Title
From Sound to Meaning: Representations of Speech in Human Cortex

Author
de Heer, Wendy Aileen

Publication Date
2015

Peer reviewed|Thesis/dissertation

From Sound to Meaning: Representations of Speech in Human Cortex

by

Wendy Aileen de Heer

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Frédéric E. Theunissen, Chair
Professor Jack L. Gallant
Professor Thomas L. Griffiths
Professor Keith A. Johnson

Summer 2015

From Sound to Meaning: Representations of Speech in Human Cortex

Abstract

From Sound to Meaning: Representations of Speech in Human Cortex

by

Wendy Aileen de Heer

Doctor of Philosophy in Psychology

University of California, Berkeley

Professor Frédéric E. Theunissen, Chair

This dissertation investigates the cortical representation of speech perception, using a combination of functional magnetic resonance imaging (fMRI) and psychoacoustical experiments.

Previous research has shown that low–level acoustical structure, phonemes, and words are processed by distinct cortical areas. However, little is known about the relationship between these different representations. To address this problem we simultaneously mapped many different representations of speech. We recorded fMRI responses from subjects listening to over two hours of natural speech. We then examined three features spaces representing the speech sounds in terms of auditory, articulatory and semantic features. We used voxel-wise modeling for each feature space combined with a novel variance-partitioning method to assess how much response variance could be explained uniquely by each model or jointly between two or three models. Validating our approach, we found that a quarter of the brain was significantly responsive to the stories, and that our models could account for up to 45% of the explainable variance in cortex and over 60% of the explainable variance in auditory areas. We also found a hierarchical set of processing steps starting in primary auditory areas and moving along the posteroventral region of the temporal lobe that are involved in the sound to word meaning transformation.

The second part of this dissertation is a psychoacoustical investigation of the modulation power spectrum (MPS) of speech. The MPS is obtained by taking the 2-dimensional Fourier transform of the speech spectrogram. We showed that comprehension of vowels and consonants is differently affected by removal of specific spectral or temporal modulations. Supplementary consonant analysis showed differences in MPS and psychoacoustical comprehension results between three groups of consonants, separated based on the manner in which they are pronounced (fricatives, stops, and sonorants). The MPS could serve as an excellent intermediate step between lower and higher levels of speech processing, and could in future studies add nuance to our previous three cortical models of speech perception.

To my parents, Patricia and Walt de Heer

Contents

Acknowledgments

Thank you to my advisor, Frédéric Theunissen, for his mathematical depth, love of science, and support throughout my years at Berkeley. The Theunissen Lab has also been a joy to work with, in particular Tyler Lee, Mike Schachter, Solveig Mouterde, Julie Elie, Channing Moore, Taffeta Elliott, Noopur Amin and Yuka Minton. While at Berkeley, I worked with talented research assistants: thank you to Alfonso Bonilla, Michelle Carney, Josh Downer, Miren Edelstein, Brittany Griffin, Shemsa Morkoch, and Jasmine Nguyen.

I could not have completed this dissertation without Jack Gallant and his lab. They helped me learn almost everything I know about fMRI, and their passion for science and excellence is inspiring. Thank you especially to Alex Huth, who was a wonderful collaborator, Natalia Bilenko, Fatma Imamoglu, James Gao, Mark Lescroart and Anwar Nuñez-Elizalde.

Thank you to my qualifying exam and dissertation committee for their insight and guidance: Rich Ivry, Tom Griffiths, Jack Gallant, and Keith Johnson.

John Schindel has provided invaluable help with navigating Berkeley's bureaucratic waters–thank you for that, and for always giving me support and encouragement at a moment's notice. Thank you also to my extended psychology and neuroscience community at Berkeley. There are far too many people to mention, but thank you in particular to Kati Markowitz for having helped to create such an amazing neuroscience community; Bob Knight, Erv Hafter, Chris Holdgraf, Yvonne Fonken, Emily Cooper, Rachel Denison and all of Time Club. Thank you to my favorite YAKers Liberty Hamilton, Asako Miyakawa, and finally Natalia Bilenko, who has helped me immeasurably both as a scientist and as a friend.

I would like to thank my undergraduate advisors from Cornell who encouraged me in my first attempts at research: Jane Marie Law, Ron Hoy, and Don Fredericksen. Thank you also to my Master's advisor from Paris, Daniel Pressnitzer, as well as my unofficial co-mentor Christian Lorenzi.

The support of friends and family was crucial during my years at Berkeley. For being such wonderful and supportive friends during these years, thank you especially to Melody Chan and Amy Katzen, Mara Joustra, Agnès Léger, Dmitri Seals, Alisa Sanchez and Rosco Sancheals, Vanessa Ferdinand, Pernille Bruhn, Ken Lin. Thank you to the House of Love, and in particular Jen Chunn, Tamera Stover, John Bicket, Gillian Boudreau, Roger Brunson, Renee Finkelstein. Thank you to Deena Varshavskaya and David Bill, Kevin Cheng and the Black Rock Academy/Friendlandians, Dr. Brainlove and all of her crew.

Thank you to Dennis Mueller for creating a wonderful home with me, that allows me to wholeheartedly work and relax, and for being an exceptional partner.

Finally, I would like to thank my parents, Patricia and Walt de Heer. They have unwaveringly supported me through all of my adventures, including this one.

Chapter 1

Introduction

While listening to speech, the human brain extracts meaning from a continuous auditory signal. This process is often described as a serial computational problem: first a sequence of phonemes is extracted from the auditory stream, then a sequence of words is extracted from the sequence of phonemes, and finally the meaning of the word sequence is inferred based on the known syntactic and semantic roles of the words (Just, Cherkassky, Aryal, & Mitchell, 2010; Rodd, Davis, & Johnsrude 2005; Fedorenko, Nieto-Castañón, & Kanwisher 2012.) This sequence of bottom up computations needed for speech comprehension supports at first glance a modular, hierarchical view of brain processing centers for language (Price, 2010; Stowe, Haverkort, & Zwarts, 2005). In this modular view, lower-level processing modules involved, for example, in the computations needed for phoneme identification would be separate from mid-level processing modules involved in lexical retrieval, and those in turn would be separate from higher-level modules involved in semantics. In support of this modular view, previous functional magnetic resonance imaging (fMRI) studies have localized cortical areas that selectively respond to the auditory spectrum (Formisano, De Martino, Bonte, & Goebel, 2008; Pasley et al., 2012) and individual phonemes (Bonte, Hausfeld, Scharke, Valente, & Formisano, 2014). Several other studies have localized cortical areas that respond to specific syntactic properties (Snijders et al., 2009) and semantic properties (Visser, Jefferies, & Ralph, 2010) of linguistic stimuli. Together these studies suggest that a network of hierarchically organized cortical areas, including superior temporal cortex, medial parietal cortex, and inferior frontal/prefrontal cortex are involved in speech perception.

At the level of auditory and auditory-association cortex, more precise models have been suggested for the initial computations that are needed for auditory object recognition of both speech and non-speech sounds. Most of these models agree that auditory processing progresses from the lower levels of the primary auditory cortex (along Heschl's gyrus) to more ventral regions on the superior temporal sulcus (STS) and from there bifurcates into two (or multiple) processing streams (e.g. reviewed in Turkeltaub & Coslett, 2010 and Bornkessel-Schlesewsky, Schlesewsky,

Small, & Rauschecker, 2015): an anteroventral stream that is implicated in auditory object recognition including word specific areas (DeWitt & Rauschecker, 2012) and a posterodorsal stream that could be involved in more dynamical aspects of speech processing such as processing temporal sequences (Belin & Zatorre, 2000) as well as in the sensori-motor transformation that represents speech in terms of articulation features (Rauschecker & Scott, 2009; Scott & Johnsrude, 2003). The exact role and precise anatomical pathways of these two (or multiple) streams remain an active area of research because of the many computational steps that are needed for speech comprehension and the fact that current data is consistent with multiple interpretations (Bornkessel-Schlesewsky et al., 2015).

Indeed one of the challenges of such research has been to design stimuli and experiments that isolate processing stages within these streams that would correspond to a specific computational step: the anatomical dissociation problem. In classical experimental designs, responses to different stimuli with similar acoustical complexity such as words and synthetic sounds (e.g. Binder et al., 2000) or words and other animal vocalizations (e.g. Leaver & Rauschecker, 2010) are obtained in block design experiments. Subtraction of responses is then used to determine brain regions that uniquely respond to particular sound classes and in this manner determine the particular features unique to speech processing. In more rigorous approaches, the variance explained by lower level acoustical features common to all sound classes are taken into account for example as additional regressors in generalized linear models. Alternatively, an adaptation paradigm can be used where one searches for brain regions that are insensitive to acoustical changes within a sound class (for example a phonetic category) but sensitive to similar changes across the sound class (Joanisse, Zevin, & McCandliss, 2007). Although such systematic approaches can be deemed to be very rigorous, they suffer from two drawbacks. First, the majority of such studies have to focus on a single stage of speech processing, since investigating each processing step requires the design of specific stimuli. Thus, the investigation of the complete sound-to-language processing stream would require a very large number of studies and resources. Second, in most cases, the stimuli including the speech stimuli are artificial and isolated and, thus, the relevance of the results to natural speech processing should be questioned. To address these shortcomings, we propose in this dissertation a novel general approach using natural speech as stimuli with a combination of multiple modeling approaches designed to investigate the cortical localization of particular computational steps. Indeed, natural speech contains richly varied spectral, phonemic, syntactic, semantic and prosodic information, making it ideal for simultaneously studying multiple stages of speech processing. Earlier studies have showed that different regions of the cortex are involved in processing the long-range and short-range temporal structure of natural speech (Lerner, Honey, Silbert, & Hasson, 2011). In another study, natural speech sounds were used to map spectral selectivity in the early auditory cortex (Moerel, De Martino, & Formisano, 2012). These results suggest that every stage of speech processing can be studied simultaneously if natural

speech stimuli are employed. Additionally, studies have found that the auditory system is tuned both at low-levels and high-levels to natural sound statistics (Escabí, Miller, & Read, 2003; Smith & Lewicki, 2006; Garcia-Lazaro et al., 2006; reviewed in Theunissen et al. 2014)–therefore using naturally spoken speech may give additional insight into speech processing that studies segmented speech would not or would at a minimum validate the results obtained with segmented speech and artificial sounds.

The challenges are therefore to find appropriate stimuli for modeling, to develop statistical tools to allow us to fit all models, and to find appropriate transformations of the speech signal that adequately represent different stages of speech processing.

In Chapter 2 of this dissertation, we examine speech perception using fMRI voxel-wise modeling with natural speech stimuli. We investigate three different models of speech: spectral, articulatory, and semantic.

In Chapter 3, we develop a method to tease apart the variance explained by the different models, which is crucial when investigating simultaneous models. This new method allows us to understand which parts of the cortex are best predicted by individual models, joint models, or the combination of all three of our speech models.

Finally, Chapter 4 of this dissertation is a psychoacoustical investigation of the modulation power spectrum (MPS). The MPS, or the 2-dimensional Fourier transform of the speech spectrogram, is a transformation of the speech signal that could serve as an excellent intermediate step between lower and higher levels of speech processing, and could in future studies add extra nuance to our previous 3 cortical models of speech perception.

Chapter 2

A hierarchy of models predicts BOLD responses to natural speech

2.1 Introduction

The ease with which we understand speech belies the complexity of speech processing in the brain. Successful speech perception requires extracting relevant spectral and temporal components of the sounds, combining phonemes into words, and extracting meaning from the words. Previous research has shown that low-level acoustical structure, phonemes, and words are processed by distinct cortical areas. However, most studies on cortical representation of speech have been limited in scope, either in design or analysis. Thus little is known about the relationship between these different representations. To address this problem, we simultaneously mapped the representation of many different representations of speech.

In this study, we further developed model-based analyses of fMRI data at individual voxel levels (Kay, Naselaris, Prenger & Gallant, 2008; Nishimoto et al. 2011; Huth, Nishimoto, Vu & Gallant, 2012) and used natural speech as stimuli to demonstrate the validity and power of this alternative approach, compared to the standard subtraction based fMRI methods described in the Introduction. The stimuli consisted of 11 narrative stories with lengths between 10 and 15 minutes, each of which was presented uninterrupted and in its entirety. The sound stimuli were then represented in three feature spaces: a low level auditory representation based on the spectral content, an intermediate level representation based on the articulatory gestures that would produce the speech sounds and a high level representation based on the semantic content of each word. Our choice of these three representations clearly jumps many potential intermediate processing steps and ignores other computations required for language (such as syntax). It allowed us, however, to get an overview of speech and language processing in the entire cortex from lowest to highest levels. Also by examining the areas of the cortex that can be sensitive to very high-level

representation (semantics) of the speech signal we were able to also question a strict modular and bottom up view of language processing, which remains controversial and arguably inconsistent with behavioral data. Both at the level of production and perception, it is clear that mid-level and high-level processing interacts with lower level processing. On the production side, phoneme sequence can change the pronunciation and acoustic structure of particular phonemes through, for example, compensation for coarticulation (Mann, 1980). On the comprehension side, it is well-known that when particular phonemes within words are masked, listeners are incapable of determining which phoneme had been masked–the high level expectation of phoneme context produced the impression that the acoustic structure of the phoneme had been perceived (Warren, 1970; Kashino, 2006). As a result of our experimental design and our analyses, we not only parceled out brain regions that are uniquely or principally involved in processing the speech sounds at three hierarchal organized levels, but also investigated the extent to which lower-level characteristics such as spectral or articulatory information can affect the processing of semantics and vice-versa.

2.2 Materials and Methods

Subjects

Functional data were collected from four male subjects (S1: age 26, S2: age 31, S3: age 26, S4: age 32), and one female subject (S5, age 31). Two of the subjects were authors on this paper (AH and WAdH). All subjects were healthy and had no reported hearing problems. The use of human subjects in this study was approved by the UC Berkeley Committee for Protection of Human Subjects.

Stimuli

The natural speech stimuli consisted of monologues taken from *The Moth Radio Hour*. In each story from *The Moth Radio Hour*, a single male or female speaker tells an autobiographical story in front of a live audience. The speakers are chosen for their story telling abilities and their stories are engaging, funny and often emotional.

For our experiments, the stimuli were categorized into a model estimation set and a validation set. The model estimation dataset consisted of ten 10- to 15-minute stories. These stories were played only once in a single continuous scan. The length of each scan was tailored to the story and also included 10 seconds of silence both before and after the story. For each subject, we chose stories told by 5 male speakers and 5 female speakers to balance the sex of the speaker. The model validation dataset consisted of a single 10-minute story. The same story was played twice for each subject in order to estimate response reliability. The validation story was told by a female speaker. In this manner, we obtained blood oxygen level dependent (BOLD)

fMRI responses from each subject while they listened to approximately 2 hours and 20 minutes of natural speech stimuli. Model estimation and validation data were collected during two (separate) 2-hour scanning sessions.

MRI data collection

MRI data were collected on a 3T Siemens TIM Trio scanner at the UC Berkeley Brain Imaging Center using a 32-channel Siemens volume coil. Functional scans were collected using a gradient echo-EPI sequence with repetition time (TR) = 2.0045s, echo time (TE) = 31ms, flip angle = 70 degrees, voxel size = 2.24 x 2.24 x 4.1 mm, matrix size = 100 x 100, and field of view = 224 x 224 mm. 32 axial slices were prescribed to cover the entire cortex. A custom-modified bipolar water excitation radiofrequency (RF) pulse was used to avoid signal from fat tissue. Anatomical data were collected using a T1-weighted MP-RAGE (Brant-Zawadzski et al., 1992) sequence on the same 3T scanner.

Sound stimuli were played over Sensimetrics S14 in-ear piezoelectric headphones. These headphones provide both high audio fidelity and some attenuation of scanner noise. Berhinger Ultra-Curve Pro hardware parametric equalizer was used to flatten the frequency response of the headphones as suggested by the manufacturer. The sampling rate of the stimuli in their digitized form was 44.1 kHz and the sounds were not filtered before presentation. Thus, the potential frequency bandwidth of the speech stimuli was limited by the frequency response of the headphones from 100 Hz to 10 kHz. The sounds were presented at comfortable hearing levels, normalized to have a peak loudness of -1dB relative to max.

fMRI data pre-processing

Each functional run was motion-corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL 4.2 (Jenkinson & Smith, 2001). All volumes in the run were then averaged to obtain a high quality template volume. FLIRT was also used to automatically align the template volume for each run to the overall template, which was chosen to be the template for the first functional run for each subject. These automatic alignments were manually checked and adjusted for accuracy. The cross-run transformation matrix was then concatenated to the motion-correction transformation matrices obtained using MCFLIRT, and the concatenated transformation was used to resample the original data directly into the overall template space.

Low-frequency voxel response drift was identified using a 2nd order Savitsky-Golay filter with a 120-second window and subtracted from the signal. After removing this time-varying mean, the response was scaled to have unit variance and in this manner obtain a z-score value.

Flatmap construction

Cortical surface meshes were generated from the T1-weighted anatomical scans using Freesurfer software (Dale et al, 1999). Five relaxation cuts were made into the surface of each hemisphere and the surface crossing the corpus callosum was removed. The calcarine sulcus cut was made at the horizontal meridian in V1 using retinotopic mapping data as a guide. Functional data were aligned to the anatomical data for surface projection using custom software written in Python.

Stimulus transcription and preprocessing

In order to construct semantic and articulation models, it was necessary to determine the timing of each specific word and phoneme in the story. To do so, we first manually transcribed all of the stories. We then used software (the Penn Phonetics Lab Forced Aligner: previously available at https://www.ling.upenn.edu/phonetics/p2fa/index.html, currently only available as an online version at http://fave.ling.upenn.edu/) to automatically convert words to phonemes and align the transcriptions to each individual story. With this procedure, the beginning and end of each word and phoneme were estimated with millisecond accuracy. We further verified the alignments by hand, using Praat (www.praat.org) visualization software, to ensure that the automatically aligned phonemes and words were correctly identified and timed.

Voxel-wise model fitting and testing

Three sound representations, more generally called feature spaces, were used to linearly predict the BOLD activity of each voxel. For a sound stimulus $s(t)$, the feature space corresponding to each model can be written $\vec{\Phi}(s(t)) = \vec{\Phi}(t)$, where $\vec{\Phi}(t)$ is a vector of features that correspond to a more or less abstract representation of the stimulus (Fig 2.1). The linear model takes into account multiple past points that are used to predict the response at a single time:

$$\hat{r}(t) = \sum_{\tau}^{n} \vec{w}_{\tau} \cdot \vec{\Phi}(t - \tau)$$

where $\hat{r}(t)$ is the predicted response of the actual bold response, $r(t)$. Thus, this linear prediction takes the form of a finite impulse response (FIR) model (Nishimoto et al. 2011; Huth et al. 2012) and is also equivalent to spatio-temporal receptive fields where space consists of the stimuli in the feature space. In this study, the prediction only used the features at four delays (t-2s, t-4s, t-6s, t-8s), and thus the equation for the FIR can be written more specifically as

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

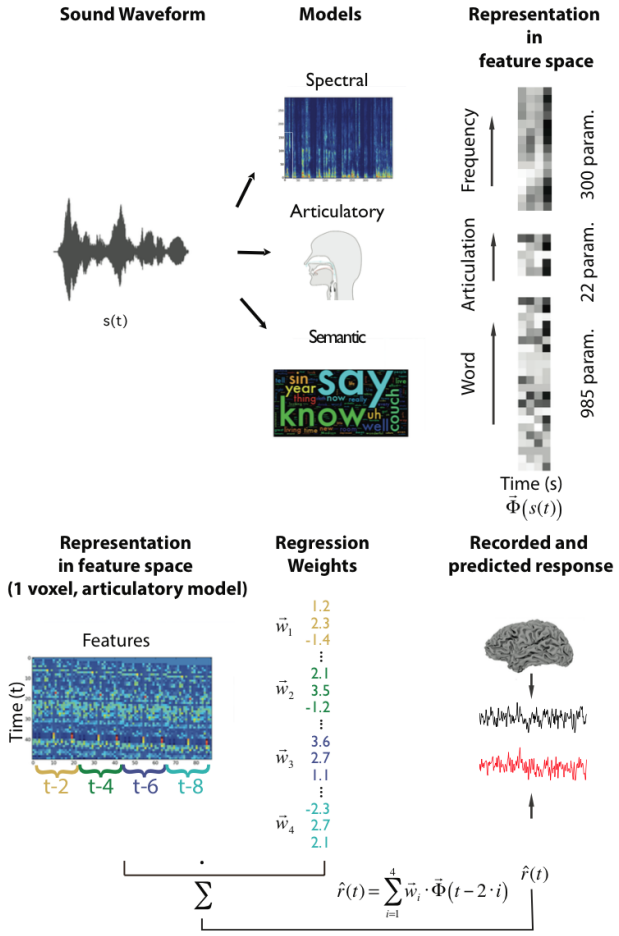

Figure 2.1: Model Construction. **Top row:** Feature spaces. Three models were used in three separate modeling efforts to predict the BOLD response of each voxel in each subject's brain: a spectral model, an articulatory model and a semantic model. Each model is realized by transforming the sound pressure waveform,$s(t)$,into a vector of values at time t, the feature space corresponding to each model, $\vec{\Phi}(s(t)) = \vec{\Phi}(t)$. The spectral features are the amplitudes of the frequency power spectrum signal calculated in a 2s window (at 300 frequency values), the articulation features are the count of which articulations (out of 22) were found in a 2s window, and the semantic features are a correlation measure of all the words from the story also counted in a 2s window with 985 most common words (see Methods for more details). **Bottom Row:** Linear Regression. The stimulus represented in a feature space (or combination of feature spaces) is then used in linear regression to obtain a prediction $\hat{r}(t)$ (red curve) of the actual bold response $r(t)$ (black curve) for each voxel in the brain. This linear regression is a linear filter with four point delays (t-2, t-4, t-6, t-8). The diagram illustrates this operation: a row in the feature matrix shown on the left corresponds to a time t of a response and shows in a color code the features (here the articulation) at times t-2, t-4, t-6 and t-8 unfolded as a single vector (a single vector). The dot product of that vector with the regression parameters (the w) yields the predicted response at time t. These parameters were obtained using ridge regression and model performance was assessed using cross-validation (see Methods).

It should also be noted that this model incorporates a -2s time delay that was not used in earlier publications from this group (Nishimoto et al. 2011; Huth et al. 2012) because auditory cortex has shorter hemodynamic delays than visual cortex. The three feature spaces investigated here were spectral frequencies (300 parameters), articulations (22 parameters), and semantics (985 parameters). The construction of these features spaces and the projection of the sound onto them are further described below. Before doing the regression, we z-scored (subtracted the mean and divided by the standard deviation) each feature channel within each story. This was done because each story includes different semantic and spectral content, and so has a different mean and variance in each feature channel. We estimated 7 models in total: each feature space independently, all pair-wise combinations of feature spaces and the combination of all three feature spaces.

The parameters of the models (i.e. the weights on feature space components for each voxel) were fitted using L2-regularized linear regression (a.k.a. ridge regression). To keep the scale of the weights consistent and prevent bias in subsequent analyses, a single value of the regularization hyperparameter was used for each feature space, for all voxels in all subjects. This regularization coefficient was found by bootstrapping the regression procedure 50 times in each subject for each model using only one feature space. In each bootstrap iteration, 800 time points (20 random blocks of 40 consecutive time points each) were removed from the training dataset and reserved for testing. Then the model parameters were estimated on the remaining 2937 time points for each of 20 possible regularization hyperparameters. These weights were used to predict responses for the 800 reserved time points and then the correlation between actual and predicted responses was found. After the bootstrapping was complete, a regularization-performance curve was obtained for each subject by averaging the bootstrap sample correlations first across the 50 samples and then across all voxels. Next, the regularization-performance curves were averaged across the five subjects and the best overall value of the regularization parameter was selected per model, for all subjects.

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

Determination of story-responsive voxels and quantification of model predictions

We determined which voxels were significantly active in response to the stories (story-responsive voxels) from the validation data set. This set consisted of BOLD responses to two repeats of the same story as describe above. We calculated the correlation between these paired BOLD responses and using the jackknifing resampling procedure estimated whether that correlation coefficient was significantly different from zero.

More precisely, the responses to the validation stimulus consisted of 290 time points. In the jackknifing procedure, we held out the same 10 timepoints from the two BOLD response trials 29 times (290/10 = 29 jackknifed sets), and then in this manner obtained 29 'delete one' estimates of the correlation coefficient. We then obtained the jackknife pseudo values of the correlation coefficient by extrapolation to infinite data size using the standard jackknife formula:

$$N\bar{X} - (N-1)\bar{X}_{[i]}$$

where $\bar{X}$ is the correlation values obtained from all the data, $\bar{X}_{[i]}$ is the correlation value for the ith 'delete one' estimate, and N is the number of jackknife samples taken (here N=29). We estimated the standard error of the correlation value for each voxel by taking the standard error of the 29 jackknife pseudo-values. To determine significance, we subtracted twice this standard error (SE) from the correlation (Corr $= \bar{X}$) value for each voxel (V): any voxel such that $Corr_S - 2SE_S > 0$ was considered to have significant activation in response to our story stimuli (and is called a story-responsive voxel). These significant 'signal' correlation coefficient values, $Corr_S$, were also saved as a measure of the maximum correlation that we could expect between the predictions of our model and the response of one trial. Next, we first estimated which voxels among story-responsive voxels had responses that could be explained significantly by our models and then quantified the goodness of fit of these predictions. Separate predicted responses were obtained for all the different models for the validation data set and for story-responsive voxels. We then estimated the average correlation between the model predictions and each of the validation trials: the model correlation or $Corr_M$. SE of $Corr_M$ were obtained by jackknifing following the same procedure are above and where, for the delete-one jackknife values, the same section of stimulus was deleted from each trial. As above, any voxel among story-responsive voxels for which $Corr_S - 2SE_S > 0$ was considered to have BOLD activity that could be significantly predicted by that particular model.

Spectral feature space

We extracted the spectral components of the story sound files using custom MATLAB software. This frequency power spectrum was estimated for each TR time point by a classical Welch method for spectral estimation density. The sound signal was first multiplied with Gaussian-shaped windows that had a standard deviation parameter of 5 ms (corresponding to a frequency resolution of 32 Hz), a length of 30 ms and with successive windows being 5 ms apart.The periodogram was calculated for each windowed signal using the amplitude square of the discrete Fourier transform and 40 of the successive estimates average to obtain a power spectrum at every 2 second time point. These power spectra consisted of 300 dimensional vector that contained the power of the signal between 25 Hz and 10 kHz in 32 Hz bands.

Articulation feature space

For the articulation feature space, we converted phonemes obtained from the alignment preprocessing into their articulatory features. Each phoneme is represented as a combination of articulatory features (manner, place and phonation for consonants; height and backness for vowels: see Table 1). We created a 22-dimensional vector with a unique pattern of articulations per phoneme, measuring manner and place and phonation for consonants, phonation, and height and backedness for vowels. We then created a binary feature vector, indicating whether the particular articulatory features were present or absent during a particular 2-second window (corresponding to the fMRI acquisition rate).

Semantic feature space

For the semantic feature space, we constructed a 985-parameter semantic feature space based on word co-occurrence statistics in a large corpus of text (Mitchell 2008; HAL; LSA). First we constructed a 10,470 word lexicon from the union of the set of all words appearing in the stimulus stories and the set of the 10,000 most common words in the training corpus. We then selected 985 basis words from the Wikipedia List of 1000 basic words (contrary to the title, this list only contains 985 unique words). This basis set was selected because it consists of common words and spans a very broad range of topics. The training corpus used to construct this model includes the transcripts of 13 Moth stories (including the 10 used as stimuli), 604 popular books, 2,405,569 wikipedia pages, and 36,333,459 user comments scraped from reddit.com. In total the 10,470 words in our lexicon appeared 1,548,774,960 times in this corpus.

Next, we constructed a word co-occurrence matrix, M, with 985 rows and 10,470 columns. Iterating through the training corpus, we added 1 to Mi,j each time word j appeared within 15 words of basis word i. A window size of 15 was selected to be large enough to suppress syntactic effects (i.e. word order) but no larger. Once the word co-occurrence matrix was complete, we log-transformed the counts, replacing Mi,j with log(1 + Mi,j). Next each row of M was z-scored to correct for differences in basis word frequency, and then each column of M was z-scored to correct for word frequency. Each column of M is now a 985-dimensional semantic vector representing one word in the lexicon. This representation tends to be semantically smooth, such that words with similar meanings (such as dog and cat) have similar vectors, but words with very different meanings (such as dog and book) have very different vectors.

The semantic feature space was constructed from the word representation of the stories: for each word-time pair (w, t) in each story we selected the corresponding column of M, creating our semantic feature vector:

$$\vec{\Phi}(t) = \left(\vec{M}_w, t\right)$$

Table 1: phoneme to articulation conversion chart

Phoneme	Articulatory Features			
B	bilabial	plosive	voiced	
CH	post-alveolar	affricate	unvoiced	
D	alveolar	plosive	voiced	
DH	dental	fricative	voiced	
F	labio-dental	fricative	unvoiced	
G	velar	plosive	voiced	
HH	glottal	fricative	unvoiced	
JH	post-alveolar	affricate	voiced	
K	velar	plosive	unvoiced	
L	alveolar	lateral	voiced	
M	bilabial	nasal	voiced	
N	alveolar	nasal	voiced	
NG	velar	nasal	voiced	
P	bilabial	plosive	unvoiced	
R	alveolar	approximant	voiced	
S	alveolar	fricative	unvoiced	
SH	post-alveolar	fricative	unvoiced	
T	alveolar	plosive	unvoiced	
TH	dental	fricative	unvoiced	
V	labio-dental	fricative	voiced	
W	velar	approximant	voiced	
Y	palatal	approximant	voiced	
Z	alveolar	fricative	voiced	
ZH	post-alveolar	fricative	voiced	
AA	low	back		
AE	low	front		
AH	mid	central		
AO	mid	back		
AW	low	central	mid	back
AY	low	central	mid	front
EH	mid	front		
ER	mid	central		
EY	mid	front		
IH	mid	front		
IY	high	front		
OW	mid	back		
OY	mid	back	high	front
UH	high	back		
UW	high	back		

These lists of vectors were then downsampled, by summing frequency bins across time, to the fMRI acquisition rate.

Models' centers of mass in the auditory area

We calculated the center of mass along two axes of the auditory area for our three different models. We first projected all of the voxels within our auditory area (defined by a custom sound localizer) onto two different axes: the anterior/posteroventral axis, and the medial/ventral axis. We then calculated the center of mass of each model and axis, for each subject:

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

where a_i is the location of the i^{th} voxel projected on the chosen axis (scaled from 0 to 1), and r is the performance (correlation or Pearson's r) of that voxel for the chosen model.

We used bootstrapping to calculate the standard errors of these calculated centers of mass. For each model, subject, and axis, we sampled 1000 points (with replacement) along the chosen auditory axis, 1000 times. We then calculated the center of mass of each sample, and calculated the standard error of the mean from these data.

2.3 Results

The goal of this study is to investigate the cortical responses to distinct features in natural speech, in order to begin to decipher the computations performed for speech comprehension in different brain regions. To do so, we examined the representation of three hierarchical sets of speech features that would allow us to follow the transformation from sound to meaning. We used spectral, articulatory and semantic feature spaces to fit models to the same BOLD data in subjects listening to recorded stories told to a live audience (see Methods for more details). We first fitted models based on each of the three feature spaces separately, in order to see how well and where each level could predict BOLD responses. The goodness of fit of the models was quantified by calculating the correlation between the predicted and actual responses in a validation set. The statistical significance of these correlations (statistically different from zero) was obtained through jackknifing (see Methods). A large fraction of the cortex was significantly activated by our stimulus. We found (Table 2) that 44% of voxels in the auditory area of the left hemisphere and 41% of voxels in the auditory area of the right hemisphere were story-responsive voxels (i.e. responded significantly to our story stimuli, see Methods). In the cortex (excluding the auditory area), 22.5% of voxels responded significantly in the left hemisphere and 21.7% of voxels responded

Table 2. Percent of cortex significantly active in response to stories, mean of 5 subjects +/- SE

ROI	Left Hem	Right Hem	Both Hems
Auditory Cortex	43.84 (±3.24)	40.55 (±3.93)	42.03 (±3.34)
Whole Cortex	23.45 (±3.39)	22.76 (±3.45)	23.11 (±3.35)
Cortex - AC	22.46 (±3.45)	21.66 (±3.44)	22.07 (±3.38)

significantly in the right hemisphere. A linear mixed-effects analyzing the percentage of story-responsive voxels with hemisphere (two levels: left and right) and cortical regions (two levels: auditory cortex, and the whole cortex excluding auditory cortex) as fixed effects and subject as random effect showed that region was significant ($p < 2e^{-16}$) but the two hemispheres were not significantly different from one another (p = 0.1629), nor was the interaction between hemisphere and region significant (p = 0.3981).

Model construction

Three feature spaces were used in a combination of modeling efforts to predict the BOLD response of each voxel in each subject's brain: a spectral feature space, an articulatory feature space and a semantic feature space (Fig. 2.1). The features are obtained by transforming the sound pressure waveform, $s(t)$ into a vector of values at time t, the feature space $\vec{\Phi}(s(t)) = \vec{\Phi}(t)$. The spectral features are the power spectra calculated in a 2s window at 300 frequency values between 25 Hz and 10 kHz (32 Hz bands). The articulatory features are a binary mask marking which articulations (out of 22 possible) were found in a 2s window. These 22 articulatory features are used in unique pattern for each phoneme of the English language (Table 1). The semantic features were derived from a word embedding space that was constructed by computing the normalized log probability of co-occurrence between each word in the stories and a set of 985 common English words (such as "above", "pencil", and "worry") across a large corpus of English text. Words related to the same semantic domain tend to occur in similar contexts, and so are assigned similar vectors in this 985-dimensional space. For example, the words "month" and "week" are very similar (the correlation between the two is 0.74), while the words "month" and "tall" are not (correlation -0.22). We used this space to transform each word that was spoken in the stories into a 985-dimensional vector. The sound in a feature space is then used in linear regression to obtain a prediction $\hat{r}(t)$ (red curve in Fig. 2.1) of the actual bold response, $r(t)$ (black curve) for each voxel in the brain. This approach is called Voxel-wise modeling (VM). Here the VM takes the form of a linear filter with four

point delays (t-2, t-4, t-6, t-8): $\hat{r}(t) = \sum_{i=1}^{4} \vec{w}_i \cdot \vec{\Phi}(t - 2 \cdot i)$.

Note that the number of parameters of the model is given by 4 times the dimensionality of the feature space. For example, the number of parameters in a model using the spectral features is 4x300 = 1200. We also generated models using all pair-wise combination of the features spaces and one model using all three feature spaces. For example if $\vec{\Phi}_{Art}(t)$ is the articulation feature space and $\vec{\Phi}_{Sem}(t)$ is the semantic feature space, a joint model using the semantic and articulation features can be written as a linear filter with: $\hat{r}(t) = \sum_{i=1}^{4} \vec{w}_{Art,i} \cdot \vec{\Phi}_{Art}(t - 2 \cdot i) + \sum_{i=1}^{4} \vec{w}_{Sem,i} \cdot \vec{\Phi}_{Sem}(t - 2 \cdot i)$ The parameters of the model, $\vec{w}$, were fit with a training data set by maximum likelihood with regularization given by Gaussian prior with zero mean on the model parameters (ridge regression). Regularization is required given the high number of parameters in our feature spaces (or combination of feature spaces). In addition, to be able to compare nested models (in Chapter 3), we forced the optimal shrinkage obtained in each ridge regression using a single feature space to be used in the regressions that combined feature spaces. This novel method is described in the Methods of Chapter 3 (Joint Ridge Regression). Model validation was performed by estimating bold responses on a separate set of stories that the ones used to fit the model parameters and quantified by calculating both the Pearson's correlation coefficient and the corresponding R^2 between the predicted and actual responses obtained for each voxel.

Model Predictions for each feature space

To determine which areas of the cortex are involved in representing the sound spectrum, articulations, and semantics, we first estimated the prediction performance values for each model based these three features spaces taken once at the time. These model performances are shown for the story-responsive voxels in Fig. 2.2. Here, the data are shown projected onto cortical flatmaps that were specially constructed for each subject. The spectral model predicts early auditory areas (Fig 2.2a) as shown by activation in the more anterior area of the region labeled AC, which is medial to the superior temporal gyrus and near Heschl's gyrus. The articulatory model predicts in early auditory areas (around Heschl's gyrus and medial superior temporal gyrus), sensory and motor mouth areas, in the left hemisphere Broca's area, and more sparsely in prefrontal areas (Fig 2.2b). The semantic model predicts activity in large areas and widespread areas of the cortex. These include the later auditory cortex (more caudal region of the auditory region indicated by AC), lateral regions of the temporal cortex, many areas in the parietal cortex and specifically along the temporal-parietal junction and in the medial parietal cortex (precuneus), and many regions of the prefrontal cortex (Fig 2.2c). These areas together have been previously defined as the semantic system (Binder et al. 2009). For all three feature spaces, there was a clear lack of predictive power in the visual cortex, the somato-sensory

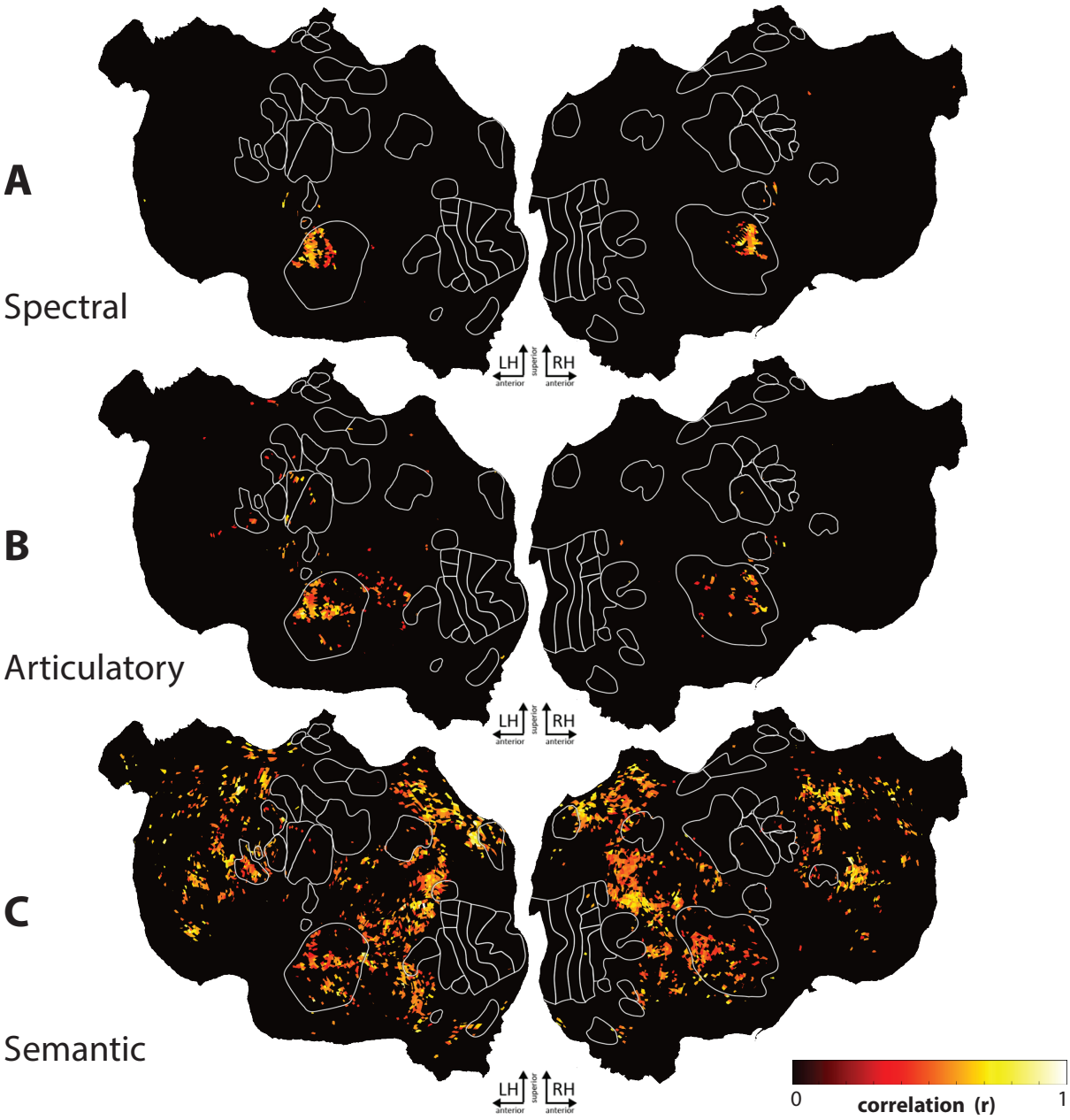

Figure 2.2: Model performance. **A**: Spectral Model Performance. Spectral model performance plotted on the flattened cortical surface of one subject. The color scale (black to red/white) is used to show the value of r (Pearson product-moment correlation coefficient) obtained by comparing the prediction of the model to the actual BOLD activity for the stories in the validation data set. All voxels for which the correlation is not significantly different from zero (assessed by jackknifing, see Methods) are shown in black. The thin white lines encircled different functional/anatomical regions of the brain obtained from localizers. **B**: Articulation Model Performance. Articulation model performance plotted on the flattened cortical surface of one subject. **C**: Semantic Model Performance. Semantic model performance plotted on the flattened cortical surface of one subject.

Table 3. percent voxels significant (of predictable voxels) per model, mean of subjects +/- SE

Model	Left Hemisphere	Right hemisphere
Spectral Model		
Auditory Cortex	41.54 (±6.22)	37.79 (±3.79)
Whole Cortex	11.85 (±1.87)	9.52 (±1.40)
Cortex – AC	9.13 (±1.71)	6.09 (±1.00)
Articulatory Model		
Auditory Cortex	59.37(±3.64)	56.57 (±3.24)
Whole Cortex	23.80 (±3.56)	17.57 (±1.14)
Cortex – AC	20.13 (±3.97)	12.60 (±0.60)
Semantic Model		
Auditory Cortex	59.88 (±2.31)	65.59 (±3.86)
Whole Cortex	45.56 (±5.40)	42.49 (±6.21)
Cortex – AC	43.55 (±6.20)	39.22 (±7.01)

cortex or the motor cortex.

The models taken together can generate significant prediction for a considerable portion of the story-responsive voxels (see Table 3 for all values). The spectral feature space generates significant predictions in 42% (left hemisphere) and 38% (right-hemisphere) of story-responsive voxels in auditory areas. However, it only generates significant predictions in 9% (left hemisphere) and 6% (right hemisphere) of story-responsive voxels in the cortex outside of auditory areas. The articulation feature space generates significant predictions in 59% (left hemisphere) and 57% (right-hemisphere) of story-responsive voxels in auditory areas. It further significantly predicts 20% (left hemisphere) and 13% (right hemisphere) of story-responsive voxels in the cortex excluding core auditory areas. Finally, the semantic feature space significantly predicts 60% (left hemisphere) and 66% (right-hemisphere) of story-responsive voxels in auditory areas. It significantly predicts 44% (left hemisphere) and 39% (right hemisphere) of story-responsive voxels in the cortex excluding auditory areas, a much greater percentage than the spectral or the articulatory models. A linear mixed-effect model comparing percentage of significantly predictable story-responsive voxels with feature space (3 levels), hemispheres (two levels), and cortical regions (two levels: auditory areas, and the whole cortex excluding auditory areas) as fixed effects and subject as random effect shows that region ($p < 2.2e^{-16}$) and feature space

Table 4. prediction correlation (r) per model (prediction performance of story-responsive voxels significantly predicted by each model), mean of subjects +/- SE

Model	Left Hemisphere	Right hemisphere
Spectral Model		
Auditory Cortex	.2311 (±.0123)	.2207 (±.0075)
Whole Cortex	.1892 (±.0055)	.1846 (±.0030)
Cortex – AC	.1679 (±.0011)	.1564 (±.0014)
Articulatory Model		
Auditory Cortex	.2243(±.0071)	.2176 (±.0072)
Whole Cortex	.1889 (±.0040)	.1879 (±.0036)
Cortex – AC	.1770 (±.0052)	.1713 (±.0010)
Semantic Model		
Auditory Cortex	.2260 (±.0079)	.2317 (±.0098)
Whole Cortex	.2230 (±.0058)	.2226 (±.0070)
Cortex – AC	.2283 (±.0071)	.2182 (±.0086)

($p < 2.2e^{-16}$) are significant, however hemisphere is not (p = 0.2262). There is also a significant interaction between region and feature space (p = 0.0007): the spectral and articulatory feature spaces predict far more voxels in auditory areas than outside auditory areas. This trend continues but is much less pronounced for the semantic feature space. However, we see no significant differences between hemispheres either in main effects or interactions.

In Table 4, we show the average correlation per region and per hemisphere of the story-responsive voxels that were significant– predictions were obtained for each feature space. A linear mixed-effect model comparing average correlation of significantly predictable story-responsive voxels (see Table 4) with feature space (three levels), hemispheres (two levels), and cortical regions (two levels: auditory areas and the entire cortex excluding auditory areas) as fixed effects and subject as random effect shows that region ($p < 10^{-4}$) and feature space ($p < 10^{-4}$) are significant, however hemisphere is not (p = 0.167). Auditory areas tend to be better predicted than non-auditory areas, and the semantic feature space overall predicts better than the articulatory or the spectral feature spaces. There is also a significant interaction between region and feature space ($p < 10^{-4}$). The spectral and the articulatory feature spaces predict better in auditory areas than outside of auditory areas, however

this is not the case for the semantic feature space.

The model predictions also show a great amount of overlap. More specifically, very similar areas of AC are well predicted by the spectral and articulatory model and the parietal and pre-frontal areas that are predicted in the articulatory model are also predicted by the semantic model. In the next Chapter, we will perform various forms of model comparisons to disentangle the information that uniquely predicted by each feature space to the information that is redundantly represented by two or three of these features.

Comparison of the three models based on single feature spaces

To compare the predictive power of models based on each of the feature spaces, we next plotted in different colors (Fig. 2.3) the model with the best correlation between actual and predicted BOLD activity: here each voxel that was significantly predicted by one of the three models, takes on a single color corresponding to the feature space that yielded the highest correlation (blue for spectral, green for articulatory, red for semantic). While the spectral model's best predictions are restricted to early auditory areas, mainly along Heschl's gyrus, the articulatory model best predicts more lateral/caudal regions of the auditory areas, which have been assigned to more secondary auditory areas. There is also a small amount of peak predictive activity in pre-motor areas, corresponding presumably to sensory and motor mouth areas. The semantic model predicts best areas in later auditory areas, other areas of the temporal cortex, the parietal cortex and much of prefrontal cortex. The auditory areas are thus a region where all three models can yield best predictions and moreover the regions of best predictions appear to be systematically mapped from auditory in lower auditory areas to semantic in higher auditory areas and articulation in the middle. That relationship is clearly seen in Fig. 2.4a that shows the best correlations only for the AC region. Specifically, the spectral model runs along Heschl's gyrus (early primary auditory cortex). The semantic model best predicts along superior temporal gyrus. The articulatory model best predicts medial to Heschl's gyrus, on the lateral portion of superior temporal gyrus. A clear hierarchical map is found along the medio-ventral axis but not the anteroventro-posteroventral axis: the spectral model performs best medially, followed by the articulatory model, and finally the semantic model that performs best on the most lateral portion of the auditory areas. To quantify this map, we projected the correlation values of each model along the medio-ventral axis and anteroventral-posteroventral axis of the auditory areas (Fig. 2.4b). Here, again, we find a functional hierarchical map along the medio-ventral axis: the spectral model performs best medially, followed by the articulatory model more laterally, and finally the semantic model performs best on the most lateral side of the auditory areas. Along the rostro-caudal axis, however, the spectral, articulatory and semantic models follow similar predictive curves with best predictions in the central regions of AC. We then calculated the centers of mass for each model. We find that the centers of mass are

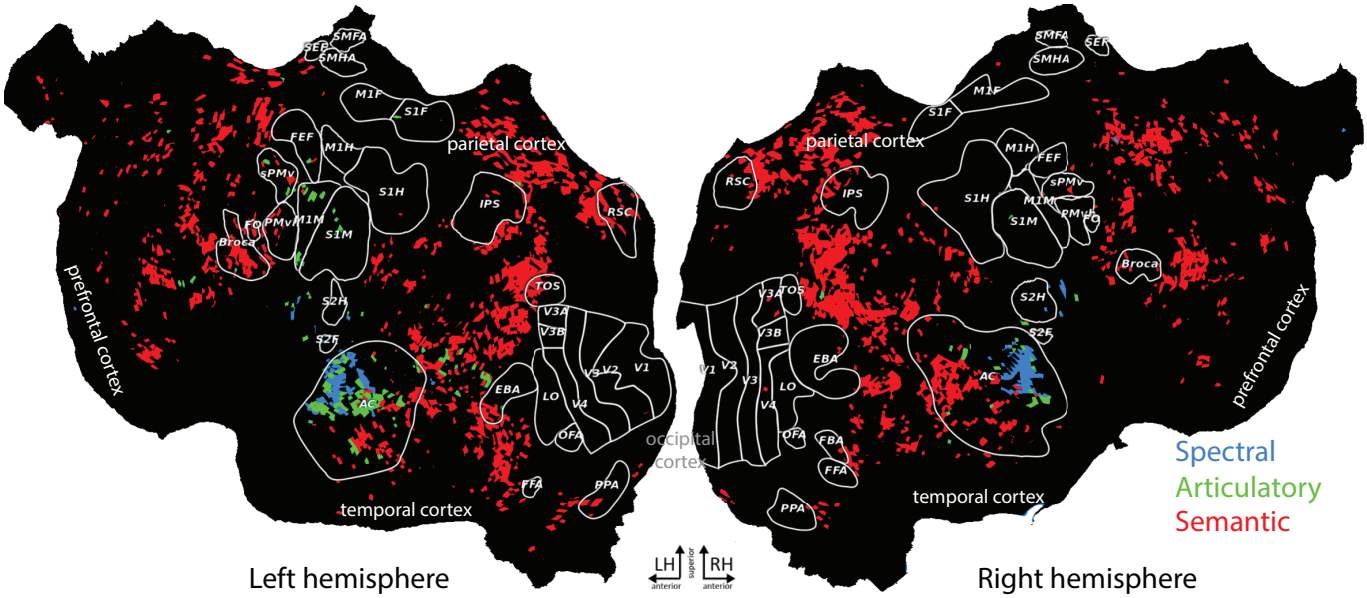

Figure 2.3: Spectral, articulatory and semantic encoding models of one subject projected onto flattened cortical surface. Each voxel is drawn with a single color corresponding to the model that yielded the best prediction: blue (spectral), green (articulatory), and red (semantic). Voxels with model predictions not significantly different from zero are shown in black. Note that since only the best-performing model color is shown for each voxel, the voxel may be well-predicted by other models as well.

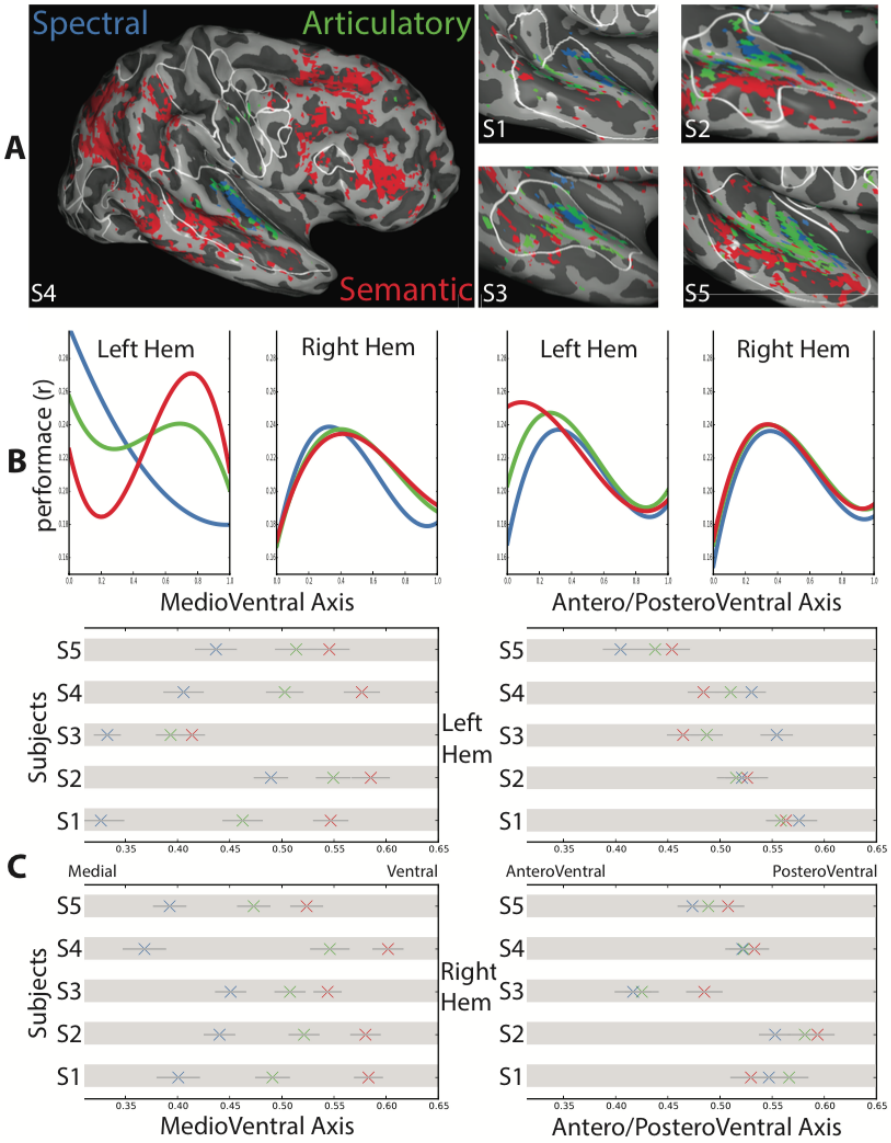

Figure 2.4: Prediction of the Three Models in Auditory Cortex **A**: Voxels best predicted by each model are shown on the cortex in the models' color: spectral (blue), articulatory (green), semantic (red). As in Figure 3, model predictions must meet a threshold of significance (2 standard deviations above zero, derived from jackknifing) in order to be included and only the best-performing model color is shown for each voxel. Each model occupies a different region within the auditory cortex, and their separation follows anatomical features. The entire right hemisphere of Subject 4 is shown, and the right auditory regions of the four other subjects are shown. **B**: Pearson's correlation coefficient for the predictions obtained from three models averaged for all voxels and all subjects at a given position along the medio-ventral axis (left two figures) and the anteroventral-posteroventral axis (right two figures). The data are smoothed using a univariate spline. **C**: The centers of mass were computed for each model, hemisphere, axis and subject. They are displayed in each model's color, with two standard errors displayed as error bars. For each figure, each line corresponds to one subject.

clearly ordered (spectral, then articulatory, then semantic) for all subjects along the medio-ventral axis, but not along the anteroventral-posteroventral axis (Fig. 2.4c). The model's order along the medio-ventral axis is significant for both the left and right hemisphere (Friedman rank-sum test: left hemisphere: Friedman chi-squared = 10, df = 2, p-value = 0.007; right hemisphere: Friedman chi-squared = 8.4, df = 2, p-value = 0.015). However, the order is not significant for the anteroventral-posteroventral axis (left hemisphere: Friedman chi-squared = 1.6, df = 2, p-value = 0.45 right hemisphere: Friedman chi-squared = 3.6, df = 2, p-value = 0.17).

2.4 Discussion

In this study, we examined neural responses in the cortical processing streams involved in language processing by modeling BOLD responses to naturally spoken stories. Consistent with previous work, we observed that a large portion of the cortex is active (approximately one fifth of the entire cortex and 42% in auditory areas) when listening to stories, implicating primary and secondary auditory areas, association areas (Bornkessel-Schlesewsky et al., 2015; Rauschecker & Scott, 2009) and much of the language network (Binder, Desai, Graves, & Conant, 2009). We then show that we are able to explain a significant fraction of the BOLD activation using encoding models and three hierarchically organized feature spaces: spectral, articulatory and semantics. For example, in auditory areas, we can obtain significant predictions in more than 60% of the story responsive voxels with the semantic model. Neither the response as assessed by the number of story-responsive voxels, nor the predictive power of our models showed a lateralization effect.

We found that these three feature spaces produce very different patterns of predicted BOLD activity across the cortical surface. The spectral model accurately predicts activity in early auditory areas such as Heschl's gyrus in A1. The articulatory model also accurately predicts early auditory areas, and additionally Broca's area and motor areas. The semantic model accurately predicts many different areas of cortex, including higher auditory areas, much of parietal cortex, and much of prefrontal cortex. Moreover semantic features tend to be represented more broadly across cortex than lower-level acoustical features and provide unique information even at level of the primary auditory cortex. The predicted activity is lateralized, with the left auditory regions predicted significantly better by lower and mid-level models than the right auditory regions.

We chose these three feature spaces to generate a first coarse functional map for low, mid and high-level computations involved in language processing from natural speech stimuli. One of the most striking aspects of our results is that the spectral and articulation feature spaces implicated very few cortical areas as solely involved in low or mid-level processing. Both are mostly restricted to primary and secondary auditory cortical areas along Heschl's gyrus (spectral) and posteroventrally on the

STS (articulation). The strong predictive power of the spectral features in early auditory cortex in response to speech is not surprising as previous studies have shown that this area is tonotopically organized when stimulated both by pure tones, and more recently natural speech (Da Costa, van der Zwaag, Miller, Clarke, & Saenz, 2013; Moerel et al., 2012).

We have shown in this chapter that it is possible to investigate different levels of speech processing using different transformations of the same stimulus. Although there appears to be a hierarchy contained within the models, in order to fully investigate this hierarchy, it is necessary to disambiguate between the kinds of variances explained by the different models. This is what we will do in Chapter 3.

Chapter 3

Disentangling unique and overlapping explained variance of a hierarchy of models predicting BOLD responses to natural speech

3.1 Introduction

In Chapter 2, we fitted models based on each of three feature spaces (spectral, articulatory, and semantic) separately, in order to see how well and where each level could predict BOLD responses. We expand on this analysis in this Chapter. In order to estimate the redundant and unique contribution of each feature space, we fitted a model that used the three features together, as well as models that used features in pairs. The goodness of fit of the models was quantified by calculating the correlation between the predicted and actual responses in a validation set. The statistical significance of these correlations (statistically different from zero) was obtained through jackknifing (see Methods, Chapter 2). Finally, by using nested models, we are able to explore the amount of variance explained uniquely by each individual model, as well as the variance explained by the intersections of models.

3.2 Materials and Methods

For this Chapter, we need additional methods in order to fit our models simultaneously and investigate the variance explained uniquely or jointly by our models. The methods for subjects, stimuli, MRI data collection, fMRI data pre-processing, flatmap construction, stimulus transcription and preprocessing, preliminary voxel-wise model fitting and testing, voxel significance, and spectral, articulatory and semantic feature spaces can be found in Chapter 2.

Joint ridge regression

In our models we are predicting responses, $\vec{r}$, from stimulus parameters $\mathbf{S}$. Here $\vec{r}$ is $n \times 1$ column vector corresponding to the BOLD response of a single voxel as a function of n discrete sampling points in time. $\mathbf{S}$ is an $n \times p$ matrix where each row corresponds to the stimulus at the same n time points as $\vec{r}$ and the columns correspond to the values describing the stimulus in its feature space. The columns can include values of $\mathbf{S}$ in past times. The maximum likelihood solution for the multiple linear regression is given by the normal equation:

$$\vec{h} = \frac{\langle Sr \rangle}{\langle SS \rangle} = \left[S^T S\right]^{-1} \left[S^T \vec{r}\right]$$

where $\vec{h}$ is the column vector of coefficients ($p \times 1$) also known as the filter. The <> stands for averaging the cross products across time (across rows). Note, more specifically, that the correct unbiased estimate of the stimulus-response cross-covariance (the numerator) and the stimulus auto-covariance (the denominator) are

$$\langle Sr \rangle = \frac{S^T \vec{r}}{n-1}$$

and

$$\langle SS \rangle = \frac{S^T S}{n-1}$$

The prediction can then be obtained by:

$$\hat{\vec{r}} = S \cdot \vec{h}$$

The inverse of the symmetric and positive definite stimulus auto-covariance matrix can easily be obtained from its eigenvalue decomposition or equivalently from the singular value decomposition (SVD) of $\mathbf{S}$. The SVD of $\mathbf{S}$, can be written as: $S = VWU^T$ where $\mathbf{V}$ is a $p \times p$ matrix of orthonormal input vectors in columns (or left singular vectors), $\mathbf{W}$ is a diagonal $p \times n$ matrix of positive single values and $\mathbf{U}$ is a $n \times n$ matrix of orthonormal output vectors in columns (or right singular vectors). The eigenvalue decomposition of $S^T S$ is then given by:

$$S^T S = VW^2 V^T$$

where W^2 is the $p \times p$ diagonal matrix of eigenvalues. To prevent over-fitting (when p is large relative to n), a regularized solution for $\vec{h}$ can be obtained by Ridge regression. The Ridge regression is the maximum a posteriori solution (or MAP solution) with a Gaussian prior on $\vec{h}$ with zero mean and covariance matrix given by $I\lambda$.

Under these assumptions, the MAP solution is:

$$\vec{h} = V(W^2 + \lambda I)^{-1} V^T S^T \vec{r}$$

If the normal equation can be interpreted as solving for $\vec{h}$ in the whitened-stimulus space (uncorrelated by the rotation given by $\mathbf{V}$ and normalized $\mathbf{W}$), the ridge regression decorrelates the stimulus space but performs a weighted normalization where the uncorrelated stimulus parameters with small variance (or small eigenvalues) are shrunken more than those with higher variance (or higher eigenvalues). The level of this relative shrinkage is controlled by the hyper-parameter λ and its optimal value is found by cross-validation (see Chapter 2, Voxel-wise model fitting and testing, Methods). In our analysis, we fit models using individual and combinations of feature spaces that have different units, a different number of parameters, p, and different degrees of correlation across those parameters. For each of these stimulus feature spaces, we also obtained a different optimal value for the ridge parameter, λ. As we explain below, we are interested in combining models to determine the overlap across stimulus representation in terms of their explanatory power: the shared variance explained. One option would be to obtain a new regression using the combined features spaces (after z-scoring) and obtain a new value of λ for that combined model. However, this approach would result in a different shrinkage in the combined model versus the component models this is particularly true when models have different number of parameters such as in the semantic and articulatory models we explored. By forcing the shrinkage in the each of the components when they are used in the combined model to be the same as when they are estimated separately, we will be able to accurately compare the combined and each of the separate models, and in this manner assess the amount of variance explained by each as well as the overlap. To do so, we performed the regression in the rotated and scaled basis obtained for each of the models. The stimulus space in that new basis set is noted with a prime in the equations below. We then perform a decorrelation of the joint stimuli but refrain from performing any additional normalization:

$$S_1'^T = \left[(W_1^2 + \lambda_1 I)^{-1/2} V^T S^T\right] \sqrt{n-1}$$

$$S_2'^T = \left[(W_2^2 + \lambda_2 I)^{-1/2} V^T S^T\right] \sqrt{n-1}$$

Having whitened the stimuli, we then need to create a correlation coefficient matrix from the covariance matrix that we can use to decorrelate the stimuli. The stimulus covariance matrix in this new stimulus space (denoted with the prime) can be obtained with

$$S'^T{}_{12} S'_{12}$$

divided by $n-1$, or:

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

where σ is variance and c is the covariance between individual parameters in each of the two feature spaces. The first index is for the feature space corresponding to model 1 or 2, and the second index runs over the parameters in that feature space. p_1 is the number of parameters in model 1 and p_2 the number of parameters in model 2. As one can notice from the form of this covariance matrix, the stimulus parameters are uncorrelated within each subset (since we already performed the de-correlation) but they are not perfectly white (because of the relative shrinkage performed by the ridge). Therefore, the variance in the diagonals is not exactly equal to 1 but slightly smaller and with decreasing values along each block diagonal. If at this stage we applied the normal equation, we would effectively remove the shrinkage performed in the ridge solution. Instead we will replace the covariance matrix with the correlation matrix. In this manner, we can decorrelate the stimulus features across the two component models while preserving the exact shrinkage that was performed in the separate ridge regressions. The correlation matrix obtained from the covariance matrix is given by:

$$Corr(S_{1,2}') = \begin{pmatrix} 1 & 0 & 0 & \frac{c_{1,1;2,1}}{\sigma_{1,1}\sigma_{2,1}} & \cdots & \frac{c_{1,1;2,p_2}}{\sigma_{1,1}\sigma_{2,p_2}} \\ 0 & \ddots & 0 & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & \frac{c_{1,p_1;2,1}}{\sigma_{1,p_1}\sigma_{2,1}} & \cdots & \frac{c_{1,p_1;2,p_2}}{\sigma_{1,p_1}\sigma_{2,p_2}} \\ \frac{c_{1,1;2,1}}{\sigma_{1,1}\sigma_{2,1}} & \cdots & \frac{c_{1,p_1;2,1}}{\sigma_{1,p_1}\sigma_{2,1}} & 1 & 0 & 0 \\ \vdots & \ddots & \vdots & 0 & \ddots & 0 \\ \frac{c_{1,1;2,p_2}}{\sigma_{1,1}\sigma_{2,p_2}} & \cdots & \frac{c_{1,p_1;2,p_2}}{\sigma_{1,p_1}\sigma_{2,p_2}} & 0 & 0 & 1 \end{pmatrix}$$

We then can calculate our combined ridge filter as

$$\vec{h}_{12}' = Corr[S'_{12}]^{-1} {S'_{12}}^{T} \vec{r} \cdot (n-1)^{-1}$$

and obtain predictions from the combined model with the equation:

$$\hat{\vec{r}} = S'_{12} \cdot \vec{h}_{12}'$$

Although we used here for clarity an example of performing joint ridge regression on two models, it is possible to extend joint ridge regression to any number of models.

Calculation of individual model variance

We used set theory to calculate the individual variance explained by each portion of the models. Given our three models, spectral, articulatory, and semantic, our aim is to find the variance explained (R^2) by each individual model alone (which has no shared variance explained with any other model); the variance explained by the intersections of the pairs of models, and the variance explained by the intersection of all three models. To do so, we used results obtained from fitting individual models and combination of two and three models and compared the R^2 of these nested models. We then used set theory to calculate the common (as a set intersection) and unique (as a set difference) variance explained (see Partitioning of variance, below, for further details).

Partitioning of variance

Given three models A, B and C, our aim is to find the variance explained (R^2) by each individual model alone (which has no shared variance explained with any other model); the variance explained by the intersections of the pairs of models, and the variance explained by the intersection of all three models (See Fig. 3.1b for a graphical represention). To do so, we will use the results obtained from fitting individual models and combination of two and three models, and compare the of these nested models. We will then use set theory to calculate the common (as a set intersection) and unique (as a set difference) variance explained. Through our model fitting, using joint ridge regression, we directly obtain the variance explained (R^2) by the three models as well as the union of all pairs of models, and the union of all three models: $A, B, C, A \cup B, A \cup C, B \cup C$, and $A \cup B \cup C$.

$$A \approx \hat{A}$$

$$B \approx \hat{B}$$

$$C \approx \hat{C}$$

$$A \cup B \approx \widehat{A \cup B}$$

$$A \cup C \approx \widehat{A \cup C}$$

$$B \cup C \approx \widehat{B \cup C}$$

$$A \cup B \cup C \approx A \cup \widehat{B \cup C}$$

$\hat{A}$, $\hat{B}$ and $\hat{C}$ correspond to the variance explained calculated directly by fitting each of the individual models. $\widehat{A \cup B}$, $\widehat{A \cup C}$, $\widehat{B \cup C}$ and $A \cup \widehat{B \cup C}$ correspond to the variance explained calculated directly by fitting each pair of models and all three models simultaneously.

Using the above values obtained directly through model fitting, we can then calculate the remaining portions of *shared* variance explained. We know from set theory that the variance explained by the union of all three models is equal to the sum of the variance explained by each model, subtracting the intersection of each pair of models and the intersection of all three models combined:

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

We can use this equation, combined with the variance explained we already calculated by fitting models separately, to calculate the shared variances explained by the intersections of sets.

$$A \cup B \cup C = \hat{A} + \hat{B} + \hat{C} - A \cap B - A \cap C - B \cap C + A \cap B \cap C$$
$$A \cap B = \hat{A} + \hat{B} - \widehat{A \cup B}$$
$$A \cap C = \hat{A} + \hat{C} - \widehat{A \cup C}$$
$$B \cap C = \hat{B} + \hat{C} - \widehat{B \cup C}$$
$$A \cap B \cap C = \widehat{A \cup B \cup C} + \hat{A} + \hat{B} + \hat{C} - \widehat{A \cup B} - \widehat{A \cup C} - \widehat{B \cup C}$$

We can then calculate the variance explained by the intersections of two models that does not include the variance explained by the intersection of all three models (shown in Fig. 3.1b).

$$(A \cap B) \backslash C = \hat{A} + \hat{B} - \widehat{A \cup B} - A \cap B \cap C$$
$$(A \cap C) \backslash B = \hat{A} + \hat{C} - \widehat{A \cup C} - A \cap B \cap C$$
$$(B \cap C) \backslash A = \hat{B} + \hat{C} - \widehat{B \cup C} - A \cap B \cap C$$

Finally, we calculate the variance explained by one model, with no overlap of variance explained by any of the other models (or the Relative Complement of each pair of models. The relative complement of BC, or BC^{RC}, is the portion of the variance explained only by model A, also shown in Fig. 3.1b):

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

Note that we use the set notation because of its simplicity and its intuitive graphical representation of the results but that one can easily rewrite these quantities in terms of R^2 and sum of errors. For example, if SS_0 is used to represent the total sum of square errors (or the SS of a zeroth order model), then we have:

$$A = R^2_A = \frac{SS_0 - SS_A}{SS_0}$$
$$B = R^2_B = \frac{SS_0 - SS_B}{SS_0}$$
$$A \cup B = R^2_{A \cup B} = \frac{SS_0 - SS_{A \cup B}}{SS_0}$$

and thus:

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

3.3 Results

Uncorrelated predictions

Although the three feature spaces represent very different transformations of the speech stimuli, these features are also correlated with each other to various extents. In addition, even when feature dimensions are independent, they might be predictive of the same fraction of the BOLD signal; this is a possibility if correlations between features are generated by cortical associations. To address this issue, we designed a methodology to estimate the fraction of the variance in the response explained solely by each feature space, as well as the fraction explained by combinations of features.

For this purpose, we fitted models with all possible combinations of features: the 3 models based on a single feature space (spectral, articulatory, semantic); the three models based of pairs of features (spectral-articulatory, spectral-semantic, articulatory-semantic), and finally a single model that used all three feature spaces together (spectral, articulatory, semantic). As mentioned above, in order to partition the variance in nested models, we used the same regularization in the models with combined feature spaces as in the models with single feature spaces. Then using set theory, we were able to calculate the variance explained uniquely by each feature space as well as the variance explained by intersections of features (see Methods). We found that the independent features or combination of features that explained the variance in the response (Fig. 3.1) were: spectral only ($Sp \backslash A \cup Se$ in Fig. 3.1b and c) in early auditory cortex, semantic only in all areas outside of auditory cortex ($Se \backslash Sp \cup A$ in Fig. 3.1b and c), the combination of all three features in the auditory cortex and some small areas outside of auditory cortex, and, finally, the combination of the semantic and the articulatory features excluding the semantic features in the mid-level areas in auditory cortex. Note that the articulatory features ($A \backslash Sp \cup Se$ in Fig. 3.1b and c) only provide very little additional predictive power; in other words, the predictions of the articulatory features are almost completely redundant with those provided by spectral features or semantic features.

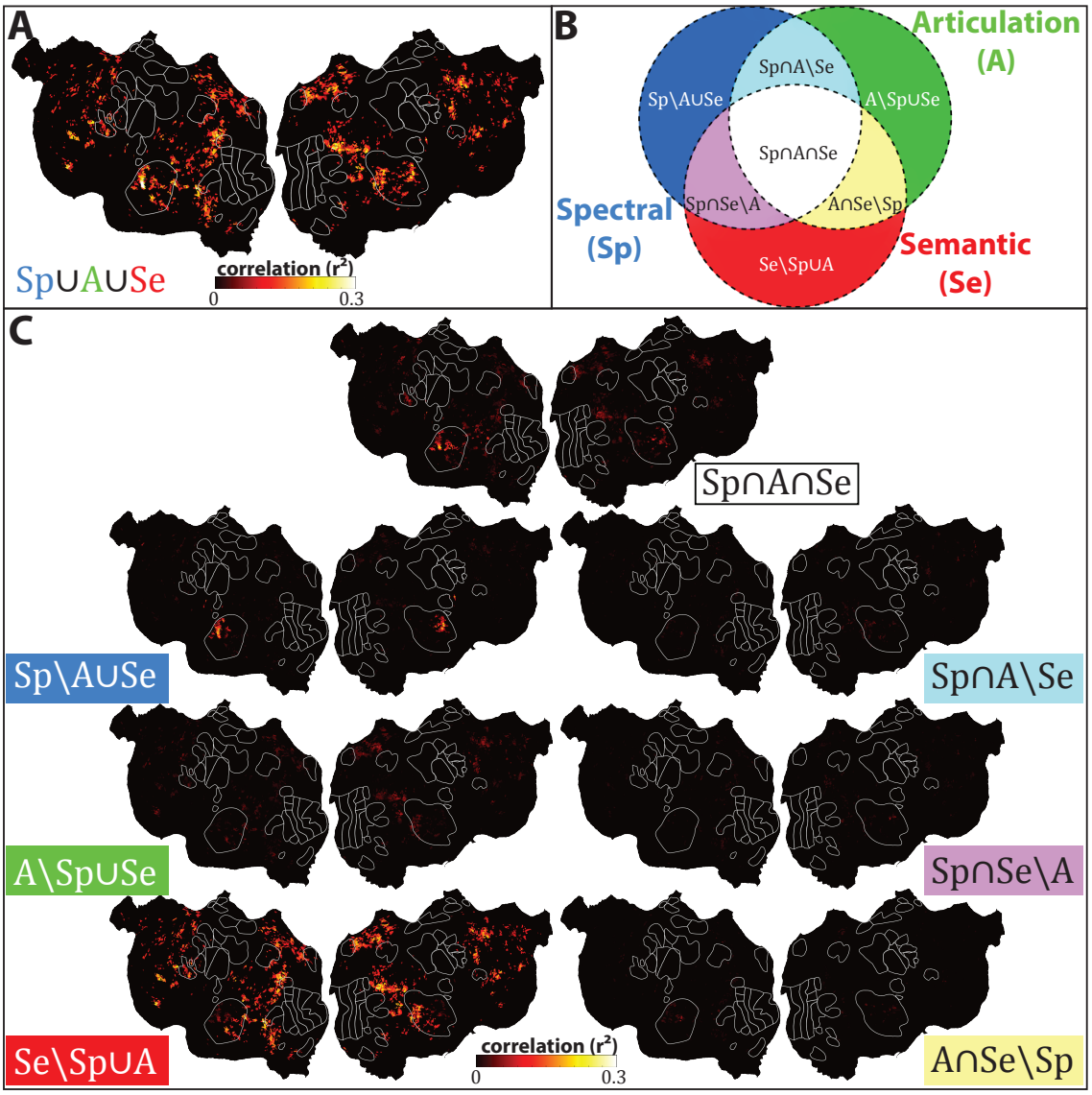

Figure 3.1: **A:** Flatmap of the correlations (R^2) between the actual and predicted BOLD activity from the union of all three models for subject S4. **B:** Cartoon Venn diagram showing variance explained by each individual part of the models and model intesections **C:** Variance explained by individual models and model intersections for subject S4 (see B as legend). Top: variance explained by the intersection of all three models. Left side, from top to bottom: variance explained by the spectral model alone; the articulatory model alone; and the semantic model alone. Right hand side, from top to bottom: variance explained by the intersection of the spectral and articulatory model (excluding the intersection of all three models); the intersection of the spectral and the semantic model (excluding the intersection of all three models); the intersection of the articulatory and the semantic model (excluding the intersection of all three models).

To further visualize the magnitude of these effects, we generated Venn diagrams showing the explained variance for these three feature spaces (Fig. 3.2, Fig. 3.3) calculated over the entire cortex (left), auditory cortex (middle) or Broca's area (right). Outside of auditory cortex, semantic features are most useful at predicting BOLD activity and the variance explained by the spectral and articulatory features largely overlaps the variance explained by the semantic feature. In contrast, the three features explain approximately a similar amount of variance in the auditory cortex with remarkably less overlap between the variance explained by the spectral and semantic features. The variance explained by the articulatory features is to a large extent also explained by either or both the spectral and semantic features; this feature space is clearly nested between the lower spectral representation and higher semantic representation.

3.4 Discussion

In Chapter 2, we established that our three features spaces (spectral, articulatory, and semantic) produce very different patterns of activity across the cortex, with the spectral model predicting activity in early auditory areas (Heschl's gyrus, A1), the articulatory model predicting activity in early auditory areas and some motor areas, and the semantic model predicting in a much wider portion of cortex, including higher auditory areas, parietal cortex, prefrontal cortex. In this Chapter, we found that the majority of the variance can be explained by the relative complement of the articulatory and semantics models (i.e. the variance explained only by the spectral model, with no overlap with any of the other models), the relative complement of the spectral and articulatory models (i.e. the variance explained only by the semantic model), the intersection between the articulatory and the semantic model, and the intersection between all three models. These results suggest that the cortical representations of different features of speech are organized into a partially overlapping hierarchy. Our results are also somewhat consistent with meta-analyses that have examined the cortical streams involved pre-lexical processing (Turkeltaub and Coslett, 2010, De Witt and Rauschecker, 2012; Bornkessel-Schlesewsky et al, 2015).

Following the theories presented in De Witt and Rauschecker (2012), pre-lexical computations require a first step involving the detection of complex spectro-temporal features performed initially by what have been called combination sensitive (CS) neurons. CS neurons have been described extensively in neurophysiological studies (e.g. Suga, O'Neill, & Manabe, 1978; Margoliash & Fortune, 1992; Rauschecker, Tian & Hauser, 1995). Cortical areas that have larger number of CS neurons would be found just ventrally from the primary auditory cortex (Tian, Reser, Durham, Kustov, & Rauschecker, 2001). CS is required for feature selection, for example to detect the presence of a particular phoneme; complex spectro-temporal receptive fields such as those observed in auditory cortex have been shown to be sufficient to reconstruct

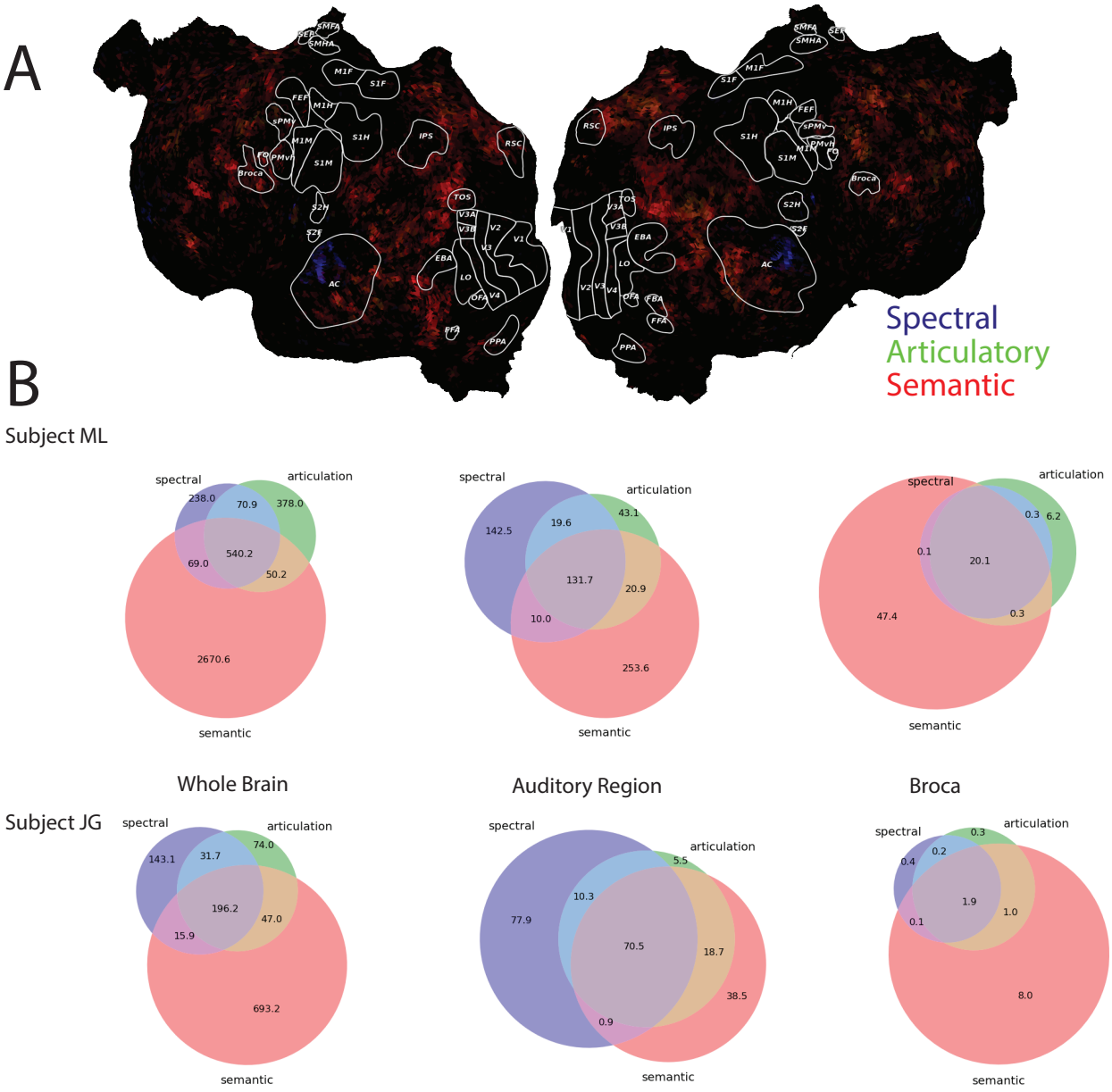

Figure 3.2: **A**: Flatmap of subject S4, with variance R^2 explained uniquely by each individual model (i.e. the variance that is not explained by any of the model intersections), indicated in blue for spectral, green for articulatory, and red for semantic. **B**: Venn diagrams of total explained variance R^2, (calculated using only significantly predicted voxels), for subject S4 (top) and subject S3 (bottom). The proportion of variance explained by each model and portion of model changes based on whether it is calculated over the entire cortex (left), auditory cortex (middle) or Broca's area (right), with the spectral model proportionally explaining more variance for auditory cortex than for other areas, and the semantic model proportionally explaining more variance for Broca's area, as well as the whole cortex than for Auditory Cortex.

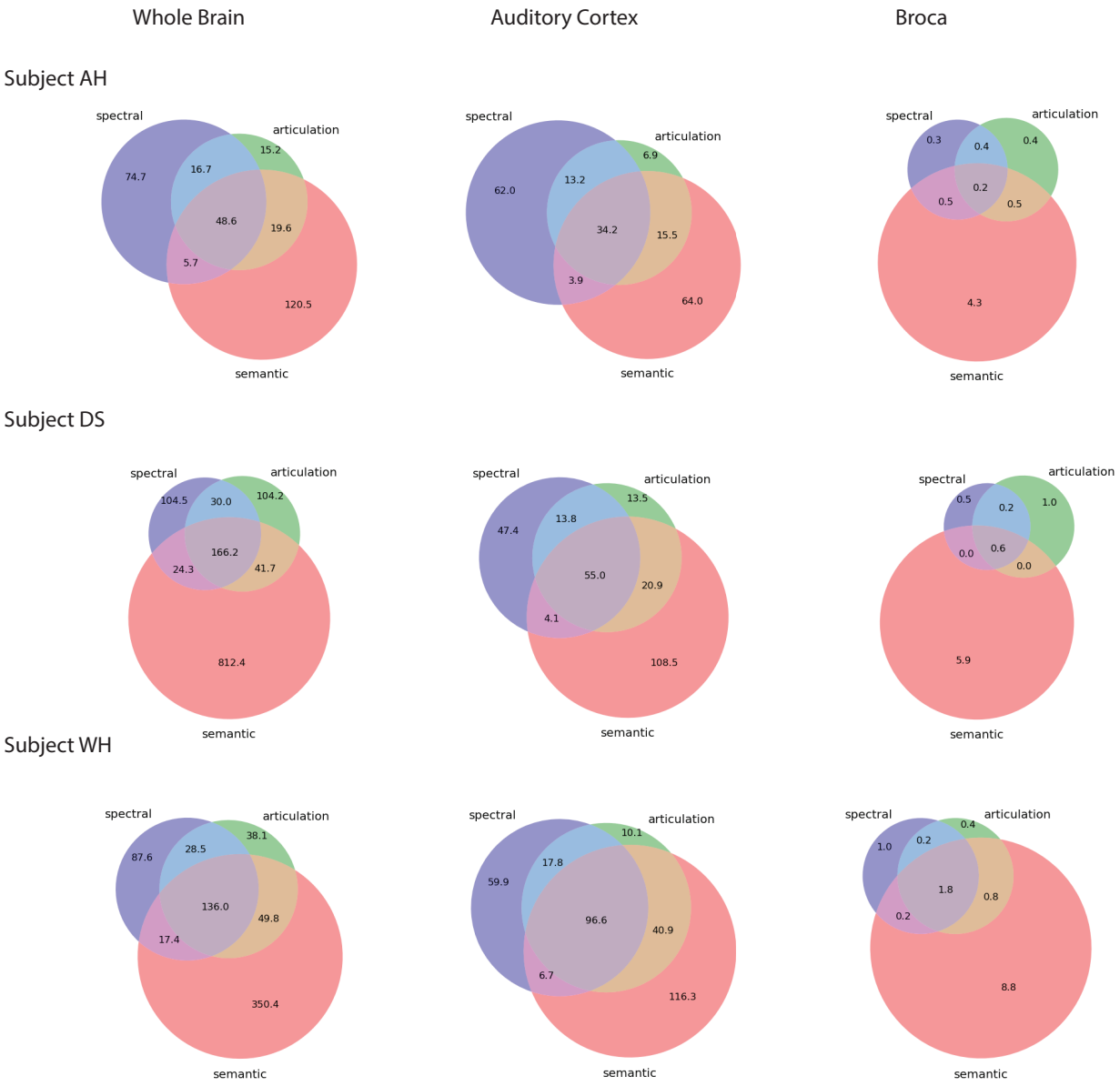

Figure 3.3: Venn diagrams of total explained variance R^2, (calculated using only significantly predicted voxels), for subject S1 (top), subject S2 (middle), and subject S5 (bottom). The proportion of variance explained by each model and portion of model changes based on whether it is calculated over the entire cortex (left), auditory cortex (middle) or Broca's area (right), with the spectral model proportionally explaining more variance for auditory cortex than for other areas, and the semantic model proportionally explaining more variance for Broca's area, as well as the whole cortex than for Auditory Cortex.

human speech (Pasley et al. 2012). This first step would then be followed by computations for categories yielding invariant representations (IR), such as similar responses for the same phoneme produced by different speakers or in different contexts. The IR cortical regions would be found on anterovental stream that has been implicated in word recognition (DeWitt & Rauschecker, 2012). Neurons with IR to communication calls have also been recently described in neurophysiological data (Elie & Theunissen, 2015). Our articulatory feature space combines both CS properties and IR properties since a given articulatory corresponds to a phoneme that is correlated in our stimuli with specific complex spectro-temporal features. At the same time, in our stories, the same phonemes (and same articulations) are produced by many speakers and in many different contexts. The fact that the variance in responses explained by the articulatory features overlaps almost completely with those explained by the spectral and semantic features is consistent with this idea: the overlap with the spectral features could correspond to CS regions and the overlap with the semantic features to IR regions. The fact that we also found a map in auditory cortex from primary regions along Heschl's gyrus with activity best explained by spectral features to ventral regions best explained first by articulatory and then by semantic features (Figs. 2.3 and 2.4) supports the idea of a mapping for this hierarchical processing stream. Future work using intermediate feature spaces between spectral and articulatory, such as the principal components or independent components of segments of speech spectrograms could be used to further distinguish CS areas from IR areas.

However, contrary to what has been reported in studies using segmented speech and artificial sounds, we found that the mapping from sounds to words occur very early on and potentially before the branching into the anteroventral and posterodorsal streams (Bornkessel-Schlesewsky et al., 2015; Turkeltaub & Coslett, 2010) that are not distinguished in our data and analysis. This discrepancy could be explained by a couple of factors both of which are a result of our use of natural speech. First, our semantic feature space combines both the abstraction from sound to word meaning that has been assigned to the anteroventral stream and the extraction of meaning derived from specific temporal sequences of words that has been assigned to the posterodorsal stream. This is the case because our semantic feature space represents the combination of word occurrences in a 2 s window and therefore also represents meaning assigned to sentences. Here again one could use additional intermediate or alternative feature spaces to more directly investigate these questions. For example one could compare a semantic representation that is sensitive to the order of the words in sentences to one that is insensitive to the meaning generated by sequences. Second, the use of segmented speech and artificial sounds in prior studies necessarily limited top-down modulation effects. Here the presence of semantic information found all the way in auditory cortex could also reflect more significant modulation that occurs when processing natural speech. It is interesting to note, that in our data, the acoustic features never outcompete the semantic features outside of auditory cortex while semantic features are not only present in much of the brain but also are best at

explaining responses in regions of the auditory cortex. Thus if there is intermingling of high-level and low-level processing, it is asymmetric with high-level processing being present in low-level areas but not the reverse suggestive of strong top-down effect. Strong top-down effects in auditory cortex as a result of attention and behavioral task relevance have clearly been demonstrated in both animal neurophysiological data (reviewed in Fritz, et al., 2007) and in human speech research (Wild et al. 2012; Peelle et al. 2013). These studies show that responses in the human auditory cortex are sensitive to the speech intelligibility. Since our subjects were clearly understanding and paying attention to the captivating stories, we hypothesize that a fraction of the predictive power of the semantic features in the auditory cortex is also a reflection of significant top-down effects.

One might be somewhat surprised at the lack of brain regions that are uniquely explained by the articulatory features. Although we also might not have expected such a complete overlap (see Fig. 3.1, Fig. 3.2), the fact that it overlaps so evenly with spectral features and semantic features and that the overlap is anatomically mapped in the auditory cortex also suggests that it was an appropriate intermediate representation. And, in future research, we might seek additional feature spaces that also share the property of strong overlap with lower and higher level features in an organized fashion. We note that we also found significant predictive power of the articulatory features (albeit overlapping with spectral and semantic features) in Broca's area and in motor mouth areas. Motor cortex response to speech has been inconsistently reported and is subject of debate in the literature (Mottonen, Dutton, & Watkins, 2013; Pulvermller, Huss, & Kherif, 2006).

The lateralization of the cortical streams involved in language processing is an active area of research with significant clinical implications. Although speech sounds are clearly processed bi-laterally in primary auditory cortex, even the relatively low pre-lexical processing steps involved in speech processing and localized to the temporal lobe could be lateralized with separate roles for the right and left hemispheres (e.g. Boemio, Fromm, Braun Poppel, 2005; Desai, Liebenthal, Waldron Binder 2008; Abrams, Nicol, Zecker Kraus, 2008). At higher levels, in brain regions principally involved in speech perception and production, the lateralization to the left hemisphere is well established (e.g. Pujol, Deus, Losilla, & Capdevila, 1999; Knecht et al., 2002). Here we found results that are not in line with these previous findings: we did not find any significant differences between the left and right hemispheres either in raw activation nor in our capacity to predict the BOLD activity using voxel wise modeling and three language related feature spaces! On one hand, it is very possible that lateralization effects as assessed by raw activation are much smaller when listening to engaging natural speech since that experience could also engage emotions, memories, visualization and other mental states that are not language specific in a strict sense. On the other hand, it was somewhat surprising and unexpected to discover that the semantic feature space yields similar predictions in both hemispheres. Lateralization might however still be present and one could imagine that the activity predicted by

our semantic feature space overlaps differentially with other stimulus features in the right and left hemispheres. Here again this could be further investigated by testing models based on other feature spaces.

Another striking result of this analysis is both the large extent and the predictive power of the semantic feature space. As mentioned above, our semantic feature space does include representation of abstraction both at the word level and the sentence level and it can therefore serve as predictive power of many different levels of processing that combines lexical, syntactic and even emotional responses. In a separate study, we have analyzed the functional mapping of semantic domains (i.e. areas with related concepts) to further parcel the large area explained by the semantic model; we found multiple functional maps of domain selectivity (Huth et al. submitted). It is possible that these separate functional semantic maps also correspond to separate streams involved in processing meaning from different information bearing structures found in human language (Bornkessel-Schlesewsky et al., 2015).

In Chapter 4, we will discuss a promising intermediate feature space that could help bridge the gap between lower and higher level models.

Chapter 4

Modulation power spectrum: comprehension of modulation filtered phonemes

4.1 Introduction

In Chapters 2 and 3, our lowest-level model, the spectral model, was a straightforward model that indicated which frequencies were present at a given point in time. Choosing such a simple representation is reasonable: tonotopy, or the systematic representation of frequencies in specific topographical areas, is a known organizing principal of the auditory system. Tonotopy is found as early in the auditory periphery as the cochlea, and continues to be present through both subcortical structures and the auditory cortex of humans and animals (Humphries, Liebenthal, & Binder, 2010; Weisz, Wienbruch, Hoffmeister & Elbert, 2004). However, this frequency representation is only one of many cortical representations that lead to the comprehension of speech. Akin to vision, which represents visual stimuli through increasingly complex transformations and representations from lower-level to higher-level visual areas (Grill-Spector & Malach, 2004), in the auditory realm there are increasingly complex cortical transformations and representations of the speech signal as it is processed in higher auditory areas. One likely candidate representation is the modulation power spectrum (MPS): the MPS is the space of spectral and temporal modulations, or, mathematically, the amplitude spectrum of the 2-dimensional Fourier transform of the speech spectrogram (Fig. 4.1) (Chi et al., 1999, Chi, Ru, & Shamma, 2005, Greenberg 1997, Elliott & Theunissen 2009). Speech is a complex signal: the power of the speech signal fluctuates both in the spectral and in the temporal domain (Fig. 4.1). Spectrotemporally complex signals such as upsweeps or downsweeps (which allow us, for example, to distinguish between /ba/ and /da/) are ubiquitous in speech. The MPS is therefore a very interesting mid-level representation to explore; both to

understand which specific spectrotemporal modulations are crucial for speech comprehension, and to determine where in the cortex these spectrotemporal modulations are represented.

Studies in recent years have indicated that there are neurons specifically tuned to modulations of sound, which have been found in subcortical auditory structures and in primary auditory cortex. fMRI studies have revealed voxels tuned to spectral and temporal modulations in human auditory cortex (Langers, Backes, & van Dijk, 2003; Schnwiesner & Zatorre, 2009; Santoro et al, 2014), and the spectrotemporal modulation space is represented in both subcortical structures and in auditory cortex in animals (Joris, Schreiner, & Rees, 2004; Mesgarani, David, Fritz, & Shamma, 2008).

In this Chapter, we investigate which spectrotemporal modulations are necessary for the comprehension of vowels and consonants. It is well known that many types of degraded speech can still be understood: subjects can comprehend sentences even with severe temporal or spectral degradation (Shannon et al., 1995). Given that only a certain portion of modulation space is used for animal and human communication (Singh & Theunissen, 2003), by definition only certain subset of the modulation power spectrum must be necessary for speech comprehension. We should therefore be able to target specific modulations that are necessary for speech comprehension. In previous work, Elliott and Theunissen (2009) have shown that filtering temporal and spectral modulations affects speech intelligibility in full sentences. Comprehension dropped off significantly when spectral modulations were filtered below 4 cycles/kHz or when temporal modulations were filtered between 1 and 7 Hz. This defines a core area of the modulation power spectrum necessary for sentence comprehension. However, context could allow subjects to fill in the blanks in a paradigm where they are hearing full sentences. In this study, we further examined the importance of specific spectrotemporal features in the speech signal by asking subjects to recognize filtered phonemes instead of sentences. Subjects were asked to recognize phonemes in a 12 (vowel) or 20 (consonant) alternative forced-choice paradigm. Subjects were presented filtered phonemes in /aCa/ (to test consonant recognition) or /hVd/ (to test vowel recognition) context. All of the filters were in the previously identified core spectrotemporal modulations for speech comprehension (Elliott and Theunissen, 2009): temporal high and low pass filters had cutoffs of 0.25, 0.75, 1.75, 3.75 and 7.75 Hz, and spectral high and low pass filters had cutoffs of 0.25, 0.75, 1.75 and 3.75 cyc/kHz.

In future work, we intend to further investigate the cortical representation of the modulation power spectrum in human subjects listening to naturally spoken stories, using the voxel-wise modeling fMRI paradigm described in Chapter 2.

4.2 Materials and Methods

Subjects

13 native American English subjects with no reported hearing loss (7 female, 6 male; mean age: 20.9; age range 18-23) were recruited via U.C. Berkeley's Research Participation Program for undergraduates. Participants were given course credit for participation. The use of human subjects in this study was approved by the UC Berkeley Committee for Protection of Human Subjects.

Speech test materials

Consonants were from the Shannon et al. (1999) stimulus set, and consisted of 20 syllables presented in the /aCa/ context: ba, cha, ba, cha, da, fa, ga, ja, ka, la, ma, na, pa, ra, sa, sha, ta, tha, va, wa, ya, za. Two male speakers (#2 and #5) and two female speakers (#1 and #3) were chosen. Vowels were from the Hillenbrand et al. (1995) stimulus set, and consisted of 12 vowels presented in the /hVd/ context: had, hawed, hayed, head, heard, heed, hid, hod, hoed, hood, hud, who'd. Two male speakers (24 and 33) and two female speakers (#15 and #26) were chosen.

Signal processing

Phonemes were filtered with custom Matlab software, using the same filtering algorithm described in Elliott and Theunissen (2009). In brief, the modulation power spectrum is the amplitude spectrum of the 2 dimensional Fourier Transform of a time-frequency representation of a sound-pressure waveform (Fig. 4.1), typically a spectrogram. The time-frequency representation that we used for our analysis is the log amplitude of a spectrogram obtained with Gaussian windows. The time-frequency scale we used to obtain the spectrograms of our sound stimuli were of 10 ms in the time domain (1000/(2p*10) 16 Hz in the frequency domain). This allows us to represent modulations of up to 50 Hz in the temporal domain and 31 cycles/kHz in the frequency domain. Temporal high pass and temporal low pass modulation filter cutoffs were chosen within the core spectral and temporal modulations also defined in Elliott and Theunissen (2009): 0.25, 0.75, 1.75, 3.75 and 7.75 Hertz. Spectral high pass and spectral low pass modulation filter cutoffs were: 0.25, 0.75, 1.75 and 3.75 cycles per kHz.

Stimulus presentation

Normalized stimuli were presented in silence at 70dB SPL over Sennheiser HD250 headphones in a soundproof booth, using custom Matlab software. The stimuli were presented in blocks of either consonants or vowels. The blocks were counterbalanced:

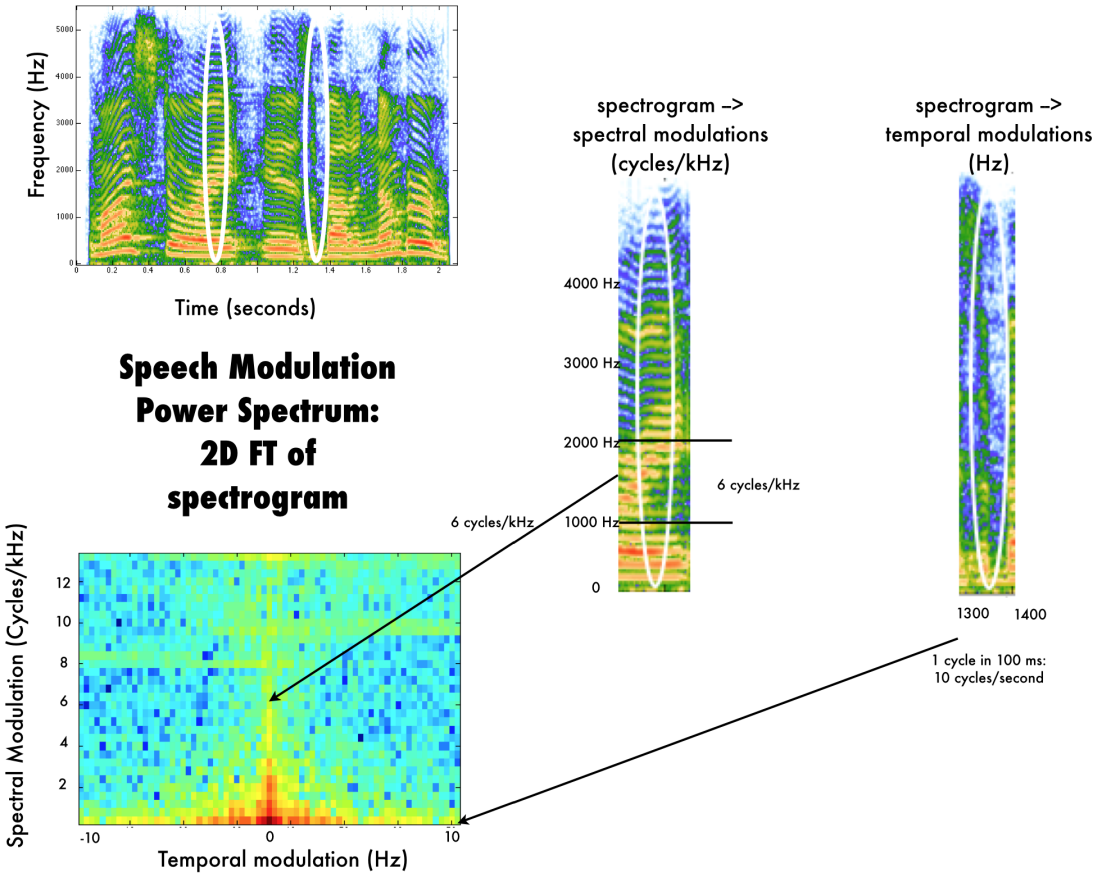

Figure 4.1: The modulation power spectrum (MPS, bottom) is the 2-dimensional Fourier Transform of a spectrogram (top). The x-axis of the MPS represents temporal modulations of the speech signal, in units of cycles per second, or Hertz (Hz). The y-axis of the MPS represents spectral modulations of the speech signal, in units of cycles per kilohertz (kHz). The first white ellipse shows a portion of the speech signal that contains almost uniquely spectral modulations (at the rate of 6 cycles/kHz). The second white ellipse shows a portion of the speech signal that contains almost uniquely temporal modulations (at the rate of 1 cycle per 100 ms, or 10 Hz)

7 subjects started with the vowel block and 6 subjects started with the consonant block. Each phoneme was presented twice for each filtered condition. Four speakers (2 males and 2 females) were chosen for the vowels and four speakers (2 males and 2 females) were chosen for the consonants. Each of the four speakers was presented an equal number of times per filtered condition. Within each block all stimuli were presented pseudo-randomly, with two repeats of each stimulus. Participants were briefly trained before being tested on the filtered sound. During their training, they were presented unfiltered phonemes in the same alternative forced choice paradigm as in the testing period (the training phonemes were spoken by different speakers than the testing phonemes). Participants would then only be tested if they obtained a perfect score in the training session. Additionally, during the testing period randomly interspersed unfiltered trials were used to verify that subjects were concentrating on the stimuli and able to perform the task.

The participants went at their own pace, with no time constraints. They listened to the sound stimulus once, then touched the screen (the experiment was presented to participants on a touchpad) corresponding to their choice of one of 12 vowel or 20 consonant options. Custom MATLAB software automatically recorded the presented stimulus and the participants' answers

Normalization of results

In order to be able to accurately compare the 20-alternative forced choice consonant condition and the 12-alternative forced choice vowel condition, we normalized the results. Both consonant and vowel data are binomial distributions, with 1/12 chance probability of being correct for vowels and 1/20 chance probability of being correct for consonants. For each filtered condition, we normalized the percent correct as follows: $\frac{(x-(n*p))}{\sqrt{n*p*(1-p)}}$ where x is the number of correct answers, n is the number of trials, and p is the probability of success per trial. The correction is corresponds removes the expected mean, and divides by the population standard deviation of the binomial distribution yielding a measure akin to a z-score. In order to be able to plot both the vowels and the consonants on the same graph, we divided this z-score by the maximum z-score value that could be obtained for 100% correct answers for the consonant test (z-max = 27.56) and for vowels (z-max = 16.53).

4.3 Results

The most important spectrotemporal modulations for phoneme comprehension were assessed psychoacoustically using two types of spectral modulation and temporal modulation filters: low-pass and high-pass. The filter range was based on previous psychoacoustical research (Elliott & Theunissen, 2009), which designated a core of modulations most important to the comprehension of sentences: modulations below

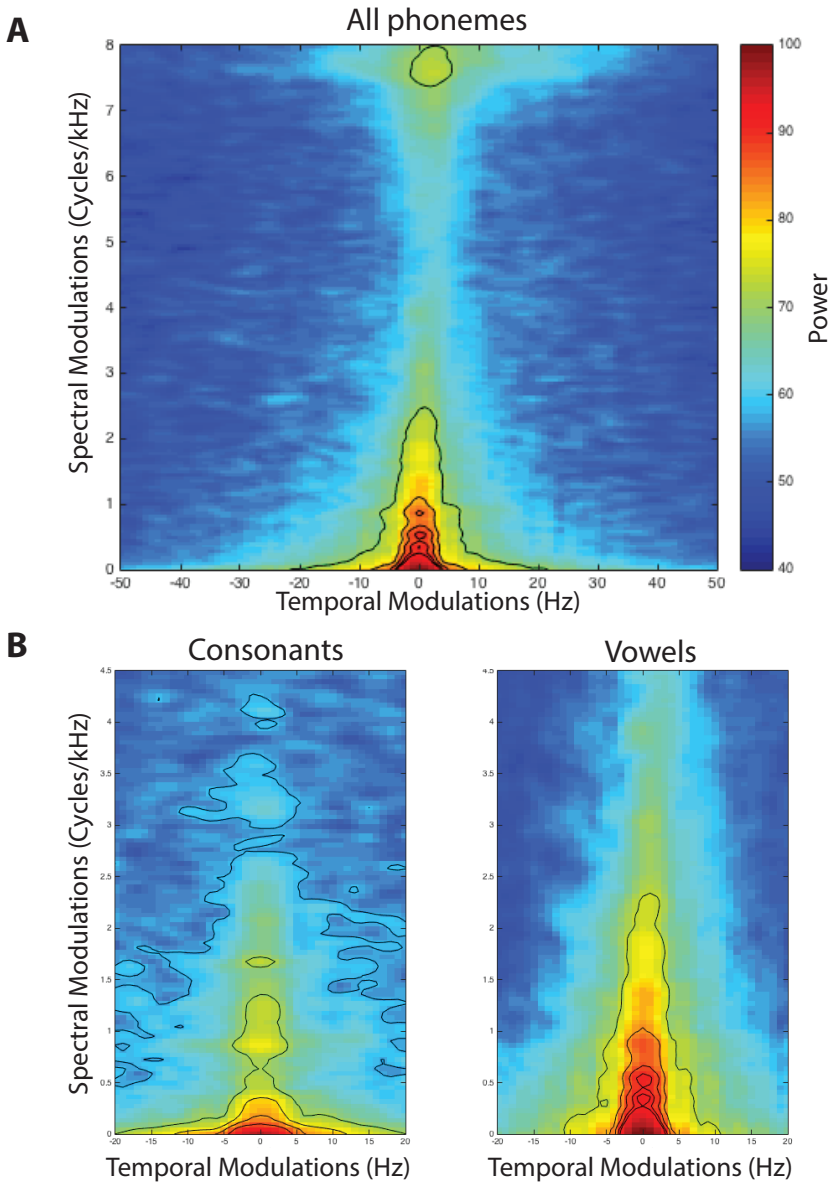

Figure 4.2: A: Modulation power spectra averaged across all phonemes (vowels and consonants), for one male speaker. Temporal modulations in Hz are represented on the x-axis. Spectral modulations in cycles/kHz are represented on the y-axis. Power is represented in color. **B:** Modulation power spectra averaged across consonants (left) and vowels (right) for one male speaker.

7 Hz in the temporal domain, and below 4 cycles/khz in the spectral domain. We first turn to the modulation power spectrum representation of consonants and vowels, to assess similarities and differences between the modulation power spectra of these two classes of phonemes.

Modulation power spectra of consonants and vowels

The average modulation power spectrum of phonemes (Fig. 4.2a) shows that the majority of spectrotemporal acoustic power lies in low spectrotemporal modulations (between -20 and 20 Hz temporally, and between 0-5 cycles/kHz spectrally)–with a notable exception of power in high spectral modulations (here, around 8 cycles/kHz), which corresponds to the pitch of the speaker's voice at approximately 125 Hz (see Elliott and Theunissen, 2009). Indeed, one of the advantages of the MPS representation is that it allows one to separate spectrotemporal energy solely attributable to the harmonic structure of voiced speech from spectrotemporal energy of the unvoiced sounds emphasizing the contribution of the filter component in the source-filter model of speech production. The average modulation power spectrum of consonants is very different than the average modulation power spectrum of vowels (Fig 2b). The average consonant MPS (Fig 4.2b, left) has power across a wide range of temporal modulations (between -20 and 20 Hz), and a smaller range of spectral modulations (between 0-2.5 cycles/kHz), whereas the average vowel MPS (Fig 4.2b, right) contains a band of focused power between -10 and 10 Hz, but spans much further in spectral modulations, up to at least 4.5 cycles/kHz.

Low-pass vowel and consonant filtering

Low-pass spectral modulation filtering

The spectral low-pass filtering results (Fig. 4.3a) show that both consonants (in green) and vowels (in blue) are impacted by low-pass spectral filtering. It appears that consonants are less affected by spectral low-pass filtering than vowels. Let us note, however, that subjects generally have more difficulty with vowel recognition, as can be seen in the unfiltered conditions: even in the unfiltered condition, the consonants outperform the vowels. For both consonants and vowels, comprehension is not greatly impacted at the two highest cutoffs (3.75 cycles/kHz and 1.75 cycles/kHz). However, there is then a steep decrease in intelligibility at 0.75 cycles/kHz and 0.25 cycles/kHz. The vowels have a steeper dropoff in intelligibility than the consonants as the low-pass filter cutoff decreases.

Low-pass temporal modulation filtering

In the low-pass temporal modulation filtering condition vowels outperform consonants considerably for the three most filtered conditions: 0.25 Hz, 0.75 Hz, and 1.75 Hz (Fig.

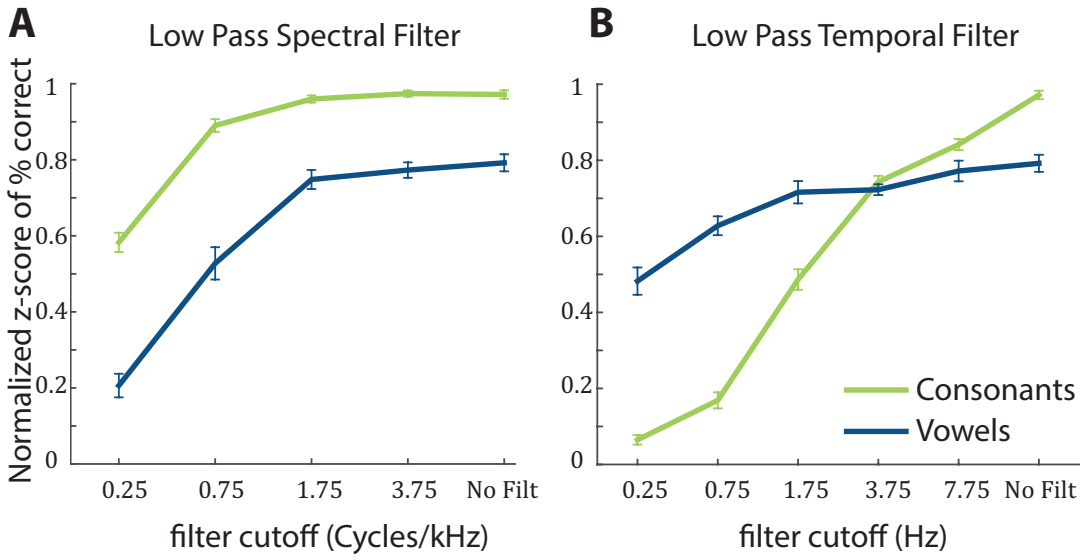

Figure 4.3: Normalized z-score of percentage of phonemes correctly identified for low-pass filtered vowels and consonants. Filter cutoffs are indicated on the x-axis. Error bars indicate standard error. **A** shows correct answers (normalized z-score) for spectrally low-pass filtered consonants (green) and vowels (blue), for all spectral filters and for no filter. **B** shows correct answers (normalized z-score) for temporally low-pass filtered consonants (green) and vowels (blue), for all temporal filters and for no filter.

4.3b). Even at 0.25 Hz, the most extreme filter we chose, there is still substantial spectral intelligibility for vowels, but no intelligibility for consonants. By 3.75 Hz the vowel intelligibility plateaus, and consonants and vowels are equally intelligible. Consonants outperform vowels again in the least filtered condition (7.75 Hz), and the unfiltered condition.

High-pass vowel and consonant filtering

High-pass spectral modulation filtering

Both consonants and vowels are similarly affected by the spectral high-pass condition (Fig. 4.4a): performance ranges from almost perfect (in the least filtered condition) to almost unintelligible (in the most filtered condition). Consonants and vowels perform equally at the most filtered condition (3.75 cycles/kHz), and the least filtered condition (0.25 cycles/kHz). Consonants outperform vowels at 0.75 cycles/kHz, and vowels outperform consonants at 1.75 cycles/kHz.

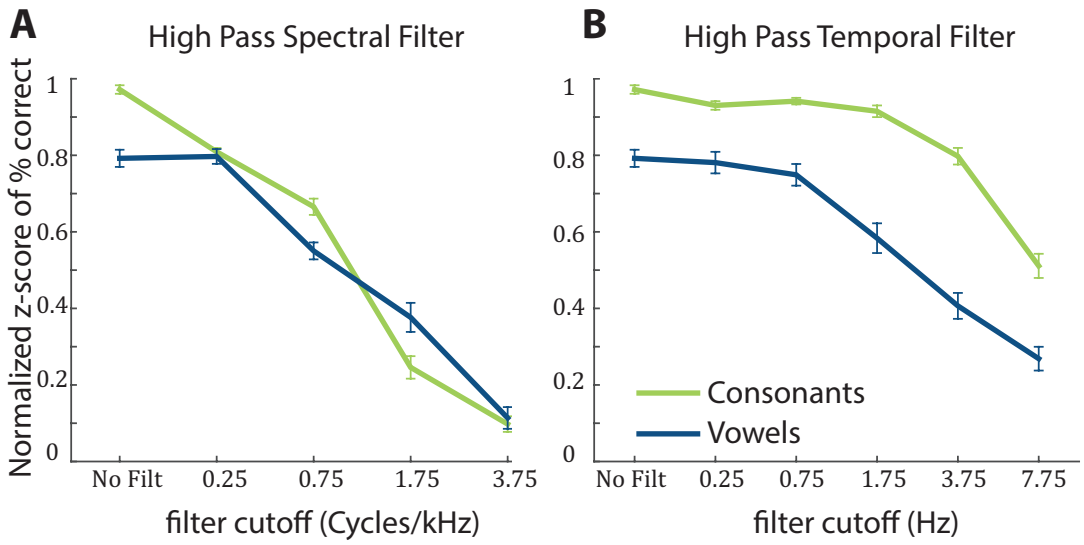

Figure 4.4: Normalized z-score of percentage of phonemes correctly identified for high-pass filtered vowels and consonants. Filter cutoffs are indicated on the x-axis. Error bars indicate standard error. **A** shows correct answers (normalized z-score) for spectrally high-pass filtered consonants (green) and vowels (blue), for all spectral filters and for no filter. **B** shows correct answers (normalized z-score) for temporally high-pass filtered consonants (green) and vowels (blue), for all temporal filters and for no filter.

High-pass temporal modulation filtering

Similar to the low-pass spectral modulation filtering, consonants consistently outperform vowels in the high-pass temporal modulation filtering (Fig. 4.4b). Consonant performance is almost unaffected by high-pass filtering until 1.75 Hz, after which there is a steep decline in intelligibility for 3.75 and 7.75 Hz. Vowel performance, however, is already affected at the least filtered condition (0.25 Hz), continues to decrease at 0.75 Hz, and declines significantly as the signal is more heavily filtered for the last three filters (1.75, 3.75 and 7.75 Hz).

Modulation Power Spectra of individual consonants and classes of consonants

Plotting the individual modulation power spectra of consonants confirms that there is a great amount of variability in the MPS of consonants (Fig. 4.5a, unlike vowels, that have quite consistent MPS (Fig. 4.6)). Consonants can be categorized into three linguistic classes based on the manner in which they are produced: sonorants (L, M, N, R, W, Y), stops (B, D, G, K, P, T), and fricatives (CH, F, J, S, SH, TH, V, Z). If the modulation power spectra of the three different classes of consonants are averaged

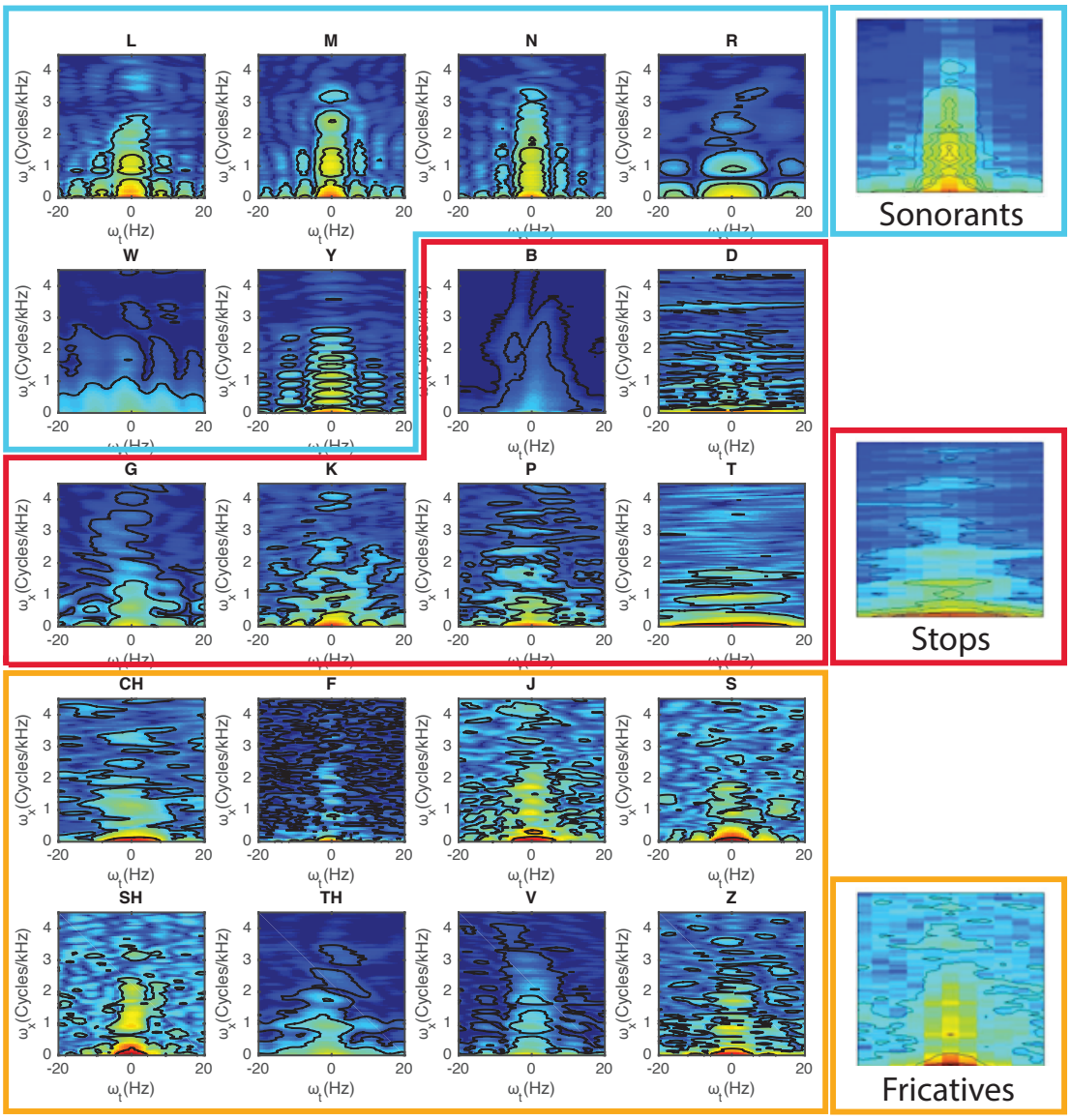

Figure 4.5: **A**: Modulation power spectra for individual consonants, spoken by a single male speaker. Consonants can be categorized as sonorants (indicated in blue), stops (indicated in green), and fricatives (indicated in red). **B**: Average modulation power spectra for all sonorants (blue), stops (green), and fricatives (red).

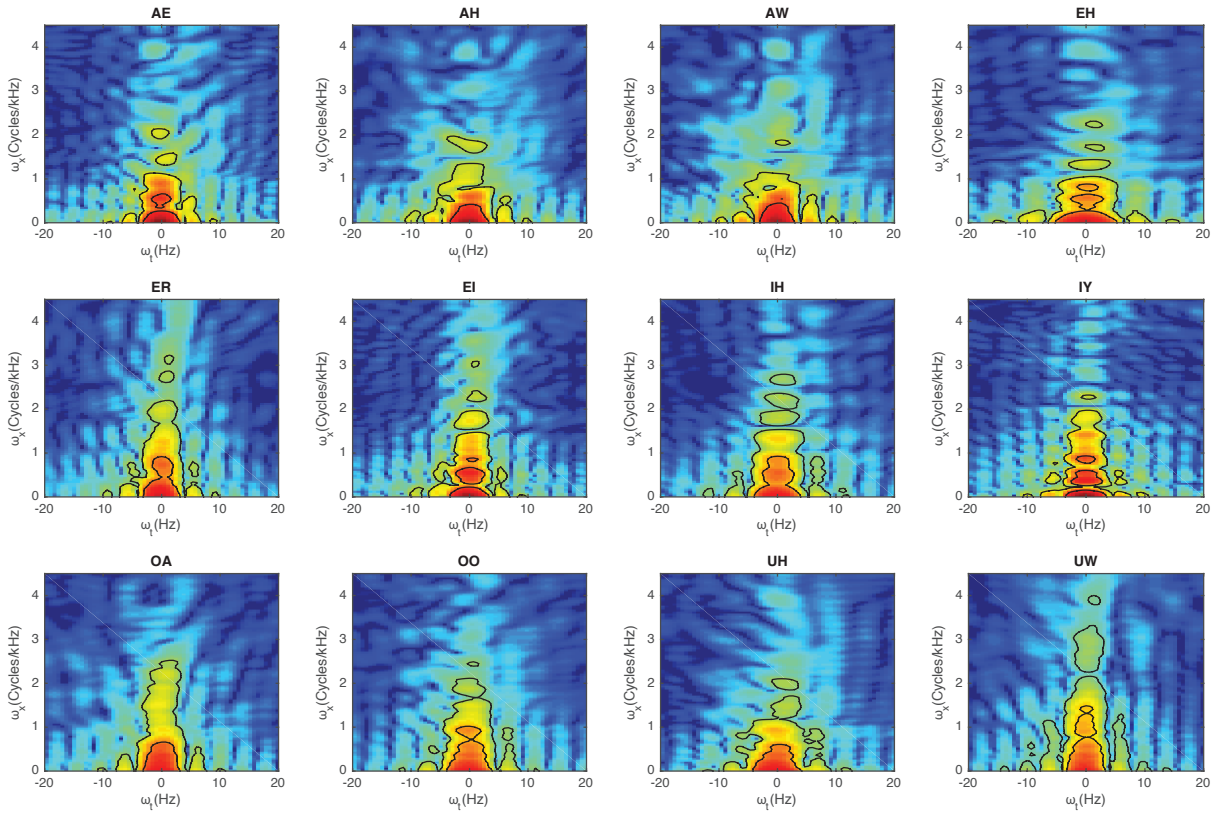

Figure 4.6: Modulation power spectra for individual vowels, spoken by a single male speaker.

(Fig. 4.5b), clear patterns emerge. The MPS of sonorants look far more 'vowel-like' than the stops or the fricatives, with greater power in higher spectral modulations, and less power in higher temporal modulations. Stops, however, contain power almost exclusively in low spectral modulations, across all temporal modulations. Finally, fricatives strike a balance between the other two classes: they contain power at higher spectral modulations, but also have power across lower spectral modulations and high temporal modulations. Fricatives also contain a large amount of power in low temporal and spectral modulations.

Differentiating between consonant classes

Given the difference in MPS between the three classes of consonants (sonorants, stops, and fricatives), it is likely that the spectrotemporal modulation filtering affects the intelligibility of the three classes of consonants differently. We therefore revisit our phoneme comprehension results, this time focusing uniquely on these three classes of phonemes.

Low-pass sonorant, stop and fricative consonant filtering

Low-pass spectral modulation filtering

For the two least-filtered conditions (3.75 cycles/kHz and 1.75 cycles/kHz), sonorants, stops and fricatives perform equally well, and are at ceiling performance (Fig. 4.7a). Performance starts to drop off at 0.75 cycles/kHz, but all three types of consonants still perform similarly. At 0.25 cycles/kHz, however, there is a clear demarcation between the three types of consonants: fricatives perform best, then stops, and finally sonorants. Note that even the worst-performing consonant class (sonorants) outperforms vowels in this condition.

Low-pass temporal modulation filtering

All three consonant classes perform equally well in the no filter condition (Fig. 4.7b), however there is a clear hierarchy for the other conditions, with sonorants this time consistently performing best of the three classes, then fricatives, and finally stops. Although the sonorants always outperform the fricatives, these two classes of consonants are almost at ceiling performance for 7.75 and 3.75 Hz, and are relatively close to one another for the other three filter cutoffs, except for the most filtered condition (0.25 Hz), where both fricatives and stops are at floor performance. The low-pass temporal filter significantly affects the intelligibility of stops for all filter cutoffs, and except for the aforementioned most filtered condition, stops perform significantly worse than the other two conditions.

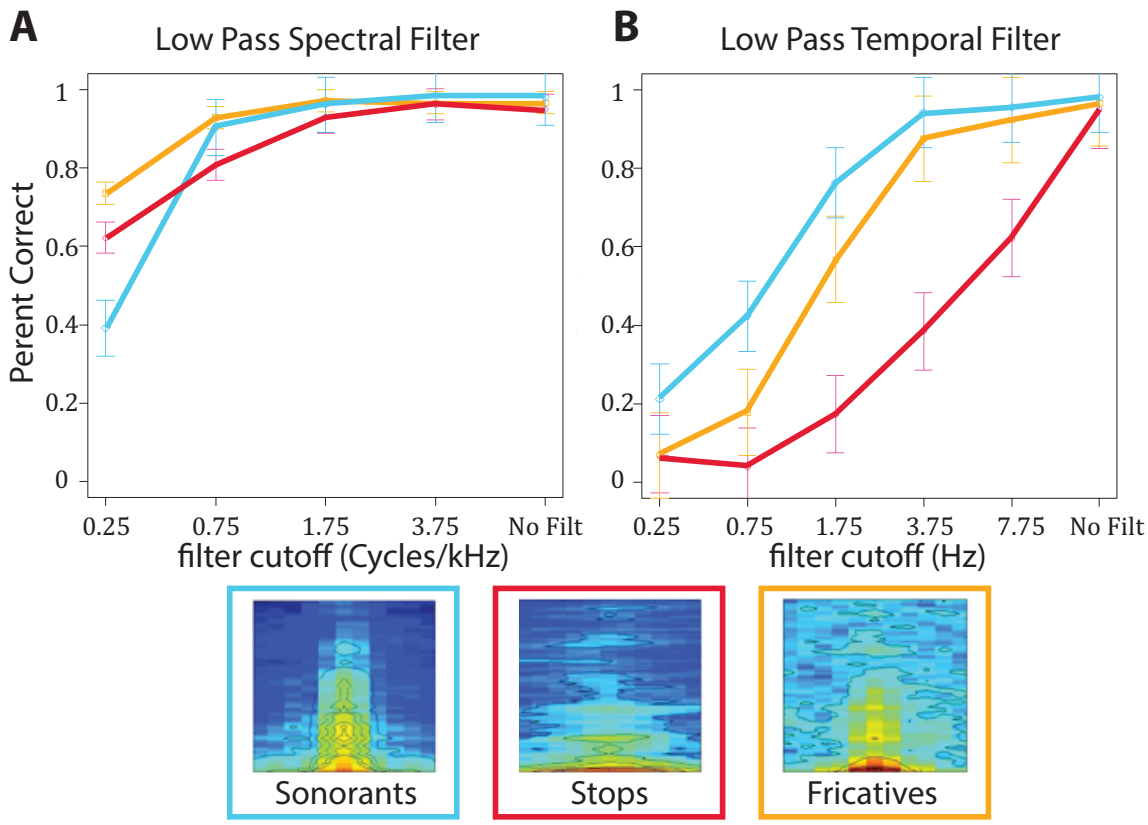

Figure 4.7: Percent phonemes correctly identified for low-pass filtered consonants. Filter cutoffs are indicated on the x-axis. Error bars indicate standard deviation. **A** shows percent correct answers for spectrally low-pass filtered sonorants (cyan), stops (red), and fricatives (orange) for all spectral filters and for no filter. **B** shows percent correct answers for temporally low-pass filtered sonorants (cyan), stops (red), and fricatives (orange) for all temporal filters and for no filter

High-pass sonorant, stop and fricative consonant filtering

High-pass spectral modulation filtering

The three consonant classes are ordered in the high-pass spectral modulation filtering condition (Fig. 4.8a), with sonorants always performing best, then stops, and finally fricatives, except for the no-filter and most filtered cutoff condition, where performance is equal for all three classes. Decline in performance is slower until after 3.75 cycles/kHz. There is a sharp dropoff in performance for all three classes of consonants between 0.75 and 1.75 cycles/kHz.

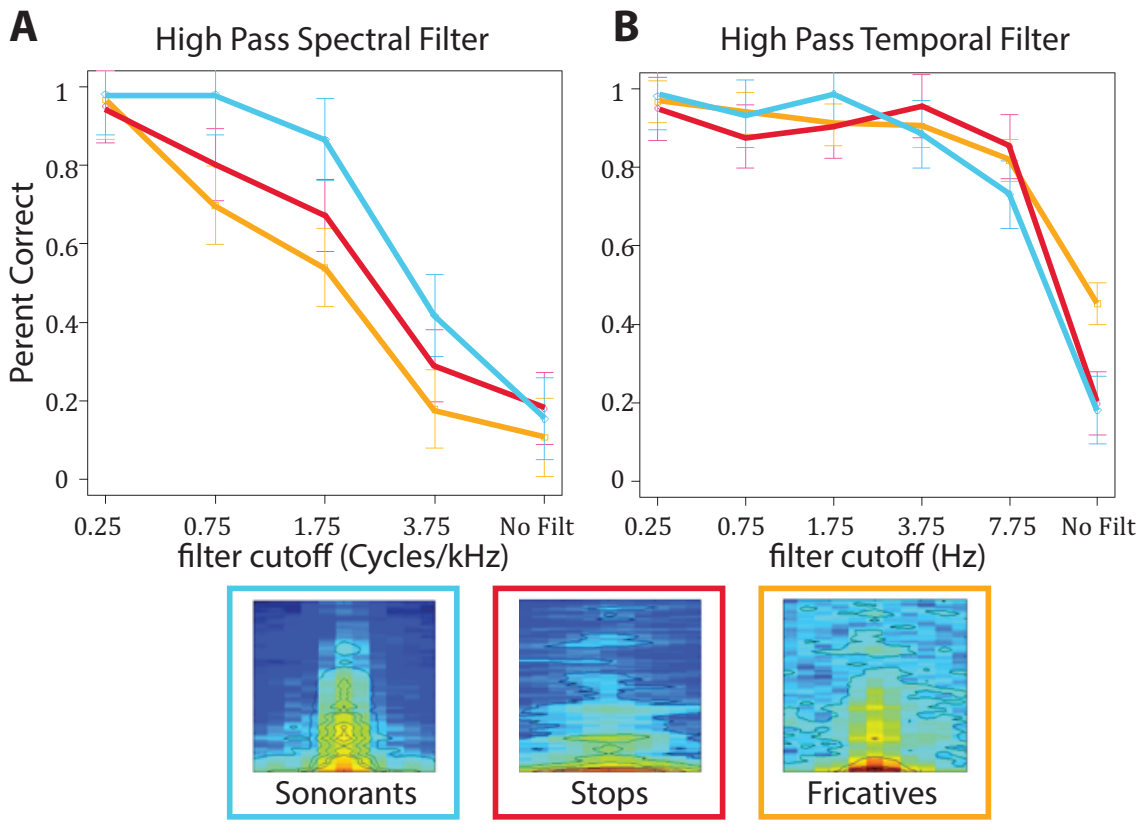

Figure 4.8: Percent phonemes correctly identified for high-pass filtered consonants. Filter cutoffs are indicated on the x-axis. Error bars indicate standard deviation. **A** shows percent correct answers for spectrally low-pass filtered sonorants (cyan), stops (red), and fricatives (orange) for all spectral filters and for no filter. **B** shows percent correct answers for temporally low-pass filtered sonorants (cyan), stops (red), and fricatives (orange) for all temporal filters and for no filter

High-pass temporal modulation filtering

The three consonant classes have almost indistinguishable performance for high pass temporal filter condition (Fig. 4.8b), except for the most filtered condition (7.25 Hz). The three classes are almost at ceiling until 1.75 Hz. There is a small decrease in performance between 1.75 Hz and 3.75 Hz, and then a very steep decrease in performance between 3.75 and 7.75 Hz. At the 7.75 Hz cutoff, the stops and sonorants classes are the worst-performers. The fricative class, although also affected, performs significantly better at this cutoff than the other two.

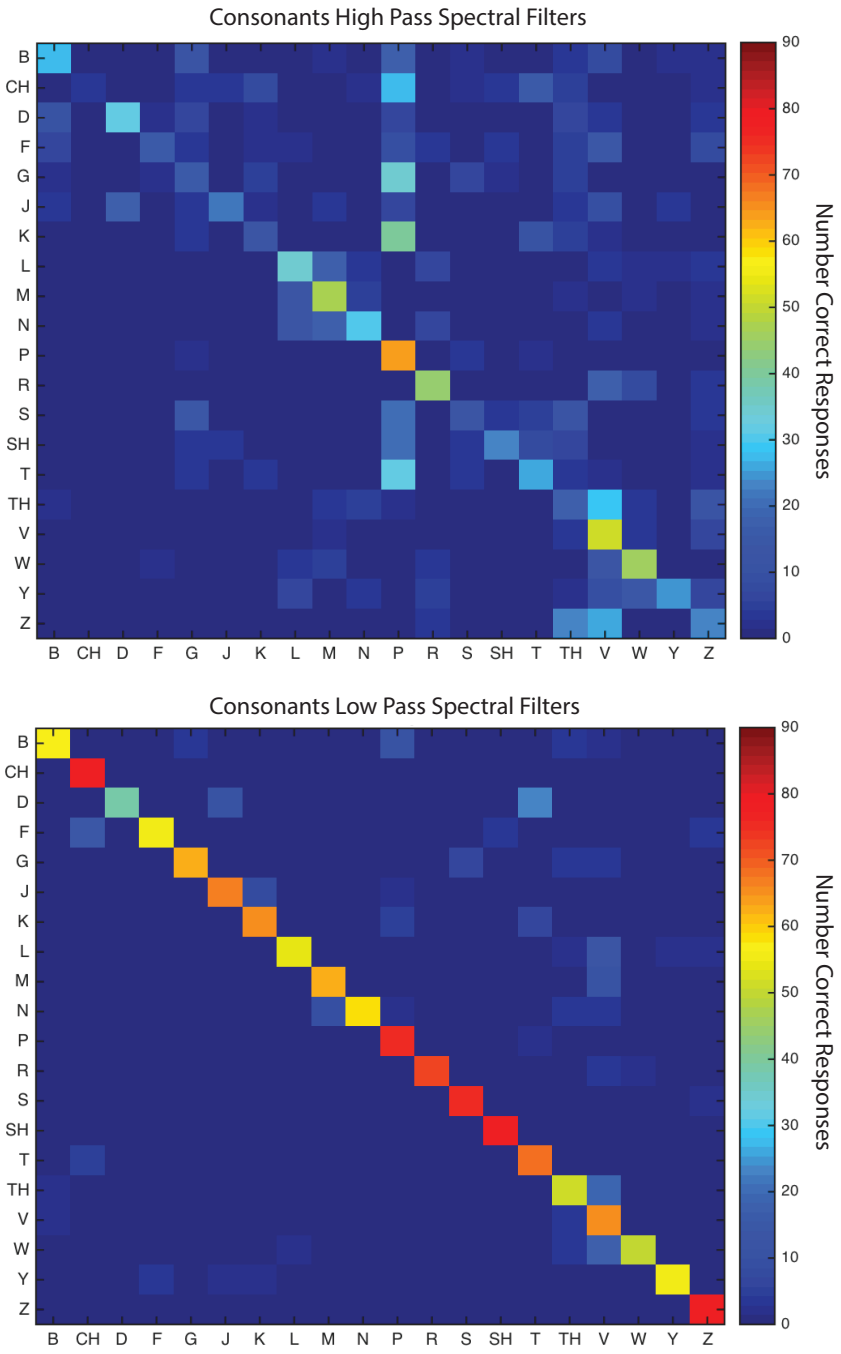

Figure 4.9: Consonant confusion matrix for the three narrowest spectral modulation filters, summed across all subjects. Rows represent stimulus presented, columns represent recorded response, and colors show the number of correct answers. The top matrix is the average confusion matrix for three high-pass spectral modulation filtering conditions (0.75, 1.75 and 3.75 cycles/kHz). The bottom matrix is the average confusion matrix for three low pass spectral modulation filtering conditions (0.25, 1.75 and 3.75 cycles/kHz).

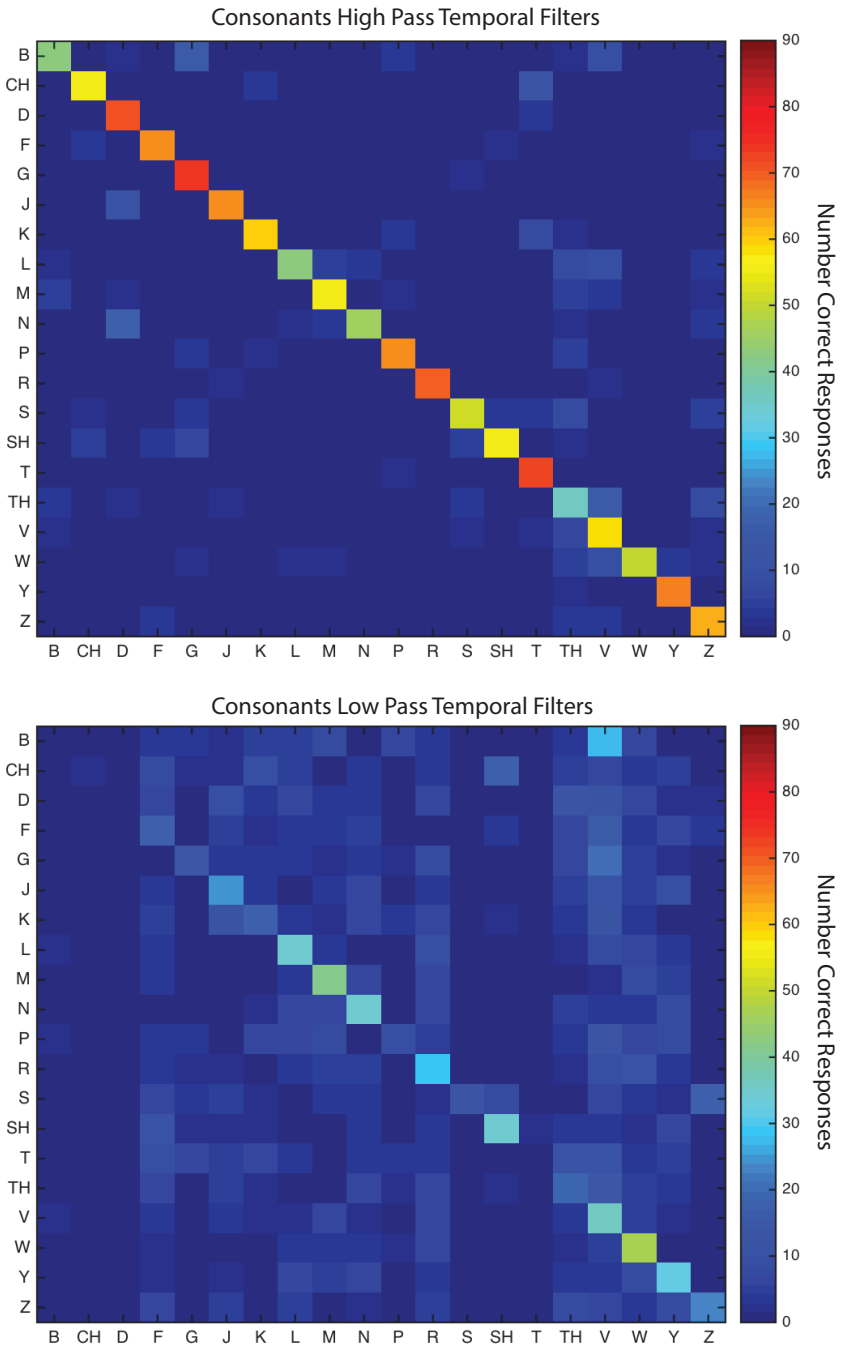

Figure 4.10: Confusion matrix for the three narrowest temporal modulation filters for consonants, summed across all subjects. Rows represent stimulus presented, columns represent recorded response, and colors show the number of correct answers. The top matrix is the average confusion matrix for three high-pass temporal modulation filtering conditions (1.75 Hz, 3.75 Hz, 7.75 Hz). The bottom confusion matrix is the average confusion matrix for three low pass temporal modulation filtering conditions (0.25, 1.75 and 3.75 Hz).

Consonant confusion matrices

In order to explore with more detail the types of mistakes subjects made with different levels of filtering, we created consonant confusion matrices for the four different types of filtering. Each time, we show the average of the three most narrow filters, where subjects made the most mistakes.

We can see in the high-pass spectral modulation filtering condition (Fig. 4.9, top) that there is a pattern of mistakes that emerges. Stops and fricatives are often confused for one another, but not for sonorants. Specifically, many stops and fricatives (CH, G, K, T, and to a lesser extent B, D, F, J, S and SH) are categorized as P. Z is equally heard as TH, V or Z, but V is usually accurately categorized. On the other hand, sonorants are often confused with one another, with L, M and N especially prone to confusion. No such pattern is revealed in the low-pass spectral modulation filtering condition (Fig. 4.9, bottom).

For the high-pass temporal modulation filtering condition (Fig. 4.10, top), the main pattern that emerges is that all classes of consonants (stops, fricatives, and sonorants) can sometimes be heard as TH. TH is itself confused with V and Z. For the low-pass temporal modulation filtering condition (Fig. 4.10, bottom), almost any consonant can be heard as almost any other consonant. Of note, people tend not to identify any consonant as B, CH, D or T.

4.4 Discussion

The goal of this experiment was to investigate which spectral and temporal modulations are crucial to the comprehension of phonemes, and by extension, to speech. To do so, we used a psychoacoustical paradigm, presenting filtered consonants and vowels to subjects (12 alternative forced choice for vowels, 20 alternative forced choice for consonants). We found that a core of spectral and temporal modulations is both necessary and sufficient for phoneme comprehension: the core is within the range of 1-7 Hz for temporal modulations, and 0-4 cycles/kHz for spectral modulations that was previously found for sentence comprehension (Elliott and Theunissen, 2009), however is even narrower than this range for specific classes of phonemes. This confirms and extends previous modulation filtering research that focused on sentence comprehension (Elliott and Theunissen, 2009). We also found that vowels and consonants are affected differently by temporal and spectral filters, and by whether the filters are high pass or low pass.

Vowels are almost unaffected by the spectral low-pass filtering until below 1.75 cycles/kHz: only the narrowest low-pass spectral modulation filters (0.75 cycles/kHz, and 0.25 cycles/kHz) greatly impacted comprehension. For the spectral high-pass condition, filtering vowels above 0.25 cycles/kHz also does not affect comprehension, although it is severely affected for filters above 0.25 cycles/kHz. This indicates

that a very narrow core of spectral modulations–between 0.25 cycles/kHz and 1.75 cycles/kHz–are necessary for vowel comprehension.

The average modulation power spectrum of vowels shows that power is focused in only a narrow range of temporal modulation frequencies, and a wider range of spectral modulation frequencies (Fig 4.2). We would therefore anticipate that temporal modulation filtering impacts vowels less than spectral modulation filtering. We did find this to be the case for the low-pass temporal filter, but not for the high-pass temporal filter, as we discuss below.

For the low-pass temporal filter, vowel comprehension declines slowly below 1.75 Hz; even at the narrowest low-pass filter of 0.25 Hz, vowels are still very intelligible. For the high-pass temporal filter, comprehension is almost unaffected up until 0.75 Hz, after which there is a steady decline in comprehension.

We have therefore found a smaller core of crucial spectrotemporal modulations for vowels than previously reported in the literature: the most important temporal modulations range between 0.75 and 1.75 Hz; the most important spectral modulations range between 0.25 and 1.75 cycles/kHz. Similar to vowels, consonants are also almost unaffected by the spectral low-pass filtering until below 1.75 cycles/kHz: only the narrowest low-pass spectral modulation filters (0.75 cycles/kHz, and 0.25 cycles/kHz) impacted comprehension, and comprehension at 0.75 cycles/kHz is still extremely high. The spectral low-pass filter for consonants is similar to the temporal low-pass filter for vowels: even at the most filtered condition (0.25 cycles/kHz), comprehension is above 50%. For the spectral high-pass condition, there is a steady decline for consonants starting at the widest filter (0.25 cycles/kHz and above). At the narrowest filter (3.75 cycles/kHz and above), consonants, like vowels, are completely unintelligible. This indicates that spectral modulations below 1.75 cycles/kHz are necessary and sufficient for consonant comprehension.

The average modulation power spectrum of consonants shows power that is focused in a wider range of spectrotemporal frequencies than vowels (Fig 4.2). We would therefore anticipate that temporal modulation filtering impacts consonants more than spectral modulation filtering, and that temporal modulation filtering impacts consonants more than it impacts vowels. We found that consonants are far less intelligible than vowels for the low-pass temporal filter, but not for the high-pass temporal filter. The low-pass temporal filter impacts consonants' intelligibility far more than the low-pass spectral filter.

For the low-pass temporal filter, comprehension declines slowly starting at our widest filter of 7.75 Hz; consonants are completely incomprehensible at the narrowest low-pass filter of 0.25 Hz. For the high-pass temporal filter, comprehension is almost unaffected up until 1.75 Hz, after which there is a small decline in comprehension at 3.75 Hz, and a much larger drop in comprehension between 3.75 and 7.75 Hz. However, even at 7.75 Hz, consonants are still quite intelligible. This indicates that the most important temporal modulations for consonants range above 1.75 Hz: we do not have an upper bound with this experiment, given that even for our narrowest

high-pass filter (above 7.75 Hz), consonants are still intelligible.

We have therefore found both a smaller and a wider core of crucial spectrotemporal modulations for consonants than previously described in the literature (Gallun & Souza, 2008; Elliott & Theunissen, 2009). The most important temporal modulations range lie above 1.75 Hz, but we were not able to determine an upper bound with this experiment; the most important spectral modulations range between 0 and 1.75 cycles/kHz.

Finally, we investigated differences in intelligibility between three classes of consonants: sonorants, stops, and fricatives. We found that the three classes were very similar for the low-pass spectral filter and the high-pass temporal filter (except for the narrowest filtering condition, for both, in which fricatives were much more intelligible than the sonorants or stops). The three classes of consonants followed very similar intelligibility curves for the low-pass temporal modulation filter and the high-pass spectral modulation filter. For the low-pass temporal modulation filter, the sonorants are always more intelligible than the fricatives, which are always more intelligible than the stops (except for the narrowest low pass temporal filter, in which the fricatives and stops are both at floor performance). For the high-pass spectral filtering condition, the sonorants are always more intelligible than the stops, which are always more intelligible than the fricatives (except for the narrowest high-pass filter, where all three classes of consonants are at floor performance). We therefore find an interesting differentiation in intelligibility between these three classes of consonants for low pass temporal filtering and high-pass spectral filtering.

In order to extend this research and specify full lower and upper bounds for spectral and temporal modulations of vowels and consonants, it would be fruitful to investigate range of temporal modulations above 7.75 Hz and spectral modulations below 0.25 cycles/kHz for consonants, and temporal modulations below 0.25 Hz for vowels.

In this Chapter, we have shown that there is a narrow range of specific spectral and temporal modulations that are both necessary and sufficient to understand phonemes. We further showed that consonants and vowels have very different modulation power spectra; furthermore, subclasses of consonants (fricatives, stops and plosives) also have specific modulation power spectra. The combination of specific modulation signatures for vowels and consonant subtypes, and the importance of specific temporal and spectral modulations for the comprehension of consonants and vowels, indicates that the MPS would be an intriguing sound space to investigate when constructing hierarchical cortical models of speech perception.

Chapter 5

Conclusion

This dissertation investigated the cortical representation of speech perception, using a combination of fMRI and psychoacoustical experiments. In Chapters 2 and 3, we demonstrated that the cortical streams involved in speech perception could be efficiently studied using fMRI of natural speech and a novel analysis that relied on nested models using a variety of high-dimensional feature spaces to represent the stimulus. We examined three features spaces representing the speech sounds in terms of auditory, articulatory and semantic features. Validating our approach, we found similar results as those described in previous research using synthetic sounds or segmented speech: there is a hierarchical set of processing steps starting in primary auditory cortex and moving along the posteroventral region of the temporal lobe that are involved in the sound to word meaning transformation. However, in natural hearing (and using our methodology) these transformations appear to occur earlier in the processing streams than what has been described in previous studies. The differences might be due to differences in the nature of the stimuli used but could also result from the very different attentional mechanisms present during natural speech that certainly engage stronger top-down processes that those involved in listening to segmented speech.

In Chapter 4, we investigated a promising intermediate feature space, the modulation power spectrum, or 2-dimensional Fourier transform of the speech spectrogram, that can be used to create an even more detailed picture of the cortical processing stream of speech. Using psychoacoustics, we showed that comprehension of vowels and consonants is differently affected by removal of specific spectral or temporal modulations. We further demonstrated that low-pass spectral and high-pass temporal modulation filtering are more detrimental to vowel than to consonant comprehension. Our findings are consistent with what we would expect given differences in the modulation power spectrum (MPS) of vowels and of consonants. Supplementary consonant analysis showed significant differences in MPS and psychoacoustical comprehension results between three groups of consonants, separated based on the manner in which they are pronounced (fricatives, stops, and sonorants).

Future work using additional intermediate feature spaces, such as the space of

modulation power spectra investigated in Chapter 4, will provide greater detail on the exact nature of computations occurring along all the cortical streams involved in language perception. Furthermore, additional experiments involving the modulation of attention while listening to natural speech should be performed to begin to understand the role of top-down processes. For the first task of better describing the cortical streams, using natural speech could be more efficient than more classical approaches. For the second task of investigating top-down processes, it will be crucial to study natural attentive mechanisms engaged during natural speech processing. In both cases, natural speech experiments will play an important role in further understanding the human language brain network.

Bibliography

Abrams, D. A., Nicol, T., Zecker, S., & Kraus, N. (2008). Right-Hemisphere Auditory Cortex Is Dominant for Coding Syllable Patterns in Speech. *Journal of Neuroscience*, *28*(15), 3958–3965.

Belin, P. & Zatorre, R. J. (2000). 'What', 'where' and 'how' in auditory cortex. *Nature Neuroscience*, *3*(10), 965–966.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex*, *19*(12), 2767–2796.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, *10*(5), 512–528.

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature Neuroscience*, *8*(3), 389–395.

Bonte, M., Hausfeld, L., Scharke, W., Valente, G., & Formisano, E. (2014). Task-Dependent Decoding of Speaker and Vowel Identity from Auditory Cortical Response Patterns. *Journal of Neuroscience*, *34*(13), 4548–4557.

Bornkessel-Schlesewsky, I., Schlesewsky, M., Small, S. L., & Rauschecker, J. P. (2015). Neurobiological roots of language in primate audition: common computational properties. *Trends in Cognitive Sciences*, *19*(3), 142–150.

Brant-Zawadzki, M., Gillan, G. D., & Nitz, W. R. (1992). Mp rage: a three-dimensional, t1-weighted, gradient-echo sequence–initial experience in the brain. *Radiology*, *182*(3), 769–775.

Chi, T., Gao, Y., Guyton, M. C., Ru, P., & Shamma, S. (1999). Spectro-temporal modulation transfer functions and speech intelligibility. *The Journal of the Acoustical Society of America*, *106*(5), 2719–2732.

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887–906.

Da Costa, S., van der Zwaag, W., Miller, L. M., Clarke, S., & Saenz, M. (2013). Tuning In to Sound: Frequency-Selective Attentional Filter in Human Primary Auditory Cortex. *Journal of Neuroscience*, *33*(5), 1858–1863.

Dale, A., Fischl, B., & Sereno, M. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage*, *9*, 179–194.

Desai, R., Liebenthal, E., Waldron, E., & Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *Journal of Cognitive Neuroscience*, *20*(7), 1174–1188.

DeWitt, I. & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences*, *109*(8), E505–E514.

Elie, J. E. & Theunissen, F. E. (2015). Meaning in the avian auditory cortex: neural representation of communication calls. *European Journal of Neuroscience*, *41*(5), 546–567.

Elliott, T. M. & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS computational biology*, *5*(3), e1000302.

Escabí, M. A., Miller, L. M., Read, H. L., & Schreiner, C. E. (2003). Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. *The Journal of neuroscience*, *23*(37), 11489–11504.

Fedorenko, E., Nieto-Castañón, A., & Kanwisher, N. (2012). Syntactic processing in the human brain: what we know, what we don't know, and a suggestion for how to proceed. *Brain and language*, *120*(2), 187–207.

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). " who" is saying" what"? brain-based decoding of human voice and speech. *Science*, *322*(5903), 970–973.

Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention—focusing the searchlight on sound. *Current Opinion in Neurobiology*, *17*(4), 437–455.

Gallun, F. & Souza, P. (2008). Exploring the role of the modulation spectrum in phoneme recognition. *Ear and hearing*, *29*(5), 800.

Garcia-Lazaro, J. A., Ahmed, B., & Schnupp, J. W. H. (2006). Tuning to Natural Stimulus Dynamics in Primary Auditory Cortex. *Current Biology*, *16*(3), 264–271.

Greenberg, S. & Kingsbury, B. E. (1997). The modulation spectrogram: in pursuit of an invariant representation of speech. In *Acoustics, speech, and signal processing, 1997. icassp-97., 1997 ieee international conference on* (Vol. 3, pp. 1647–1650). IEEE.

Grill-Spector, K. & Malach, R. (2001). fMR-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta psychologica*, *107*(1-3), 293–321.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099–3111.

Humphries, C., Liebenthal, E., & Binder, J. R. (2010). Tonotopic organization of human auditory cortex. *Neuroimage*, *50*(3), 1202–1211.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224.

Jenkinson, M. & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical image analysis*, *5*(2), 143–156.

Joanisse, M. F., Zevin, J. D., & McCandliss, B. D. (2007). Brain Mechanisms Implicated in the Preattentive Categorization of Speech Sounds Revealed Using fMRI and a Short-Interval Habituation Trial Paradigm. *Cerebral Cortex*, *17*(9), 2084–2093.

Jones, E. et al., Oliphant, T. et al., Peterson, P., et al. (2007). Scipy: open source scientific tools for python, 2001–. *URL http://www. scipy. org*, *73*, 86.

Joris, P., Schreiner, C., & Rees, A. (2004). Neural processing of amplitude-modulated sounds. *Physiological reviews*, *84*(2), 541–577.

Just, M. A., Cherkassky, V. L., Aryal, S., & Mitchell, T. M. (2010). A Neurosemantic Theory of Concrete Noun Representation Based on the Underlying Brain Codes. *PLoS ONE*, *5*(1), e8622.

Kashino, M. (2006). Phonemic restoration: The brain creates missing speech sounds. *Acoustical Science and Technology*, *27*(6), 318–321.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355.

Knecht, S., Flöel, A., Dräger, B., Breitenstein, C., Sommer, J., Henningsen, H., ... Pascual-Leone, A. (2002). Degree of language lateralization determines susceptibility to unilateral brain lesions. *Nature Neuroscience*, 1–5.

Langers, D. R., Backes, W. H., & van Dijk, P. (2003). Spectrotemporal features of the auditory cortex: the activation in response to dynamic ripples. *Neuroimage*, *20*(1), 265–275.

Leaver, A. M. & Rauschecker, J. P. (2010). Cortical Representation of Natural Complex Sounds: Effects of Acoustic Features and Auditory Object Category. *Journal of Neuroscience*, *30*(22), 7604–7612.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *Journal of Neuroscience*, *31*(8), 2906–2915.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*(5), 407–412.

Margoliash, D. & Fortune, E. S. (1992). Temporal and harmonic combination-sensitive neurons in the zebra finch's HVc. *The Journal of neuroscience*.

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *The Journal of the Acoustical Society of America*, *123*(2), 899–11.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191–1195.

Moerel, M., De Martino, F., & Formisano, E. (2012). Processing of Natural Sounds in Human Auditory Cortex: Tonotopy, Spectral Tuning, and Relation to Voice Sensitivity. *Journal of Neuroscience*, *32*(41), 14205–14216.

Mottonen, R., Dutton, R., & Watkins, K. E. (2013). Auditory-Motor Processing of Speech Sounds. *Cerebral Cortex*, *23*(5), 1190–1197.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, *21*(19), 1641–1646.

Oliphant, T. E. (2006). *A guide to numpy*. Trelgol Publishing USA.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., ... Chang, E. F. (2012). Reconstructing Speech from Human Auditory Cortex. *PLoS Biology*, *10*(1), e1001251.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension. *Cerebral Cortex*, *23*(6), 1378–1387.

Price, C. J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, *1191*(1), 62–88.

Pujol, J., Deus, J., Losilla, J. M., & Capdevila, A. (1999). Cerebral lateralization of language in normal left-handed people studied by functional mri. *Neurology*, *52*(5), 1038–1038.

Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F. M., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, *103*(20), 7865–7870.

Rauschecker, J. P. & Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718–724.

Rauschecker, J. P., Tian, B., & Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, *268*(5207), 111–114.

Rodd, J. M. (2004). The Neural Mechanisms of Speech Comprehension: fMRI studies of Semantic Ambiguity. *Cerebral Cortex*, *15*(8), 1261–1269.

Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fmri studies of semantic ambiguity. *Cerebral Cortex*, *15*(8), 1261–1269.

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.

Schönwiesner, M. & Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences*, *106*(34), 14611–14616.

Scott, S. K. & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in neurosciences*, *26*(2), 100–107.

Shannon, R. V., Jensvold, A., Padilla, M., Robert, M. E., & Wang, X. (1999). Consonant recordings for speech testing. *The Journal of the Acoustical Society of America*, *106*(6), L71–L74.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304.

Singh, N. C. & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, *114*(6), 3394–3411.

Smith, E. C. & Lewicki, M. S. (2006). Efficient auditory coding. *Nature*, *439*(7079), 978–982.

Snijders, T. M., Vosse, T., Kempen, G., Van Berkum, J. J. A., Petersson, K. M., & Hagoort, P. (2009). Retrieval and Unification of Syntactic Structure in Sentence Comprehension: an fMRI Study Using Word-Category Ambiguity. *Cerebral Cortex*, *19*(7), 1493–1503.

Stowe, L. A., Haverkort, M., & Zwarts, F. (2005). Rethinking the neurological basis of language. *Lingua*, *115*(7), 997–1042.

Suga, N., O'Neill, W. E., & Manabe, T. (1978). Cortical neurons sensitive to combinations of information-bearing elements of biosonar signals in the mustache bat. *Science*, *200*(4343), 778–781.

Theunissen, F. E. & Elie, J. E. (2014). Neural processing of natural sounds. *Nature Publishing Group*, *15*(6), 355–366.

Tian, B., Reser, D., Durham, A., Kustov, A., & Rauschecker, J. P. (2001). Functional specialization in rhesus monkey auditory cortex. *Science*, *292*(5515), 290–293.

Turkeltaub, P. E. & Coslett, H. B. (2010). Localization of sublexical speech perception components. *Brain and Language*, *114*(1), 1–15.

Visser, M., Jefferies, E., & Ralph, M. (2010). Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature. *Journal of Cognitive Neuroscience*.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*.

Weisz, N., Wienbruch, C., Hoffmeister, S., & Elbert, T. (2004). Tonotopic organization of the human auditory cortex probed with frequency-modulated tones. *Hearing research*, *191*(1), 49–58.

Wild, C. J., Davis, M. H., & Johnsrude, I. S. (2012). Human auditory cortex is sensitive to the perceived clarity of speech. *NeuroImage*, *60*(2), 1490–1502.