Published in final edited form as:

Hear Res. 2021 September 01; 408: 108307. doi:10.1016/j.heares.2021.108307.

Speech-Driven Spectrotemporal Receptive Fields Beyond the Auditory Cortex

Jonathan H. Venezia[1,2], **Virginia M. Richards**[3], **Gregory Hickok**[3]

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

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

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

Abstract

We recently developed a method to estimate speech-driven spectrotemporal receptive fields (STRFs) using fMRI. The method uses spectrotemporal modulation filtering, a form of acoustic distortion that renders speech sometimes intelligible and sometimes unintelligible. Using this method, we found significant STRF responses only in classic auditory regions throughout the superior temporal lobes. However, our analysis was not optimized to detect small clusters of STRFs as might be expected in non-auditory regions. Here, we re-analyze our data using a more sensitive multivariate statistical test for cross-subject alignment of STRFs, and we identify STRF responses in non-auditory regions including the left dorsal premotor cortex (dPM), left inferior frontal gyrus (IFG), and bilateral calcarine sulcus (calcS). All three regions responded more to intelligible than unintelligible speech, but left dPM and calcS responded significantly to vocal pitch and demonstrated strong functional connectivity with early auditory regions. Left dPM's STRF generated the best predictions of activation on trials rated as unintelligible by listeners, a hallmark auditory profile. IFG, on the other hand, responded almost exclusively to intelligible speech and was functionally connected with classic speech-language regions in the superior temporal sulcus and middle temporal gyrus. IFG's STRF was also (weakly) able to predict activation on unintelligible trials, suggesting the presence of a partial 'acoustic trace' in the region. We conclude that left dPM is part of the human dorsal laryngeal motor cortex, a region previously shown to be capable of operating in an 'auditory mode' to encode vocal pitch. Further,

Corresponding Author: Jonathan H. Venezia, PhD, 11201 Benton St., Mail Code 151, Loma Linda, CA 92357, 909-825-7084 x7292, jonathan.venezia@va.gov.

Author Contributions

JHV: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, software, validation, visualization, writing – original draft, writing – review and editing.

VMR: Formal analysis, funding acquisition, project administration, supervision, validation, writing – review and editing.

GH: Conceptualization, funding acquisition, project administration, supervision, writing – review and editing.

Declaration of Competing Interests

None.

Ethics Statement

This work was carried out in accordance with the Declaration of Helsinki. All subjects provided written informed consent. This study was approved by the University of California, Irvine Institutional Review Board.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

given previous observations that IFG is involved in syntactic working memory and/or processing of linear order, we conclude that IFG is part of a higher-order speech circuit that exerts a top-down influence on processing of speech acoustics. Finally, because calcS is modulated by emotion, we speculate that changes in the quality of vocal pitch may have contributed to its response.

Keywords

fMRI; Pitch; Premotor; Spectrotemporal Modulations; Speech Intelligibility; STRF

1. Introduction

Two decades ago, Buchsbaum et al. (2001) used functional magnetic resonance imaging (fMRI) to characterize a left-hemisphere network of regions in the posterior temporal and inferior frontal lobes that responded reliably during both perception and covert production of speech. The posterior superior temporal sulcus (pSTS) and dorsal speech-premotor cortex (dPM) responded relatively more during perception and production, respectively, while a region in the posterior Sylvian fissure (Spt) and the pars opercularis of the inferior frontal gyrus (pOP) responded roughly equally to perception and production of speech. The results were interpreted largely as evidence that auditory regions (pSTS) interface with articulatory regions (dPM) via an intermediary auditory-motor network (Spt to pOP) in service of speech production. The same network was later identified using sub-vocal repetition of speech and tonal melodies (Hickok, Buchsbaum, Humphries, & Muftuler, 2003). Wilson et al. (2004) independently demonstrated that listening to speech activates Spt, pOP and dPM. In contrast to the earlier studies, the authors interpreted their finding as evidence for a role of the motor cortex in speech perception. Indeed, they showed that overlapping regions of dPM were activated during perception and production of speech. The studies by Buchsbaum et al. (2001) and Wilson et al. (2004) both noted significant activation to heard speech in dPM for every single subject. A later meta-analysis by Buchsbaum and colleagues (Buchsbaum et al., 2011) found the cross-subject peak (overlap analysis) of regions activating to both speech perception and production was located in the dPM at Montreal Neurological Institute (MNI) coordinate [−51, −9, 42]. The average center of mass for perception-related activations in dPM noted by Wilson et al. (2004) was at MNI coordinate [−50, −6, 47] (see Figure 7, Discussion).

Despite the similarity of these findings, the study by Wilson et al. (2004) was cited more widely among studies on speech perception and became a major precipitant to years of renewed interest in the role of the motor cortex in speech perception (cf., D’Ausilio, Craighero, & Fadiga, 2012; Peelle, 2012; J. E. Peelle, I. S. Johnsrude, & M. H. Davis, 2010; Poeppel & Assaneo, 2020; Pulvermuller & Fadiga, 2010; Skipper, Devlin, & Lametti, 2017; Tremblay & Small, 2011). Overall, this more recent research has coalesced around analysis-by-synthesis (Bever, 2010; Skipper, Nusbaum, & Small, 2005) and Bayesian inference (Moulin-Frier, Diard, Schwartz, & Bessière, 2015) frameworks in which heard (or seen) speech is, at some level, re-encoded in terms of the articulatory commands used to generate the speech signal in order to form perceptual hypotheses that constrain the analysis and interpretation of incoming speech sounds. While several researchers have dismissed the notion that such mechanisms are a core component of speech perception (Hickok, 2010; Holt & Lotto, 2008; Rogalsky et al., 2020; Scott, McGettigan, & Eisner, 2009; Venezia & Hickok, 2009), it has been acknowledged that motor speech circuits likely play a small, modulatory role in certain listening situations (Stokes, Venezia, & Hickok, 2019).

An alternative computational account of the motor system's involvement in speech perception is that some motor regions process heard speech in terms of its auditory features, without necessarily translating heard speech into motor commands. Specifically, recent studies show that pre- and primary-motor speech regions fail to discriminate among heard syllables in terms of their articulatory features (i.e., place of articulation; Arsenault & Buchsbaum, 2016; Cheung, Hamilton, Johnson, & Chang, 2016). However, a region within dPM discriminates among heard syllables based on features that are acoustically well defined (i.e. manner of articulation), shows auditory-like tuning to the spectrotemporal modulations in continuous speech, and tracks time-varying acoustic speech features including the spectral envelope, rhythmic phrasal structure, and pitch contour (Berezutskaya, Baratin, Freudenburg, & Ramsey, 2020; Cheung et al., 2016). Chang and colleagues (Breshears, Molinaro, & Chang, 2015; Dichter, Breshears, Leonard, & Chang, 2018) have shown convincingly that this region is part of the human dorsal laryngeal motor cortex (dLMC; cf., Brown, Ngan, & Liotti, 2008) and has both auditory and motor representations of vocal pitch. Among primates, the dLMC is unique to humans and likely evolved to provide voluntary control of the larynx (Belyk & Brown, 2017; Dichter et al., 2018; Simonyan, 2014). It has been suggested that the dLMC contains sulcal (BA 4) and gyral (BA 6) components (Brown, Yuan, & Belyk, 2020). The seminal study by Brown and colleagues (2008) puts the left-hemisphere peak of the gyral component at MNI coordinate [−54, −2, 46] (see Figure 7, Discussion).

To summarize, we and others have long recognized dPM as an important node in the auditory-motor "dorsal stream" that responds during both perception and production (Chen, Penhune, & Zatorre, 2009; Hickok & Poeppel, 2004). Initially, we dismissed dPM's response to heard speech as somewhat trivial given that feedforward auditory-to-motor activation is expected within the dorsal stream where auditory representations serve as the "targets" for speech production (Venezia & Hickok, 2009). However, recent research suggests that speech is represented differently in dPM depending on whether speech is perceived or produced. This may relate to the fact that dPM is located within the dLMC, which some researchers have identified as playing crucial role in integrating phonation with other aspects of speech motor control (Belyk & Brown, 2017; Brown et al., 2020). We have independently suggested that dPM is part of the cortical circuit for laryngeal motor control (Hickok, 2017). Therefore, one hypothesis is that dPM responds to multiple aspects of speech motor control during production (Belyk et al., 2020; Cheung et al., 2016) but responds preferentially to pitch/voicing during perception (Cheung et al., 2016), which may owe to the presence of non-identical subpopulations of neurons coding auditory and motor features, respectively, as is typical of brain regions involved in sensorimotor integration (Sakata, Taira, Murata, & Mine, 1995).

We recently developed a method called Auditory Bubbles to estimate speech-driven spectrotemporal receptive fields (STRFs) using fMRI (Venezia, Thurman, Richards, &

Hickok, 2019b). Using this method, we detected widespread and reliable responses to spectrotemporal speech features in the auditory core and immediate surrounds, superior temporal gyrus/sulcus, and planum temporale. We did not detect such responses beyond these classic auditory-speech regions, e.g., in inferior frontal or speech motor regions. However, spectrotemporal responses were quantified at the second (group) level using univariate statistical methods that were not optimized to detect less robust spectrotemporal responses in focal brain regions such as the dPM/dLMC (cf., Cheung et al., 2016). Here, we develop a new multivariate analysis for Auditory Bubbles – and applicable to any related feature-encoding method – that maximizes sensitivity to spectrotemporal responses while maintaining full statistical rigor at the second level. Note, here and elsewhere we use 'multivariate' to refer to statistical analysis of linear-encoding-model 'weights' associated with stimulus features in a multidimensional space, not to incorporation of information from multiple neighboring voxels such as in multivoxel pattern analysis. Using this technique, we re-analyze our original dataset with the intention of revealing STRF-like responses in inferior frontal speech-motor regions. To preview, we indeed find such responses in the two frontal regions previously identified as responding to heard speech: left dPM and left inferior frontal gyrus (IFG). Of note, Auditory Bubbles estimates STRFs in the spectrotemporal modulation domain (Figure 1). To improve the efficiency of STRF estimation, Auditory Bubbles introduces acoustic variation into the speech signal via spectrotemporal modulation filtering, rendering the filtered speech sometimes intelligible and sometimes unintelligible. Thus, Auditory Bubbles is uniquely positioned to answer several outstanding questions about the computations performed in left dPM and IFG during perception of continuous speech: (a) what acoustic-speech information are these regions sensitive to and how does this relate to information encoded at early and late stages of auditory processing in classic temporal lobe regions?; (b) do these patterns interact with speech intelligibility?; and (c) how well can spectrotemporal response properties be used to predict activation in inferior frontal versus temporal speech regions for speech that is (un-)intelligible?

In principle, responses to heard speech in IFG and dPM could reflect auditory, motor, or auditory-motor processing. We aimed to characterize these responses in the STRF domain using a series of follow-up analyses to disentangle fundamentally auditory responses from motor, auditory-motor, or other higher-level (i.e., speech-specific) responses. Our confidence in interpreting a response as auditory increases if the following hold true: (i) a significant STRF-like response is observed, given that, generally, responses to speech in the spectrotemporal modulation domain are more likely to reflect processing of auditory-phonetic speech cues than processing at the articulatory-phonetic, phonological, or some other higher level; (ii) within the STRF, a significant response to vocal pitch is observed, given that vocal pitch is acoustically well defined in the spectrotemporal modulation domain (Figure 1); and (iii) an STRF-like response is observed *within* trials rated as unintelligible by the listener (i.e., when speech is not processed as speech *per se*), given that the ability to extract speech-motor or other speech-specific cues is greatly reduced when the presence of such cues in the signal is, by definition, diminished. Finally, for a given non-auditory region, confidence that its response is auditory-driven is further increased if that region is functionally connected to early auditory regions rather than higher-level auditory regions that respond selectively to intelligible speech.

Briefly, we find that responses in left IFG are primarily driven by spectotemporal features that mediate speech intelligibility (i.e., its properties mirror a classic intelligible vs. unintelligible contrast; Narain et al., 2003; Okada et al., 2010), while those in left dPM are driven by additional spectrotemporal features including, most notably, features associated with vocal pitch, though both regions are better activated by intelligible than unintelligible speech. In addition, both regions show some level of response to surface-level acoustic characteristics within trials rated as unintelligible by listeners, but this response is more pronounced in left dPM compared to left IFG. Interestingly, sensitivity to vocal pitch in left dPM emerges only within unintelligible trials. Finally, though both regions show functional connectivity with regions in the auditory cortex, left dPM is maximally correlated with regions in and around Heschl's gyrus while left IFG is maximally correlated with downstream speech regions in the superior temporal gyrus/sulcus and middle temporal gyrus. We also report evidence that the bilateral early visual cortex shows STRF-like responses and patterns of functional connectivity similar to left dPM. In the Discussion, we clarify how these response profiles differ and speculate as to the role, if any, of visual regions in auditory speech processing.

2. Materials and Methods

2.1 Overview

A detailed description of the materials and methods is given by Venezia et al. (2019b). Here, we recapitulate only those methodological details necessary to understand the study, but we describe in full detail the present re-analysis of the data. In this overview section, we describe the Auditory Bubbles method and its use in fMRI to estimate speech-driven spectrotemporal receptive fields (STRFs). Auditory Bubbles falls within the class of techniques referred to as 'relative weights' or 'reverse correlation' in the psychoacoustic and/or neurophysiological literatures (Klein, Depireux, Simon, & Shamma, 2000; Matiasek & Richards, 1999), and as 'encoding models' in the fMRI literature (Naselaris, Kay, Nishimoto, & Gallant, 2011). In the context of neurophysiological responses to complex natural sounds such as speech, this class of methods is used to estimate STRFs (Theunissen & Elie, 2014). In typical STRF estimation, natural variation in the acoustic envelope of individual spectral bands is mapped to a time-locked neurophysiological response (e.g., a spike train) using a deconvolutional linear model. The STRF comprises the model weights, which describe the spectrotemporal patterns that best activate a given unit of analysis (neuron, voxel, electrode, etc.). The STRF can be visualized as an 'impulse-response spectrogram' with a time axis (abscissa) centered on time zero (impulse onset) and a frequency axis (ordinate) spanning the range of frequencies present in the stimulus. Alternatively, the STRF can be transformed into rate-scale space via the two-dimensional Fourier transform (Theunissen & Elie, 2014), such that best responses are described in terms of temporal (abscissa) and spectral (ordinate) modulations.

In fMRI, rate-scale STRFs must be estimated directly in the rate-scale space because sampling of the neurophysiological signal is much slower than the relevant temporal modulation rates in natural sounds (and sampling is often temporally sparse as well; Santoro et al., 2014; Venezia et al., 2019b). In other words, the acoustic stimulus must first be transformed to rate-scale space prior to the linear modeling phase of STRF estimation. This is the approach taken by Auditory Bubbles, though we take the additional step of introducing artificial acoustic variation into the stimulus by applying a filter in the rate-scale domain. This improves the efficiency of STRF estimation while introducing some acoustic distortion, where the latter has advantages (e.g., allows for a calibrated behavioral task) and disadvantages (e.g., makes the stimulus sound less natural). Here, we refer to the rate-scale representation of sounds as the modulation power spectrum (MPS). Briefly, Auditory Bubbles applies a binary filter to the MPS of acoustic speech stimuli – typically sentences – such that contiguous segments of the MPS are either retained in the stimulus or removed. The shape of the filter, which determines the MPS segments that will be retained/removed, is determined quasi-randomly. A different filter is applied to each sentence in the stimulus set, and the filtered sentences are presented sequentially to a listener in the MRI scanner. After scanning, the filters themselves – each summarizing the spectrotemporal patterns retained in each sentence – are weighted by the response elicited by the corresponding sentence and summed to produce a STRF. Essentially, the weighted sum is a linear model (cf., Venezia, Hickok, & Richards, 2016; Venezia, Leek, & Lindeman, 2020; Venezia, Martin, Hickok, & Richards, 2019a) whose predictors are the binary filters and whose criterion is the sentence-by-sentence fMRI response magnitude (i.e., event-related beta time series). A schematic of this process is shown in Figure 1C. Notably, Auditory Bubbles obtains the MPS from sentence spectrograms with a linear frequency axis, which results in speech energy clustering in two locations of the MPS – a high-spectral-modulation-rate region associated with vocal pitch and a low-spectral-modulation-rate region associated with phonetic content (roughly formant-scale features; Figure 1A). Thus, STRFs estimated with Auditory bubbles also cluster in these two regions of the MPS.

2.2 Participants

There were ten participants (mean age = 26, range = 20–33, 2 females) all of whom were right-handed, native speakers of American English with self-reported normal hearing and normal or corrected-to-normal vision. The participants all gave their informed consent in accordance with the University of California, Irvine Institutional Review Board guidelines.

2.3 Stimuli

The stimuli 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 and the waveforms were zero-padded to an equal duration of 3.29 s. The stimuli were then filtered to remove randomly selected segments of the MPS (Figure 1B). Specifically, a log spectrogram was obtained using Gaussian windows with a 4.75 ms-33.5 Hz time-frequency scale and the MPS was then obtained as the modulus of the two-dimensional Fourier transform of the spectrogram. A binary “bubbles” filter was then multiplied with the MPS and the filtered waveform was obtained via inverse Fourier transform and iterative inversion of the resultant magnitude spectrogram (Griffin & Lim, 1984; Theunissen & Elie, 2014). Specifically, a 2D image of the same dimensions as the MPS was created with a set number of randomly selected pixel locations assigned the value 1 and the remainder of pixels assigned the value 0. A symmetric Gaussian low-pass filter (sigma = 7 pixels) was then 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 several contiguous regions with value 1. A second Gaussian low-pass filter (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 corresponds to the number of "bubbles" in the filter. It is noteworthy that the low-pass form of the bubbles filters propagates to STRFs estimated from those filters (see Section 2.7.1), resulting in smoother STRFs than might be obtained from methods such as autocorrelation-normalized reverse correlation or boosting (David, Mesgarani, & Shamma, 2007; Theunissen, Sen, & Doupe, 2000). 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. Separate sets of filtered stimuli were created using different numbers of bubbles (20–100 in steps of five), producing a total of 7684 filtered sentences. All stimuli were generated offline and stored prior to the experiment. The average proportion of the MPS revealed to the listener is ~ 0.25 for 20 bubbles and ~ 0.7 for 100 bubbles. Examples of filtered stimuli with 50 bubbles are given in Supplementary Audio 1–3 (each presents a filtered sentence followed by the unfiltered source).

2.4 Procedure

Participants were presented with filtered sentences during sparse acquisition fMRI scanning. On each trial, 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 using an up-down staircase such that participants rated sentences as intelligible on ~ 50% of trials. Participants performed 45 trials per run along with 5 trials of a visual baseline task that were quasi-randomly intermixed. We originally included the visual baseline task so that contrasts such as 'speech vs. baseline' would include a meaningful baseline condition rather than rest, but here the baseline condition is irrelevant to, and therefore not included in, all subsequent analyses. 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). Acoustic stimuli were amplified using a Dayton DTA-1 portable amplifier and presented diotically over Sensimetrics S14 piezoelectric earphones. Participants adjusted the volume to a comfortable level slightly above that of conversational speech (~75–80 dB SPL).

2.5 Image Acquisition Parameters

Images were acquired at the University of California, Irvine Neuroscience Imaging Center using a Philips Achieva 3T MRI scanner with a 32-channel sensitivity encoding (SENSE) head. A sparse, gradient-echo EPI acquisition sequence was employed (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 × 1.875 × 3 mm, matrix = 128 × 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 × 256).

2.6 Preprocessing

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 motion correction and co-registration to the T1 image. The functional data were then mapped to a merged, standard-topology surface mesh using Freesurfer v5.3 (Fischl, 2012), AFNI, and the “surfing” toolbox v0.6 (https://github.com/nno/surfing; Oosterhof, Wiestler, Downing, & Diedrichsen, 2011). The surface-space data were smoothed to a target level of 4 mm full width at half maximum and scaled to have a mean of 100 across time points subject to a range of 0–200. At each node in the cortical surface mesh, a beta time series was obtained using the least squares separate (LSS) technique (AFNI 3dLSS; Mumford, Turner, Ashby, & Poldrack, 2012). The original surface mesh was produced using linear icosahedral tessellation with 128 edge divides, resulting in a two-hemisphere mesh with 327684 nodes. To reduce computational burden, the surface mesh was subsampled for eight decimation iterations using the function surfing_subsample_surface in the surfing toolbox. This produced a new surface mesh with 78812 nodes. Subsequent analyses were performed on the subsampled surface.

2.7 Analysis

2.7.1 STRF Estimation—The original dimensions of the binary bubbles filters were 165 × 215 pixels, corresponding to 0–16 cyc/kHz and 0–50 Hz on the MPS. For analysis, pixels at MPS locations > 6 cyc/kHz and > 20 Hz were discarded, resulting in filters of size 67 × 86. These were then vectorized to produce feature vectors of length 5762. To reduce the dimensionality of the feature space and eliminate covariance among the individual features, an experiment-wide feature matrix was generated by stacking the feature vectors across trials and participants. This feature matrix was then submitted to principal component analysis (PCA). It was determined that the first 104 components in the orthogonal PCA space explained > 95% of the variance in the original feature space. These components were thus retained, and the remainder discarded. The resultant, experiment-wide matrix of PCA scores, which represented the trial-by-trial bubbles feature vectors in the PCA-reduced space, was then split by participant, resulting in 10 participant-level feature matrices. For a given participant, this yielded a feature matrix, *F*, with number of rows equal to number of trials completed by that participant and number of columns equal to the number of the dimensions in the PCA-reduced feature space (104). At each cortical surface node (voxel-like unit), the beta time series containing a single estimate of activation magnitude (percent signal change) for each bubbles-filtered sentence was z-scored, resulting in a criterion matrix, *C*, with number of rows equal to the number of trials and number of columns equal to the number of surface nodes (78812). For each participant, a STRF matrix (i.e., STRF brain volume) of dimensions 78812 × 252 was obtained as:

$$STRF = F^T C \tag{1}$$

In other words, at each cortical surface node a STRF was calculated as the sum of each feature across trials weighted by the z-scored activation on the corresponding trial. Of note, STRFs can be obtained in the original, full-resolution feature space by multiplying each element of the PCA-reduced STRF with the length-5762 vector of principal component weights for the corresponding principal component. We performed this backward projection into the original space to generate STRF plots and summary measures (2.7.4), computed statistics for second-level STRF alignment in the PCA-reduced space (2.7.3).

2.7.2 STRF Decomposition by Intelligibility—A residualizing procedure was used to decompose STRFs into separate components reflecting the neural response to trials with the same (what we call "within" STRFs) versus different (what we call "between" STRFs) intelligibility ratings. To accomplish this, a regression model with a single effect-coded predictor reflecting the behavioral response on each trial (1 = rated as intelligible, −1 = rated as unintelligible) was fitted at each cortical surface node where the criterion variable was the z-scored beta time series at that node. Two output time series were obtained from the regression: (1) the predicted values, $\hat{y}$, reflecting the mean activation to intelligible vs. unintelligible trials, respectively; and (2) the residuals, r, reflecting activation across trials with the main effect of intelligibility removed. Across all cortical surface nodes, this yielded two new criterion matrices, $C_{\hat{y}}$ and C_r. The matrix C_r was further divided into separate matrices, C_{r_intel} and $C_{r\_unintel}$, reflecting activation across trials rated as intelligible and unintelligible, respectively. The feature matrix, F, was similarly decomposed into F_{intel} and $F_{unintel}$. Three additional STRF matrices were then obtained as follows:

$$STRF_{between} = F^T C_{\hat{y}} \tag{2}$$

$$STRF_{within\_intel} = F^T_{intel} C_{r\_intel} \tag{3}$$

$$STRF_{within\_unintel} = F^T_{unintel} C_{r\_unintel} \tag{4}$$

This decomposition has the advantage that STRFs obtained from Eq. 1 must be equal to the sum of STRFs obtained from Eqs. 2–4. However, we cannot immediately conclude that "between" effects of intelligibility reflect high-level speech or linguistic processing, nor can we immediately conclude that such effects reflect general acoustic tuning. What is certain is that "between" STRFs must, by definition, reflect those MPS regions that support intelligibility behaviorally (Venezia et al., 2016; Venezia et al., 2019a; Venezia et al., 2019b). Conversely, it is likely that "within" STRFs reflect general acoustic tuning at some level, although such tuning may still be modulated by the experimental setting (i.e., stimuli that are exclusively speech and a task that requires subjective ratings of the speech according to intelligibility). Moreover, the "within" STRFs must, by definition, be biased away from those MPS regions described by the "between" STRFs.

2.7.3 Second-Level Assessment of STRF Alignment—The objective of STRF analysis at the second level is to determine whether the spectrotemporal response patterns exhibited at a given cortical surface node are reliable across subjects. Previously, we took the approach of calculating a t-statistic separately for each STRF feature, applying false discovery rate correction across features (Benjamini & Hochberg, 1995), clustering the data at corrected $p < 0.05$, and thresholding at the cluster level by obtaining a null distribution of max cluster sizes after sign-flipping the individual-participant STRFs in random order (Venezia et al., 2019b). With 10 participants, we previously used all possible sign-flip orders to form a null distribution of length 1022 (2^10 – 2). This method was sufficiently powered to detect large clusters of voxels with STRFs in the auditory cortex and immediate surrounds, but not small clusters that might be expected in nonspeech or speech-motor brain regions.

Here, we take an approach similar to Das and colleagues (2020) who used the multivariate, one-sample Hotelling T^2 test with threshold free cluster enhancement (TFCE; Smith & Nichols, 2009) to assess significance of an entire feature vector at the second level. In our case, this corresponds to a test for a reliable pattern of STRF responses across participants at a given cortical surface node after correcting for multiple comparisons. However, we will replace the Hotelling T^2 with a recently developed nonparametric alternative (Wang, Peng, & Li, 2015) that is appropriate for cases where the dimensionality of the feature space is high relative to the number of observations, and where the data may not be normally distributed. For a data matrix with n observations of p variables, the asymptotic null distribution ($n, p \to \infty$) of the test statistic is standard normal. Thus, for the present application, large positive values indicate the individual-participant STRFs at a given cortical surface node are reliably concentrated in a similar region of the unit hypersphere and large negative values indicate that individual-participant STRFs are uniformly distributed about the unit hypersphere. This test statistic, which we shall call Z, was calculated at each cortical surface node. A null distribution of Z images was formed by first randomly permuting the trial order (rows) of the bubbles feature matrix, F, and then obtaining Z (Eq. 1) at each surface node (Stelzer, Chen, & Turner, 2013). In all, 10,000 permutations resulted in a null distribution of 10,000 Z images. Figure 2 compares the empirical null distribution of Z to the standard normal distribution. Relative to the standard normal, the empirical null distribution of Z is right-skewed and highly kurtotic, which is the expected behavior under the null when n or p is not sufficiently large. However, this is not problematic given we adopted TFCE, a nonparametric approach to cluster-level statistical testing and multiple comparison correction. Specifically, the original Z image and the null Z images were submitted to TFCE using CoSMoMVPA v1.1.0 (Oosterhof, Connolly, & Haxby, 2016). A node-wise, uncorrected p-value image was then obtained as the proportion of null TFCE values exceeding the true TFCE value at each node. The resultant p-value image was then corrected for multiple comparisons using the false discovery rate procedure (Benjamini & Hochberg, 1995; Chumbley, Worsley, Flandin, & Friston, 2010; Leblanc, Dégeilh, Daneault, Beauchamp, & Bernier, 2017).

We chose TFCE for cluster-level analysis because: (i) TFCE obviates the requirement to choose an initial cluster forming threshold; (ii) TFCE maintains sensitivity to small local clusters with strong activation and large, spatially diffuse clusters with relatively weak activation, and (iii) TFCE is nonparametric (Smith & Nichols, 2009), which was important given the non-Gaussian form of the null distribution of the test statistic (Figure 2). A one-tailed test was employed because we were only interested in detecting STRFs with reliably similar – as opposed to reliably dissimilar – response properties across participants. Our primary interest was to apply this analysis to STRFs derived from Eq. 1. However, uncorrected Z maps were also obtained for STRFs derived from Eqs. 3–4 ("within" STRFs). These maps must be interpreted with caution because "within" STRFs are necessarily biased, as described above. No Z map was generated for "between" STRFs because the values of Z would be overwhelmed by bias resulting in unreliable statistics. For "between" STRFs, a map of the maximum STRF amplitude was obtained after averaging the STRFs across participants.

2.7.4 Region of Interest Analyses—Our multivariate analysis of STRF alignment with TFCE revealed five significant clusters outside the auditory cortex: left dPM, left IFG, calcS, right ventral speech-premotor cortex (vPM) and right supplementary motor area (SMA). For each of these regions, we obtained the mean value of Z across all nodes in the region for the overall $STRF$ (Eq. 1), $STRF_{within\_intel}$ (Eq. 3), and $STRF_{within\_unintel}$ (Eq. 4). We also obtained the mean across all nodes in the region of the peak pixel intensity of $STRF_{between}$ (Eq. 2). We then obtained region level averages of $STRF_{between}$, $STRF_{within\_intel}$, and $STRF_{within\_unintel}$. At each pixel in these region-level average STRFs, we obtained an importance index by calculating the sum, pixel by pixel, of the absolute pixel intensity across the STRF subtypes ($STRF_{between}$, $STRF_{within\_intel}$, and $STRF_{within\_unintel}$) and expressing the contribution to this sum from each subtype as a proportion, resulting in three "proportion images", one for each STRF subtype. We then multiplied these proportion images, pixel by pixel, with the absolute value of the region-level average overall $STRF$ (Eq. 1). This produced three "importance images" for the three STRF subtypes. To obtain an overall measure of importance, the mean across-pixel importance was obtained each STRF subtype, and these means were scaled to sum to one. In other words, the importance index expresses the relative contribution of each decomposed STRF (Eqs. 2–4) to the overall STRF (Eq. 1). These overall importance indices were obtained separately for each region of interest. Finally, for each region of interest, we obtained a leave-one-out (LOO) estimate of the overall $STRF$'s ability to predict the region's beta time series within intelligible and unintelligible trials, respectively. Specifically, for each cortical surface node in a given region, the beta time series, y, was obtained and a LOO predicted time series, $\hat{y}$, was calculated as follows. In each LOO iteration, the beta value, y_{LOO}, and feature vector, F_{LOO}, from a single trial were held out and $STRF_{LOO}$ was produced from the remaining data using Eq. 1. The held-out data point was then predicted using:

$$\hat{y}_{LOO} = STRF_{LOO}^{T} F_{LOO} \tag{5}$$

Iterating over all trials in order, $\hat{y}$ was obtained by appending $\hat{y}_{LOO}$ to the accumulated $\hat{y}$ after each iteration. A Bayesian hierarchical linear mixed model was then employed to determine the ability of $\hat{y}$ to predict y across participants. The form of the model was:

$$y = intel + \hat{y}_{intel1} + \hat{y}_{intel0} + (intel + \hat{y}_{intel1} + \hat{y}_{intel0} \mid sub) \tag{6}$$

where $intel$ is a binary predictor encoding trial-by-trial intelligibility ratings (1 = intelligible, 0 = unintelligible), $\hat{y}_{intel1}$ is the product of $intel{=}{=}1$ and $\hat{y}$, and $\hat{y}_{intel0}$ is the product of *intel==0* and $\hat{y}$. Prior to fitting the model, *y* and $\hat{y}$ were z-scored within each participant. In Eq. 6, terms to the right of equality are fixed (non-parenthetical; group level) or random (parenthetical; participant level). This model is equivalent to a model with a main effect of *intel* and the two-way interaction of *intel* and $\hat{y}$, but in Eq. 6 the interaction is instead coded as simple effects of $\hat{y}$ within intelligible and unintelligible trials, respectively. Note, here and in subsequent Eqs. 7–8, we use *R* model formula notation in which the names of the predictor variables in the fixed and random effects design matrices are specifically enumerated, but the associated regression coefficient terms (e.g., β_1*intel $+ \beta_2 * \hat{y}_{intel1}$, etc.) are suppressed. The model was fitted using the *BayesFactor* package v0.9.12–4.2 in *R* v3.6.1 (R Core Team, 2019). The prior scale was set to 'medium' for fixed effects and 'nuisance' for random effects. The Bayes factor was obtained for the fixed effects of $\hat{y}_{intel1}$ and $\hat{y}_{intel0}$ using a model comparison approach (Morey & Rouder, 2011). That is, separate null models were fitted without the fixed effects $\hat{y}_{intel1}$ and $\hat{y}_{intel0}$, respectively, and a Bayes factor was obtained showing the ratio of the evidence in favor of the full model (Eq. 6) versus each null model. We refer to these quantities as BF_{intel} and $BF_{unintel}$, where, controlling for "between" effects, values above one reflect increasing evidence that $\hat{y}$ has some predictive validity with respect to *y* and values below one reflect increasing evidence that $\hat{y}$ has no predictive validity with respect to *y*. The direction of this predictive validity is ambiguous in the Bayes factor and must be determined from the sign of the fixed effects regression coefficients. To generate region-level summaries, the geometric means of BF_{intel} and $BF_{unintel}$ were obtained across all nodes in the region. For non-auditory regions of interest, we also obtained the 80th percentile of BF_{intel} and $BF_{unintel}$ across all nodes in the region to identify cases where, despite a relatively low geometric mean, a nontrivial proportion (at least 20%) of nodes demonstrated high predictive validity.

2.7.5 Beta Time Series Functional Connectivity—

For left dPM, left IFG, and calcS, we were interested in whether and which auditory cortical regions were functionally connected with these non-auditory regions of interest. Therefore, we performed a whole brain beta time series functional connectivity analysis for each of these three regions of interest. First, the mean beta time series from a given region of interest was extracted and z-scored within participants. A first-level linear regression model was then fitted as:

$$y_n = intel + y_{seed} + intel * y_{seed} \tag{7}$$

where y_n is the z-scored beta time series at a given cortical surface node, *intel* is an effect-coded predictor reflecting intelligibility ratings (1 = intelligible, −1 = unintelligible), y_{seed} is the z-scored beta time series from the seed region, and '*' denotes the two-way interaction. No intercept was included in the model because y_n was z-scored. The terms of interest in Eq. 7 were y_{seed} and $intel * y_{seed}$. The first-level regression coefficients for these terms were thus analyzed at the second level via one-sample t-test with false discovery rate correction (Benjamini & Hochberg, 1995). Second-level connectivity maps were thresholded at corrected $p < 0.01$. For $intel * y_{seed}$, no significant surface nodes were detected after correction for multiple comparisons. Before concluding that no significant interaction of connectivity by intelligibility rating was present, we ran an additional Bayesian hierarchical model of the form:

$$y_n = intel + y_{seed} + intel * y_{seed} + (intel + y_{seed} + intel * y_{seed} \mid sub) \quad (8)$$

This model can be interpreted just as Eq. 7, except it has been extended to a hierarchical (i.e., first-plus-second level) framework as in Eq. 6. The model was analyzed using the *BayesFactor* package just as explained for Eq. 6. The term of interest was the fixed effect of $intel * y_{seed}$. Therefore, at each cortical surface node we produced a Bayes factor showing the ratio of the evidence in favor of the full model (Eq 8) and a null model without the fixed effect of $intel * y_{seed}$. This interaction Bayes Factor map was thresholded at |logBF| > 1.6 ($BF_{full} > 5$ or $BF_{null} > 5$) for visualization.

For each of the three seed regions of interest – left dPM, left IFG, calcS – subregions of the auditory cortex were identified containing only those surface nodes that (a) were significantly correlated with the seed at the second level (y_{seed} in EQ. 7; false-discovery-rate-corrected p < 0.01) and (b) had second-level correlations with the seed more than one standard deviation above the mean among auditory-cortical surface nodes with significant STRF alignment. Further, for left dPM and left IFG, subregions of the auditory cortex were identified containing only those surface nodes whose second-level interaction of seed connectivity by intelligibility rating were significant (Eq. 8, $BF_{full} > 5$). For each of these auditory-cortical subregions, average STRFs were calculated using Eqs. 1–4, and all summary measures described in the 'Region of Interest Analyses' subsection were tabulated.

3. Results

3.1 Significant STRF Alignment Beyond the Auditory Cortex

The results of our primary second-level analysis, which probed for reliable cross-subject alignment of STRFs (see Eq. 1), are shown in Figure 3A. In the auditory cortex and surrounds, this analysis replicated our previous results (Venezia et al., 2019b) with significant STRF alignment in the bilateral supratemporal plane, planum temporale, and superior temporal gyrus/sulcus (STG/S). Additionally, a new cluster emerged in the right pSTS that was disjoint from the main cluster of significant STRF responses in the right auditory cortex. This suggests that our analytic method was sensitive enough to detect relatively small clusters with strong, selective responses to the spectrotemporal modulations in speech. Indeed, unlike our previous result, we also detected small clusters of significant STRF responses outside the auditory cortex, with notable clusters in the left dPM, left IFG, right SMA, right vPM, and bilateral calcS.

For those cortical surface nodes with significant cross-subject STRF alignment in Figure 3A, Figure 3B–D plots maps of "decomposed" STRF properties – specifically the maximum amplitude of the "between" STRF derived from the contrast of intelligible vs. unintelligible speech (Eq. 2), and the strength of STRF alignment within trials rated as intelligible (Eq. 3) and unintelligible (Eq. 4), respectively. The largest between effects (Figure 3B) were observed in classic auditory speech regions including the planum temporale, anterior and posterior STG, and dorsal STS bilaterally, as well as the left ventral pSTS and middle temporal gyrus. Outside the auditory cortex, the largest between effects were observed in the left IFG and right SMA (in fact, right SMA showed a large *negative* between effect as will be described below). All regions were modulated at least somewhat by intelligibility. In general, STRF alignment was poor within trials rated as intelligible (Figure 3C), with notable exceptions in the posterior planum temporale and anterior STS bilaterally. Among non-auditory areas, modest STRF alignment within intelligible trials was observed only in the calcS. A different pattern was observed for unintelligible trials (Figure 3D), where strong alignment was observed throughout the supratemporal plane and STG bilaterally. In non-auditory areas, the strongest alignment within unintelligible trials was observed in the left dPM, and alignment increased in strength within left dPM along a gradient from anterior to posterior (i.e., with maximum strength nearest to the central sulcus). Alignment ranging in strength from modest to moderate was also observed in the calcS, but STRFs in the other non-auditory areas were generally poorly aligned within unintelligible trials.

To visualize STRF responses in non-auditory regions, we obtained regional averages of second-level STRFs (overall and decomposed) for left dPM (Figure 4A), left IFG (Figure 4B), calcS (Figure 4C), right SMA (Figure 4D), and right vPM (Figure 4E). The overall STRFs ($STRF$, Figure 4, left column) are annotated with regional STRF summary measures including strength of alignment (Z) and ability to predict activation on held out data separately for intelligible (BF_{intel}) and unintelligible ($BF_{unintel}$) trials. The decomposed STRFs (Figure 4, right three columns) are annotated with regional STRF summary measures including max amplitude ($STRF_{between}$, second column), strength of alignment (Z; $STRF_{within\_intel}$ and $STRF_{within\_unintel}$, right two columns), and relative importance (all decomposed STRFs, right three columns). Several patterns are noteworthy. First, all five non-auditory regions were strongly driven by MPS patterns associated with speech intelligibility (0–10 Hz, 0–2 cyc/kHz) as visible in $STRF$ and $STRF_{between}$. However, this pattern was positive (intelligible > unintelligible) in left dPM, left IFG, and calcS, and negative (unintelligible > intelligible) in right SMA and right vPM. A negative STRF region reflects increased activation when that region of the MPS was removed from the signal. In other words, right SMA and right vPM activated more strongly when speech was more distorted, indicating a top-down or difficulty-driven response profile. Since this profile was not of primary interest, right SMA and right vPM will not be discussed further. Second, among the three regions with positive STRFs, left IFG responded most prominently to MPS patterns associated with intelligibility, while left dPM and calcS responded to MPS patterns associated with intelligibility and MPS patterns associated with vocal pitch (see Figure 1A). Indeed, left IFG had a larger amplitude in and more relative importance ascribed to $STRF_{between}$ compared to left dPM and calcS. Third, left dPM and calcS showed a non-trivial level of cross-subject alignment in $STRF_{within\_unintel}$, while left IFG did not; this was driven largely by reliable responses to vocal pitch, but to a lesser extent by responses to low spectral modulation rates (0–2 cyc/kHz) at relatively high temporal modulation rates (5–15 Hz). Interestingly, calcS showed a similar response profile in $STRF_{within\_intel}$, but left dPM and left IFG, like most of the auditory cortex, did not demonstrate strong STRF responses within intelligible trials. Relative importance was spread rather evenly across $STRF_{between}$, $STRF_{within\_intel}$, and $STRF_{within\_unintel}$ in calcS; in left IFG relative importance was weighted heavily toward $STRF_{between}$, while in left dPM relative importance was weighted heavily toward $STRF_{between}$ and $STRF_{within\_unintel}$. The only regions in which LOO-predicted activation was significantly associated with true activation within unintelligible trials was left dPM (geometric mean of $BF_{unintel}$ = 4.1, 80th percentile = 43.6). However, this association was also present in a nontrivial proportion of surface nodes in left IFG (80th percentile of $BF_{unintel}$ = 7.7). LOO-predicted activation was significantly associated with true activation within intelligible trials in calcS (geometric mean of BF_{Intel} = 8.4, 80th percentile = 91.0) and in a nontrivial proportion of surface nodes in left dPM (80th percentile of BF_{Intel} = 13.9). However, inspection of the fixed effects regression coefficients in the LOO-predictive models revealed this association to be *negative* within intelligible trials. This means that *STRF* weights in calcS and left dPM should be sign-reversed in at least some subregion of the MPS to yield accurate predictions within intelligible trials.

3.2 Beta Time Series Functional Connectivity

The primary motivation behind our functional connectivity analysis was to determine for each of the non-auditory regions with positive-valued STRFs – left dPM, left IFG, and calcS – whether and which areas of the auditory cortex were significantly correlated in their activation patterns across the course of the experiment. Secondarily, we wanted to know whether such patterns of functional connectivity were modulated by intelligibility. To these ends, we conducted whole brain regressions at the first level with non-auditory seed time series, binary intelligibility judgments, and their two-way interaction as predictors. The seed time series and intelligibility-by-seed interaction were the predictors of interest. In fact, the seed was significantly correlated with activation time series in the auditory cortex for all three regions of interest (Figure 5A–C). For left dPM and calcS, maximal correlations were observed in Heschl’s gyrus/sulcus and the immediately adjacent STG. Conversely, for left IFG maximal correlations were observed in the lateral/ventral STG, STS, and middle temporal gyrus. Moreover, left dPM and calcS were correlated with widespread sensorimotor networks across much of the cortical surface, while left IFG was correlated with a more restricted network of classic peri-Sylvian and inferior frontal speech-language networks. Together, these findings suggest that left dPM and calcS operate at a lower hierarchical level of information processing than left IFG.

For the intelligibility-by-seed interaction, no significant effects were detected at the chosen threshold of FDR-corrected $p < 0.01$. Therefore, we conducted a follow-up Bayesian analysis on the interaction effect to determine the extent to which the evidence favored the null hypothesis (no interaction) versus the alternative hypothesis (presence of an interaction). The results of a Bayesian hierarchical regression model are plotted for left dPM and left IFG in Figure 5D and Figure 5E, respectively. Specifically, the whole-brain plots show maps of the log Bayes factor (logBF) for the fixed (group-level) effect of the two-way interaction, where negative values reflect evidence in favor of the null and positive values reflect evidence in favor of the alternative. The maps are thresholded at $|\text{logBF}| > 1.6$ ($BF_{full} > 5$ or $BF_{null} > 5$). In general, evidence favored the null (negative logBF) across most of the cortical surface for both left dPM and left IFG. However, evidence in favor of the alternative was observed in some regions of the auditory cortex for both left dPM and left IFG. For left dPM, these effects were maximal in the left STG just lateral to Heschl’s gyrus. For left IFG, these effects were maximal in regions of the STG lateral and posterior to those observed for left dPM. In both cases, inspection of the regression coefficients showed that significant interactions were driven by relatively greater functional connectivity with the seed on unintelligible versus intelligible trials. Similar patterns were observed for calcS (not shown) but without any substantial evidence in favor of the alternative within the auditory cortex. In general, widespread evidence in favor of the null suggests that effects were dominated by intrinsic connectivity with the seeds in the broader functional context of speech processing (regardless of intelligibility).

3.3 STRFs in Auditory Regions Identified via Functional Connectivity

To visualize the STRF properties of the auditory regions that were maximally functionally connected to our non-auditory regions of interest (left dPM, left IFG, and calcS), we obtained second level average STRFs for auditory regions that (a) were significantly correlated with the seed at the second level and (b) had second-level correlations with the seed more than one standard deviation above the mean among those auditory-cortical surface nodes with significant cross-subject STRF alignment. These regional average STRFs are plotted in Figure 6A–C, each annotated with regional summary measures as in Figure 4. It is immediately apparent that the results pattern just as in Figure 4. Specifically, just as for left IFG, the auditory regions most correlated with left IFG responded primarily to MPS regions associated with speech intelligibility as indicated by increased relative importance/amplitude for $STRF_{between}$ and relatively decreased (but still significant) alignment for $STRF_{within\_unintel}$. On the other hand, just as for left dPM and calcS, the auditory regions most correlated with left dPM and calcS responded to MPS regions associated with both intelligibility and vocal pitch. In this case, pitch responses were driven almost entirely by $STRF_{within\_unintel}$, which mirrors the pattern in left dPM but not calcS which also responded to pitch in $STRF_{within\_intel}$. In other words, auditory regions that responded strongly to pitch (Figure 6A–B) only did so within trials rated as unintelligible by the listeners. The strength of this response was much greater in auditory regions than in left dPM and calcS (Z ~= 10 vs. Z ~= 2). Auditory regions that did not respond strongly to pitch (Figure 6C) were similar to left dPM in terms of the strength of cross-subject STRF alignment within unintelligible trials (Figure 4A; Z = 2.6 vs. Z = 2.4). For all these auditory regions (Figure 6A–C), there was a strong effect of speech intelligibility (large amplitude for $STRF_{between}$) yet negligible STRF responses within intelligible trials ($STRF_{within\_intel}$). Correspondingly, LOO predictions for all auditory regions were significant within unintelligible (all $BF_{unintel}$ > 1e3) but not intelligible trials (all BF_{intel} <= 1). In general, $BF_{unintel}$ was orders of magnitude larger in these auditory regions than in their corresponding non-auditory regions, though this effect was more pronounced for the early auditory regions connected to left dPM and calcS than for the downstream auditory regions connected to left IFG. This suggests $BF_{unintel}$ is a good proxy of the extent to which a given region has auditory-like properties (i.e., acoustically driven STRF responses that explain a relatively large proportion of signal variance).

We also obtained regional average STRFs for those auditory regions whose functional connectivity with the seed region (in this case, left dPM and left IFG) was significantly modulated by speech intelligibility (Figure 6D–E). These regions, located in more dorsal versus more ventral regions of the STG lateral to Heschl’s gyrus for left dPM and left IFG, respectively, showed a very similar response profile to the auditory regions that were maximally correlated with left dPM (Figure 6A). Specifically, STRFs exhibited responses to MPS regions associated with both intelligibility and vocal pitch, strong responses were observed for $STRF_{within\_unintel}$ but not $STRF_{within\_intel,}$ and strong effects of intelligibility (large amplitude for $STRF_{between}$) were counterbalanced by a large relative importance for $STRF_{within\_unintel}$. As with the auditory areas defined by the main effect of connectivity, LOO predictions were significant for unintelligible trials (both $BF_{unintel}$ > 5e7) but not intelligible trials (both BF_{intel} <= 0.9). These effects are unsurprising given that the interaction of functional connectivity with intelligibility was driven by relatively stronger connectivity within unintelligible trials. In other words, we would expect such effects to localize to relatively early auditory regions whose STRF responses are more acoustically driven. However, pitch responses and strength of alignment in $STRF_{within\_unintel}$ were both relatively larger in auditory regions for which this interaction was present with left dPM (Figure 6D) than with left IFG (Figure 6E), which suggests that while both left dPM and left IFG receive input from acoustically driven regions of auditory cortex, left dPM receives such input from earlier and more strongly acoustically driven auditory regions. However, $BF_{unintel}$ was extremely large for both interaction-defined auditory regions, suggesting a predominantly acoustically driven response in both cases.

4. Discussion

In the present study, we re-analyzed fMRI data from which speech-driven spectrotemporal receptive fields (STRFs) could be estimated in young, normal-hearing listeners (Venezia et al., 2019b). Specifically, we developed a multivariate analysis combined with threshold free cluster enhancement to increase sensitivity to cross-subject STRF alignment at the second level, with the intention of revealing STRF responses beyond classical auditory-speech regions in the superior temporal lobe. Indeed, our primary analysis replicated our earlier findings in superior temporal regions (Venezia et al., 2019b) while identifying five non-auditory regions with reliable STRF alignment at the second level (Figure 3): left dorsal speech-premotor cortex (dPM), left inferior frontal gyrus (IFG), bilateral calcarine sulcus (calcS), right supplementary motor area (SMA), and right ventral speech-premotor cortex (vPM). Of these, right SMA and right vPM were shown to have a negative response to MPS regions associated with intelligible speech (Figure 4), suggesting a top-down profile in which these regions activate with increasing distortion of the acoustic speech signal and thus with increasing difficulty of speech recognition. Though such a response profile is surely of interest in the context of speech recognition in background noise (cf., Binder, Liebenthal, Possing, Medler, & Ward, 2004), it does not reflect canonical, auditory-like STRF responses which were of primary interest here. Therefore, these regions will not be discussed further. However, it is notable that this finding dovetails with a growing body of evidence suggesting that motor speech areas play a role in the processing of degraded speech (D’Ausilio et al., 2012; Du, Buchsbaum, Grady, & Alain, 2014, 2016; Erb & Obleser, 2013; Evans & Davis, 2015; J. E. Peelle, I. Johnsrude, & M. H. Davis, 2010; Szenkovits, Peelle, Norris, & Davis, 2012; Wild et al., 2012). The remaining three non-auditory regions – left dPM, left IFG, and calcS – showed auditory-like STRF profiles though with several important differences between them.

In our previous publication of this data set (Venezia et al., 2019b), we found that a defining characteristic of early auditory areas in the supratemporal plane was sensitivity to MPS regions associated with vocal pitch in addition to MPS regions associated with speech intelligibility, whereas intelligible speech much more selectively drove responses in downstream regions of the STG/S. Here, we observed similar patterns in non-auditory regions, with left dPM and calcS responding well to vocal pitch and intelligible speech while left IFG responded more selectively to intelligible speech (Figure 4). Interestingly, calcS responded to vocal pitch within both intelligible and unintelligible trials, while left dPM responded to vocal pitch only within unintelligible trials (Figure 4, compare $STRF_{within\_intel}$ to $STRF_{within\_unintel}$). However, the STRFs estimated within calcS were not able to generate good leave-one-out (LOO) predictions of activation within unintelligible trials, a hallmark auditory profile. On the other hand, the STRFs estimated within left dPM were able to generate good LOO predictions within unintelligible trials (geometric mean of $BF_{unintel}$ = 4.1, 80th percentile = 43.6), and the STRFs estimated within left IFG were able to generate marginally good LOO predictions within unintelligible trials (geometric mean of $BF_{unintel}$ = 1.4, 80th percentile = 7.7).

In our beta time series functional connectivity analysis, left dPM and calcS were maximally connected with early auditory regions in the supratemporal plane, while left IFG was maximally correlated with downstream regions in the STG/S and middle temporal gyrus (Figure 5, A–C). As expected, and consistent with our prior publication (Venezia et al., 2019b), the STRFs in these auditory regions mirrored those in their non-auditory counterparts: sensitivity to vocal pitch and intelligible speech in the auditory regions maximally connected with left dPM and calcS, and more selective responses to intelligible speech in the auditory regions maximally connected with left IFG (Figure 6, A–C). All of these auditory regions showed strong STRF responses within unintelligible but not intelligible trials, and their STRFs generated good LOO predictions within unintelligible but not intelligible trials. However, these effects were stronger in the early auditory regions connected with left dPM and calcS than in the STG/S regions connected with left IFG. Interestingly, connectivity with auditory regions was modulated by intelligibility for left dPM and left IFG but not calcS. Regions of the STG just lateral to Heschl's gyrus were more strongly connected with left dPM and left IFG within unintelligible trials (Figure 5, D–E). These effects localized to slightly more dorsal STG regions for left dPM relative to left IFG. In general, these STG regions responded similarly to the early auditory regions connected with left dPM and calcS, though the response to vocal pitch was larger in STG regions showing an intelligibility-by-connectivity interaction with left dPM relative to those showing such an interaction with left IFG (Figure 6, D–E). The STRFs in these regions generated very good LOO predictions of activation but only within unintelligible trials.

Taken together, these results suggest that only left dPM demonstrated a clear "auditory-like" response profile. Like early auditory regions, left dPM responded to both vocal pitch and intelligible speech, responded to vocal pitch primarily within unintelligible trials, demonstrated strong cross-subject STRF alignment within unintelligible but not intelligible trials, and generated good LOO predictions within unintelligible trials. Moreover, left dPM was maximally connected with early auditory regions and showed an increase in connectivity with certain early auditory regions on unintelligible trials during which vocal pitch more strongly activated those auditory regions. On the other hand, left IFG did not respond strongly to vocal pitch, generated only marginally good LOO predictions within unintelligible trials, and was maximally connected with downstream auditory regions that did not strongly encode vocal pitch. Finally, calcS responded to vocal pitch and was maximally connected with early auditory areas that strongly responded to pitch, but unlike in auditory regions this pitch response was not restricted to unintelligible trials. Moreover, calcS generated poor LOO predictions within unintelligible trials, which suggests that while activation in calcS was modulated by the relative presence or absence of vocal pitch, this modulation accounted for a trivial proportion of the overall signal variance. Like their auditory counterparts, all three non-auditory regions were strongly modulated by speech intelligibility (Figures 4/6, $STRF_{between}$). Below, we explore what these patterns suggest about the computational roles played by these non-auditory regions during perception of continuous speech.

4.1 Left dPM: Sensitivity to Vocal Pitch in the Dorsal Laryngeal Motor Cortex

In the Introduction, we identified two prominent fMRI studies showing that a region of the left dorsal premotor cortex activates during speech perception and production (Buchsbaum et al., 2001; Wilson et al., 2004). We also noted subsequent work showing that the same region is part of a precentral/central complex that forms the human dorsal laryngeal motor cortex or ‘dLMC’ (Brown et al., 2008). In the present study, we found that left dPM exhibits auditory-like STRF properties including a substantial response to vocal pitch. In Figure 7, we reproduce the lateral left hemisphere plot from Figure 3A showing regions with significant cross-subject STRF alignment, here overlaid with 2mm spheres at the peak dorsal premotor coordinates from Buchsbaum et al. (2001), Wilson et al. (2004), and Brown et al. (2008). Considering the different samples and methods employed in these studies together with the fact that second level analysis introduces a degree of anatomical uncertainty, the proximity of activations across studies is, in our view, compelling support for a common underlying mechanism. Indeed, we suggest that left dPM, as identified in the present study, is the gyral component of the left dLMC.

Recent work suggests there is a dual representation of the larynx in human motor cortex, with a more ventral, premotor representation in the posterior frontal operculum (vLMC), and a more dorsal, primary motor representation in the central sulcus (dLMC) with close proximity to the primary motor representation of the lips (Bouchard, Mesgarani, Johnson, & Chang, 2013; Breshears et al., 2015; Chang, Niziolek, Knight, Nagarajan, & Houde, 2013; Dichter et al., 2018; Eichert, Papp, Mars, & Watkins, 2020; Galgano & Froud, 2008; Grabski et al., 2012; Olthoff, Baudewig, Kruse, & Dechent, 2008; Terumitsu, Fujii, Suzuki, Kwee, & Nakada, 2006). Current theory (Belyk & Brown, 2017; Brown et al., 2020; Eichert et al., 2020) poses that human vLMC is evolutionarily homologous with the laryngeal motor cortex in nonhuman primates, which lack the sophisticated vocal learning and voluntary laryngeal motor control abilities of humans, and whose laryngeal motor cortex is entirely confined to the premotor vLMC. The dLMC is thus uniquely human, which implies an evolutionary ‘duplication-and-migration’ (Belyk & Brown, 2017) allowing for voluntary laryngeal control. Further, it has been suggested that dLMC preferentially controls the extrinsic muscles of the larynx, which move the entire larynx vertically within the airway and play a key role in complex pitch modulations of the sort common in human speech and singing; conversely, the vLMC preferentially controls the intrinsic muscles of the larynx, which perform rapid tensioning and relaxation of the vocal folds at the onset and offset of vocalizations (i.e., serve as a coarse ‘on-off’ switch for phonation; Eichert et al., 2020; but see Belyk et al., 2020; Belyk & Brown, 2014). Interestingly, stimulation of vLMC produces speech arrest (Chang et al., 2017) while stimulation of dLMC produces involuntary vocalization and disruption of fluent speech (Belkhir et al., 2021; Dichter et al., 2018), and dLMC appears to be selectively involved in producing fast pitch changes of the sort that generate stress patterns within a sentence (Dichter et al., 2018).

While dLMC localizes most prominently to BA 4 deep in the central sulcus (Simonyan, 2014), numerous studies have now shown that it extends into BA 6 in a neighboring region of the dPM (cf., Brown et al., 2020). This gyral component of dLMC seems to be preferentially engaged during perception. As noted early on by Wilson et al. (2004), although sensory speech activations overlap with motor speech activations in BA 4, peak sensory activation occurs about 5 mm anterior to peak motor activations (i.e, in BA 6). Intracranial studies have consistently implicated this gyral component of dLMC in speech perception and production (Belkhir et al., 2021; Berezutskaya et al., 2020; Bouchard et al., 2013; Breshears et al., 2015; Chang et al., 2017; Chang et al., 2013), particularly with respect to encoding and production of vocal pitch (Cheung et al., 2016; Dichter et al., 2018). This work dovetails with the present study in which we have identified a region of left dPM that: (a) co-locates with gyral dLMC (Figure 7); (b) responds vigorously to vocal pitch (Figure 4); (c) shows maximum functional connectivity with core regions of the auditory cortex that also respond vigorously to vocal pitch; and (d) shows strong acoustically driven responses within speech trials rated as unintelligible.

One curious aspect of the present results is that left dPM responded to pitch selectively within unintelligible trials (Figure 4). While it is tempting to attribute this to acoustic biases introduced by separating the trials according to intelligibility judgments (i.e., we should, by definition, expect lower variance and therefore lower within-category STRF weights for acoustic features that contributed strongly to intelligibility judgments), such biases could not have affected estimation of the response to vocal pitch because vocal pitch did not contribute strongly to intelligibility judgments (CITE Venezia et al 2019). Moreover, it is notable that auditory regions responding to vocal pitch also did so selectively within unintelligible trials (Figure 6). In that sense, the broader issue is to determine why general auditory sensitivity to vocal pitch occurred selectively within trials rated as unintelligible. We suggest this pattern was driven by the intelligibility judgment task itself. Specifically, listeners were asked to make a binary, yes/no rating of intelligibility on each trial. Therefore, within trials rated as intelligible, the brain was able to tune in selectively to those acoustic features that are relevant to perception of intelligible speech (Figures 4/6, the “bulls-eye” in $STRF_{between}$). Those features were adequately, or perhaps even completely, characterized by $STRF_{between}$ and so little to no signal was present in $STRF_{within-intel}$ (i.e., the brain responded rather uniformly within intelligible trials, at least with respect to the acoustic features that support intelligibility). On unintelligible trials, when those features were, by definition, less available, or unavailable in the signal, the brain keyed into other acoustic features of interest, for example vocal pitch or faster temporal modulations (> 10 Hz). Indeed, we suggest that regions responding to vocal pitch within unintelligible trials – including left dPM and core auditory regions – are the same regions that would respond to vocal pitch on intelligible trials if vocal pitch were relevant to the task (e.g., speech recognition with two co-located, simultaneous talkers).

Left dPM deviated from core auditory regions in that its STRF-derived LOO predictions of activation were correlated with true activation within both intelligible and unintelligible trials (Figure 4, first column). However, as noted in the Results, for intelligible trials this association was driven by LOO predictions that were significantly negatively correlated with true activation. In other words, STRF responses in left dPM shifted from a bottom-up profile (positive STRF weights) to a top-down profile (negative STRF weights) within intelligible trials, such that left dPM responded more to intelligible speech when that speech was partially distorted. It is unlikely that this top-down response was related to vocal pitch given that STRF weights on vocal pitch were positive overall for $STRF_{within\_intel}$ in left dPM (Figure 4, third column). Like the entirety of auditory cortex, it appears that left dPM ‘re-tuned’ on intelligible trials to respond selectively to MPS regions associated with intelligibility, except that left dPM responded more strongly when those MPS regions were only partially represented in the stimulus. This suggests that responses in left dPM are highly dynamic, which reinforces our claim that left dPM is, in fact, the gyral component of dLMC, where response properties have been shown to shift adaptively depending on whether speech is perceived or produced (Cheung et al., 2016). Consistent with a top-down role and a role in perceptual processing of vocal pitch, it has been suggested that the left dPM contributes to speech comprehension by generating temporal predictions on the phrasal scale (Keitel, Gross, & Kayser, 2018). However, disruption or damage to left dPM appears to have very small effects on comprehension (Hickok et al., 2008; Krieger-Redwood, Gaskell, Lindsay, & Jefferies, 2013; Rogalsky et al., 2020; Rogalsky, Love, Driscoll, Anderson, & Hickok, 2011), which calls into question a core role for such a predictive mechanism in comprehension.

4.2 Left IFG: Response to Structured Speech Input

Significant cross-subject STRF alignment in the left IFG was observed with a peak response at the boundary between the pars opercularis and pars triangularis (BA 44/45), MNI coordinate [−50, 19, 3]. At the time of this writing, maximal associations with the coordinate nearest this peak in NeuroSynth (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011) were ‘inferior frontal,’ ‘language,’ ‘syntactic,’ ‘semantic,’ ‘nouns,’ ‘word,’ and ‘verb’ (retrieved from https://neurosynth.org/locations/−48_20_2_6/ on 02/03/2021). It is therefore unsurprising that left IFG responded primarily to intelligible versus unintelligible speech (Figure 4, $STRF_{between}$) – proportionally more so than left dPM, calcS, or classic auditory speech regions (Figures 4/6) – and was functionally connected with a classic peri-Sylvian speech-language network (Figure 5). However, two findings suggest the presence of at least a weak ‘acoustic trace’ of the input speech signal in left IFG. First, LOO predicted activation was accurate within unintelligible trials in a nontrivial proportion of IFG surface nodes (80th percentile of $BF_{unintel} = 7.7$), though much less accurate than left dPM or classic auditory speech regions (Figures 4/6). Second, functional connectivity between left IFG and the lateral STG was significantly increased on unintelligible trials (Figure 5E), mirroring the pattern shown by left dPM, though left IFG showed this pattern with STG regions that were relatively more lateral (and presumably downstream) compared to those STG regions showing the same effect with left dPM (Figure 6D–E).

Matchin and Hickok (2020) suggest that left IFG at the BA 44/45 boundary is involved primarily in the computation of linear order at the morphemic level during speech production. Matchin et al. (2017) further suggest that left IFG plays a role in speech comprehension by computing top-down predictions of phrasal nodes or using linearization as a form of syntactic working memory. If these suggestions are correct, one potential hypothesis is that left IFG retains some association with regions of auditory cortex that extract and represent linear sequences of speech sounds (i.e., STG). Under this hypothesis, left IFG would be primarily activated by intelligible speech, though also for a small subset of unintelligible trials on which speech is semi-intelligible and for which phrase level predictions or working memory traces have immediate consequences for parsing the degraded acoustic speech signal (Keitel et al., 2018; Poeppel & Assaneo, 2020). In other words, this morphosyntactic sequencing-based account could explain why left IFG appears to have an STRF-like response (primarily driven by the intelligible versus unintelligible contrast) and why its STRF is able to predict some, but only relatively little, of the variance in left IFG activation on unintelligible trials.

4.3 calcS: Modulation of Visual Cortex by Voice

Though we neither strongly expected nor were primarily interested in cross-modal responses to heard speech, we nonetheless observed significant cross-subject STRF alignment in the bilateral calcS with a peak response localized to the left calcS at MNI coordinate [−9, −88, 10] in BA 17/area V1. Like left dPM, calcS responded significantly to vocal pitch within unintelligible trials, showed maximal functional connectivity with core auditory regions in the supratemporal plane, and responded more to intelligible versus unintelligible speech. However, unlike left dPM, calcS responded to vocal pitch within intelligible trials and did not show increased connectivity with early auditory regions during unintelligible trials. Moreover, STRF-derived LOO predictions of activation in calcS were poor within unintelligible trials. Overall, these results suggest that activation in calcS was modulated weakly and nonspecifically by vocal pitch. They further suggest a mechanism unrelated to speech processing *per se*.

Previous studies have observed acoustically driven activation in early visual areas for both speech (Calvert, Hansen, Iversen, & Brammer, 2001) and nonspeech (Brewer, Barton, Venezia, Saberi, & Hickok, 2013; Hertz & Amedi, 2010; Watkins, Shams, Tanaka, Haynes, & Rees, 2006) sounds (and see Frostig, Chen-Bee, Johnson, & Jacobs, 2017). In the speech domain, unimodal auditory and cross-modally enhanced activation in calcS has been observed when subjects make judgments about gender and person identity from face-voice stimuli (Joassin, Maurage, & Campanella, 2011a; Joassin et al., 2011b), suggesting that calcS plays a role in processing of voice. Interestingly, regions of the auditory cortex that are modulated by vocal emotion appear to be both structurally and functionally connected with early visual areas (Ethofer et al., 2012). Cross-modal enhancement of early visual evoked responses by vocal emotion has also been observed (Brosch, Grandjean, Sander, & Scherer, 2009). When emotion is conveyed by music, functional connectivity between calcS and auditory cortex is modulated by emotional valence (Koelsch, Skouras, & Lohmann, 2018). Moreover, background acoustic noise has been shown to interact with neural responses to different musically conveyed emotions in calcS (Skouras, Gray, Critchley, & Koelsch, 2013). These music-based studies show that calcS is preferentially modulated by joyful compared to fearful musical excerpts (Koelsch & Skouras, 2014; Koelsch et al., 2018; Skouras et al., 2013). Anecdotally, we can report that bubbles-filtered speech, as employed in the present study, can sound ominous and unpleasant when MPS regions encoding vocal pitch are removed from the signal. However, removal of vocal pitch does not affect intelligibility, which may explain why calcS responded to vocal pitch within intelligible and unintelligible trails. We therefore speculate that vocal emotion is driving STRF responses in calcS, which, given the fact that vocal emotion was not relevant to the task, could explain why calcS was only weakly modulated by vocal pitch.

Just as with acoustically conveyed emotion, it has long been known that calcS and surrounding early visual regions respond selectively to visually conveyed emotions (Pessoa, McKenna, Gutierrez, & Ungerleider, 2002; Vuilleumier, Richardson, Armony, Driver, & Dolan, 2004). More recent work suggests that emotional arousal is signaled using a multisensory code and that neural representations of emotionally significant stimuli in early auditory and visual cortices take a common form (Sievers, Lee, Haslett, & Wheatley, 2019; Sievers et al., 2018). Indeed, early auditory and visual cortices are anatomically connected to one another (Smiley et al., 2007) and to the amygdala (Amaral, Behniea, & Kelly, 2003; Hackett, 2011). In the visual domain, a recent model suggests that emotion influences visual encoding in calcS primarily by modulating attention (Zhang, Japee, Safiullah, Mlynaryk, & Ungerleider, 2016). Therefore, in the context of our auditory bubbles task, we predict a more substantial STRF response to vocal pitch in calcS when the task demands attention to the voice (e.g., with competing talkers) or when the task explicitly involves extraction of vocal emotion (or both).

4.4 Caveats

A potential weakness of the Auditory Bubbles technique is that it introduces extrinisic distortion into the speech signal, thus rendering speech sometimes intelligible and other times unintelligible, or sometimes natural sounding and other times unnatural sounding. In the present study, we have attempted to harness this potential weakness as a strength by decomposing STRFs according to intelligibility judgments and showing that the functional properties of neural speech circuits change depending on the mode of speech processing. However, this effect is likely to be dependent on the task and the ratio of intelligible to unintelligible trials (Wu, Stangl, Zhang, Perkins, & Eilers, 2016). Moreover, we cannot ascertain with certainty whether STRF responses to MPS regions correlated with intelligibility is driven by categorical effects of intellibility or a faithful representation of the underlying acoustic patterns associated with intelligibility. In other words, we would expect to measure a significant STRF-like response using Auditory Bubbles at any brain location for which the computation performed by that region is strongly correlated with speech intelligibility (from acoustics on up to syntax and domain general processing). Therefore, in terms of STRF responses in non-auditory regions, we have attempted to buttress our measurements of the STRFs themselves by examining patterns of functional connectivity of non-auditory regions with classic auditory-speech regions for which computational accounts are more readily available in the published literature (Hackett, 2011; Hickok & Poeppel, 2007; Rauschecker & Scott, 2009). This approach has been informative, but ultimately we cannot with certainty discount even the most extreme alternative explanation which is that STRF responses in non-auditory regions are entirely epiphenomenal, driven, for example, by obligatory feedforward connections between auditory and non-auditory regions or increased engagement of the listener on intelligible versus unintelligible trials or natural-sounding versus unnatural-sounding trials. Future work should examine how STRFs change at different average performance levels and with stimuli that are acoustically comparable to speech but intrinsically unintelligible (e.g., foreign language speech).

Another point that deserves some discussion is our use of subjective, yes-no ratings to determine whether trials were intelligible versus unintelligible. Indeed, this forced listeners to label partially intelligible sentences as entirely (un-)intelligible. To the extent that such partially intelligible sentences were present in the stimulus set, we should expect that neural responses within categories labeled as (un-)intelligible were not entirely uniform with respect to intelligibility. As noted above, the presence of partially intelligible sentences within the subset of items labeled as intelligible by the listeners likely explains why STRF-derived LOO predicted activation was anti-correlated with true activation for intelligible trials in dPM and calcS. We also noted that the presence of partially intelligible sentences within the subset of items labeled as unintelligible by the listeners could explain why STRF-derived LOO predicted activation was positively, though weakly, correlated with true activation for unintelligible trials in IFG. However, we believe the total influence of these partially intelligible sentences on our results was minimal. Specifically, we showed in a previous behavioral study (Venezia et al., 2016) using the exact same stimulus set as the present study, but with objective rather than subjective intelligibility scores (keywords correctly reported for each sentence), that the most common responses (> 50% of trials) produced by bubbles-filtered sentences were either entirely intelligible (all keywords correct) or entirely unintelligible (no keywords correct). Like the present study, that study also used an adaptive procedure to maintain overall intelligibility at 50% (keywords correct, in that case). The mean number of bubbles to reach equilibrium in the keyword task was 53.7 (SD = 8.5, N = 10), and the mean number of bubbles to reach equilibrium in the present study, which used subjective yes/no intelligibility ratings, was 51.2 (SD = 10.9, N = 10). This suggests that both tasks produced very similar perceptual experiences of the stimuli, insofar as similar numbers of bubbles indicate similar amounts of acoustic distortion, so we can infer that the majority of trials rated as (un-)intelligible in the present study were objectively (un-)intelligible.

5. Conclusions

Using Auditory Bubbles and a multivariate analysis of speech-driven STRFs at the second level, we have identified significant cross-subject STRF alignment in left dPM, left IFG, and calcS. We posit that, among these, left dPM is the most ‘auditory-like,’ with a STRF that robustly encodes vocal pitch and predicts a significant amount of the variance in neural activation within unintelligible trials. Indeed, we suggest that left dPM is located within the human dorsal laryngeal motor cortex. Future work is required to examine the relation between sensory- and motor-speech-related activations in this region within the same group of subjects. The left IFG, on the other hand, is the least ‘auditory-like,’ with STRF responses driven primarily by the difference in activation on intelligible versus unintelligible trials, though left IFG remains weakly responsive to speech acoustics and functionally connected with some early auditory areas. We therefore suggest that left IFG plays a top-down role in speech comprehension based on computation of linear order or syntactic working memory that can influence bottom-up sensory processing. Finally, calcS is somewhat ‘auditory-like,’ showing a significant response to vocal pitch and functional connectivity with early auditory areas, but its pitch response is nonspecific and its STRF explains very little variance in neural activation. We therefore conclude that calcS is weakly modulated by vocal pitch, perhaps related to a role in cross-modal processing of emotional cues from sensory signals including speech.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the U.S. Department of Veterans Affairs, Veterans Health Administration, Rehabilitation Research & Development Service Award IK2RX002702 to J.H.V. This material is the result of work supported with resources from 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 U.S. government. 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).

Data Availability

The data and analysis scripts used to produce this manuscript can be obtained at https://osf.io/4p9vh/.

Abbreviations:

AC	auditory cortex
BA	Brodmann area
BF	Bayes factor
calcS	calcarine sulcus
dLMC	dorsal laryngeal motor cortex
dPM	dorsal premotor cortex
EPI	echo planar imaging
FDR	false discovery rate
IFG	inferior frontal gyrus
LOO	leave one out

LSS	least squares separate
MNI	Montreal Neurological Institute
MPS	modulation power spectrum
pOP	pars opercularis
SENSE	sensitivity encoding
SMA	supplementary motor area
STG	superior temporal gyrus
STRF	spectrotemporal receptive field
STS	superior temporal sulcus
TA	acquisition time
TE	echo time
TFCE	threshold free cluster enhancement
TR	repetition time
vPM	ventral premotor

References

Amaral DG, Behniea H, & Kelly JL (2003). Topographic organization of projections from the amygdala to the visual cortex in the macaque monkey. Neuroscience, 118(4), 1099–1120. [PubMed: 12732254]

Arsenault JS, & Buchsbaum BR (2016). No evidence of somatotopic place of articulation feature mapping in motor cortex during passive speech perception. Psychonomic Bulletin & Review, 23(4), 1231–1240. [PubMed: 26715582]

Belkhir JR, Fitch WT, Garcea FE, Chernoff BL, Sims MH, Navarrete E, … Pilcher WH (2021). Direct electrical stimulation evidence for a dorsal motor area with control of the larynx. Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation, 14(1), 110–112.

Belyk M, Brown R, Beal DS, Roebroeck A, McGettigan C, Guldner S, & Kotz SA (2020). Human specific neurophenotype integrates laryngeal and respiratory components of voice motor control.

Belyk M, & Brown S (2014). Somatotopy of the extrinsic laryngeal muscles in the human sensorimotor cortex. Behavioural Brain Research, 270, 364–371. [PubMed: 24886776]

Belyk M, & Brown S (2017). The origins of the vocal brain in humans. Neuroscience and Biobehavioral Reviews, 77, 177–193. [PubMed: 28351755]

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.

Berezutskaya J, Baratin C, Freudenburg ZV, & Ramsey NF (2020). High-density intracranial recordings reveal a distinct site in anterior dorsal precentral cortex that tracks perceived speech. Human Brain Mapping, 41(16), 4587–4609. [PubMed: 32744403]

Bever TGP, D. (2010). Analysis by synthesis: A (re-) emerging program of research for language and vision. Biolinguistics, 4(2–3), 174–200.

Binder JR, Liebenthal E, Possing ET, Medler DA, & Ward BD (2004). Neural correlates of sensory and decision processes in auditory object identification. Nature Neuroscience, 7(3), 295–301. [PubMed: 14966525]

Bouchard KE, Mesgarani N, Johnson K, & Chang EF (2013). Functional organization of human sensorimotor cortex for speech articulation. Nature, 495(7441), 327–332. [PubMed: 23426266]

Breshears JD, Molinaro AM, & Chang EF (2015). A probabilistic map of the human ventral sensorimotor cortex using electrical stimulation. Journal of Neurosurgery, 123(2), 340–349. [PubMed: 25978714]

Brewer AA, Barton B, Venezia JH, Saberi K, & Hickok G (2013). Cross sensory activation of ‘clover leaf’ clusters in human auditory and visual cortex. Annual Meeting of the Cognitive Neuroscience Society, San Francisco, CA.

Brosch T, Grandjean D, Sander D, & Scherer KR (2009). Cross-modal emotional attention: emotional voices modulate early stages of visual processing. Journal of Cognitive Neuroscience, 21(9), 1670–1679. [PubMed: 18767920]

Brown S, Ngan E, & Liotti M (2008). A larynx area in the human motor cortex. Cerebral Cortex, 18(4), 837–845. [PubMed: 17652461]

Brown S, Yuan Y, & Belyk M (2020). Evolution of the speech-ready brain: The voice/jaw connection in the human motor cortex. Journal of Comparative Neurology.

Buchsbaum BR, Baldo J, Okada K, Berman KF, Dronkers N, D’Esposito M, & Hickok G (2011). Conduction aphasia, sensory-motor integration, and phonological short-term memory - An aggregate analysis of lesion and fMRI data. Brain and Language, 119(3), 119–128. doi:10.1016/j.bandl.2010.12.001 [PubMed: 21256582]

Buchsbaum BR, Hickok G, & Humphries C (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. Cognitive Science, 25(5), 663–678. doi:10.1207/s15516709cog2505_2

Calvert GA, Hansen PC, Iversen SD, & Brammer MJ (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. Neuroimage, 14(2), 427–438. doi:10.1006/nimg.2001.0812 [PubMed: 11467916]

Chang EF, Breshears JD, Raygor KP, Lau D, Molinaro AM, & Berger MS (2017). Stereotactic probability and variability of speech arrest and anomia sites during stimulation mapping of the language dominant hemisphere. Journal of Neurosurgery, 126(1), 114–121. [PubMed: 26894457]

Chang EF, Niziolek CA, Knight RT, Nagarajan SS, & Houde JF (2013). Human cortical sensorimotor network underlying feedback control of vocal pitch. Proceedings of the National Academy of Sciences, 110(7), 2653–2658.

Chen J, Penhune V, & Zatorre R (2009). The role of auditory and premotor cortex in sensorimotor transformations. Annals of the New York Academy of Sciences, 1169(1), 15–34. [PubMed: 19673752]

Cheung C, Hamilton LS, Johnson K, & Chang EF (2016). The auditory representation of speech sounds in human motor cortex. Elife, 5, e12577. [PubMed: 26943778]

Chumbley J, Worsley K, Flandin G, & Friston K (2010). Topological FDR for neuroimaging. Neuroimage, 49(4), 3057–3064. [PubMed: 19944173]

Cox RW (2012). AFNI: what a long strange trip it’s been. Neuroimage, 62(2), 743–747. [PubMed: 21889996]

D’Ausilio A, Craighero L, & Fadiga L (2012). The contribution of the frontal lobe to the perception of speech. Journal of Neurolinguistics, 25(5), 328–335.

Das P, Brodbeck C, Simon JZ, & Babadi B (2020). Neuro-current response functions: A unified approach to MEG source analysis under the continuous stimuli paradigm. Neuroimage, 211, 116528. [PubMed: 31945510]

David SV, Mesgarani N, & Shamma SA (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. Network: Computation in Neural Systems, 18(3), 191–212.

Dichter BK, Breshears JD, Leonard MK, & Chang EF (2018). The control of vocal pitch in human laryngeal motor cortex. Cell, 174(1), 21–31. e29. [PubMed: 29958109]

Du Y, Buchsbaum BR, Grady CL, & Alain C (2014). Noise differentially impacts phoneme representations in the auditory and speech motor systems. Proceedings of the National Academy of Sciences, 111(19), 7126–7131.

Du Y, Buchsbaum BR, Grady CL, & Alain C (2016). Increased activity in frontal motor cortex compensates impaired speech perception in older adults. Nature Communications, 7(1), 1–12.

Eichert N, Papp D, Mars RB, & Watkins KE (2020). Mapping human laryngeal motor cortex during vocalization. Cerebral Cortex, 30(12), 6254–6269. [PubMed: 32728706]

Erb J, & Obleser J (2013). Upregulation of cognitive control networks in older adults' speech comprehension. Frontiers in Systems Neuroscience, 7, 116. [PubMed: 24399939]

Ethofer T, Bretscher J, Gschwind M, Kreifelts B, Wildgruber D, & Vuilleumier P (2012). Emotional voice areas: anatomic location, functional properties, and structural connections revealed by combined fMRI/DTI. Cerebral Cortex, 22(1), 191–200. [PubMed: 21625012]

Evans S, & Davis MH (2015). Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. Cerebral Cortex, 25(12), 4772–4788. [PubMed: 26157026]

Fischl B (2012). FreeSurfer. Neuroimage, 62(2), 774–781. [PubMed: 22248573]

Frostig RD, Chen-Bee CH, Johnson BA, & Jacobs NS (2017). Imaging Cajal's neuronal avalanche: how wide-field optical imaging of the point-spread advanced the understanding of neocortical structure–function relationship. Neurophotonics, 4(3), 031217. [PubMed: 28630879]

Galgano J, & Froud K (2008). Evidence of the voice-related cortical potential: An electroencephalographic study. Neuroimage, 41(4), 1313–1323. [PubMed: 18495493]

Grabski K, Lamalle L, Vilain C, Schwartz JL, Vallée N, Tropres I, … Sato M (2012). Functional MRI assessment of orofacial articulators: neural correlates of lip, jaw, larynx, and tongue movements. Human Brain Mapping, 33(10), 2306–2321. [PubMed: 21826760]

Griffin DW, & Lim JS (1984). Signal estimation from modified short-time Fourier transform. Acoustics, Speech and Signal Processing, IEEE Transactions on, 32(2), 236–243.

Hackett TA (2011). Information flow in the auditory cortical network. Hearing Research, 271(1), 133–146. [PubMed: 20116421]

Hertz U, & Amedi A (2010). Disentangling unisensory and multisensory components in audiovisual integration using a novel multifrequency fMRI spectral analysis. Neuroimage, 52(2), 617–632. [PubMed: 20412861]

Hickok G (2010). The role of mirror neurons in speech perception and action word semantics. Language and Cognitive Processes, 25(6), 749–776.

Hickok G (2017). A cortical circuit for voluntary laryngeal control: Implications for the evolution language. Psychonomic Bulletin & Review, 24(1), 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(5), 673–682. doi:10.1162/jocn.2003.15.5.673 [PubMed: 12965041]

Hickok G, Okada K, Barr W, Pa J, Rogalsky C, Donnelly K, … Grant A (2008). Bilateral capacity for speech sound processing in auditory comprehension: evidence from Wada procedures. Brain and Language, 107(3), 179–184. [PubMed: 18976806]

Hickok G, & Poeppel D (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition, 92(1–2), 67–99. doi:10.1016/j.cognition.2003.10.011 [PubMed: 15037127]

Hickok G, & Poeppel D (2007). The cortical organization of speech processing. Nature Reviews Neuroscience, 8(5), 393–402. [PubMed: 17431404]

Holt LL, & Lotto AJ (2008). Speech perception within an auditory cognitive science framework. Current Directions in Psychological Science, 17(1), 42–46. [PubMed: 19060961]

Joassin F, Maurage P, & Campanella S (2011a). The neural network sustaining the crossmodal processing of human gender from faces and voices: An fMRI study. Neuroimage, 54(2), 1654–1661. [PubMed: 20832486]

Joassin F, Pesenti M, Maurage P, Verreckt E, Bruyer R, & Campanella S (2011b). Cross-modal interactions between human faces and voices involved in person recognition. Cortex, 47(3), 367–376. [PubMed: 20444445]

Keitel A, Gross J, & Kayser C (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. PLoS Biology, 16(3), e2004473. [PubMed: 29529019]

Klein DJ, Depireux DA, Simon JZ, & Shamma SA (2000). Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. Journal of Computational Neuroscience, 9(1), 85–111. [PubMed: 10946994]

Koelsch S, & Skouras S (2014). Functional centrality of amygdala, striatum and hypothalamus in a "small-world" network underlying joy: An fMRI study with music. Human Brain Mapping, 35(7), 3485–3498. [PubMed: 25050430]

Koelsch S, Skouras S, & Lohmann G (2018). The auditory cortex hosts network nodes influential for emotion processing: An fMRI study on music-evoked fear and joy. PloS One, 13(1), e0190057. [PubMed: 29385142]

Krieger-Redwood K, Gaskell MG, Lindsay S, & Jefferies E (2013). The selective role of premotor cortex in speech perception: a contribution to phoneme judgements but not speech comprehension. Journal of Cognitive Neuroscience, 25(12), 2179–2188. [PubMed: 23937689]

Leblanc É, Dégeilh F, Daneault V, Beauchamp MH, & Bernier A (2017). Attachment security in infancy: A preliminary study of prospective links to brain morphometry in late childhood. Frontiers in Psychology, 8, 2141. [PubMed: 29312029]

Matchin W, Hammerly C, & Lau E (2017). The role of the IFG and pSTS in syntactic prediction: Evidence from a parametric study of hierarchical structure in fMRI. Cortex, 88, 106–123. [PubMed: 28088041]

Matchin W, & Hickok G (2020). The cortical organization of syntax. Cerebral Cortex, 30(3), 1481–1498. [PubMed: 31670779]

Matiasek MR, & Richards VM (1999). Relative weights for three different psychophysical tasks. The Journal of the Acoustical Society of America, 106(4), 2209–2210.

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

Morey RD, & Rouder JN (2011). Bayes factor approaches for testing interval null hypotheses. Psychological Methods, 16(4), 406. [PubMed: 21787084]

Moulin-Frier C, Diard J, Schwartz J-L, & Bessière P (2015). COSMO ("Communicating about Objects using Sensory–Motor Operations"): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems. Journal of Phonetics, 53, 5–41.

Mumford JA, Turner BO, Ashby FG, & Poldrack RA (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. Neuroimage, 59(3), 2636–2643. [PubMed: 21924359]

Narain C, Scott SK, Wise RJ, Rosen S, Leff A, Iversen S, & Matthews P (2003). Defining a left-lateralized response specific to intelligible speech using fMRI. Cerebral Cortex, 13(12), 1362–1368. [PubMed: 14615301]

Naselaris T, Kay KN, Nishimoto S, & Gallant JL (2011). Encoding and decoding in fMRI. Neuroimage, 56(2), 400–410. [PubMed: 20691790]

Okada K, Rong F, Venezia J, Matchin W, Hsieh IH, Saberi K, … Hickok G (2010). Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cerebral Cortex, 20(10), 2486–2495. doi:10.1093/cercor/bhp318 [PubMed: 20100898]

Olthoff A, Baudewig J, Kruse E, & Dechent P (2008). Cortical sensorimotor control in vocalization: a functional magnetic resonance imaging study. The Laryngoscope, 118(11), 2091–2096. [PubMed: 18758379]

Oosterhof NN, Connolly AC, & Haxby JV (2016). CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. Frontiers in Neuroinformatics, 10, 27. [PubMed: 27499741]

Oosterhof NN, Wiestler T, Downing PE, & Diedrichsen J (2011). A comparison of volume-based and surface-based multi-voxel pattern analysis. Neuroimage, 56(2), 593–600. [PubMed: 20621701]

Peelle JE (2012). The hemispheric lateralization of speech processing depends on what "speech" is: a hierarchical perspective. Frontiers in Human Neuroscience, 6, 309. [PubMed: 23162455]

Peelle JE, Johnsrude I, & Davis MH (2010). Hierarchical processing for speech in human auditory cortex and beyond. Frontiers in Human Neuroscience, 4, 51. [PubMed: 20661456]

Peelle JE, Johnsrude IS, & Davis MH (2010). Hierarchical processing for speech in human auditory cortex and beyond. Frontiers in Human Neuroscience, 4.

Pessoa L, McKenna M, Gutierrez E, & Ungerleider LG (2002). Neural processing of emotional faces requires attention. Proceedings of the National Academy of Sciences, 99(17), 11458–11463.

Poeppel D, & Assaneo MF (2020). Speech rhythms and their neural foundations. Nature Reviews Neuroscience, 1–13. [PubMed: 31796912]

Pulvermuller F, & Fadiga L (2010). Active perception: sensorimotor circuits as a cortical basis for language. Nature Reviews: Neuroscience, 11(5), 351–360. doi:10.1038/nrn2811 [PubMed: 20383203]

Rauschecker JP, & Scott SK (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Neuroscience, 12(6), 718–724. [PubMed: 19471271]

Rogalsky C, Basilakos A, Rorden C, Pillay S, LaCroix AN, Keator L, … Fridriksson J (2020). The Neuroanatomy of Speech Processing: A Large-Scale Lesion Study. bioRxiv.

Rogalsky C, Love T, Driscoll D, Anderson SW, & Hickok G (2011). Are mirror neurons the basis of speech perception? Evidence from five cases with damage to the purported human mirror system. Neurocase, 17(2), 178–187. [PubMed: 21207313]

Sakata H, Taira M, Murata A, & Mine S (1995). Neural mechanisms of visual guidance of hand action in the parietal cortex of the monkey. Cerebral Cortex, 5(5), 429–438. [PubMed: 8547789]

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(1), e1003412. [PubMed: 24391486]

Scott SK, McGettigan C, & Eisner F (2009). A little more conversation, a little less action—candidate roles for the motor cortex in speech perception. Nature Reviews Neuroscience, 10(4), 295–302. [PubMed: 19277052]

Sievers B, Lee C, Haslett W, & Wheatley T (2019). A multi-sensory code for emotional arousal. Proceedings of the Royal Society B, 286(1906), 20190513. [PubMed: 31288695]

Sievers B, Parkinson C, Kohler PJ, Hughes J, Fogelson SV, & Wheatley T (2018). Visual and auditory brain areas share a neural code for perceived emotion. BioRxiv, 254961.

Simonyan K (2014). The laryngeal motor cortex: its organization and connectivity. Current Opinion in Neurobiology, 28, 15–21. [PubMed: 24929930]

Skipper JI, Devlin JT, & Lametti DR (2017). The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. Brain and Language, 164, 77–105. doi:10.1016/j.bandl.2016.10.004 [PubMed: 27821280]

Skipper JI, Nusbaum HC, & Small SL (2005). Listening to talking faces: motor cortical activation during speech perception. Neuroimage, 25(1), 76–89. doi:10.1016/j.neuroimage.2004.11.006 [PubMed: 15734345]

Skouras S, Gray M, Critchley H, & Koelsch S (2013). fMRI scanner noise interaction with affective neural processes. PloS One, 8(11), e80564. [PubMed: 24260420]

Smiley JF, Hackett TA, Ulbert I, Karmas G, Lakatos P, Javitt DC, & Schroeder CE (2007). Multisensory convergence in auditory cortex, I. Cortical connections of the caudal superior temporal plane in macaque monkeys. Journal of Comparative Neurology, 502(6), 894–923.

Smith SM, & Nichols TE (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage, 44(1), 83–98. [PubMed: 18501637]

Stelzer J, Chen Y, & Turner R (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. Neuroimage, 65, 69–82. [PubMed: 23041526]

Stokes RC, Venezia JH, & Hickok G (2019). The motor system's [modest] contribution to speech perception. Psychonomic Bulletin & Review, 26(4), 1354–1366. doi:10.3758/s13423-019-01580-2 [PubMed: 30945170]

Szenkovits G, Peelle JE, Norris D, & Davis MH (2012). Individual differences in premotor and motor recruitment during speech perception. Neuropsychologia, 50(7), 1380–1392. [PubMed: 22521874]

Team, R. C. (Producer). (2019). R: A Language and Environment for Statistical Computing. R. Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Terumitsu M, Fujii Y, Suzuki K, Kwee IL, & Nakada T (2006). Human primary motor cortex shows hemispheric specialization for speech. Neuroreport, 17(11), 1091–1095. [PubMed: 16837833]

Theunissen FE, & Elie JE (2014). Neural processing of natural sounds. Nature Reviews Neuroscience, 15(6), 355–366. [PubMed: 24840800]

Theunissen FE, Sen K, & Doupe AJ (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. Journal of Neuroscience, 20(6), 2315–2331. [PubMed: 10704507]

Tremblay P, & Small SL (2011). On the context-dependent nature of the contribution of the ventral premotor cortex to speech perception. Neuroimage, 57(4), 1561–1571. [PubMed: 21664275]

Venezia JH, & Hickok G (2009). Mirror Neurons, the Motor System and Language: From the Motor Theory to Embodied Cognition and Beyond. Language and Linguistics Compass, 3(6), 1403–1416. doi:10.1111/j.1749-818X.2009.00169.x

Venezia JH, Hickok G, & Richards VM (2016). Auditory “bubbles”: Efficient classification of the spectrotemporal modulations essential for speech intelligibility. Journal of the Acoustical Society of America, 140(2), 1072. doi:10.1121/1.4960544

Venezia JH, Leek MR, & Lindeman MP (2020). Suprathreshold Differences in Competing Speech Perception in Older Listeners With Normal and Impaired Hearing. Journal of Speech, Language, and Hearing Research, 63(7), 2141–2161.

Venezia JH, Martin A-G, Hickok G, & Richards VM (2019a). Identification of the Spectrotemporal Modulations That Support Speech Intelligibility in Hearing-Impaired and Normal-Hearing Listeners. Journal of Speech, Language, and Hearing Research, 62(4), 1051–1067.

Venezia JH, Thurman SM, Richards VM, & Hickok G (2019b). Hierarchy of speech-driven spectrotemporal receptive fields in human auditory cortex. Neuroimage, 186, 647–666. doi:10.1016/j.neuroimage.2018.11.049 [PubMed: 30500424]

Vuilleumier P, Richardson MP, Armony JL, Driver J, & Dolan RJ (2004). Distant influences of amygdala lesion on visual cortical activation during emotional face processing. Nature Neuroscience, 7(11), 1271–1278. [PubMed: 15494727]

Wang L, Peng B, & Li R (2015). A High-Dimensional Nonparametric Multivariate Test for Mean Vector. J Am Stat Assoc, 110(512), 1658–1669. doi:10.1080/01621459.2014.988215 [PubMed: 26848205]

Watkins S, Shams L, Tanaka S, Haynes J-D, & Rees G (2006). Sound alters activity in human V1 in association with illusory visual perception. Neuroimage, 31(3), 1247–1256. [PubMed: 16556505]

Wild CJ, Yusuf A, Wilson DE, Peelle JE, Davis MH, & Johnsrude IS (2012). Effortful listening: the processing of degraded speech depends critically on attention. Journal of Neuroscience, 32(40), 14010–14021. [PubMed: 23035108]

Wilson SM, Saygin AP, Sereno MI, & Iacoboni M (2004). Listening to speech activates motor areas involved in speech production. Nature Neuroscience, 7(7), 701–702. doi:10.1038/nn1263 [PubMed: 15184903]

Wu Y-H, Stangl E, Zhang X, Perkins J, & Eilers E (2016). Psychometric functions of dual-task paradigms for measuring listening effort. Ear and Hearing, 37(6), 660. [PubMed: 27438866]

Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, & Wager TD (2011). Large-scale automated synthesis of human functional neuroimaging data. Nature methods, 8(8), 665–670. [PubMed: 21706013]

Zhang X, Japee S, Safiullah Z, Mlynaryk N, & Ungerleider LG (2016). A normalization framework for emotional attention. PLoS Biology, 14(11), e1002578. [PubMed: 27870851]

Highlights

- Speech-driven spectrotemporal receptive fields (STRFs) are estimated using fMRI with spectrotemporal modulation filtering of continuous speech (“Auditory Bubbles”)
- A multivariate analysis of cross-subject STRF alignment is developed to maintain sensitivity to small clusters of STRFs
- Significant STRF alignment is observed outside traditional auditory cortex
- Left dorsal speech-premotor cortex (dPM) and bilateral calcarine sulcus (calcS) respond to acoustic speech features associated with speech intelligibility and vocal pitch
- Left inferior frontal gyrus (IFG) responds only to features associated with intelligibility
- dPM and calcS are maximally functionally connected with early auditory cortex; IFG is maximally connected with superior temporal gyrus/sulcus and middle temporal gyrus
- STRFs in dPM predict activation on trials for which speech is rated as unintelligible by listeners, a hallmark auditory profile
- We posit that dPM is part of the laryngeal motor cortex capable of processing speech in an ‘auditory mode’

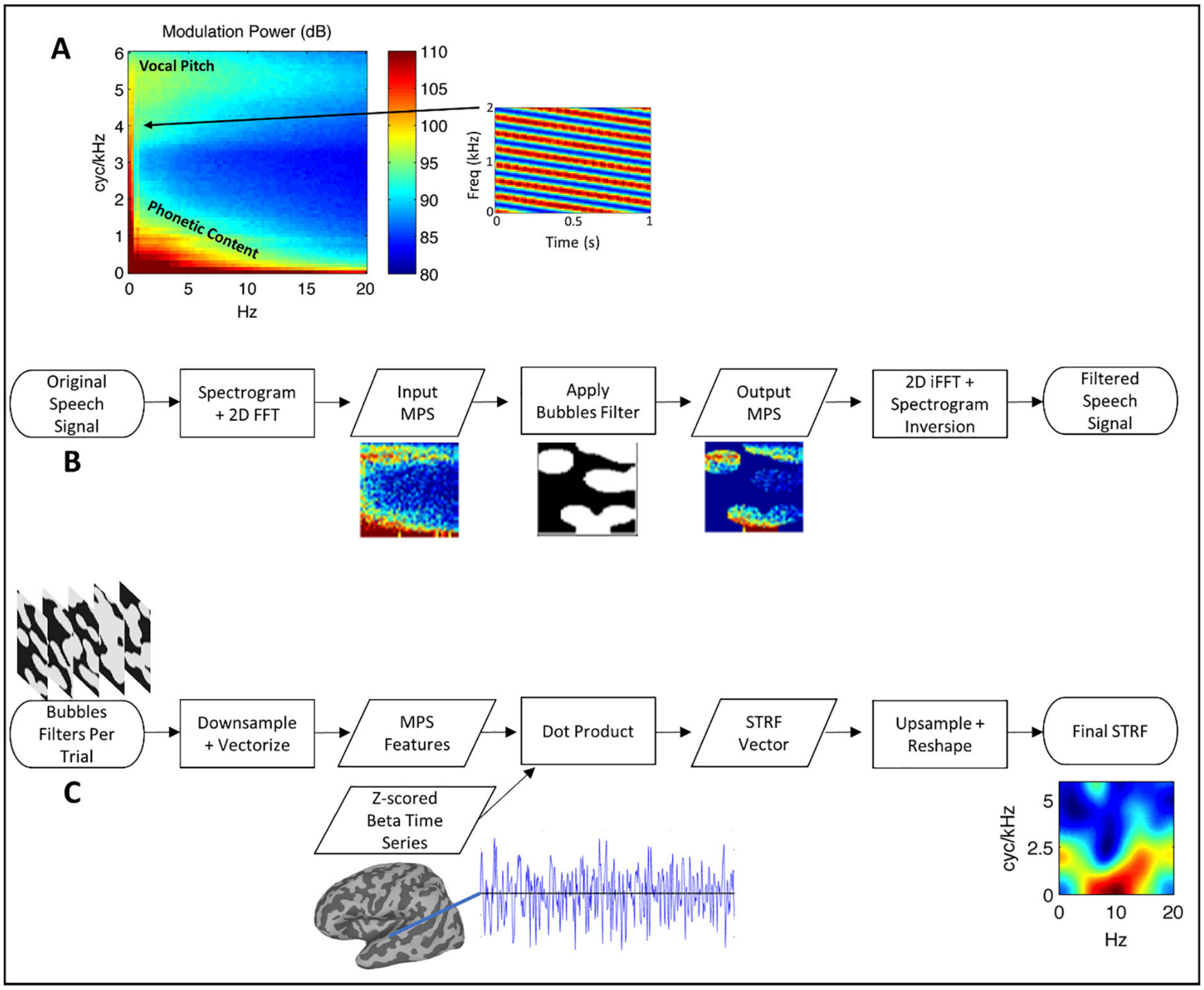

Figure 1.

(A) Average MPS of 452 IEEE sentences spoken by a female talker. The MPS describes the speech spectrogram as a weighted sum of spectrotemporal ripples containing energy at a unique combination of temporal (abscissa, Hz) and spectral (ordinate, cyc/kHz) modulation rate. An example of a downward-sweeping ripple (2Hz, 4 cyc/kHz) is shown at right. In this figure and throughout, energy at a given pixel location in the MPS reflects the average of downward- and upward-sweeping ripples at that location. Modulation energy clusters into two discrete regions: a high-spectral-modulation-rate region corresponding to vocal pitch (F0 Harmonics) and a low-spectral-modulation-rate region corresponding the resonant frequencies of the vocal tract (formants/phonetic content). Color scale = dB (arb. ref). (B) Flow-chart schematic of the signal processing steps involved in Bubbles filtering. (C) Flow-chart schematic of the Bubbles fMRI analysis (see also Eq. 1).

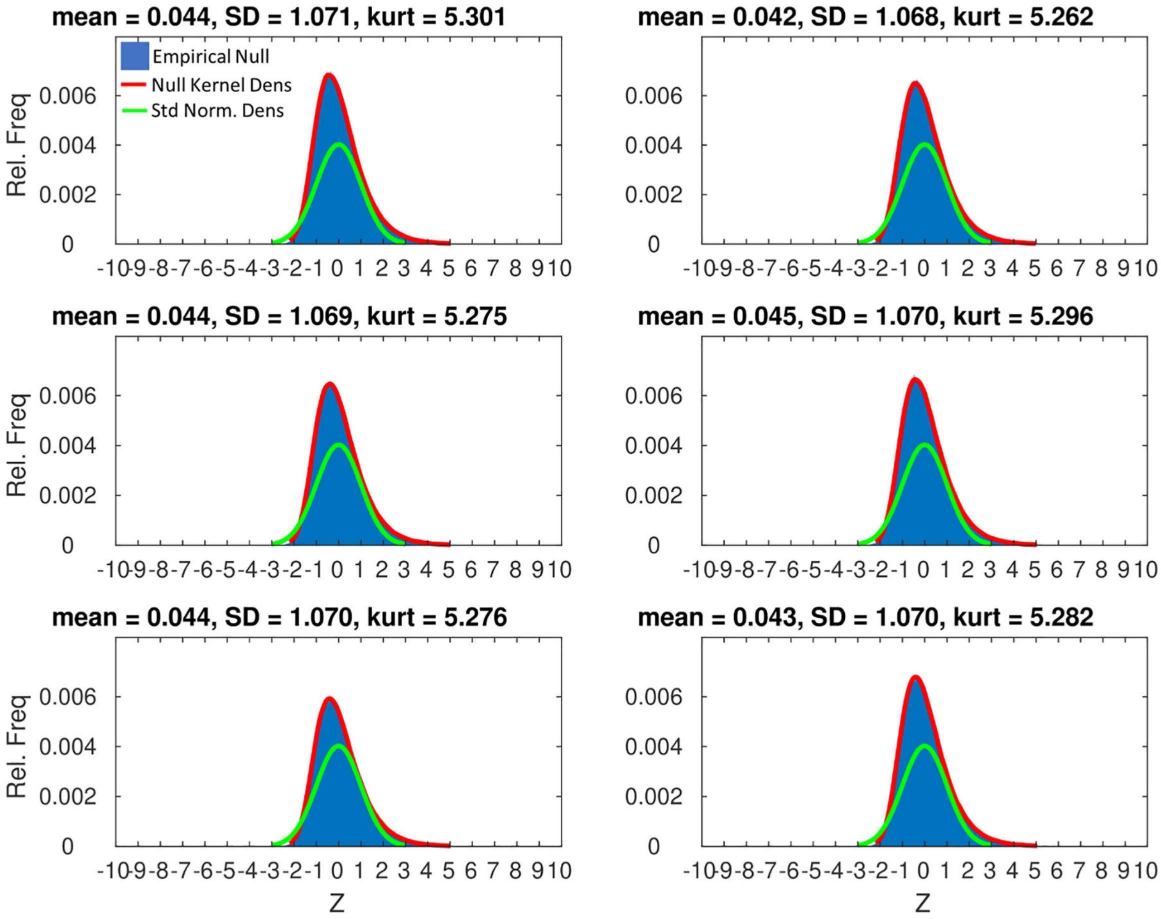

Figure 2.

For six samples of 100 cortical surface nodes selected at random from across the whole brain (panels), the empirical null distribution of Z (10,000 draws after random permutation of the trial order of bubbles feature vectors) is plotted as a relative frequency histogram (blue) overlayed with a kernel density estimate of the continuous probability density function of Z (red) and the standard normal probability density function (green). Above each panel is shown the mean, standard deviation (SD), and kurtosis (kurt) of the null distribution of Z for the corresponding sample. The empirical null is right-skewed and more kurtotic compared to the standard normal, which has a kurtosis of 3.

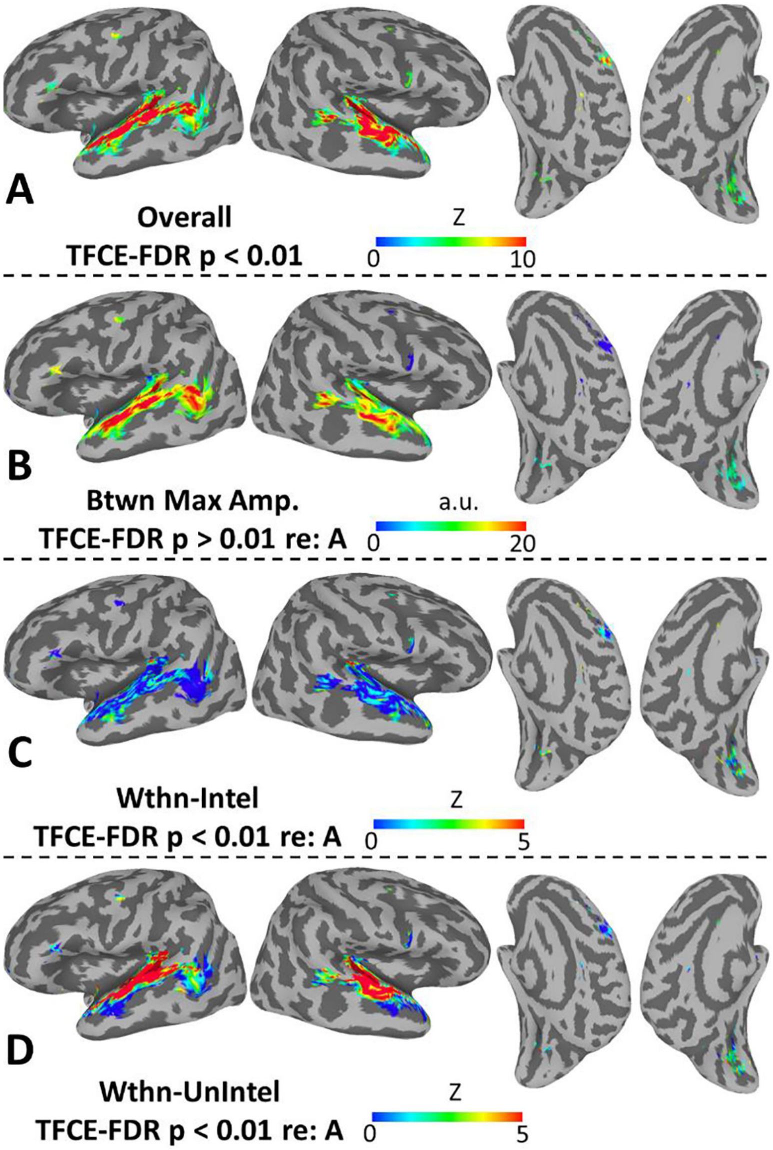

Figure 3.

(A) Maps show the strength of cross-subject STRF alignment at the second level as quantified with multivariate test statistic, Z (see Materials and Methods; only significant values shown, TFCE-FDR-corrected p < 0.01). (B) Maps show the second level mean of the maximum amplitude of $\mathrm{STRF}_{\mathrm{between}}$, restricted to surface nodes identified as significant in A. (C) Maps show the strength of cross-subject STRF alignment within trials rated as intelligible by the listeners, restricted to surface nodes identified as significant in A. (D) Maps show the strength of cross-subject STRF alignment within trials rated as unintelligible by the listeners, restricted to surface nodes identified as significant in A. All plots use a standard topology inflated surface derived from the Colin N27 template in MNI space.

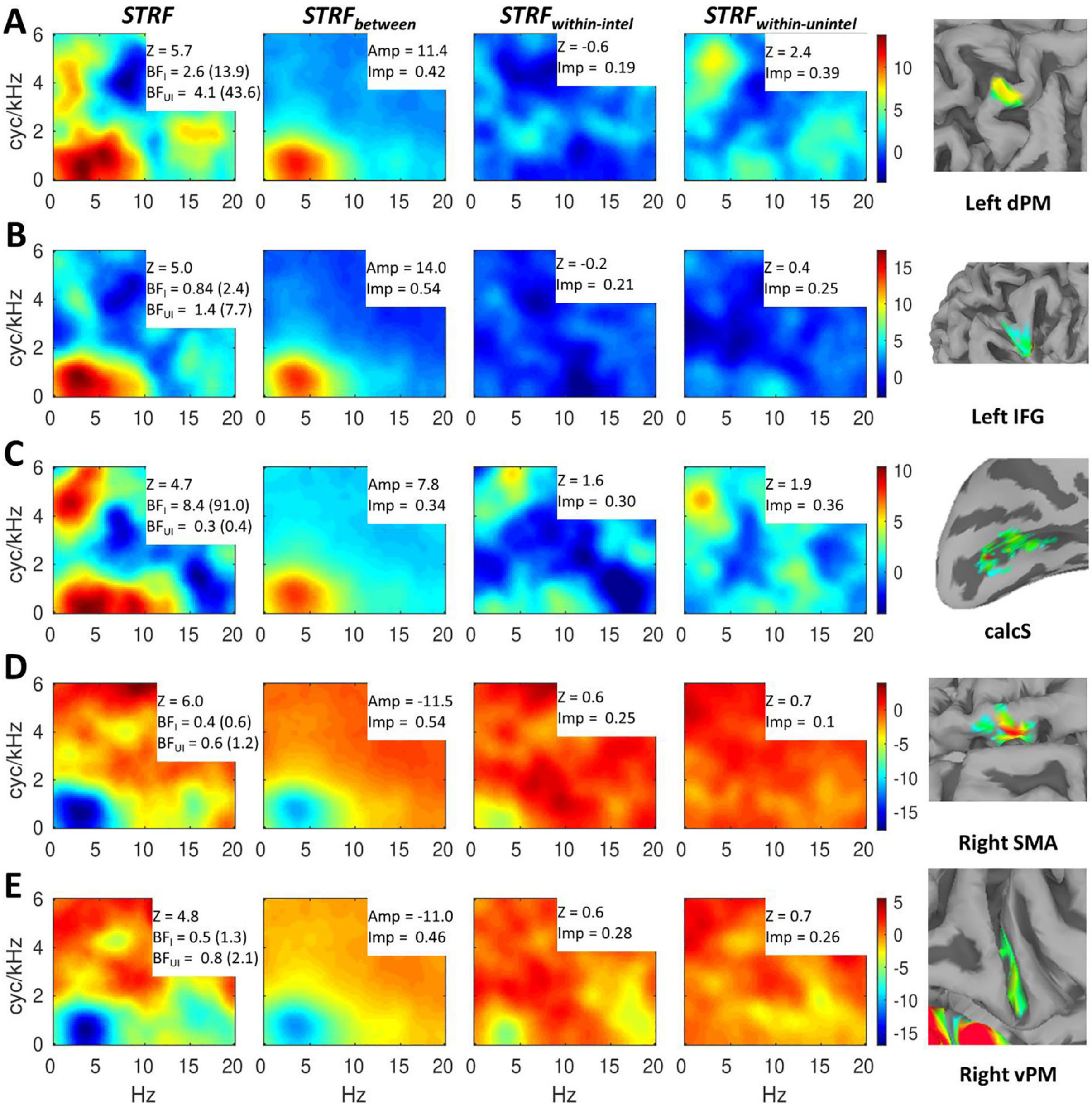

Figure 4.

Speech-driven STRFs in non-auditory regions. A row (A-E) is dedicated to each region (right insets: zoomed versions of Figure 3A on semi-inflated or inflated standard topology surfaces). Overall responses (STRF) are shown in the first column, while effects of intelligibility ($\text{STRF}_{\text{between}}$) and responses within intelligible ($\text{STRF}_{\text{intel}}$) and unintelligible ($\text{STRF}_{\text{unintel}}$) trials are shown in the second through fourth columns, respectively. For each region (A-E), the color scale is based on the overall STRF min/max and applied to each of the decomposed STRFs at right to allow a visualization of the relative contribution of each component STRF to the overall STRF. The overall STRFs are annotated with second-level alignment strength (Z) and Bayes Factors (geometric mean across surface nodes, 80th percentile given parenthetically) indicating quality of LOO predictions within intelligible ($\text{BF}_{\text{I}} = \text{BF}_{\text{intel}}$) and unintelligible ($\text{BF}_{\text{UI}} = \text{BF}_{\text{unintel}}$) trials. The decomposed STRFs are annotated with the max amplitude (Amp; $STRF_{between}$) or second-level alignment strength (Z; $STRF_{intel}$ and $STRF_{unintel}$) and the proportional relative importance (Imp).

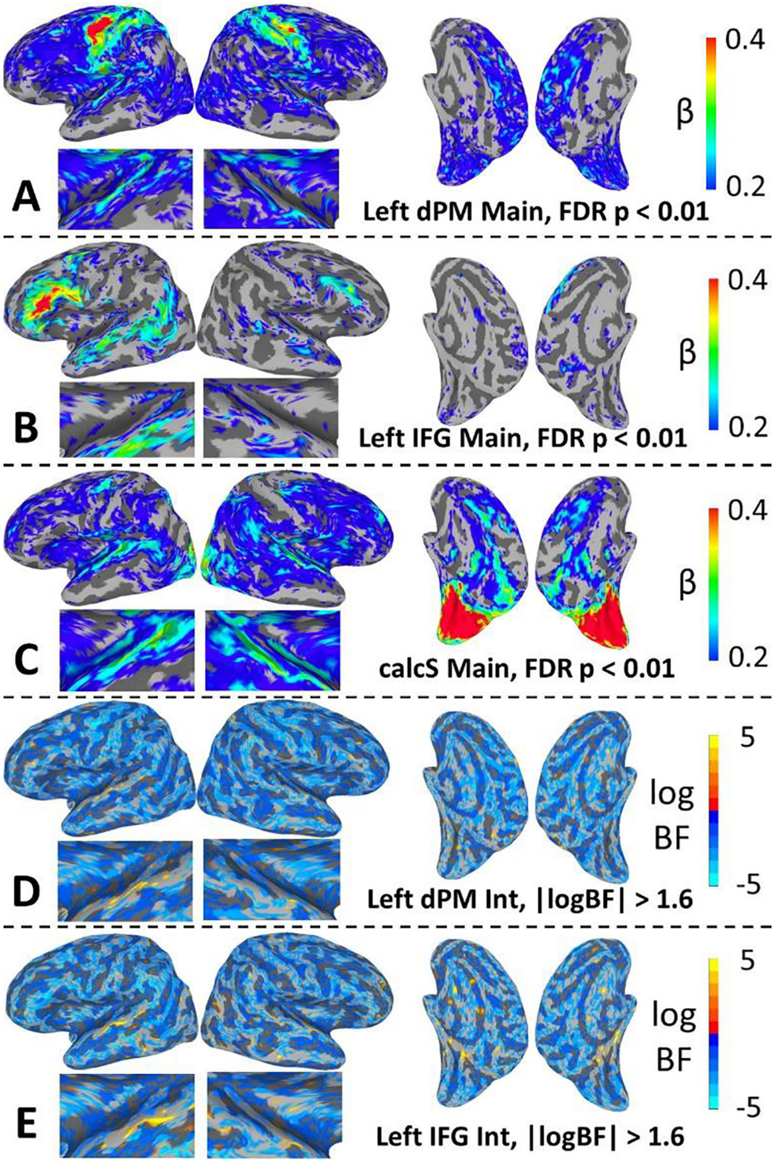
0.4
β
0.2
A
Left dPM Main, FDR p < 0.01
0.4
β
0.2
B
Left IFG Main, FDR p < 0.01
0.4
β
0.2
C
calcS Main, FDR p < 0.01
5
log
BF
-5
D
Left dPM Int, |logBF| > 1.6
5
log
BF
-5
E
Left IFG Int, |logBF| > 1.6

Figure 5.

(A-C) Maps show the average regression coefficient (β) for the main effect of seed connectivity (two-way, FDR-corrected p < 0.01). Note the color scale: nearly all (>99%) significant effects were positive; the color scale was selected to allow visualization of the full range of these positive effects across brain regions. That is, both cool and warm colors denote positive functional connectivity. (D-E) Maps show the log Bayes Factor for the intelligibility-by-seed-connectivity interaction (thresholded at |logBF| > 1.6). Negative values (cool colors) indicate support for the null hypothesis (no interaction) and positive values indicate support for the alternative hypothesis (interaction present). Plots zoomed on the auditory cortex are shown beneath the lateral surface plots. All plots use a standard topology inflated surface derived from the Colin N27 template in MNI space.

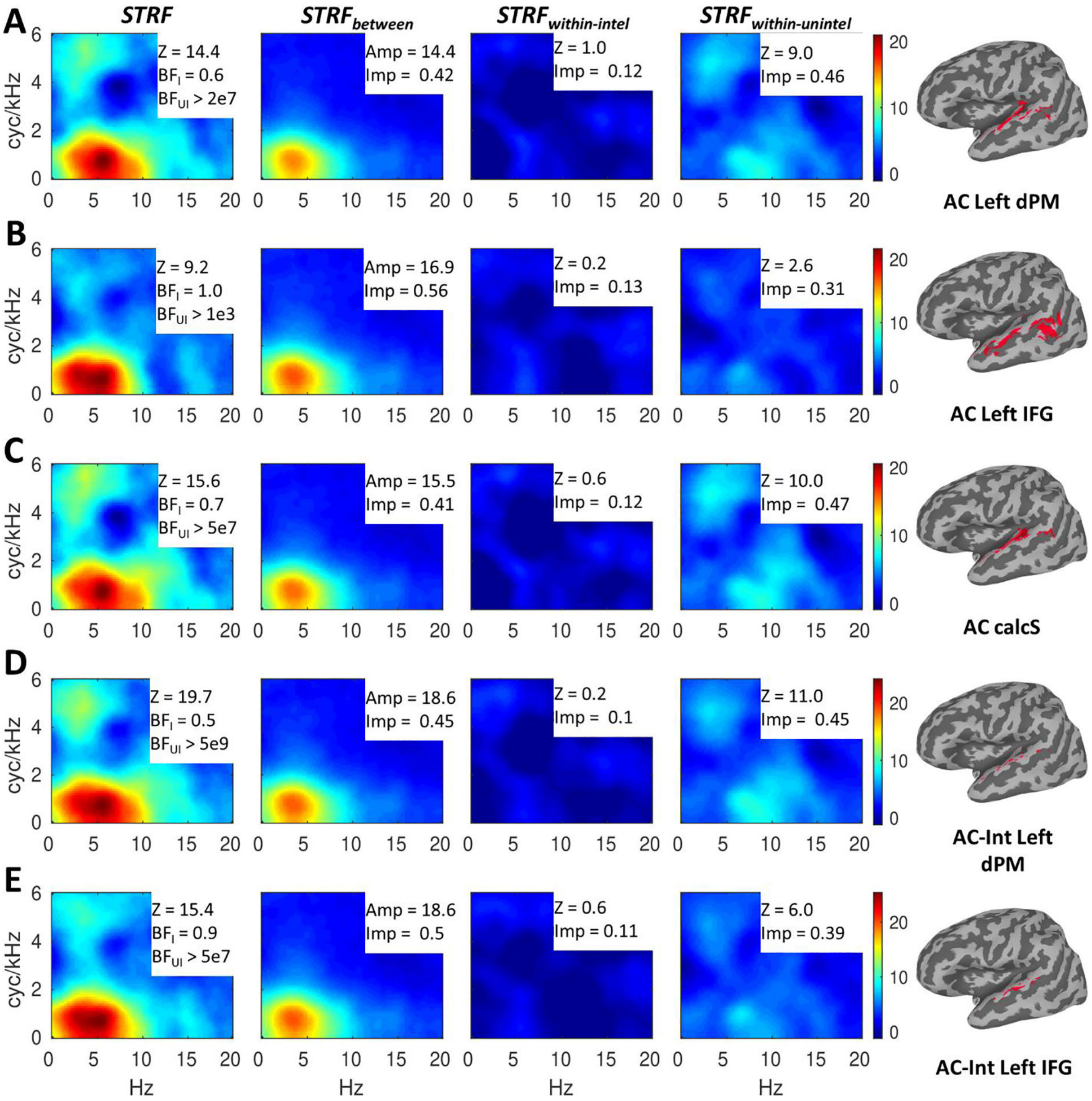

Figure 6.

Speech-driven STRFs in auditory regions defined by functional connectivity with non-auditory regions. A row (A-E) is dedicated to each auditory region (right insets: region mask displayed on inflated standard topology surfaces and labeled with the corresponding seed region). Regions defined by the main effect of seed connectivity (A-C) are labeled as ‘AC’ and regions defined by the interaction of seed connectivity with speech intelligibility (D-E) are labeled as ‘AC-Int.’ Overall responses (STRF) are shown in the first column, while effects of intelligibility ($\mathrm{STRF}_{\mathrm{between}}$) and responses within intelligible ($\mathrm{STRF}_{\mathrm{intel}}$) and unintelligible ($\mathrm{STRF}_{\mathrm{unintel}}$) trials are shown in the second through fourth columns, respectively. For each region (A-E), the color scale is based on the overall STRF min/max and applied to each of the decomposed STRFs at right to allow a visualization of the relative contribution of each component STRF to the overall STRF. The overall STRFs are annotated with second-level alignment strength (Z) and Bayes Factors (geometric mean across surface nodes) indicating quality of LOO predictions within intelligible ($BF_I = BF_{intel}$) and unintelligible ($BF_{UI} = BF_{unintel}$) trials. The decomposed STRFs are annotated with the max amplitude (Amp; $STRF_{between}$) or second-level alignment strength (Z; $STRF_{intel}$ and $STRF_{unintel}$) and the proportional relative importance (Imp).

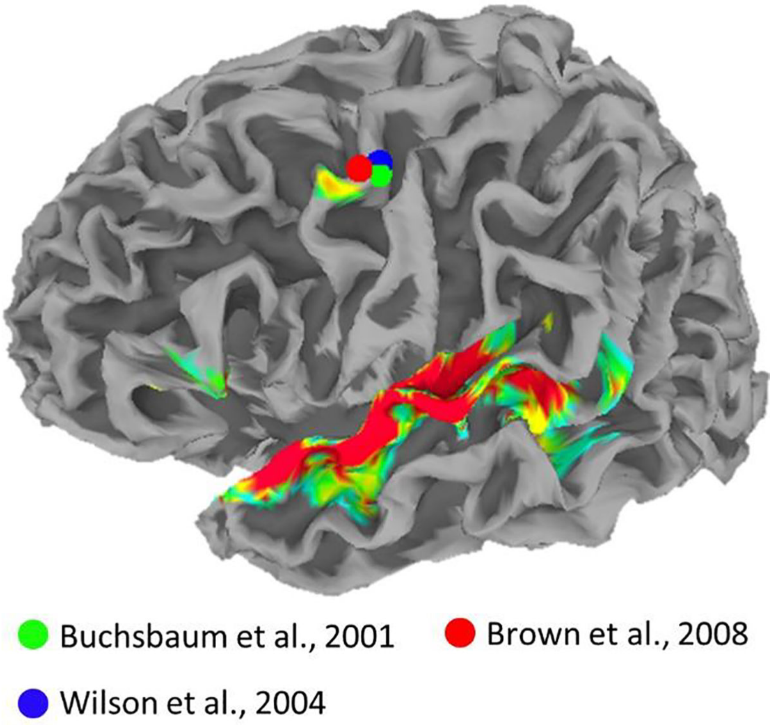

Figure 7.

Heatmap is a replication of Figure 4A. Spheres show peak dorsal premotor coordinates from previous studies showing activation during sensory and motor speech tasks (Buchsbaum et al., 2001, green; Wilson et al., 2004, blue) and during laryngeal motor tasks (Brown et al, 2008, red). Plots are overlaid on the standard topology white-matter-boundary surface derived from the Colin27 template in MNI space.