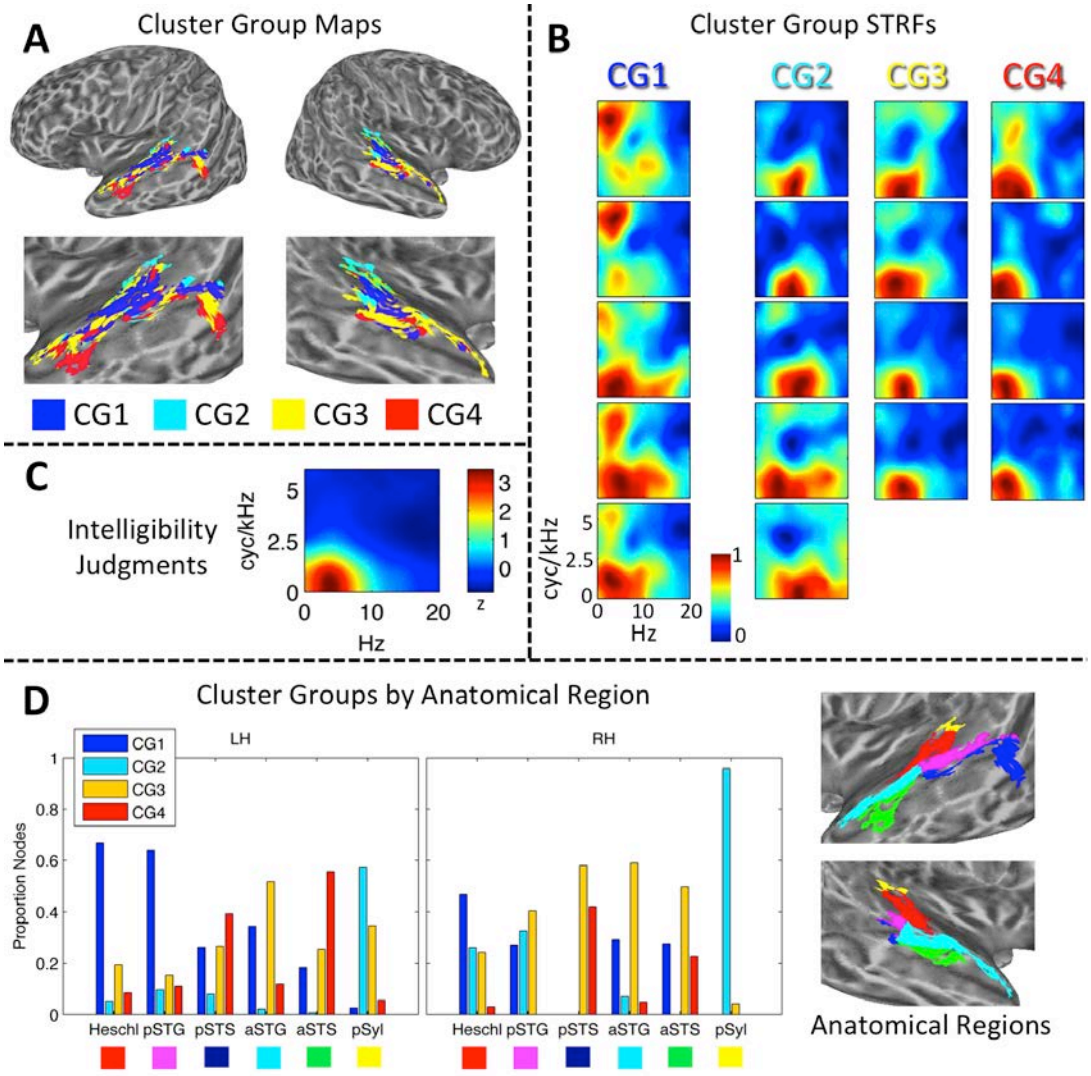

Figure 3. (A) Maps of STRF Cluster Groups in Auditory Cortex. *Cluster Groups are plotted by color on cortical surface renderings of the left and right hemispheres. Zoomed renderings of the temporal lobe are shown beneath whole-brain plots. Cluster Group 1 (CG1, blue) is located primarily in the supratemporal plane and posterior STG. Cluster Group 2 (CG2, cyan) is located primarily in medial supratemporal reigons. Cluster Groups 3 and 4 (CG3/4, yellow/red) are located primarily in the posterior and anterior STG/STS.* ***(B) STRF-Cluster Patterns.*** *For each of the 18 STRF clusters identified by GMM analysis, the cluster-average group-level (t-score) STRF is plotted. STRF magnitudes have been normalized to the range [0, 1]. Larger values are associated with STMs that produced relatively more BOLD activation. STRFs are organized by Cluster Group (CG1-4) in columns running from left to right. STRFs associated with CG1 respond to a broad range of STMs. STRFs associated with CG2 respond especially to high temporal modulation rates. STRFs associated with CG3/4 respond to STMs important for intelligibility (see C).* ***(C) Behavioral Classification Image for Intelligibility Judgments.*** *This plot is essentially a 'behavioral STRF', derived entirely from button-press responses (yes-no intelligibility judgments) rather than neural activity. The z-scored group-level behavioral classification image*

is shown. Larger values are associated with STMs that contribute relatively more to intelligibility. Temporal modulations from 2-7 Hz and spectral modulations less than 1 cyc/kHZ contribute maximally. ***(D) Distribution of Cluster Groups within Anatomically Defined Regions.*** *The proportion of cortical surface nodes belonging to CG1-4 is plotted for six anatomical regions of interest in the left (LH) and right (RH) hemispheres: Heschl = Heschl's gyrus/sulcus, pSTG/S = posterior STG/S, aSTG/S = anterior STG/S, pSyl = posterior Sylvian cortex. Colored boxes beneath region labels correspond to the colors of the anatomical regions plotted on zoomed cortical surface renderings at right. Only significantly tuned cortical surface nodes are labeled.*

The exact distribution of the four cluster groups within different anatomical regions, defined using the Destrieux atlas (Destrieux et al., 2010) as described above, is shown in Fig. 3D. A multinomial regression model showed significant main effects of hemisphere (analysis of deviance, type III SS; $\chi^2(3) = 174.6$, $p < 0.001$) and anatomical region ($\chi^2(15) = 1528.1$, $p < 0.001$), and a significant two-way interaction ($\chi^2(15) = 229.6$, $p < 0.001$). The simple main effect of anatomical region remained significant within each hemisphere (left: $\chi^2(15) = 1528.1$, $p < 0.001$; right: $\chi^2(15) = 1429.8$, $p < 0.001$). In both hemispheres, a majority of the nodes in Heschl's gyrus/sulcus belonged to Cluster Group 1 (Fig. 3D, 'Heschl'), while the other three cluster groups each accounted for about 5-20% of the nodes. A similar pattern was observed in the left hemisphere posterior STG (Fig. 3D, 'pSTG') with Cluster Group 1 dominating (> 60%), while in the right hemisphere Cluster Groups 1-3 each accounted for 20-40% of nodes. The posterior STS (Fig. 3D, 'pSTS') had a mixed profile in the left hemisphere, with Cluster Groups 3 and 4 together accounting 65% of the nodes, and Cluster Group 1 accounting for 25% of the nodes. Most nodes on the dorsal bank of the pSTS bordering the STG belonged to Cluster Groups 1 and 2, though a small group of nodes in the dorsal mid-posterior STS belonged to Cluster Group 4. A group of nodes in the ventral pSTS bordering the middle temporal gyrus belonged to Cluster Groups 3 and 4. In other words, the anatomically-defined pSTS in the left hemisphere seemed to encompass multiple functionally distinct subregions. In the right hemisphere, the pSTS was dominated by Cluster Groups 3 and 4 (~ 70%), but there were very few significantly tuned nodes in the right pSTS overall. The anterior STG (Fig. 3D, 'aSTG') in the left hemisphere was dominated by Cluster Group 3 (~ 55%), followed by Cluster Groups 1 (~ 35%) and 4 (~ 10%). A similar pattern was observed in the right aSTG. The anterior STS was dominated by Cluster Groups 3 and 4 in both hemispheres, though both hemispheres also had a

non-trivial contribution from Cluster Group 1. The posterior Sylvian region was dominated by Cluster Group 2 in both hemispheres (Fig. 3D, 'pSyl').

To summarize, Cluster Groups 1 and 2 were located in supratemporal regions and the posterodorsal lateral temporal lobe, and a transition to Cluster Groups 3 and 4 occurred moving laterally (e.g., STG to STS) and anteriorly (e.g., pSTG to aSTG). The posterior- and anterior-most regions of the ventral STS were dominated by Cluster Groups 3 and 4, and thus likely represent the highest levels of processing in the auditory cortex. It is important to note, again, that clusters and cluster groups were defined entirely on a functional basis – that is, the assignment of surface nodes to a cluster or cluster group was made entirely based on STRF patterns, with no restrictions based on anatomical location. The anatomical organization of clusters and cluster groups emerged naturally from their functional similarity. Though some amount of spatial correlation is expected due to both intrinsic and extrinsic spatial smoothing of the fMRI data (Hagler Jr. et al., 2006) – which increases the likelihood that neighboring nodes will demonstrate similar STRF profiles – the level of smoothness in our data (4 mm FWHM) does not account for the large-scale functional-anatomic organization of STRFs observed.

3.2 Effect of intelligibility on STRFs

As described in the Methods (2.8), it is possible that STRFs estimated from the full dataset (intelligible and unintelligible trials) were biased due to a global effect of intelligibility (see, e.g., Fig. 8A below). Therefore, we obtained STRFs with the effect of intelligibility removed ($STRF_{Unbiased}$) by calculating separate STRFs for intelligible ($STRF_{Intel}$) and unintelligible ($STRF_{Unintel}$) trials and averaging them. Crucially, the neural signal was normalized prior to this calculation to remove differences in the trial-by-trial mean and variance of intelligible and unintelligible trials, respectively (as is commonly done in pre-processing for MVPA; Coutanche, 2013). For each of the 18 STRF clusters identified by GMM clustering (3.1), we calculated cluster-level estimates of $STRF_{Unbiased}$. The overall patterns of cluster-level STRFs estimated from the full dataset (Fig. 3B) were maintained for cluster-level estimates of $STRF_{Unbiased}$ (Fig. 4). Indeed, the mean cluster-by-cluster correlation between the original

cluster-level STRFs and the cluster-level estimates of $STRF_{Unbiased}$ was r = 0.89 (± 0.01 SEM). However, the mean cluster-by-cluster correlation between the original cluster-level STRFs and the behavioral classification image for intelligibility (Fig. 3C) was r = 0.72 (± 0.05 SEM), while the mean value of this correlation was only r = 0.39 (± 0.05 SEM) for $STRF_{Unbiased}$, and the difference between these correlations was significant (t_{17} =14.1, p < 0.001). This suggests that the global effect of intelligibility was effectively removed from $STRF_{Unbiased}$. The primary difference between the original STRFs and the unbiased STRFs was a relatively enhanced representation of pitch-related STMs compared intelligibility-related STMs in the unbiased STRFs (compare Fig. 4 to Fig. 3B). Of note, $STRF_{Intel}$ and $STRF_{Unintel}$ were not examined individually because the stimulus space was not randomly sampled by bubbles within each class of trials (intelligible, unintelligible; Fig. 3C).

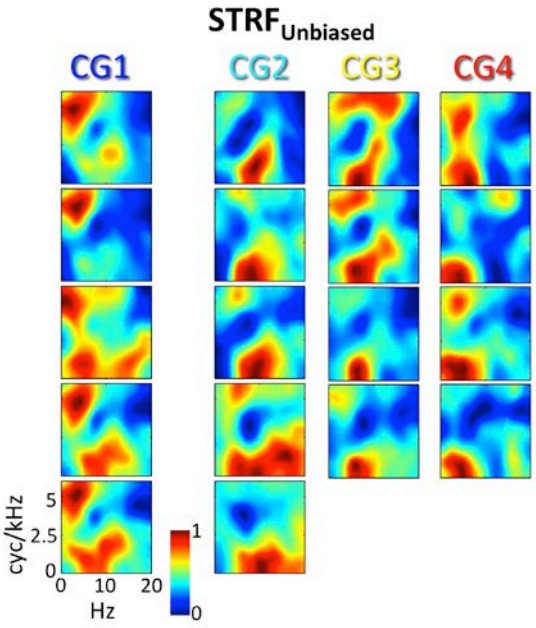

Figure 4. *Cluster level STRFs with the global effect of intelligibility removed ($STRF_{Unbiased}$). For each of the 18 STRF clusters identified by GMM analysis, the cluster-average group-level (t-score) $STRF_{Unbiased}$ is plotted. STRF magnitudes have been normalized to the range [0, 1]. Larger values are associated with STMs that produced relatively more BOLD activation. STRFs are organized by Cluster Group (CG1-4) in columns running from left to right. Compare to Fig. 3B.*

3.3 STRF peak modulation tuning

Cluster-level STRFs represent the mean response across all cortical surface nodes included in that cluster. Because clusters and cluster groups are defined based on the similarity of responses among their constituent nodes, we might expect little variability in STM tuning within a given cluster group. On the other hand, it is possible that STRFs within a cluster group are bound by some general feature but still vary in terms of the specific information encoded at each cortical node. To examine STRF variability within and between cluster groups, we estimated temporal and spectral peak modulation frequencies (tPMF and sPMF, respectively) from the group-level STRF of each positively tuned auditory-cortical surface node. The predominant tPMFs were between ~2-8 Hz and this range was distributed throughout the auditory cortex (Fig. 5A, tPMF). There was no clear evidence of spatially organized topographic gradients of tPMF within individual cortical regions. However, a clear pattern emerged that respected the defining characteristics of the cluster group STRFs as observed in Fig. 3B – i.e., a broad range of tPMFs was represented in Cluster Group 1, high-rate tPMFs (> 6 Hz) were preferentially represented in Cluster Group 2, and Cluster Groups 3 and 4 were distinguished by a relative shift toward higher tPMFs in Cluster Group 3 (Fig. 5B, top; see interquartile ranges). This pattern was similar across the left and right hemispheres. The range of spectral modulations represented was more restricted, with predominantly low sPMFs (< 1.5 cyc/kHz) found throughout the auditory cortex (Fig. 5A, sPMF). High sPMFs (4-6 cyc/kHz, vocal pitch range) were restricted primarily to the supratemporal plane and Heschl's gyrus. Indeed, the distribution of sPMFs (Fig 5B, bottom) was overwhelmingly weighted toward low sPMFs in Cluster Groups 2-4, but the distribution in Cluster Group 1 had considerably more weight on higher sPMFs (Fig. 5B, bottom; see interquartile ranges) . This again paralleled the pattern of cluster group STRFs observed in Fig. 3B. The left and right hemispheres were again similar. Overall, STRFs within each cluster group responded to a range of STMs, but the groups were clearly distinguishable based on the distributions of STMs represented.

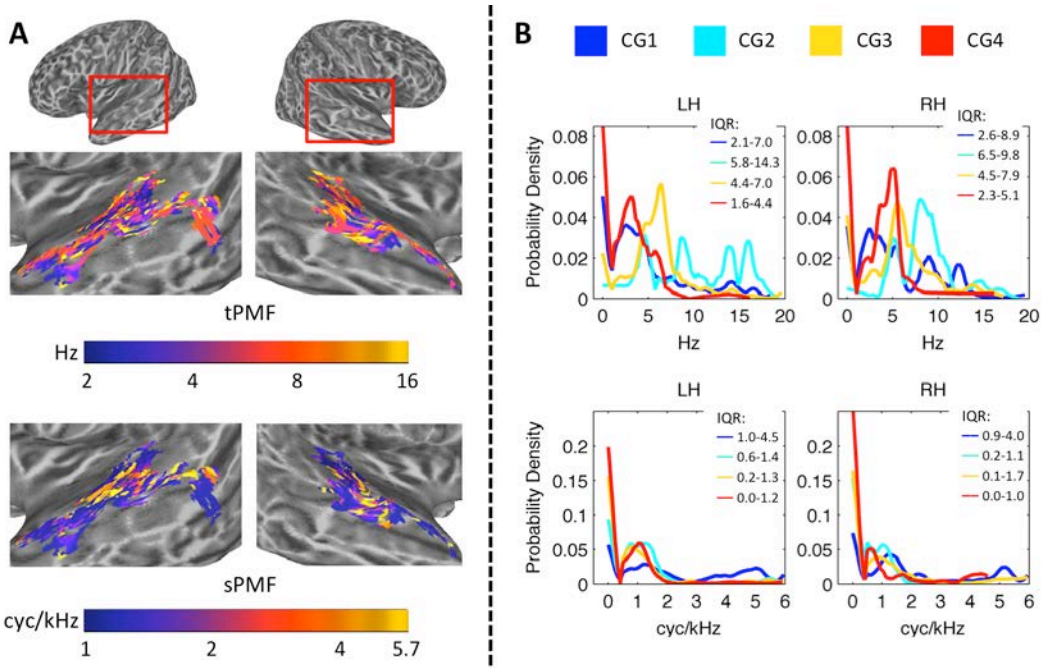

Figure 5. (A) Cortical Maps of Peak Modulation Frequencies. *Node-wise maps of temporal peak modulation frequency (tPMF, Hz) and spectral peak modulation frequency (sPMF, cyc/kHz) are displayed on inflated cortical surface renderings of the left and right temporal lobes. The renderings have been zoomed in as indicated by the red boxes at the top of the figure. Color scales are logarithmic.* ***(B) Probability Density of tPMF and sPMF Within Cluster Groups.*** *Empirical cumulative distribution functions (eCDFs; Kaplan-Meier method) for tPMF (Hz, top) and sPMF (cyc/kHz, bottom) were generated. Empirical probability density functions (ePDFs) were obtained by taking the derivative of the eCDFs. The ePDFs are plotted for each cluster group (colored lines, see legend) separately for the left (LH) and right (RH) hemispheres. The interquartile ranges (25th percentile - 75th percentile) of each distribution are indicated at the top right of each panel (IQR). The ordinate is the estimated probability density.*

In addition to qualitatively describing the distribution of group-level PMFs, we wanted to test quantitatively whether STRF peak modulation frequencies were reliably organized by cluster group across subjects and cortical surface nodes, and whether such organization varied by hemisphere. To accomplish this, we calculated the best temporal and spectral modulation frequencies (tBMF and sBMF, respectively) of individual-participant STRFs at each significantly tuned auditory-cortical surface node. For a given participant, the tBMF and sBMF at a given cortical surface node were selected jointly by identifying the 2D peak of the individual-participant STRF. The un-aggregated estimates of tBMF and sBMF across all surface nodes and all participants were transformed to an octave scale and entered as the dependent variables in separate linear mixed effects (LME) analyses with *hemisphere* (left, right), *cluster group* (1-4, as defined on the group data), and their interaction included as fixed factors, and *participant*

included as a random factor (see 2.12 for a comprehensive description of the random effects structure). For tBMF, there was a significant main effect of cluster group ($F_{3,26.7} = 9.98$, $p < 0.001$) but no significant main effect of hemisphere ($F_{1,9.0} = 2.29$, $p = 0.16$) and no significant interaction ($F_{3,25.1} = 1.49$, $p = 0.24$). Crucially, the pattern of cluster-group differences matched the characteristics of the aggregate STRFs for each cluster group (see 2.12 and 4.2 for a discussion of the importance of this finding): the highest tBMFs were found in Cluster Group 2, followed by Cluster Group 3, and Cluster Groups 1 and 4 had the lowest tBMFs (Figure 5A, top). The right hemisphere had higher tBMFs overall, particularly in Cluster Group 2, although this effect was not statistically significant. For sBMF, there was a significant main effect of cluster group ($F_{3,26.8} = 3.72$, $p < 0.05$) and a significant cluster group x hemisphere interaction ($F_{3,26.0} = 4.99$, $p < 0.01$), but no significant effect of hemisphere ($F_{1,9.1} = 0.55$, $p = 0.48$). The simple main effect of cluster group remained significant in the right hemisphere ($F_{3,26.7} = 5.37$, $p < 0.01$) and at the trend level in the left hemisphere ($F_{3,26.6} = 2.73$, $p = 0.06$). Again, the pattern of cluster-group differences matched the characteristics of the aggregate STRFs for each cluster group (see 2.12 and 4.2 for a discussion of importance): the highest sBMFs were found in Cluster Group 1 (Fig. 5A, bottom). In the left hemisphere, sBMFs decreased progressively from Cluster Group 2 to 4, but in the right hemisphere sBMFs were relatively higher in Cluster Groups 3 and 4 compared to Cluster Group 2.

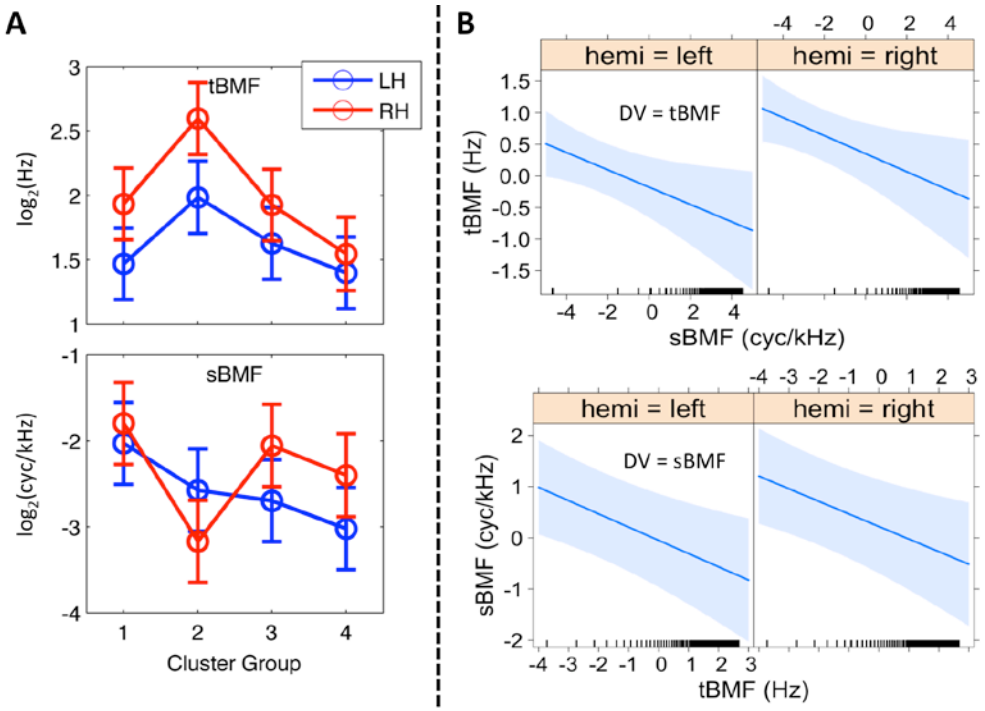

Figure 6. Linear Mixed Effects Models: Best Modulation Frequency. (A) Effect of Cluster Region. *The mean of the fitted values produced by the LME model for temporal (tBMF, top) and spectral (sBMF, bottom) best modulation frequencies (octave scale, ordinate) are plotted for Cluster Groups 1-4 (abscissa) in the left (blue) and right (red) hemispheres. Error bars reflect ± 1 SEM. Spectral BMFs are negative because nodes with an sBMF of 0, of which there were many, were set to 0.01 (-6.6 on the octave scale).* ***(B) Covariation between tBMF and sBMF.*** *Results of linear mixed effects regression of sBMF on tBMF (top) and tBMF on sBMF (bottom) by hemisphere are plotted as fitted lines (bold blue) with 95% confidence regions (light blue shading). BMFs have been mean-centered and transformed to the octave scale (i.e., axes show distance from the mean t/sBMF in octaves). Ticks above the abscissa indicate the values of the covariate at which data were actually observed.*

To test for a systematic relationship between tBMF and sBMF across cortical surface node locations and hemispheres, we conducted an additional LME analysis with tBMF as the dependent variable and fixed effects of *hemisphere*, sBMF (continuous covariate), and the interaction of sBMF by hemisphere. *Participant* was included as a random factor including a within-participant random slope term for sBMF (see 2.12 for a comprehensive description of the random effects structure). An analogous model with sBMF as the dependent variable and tBMF as the covariate was also estimated. BMFs were again transformed to an octave scale prior to analysis. Prominent models of speech processing advocated by Zatorre (Zatorre et al., 2002) and Poeppel (2003) predict: (i) a tradeoff in spectral and temporal resolution such that fast temporal modulations are encoded by neural populations that with poor spectral resolution and vice versa (i.e., a negative relationship between tBMF and sBMF); and (ii) a hemispheric

asymmetry in which greater temporal resolution is achieved by the left hemisphere (i.e., higher tBMFs are encoded in the left hemisphere and, by extension, higher sBMFs are encoded in the right hemisphere). In fact, with tBMF as the dependent variable, we observed a trend-level linear relation with sBMF in which a one-octave increase in sBMF predicted a 0.09 octave decrease in tBMF (beta = -0.09; $F_{1,9.0} = 4.83$, $p = 0.06$). We also found a trend-level effect of hemisphere in which tBMFs were, on average, 0.54 of an octave higher in the *right* hemisphere (beta = 0.54, $F_{1,9.0} = 4.84$, $p = 0.06$), but no significant interaction in the effect of sBMF by hemisphere ($F_{1,9.0} = 1.05$, $p = 0.33$). With sBMF as the dependent variable, there was a significant negative linear relation with tBMF (beta = -0.15, $F_{1,9.0} = 5.14$, $p < 0.05$), but there was no significant effect of hemisphere (beta = 0.25, $F_{1,9.0} = 0.87$, $p = 0.37$) and no interaction in the effect of tBMF by hemisphere ($F_{1,9.0} = 1.28$, $p = 0.29$). We then restricted the analysis to surface nodes belonging to Cluster Group 1 (Fig. 6B) because this cluster encompasses the lowest levels of auditory processing and represents the greatest range of BMFs. Within Cluster Group 1 (Fig. 6B), the negative linear relation between tBMF and sBMF was significant and became stronger (DV = tBMF: beta = -0.14, $F_{1,9.0} = 6.0$, $p < 0.05$; DV = sBMF: beta = -0.25, $F_{1,8.8} = 13.3$, $p < 0.01$). The effect of hemisphere (DV = tBMF: beta = 0.53, $F_{1,16.0} = 4.61$, $p < 0.05$; DV = sBMF: beta = 0.27, $F_{1,9.0} = 0.63$, $p = 0.45$) and the two-way interaction (DV = tBMF: $F_{1,10.7} = 0.04$, $p = 0.85$; DV = sBMF: $F_{1,8.5} = 0.01$, $p = 0.91$) patterned just as with the full dataset. Overall, we observed a significant negative relation between tBMF and sBMF as predicted, but this relation was not driven by hemispheric asymmetries. In fact, both tBMF and sBMF were, if anything, higher in the right hemisphere on average. Therefore, finer temporal resolution was not observed in the left hemisphere and no interhemispheric tradeoff in spectral and temporal resolution was revealed.

3.4 Processing specializations within cluster groups

The preceding sections 3.1 and 3.3 have shown that: (a) speech-driven STRFs in the auditory cortex are organized in a manner consistent with a hierarchical interpretation, as revealed by data-driven clustering; (b) cluster groups can be distinguished based on broad differences in their STM tuning; and (c)

significant variability in peak STM tuning is nonetheless present within each cluster group. This raises the possibility that subsets of STRFs within a given level of the cortical hierarchy are specialized for processing specific speech information. Indeed, Fig. 3B demonstrates that individual clusters within a given cluster group can vary considerably even while maintaining the defining characteristics of that cluster group. Here, we focus on three individual clusters from within the cluster groups plotted in Fig. 3B that are strongly suggestive of processing specializations within levels of the functionally defined cortical speech hierarchy.

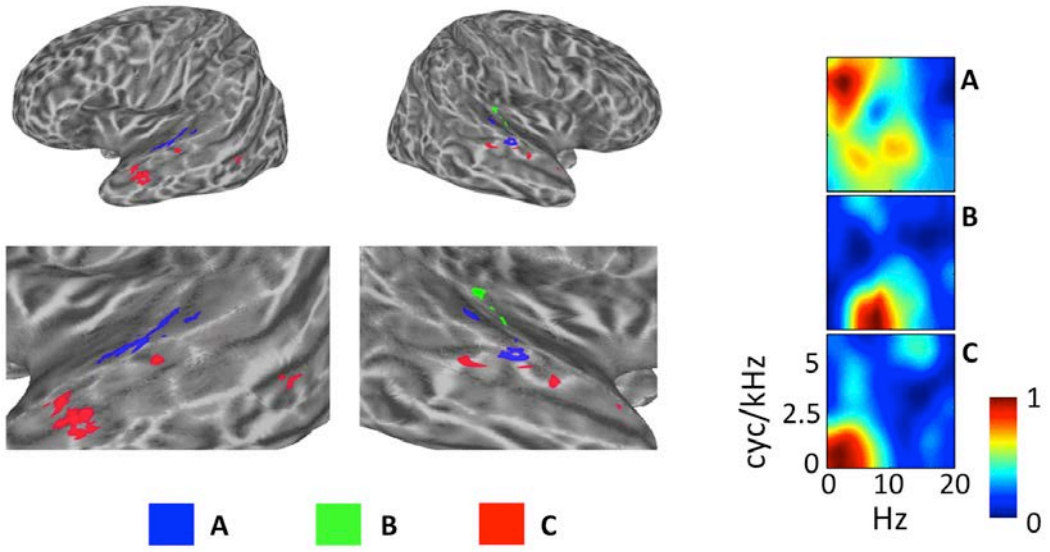

Figure 7. Evidence of STRF Specializations Within Cluster Groups*. Individual clusters of interest (A-C) are plotted on inflated cortical surface renderings of the left and right hemispheres (figure left). Zoomed surface renderings of the temporal lobes are shown beneath the whole-brain plots. The cluster-average group-level (t-score) STRFs are also plotted with magnitudes normalized to the range [0, 1] (figure right).* ***(A, blue)*** *From Cluster Group 1, this cluster on lateral Heschl's Gyrus and the neighboring STG responds best to STMs at high cyc/kHz ("pitch" STMs).* ***(B, Green)*** *From Cluster Group 2, this cluster located entirely in the right auditory cortex responds best to STMS at high temporal modulation rates (Hz).* ***(C, Red)*** *From Cluster Group 4, this cluster located prominently in the left anterior temporal lobe responds best to STMs important for intelligibility, particularly at very low temporal modulation rates (<= 3 Hz).*

Within Cluster Group 1, the defining feature of STRFs was a broad response spanning both pitch and formant regions of the speech MPS. However, certain clusters within Group 1 were tuned relatively more selectively to pitch STMs compared to formant STMs. The single cluster with the largest relative pitch response (i.e., largest relative t-score for STMs above 4 cyc/kHz) was located primarily in lateral Heschl's gyrus and the immediately neighboring STG, bilaterally (Fig. 7A). This region has been

implicated previously in human pitch processing (Griffiths, 2003). Within Cluster Group 2, the defining feature of STRFs was an increased response to high temporal modulation rates (> 6 Hz). One cluster among these showed a relatively selective response to these high rates. This cluster was located exclusively in regions of the right auditory cortex (Fig. 7B). This suggests some degree of hemispheric lateralization for temporal processing (Poeppel, 2003). Finally, within Cluster Group 4, the defining feature of STRFs was a very strong correlation with the behavioral classification image for intelligibility (per-cluster Pearson r of 0.83-0.96). However, among these, one cluster responded particularly well to low temporal modulation rates (essentially low pass in the temporal modulation domain). This cluster was located primarily in the anterior STS bilaterally, with particularly strong representation in the left hemisphere (Fig. 7C). The anterior temporal lobe has been implicated in several functions relevant to the analysis of intelligible speech at long time scales including prosodic, syntactic, and combinatorial semantic analysis (Humphries et al., 2005; Rogalsky and Hickok, 2008; Wilson et al., 2014). Overall, these results suggest that processing specializations occur within levels of the auditory cortical hierarchy. Crucially, the hierarchy is described presently in terms of neural tuning within an acoustic domain (i.e., the STM domain), and such tuning appears to underlie processing specializations within low (e.g. Cluster Groups 1 and 2) and high (e.g., Cluster Group 4) levels of the hierarchy.

3.5 Cortical Maps of Speech Intelligibility

To determine the extent to which different brain regions were involved in processing intelligible speech, correlations between behavioral classification images and neural STRFs were calculated at each cortical surface node separately for each participant. These correlations described the degree to which a cortical surface node was activated most strongly when the speech information most important to behavioral intelligibility was present in the stimulus. In addition, to facilitate direct comparison to previous studies using standard subtraction contrast methods, trials were sorted according to button press responses in the yes-no intelligibility judgment task (i.e., into “intelligible” and “unintelligible” trials) and a mean activation contrast value (intelligible vs. unintelligible) was calculated at each node. Second-level

correlation and contrast intelligibility maps were calculated (one-sample t-test; corrected $p < 0.05$) and compared.

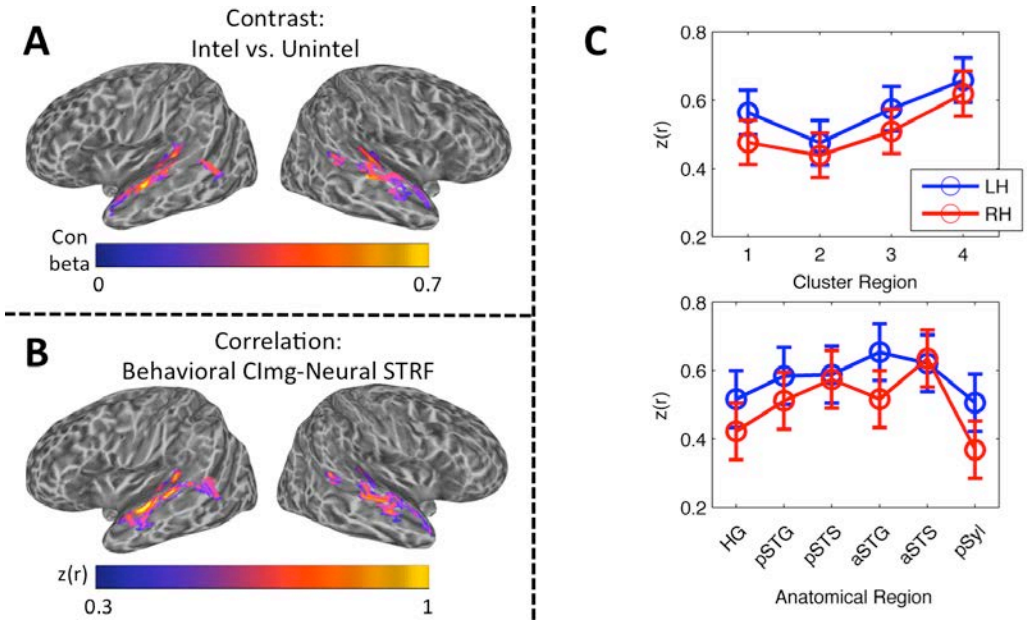

Figure 8. (A) Contrast Map of Speech Intelligibility. *The group mean contrast beta (intelligible trials vs. unintelligible trials) is plotted on cortical surface renderings of the left and right hemispheres. Whole-brain analysis, wild-bootstrap-corrected $p < 0.05$.* ***(B) Correlation Map of Speech Intelligibility.*** *The group mean Fisher z-transformed correlation, z(r), between behavioral classification images for intelligibility and neural STRFs is plotted on cortical surface renderings of the left and right hemispheres. Whole-brain analysis, wild-bootstrap-corrected $p < 0.05$.* ***(C) Linear Mixed Effects Analysis of Intelligibility Correlation Values.*** *The mean of the LME-fitted values of the Fisher z-transformed correlation, z(r), between behavioral classification images for intelligibility and neural STRFs is plotted across cluster regions (top) and anatomical regions (bottom) in the left (blue) and right (red) hemispheres. Error bars reflect ± 1 SEM.*

Cortical maps of intelligibility based on activation contrast (Fig. 8A) and STRF-based correlation (Fig. 8B) were broadly similar. Each map primarily emphasized the STG/STS and Heschl's sulcus bilaterally. In the contrast maps, the largest differences were observed in Heschl's sulcus and the mid-anterior STG/S, and the magnitude of activation differences was similar across the left and right hemispheres. In the correlation maps, the largest correlations were observed in Heschl's sulcus, all along the STG, and in the anterior STS; the magnitude of correlations was larger in the left hemisphere. There were 5465 significant nodes in the contrast map (48.0% left hemisphere) and 5225 significant nodes in the correlation map (54.1% left hemisphere). The correlation map overlapped 63.4% with the contrast map. The contrast map overlapped 65.6%, and the correlation map overlapped 74.6%, with the map of

significantly tuned STRFs in the auditory cortex. Neither the contrast nor the correlation map contained nodes outside the temporal lobe.

In general, the correlation procedure yields a value that is more easily interpretable than the contrast procedure – i.e., rather than a mean difference in activation magnitude between two conditions, the correlation map shows the degree to which a cortical surface node responds selectively to acoustic patterns that are relevant to behavioral intelligibility. Thus, to examine how the representation of intelligible speech relates to STRF-cluster-based vs. anatomically-based organization of the auditory cortex, correlations measured at individual cortical surface nodes within individual participants were entered as the dependent variable in two second-level LME models: (1) a cluster group model with *hemisphere* (left, right) and *cluster group* (1-4, as defined on the group data), and their interaction as fixed effects, and *participant* as a random effect (see 2.12 for comprehensive random effects structure); and (2) an anatomical model with *hemisphere* (left, right) and *anatomical region* (Heschl's gyrus/sulcus, posterior STG, posterior STS, anterior STG, anterior STS, and posterior Sylvian cortex), and their interaction as fixed effects, and *participant* as a random effect (see 2.12 for comprehensive random effects structure). The LME analysis was restricted to auditory-cortical nodes with that were significantly tuned in the group-level STRF analysis. Correlation values were not aggregated across cortical surface nodes within a given region.

For the cluster group model, there was a significant main effect of cluster group ($F_{3,27.0} = 4.55$, $p < 0.05$), but no significant main effect of hemisphere ($F_{1,9.0} = 1.09$, $p = 0.32$) and no significant interaction ($F_{3,26.6} = 0.19$, $p = 0.90$). Like the LMEs carried out on tBMF and sBMF, the pattern of cluster-group differences matched the characteristics of the aggregate STRFs for each cluster group (see 2.12 and 4.2 for a discussion of importance): intelligibility correlations were largest in Cluster Group 4, followed by Cluster Groups 3, 1 and 2 in that order (Fig. 8C, top). There appeared to be a trend toward larger correlations in the left hemisphere, although this effect was not statistically significant. For the anatomical model, there was no significant effect of region ($F_{5,44.9} = 1.71$, $p = 0.15$), hemisphere ($F_{1,9.0} = 1.20$, $p = 0.30$) or their interaction ($F_{5,44.8} = 0.69$, $p = 0.63$). The pattern of correlations across anatomical

regions (Fig. 8C, bottom) should be interpreted with caution due to the lack of a significant effect. Indeed, organization by cluster group appeared to provide a better characterization of the data than organization by anatomical region. A post-hoc LME model containing fixed and random effects of both cluster group and anatomical region showed that removal of the fixed effect of cluster group had a greater effect on model fit (likelihood ratio = 9.13) than removal of the fixed effect of anatomical region (likelihood ratio = 6.14), despite the fixed effect of cluster group accounting for fewer degrees of freedom (3 df) than the fixed effect of anatomical region (5 df). This demonstrates that, although the cluster groups are significantly associated with particular anatomical regions (Fig. 3D), this association is not perfect. In other words, certain anatomical regions (e.g., the left pSTS) are mixed with respect to the distribution of cluster groups, and these regions therefore encompass multiple hierarchical levels of processing or, at least, process a broader range of speech information.

4. Discussion

In the present fMRI study, a classification image procedure ("bubbles") was used to estimate speech-driven STRFs in the modulation-power-spectrum domain from single-trial BOLD response amplitudes. These STRFs were estimated for a group of healthy, normal-hearing participants at each node in a standard-topology cortical surface model. Data-driven clustering was used to define groups of STRFs with similar response properties. The clustering procedure recovered an organization consistent with hierarchical interpretations of cortical speech processing. Specifically, STRF clusters representing a broad range of spectrotemporal features were located in early auditory regions of the supratemporal plane, while STRF clusters representing the spectrotemporal features most important for intelligibility were located in later auditory regions of the lateral temporal lobe. Although clusters were defined at the group-level using a "t-score" approach, an LME analysis of individual-participant STRF scalar metrics (Joosten and Neri, 2012), e.g., best modulation frequency, showed that the STRF-tuning patterns of group-defined clusters were reliable across individual participants and cortical surface nodes.

The notion of a cortical hierarchy for processing auditory speech is not new, but our method is unique in that it reveals precisely what acoustic information is processed within and between levels of the hierarchy. For example, we find that among later (intelligibility-focused) regions, faster temporal modulations corresponding roughly to syllable or phoneme length units are processed primarily in the anterior STG and posterior STS, while slower temporal modulations corresponding roughly to suprasegmental units are processed primarily in the anterior STS. Moreover, we find specializations for processing pitch in lateral Heschl's gyrus, and at least a qualitative hemispheric preference for processing fast temporal modulation rates in the right hemisphere and slow temporal modulations essential for intelligibility in the left hemisphere. Overall, these results provide a much more nuanced characterization of the cortical speech hierarchy compared to existing data. In the sections below (4.1-4.5), we discuss these and other major findings and their theoretical implications.

4.1 Hierarchical organization of speech-driven STRFs in the auditory cortex

Data-driven analysis identified four groups of STRF clusters defined by their similar within-group functional properties. We envision these cluster groups as capturing different levels of processing within a feed-forward cortical speech hierarchy that progresses from detailed spectrotemporal processing (Cluster Groups 1 and 2) to more abstracted processing of acoustic patterns specific to speech (Cluster Groups 3 and 4). As we will describe, different STRF patterns corresponding to different acoustic (vocal harmonics vs. formants; transient vs. sustained events) or linguistic (phonemes vs. syllables/words) speech cues are represented separately within different hierarchical levels.

The lowest-level group of STRF clusters, Cluster Group 1, represented a broad range of spectrotemporal modulations spanning both the "pitch" and "formant" regions of the speech MPS (Fig. 1A). Cluster Group 1 was located primarily in Heschl's gyrus/sulcus and the immediately neighboring posterior STG (Fig. 3A/D). In the aggregate, the STRFs in Cluster Group 1 appeared to behave as a simple "energy detector" (Fig. 3B, CG1), i.e., they responded to modulation energy in the speech signal regardless of the particular pattern presented to the listener (see also, Santoro et al., 2017). However,

examination of node-wise peak modulation frequencies revealed that STRFs in Cluster Group 1 were individually tuned to particular temporal and spectral modulation rates spanning a wide range (Fig. 5B). Moreover, certain individual STRF clusters within Cluster Group 1 were tuned relatively more selectively to the pitch or formant regions of the MPS. Therefore, the data suggest that the broad representation of spectrotemporal features in Cluster Group 1 reflects the integrated activity of neuronal subpopulations tuned to relatively narrow patterns within the speech MPS. Indeed, bandpass modulation tuning across a range of best modulation frequencies has been observed in early auditory-cortical regions in a number of species, particularly for temporal modulations (Bieser and Müller-Preuss, 1996; Liang et al., 2002; Miller et al., 2002; Schreiner and Urbas, 1988; Scott et al., 2011; Woolley et al., 2005). Thus, our finding is consistent with the view that early auditory cortex is essentially an STM filterbank (Chi et al., 1999; Chi et al., 2005).

The next, perhaps intermediate, group of STRF clusters was Cluster Group 2, which was defined primarily by an increased response to high temporal modulation rates (Fig. 3B, CG2; Fig. 5B). Cluster Group 2 was located primarily in posteromedial aspects of the supratemporal plane bordering the outer edges of Cluster Group 1 (Fig. 3A). The vast majority of cortical surface nodes in the posterior Sylvian region were contained in Cluster Group 2 (Fig. 3D). Both human and animal data support the notion of a posterior to anterior temporal processing gradient in which cells with the shortest temporal integration windows are located in the posterior temporal lobe (Bendor and Wang, 2008a; Hullett et al., 2016). The posterior Sylvian region, in particular, has been implicated in the processing of temporal order and in sound sequencing (Bernasconi et al., 2010; Bernasconi et al., 2011; Hickok et al., 2011), potentially necessitating a need to represent speech features on a relatively short time scale. This may be particularly important for extracting information relevant for accessing the phonological form of the speech signal at a relatively fine temporal scale.

Cluster Groups 3 and 4 correspond to the highest level of the presumed cortical speech hierarchy. The STRF clusters in these groups responded selectively to the MPS region most important for speech intelligibility (Fig. 3B/C). Cluster Groups 3 and 4 accounted for a majority of the cortical surface nodes

in the posterior STS, anterior STG, and anterior STS (Fig. 3D), regions implicated previously in high-level processing of intelligible speech (Evans et al., 2014; Narain et al., 2003; Okada et al., 2010; Scott et al., 2000). Functionally, Cluster Groups 3 and 4 were distinguished by a fairly substantial difference in STRF tuning in the left hemisphere – namely, the temporal modulation tuning of STRFs in Cluster Group 3 was shifted up roughly an octave compared to STRFs in Cluster Group 4 (median = 3.0 vs. 6.1, respectively; interquartile range = 1.6 - 4.4 vs. 4.4 - 7.0, respectively); the shift was also present in the right hemisphere but was somewhat more modest (median: 4.0 vs. 5.8, interquartile range: 2.3 – 5.1 vs. 4.4 – 8.0). Translated to speech units, the shift in the left hemisphere corresponds roughly to the timescale of phonemes versus syllables/words or even short phrases. It is tempting to hypothesize a correspondence between Cluster Group 3 STRFs as playing a role in processing shorter duration phonological information on the order of phonemes to syllables and a correspondence between Cluster Group 4 and processing higher-order linguistic chunks from words to phrases. Anatomically, in the left hemisphere Cluster Group 3 was represented most prominently in the anterior dorsal STG and posterior STS while Cluster Group 4 was represented most prominently in the anterior STS and ventral posterior STS. In the right hemisphere, Cluster Group 3 was represented in the anterior and posterior dorsal and lateral STG; Cluster Group 4 was represented in the both anterior and posterior STS. This result may account for differences in existing hierarchical models of speech processing that place relatively more emphasis on posterior versus anterior temporal lobe regions, or vice versa (Bernstein and Liebenthal, 2014; Hickok and Poeppel, 2007; Rauschecker and Scott, 2009). That is, both anterior and posterior STG/STS regions play an important role in extracting intelligible information from speech, but at different levels of analysis. The STRFs in Cluster Group 4 correlated most strongly with the group-level behavioral classification image for intelligibility and were located most prominently in the anterior temporal lobe and the ventral-most aspect of the left posterior STS (i.e., abutting the middle temporal gyrus); this makes sense if Cluster Group 4 represents the top of the processing hierarchy either acoustically or linguistically.

In addition to evidence suggesting a hierarchical organization of cluster groups, there was evidence for processing specializations (i.e., differences in STRF patterns) within each cluster group (Fig. 7). As mentioned, STRFs in Cluster Group 1 showed varying degrees of selectivity for spectrotemporal modulations in the pitch region of the MPS, with a cluster in lateral Heschl's gyrus/STG demonstrating maximum selectivity for pitch. Both animal and human studies have identified a similar "pitch region" just lateral to primary auditory cortex (Bendor and Wang, 2006; Griffiths, 2003; Penagos et al., 2004). In Cluster Group 2, STRFs showed varying amounts of selectivity for high temporal modulation rates, with a cluster in the right auditory cortex demonstrating this selectivity most clearly. In general, we observed a tendency for right hemisphere regions to respond to slightly higher temporal modulation rates (Fig. 6A, top).

We cannot be certain whether the observed STRF properties reflect sensitivity to particular acoustic patterns or the higher-level processing of information extracted from those patterns. For example, we would expect the operation of linguistic computations (e.g., syntax and semantics) to be correlated with the acoustic patterns feeding into those mechanisms. This is particularly true in the context of the bubbles technique where linguistic processes fail to engage when the acoustic patterns that support intelligibility are filtered from the signal. A recent study (de Heer et al., 2017) compared fMRI-encoding models based on spectrotemporal, articulartory-phonological, and semantic features using a variance partitioning scheme to estimate which model best accounted for activation patterns during continuous listening to narrative speech. In fact, the spectrotemporal model best accounted for activation patterns only in a restricted region of the early auditory cortex, which would be circumscribed by Cluster Group 1 as measured here. Activity in later regions comparable to Cluster Groups 3 and 4 was best described by the articulatory-phonological and semantic models. This finding is in agreement with Schönwiesner and Zatorre's (2009) data, which show that only regions in the primary and secondary auditory cortices respond to synthetic STM stimuli. However, there is reason to believe that, even at high levels of processing, tuning to particular spectrotemporal patterns remains an important organizational principle of the auditory cortex. Recent work by Chang and colleagues shows that high-level information

about phonetic features, pitch contour, and talker identity is encoded and intermixed within neuronal populations of the STG that are also topographically organized in terms of spectrotemporal modulation tuning (Hullett et al., 2016; Mesgarani et al., 2014; Tang et al., 2017). Work by Poeppel and colleagues shows that temporal lobe regions up to and including the STS respond to speech-specific temporal patterns even when stimuli are unintelligible (Boemio et al., 2005; Overath et al., 2015). In the present study, we observed a range of spectrotemporal modulation tuning patterns even at the highest levels of the presumed auditory cortical hierarchy (Cluster Groups 3 and 4; Figs. 3B/5B), which suggests that a partially abstracted "acoustic trace" is maintained at these later processing stages. We also found similar STRF patterns after correcting for global effects of intelligibility (Fig. 4). Together, these findings suggest that the organization observed here for speech may also apply to cortical-acoustic analysis of other (non-speech) sounds, a notion supported by recent fMRI encoding studies showing similar patterns of organization – broad responses in early auditory regions, sensitivity to specific features in STG/S, pitch regions in lateral HG, etc. – for non-speech natural sounds (De Angelis et al., 2017; Moerel et al., 2012; Santoro et al., 2017). However, Santoro et al. (2017) show that STRFs estimated from a wide range of natural sounds including speech are different when speech stimuli are left out of the STRF computation. This suggests that cortical-acoustic analyses shift when processing occurs in a linguistic context, but even if one were inclined to argue that higher-order STRFs were driven by linguistic computation, the fact that particular ranges of STMs are represented in different groups of STRFs would provide rather compelling evidence regarding the level of linguistic computation being carried out by each group.

4.2 Consistency of STRF Cluster Organization Across Participants and Surface Nodes

Clustering analysis was performed on auditory-cortical STRFs after averaging across participants at each node of a standard topology cortical surface model. Moreover, the STRFs representing each cluster (Fig. 3) reflect an aggregate pattern across many constituent cortical surface nodes. Some degree of inter-participant and inter-node variability should be expected, so it was important to determine whether the group-level data were representative of the STRF patterns observed at individual cortical

surface nodes in individual participants. Neri and colleagues (Joosten and Neri, 2012; Neri, 2010; Neri and Levi, 2008) caution against drawing strong conclusions from qualitative inspection of aggregate classification images (e.g., STRFs) because, in cases where significant individual variability is present, the aggregate patterns may not be representative of any given individual. Rather, they suggest conclusions should be drawn from quantitative analysis of individual classification images, namely by extracting a scalar metric that summarizes the shape of the classification image and analyzing the metric statistically. This approach is what motivated our decision to analyze STRFs at the group level by calculating a "t-score" version of the STRF at each cortical surface node. The t-scoring procedure allowed us to identify aggregate STRFs for which at least a subset of STRF features was consistent across participants. Only such STRFs were entered in the GMM clustering analysis.

However, the t-scoring procedure alone did not ensure that wholesale STRF patterns observed at the cluster-group level (Fig. 3) would reflect the underlying STRF patterns in individual participants or surface nodes. Therefore, we adopted Neri and colleagues' approach of extracting scalar metrics from individual STRFs and analyzing them quantitatively. Linear mixed effects (LME) models were constructed with individual-participant STRF scalar metrics – tBMF, sBMF, and behavioral-neural intelligibility correlation – as the dependent variables to determine if these metrics would pattern according to the definitional features of Cluster Groups 1-4 as defined at the group level. In fact, all three scalar metrics behaved as predicted: tBMF was highest in Cluster Group 2 (Fig. 5A), sBMF was highest in Cluster Group 1 (Fig. 5A), and the behavioral-neural intelligibility correlation was highest in Cluster Group 4 (Fig. 8C). Cluster group 3 was intermediate in terms of tBMF and the behavioral-neural intelligibility correlation. The difference across cluster groups was statistically significant in all cases. Crucially, the inputs to the LME models were the *un-aggregated* scalar metrics from every auditory-cortical node across every participant, and participant-level variance was explicitly accounted for by including the appropriate random effects terms in the LME models (2.12). Thus, we can be reasonably confident in the reliability of STRF patterns across participants and cortical surface nodes. While this does not guarantee or even suggest that similar clusters of STRFs would be identified if we analyzed the

data of each participant separately, it provides evidence that the large-scale STRF organization determined from the group data did not arise spuriously from the aggregation of highly-variable and/or disorganized STRFs across participants and surface nodes. However, since STRF clusters were identified from the same data used in LME modeling, we cannot make strong claims regarding the likelihood of these findings to generalize beyond the present study.

To examine whether analyzing each participant separately would produce qualitatively similar results to analysis of the group data, we repeated the clustering analysis on each individual participant using the cluster model that provided the best description of the group data (K = 18 clusters, "VVV" covariance structure; see 2.10). Since cluster labels are arbitrarily assigned, the individual participant clusters were relabeled to maximize the cluster-by-cluster correlation with the group data (Kuhn, 1955), and then partitioned into four cluster groups using the same mapping as for the group data. Crucially, this procedure did not ensure that a similar number of cortical surface nodes would be assigned to each cluster group in individual participants compared to the group, nor did it ensure that the cluster groups would be assigned to similar anatomical regions in individual participants compared to the group. Yet, we found a remarkably similar anatomical distribution of Cluster Groups 1-4 when defined on individual participant data (Fig. 9) compared to the group data (Fig. 3). The average node-by-node percent agreement of cluster group assignments when comparing individual participants to the group was 49.1 % (± 2.0 % SEM).

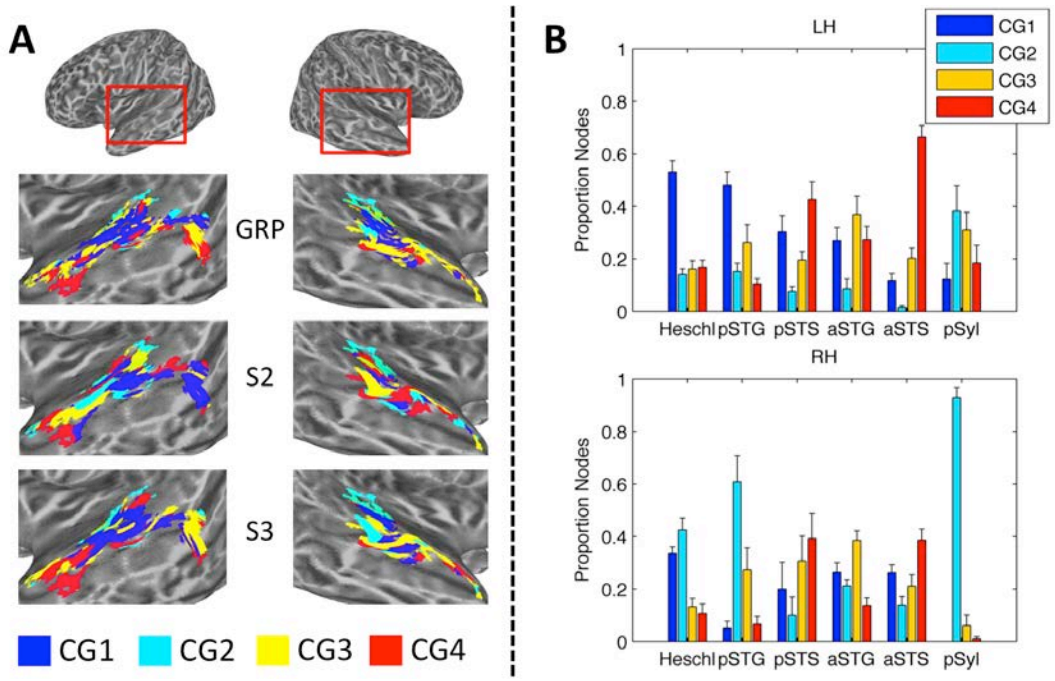

Figure 9. ***(A) Cluster-Group Maps at the Group Level and in Representative Individual Participants.*** *Cluster Groups are plotted by color on cortical surface renderings of the left and right hemispheres. Separate maps are shown for the group-level data (GRP), and for the two individual participants with the lowest (S2) and highest (S3) percent agreement with the group.* ***(B) Distribution of Individual-Participant Cluster Groups within Anatomically Defined Regions.*** *The across-participant average proportion of cortical surface nodes belonging to Cluster Group 1-4 is plotted for six anatomical regions of interest in the left (LH) and right (RH) hemispheres: Heschl = Heschl's gyrus/sulcus, pSTG/S = posterior STG/S, aSTG/S = anterior STG/S, pSyl = posterior Sylvian cortex. Error bars = ± 1 SEM. Compare to Fig. 3D for group-level distributions.*

4.3 Cortical Maps of STRF Peak Modulation Frequencies

We observed a broad trend in which STRFs tuned to relatively fast temporal modulation rates (6-10 Hz) were located in posteromedial auditory regions (Cluster Group 2), while STRFs tuned to relatively slow temporal modulation rates (1-4 Hz) were located in anterolateral auditory regions (Cluster Group 4). However, examination of node-wise cortical maps of tPMF (Fig. 5A, top) did not reveal an organized high-to-low gradient moving from posterior to anterior regions. Rather, there was a range of tPMFs within each cluster group (Fig. 5B, top), with hierarchical distinctions suggested by differing *distributions* of temporal modulations by cluster group. The lack of a within-hierarchical-level spatial gradient stands in apparent contrast to a previous finding (Barton et al., 2012), although in that work identification of topographic mapping was performed for core and belt subfields using individual-participant data and

multiple maps were identified. We simply may not have the spatial resolution to resolve such internal structure.

In contrast to the temporal modulation maps, we found a clear organization for spectral modulations in which high spectral modulation rates (4-6 cyc/kHz) were represented most prominently in early regions of the supratemporal plane, and later auditory regions became increasingly "low-pass" (< 2 cyc/kHz) in their representation of spectral modulations (Fig. 5B, bottom). This pattern was essentially born out in node-wise maps of sBMF (Fig. 5A, bottom), although there was evidence for some additional representation of high spectral modulation rates in the mid and posterior STG.

Together, these results agree strongly with data based on responses to synthetic STM sounds (Schönwiesner & Zatorre, 2009), and they broadly comport with an existing model of auditory cortical organization based on non-human primate data in which temporal integration windows increase along a gradient from caudal to rostral auditory areas, and spectral integration windows increase along a gradient from medial to lateral auditory areas (Bendor and Wang, 2008b). A similar organization was revealed by a recent fMRI study of human natural sound processing, which, using a modulation-based encoding model, characterized the organization of auditory cortex in terms of a tradeoff between spectral and temporal resolution – namely, regions posterior and lateral to Heschl's gyrus encoded relatively coarse spectral information (low spectral modulation rates) and fine temporal information (high temporal modulation rates), while regions located on and immediately anteroventral to Heschl's gyrus encoded fine spectral information with and coarse temporal information (Santoro et al., 2014). Indeed, we found a strong negative correlation between the best temporal modulation rate and best spectral modulation rate of cortical nodes in these early auditory regions (Fig. 6; but, see 4.4 below). A recent human ECoG study derived speech-driven STRFs from electrodes placed throughout the STG and found that electrodes in the posterior STG responded best to high temporal modulation rates and low spectral modulation rates, while electrodes in the anterior STG responded best to high spectral modulation rates and low temporal modulation rates (Hullett et al., 2016).

4.4 Relation to Hemispheric Lateralization of Cortical Speech Processing: The Spectral-Temporal and Asymmetric Sampling in Time Models

There is a fairly entrenched notion that speech is processed preferentially in the left hemisphere while pitch and prosody (e.g., music) are processed preferentially in the right hemisphere (cf., Price et al., 2005). Zatorre and colleagues (Zatorre and Belin, 2001; Zatorre et al., 2002) suggest that such asymmetries arise from differences in early spectrotemporal processing of sound features in the auditory cortex. Their spectral-temporal model asserts that temporal features are processed predominantly in the left hemisphere, while spectral features are processed predominantly in the right hemisphere. A related speech-centric model – Poeppel's asymmetric sampling in time (AST) model (Poeppel, 2003) – suggests that hemispheric asymmetries arise as a consequence of the way auditory representations are analyzed in the time domain. According to AST, the left hemisphere extracts information preferentially from a short time window (25-50 ms or 20-40 Hz), lending itself to analysis on a scale appropriate for detecting rapid formant transitions, while the right hemisphere extracts information from a longer time window (125-300 ms or 3-8 Hz), lending itself to analysis on a syllabic scale. Poeppel further suggests that right hemisphere specializations for processing spectral information can be explained in terms of that hemisphere's longer analysis window – that is, greater spectral resolution is achieved with an increasing integration time constant. Thus, both the spectral-temporal and AST models provide two predictions in the context of the present study: (1) speech-driven STRFs tuned to high temporal modulation rates (fine temporal resolution) will also be tuned to low spectral modulation rates (poor spectral resolution), and vice versa; (2) a greater preponderance of STRFs tuned to high temporal modulation rates will be located in the left hemisphere and STRFs tuned to low temporal modulation rates (and high spectral modulation rates) will be located in the right hemisphere.

We tested these predictions directly by examining the relation between best temporal and best spectral modulation frequency (tBMF and sBMF, respectively) across all significantly tuned auditory-cortical surface nodes and across hemispheres. In fact, there was a significant negative linear relation

(i.e., in the predicted direction) between tBMF and sBMF within Cluster Group 1 (Fig. 6B), and there was no difference in the strength of this relation between hemispheres. There was also a significant main effect of hemisphere on tBMF, but it was in the opposite direction of that predicted by lateralization models – namely, higher temporal modulation rates tended to be represented in the *right* hemisphere. Therefore, while tBMF and sBMF are related within each hemisphere, we find, if anything, the opposite between-hemisphere relation as that predicted by lateralization models.

A rather simple explanation appears to account for the significant (within-hemisphere) relation between tBMF and sBMF within Cluster Group 1. Recall that Cluster Group 1 responds essentially like an STM filterbank, and it is the only cluster group that responds significantly to high spectral modulation rates (pitch). Figure 1A (left panel) displays the boundary containing 80% of the power in the speech modulation spectrum. This boundary reveals that speech energy at high spectral modulation rates tends to be located at low temporal modulation rates and vice versa. Indeed, this relation holds for many animal vocalizations including human speech (Elliott and Theunissen, 2009). We therefore suggest that the natural modulation statistics of speech are reflected straightforwardly in the outputs of spectrotemporal modulation filters in Cluster Group 1. Zatorre and colleagues (Zatorre et al., 2002) describe an 'acoustic uncertainty principle' in which there is a tradeoff between the precision that can be achieved in the time and frequency domains when analyzing an acoustic event. We cannot rule out the notion that this spectral-temporal tradeoff reflects an intrinsic organizing principle of the auditory cortex, in which case it is possible that the human vocal apparatus and its associated motor control circuits have adapted to shape speech acoustics to match this pattern (Giraud et al., 2007).

Regarding the possible right hemisphere preference for processing faster temporal modulation rates, it is unclear whether this might reflect a right hemisphere specialization for processing (slightly) more fine-grained temporal features, or a left hemisphere specialization for processing high-level components of intelligible speech (e.g. words, phrases; Peelle, 2012; Specht, 2013), which tend to come across at relatively slower rates. While we found clear evidence for specialized processing of slow rates in a STRF cluster from Cluster Group 4 (Fig. 7C), this cluster was localized to the STS bilaterally. In

fact, the broad organization of STRF clusters was remarkably similar across the hemispheres (Fig. 3A), as were node-wise maps of STRF peak modulation frequencies (Fig. 5A). Therefore, we assert that our data do not support the existence of broad hemispheric differences in spectrotemporal processing. We should note that we did not test for hemispheric differences in temporal processing above 20 Hz. This is because there is very little speech modulation energy above 20 Hz (Fig. 1A), so the bubbles technique was unlikely to identify consistent responses at such high modulation rates. However, one of the critical windows in AST is 20-40 Hz, so we necessarily failed to detect any differences within that range. Electrophysiological recording techniques capable of detecting synchronized or phase-locked neuronal activity on a fine time scale are perhaps better suited to exploring speech processing in that time window.

4.5 Cortical Maps of Speech Intelligibility: Left-Right and Anterior-Posterior Asymmetries

Previous imaging studies using standard subtraction-contrast and multivariate analysis methods have yielded somewhat conflicting interpretations of the cortical organization for processing intelligible speech. Two early studies localized processing of intelligible speech to the left anterior temporal lobe (Scott et al., 2000; Scott and Johnsrude, 2003), while more recent studies have observed bilateral activation to intelligible speech in the STG/S with a greater extent of activation in the left hemisphere, particularly in posterior temporal lobe regions (Davis and Johnsrude, 2003; Evans et al., 2014; McGettigan et al., 2012; Okada et al., 2010; Scott et al., 2006). Multivariate analysis of activation patterns within these bilateral regions suggests that patterns in the left hemisphere are maximally informative regarding the distinction between intelligible and unintelligible speech (Evans et al., 2014; McGettigan et al., 2012; Okada et al., 2010), yet these studies disagree on the relative contributions of posterior (Okada et al., 2010) versus anterior (Evans et al., 2014) temporal lobe regions. It should be noted that all of these studies used continuous speech (i.e., sentences) as stimuli, so “intelligibility” encompasses acoustic, phonetic, lexical-semantic, and syntactic/combinatorial semantic processing. However, studies using sublexical stimuli (i.e., comparing “phonetic” to “surface acoustic” processing) find similar patterns: a mix of left-lateralized (Liebenthal et al., 2005; Specht et al., 2009) and bilateral

(Evans and Davis, 2015; Vaden et al., 2010) effects in the STG/S, with a left hemisphere bias when effects were bilateral, and some disagreement over the relative contributions of posterior (Vaden et al., 2010) versus anterior (Liebenthal et al., 2005; Specht et al., 2009) superior temporal lobe regions.

Here, we assessed processing of intelligible speech in two ways: (1) using the standard subtraction contrast method (i.e., testing for a mean activation difference on intelligible vs. unintelligible trials); and (2) by testing directly for a correlation between STRFs estimated at each cortical surface node and the behavioral classification images for intelligibility (i.e, "behavioral STRFs") estimated for each participant. Both methods (Fig. 8A/B) revealed essentially bilateral activation of superior temporal lobe regions (STG/S): 48.0% and 54.1% of significant nodes were located in the left hemisphere for methods 1 and 2, respectively. An examination of the strength of the neural-behavioral correlations from method 2 turned up somewhat subtle (i.e., qualitative) evidence of hemispheric asymmetries with the left hemisphere yielding higher overall correlation values than the right (Fig. 8B). Within the left hemisphere, the largest relative correlations were observed in the anterior STG, mid-posterior STG, and Heschl's sulcus. Strong correlations were also observed in the left posterior STS. Within the right hemisphere, the largest relative correlations were observed in the mid-anterior STG/S.

Thus, while the overwhelming tendency in the present data is for intelligible speech to be processed and represented bilaterally (Figs. 3, 5, 8), and for both posterior and anterior regions of the STG/S to process speech at a relatively high (abstracted) level (Figs. 3D, 8C), there are modest biases favoring anterior over posterior regions and left over right hemisphere. In a sense, these intelligibility biases fall out naturally from the organization of the cortical speech hierarchy. Regions at the top level of the hierarchy (e.g., those coding word- or phrase-level information) depend on accurate encoding of speech information at lower processing levels (e.g., those regions coding syllables, phonemes, or their underlying spectrotemporal patterns). A failure to extract intelligible speech information at any level of processing will propagate up to the highest level and, therefore, lead to a near-perfect readout of intelligibility failures (and successes) in the activation patterns of regions like the anterior STG/S.

Activation in potentially lower-level regions (posterior STG/S) will reflect a partial readout of intelligibility failures, yielding the apparent gradients in the intelligibility correlations.

One early auditory region, at least anatomically speaking, that showed a strong intelligibility response was a portion of left Heschl's sulcus (Fig. 8B). This region was classified as belonging to Cluster Group 4 – the highest-level cluster from a functional standpoint – as was a small group of analogous cortical surface nodes in right Heschl's sulcus (Fig. 3A). Presumably, processing in this region reflects a strictly acoustic representation of the MPS features that support intelligible speech. This conclusion is supported by previous work suggesting that Heschl's sulcus plays a role in error correction during speech production. These studies show that Heschl's sulcus responds selectively to sublexical speech sounds (Formisano et al., 2008; Jäncke et al., 2002), activates during monitoring of overt speech feedback during production (van de Ven et al., 2009), and is suppressed when the likelihood of making a speech production error is reduced (Christoffels et al., 2011). Thus, the suggestion is that Heshcl's sulcus hosts the high-level acoustic targets (Guenther, 2006; Hickok, 2012) for speech production. Alternatively, it is possible that this region is processing a different slow modulation rate linguistic cue, namely prosody and/or stress patterns, that impact intelligibility. This is a topic for future work.

5. Conclusions

Data-driven clustering of speech-driven STRFs recovered a hierarchy of cortical speech processing in which early auditory areas in the supratemporal plane faithfully reconstructed the speech signal while later areas in the lateral temporal lobe gradually abstracted over earlier representations to emphasize the speech features important for intelligibility. Crucially, unlike previous imaging work, STRF-based analysis revealed the precise nature of speech representations throughout the cortical hierarchy, including evidence of processing specializations within and between different hierarchical levels. Particular cortical regions were specialized for processing different subsets of acoustic speech information within or outside the range of features that support intelligibility. A general trend was observed in which posteromedial regions of the supratemporal plane processed fine temporal information,

while anterolateral regions of the temporal lobe processed coarse temporal information. Similarly, medial supratemporal regions processed fine spectral information while lateral temporal regions processed coarse spectral information. This broad organization was nearly identical between the left and right hemispheres, though the left hemisphere showed a slight preference for processing the slow spectrotemporal modulations associated with intelligible speech. The left hemisphere also showed somewhat stronger correlations between STRF patterns the behaviorally-determined pattern of spectrotemporal modulations that underlie speech intelligibility. These correlations also tended to be somewhat stronger in anterior than posterior temporal lobe regions. Overall these findings are consistent with a bilateral but modestly asymmetric model of cortical speech processing with posterior-lateral regions preferentially processing phonological level information and anterior-lateral regions preferentially processing speech at the word and phrase level. No significant STRF tuning was observed outside the auditory cortex.

Acknowledgements

Research reported in this publication was supported by the National Institute on Deafness and Other Communication Disorders under Award Numbers R21 DC013406 (MPIs: VMR and Yi Shen) and R01 DC03681 (PI: GH). During this investigation, JHV was supported by the National Institute on Deafness and Other Communication Disorders under Award Numbers R01 DC000626 (PI: Marjorie R. Leek) and T32 DC010775 from the University of California, Irvine, CA, USA. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This material is the result of work supported with resources and the use of facilities at the VA Loma Linda Healthcare System, Loma Linda, CA. The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the US Government. Borja

Sànchez and Allison-Graham Martin performed quality control on the imaging data as undergraduate research assistants at the University of California, Irvine.

References

Barton, B., Venezia, J.H., Saberi, K., Hickok, G., Brewer, A.A., 2012. Orthogonal acoustic dimensions define auditory field maps in human cortex. Proceedings of the National Academy of Sciences 109, 20738-20743.

Bendor, D., Wang, X., 2006. Cortical representations of pitch in monkeys and humans. Curr Opin Neurobiol 16, 391-399.

Bendor, D., Wang, X., 2008a. Neural response properties of primary, rostral, and rostrotemporal core fields in the auditory cortex of marmoset monkeys. J Neurophysiol 100, 888-906.

Bendor, D., Wang, X., 2008b. Neural response properties of primary, rostral, and rostrotemporal core fields in the auditory cortex of marmoset monkeys. J Neurophysiol 100, 888-906.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 289-300.

Bernasconi, F., Grivel, J., Murray, M.M., Spierer, L., 2010. Interhemispheric coupling between the posterior sylvian regions impacts successful auditory temporal order judgment. Neuropsychologia 48, 2579-2585.

Bernasconi, F., Manuel, A.L., Murray, M.M., Spierer, L., 2011. Pre-stimulus beta oscillations within left posterior sylvian regions impact auditory temporal order judgment accuracy. International Journal of Psychophysiology 79, 244-248.

Bernstein, L.E., Liebenthal, E., 2014. Neural pathways for visual speech perception. Frontiers in neuroscience 8.

Bieser, A., Müller-Preuss, P., 1996. Auditory responsive cortex in the squirrel monkey: neural responses to amplitude-modulated sounds. Experimental Brain Research 108, 273-284.

Binder, J., Frost, J., Hammeke, T., Bellgowan, P., Springer, J., Kaufman, J., Possing, E., 2000. Human temporal lobe activation by speech and nonspeech sounds. Cerebral Cortex 10, 512-528.

Bitterman, Y., Mukamel, R., Malach, R., Fried, I., Nelken, I., 2008. Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. Nature 451, 197-201.

Bizley, J.K., Cohen, Y.E., 2013. The what, where and how of auditory-object perception. Nature Reviews Neuroscience 14, 693.

Bizley, J.K., Walker, K.M.M., Silverman, B.W., King, A.J., Schnupp, J.W.H., 2009. Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. Journal of Neuroscience 29, 2064-2075.

Boemio, A., Fromm, S., Braun, A., Poeppel, D., 2005. Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nat Neurosci 8, 389-395.

Bornkessel-Schlesewsky, I., Schlesewsky, M., Small, S.L., Rauschecker, J.P., 2015. Neurobiological roots of language in primate audition: common computational properties. Trends in Cognitive Sciences 19, 142-150.

Brugge, J.F., Merzenich, M.M., 1973. Response of neurons in auditory cortex of the macaque monkey to monaural and binaural stimulation. J Neurophysiol.

Brugge, J.F., Nourski, K.V., Oya, H., Reale, R.A., Kawasaki, H., Steinschneider, M., Howard, M.A., 2009. Coding of repetitive transients by auditory cortex on Heschl's gyrus. J Neurophysiol 102, 2358-2374.

Brugge, J.F., Volkov, I.O., Oya, H., Kawasaki, H., Reale, R.A., Fenoy, A., Steinschneider, M., Howard, M.A., 2008. Functional localization of auditory cortical fields of human: click-train stimulation. Hear Res 238, 12-24.

Camalier, C.R., D'Angelo, W.R., Sterbing-D'Angelo, S.J., Lisa, A., Hackett, T.A., 2012. Neural latencies across auditory cortex of macaque support a dorsal stream supramodal timing advantage in primates. Proceedings of the National Academy of Sciences 109, 18168-18173.

Chevillet, M., Riesenhuber, M., Rauschecker, J.P., 2011. Functional correlates of the anterolateral processing hierarchy in human auditory cortex. Journal of Neuroscience 31, 9345-9352.

Chi, T., Gao, Y., Guyton, M.C., Ru, P., Shamma, S., 1999. Spectro-temporal modulation transfer functions and speech intelligibility. The Journal of the Acoustical Society of America 106, 2719-2732.

Chi, T., Ru, P., Shamma, S.A., 2005. Multiresolution spectrotemporal analysis of complex sounds. The Journal of the Acoustical Society of America 118, 887-906.

Christoffels, I.K., van de Ven, V., Waldorp, L.J., Formisano, E., Schiller, N.O., 2011. The sensory consequences of speaking: parametric neural cancellation during speech in auditory cortex. PLoS ONE 6, e18307.

Coutanche, M.N., 2013. Distinguishing multi-voxel patterns and mean activation: Why, how, and what does it tell us? Cognitive, Affective, & Behavioral Neuroscience 13, 667-673.

Cox, R.W., 2012. AFNI: what a long strange trip it's been. Neuroimage 62, 743-747.

D'Ausilio, A., Craighero, L., Fadiga, L., 2012. The contribution of the frontal lobe to the perception of speech. Journal of Neurolinguistics 25, 328-335.

Davis, M.H., Johnsrude, I.S., 2003. Hierarchical processing in spoken language comprehension. J Neurosci 23, 3423-3431.

De Angelis, V., De Martino, F., Moerel, M., Santoro, R., Hausfeld, L., Formisano, E., 2017. Cortical processing of pitch: Model-based encoding and decoding of auditory fMRI responses to real-life sounds. Neuroimage, online Nov 13.

de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., Theunissen, F.E., 2017. The hierarchical cortical organization of human speech processing. Journal of Neuroscience 37, 6539-6557.

de la Mothe, L.A., Blumell, S., Kajikawa, Y., Hackett, T.A., 2006. Cortical connections of the auditory cortex in marmoset monkeys: core and medial belt regions. Journal of Comparative Neurology 496, 27-71.

Destrieux, C., Fischl, B., Dale, A., Halgren, E., 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage 53, 1-15.

Ding, N., Simon, J.Z., 2012. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 107, 78-89.

Elliott, T.M., Theunissen, F.E., 2009. The modulation transfer function for speech intelligibility. PLoS computational biology 5, e1000302.

Evans, S., 2017. What has replication ever done for us? Insights from neuroimaging of speech perception. Front Hum Neurosci 11.

Evans, S., Davis, M.H., 2015. Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. Cerebral Cortex 25, 4772-4788.

Evans, S., Kyong, J.S., Rosen, S., Golestani, N., Warren, J.E., McGettigan, C., Mourão-Miranda, J., Wise, R.J.S., Scott, S.K., 2014. The pathways for intelligible speech: multivariate and univariate perspectives. Cerebral Cortex 24, 2350-2361.

Felleman, D.J., Van Essen, D.C., 1991. Distributed hierarchical processing in the primate cerebral cortex. Cerebral cortex (New York, NY: 1991) 1, 1-47.

Fischl, B., 2012. FreeSurfer. Neuroimage 62, 774-781.

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.

Foxe, J.J., Schroeder, C.E., 2005. The case for feedforward multisensory convergence during early cortical processing. Neuroreport 16, 419-423.

Ghazanfar, A.A., Schroeder, C.E., 2006. Is neocortex essentially multisensory? Trends in Cognitive Sciences 10, 278-285.

Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T.E., Frackowiak, R.S., Laufs, H., 2007. Endogenous cortical rhythms determine cerebral specialization for speech perception and production. Neuron 56, 1127-1134.

Gosselin, F., Schyns, P.G., 2001. Bubbles: a technique to reveal the use of information in recognition tasks. Vision research 41, 2261-2271.

Griffin, D., Lim, J., 1984. Signal estimation from modified short-time Fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing 32, 236-243.

Griffiths, T.D., 2003. Functional imaging of pitch analysis. Ann N Y Acad Sci 999, 40-49.

Griffiths, T.D., Warren, J.D., 2004. What is an auditory object? Nature Reviews Neuroscience 5, 887-893.

Guenther, F.H., 2006. Cortical interactions underlying the production of speech sounds. J Commun Disord 39, 350-365.

Hackett, T.A., 2011. Information flow in the auditory cortical network. Hear Res 271, 133-146.

Hackett, T.A., Lisa, A., Camalier, C.R., Falchier, A., Lakatos, P., Kajikawa, Y., Schroeder, C.E., 2014. Feedforward and feedback projections of caudal belt and parabelt areas of auditory cortex: refining the hierarchical model. Frontiers in neuroscience 8.

Hackett, T.A., Stepniewska, I., Kaas, J.H., 1998. Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. Journal of Comparative Neurology 394, 475-495.

Hagler Jr, D.J., Saygin, A.P., Sereno, M.I., 2006. Smoothing and cluster thresholding for cortical surface-based group analysis of fMRI data. Neuroimage 33, 1093-1103.

Hickok, G., 2012. Computational neuroanatomy of speech production. Nature Reviews Neuroscience 13, 135-145.

Hickok, G., 2017. A cortical circuit for voluntary laryngeal control: Implications for the evolution language. Psychonomic Bulletin & Review 24, 56-63.

Hickok, G., Buchsbaum, B., Humphries, C., Muftuler, T., 2003. Auditory–Motor Interaction Revealed by fMRI: Speech, Music, and Working Memory in Area Spt. Journal of Cognitive Neuroscience 15, 673-682.

Hickok, G., Houde, J., Rong, F., 2011. Sensorimotor integration in speech processing: computational basis and neural organization. Neuron 69, 407-422.

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

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat Rev Neurosci 8, 393-402.

Hilgetag, C.C., O'Neill, M.A., Young, M.P., 2000. Hierarchical organization of macaque and cat cortical sensory systems explored with a novel network processor. Philosophical Transactions of the Royal Society of London B: Biological Sciences 355, 71-89.

Holdgraf, C.R., De Heer, W., Pasley, B., Rieger, J., Crone, N., Lin, J.J., Knight, R.T., Theunissen, F.E., 2016. Rapid tuning shifts in human auditory cortex enhance speech intelligibility. Nature Communications 7, 13654.

Howard, M.A., Volkov, I.O., Mirsky, R., Garell, P.C., Noh, M.D., Granner, M., Damasio, H., Steinschneider, M., Reale, R.A., Hind, J.E., 2000. Auditory cortex on the human posterior superior temporal gyrus. Journal of Comparative Neurology 416, 79-92.

Hullett, P.W., Hamilton, L.S., Mesgarani, N., Schreiner, C.E., Chang, E.F., 2016. Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. The Journal of Neuroscience 36, 2014-2026.

Humphries, C., Love, T., Swinney, D., Hickok, G., 2005. Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. Hum Brain Mapp 26, 128-138.

Jancke, L., Loose, R., Lutz, K., Specht, K., Shah, N.J., 2000. Cortical activations during paced finger-tapping applying visual and auditory pacing stimuli. Brain Res Cogn Brain Res 10, 51-66.

Jäncke, L., Wüstenberg, T., Scheich, H., Heinze, H.J., 2002. Phonetic perception and the temporal cortex. Neuroimage 15, 733-746.

Joosten, E.R.M., Neri, P., 2012. Human pitch detectors are tuned on a fine scale, but are perceptually accessed on a coarse scale. Biological cybernetics, 1-18.

Kaas, J.H., Hackett, T.A., 1998. Subdivisions of AuditoryCortex and Levels of Processing in Primates. Audiology and Neurotology 3, 73-85.

Kaas, J.H., Hackett, T.A., 2000. Subdivisions of auditory cortex and processing streams in primates. Proceedings of the National Academy of Sciences 97, 11793-11799.

Kikuchi, Y., Horwitz, B., Mishkin, M., 2010. Hierarchical auditory processing directed rostrally along the monkey's supratemporal plane. Journal of Neuroscience 30, 13021-13030.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., Broussard, C., 2007. What's new in Psychtoolbox-3. Perception 36, 1.1-16.

Kowalski, N., Depireux, D.A., Shamma, S.A., 1996. Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. J Neurophysiol 76, 3503-3523.

Kuhn, H.W., 1955. The Hungarian method for the assignment problem. Naval Research Logistics (NRL) 2, 83-97.

Kuśmierek, P., Rauschecker, J.P., 2009. Functional specialization of medial auditory belt cortex in the alert rhesus monkey. J Neurophysiol 102, 1606-1622.

Lakatos, P., Pincze, Z., Fu, K.-M.G., Javitt, D.C., Karmos, G., Schroeder, C.E., 2005. Timing of pure tone and noise-evoked responses in macaque auditory cortex. Neuroreport 16, 933-937.

Lalor, E.C., Foxe, J.J., 2010. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur J Neurosci 31, 189-193.

Lashkari, D., Vul, E., Kanwisher, N., Golland, P., 2010. Discovering structure in the space of fMRI selectivity profiles. Neuroimage 50, 1085-1098.

Leaver, A.M., Rauschecker, J.P., 2010. Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. Journal of Neuroscience 30, 7604-7612.

Liang, L., Lu, T., Wang, X., 2002. Neural representations of sinusoidal amplitude and frequency modulations in the primary auditory cortex of awake primates. J Neurophysiol 87, 2237-2261.

Liebenthal, E., Binder, J.R., Spitzer, S.M., Possing, E.T., Medler, D.A., 2005. Neural substrates of phonemic perception. Cereb Cortex 15, 1621-1631.

Liegeois-Chauvel, C., Musolino, A., Badier, J.M., Marquis, P., Chauvel, P., 1994. Evoked potentials recorded from the auditory cortex in man: evaluation and topography of the middle latency components. Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section 92, 204-214.

Liegeois-Chauvel, C., Musolino, A., Chauvel, P., 1991. Localization of the primary auditory area in man. Brain 114, 139-153.

Luke, S.G., 2017. Evaluating significance in linear mixed-effects models in R. Behavior research methods 49, 1494-1502.

Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer research 27, 209-220.

McGettigan, C., Evans, S., Rosen, S., Agnew, Z.K., Shah, P., Scott, S.K., 2012. An application of univariate and multivariate approaches in FMRI to quantifying the hemispheric lateralization of acoustic and linguistic processes. Journal of Cognitive Neuroscience 24, 636-652.

McLachlan, G., Peel, D., 2004. Finite mixture models. John Wiley & Sons.

Measurements, I.S.o.S., 1969. IEEE Recommended Practices for Speech Quality Measurements. IEEE Transactions on Audio and Electroacoustics 17, 227-246.

Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485, 233-236.

Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F., 2014. Phonetic feature encoding in human superior temporal gyrus. Science 343, 1006-1010.

Miller, L.M., Escabí, M.A., Read, H.L., Schreiner, C.E., 2002. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol 87, 516-527.

Moerel, M., De Martino, F., Formisano, E., 2012. Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. Journal of Neuroscience 32, 14205-14216.

Mumford, J.A., Turner, B.O., Ashby, F.G., Poldrack, R.A., 2012. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. Neuroimage 59, 2636-2643.

Narain, C., Scott, S.K., Wise, R.J., Rosen, S., Leff, A., Iversen, S., Matthews, P., 2003. Defining a left-lateralized response specific to intelligible speech using fMRI. Cerebral Cortex 13, 1362-1368.

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. Neuroimage 56, 400-410.

Neri, P., 2010. Visual detection under uncertainty operates via an early static, not late dynamic, non-linearity. Frontiers in computational neuroscience 4.

Neri, P., Levi, D.M., 2008. Evidence for joint encoding of motion and disparity in human visual perception. J Neurophysiol 100, 3117-3133.

Nourski, K.V., Brugge, J.F., Reale, R.A., Kovach, C.K., Oya, H., Kawasaki, H., Jenison, R.L., Howard, M.A., 2013. Coding of repetitive transients by auditory cortex on posterolateral superior temporal gyrus in humans: an intracranial electrophysiology study. J Neurophysiol 109, 1283-1295.

Nourski, K.V., Steinschneider, M., McMurray, B., Kovach, C.K., Oya, H., Kawasaki, H., Howard, M.A., 2014. Functional organization of human auditory cortex: Investigation of response latencies through direct recordings. Neuroimage 101, 598-609.

Nourski, K.V., Steinschneider, M., Oya, H., Kawasaki, H., Jones, R.D., Howard, M.A., 2012. Spectral organization of the human lateral superior temporal gyrus revealed by intracranial recordings. Cerebral Cortex 20, 340-352.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.H., Saberi, K., Serences, J.T., Hickok, G., 2010. Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cereb Cortex 20, 2486-2495.

Oosterhof, N.N., Wiestler, T., Downing, P.E., Diedrichsen, J., 2011. A comparison of volume-based and surface-based multi-voxel pattern analysis. Neuroimage 56, 593-600.

Overath, T., McDermott, J.H., Zarate, J.M., Poeppel, D., 2015. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat Neurosci 18, 903-911.

Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N.E., Knight, R.T., Chang, E.F., 2012. Reconstructing speech from human auditory cortex. PLoS-Biology 10, 175.

Peelle, J.E., 2012. The hemispheric lateralization of speech processing depends on what "speech" is: a hierarchical perspective. Front Hum Neurosci 6, 309.

Peelle, J.E., Johnsrude, I.S., Davis, M.H., 2010. Hierarchical processing for speech in human auditory cortex and beyond. Front Hum Neurosci 4.

Penagos, H., Melcher, J.R., Oxenham, A.J., 2004. A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. J Neurosci 24, 6810-6815.

Perrachione, T.K., Ghosh, S.S., 2013. Optimized design and analysis of sparse-sampling fMRI experiments. Frontiers in neuroscience 7.

Petersen, S.E., Fox, P.T., Posner, M.I., Mintun, M., Raichle, M.E., 1989. Positron emission tomographic studies of the processing of singe words. Journal of Cognitive Neuroscience 1, 153-170.

Poeppel, D., 2003. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. Speech Communication 41, 245-255.

Poeppel, D., Emmorey, K., Hickok, G., Pylkkänen, L., 2012. Towards a new neurobiology of language. Journal of Neuroscience 32, 14125-14131.

Pollok, B., Krause, V., Butz, M., Schnitzler, A., 2009. Modality specific functional interaction in sensorimotor synchronization. Hum Brain Mapp 30, 1783-1790.

Price, C., Thierry, G., Griffiths, T., 2005. Speech-specific auditory processing: where is it? Trends in Cognitive Sciences 9, 271-276.

Pulvermuller, F., Fadiga, L., 2010. Active perception: sensorimotor circuits as a cortical basis for language. Nat Rev Neurosci 11, 351-360.

Rauschecker, J.P., 1998. Cortical processing of complex sounds. Curr Opin Neurobiol 8, 516-521.

Rauschecker, J.P., Scott, S.K., 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci 12, 718-724.

Rauschecker, J.P., Tian, B., 2004. Processing of band-passed noise in the lateral auditory belt cortex of the rhesus monkey. J Neurophysiol 91, 2578-2589.

Rauschecker, J.P., Tian, B., Hauser, M., 1995. Processing of complex sounds in the macaque nonprimary auditory cortex. Science, 111-114.

Rauschecker, J.P., Tian, B., Pons, T., Mishkin, M., 1997. Serial and parallel processing in rhesus monkey auditory cortex. The Journal of comparative neurology 382, 89-103.

Recanzone, G.H., Guard, D.C., Phan, M.L., 2000. Frequency and intensity response properties of single neurons in the auditory cortex of the behaving macaque monkey. J Neurophysiol 83, 2315-2331.

Reddy, C.G., Dahdaleh, N.S., Albert, G., Chen, F., Hansen, D., Nourski, K., Kawasaki, H., Oya, H., Howard Iii, M.A., 2010. A method for placing Heschl gyrus depth electrodes. Journal of neurosurgery 112, 1301-1307.

Riesenhuber, M., Poggio, T., 2002. Neural mechanisms of object recognition. Curr Opin Neurobiol 12, 162-168.

Rockland, K.S., Ojima, H., 2003. Multisensory convergence in calcarine visual areas in macaque monkey. International Journal of Psychophysiology 50, 19-26.

Rogalsky, C., Hickok, G., 2008. Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. Cerebral Cortex 19, 786-796.

Saad, Z.S., Reynolds, R.C., 2012. Suma. Neuroimage 62, 768-773.

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 computational biology 10, e1003412.

Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., Formisano, E., 2017. Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. Proceedings of the National Academy of Sciences, 201617622.

Schönwiesner, M., Zatorre, R.J., 2009. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proceedings of the National Academy of Sciences 106, 14611-14616.

Schreiner, C.E., Urbas, J.V., 1988. Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields. Hear Res 32, 49-63.

Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S., Puce, A., 2008. Neuronal oscillations and visual amplification of speech. Trends in Cognitive Sciences 12, 106-113.

Scott, B.H., Malone, B.J., Semple, M.N., 2011. Transformation of temporal processing across auditory cortex of awake macaques. J Neurophysiol 105, 712-730.

Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123, 2400-2406.

Scott, S.K., Johnsrude, I.S., 2003. The neuroanatomical and functional organization of speech perception. Trends in neurosciences 26, 100-107.

Scott, S.K., Rosen, S., Lang, H., Wise, R.J.S., 2006. Neural correlates of intelligibility in speech investigated with noise vocoded speech—a positron emission tomography study. The Journal of the Acoustical Society of America 120, 1075-1083.

Scott, S.K., Wise, R.J.S., 2003. PET and fMRI studies of the neural basis of speech perception. Speech Communication 41, 23-34.

Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E., 2016. mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. The R Journal 8, 289.

Serre, T., Oliva, A., Poggio, T., 2007. A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences 104, 6424-6429.

Shamma, S., 2001. On the role of space and time in auditory processing. Trends in Cognitive Sciences 5, 340-348.

Simonyan, K., 2014. The laryngeal motor cortex: its organization and connectivity. Curr Opin Neurobiol 28, 15-21.

Singmann, H., Kellen, D., 2017. An Introduction to Mixed Models for Experimental Psychology. New Methods in Neuroscience and Cognitive Psychology. Psychology Press Hove.

Smith, F.W., Muckli, L., Brennan, D., Pernet, C., Smith, M.L., Belin, P., Gosselin, F., Hadley, D.M., Cavanagh, J., Schyns, P.G., 2008. Classification images reveal the information sensitivity of brain voxels in fMRI. Neuroimage 40, 1643-1654.

Specht, K., 2013. Mapping a lateralization gradient within the ventral stream for auditory speech perception. Front Hum Neurosci 7, 629.

Specht, K., Osnes, B., Hugdahl, K., 2009. Detection of differential speech-specific processes in the temporal lobe using fMRI and a dynamic "sound morphing" technique. Hum Brain Mapp 30, 3436-3444.

Tang, C., Hamilton, L.S., Chang, E.F., 2017. Intonational speech prosody encoding in the human auditory cortex. Science 357, 797-801.

Theunissen, F.E., Elie, J.E., 2014. Neural processing of natural sounds. Nature Reviews Neuroscience 15, 355-366.

Town, S.M., Bizley, J.K., 2013. Neural and behavioral investigations into timbre perception. Frontiers in systems neuroscience 7, 88.

Tremblay, P., Small, S.L., 2011. On the context-dependent nature of the contribution of the ventral premotor cortex to speech perception. Neuroimage 57, 1561-1571.

Vaden, K.I., Muftuler, L.T., Hickok, G., 2010. Phonological repetition-suppression in bilateral superior temporal sulci. Neuroimage 49, 1018-1023.

van de Ven, V., Esposito, F., Christoffels, I.K., 2009. Neural network of speech monitoring overlaps with overt speech production and comprehension networks: a sequential spatial and temporal ICA study. Neuroimage 47, 1982-1991.

Venezia, J.H., Hickok, G., Richards, V.M., 2016. Auditory "bubbles": Efficient classification of the spectrotemporal modulations essential for speech intelligibility. The Journal of the Acoustical Society of America 140, 1072-1088.

Vigneau, M., Beaucousin, V., Hervé, P.-Y., Jobard, G., Petit, L., Crivello, F., Mellet, E., Zago, L., Mazoyer, B., Tzourio-Mazoyer, N., 2011. What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing?: Insights from a meta-analysis. Neuroimage 54, 577-593.

Wessinger, C.M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., Rauschecker, J.P., 2001. Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. Journal of Cognitive Neuroscience 13, 1-7.

Wilson, S.M., DeMarco, A.T., Henry, M.L., Gesierich, B., Babiak, M., Mandelli, M.L., Miller, B.L., Gorno-Tempini, M.L., 2014. What role does the anterior temporal lobe play in sentence-level processing? Neural correlates of syntactic processing in semantic variant primary progressive aphasia. Journal of Cognitive Neuroscience 26, 970-985.

Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nat Neurosci 7, 701-702.

Woods, D.L., Herron, T.J., Cate, A.D., Yund, E.W., Stecker, G.C., Rinne, T., Kang, X., 2010. Functional properties of human auditory cortical fields. Frontiers in systems neuroscience 4.

Woolley, S.M., Fremouw, T.E., Hsu, A., Theunissen, F.E., 2005. Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. Nat Neurosci 8, 1371-1379.

Zatorre, R.J., Belin, P., 2001. Spectral and temporal processing in human auditory cortex. Cerebral Cortex 11, 946-953.

Zatorre, R.J., Belin, P., Penhune, V.B., 2002. Structure and function of auditory cortex: music and speech. Trends in Cognitive Sciences 6, 37-46.

Figure Captions

Figure 1. (A) Speech Modulation Power Spectrum. Left: Average MPS of 452 sentences spoken by a single female talker. The MPS describes speech as a weighted sum of spectrotemporal ripples containing energy at a unique combination of temporal (Hz; abscissa) and spectral (cycles/kHz; ordinate) modulation rate. Modulation energy (dB, arb. ref; color scale) clusters into two discrete regions: a high-spectral-modulation-rate region corresponding to finely spaced harmonics of the fundamental (a "pitch region") and a low-spectral-modulation-rate region corresponding to coarsely spaced resonant frequencies of the vocal tract (a "formant region"). The black contour line indicates the modulations accounting for 80% of the total modulation power. A spectrogram of an example spectrotemporal ripple (2 Hz, 4 cyc/kHz) is shown beneath. Right: Coefficient of variation across the 452 sentences (sd/mean), expressed as a percentage (color scale). Plotted on the same axes as the MPS. There is relatively little variation across utterances (maximum CV ~7%). **(B) Bubbles Procedure**. Bubbles (middle) are applied to an image of a face (left) and the MPS of an individual sentence (right). In either case, bubbles reduce the information in the stimulus. Different random bubble patterns are applied across trials of an experiment. For auditory bubbles, we in practice use a binary masker with bubbles that are larger than those shown in the example.

Figure 2. Bubbles Analysis Schematic. A BOLD activation time-course from a single voxel in left Heschl's gyrus of a representative subject is shown (blue line). The time-course plots the z-scored time-series of single-trial activation magnitudes (beta; ordinate) evoked by "bubble-ized" sentences (Sentence No., abscissa). Example bubble patterns (black-and-white panels) associated with sentences that evoked relatively large (top) and small (bottom) activations are plotted and identified by their sentence number. Z-scored activation magnitudes associated with these examples are shown next to the corresponding point in the activation time-course. Bubbles are applied to the MPS of each sentence as shown in Fig. 1. White pixels show regions of the MPS that are transmitted to the listener, while black pixels show regions of the MPS that are removed. Each bubble pattern is multiplied by its associated z-score, and the series of

bubble patterns is summed pixel-by-pixel. The resulting summed image is then blurred (Gaussian filter with sigma = 5 pixels) and scaled by the across-pixel standard deviation (sd_{px}). The result is a STRF (top right) showing which regions of the MPS best activated this voxel. The STRF color scale is in across-pixel standard deviation units, where large positive values (yellow-red) correspond to regions of the MPS that evoked relatively large activations.

Figure 3. (A) Maps of STRF Cluster Groups in Auditory Cortex. Cluster Groups are plotted by color on cortical surface renderings of the left and right hemispheres. Zoomed renderings of the temporal lobe are shown beneath whole-brain plots. Cluster Group 1 (CG1, blue) is located primarily in the supratemporal plane and posterior STG. Cluster Group 2 (CG2, cyan) is located primarily in medial supratemporal reigons. Cluster Groups 3 and 4 (CG3/4, yellow/red) are located primarily in the posterior and anterior STG/STS. **(B) STRF-Cluster Patterns.** For each of the 18 STRF clusters identified by GMM analysis, the cluster-average group-level (t-score) STRF is plotted. STRF magnitudes have been normalized to the range [0, 1]. Larger values are associated with STMs that produced relatively more BOLD activation. STRFs are organized by Cluster Group (CG1-4) in columns running from left to right. STRFs associated with CG1 respond to a broad range of STMs. STRFs associated with CG2 respond especially to high temporal modulation rates. STRFs associated with CG3/4 respond to STMs important for intelligibility (see C). **(C) Behavioral Classification Image for Intelligibility Judgments.** This plot is essentially a 'behavioral STRF', derived entirely from button-press responses (yes-no intelligibility judgments) rather than neural activity. The z-scored group-level behavioral classification image is shown. Larger values are associated with STMs that contribute relatively more to intelligibility. Temporal modulations from 2-7 Hz and spectral modulations less than 1 cyc/kHZ contribute maximally. **(D) Distribution of Cluster Groups within Anatomically Defined Regions.** The proportion of cortical surface nodes belonging to CG1-4 is plotted for six anatomical regions of interest in the left (LH) and right (RH) hemispheres: Heschl = Heschl's gyrus/sulcus, pSTG/S = posterior STG/S, aSTG/S = anterior STG/S, pSyl = posterior Sylvian cortex. Colored boxes beneath region labels correspond to the colors of

the anatomical regions plotted on zoomed cortical surface renderings at right. Only significantly tuned cortical surface nodes are labeled.

Figure 4. Cluster level STRFs with the global effect of intelligibility removed ($STRF_{Unbiased}$). For each of the 18 STRF clusters identified by GMM analysis, the cluster-average group-level (t-score) $STRF_{Unbiased}$ is plotted. STRF magnitudes have been normalized to the range [0, 1]. Larger values are associated with STMs that produced relatively more BOLD activation. STRFs are organized by Cluster Group (CG1-4) in columns running from left to right. Compare to Fig. 3B.

Figure 5. (A) Cortical Maps of Peak Modulation Frequencies. Node-wise maps of temporal peak modulation frequency (tPMF, Hz) and spectral peak modulation frequency (sPMF, cyc/kHz) are displayed on inflated cortical surface renderings of the left and right temporal lobes. The renderings have been zoomed in as indicated by the red boxes at the top of the figure. Color scales are logarithmic. **(B) Probability Density of tPMF and sPMF Within Cluster Groups.** Empirical cumulative distribution functions (eCDFs; Kaplan-Meier method) for tPMF (Hz, top) and sPMF (cyc/kHz, bottom) were generated. Empirical probability density functions (ePDFs) were obtained by taking the derivative of the eCDFs. The ePDFs are plotted for each cluster group (colored lines, see legend) separately for the left (LH) and right (RH) hemispheres. The interquartile ranges (25^{th} percentile - 75^{th} percentile) of each distribution are indicated at the top right of each panel (IQR). The ordinate is the estimated probability density.

Figure 6. Linear Mixed Effects Models: Best Modulation Frequency. (A) Effect of Cluster Region. The mean of the fitted values produced by the LME model for temporal (tBMF, top) and spectral (sBMF, bottom) best modulation frequencies (octave scale, ordinate) are plotted for Cluster Groups 1-4 (abscissa)

in the left (blue) and right (red) hemispheres. Error bars reflect ± 1 SEM. Spectral BMFs are negative because nodes with an sBMF of 0, of which there were many, were set to 0.01 (-6.6 on the octave scale). **(B) Covariation between tBMF and sBMF.** Results of linear mixed effects regression of sBMF on tBMF (top) and tBMF on sBMF (bottom) by hemisphere are plotted as fitted lines (bold blue) with 95% confidence regions (light blue shading). BMFs have been mean-centered and transformed to the octave scale (i.e., axes show distance from the mean t/sBMF in octaves). Ticks above the abscissa indicate the values of the covariate at which data were actually observed.

Figure 7. Evidence of STRF Specializations Within Cluster Groups. Individual clusters of interest (A-C) are plotted on inflated cortical surface renderings of the left and right hemispheres (figure left). Zoomed surface renderings of the temporal lobes are shown beneath the whole-brain plots. The cluster-average group-level (t-score) STRFs are also plotted with magnitudes normalized to the range [0, 1] (figure right). **(A, blue)** From Cluster Group 1, this cluster on lateral Heschl's Gyrus and the neighboring STG responds best to STMs at high cyc/kHz ("pitch" STMs). **(B, Green)** From Cluster Group 2, this cluster located entirely in the right auditory cortex responds best to STMS at high temporal modulation rates (Hz). **(C, Red)** From Cluster Group 4, this cluster located prominently in the left anterior temporal lobe responds best to STMs important for intelligibility, particularly at very low temporal modulation rates (< 3 Hz).

Figure 8. (A) Contrast Map of Speech Intelligibility. The group mean contrast beta (intelligible trials vs. unintelligible trials) is plotted on cortical surface renderings of the left and right hemispheres. Whole-brain analysis, wild-bootstrap-corrected $p < 0.05$. **(B) Correlation Map of Speech Intelligibility.** The group mean Fisher z-transformed correlation, z(r), between behavioral classification images for intelligibility and neural STRFs is plotted on cortical surface renderings of the left and right hemispheres. Whole-brain analysis, wild-bootstrap-corrected $p < 0.05$. **(C) Linear Mixed Effects Analysis of Intelligibility Correlation Values.** The mean of the LME-fitted values of the Fisher z-transformed

correlation, z(r), between behavioral classification images for intelligibility and neural STRFs is plotted across cluster regions (top) and anatomical regions (bottom) in the left (blue) and right (red) hemispheres. Error bars reflect ± 1 SEM.

Figure 9. (A) Cluster-Group Maps at the Group Level and in Representative Individual Participants. Cluster Groups are plotted by color on cortical surface renderings of the left and right hemispheres. Separate maps are shown for the group-level data (GRP), and for the two individual participants with the lowest (S2) and highest (S3) percent agreement with the group. **(B) Distribution of Individual-Participant Cluster Groups within Anatomically Defined Regions.** The across-participant average proportion of cortical surface nodes belonging to Cluster Group 1-4 is plotted for six anatomical regions of interest in the left (LH) and right (RH) hemispheres: Heschl = Heschl's gyrus/sulcus, pSTG/S = posterior STG/S, aSTG/S = anterior STG/S, pSyl = posterior Sylvian cortex. Error bars = ± 1 SEM. Compare to Fig. 3D for group-level distributions.