A Dissertation
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Neuroscience

By

Mark A. Chevillet, B.S.

Washington, DC
August 31, 2011
Current theories of speech perception in the human brain propose two separate, hierarchically organized cortical processing streams, a ventral and a dorsal stream. The ventral stream is generally believed to mediate acoustic decoding of the speech signal and, ultimately, link sounds to meanings. The function of the dorsal stream is less well understood, however, and has been a matter of some debate. It has recently been postulated that the dorsal stream may play a key role in sensorimotor integration linking speech sounds to motor articulations. To further investigate these theories we first examined the functional organization of human auditory cortex. Using functional magnetic resonance imaging (fMRI), robust hierarchical organization was identified in human auditory cortex. Whereas primary auditory cortex responded to all sounds tested, the surrounding regions responded only to sufficiently complex sounds; a surrounding belt region responded to both band-passed noise bursts and phonemes (but not pure tones), and a more distant parabelt region responded only to phonemes.

We further probed the neural representations of phonemes in the human brain using a novel fMRI rapid adaptation (fMRI-RA) paradigm. In fMRI-RA, two stimuli are presented in each trial, and the resulting BOLD-signal is thought to reflect the dissimilarity between neuronal activation patterns for the two stimuli. By pairing speech sounds of comparable acoustic
dissimilarity from either the same or a different phonetic category we could dissociate neuronal selectivity for acoustic-features vs. phonetic categories. Our results support a model of speech processing in which a ventral stream represents sounds in an acoustic feature-based hierarchy and links them to task-relevant meanings, while a dorsal stream automatically links speech sounds to their motor-articulations via separate sensorimotor representations of speech sounds and articulatory phoneme categories.
ACKNOWLEDGEMENTS

I am thankful and proud to have had the opportunity to pursue my graduate studies at Georgetown, and specifically within the Interdisciplinary Program in Neuroscience. Special thanks are in order for all the people who have spent a great deal of time and effort making the IPN such a great program, especially Karen Gale, Bill Rebeck, Barry Wolfe and Becky Hoxter.

My deepest gratitudes are of course due to my mentors, Max Riesenhuber and Josef Rauschecker, for their insights, guidance, and willingness to tackle the challenges involved in dual-mentoring a graduate student. This arrangement provided a rich training environment, in which I gained detailed knowledge of both the auditory and visual sensory systems. This in turn provided us all the unique opportunity to make detailed comparisons between these systems at a level that simply would not have been possible in either lab alone.

Special thanks are also due to Guinevere Eden for chairing my committee, and providing exceptional support and guidance, and to the rest of my committee for their contributions and suggestions: Michael Ullman, John VanMeter, Jagmeet Kanwal and Peter Bandettini.

I also want to thank all my labmates in the Riesenhuber and Rauschecker labs, and classmates in the IPN for their insights, friendship and general willingness to listen to my rambling. In particular, these studies could not have been done without the exceptional help I received from Xiong Jiang and Amber Leaver regarding experimental design and data analysis. I would also like to thank everyone down at CFMI for their help and tolerance for ever-changing experimental setups.
Finally, I am most grateful to my family and friends who provided ever-increasing levels of love and support, especially my dear wife Jeanne and our vibrant twin toddlers Chloe and Noah!
Chapter IV. Discussion

A. Summary of goals ................................................................. 64

B. Summary of results ............................................................... 65

C. Interpretation ........................................................................ 66

D. Hierarchical vs. ‘modular’ processing ....................................... 68

E. What kinds of invariance do we expect? .................................. 70

F. Transfer of learning ............................................................... 72

Bibliography .............................................................................. 74
List of Figures and Tables

Figure 1. Three functionally discrete regions identified in human auditory cortex ...... 15
Figure 2. Progression of simple to complex selectivity in the antero-lateral direction 18
Figure 3. Functional organization in individual subjects........................................ 22
Figure 4. Quantification of single-subjects analysis............................................. 24
Figure 5. Individual subjects analysis compared to group analysis..................... 27
Figure 6. Organization based on spectral complexity....................................... 29
Figure 7. Behavioral performance outside the scanner...................................... 49
Figure 8. Selecting stimuli for fMRI conditions................................................. 54
Figure 9. Neural selectivity for acoustic features in left auditory cortex............... 55
Figure 10. Neural selectivity for category labels in left premotor cortex............. 57
Figure 11. Signal in the left premotor ROI....................................................... 60
Table 1. Results of contrasts in the group random-effects analysis............... 19
Table 2. Activation volume increases with increasing spectral complexity......... 21
CHAPTER I

INTRODUCTION

A. Speech Recognition as Perceptual Categorization

Object recognition across all sensory modalities is a crucial cognitive task that requires mapping variable patterns of sensory input to their specified meaning. A particular challenge in object recognition is that different sounds can have the same meaning (e.g., the same word spoken by different speakers) and some words that sound similar might have different meanings (e.g., “date” vs. “gate”). These requirements are common to recognition problems across sensory domains, and are known as specificity and invariance, respectively. In audition, higher primates have highly developed mechanisms to recognize complex sounds and use them for auditory communication, the most notable example of which is human speech recognition. While appearing effortless, even when attention is shared with other demanding tasks (e.g., carrying on a conversation while driving through a crowded city), recognizing sounds (including speech) remains a computationally difficult problem and the underlying neural mechanisms are only poorly understood. However, posing sound recognition (and in particular, speech recognition) as a specific case of perceptual categorization, as others have done (Holt and Lotto, 2010), provides a rich framework in which to investigate plausible mechanisms for categorical perception.

In particular, a computational strategy has been shown to solve perceptual
categorization problems by using two separate stages of processing (Riesenhuber and Poggio, 2000):

1. a stimulus-selective, hierarchical neural representation

2. task representations separate from the sensory representation

The first stage of this model is a hierarchical sensory-feature representation, wherein increasing degrees of specificity and invariance are gradually built up by successively combining simpler features detected by lower layers. This hierarchy provides a continuous neural representation of stimuli based on specific physical features, but invariant to simple transformations (such as change in size and location in vision). The second stage of the model contains task circuits that can divide the continuous sensory representation according to specific tasks (i.e. categorization, discrimination, etc.). This architecture is computationally appealing, in that it allows a single perceptual representation to support multiple different tasks. Take, for example, a case of four sounds including a male and female each speaking the words “hid” and “head.” One category circuit may divide the sounds according to the meaning of the spoken word, while another may divide them by the gender of the speaker. Whereas biasing the perceptual representation to facilitate one task would interfere with the other, the two-stage architecture keeps these tasks separated and thus avoids interference.

In addition to the theoretical support for this model, this strategy has also been implemented as a computational model to quantitatively test its recognition capabilities. Initial simulations demonstrated that a hierarchical feature
representation is capable of accurately recognizing specific objects, invariant to certain stimulus transformations (Riesenhuber and Poggio, 1999a), even in the presence of other objects (Riesenhuber and Poggio, 1999b). Further work also showed that the two-stage architecture qualitatively matched the recognition behaviors of human subjects (Jiang et al., 2006; Serre et al., 2007).

Evidence supporting the use of this computational strategy in biological sensory systems has emerged from recent studies of the visual systems of macaques (Op de Beeck et al., 2001; Thomas et al., 2001; Freedman et al., 2003) and humans (Jiang et al., 2007). In the case of vision, the well-known ventral pathway (or “what” stream) is thought to provide a continuous, hierarchical representation of objects (first-stage), whereas neurons supporting specific cognitive tasks (second-stage), such as categorization, are thought to be separately located in parts of prefrontal cortex (Freedman et al., 2001; Jiang et al., 2007).

While the theory has emerged in synchrony with our understanding of the visual system, the computations proposed need not be domain-specific. In the next section I will discuss current models of auditory recognition where a similar functional organization is proposed, providing the opportunity to test whether a second sensory (auditory) system utilizes such a two-stage model to solve an important problem of perceptual categorization.
B. Current Theories of Speech Processing

Detailed anatomical (Kaas and Hackett, 2000b) and physiological (Tian et al., 2001a) studies in non-human primates provided early evidence that the auditory system shares key organizational principles with the visual system. Namely, that both systems exhibit two separate processing pathways, and that these pathways are hierarchically organized. These studies identified a ventral pathway, projecting from primary auditory cortex through anterior aspects of the superior temporal lobe and on into inferior regions of prefrontal cortex, and suggested that it functions to recognize sounds (i.e., “what” stream). A separate dorsal pathway was also identified, projecting from the primary auditory cortex through the posterior superior temporal (pST) areas into inferior parietal (IPL), premotor (PMC) and dorsolateral prefrontal regions (DLPFC), which was implicated in processing the spatial location of sounds (i.e., “where” stream).

Theories of speech processing, however, took issue with this proposed functional organization claiming that it was in conflict with lesion data (and more recently, functional imaging data) from humans implicating the involvement of pST in speech processing. Given that both functions in pST are strongly supported by experimental evidence, more recent accounts of speech processing reconcile this conflict by expanding the interpretation of dorsal stream function to include more general processes of sensorimotor integration (Hickok et al., 2011; Rauschecker, 2011), a function that has been similarly proposed in the visual
dorsal stream (Goodale and Milner, 1992; Kravitz et al., 2011). By this account, it has also been proposed that the dorsal stream supports “internal models” (Wolpert et al., 1995) of receptive and productive speech (Rauschecker and Scott, 2009b). In the case of speech, a “forward model” refers to predicting the sensory result of an intended motor articulation (i.e. how an utterance will sound), whereas an “inverse model” refers to predicting the motor articulation that caused a particular speech sound.

In addition to sharing dual-pathway organization, the auditory and visual cortical systems also exhibit hierarchical organization in both anatomy and function. In the ventral auditory pathway, projections from primary auditory cortex reach the surrounding belt, and projections from belt then extend further to parabelt regions. However, parabelt receives comparably few projections directly from primary auditory cortex. This anatomical hierarchy evokes physiological correlates as well, in that neurons in early auditory areas respond well to simple sounds while those in higher auditory cortical regions respond selectively to increasingly complex sounds (see Chapter II for more detail). This property is often referred to as “combination sensitivity,” (Suga et al., 1979b; Margoliash and Fortune, 1992b). Finally, projections from parabelt reach beyond auditory cortex into prefrontal regions, including the inferior frontal gyrus (IFG). As in vision, prefrontal function in the auditory system has been tied to task execution (Freedman et al., 2001; Russ et al., 2007). While our understanding of the dorsal stream is somewhat less detailed, it is thought that the same principle of
hierarchical organization should apply given that it has also been shown in the visual system (Konen and Kastner, 2008).

C. Investigational overview

Given the structural and functional similarities reported across auditory and visual systems, it is possible that these architectures support similar computations across domains as well. Taken together with the success of the two-stage model in the visual domain, it would seem pertinent to ask whether or not such a framework can be used to understand auditory recognition as well. While this argument applies to auditory recognition in general, the experiments presented in Chapters II and III focus on the specific topic of speech recognition because it is natural and familiar to human subjects, and can evoke robust categorization behaviors (Liberman et al., 1957).

The experiments presented in the following chapters were designed to test whether or not the auditory system exhibited evidence of a hierarchical, stimulus-selective neural representation and separate “task” representations as proposed by the two-stage model of vision. In Chapter II, I will present an fMRI experiment in which the cortical responses to increasingly complex stimuli reveal clearly hierarchical functional organization in auditory cortex. In Chapter III, I will then present a second fMRI study that includes behavioral experiments. This study suggests that neurons in the higher layers of the auditory hierarchy respond selectively to speech stimuli based on their acoustic-features, while the neurons
in higher cortical regions (in this case, left PMC) respond to the categorically defined meanings of speech stimuli. Finally, Chapter IV discusses how the results of these two studies provide key points of convergence to the computational models described, and explores several important implications of this framework in the auditory domain to be tested by future studies.
CHAPTER II

FUNCTIONAL LOCALIZATION OF THE VENTRAL AUDITORY HIERARCHY

A. Summary

Converging evidence supports the hypothesis that an antero-lateral processing pathway mediates sound identification in auditory cortex, analogous to the role of the ventral cortical pathway in visual object recognition. Studies in nonhuman primates have characterized the antero-lateral auditory pathway as a processing hierarchy, composed of three anatomically and physiologically distinct initial stages: core, belt and parabelt. In humans, potential homologues of these regions have been identified anatomically, but reliable and complete functional distinctions between them have yet to be established. Because the anatomical locations of these fields vary across subjects, investigations of potential homologues between monkeys and humans require these fields to be defined in single subjects. Using functional MRI, we presented three classes of sounds (tones, band-passed noise bursts, and conspecific vocalizations), equivalent to those used in previous monkey studies. In each individual subject, three regions showing functional similarities to macaque core, belt and parabelt were readily identified. Furthermore, the relative sizes and locations of these regions were consistent with those reported in human anatomical studies. Our results demonstrate that the functional organization of the antero-lateral processing pathway in humans is largely consistent with that of nonhuman primates. Because our scanning sessions last only 15 min/subject, they can be run in
conjunction with other scans. This will enable future studies to characterize functional modules in human auditory cortex at a level of detail previously possible only in visual cortex. Furthermore, the approach of employing identical schemes in both humans and monkeys will aid in establishing potential homologies between them.

B. Introduction

Convergent evidence indicates that sound recognition in the macaque is mediated by an antero-laterally directed cortical pathway. This pathway forms an anatomical hierarchy: projections extend from auditory core cortex via antero-lateral belt to rostral parabelt and ventrolateral prefrontal cortex (Romanski et al., 1999; Kaas and Hackett, 2000a; Hackett et al., 2001). This hierarchy is also expressed functionally; whereas core neurons respond selectively to pure tones (Merzenich and Brugge, 1973), belt responses require band-passed noise (BPN) (Rauschecker et al., 1995; Petkov et al., 2006; Kusmierek and Rauschecker, 2009), and parabelt neurons respond to more complex sounds such as species-specific vocalizations (Poremba et al., 2003; Petkov et al., 2008b; Kikuchi et al., 2010). Ventrolateral prefrontal areas respond strongly to vocalizations (Romanski and Goldman-Rakic, 2002; Averbeck and Romanski, 2006) and possibly task-related aspects of sounds (Cohen et al., 2009).
This anatomical hierarchy is thought to be conserved across species: human core (corresponding to Brodmann area (BA) 41 (Hackett et al., 2001)), belt and parabelt regions (Sweet et al., 2005) are reliably identified using the same techniques as in the monkey. Functionally, human core and belt have been dissociated using stimulus preferences defined in the macaque (Wessinger et al., 2001). However, while more antero-lateral regions of human auditory cortex are known to respond well to species-specific vocalizations (i.e., human speech and voices) (Belin et al., 2000; Binder et al., 2000b; Scott et al., 2000a; Leaver and Rauschecker, 2010b), a clear functional delineation of belt and parabelt in humans has not yet been performed. Demonstration of such a boundary would argue for at least three discrete processing stages in human auditory cortex, an important step toward assessing potential homologies between human and macaque. Furthermore, the inability to delineate belt from parabelt regions in humans restricts our ability to interpret other human neuroimaging data. For example, several studies demonstrate selectivity for several types of speech sounds in antero-lateral auditory regions (Obleser et al., 2006b; Joanisse et al., 2007b; Leaver and Rauschecker, 2010b), but at present it is unclear whether these regions belong to belt or parabelt.

Besides the fact that a clear belt/parabelt boundary has not been established in humans, it is also well known that the anatomical locations of auditory subregions vary widely between individuals (Penhune et al., 1996; Morosan et al., 2001; Rademacher et al., 2001; Sweet et al., 2005). When investigating processing in
these subregions, then, using a functional localization scheme would assure data to be functionally consistent across studies. Human visual cortex studies routinely identify regional boundaries by means of visual field mapping (Engel et al., 1997; Sereno et al., 2001; Wandell et al., 2007). In auditory cortex, however, frequency gradient reversals can demarcate cochleotopic areas only at the same hierarchical level (Merzenich and Brugge, 1973), but not between levels. The latter boundaries correlate instead with bandwidth and spectro-temporal complexity (Rauschecker et al., 1995). Here, we use the term complexity to refer to the likely information content of a particular stimulus class relative to another. Thus, we predicted the borders between core, belt, and parabelt could be functionally localized in single human subjects using spectro-temporal stimulus preferences defined in the macaque.

C. Experimental Procedures

Subjects. The Georgetown University Institutional Review Board gave approval for all experiments, and subjects gave written informed consent before participating. A total of thirteen subjects participated in this investigation (18-36 years of age, 5 males). All subjects were right-handed, reported no history of hearing problems or neurological illness, and spoke American English as their native language.

Stimuli. Pure-tone (PT), band-passed noise (BPN) and white noise (WN) stimuli were generated using MATLAB (Mathworks). PT and BPN stimuli had center
frequencies of 500, 2000 and 8000 Hz, and BPN stimuli had a bandwidth of 1 octave. Limiting the species-specific vocalization category to vowels (VOW) allowed all three sound classes to be well matched in duration and temporal dynamics. Additionally, VOW stimuli provide a logical progression in spectral complexity in that they contain combinations of BPN, analogous to BPN containing combinations of PT (Figure 1A). VOW stimuli were isolated from a database of recordings available online (Hillenbrand et al., 1995). The vowel portions of these waveforms were identified, and sections that were largely steady-state were selected on the basis of visual inspection. All stimuli were root-mean-square (RMS) normalized in amplitude and were presented as a short block (i.e. 166 ms of sound followed by 166 ms of silence, repeated for 3 s). Six subjects were presented with PT, BPN and VOW stimuli; seven subjects were presented with PT, BPN, VOW and WN stimuli.

**Data Acquisition.** All MRI data were acquired at the Center for Functional and Molecular Imaging at Georgetown University on a 3.0-Tesla Siemens Trio Scanner using whole-head echo-planar imaging (EPI) sequences (Flip Angle = 90°, TE = 30 ms, FOV = 205, 64x64 matrix). In six subjects a single-channel CP head coil was used, and in the remaining seven subjects a 12-channel head coil was used. A sparse sampling paradigm (Edmister et al., 1999) was used in which image acquisition is clustered into a portion of the total repetition time (TR = 7500 ms; TA = 2000 ms). Stimuli were presented during the remaining silent portion of the TR (5500 ms), starting 1000 ms after the offset of a volume acquisition. Following functional scans, high-resolution (1x1x1 mm³) anatomical images
(MPRAGE) were acquired. Auditory stimuli were presented through STAX electrostatic earphones at a comfortable listening volume (~65-70 dB) worn inside ear defenders giving ~26 dB attenuation. One stimulus block was presented on each trial. While in the scanner, subjects were asked to perform a passive listening task in which they simply attended to the presented stimuli. This task was chosen over an oddball task to maintain the short scan duration. Each condition was presented 30 times in random order, for a total of 120 volumes (PT, BPN, VOW and SIL conditions), or 150 volumes (PT, BPN, VOW, WN and SIL conditions).

Data Analysis. Data were analyzed using BrainVoyager QX 2.1.0 (Brain Innovation), SPSS and Matlab (The Mathworks). Functional images were first co-registered to their corresponding anatomical images, and then both sets of images were spatially normalized to the Talairach atlas. The first two volumes (14 s) of each functional run were discarded, and the remaining images were preprocessed for motion correction, linear trend removal and spatial smoothing using an 8-mm Gaussian kernel. Whole-head statistical maps were calculated for each subject individually using fixed-effects general linear models and then submitted to a second-level random-effects group analysis. Activation for each individual condition is reported relative to the baseline value, defined as the average value for each voxel over all acquisitions.
D. Results

Group analysis identifies three functionally discrete regions in auditory cortex

In a random-effects group analysis, we included scans from all thirteen subjects for PT, BPN and VOW. To visualize the activation pattern evoked by each stimulus condition, we produced activation maps for each condition relative to baseline. Because statistical maps vary with the chosen statistical threshold, we chose to set our threshold such that the volume of activation to the PT condition was consistent with values reported for auditory core (~1800 mm$^3$ averaged across hemispheres) in previous human anatomical studies (Penhune et al., 1996; Rademacher et al., 2001). This chosen threshold was restricted to be at least p(FDR) < 0.05, though it ultimately was much more restrictive. This approach has also been used in several recent functional neuroimaging studies of nonhuman primates (Petkov et al., 2006; Petkov et al., 2008a; Petkov et al., 2008b).

We then applied the same threshold to activation maps for BPN and VOW. To determine whether a functional hierarchy as identified previously in nonhuman primates could be identified, we created regions of interest (ROI) based on logical combinations of PT, BPN and VOW responses (Fig. 1B). First, since auditory core cortex is known to respond well to PT, we defined the set of all PT-responsive voxels to be auditory core. These voxels, shown in blue in
Figure 1. Three functionally discrete regions identified in human auditory cortex. (A) Example stimuli used. Subjects were presented with pure tones (PT), band-passed noises (BPN), and species-specific vocalizations (vowel-like speech sounds, VOW), based on the ability to differentiate core, belt and parabelt using these stimulus classes in previous nonhuman primate studies. (B) Random-effects group analysis (n = 13). A hierarchy of three functionally separable regions is clearly visible in each hemisphere. PT activation (blue) was centered on Talairach coordinates [48, -20, 7] and [-45, -24, 8]. Adjacent voxels responded to BPN but not PT (yellow), and further regions of cortex responded to VOW, but not BPN or PT. Minimum t value for all conditions is 9.8.
Figure 1B, were located primarily in bilateral Heschl’s gyri, in good agreement with anatomical expectations based on coordinates (Penhune et al., 1996; Rivier and Clarke, 1997). We also calculated the overlap of these voxels with cytoarchitectonic probability maps of three areas (Te 1.0, Te 1.1, Te 1.2), which together are thought to represent BA 41 (consistent with definitions for human “core”; (Hackett et al., 2001), and are distributed with the Anatomical Toolbox for SPM (Morosan et al., 2001; Rademacher et al., 2001; Eickhoff et al., 2005; Eickhoff et al., 2007). We found that as much as 76% of our group-defined PT responsive voxels overlapped with the conjunction of the three maps. More specifically, 75% were still found to overlap with the conjunction of only regions Te 1.0 and Te 1.1. Very few PT-responsive voxels extended into region Te 1.2, arguing that this subfield may be more consistent with belt than core (see Discussion). In a second ROI definition, we included all voxels responding to BPN, and then removed any voxels shared with PT activation (thus already assigned to auditory core cortex). This logical combination (BPN NOT PT) is the behavior to be expected from “belt” regions. These voxels were located in cortex adjacent (both medial and lateral) to our defined auditory core cortex ROI. Finally, we created a third region consisting of those voxels that responded to VOW, and then excluded all voxels already assigned to one of the other ROI (VOW NOT (BPN OR PT)). This response profile represents the behavior expected from rostral “parabelt” regions, in particular those responsive to VOW. These voxels were located in cortex adjacent to the belt ROI. Rather than
expanding symmetrically, however, these voxels were located primarily anterior, lateral and ventral to the adjacent “belt” ROI.

We further tested whether this progression from simple to complex selectivity in the antero-lateral direction was significant by running statistical comparisons directly between the conditions in our group study. Figure 2 shows the results of contrasts for both BPN > PT and VOW > BPN, and the sizes and coordinates for the resulting clusters are reported in Table 2. These contrasts yielded large significant activations at p(FDR) < 0.05. As can be seen in the series of slices, the BPN > PT contrast (yellow) yields activation primarily restricted to both the medial and lateral extents of Heschl’s gyrus in each hemisphere. This activation overlaps our PT-responsive region in the previous analysis. This indicates that complex stimuli such as BPN can evoke greater activation in core than PT stimuli. We interpret this to indicate that more neurons in core, each responding to select frequencies contained in the BPN, are recruited by the spectrally wider BPN. Additionally, the VOW > BPN contrast (red) yields activation further lateral, anterior and ventral to the BPN > PT contrast, also seen in the coordinates listed in Table 1. The two contrasts share very little overlap (orange).

Hierarchical organization is consistent across subjects

While the group analysis supports the existence of a tiered hierarchy and a general progression from dorso-medial to antero-lateral auditory cortex, it is well known that the specific anatomical locations of auditory cortical fields vary widely
Figure 2. Progression from simple to complex selectivity in the antero-lateral direction. A BPN > PT contrast (yellow) yielded activation largely restricted to areas adjacent to the medial and lateral extents of Heschl’s gyrus in each hemisphere. The VOW > BPN contrast yielded significant activation in antero-lateral aspects of the superior temporal gyrus. The two contrasts share little overlap (orange). Thresholds were $p$(FDR) < 0.05 for both contrasts, and the sizes and Talairach coordinates of the resulting clusters are reported in Table 2.
Table 1. Results of contrasts in the group random-effects analysis. Threshold is $p$(FDR)$<0.05$ for BPN $>$ PT and VOW $>$ BPN contrasts. Threshold for VOW $>$ WN is $p$(uncorr) $<0.007$, corrected to $P<0.05$ using a random Monte-Carlo simulation and cluster-extent threshold of 5 contiguous voxels. The different correction is likely due to the small sample size for a random-effects analysis (only 7 subjects were presented WN stimuli).

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Size ($mm^3$)</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPN $&gt;$ PT</td>
<td>3046</td>
<td>-38</td>
<td>-25</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>3934</td>
<td>47</td>
<td>-21</td>
<td>12</td>
</tr>
<tr>
<td>VOW $&gt;$ BPN</td>
<td>11928</td>
<td>-56</td>
<td>-19</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>11884</td>
<td>53</td>
<td>-17</td>
<td>2</td>
</tr>
<tr>
<td>VOW $&gt;$ WN</td>
<td>1940</td>
<td>-56</td>
<td>-15</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>265</td>
<td>56</td>
<td>-6</td>
<td>-4</td>
</tr>
<tr>
<td></td>
<td>168</td>
<td>48</td>
<td>-24</td>
<td>1</td>
</tr>
</tbody>
</table>
across subjects (Penhune et al., 1996; Rademacher et al., 2001; Sweet et al., 2005). This raises the question of whether the group analysis is representative of the functional hierarchy also in individual subjects. We therefore reanalyzed our data, localizing regions of interest (ROI) selective for the different sound classes in each subject individually, followed by a group analysis of the ROI parameters to test if the relative functional organization was consistent across subjects.

To identify the auditory hierarchy in individual subjects, we created maps similar to the group analysis of Figure 1B for each of our thirteen subjects. The statistical threshold was set independently for each subject using the same criterion as for the group (i.e., ~1800 mm$^3$ averaged for the PT condition across hemispheres). Thresholds were set via minimum t statistic, and the values for each subject are reported in Table 2. In all but one subject, the chosen t statistic corresponded to at least p(FDR) < 0.05, but for most subjects the threshold was much more restrictive. Three representative examples of these maps are shown as a series of axial sections in Figure 3. In the resulting images, bilateral activation was observed in each subject for all three experimental conditions. Visual inspection of these maps confirmed that the functional organization in all imaged subjects qualitatively matched that of Figure 1B, but significant variability in locations and size can be seen across subjects. To quantify these individual data sets, we first calculated activation volume as a function of stimulus condition and hemisphere for each subject. For those cases in which more than one cluster per hemisphere was observed, volume was recorded as the sum over all clusters within each hemisphere. The resulting values for each subject are
Table 2. Activation volume increases with increasing spectral complexity. The threshold for each subject was chosen such that the volume of their average PT activation across hemispheres was consistent with average human auditory core volumes reported in previous anatomical studies.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Threshold (min t)</th>
<th>PT-LT (mm$^3$)</th>
<th>PT-RT (mm$^3$)</th>
<th>BPN-LT (mm$^3$)</th>
<th>BPN-RT (mm$^3$)</th>
<th>VOW-LT (mm$^3$)</th>
<th>VOW-RT (mm$^3$)</th>
<th>WN-LT (mm$^3$)</th>
<th>WN-RT (mm$^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.7</td>
<td>2125</td>
<td>1487</td>
<td>5706</td>
<td>5851</td>
<td>7573</td>
<td>7300</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>2</td>
<td>8.4</td>
<td>1814</td>
<td>1802</td>
<td>5654</td>
<td>4579</td>
<td>8817</td>
<td>7189</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>3</td>
<td>4.5</td>
<td>1726</td>
<td>1924</td>
<td>5337</td>
<td>7993</td>
<td>11501</td>
<td>12284</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>6.7</td>
<td>2279</td>
<td>1444</td>
<td>4822</td>
<td>3531</td>
<td>6729</td>
<td>5899</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>5</td>
<td>9.4</td>
<td>792</td>
<td>2777</td>
<td>2189</td>
<td>4052</td>
<td>5890</td>
<td>8063</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6</td>
<td>4.4</td>
<td>2077</td>
<td>1600</td>
<td>4869</td>
<td>4861</td>
<td>14467</td>
<td>13393</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>7</td>
<td>9.9</td>
<td>3021</td>
<td>586</td>
<td>3896</td>
<td>1524</td>
<td>8302</td>
<td>2764</td>
<td>3200</td>
<td>2386</td>
</tr>
<tr>
<td>8</td>
<td>10.1</td>
<td>2747</td>
<td>862</td>
<td>3039</td>
<td>743</td>
<td>8440</td>
<td>4916</td>
<td>4199</td>
<td>3460</td>
</tr>
<tr>
<td>9</td>
<td>3.0</td>
<td>1456</td>
<td>2056</td>
<td>10329</td>
<td>13871</td>
<td>21403</td>
<td>27006</td>
<td>9898</td>
<td>12572</td>
</tr>
<tr>
<td>10</td>
<td>7.6</td>
<td>2346</td>
<td>1262</td>
<td>6585</td>
<td>3156</td>
<td>7543</td>
<td>4891</td>
<td>9137</td>
<td>5757</td>
</tr>
<tr>
<td>11</td>
<td>5.5</td>
<td>2816</td>
<td>754</td>
<td>3679</td>
<td>1140</td>
<td>9350</td>
<td>6308</td>
<td>2685</td>
<td>2327</td>
</tr>
<tr>
<td>12</td>
<td>5.2</td>
<td>2341</td>
<td>1135</td>
<td>2643</td>
<td>2653</td>
<td>6913</td>
<td>6161</td>
<td>2456</td>
<td>2958</td>
</tr>
<tr>
<td>13</td>
<td>4.4</td>
<td>2360</td>
<td>1226</td>
<td>5784</td>
<td>6923</td>
<td>9113</td>
<td>14192</td>
<td>5016</td>
<td>5398</td>
</tr>
<tr>
<td>Mean</td>
<td>6.7</td>
<td>2146</td>
<td>1455</td>
<td>4964</td>
<td>4683</td>
<td>9695</td>
<td>9259</td>
<td>5227</td>
<td>4980</td>
</tr>
<tr>
<td>Std Err</td>
<td>0.7</td>
<td>166</td>
<td>165</td>
<td>582</td>
<td>974</td>
<td>1155</td>
<td>1763</td>
<td>1159</td>
<td>1368</td>
</tr>
</tbody>
</table>
Figure 3. Functional organization in individual subjects. The three functionally differentiable regions identified in the group were also consistently identified in all thirteen subjects in each hemisphere. Three representative subjects are shown here. Thresholds varied by subject and are reported in Table 2.
reported in Table 2, and the average values are plotted in Figure 4A. Repeated-measures ANOVA indicated a significant main effect of Stimulus Condition ($p < 0.001$). Post-hoc paired t-tests confirmed highly significant ($p < 0.001$) differences in activation volumes between all three stimulation conditions. While there was no main effect of hemisphere, post-hoc paired t-tests indicated a significant ($p < 0.01$) difference between hemispheres for PT, but not for BPN or VOW. For PT, the left-hemisphere volume was greater than the right-hemisphere volume, consistent with previous anatomical (Penhune et al., 1996; Rademacher et al., 2001) and functional (Wessinger et al., 2001) results. Inspection of Table 2 also confirms that the activation volume between conditions increases from PT to BPN to VOW in both hemispheres for all subjects, with the only exception being the right hemisphere in subject 8.

Next, we quantified the spatial overlap between these activation patterns by calculating the percentage of active voxels shared between each of the stimulus conditions. Because no significant main effect of hemisphere was observed in the preceding analysis, data were collapsed across hemispheres. In each subject, we defined three ROI, one for each of the three testing conditions and containing all voxels active for the condition. Importantly, here we include all voxels responding to each condition, and do not apply any of the logical operations described for Figure 1. The matrix shown in Figure 4B reports the percentage of voxels within each of these regions of interest (indicated on the horizontal axis) that are also activated by the stimulus condition indicated on the vertical axis. Thus, the far left column of this matrix indicates that a very high
Figure 4. Quantification of single-subjects analysis. (A) Activation volume by stimulus condition across subjects. (B) Quantification of the spatial overlap between stimulus conditions. Matrix values indicate the percentage of voxels within the region of interest displayed along the horizontal axis that are also activated by the conditions shown along the vertical axis. Dark colors indicate a low percentage of overlap between conditions (i.e. few active voxels in common), while light colors indicate a high percentage of overlap. (C) Average location of active voxels for each condition across subjects (medial-lateral coordinates are reported as absolute values because both hemispheres were collapsed). * = p < 0.05, ** = p < 0.01, *** = p < 0.001.
percentage of PT-responsive voxels are also activated by both BPN and VOW stimuli, as might be predicted for auditory core. The middle column indicates that only a low percentage of BPN-responsive voxels responded to PT (bottom), as predicted for belt areas, whereas a very high percentage of these voxels were also activated by VOW stimuli (top). Finally, the far right column shows that PT and BPN stimuli activated only small subsets of VOW-responsive voxels (bottom and middle). This pattern of responses strongly supports the conclusion that these areas form a functional hierarchy: early stages are activated by both simple and complex stimuli (the latter containing simple features that can drive neurons in these early areas), whereas activation of later stages requires sufficiently complex stimuli.

Finally, we determined whether the stimulus condition had a significant effect on the location of activation in individual subjects. For each subject, we calculated the mean Talairach coordinates over all voxels activated by each stimulus condition. Absolute values were used for X-coordinate locations to account for the sign difference across hemispheres. These values are plotted in Figure 4C. Repeated-measures ANOVA with factors for Stimulus Condition, Hemisphere, and Coordinate indicated significant main effects of Stimulus Condition (p < 0.05) and Coordinate (p<0.05). No significant main effect was observed for Hemisphere nor for the interaction Condition*Hemisphere. The data were then collapsed across hemisphere, and post-hoc paired t-tests confirmed that the location of VOW activation was significantly more lateral (p < 0.05), rostral (p < 0.01) and ventral (p < 0.05) to the location of PT activation. No
significant displacement was observed between PT and BPN conditions in any direction, reflecting the fact that BPN responses were observed surrounding PT responses on all sides. These results are consistent with the assumption that a ventral auditory processing stream extends from auditory core areas of cortex via anterior auditory belt to anterior and lateral aspects of the superior temporal gyrus (i.e., rostral parabelt).

While the relative functional organization across subjects is largely consistent with the results of our group random-effects analysis, we noted a wide variation in the absolute locations of the ROI across subjects. Examples of this discrepancy are shown in Figure 5A in a series of slices for the same subjects as in Figure 3. This illustrates that the group ROI captures only a small portion of the variability of ROI across subjects. To quantify how well group-defined and individual ROI matched, we calculated for each subject the fraction of voxels in the individually defined ROI that were also present in the group ROI for the PT, BPN, and VOW conditions (Figure 5B). The resulting matrix shows an average 24% overlap between each subject’s ROI and the group ROI for PT, 51% for BPN and 56% for VOW. These results show that using a localizer-based approach when studying higher-level auditory processing in humans can provide an up to fourfold improvement in accounting for inter-individual variability.
Figure 5. Individual subjects analysis compared to group analysis. (A) In three example subjects, the mismatch between PT activation defined by the group analysis and the single-subject is apparent. (B) Fraction of voxels in single-subject ROIs that are also present in the group-defined ROI. Minimum t value was 9.8 for the group, and single-subject values are as listed in Table 2.
Organization reflects sensitivity to spectral combinations, not just increasing bandwidth

While our stimuli were chosen to increase in spectral complexity to demonstrate hierarchical processing, they also displayed a simple increase in spectral bandwidth. This leaves open the possibility that any increases in activation volume we observed for PT, BPN and VOW were simply a result of increasing bandwidth, rather than increasing spectral complexity. If this were correct, further increasing bandwidth to white noise (WN) should yield maximal activation. In a control experiment we tested this possibility by adding WN stimuli, which have maximal bandwidth but minimal spectral complexity.

In a random-effects group analysis of the seven subjects presented with WN stimuli the VOW > WN contrast was not significant at p(FRD) < 0.05. This was likely due to the small sample size. Alternatively, the VOW > WN contrasts with an uncorrected p < 0.007 and then corrected to P < 0.05 based on cluster-extent (1,000 iteration Monte-Carlo simulation of the data resulting in a minimum cluster size of 5 contiguous voxels) did yield significant activation in antero-lateral auditory cortex. The resulting cluster sizes and coordinates are listed in Table 2. These clusters are similar in location to those of the VOW > BPN contrast.

We then conducted single-subject analyses of these scans including WN, and the results are shown in Figure 6. After quantifying the volume of activation by stimulus condition in each individual subject, repeated-measures ANOVA confirmed a significant effect of stimulus condition (p < 0.001). Again we found no
Figure 6. Organization based on spectral complexity. In an additional control experiment activation by WN was compared with activation by PT, BPN and VOW. The results demonstrate that the regional differences do not merely depend on increasing bandwidth, but depend instead on spectral complexity, similar to the monkey. Activation volume across subjects (n = 7) for each stimulus condition is shown. Minimum t values for each subject are those listed in Table 2. * = p < 0.05, ** = p < 0.01, *** = p < 0.001.
significant effect of hemisphere, so data were collapsed across hemisphere. As in Figure 4A, activation volume increased from PT to BPN to VOW. By contrast, WN stimuli activated a significantly smaller volume than VOW stimuli (p < 0.05, post-hoc paired t-test). These results demonstrate that the findings above do not simply reflect a monotonic increase in preferred bandwidth from PT to WN, but more likely reflect increases in preferred spectral complexity that are relevant for the extraction or identification of particular types of sound.

E. Discussion

Building on previous work that demonstrated similar anatomical organization between macaque and human auditory cortex (Hackett et al., 2001; Sweet et al., 2005), we investigated whether common functional organization could be identified in the two species as well. In a sparse-sampling fMRI study, we presented human subjects with the same sound classes typically associated with the core, belt and parabelt regions in studies of the macaque: pure tones (PT), band-passed noise (BPN) and species-specific vocalizations (VOW). We found that PT elicited activation primarily within the caudo-medial extent of Heschl’s Gyrus. BPN bursts also activated this region, as well as adjacent rostro-medial and dorso-lateral cortex. Finally, vowel sounds (VOW) activated both of the preceding regions, as well as antero-lateral regions in the superior temporal plane (“planum temporale”), superior temporal gyrus and superior temporal sulcus. We found this organization to be robust and consistent in both group and single-subject analyses.
The activations we report are more restricted in their extent through the superior temporal lobe than have previously been reported in macaques for similar stimuli (Poremba et al., 2003). This is due to our use of statistical thresholds beyond the customarily used $p < 0.05$ (corrected for multiple comparisons). While we expected to observe the strongest responses to PT in core, it is possible for PT to still evoke (smaller) activations in other regions. This was our motivation in scaling the threshold for each subject such that their PT activation volume matched anatomical expectations.

Overall, the locations and sizes of the regions identified by our functional analyses are in good agreement with those reported for human core, belt and parabelt in recent anatomical studies (Hackett et al., 2001; Sweet et al., 2005). Additionally, we found that 75% of our PT-responsive voxels (which we define as core) overlapped with areas Te 1.0 and Te 1.1 (Morosan et al., 2001), but very few with area Te 1.2. This result adds support that this area is more consistent with the human belt, rather than core, which has been argued both anatomically (Wallace et al., 2002) and functionally (Hall et al., 2006).

Given the observed anisotropy in VOW activation, which favors anterior and lateral directions, we infer that this activation more specifically represents the rostral subdivision of the parabelt described previously in the monkey (Hackett et al., 1998). By contrast, BPN activation was more symmetrical between medial and lateral belt, consistent with findings in the macaque (Kusmierek and Rauschecker, 2009). The lack of any significant left-hemisphere lateralization for our speech stimuli (VOW) was somewhat surprising, though compatible with a
recent meta-analysis of fMRI studies using speech sounds (Turkeltaub and Coslett, 2010). The authors report equivalent bilateral activity for sublexical speech segments with left lateralization only being driven by tasks requiring explicit attention to phonology. Thus, the lack of any left lateralization for VOW stimuli in the present data may be explained by the passive listening task used.

In addition to being speech sounds, the VOW stimuli we presented were also human voices. This raises the question of the extent to which our VOW-selective responses compare to the “Temporal Voice Areas” (TVA) documented in previous results (Belin et al., 2000). As could be expected, the results of our VOW > BPN contrast are in good agreement with the locations of the TVA bilaterally (Talairach coordinates in our study [-56 -19 3] and [53 -17 2] compared with (Belin et al., 2000) [54 -13 4] and [-60 -23 6]). Additionally, the coordinates from (Belin et al., 2000) both fall within our reported parabelt region. We interpret this agreement to reflect the increased spectro-temporal complexity required to activate further antero-lateral regions of auditory cortex, with the condition that this complexity is relevant for vocal communication.

Expanding on prior work (Wessinger et al., 2001), we demonstrated a series of hierarchically related auditory cortical fields in the human. Early stages of this hierarchy are activated by both simple and complex stimuli, whereas activation of later stages requires stimuli containing sufficiently complex combinations of features. In such a scheme, neurons in primary and primary-like core areas of auditory cortex (while narrowly tuned to specific frequencies) are not necessarily exclusively responsive to tonal stimuli alone but respond to more
complex sounds as well, as long as these sounds contain energy within the neurons tuning range (Rauschecker and Tian, 2004). Including WN stimuli in some of our scans contributed an important control condition by showing that WN stimuli activate a smaller cortical volume than VOW stimuli. This demonstrated that the increases in activation volume observed were due to spectral complexity rather than a simple increase in the number and range of frequencies present. This selectivity of higher-order areas is due to non-linear integration in frequency and time (“combination sensitivity”) (Suga et al., 1979a; Margoliash and Fortune, 1992a; Rauschecker and Scott, 2009a).

In the present study, potential functional correlates of core, belt and rostral parabelt regions were identified robustly in single subjects using a relatively short duration scan (~15 minutes), making the paradigm suitable as a functional localizer for studies with other types of complex stimuli. The need for auditory functional localization methods is evident from the existing literature (Formisano et al., 2003; Petkov et al., 2006; Fedorenko et al., 2010), and highlighted by the variability in ROI size and location in individual subjects that was not well captured by the group analysis. This variability in functional location severely limits the ability of auditory fMRI studies to investigate the sound recognition process at different levels of auditory processing. Our paradigm overcomes this limitation and will enable future studies to characterize functional modules in human auditory cortex at each stage of the auditory hierarchy, a level of detail previously possible only in visual cortex. Furthermore, the approach of employing identical schemes in both humans and monkeys will aid with
establishing potential homologies between them. In the case of monkeys, however, our human speech sounds (VOW) would presumably be replaced with con-specific vocalizations (perhaps coos which exhibit spectro-temporal complexity beyond BPN similar to VOW).

The direct contrasts between conditions (i.e. BPN > PT, VOW > BPN and VOW > WN) ensure that the progression of selectivity we report is statistically significant. However, when localizing the core, belt and parabelt regions in individual subjects, we still prefer to define these based on the logical combinations shown in Figures 1 and 3. The reason is that we expect the “boundary” to have some finite width. Defining the boundary using only significantly different voxels will necessarily ignore voxels that are in between regions where responses are in transition from one sound class to the next. This method is comparable to that used to define the core/belt boundary in electrophysiology studies from our lab (Tian et al., 2001b; Kusmierek and Rauschecker, 2009), dividing regions logically based on a preference index.

Several previous studies have attempted to map auditory fields via tono- or cochleotopic organization (Formisano et al., 2003; Talavage et al., 2004; Humphries et al., 2010). While such studies have indeed been able to confirm the existence of multiple cochleotopic areas in the auditory cortex of humans, the monkey literature would suggest that these areas may correspond to subfields within one level, e.g. core (i.e. A1, R and RT) or belt (AL, ML, CL), but cannot distinguish boundaries between adjacent levels of the hierarchy. The boundaries between hierarchical levels are instead predicted to correlate with differences in
preferred bandwidth, or perhaps spectral complexity (Rauschecker et al., 1995). The work of Wessinger et al. (2001) provided early evidence for a core/belt boundary in the human. Our present results confirmed this finding, and provided a fundamentally new contribution by defining a belt/parabelt boundary. It would be useful in future studies to combine tonotopic mapping with our established method of bandwidth/complexity mapping to identify individual subregions within each level of the hierarchy.

While our results define auditory cortical fields based on varying spectral complexity, the structure and function of auditory cortex must also reflect temporal complexity. This was not tested in the present study since our stimuli were intentionally chosen to be steady-state, allowing stimulus durations to be precisely controlled across conditions. However, similar hierarchical organization reflecting temporal complexity has indeed been previously reported (Patterson et al., 2002). Combining all aspects of complex sounds, both spectrally and temporally, should deliver a complete picture of functional mapping in at least the initial stages of auditory processing in humans.
CHAPTER III

EVIDENCE FOR A “TWO-STAGE” ACCOUNT OF HUMAN PHONEME CATEGORIZATION

A. Summary

Current theories of receptive speech processing include a ventral stream responsible for speech recognition (i.e., maps sounds to their semantic meanings), and a dorsal stream that has recently been proposed to mediate sensorimotor integration (i.e., maps sounds to the motor-articulations likely to have produced them). While these theories have gained increasing support, the neural computations underlying each of these functions, particularly that of the dorsal stream, remain largely unknown. However, previous clinical work has provided useful insight showing that aural word comprehension doubly dissociates from syllable identification, suggesting that phoneme identities (i.e., categories) should be represented in the dorsal stream, and not in the ventral stream. We tested this prediction by probing the neuronal tuning properties underlying phoneme representations in the dorsal and ventral streams using a novel fMRI rapid adaptation paradigm (fMRI-RA), in which larger BOLD-signals are measured for pairs of stimuli that are increasingly dissimilar. In this paradigm, we presented subjects with pairs of “morphed” speech sounds (from one male and one female voice morphing from /da/ to /ga/), for which acoustic differences can be dissociated from category differences. We found adaptation only for same-stimulus pairs in a late stage of the ventral stream (left anterior middle
temporal gyrus), suggesting that neurons in this region are tightly tuned to the acoustic structure of the sounds. In the dorsal stream, we found two representations of phonemes. First, in left posterior superior temporal gyrus, we again found adaptation only to identical pairs of stimuli, suggesting that neurons in this region may be responsible for matching specific acoustic features to the low-level motor-articulations by which they are produced. Second, in the left premotor cortex, we found adaptation for both identical pairs, and pairs that were acoustically different but belonged to the same category. Importantly, we tested different-category pairs of two different degrees of acoustic difference and both produced the same BOLD-signal, indicating that neurons in this region are tightly tuned to the category membership of phonemes rather than their acoustic structure. These results are consistent with the proposed functions of dorsal and ventral streams given that phonemes are not mapped to a specific meaning, but are explicitly mapped to specific motor-articulations. These results provide strong support for dual-stream theories and their proposed functions of the dorsal and ventral streams, as well as extend them by providing details about plausible underlying neural computations.

B. Introduction

Current theories of human speech processing commonly describe the involvement of dual (i.e., ventral and dorsal) hierarchical processing streams (Belin and Zatorre, 2000; Scott and Johnsrude, 2003; Hickok and Poeppel, 2007;
Rauschecker and Scott, 2009b). The ventral auditory pathway is widely believed to be responsible for recognizing sounds, such as speech. The ability to map highly variable acoustic signals to invariant meanings emerges in the ventral stream as a result of hierarchical processing, in which simple features are successively combined in order to increase selectivity to complex sounds, as well as invariance to certain transformations. While evidence of this hierarchy is now widely available (Binder et al., 2000a; Rauschecker and Scott, 2009b; Leaver and Rauschecker, 2010a), the nature of the neural representation of speech sounds in the hierarchy remains poorly understood. In particular, while previous studies have shown that portions of the ventral auditory pathway respond better to phonemes than to various control sounds (for a review, see (Turkeltaub and Coslett 2010)), it is currently unknown whether this representation is based on the acoustic-features of these sounds or the phonetic categories to which they belong.

Importantly, a double dissociation has previously been demonstrated between aural word comprehension (i.e., speech recognition) and the ability to identify and discriminate syllables (Miceli et al., 1980; Baker et al., 1981). This suggests that categorical phoneme representations should be located outside the ventral auditory pathway. Interestingly, the dorsal auditory stream has been increasingly implicated in aspects of sensorimotor integration (Hickok and Poeppel, 2007; Rauschecker and Scott, 2009b; Hickok et al., 2011; Rauschecker, 2011). More specifically, it has been proposed that one function of the dorsal stream is to
estimate an “inverse model” of speech (Rauschecker and Scott, 2009b), whereby incoming speech sounds are mapped to the motor-articulations likely to have produced them. We thus hypothesized that acoustic-feature representations of phonemes would be located in the ventral stream, whereas phoneme-category representations would be located in the dorsal stream.

More specifically, premotor cortex (PMC) has often been implicated in processing phoneme categories, and is a primary target of projections from postero-dorsal auditory cortex (Rauschecker and Scott, 2009b). Functional imaging studies have reported left hemisphere PMC activation when subjects listen to speech sounds (Turkeltaub and Coslett 2010), produce speech, or both (Wilson et al., 2004; Pulvermüller et al., 2006). Additionally, temporary deactivation of this region by transcranial magnetic stimulation (TMS) disrupts syllable identification (Meister et al., 2007) and prolongs reaction times for phoneme discrimination (Sato et al., 2009). Given these results, and because the motor-articulation programs for producing phonemes differing in place of articulation are categorically different (i.e., /da/ is produced by the tongue and the roof of the mouth, whereas /ga/ is produced with the back of the throat), left PMC would seem to be a viable candidate region in which to find categorical phoneme representations.

Prior fMRI studies investigating the neural representation of phonemes have typically examined the average BOLD-contrast responses to phonemes vs. non-
speech control sounds (Liebenthal et al.; Obleser et al., 2006a; Obleser et al., 2007; Turkeltaub and Coslett 2010). However, while such comparisons do reveal phoneme representations, they are confounded with respect to inferring the underlying neuronal tuning properties. This is because both the density of selective neurons and the broadness of their tuning contribute to the average BOLD-signal level in a voxel; a given BOLD-contrast signal in a voxel could arise from a small number of strongly activated units with high selectivity, or from a large number of broadly tuned neurons with low activity. By contrast, it has been suggested that fMRI rapid adaptation (fMRI-RA) techniques are capable of probing neuronal tuning more directly and selectively (for a recent review, see (Grill-Spector, 2006)). In the fMRI-RA design two stimuli are presented sequentially in each trial, and the BOLD-contrast response to the pair is taken to reflect similarity of the neuronal activation patterns corresponding to the two individual stimuli. The lowest response is expected for two stimuli activating identical neuronal populations. The signal increases parametrically as the two stimuli activate increasingly dissimilar neuronal populations, with maximum signal expected if the two stimuli activate disjoint groups of neurons (Jiang et al., 2006).

To directly probe the neuronal tuning underlying the neural representation of phonemes, we generated a place of articulation continuum, between natural utterances of /da/ and /ga/, using a morphing system capable of finely and parametrically manipulating stimulus acoustics (Kawahara and Matsui, 2003). Similar to our previous visual experiments involving visual morphing between
categories (Freedman et al., 2001; Jiang et al., 2007), this allowed us to dissociate category selectivity, which requires neurons to respond similarly to dissimilar stimuli from the same category as well as respond differently to similar stimuli belonging to different categories (Freedman et al., 2001), from mere tuning to acoustic differences, where neuronal responses are simply a function of acoustic dissimilarity (i.e., without the sharp transition at the category boundary that is a hallmark of perceptual categorization).

Using this combination of sound morphing and fMRI-RA, we found evidence of three phoneme representations. In the ventral auditory pathway, we found only one phoneme representation, in the left anterior middle temporal gyrus. The BOLD-signal in this region was modulated by differences in the acoustic-features between paired stimuli, regardless of whether they belonged to different phonetic categories. In the dorsal stream, however, we found evidence of two different phoneme representations: one in the left posterior superior temporal gyrus that was modulated by changes in acoustic-features but not category differences, and another in left premotor cortex that was modulated only by differences between phonetic categories. Importantly, this left PMC region produced the same signal modulation for pairs of stimuli from different phonetic categories with two different degrees of acoustic difference, while producing no modulation for pairs separated by an identical degree of acoustic difference but within the same category. These results strongly support and extend current dual-stream auditory models by providing evidence that phonemes are represented categorically in the
dorsal stream, where they are likely mapped to their corresponding motor-articulations. In the ventral stream, however, phonemes are instead represented according to their acoustic structure, reflecting the fact that they have no categorically specified semantic meaning.

C. Experimental Procedures

Participants. Sixteen subjects participated in this study (7 females, 18-32 years of age). The Georgetown University Institutional Review Board approval all experimental procedures, and all subjects gave written informed consent before participating. All subjects were right-handed, reported no history of hearing problems or neurological illness, and spoke American English as their native language. One subject (female) was excluded from the imaging study based on the behavioral pretest showing no evidence of categorical phoneme perception (see below). Imaging data from one additional subject (male) was excluded from subsequent analyses due to excessive head motion, leaving a total of fourteen subjects in the imaging study.

Stimuli. Speech stimuli were recordings (44.1 kHz sampling frequency) of natural “da” and “ga” utterances (Shannon et al., 1999). Two phonetic continua (or “morphs”) were generated using the Matlab toolbox STRAIGHT (Kawahara and Matsui, 2003). Sounds were generated at 0.5% intervals between “da” and “ga” prototypes for one male and one female voice. Stimuli were additionally
generated up to 25% beyond each prototype, for a total of 301 stimuli per morphline. All stimuli were then resampled to 48 kHz, trimmed to 300 ms duration and root-mean-squared (RMS) normalized in amplitude. A linear amplitude ramp of 10 ms duration was applied to sound offsets to avoid auditory artifacts. Amplitude ramps were not applied to the onsets, however, to avoid interfering with the natural features of the consonant sound.

**Discrimination Behavior.** To identify participants’ individual category boundaries while minimizing the risk that they would covertly categorize sounds in the scanner, participants completed a discrimination test prior to scanning. Participants’ discrimination thresholds were measured at 10% intervals along each continuum, for stimuli displaced in each direction. An adaptive staircase algorithm (QUEST) from the psychophysics toolbox for Matlab (psychtoolbox) was used to adjust the difference between paired stimuli based on subject performance (Kuhl, 1981). On each trial, participants heard two sounds (each 300 ms in duration, separated by 50 ms) and were asked to report as quickly and accurately as they could whether the two sounds were exactly the same, or in any way different. A maximum time of 3000 ms was allowed for a response before the next trial started. In half the trials paired stimuli were identical and in the other half they were different. Of the different pairs, half were for displacement in one direction along the continuum, and half were in the opposing direction. The adaptive staircase algorithm adjusts its model of the psychometric function (i.e. the category boundary location) based on subject responses to
different pairs only (same pairs do not contribute), making differences larger if the subject answers incorrectly and smaller if the subject answers correctly. Subjects completed 28 trials per condition, for 20 conditions (10% intervals from 0-90% with displacements toward 100%, and 10% intervals from 10-100% with displacement toward 0%), yielding 560 total trials.

**Identification Behavior.** After subjects completed the fMRI experiment, they were asked to categorize the auditory stimuli along each continuum to confirm the location of participants' category boundary and to measure the sharpness of their category boundary. Categorization was tested at 10% intervals from -25% (i.e. 25% past “da” away from “ga”) to 125% (25% past “ga” away from “da”). On each trial, subjects were presented with a single sound and given up to 3 seconds to indicate as quickly and accurately as they could whether they heard “da” or “ga.” Subjects completed 20 trials per condition for 15 conditions, for a total of 300 trials per morph. We then fit the resulting data with a sigmoid function to estimate the boundary location as well as boundary sharpness for each subject. The sigmoid was calculated as

\[
f(x) = \frac{1}{1 + \exp\left(-\frac{(x-\alpha)}{\beta}\right)},
\]

where \(x\) is the location along the morphline, \(\alpha\) is the location of the boundary along the morphline and \(\beta\) is the steepness of the boundary (smaller values of \(\beta\) result in sharper boundaries).
**Event-related fMRI-RA experiment.** To observe responses largely independent of overt phonemic processing, we scanned subjects while they performed an attention-demanding distractor task. Auditory stimuli were presented using Presentation (Neurobehavioral Systems) via STAX electrostatic earphones at a comfortable listening volume (~65-70 dB) worn inside ear protection (Bilsom Thunder T1) giving ~26 dB attenuation. In each trial, subjects heard pairs of sounds presented sequentially (each 300 ms duration), separated by 50 ms. The delay of 50 ms was chosen based on existing literature using similar paradigms (Joa nisse et al., 2007a, Myers et al., 2009). Each of these sounds persisted slightly longer in one ear than in the other (i.e. 10 ms offset ramp applied to each channel, offset by ~30 ms between channels). The subject was asked to listen for these offsets, and to report whether the two sounds persisted longer in either the same or different ears. Subjects held two response buttons and the letters “S” and “D,” indicating “same” and “different” respectively, were presented simultaneously on opposing sides of a computer screen to indicate which button to press. Their order was alternated on each run, to disentangle activation due to categorical decisions from categorical motor activity (i.e. to average out the motor responses). Images were collected for 6 runs, each run lasting 669 s. Trials lasted 12 s each, yielding 4 volumes per trial, and there were two silent trials (i.e. 8 images) at the beginning and end of each run. The first 4 images of each run were discarded and analyses were performed on the other 50 trials—10 each of the five different conditions defined by the change of shape and category.
between the two cars: M0, same category and same shape; M3\textsubscript{within}, same category and 33% shape change; M3\textsubscript{between}, different category and 33% shape change; M6, different category and 67% shape change; and null trials (Fig. X). Trial order was randomized and counterbalanced using M sequences (Buracas and Boynton, 2002), and the number of presentations was equalized for all stimuli in each experiment.

**fMRI Data Acquisition.** All MRI data were acquired at the Center for Functional and Molecular Imaging at Georgetown University on a 3.0-Tesla Siemens Trio Scanner using whole-head echo-planar imaging (EPI) sequences (Flip Angle = 90°, TE = 30 ms, FOV = 205, 64x64 matrix) with a 12-channel head coil. In both fMRI-adaptation experiments, a clustered acquisition paradigm (TR = 3000 ms, TA = 1500 ms) was used such that each image was followed by an equal duration of silence before the next image was acquired. Stimuli were presented after every fourth volume, yielding a trial time of 12000 ms. In all functional scans, 28 axial slices were acquired in descending order (thickness = 3.5 mm, 0.5 mm gap; in-plane resolution = 3.0 X 3.0mm\(^2\)). Following functional scans, high-resolution (1x1x1 mm\(^3\)) anatomical images (MPRAGE) were acquired.

**fMRI Data Analysis.** Data were analyzed using the software package SPM2 (http://www.fil.ion.ucl.ac.uk/spm/software/spm2/). After discarding images from the first twelve seconds of each functional run, EPI images were temporally
corrected to the middle slice, spatially realigned resliced to 2 x 2 x 2 mm³, and normalized to a standard MNI reference brain in Talairach space. Images were then smoothed using an isotropic 6 mm Gaussian kernel.

For whole-brain analyses, a high-pass filter (1/128 Hz) was applied to the data. We then modeled fMRI responses with a design matrix comprising the onset of predefined non-null trial types (M0, M3within, M3between and M6), six movement parameters and the global mean as regressors using a standard canonical hemodynamic response function (HRF). The parameter estimates of the HRF for each regressor were calculated for each voxel. The contrasts for each trial type against baseline at the single-subject level were computed and entered into a second-level model (ANOVA) in SPM2 (participants as random effects) with additional smoothing (6 mm). For all whole-brain analyses, a threshold of at least p < 0.001 (uncorrected) and at least 20 contiguous voxels was used unless otherwise mentioned.

D. Results

Behavior Experiments

We located each subjects’ category boundary for the two morphs using two behavior experiments. To minimize the likelihood of subjects covertly categorizing the sounds during the scans, we did not have them overtly
categorize the sounds before scanning. Instead, subjects were asked to report whether pairs of sounds were exactly the same, or in any way different. Discrimination was tested at ten percent intervals along each morphline, and the morph-distance between paired sounds was varied according to an adaptive staircase design in each direction of the morphline, similar to a design used in the chinchilla (Kuhl, 1981). This allowed us to measure the just-noticeable difference (JND) at each location (for each morph direction), which should have minimum value at the category boundary. In all but one subject, we observed clear minima in the JND for morph-differences in each direction. The subject with no detectable boundary was excluded from further participation and analyses. The boundary was inferred to be halfway between the two smallest JND measurements within 30% of the suspected boundary location in each direction. After scanning, subjects participated in a second behavior experiment where they were asked to overtly categorize each sound in order to explicitly confirm the location of the category boundary. Average discrimination and categorization results for the group are shown in Figure 7, indicating a clear correspondence between the JND minima and the explicit category boundary. There was only one case in which the measured category boundary did not align correctly with the
Figure 7. Behavioral performance outside the scanner. Before participating in scans, subjects first participated in a behavioral discrimination task. The task served both as an initial screen to ensure subjects exhibited categorical perception for our stimuli, and to locate the perceptual category boundary. At ten percent intervals along each morphline (one male voice, one female voice), we measured the just-noticeable-difference (JND) in each direction, i.e. toward “da” (blue) and toward “ga” (red). The halfway point between the minima for these two curves accurately predicts the category boundary, measured directly after scanning (black). Given that our morphlines have finite extents (25% past each endpoint, or 150% total), the green lines mark the maximum possible difference at each morphstep. Thus the fact that, in each direction, the JND decreases beyond the category boundary is an artifact that simply reflects the maximum possible difference for those steps.
JND minima (one subject for the male morphline, boundary shifted toward /ga/ ~10% from prediction of JND minima).

During the scans, subjects were asked to do an attentionally-demanding direction-matching task. In each imaging trial, each of the paired sounds persisted slightly longer (~30ms) in one ear than in the other. Subjects were asked to report whether the two sounds persisted longer in the same ear or different ears. The nature of this task requires that subjects listen closely to all sounds presented, but is largely independent of the category nature of the sounds. The average performance across subjects was 71%, indicating that the task was attention-demanding, which minimizes the chance that subjects covertly categorized the stimuli in addition to doing the direction task.

fMRI-RA Experiment

To probe neuronal selectivity using fMRI, we adopted an event-related fMRI-RA paradigm (Kourtzi and Kanwisher, 2001; Jiang et al., 2006) in which a pair of sounds of varying acoustic similarity was presented in each trial. The fMRI-RA approach is motivated by findings from IT monkey electrophysiology experiments reporting that when pairs of stimuli were presented sequentially, the second stimulus evokes a smaller neural response that the first (Miller et al., 1993; Lueschow et al., 1994). Additionally, it has been suggested that the degree of adaptation depends on stimulus similarity, with repetitions of the same stimulus
causing the greatest suppression. In the fMRI version of this experiment, the BOLD-contrast response to a pair of stimuli presented in rapid succession was measured for pairs differing in specific perceptual aspects (e.g., viewpoint or shape), and the combined response level was assumed to predict stimulus representational dissimilarity at the neural level (Murray and Wojciulik, 2004; Grill-Spector, 2006). Indeed, we (Jiang et al., 2006; Jiang et al., 2007) and others (Murray and Wojciulik, 2004; Fang et al., 2007; Gilaie-Dotan and Malach, 2007) have recently provided evidence that parametric variations in shape, orientation, or viewpoint, stimulus parameters putatively associated with neuronal tuning properties in specific visual brain areas, are reflected in systematic modulations of the BOLD-contrast response. This suggests that fMRI-RA can be used as an indirect measure of neural population tuning (Grill-Spector, 2006).

It follows that if a region of the auditory system exhibits sharp neuronal selectivity to speech sounds, then the overlap of neuronal activations caused by two sequentially presented speech sounds differing by a fixed amount of acoustic change would be equal, regardless of whether the paired sounds belong to the same or different categories, resulting in both pairs evoking similarly higher combined BOLD-contrast responses over pairs of identical stimuli in speech-selective regions. Previous studies have suggested that the left anterior superior temporal gyrus plays a central role in representing speech sounds (Binder et al., 2000a; Scott et al., 2000b; Liebenthal et al., 2005) and we therefore hypothesized that acoustic selectivity for phonemes should be detected in this
area. We scanned participants using an event-related fMRI-RA paradigm with an offset delay matching task (see Experimental Methods) for which category pairing was irrelevant, thus avoiding potentially confounding influences due to difference in task difficulty across pairing conditions and other potential confounds caused by top-down effects of the task itself (Grady et al., 1996; Sunaert et al., 2000; Freedman et al., 2003).

In this fMRI-RA paradigm, we presented subjects with pairs of stimuli that varied in acoustic and categorical dissimilarity. We used an auditory morphing technique (Kawahara and Matsui, 2003) to create two phonetic continua (one male voice and one female voice) between two natural sound recordings (/da/ and /ga/). This approach allowed us to dissociate category selectivity, which requires neurons to respond similarly to dissimilar stimuli from the same category as well as respond differently to similar stimuli belonging to different categories (Freedman et al., 2003), from mere tuning to acoustic differences, where neuronal responses are a function of acoustic dissimilarity, without the sharp transition at the category boundary that is a hallmark of perceptual categorization. An example of the male morph from /da/ to /ga/ is shown in Figure 8.

We then created pairs of identical sounds (condition M0) and pairs of sounds differing by 33.33% acoustic change (where 100% corresponds to the difference between the original /da/ and /ga/ utterances), with both sounds in a pair either
belonging to the same category, M3\textsubscript{within}, or to different categories, M3\textsubscript{between} (Figure 8). This made it possible to attribute possible signal differences between M3\textsubscript{within} and M3\textsubscript{between} to an explicit representation of the phonetic categories. Morphlines were also extended beyond the prototypes (25% in each direction) so that the actual stimuli used to create the stimulus pairs for each subject would span 100% of the difference between /da/ and /ga/, but could be shifted so that they centered on the category boundary for each subject, measured behaviorally before scanning (see Behavior Experiments above).

Using parametrically morphed stimuli also allowed us to identify ROI using contrasts between stimuli that are well controlled in terms of their low-level features. This avoids the problem inherent with contrasts against “control” stimuli, which are difficult to define, and to control against categorical differences in low-level features between categories, and can greatly affect ROI location (Turkeltaub and Coslett, 2010). To locate cortical regions sensitive to acoustic differences regardless of category label, we examined the results of the contrast M6 + M3\textsubscript{between} + M3\textsubscript{within} > M0 (i.e. all “different-sound” conditions vs. “same-sound” condition). This contrast yields two significant clusters (Figure 9), both in the left hemisphere, located in the anterior middle temporal gyrus (MNI coordinate [-56 -6 -22]), and the posterior superior temporal gyrus (MNI coordinate [-56 -50 8]). We then extracted the percent signal change for each condition from each of these clusters. Post-hoc paired t-tests (single tailed) show that M3\textsubscript{within}, M3\textsubscript{between} and M6 conditions are each significantly higher than the
Figure 8: Selecting stimuli for fMRI conditions. Pairing conditions used in the scanner. Pairs are arranged to align with the category boundary as measured via discrimination for each individual subject (Figure 7), shown here at the 50% morph point between the two prototypes. M0 pairs are always constructed using the endpoint stimuli in order to counterbalance the total number of times each of the four stimuli are presented throughout the duration of the experiment.
Figure 9: Neural selectivity for acoustic features in left auditory cortex. Whole-brain contrasts of all different-sound pairs with greater signal than same-sound pairs (M6 + M3\textsubscript{between} + M3\textsubscript{within} > M0), masked by expected adaption (M6 > M0), reveals only two significant clusters (p < 0.05 corrected for cluster-extent): left anterior middle temporal gyrus (MNI -56 -6 -22) and left posterior superior temporal gyrus (MNI -56 -50 8), shown in the middle in red, on a standard reconstructed cortical surface. Extracting percent signal change for each testing condition in these clusters shows full release from adaptation for all different-sound pairs (i.e. M3\textsubscript{within}, M3\textsubscript{between} and M6, shown to the right in blue bar graphs). Asterisks indicate significant differences from the M0 condition (p < 0.001 or better, post-hoc paired t-test).
M0 condition in both the aMTG and pSTG. That all three conditions containing acoustic differences are released from adaptation suggests that neurons in these locations are not only selective for the acoustic features of the incoming sounds, but that they are also quite sharply tuned such that sounds only 33% different share little overlap in their neural representation.

**Category selectivity in left PMC**

We then wanted to locate cortical regions selective for the phoneme categories, which we did using a M3\_between > M3\_within contrast. This contrast yielded only one significant cluster (Figure 10), in the left middle frontal gyrus (MNI coordinate [-5 4 44], BA 6). Post-hoc paired t-tests (single tailed) showed that M3\_between and M6 conditions were each significantly higher than both the M0 and M3\_within conditions, with no significant differences between M0 and M3\_within nor between M3\_between and M6. This result suggests that neurons in this cluster are selective for the phonetic categories of the incoming sounds, and release fully from adaptation for pairs in which the sounds belong to different categories, with no additional signal increase for between-category conditions with greater acoustic differences (i.e., no significant difference between M3\_between and M6 conditions).

**Functional connection between pST and PMC**

To determine whether the category signals observed in PMC were the result of ventral or dorsal stream processing, we examined the functional connectivity between each of the acoustically tuned ROI detected in auditory cortex (i.e.,
Figure 10: Neural selectivity for category labels in left premotor cortex. Whole-brain contrasts of categorically different pairs with greater signal than categorically identical pairs with equal acoustic difference ($M_{3\text{between}} > M_{3\text{within}}$), masked by expected adaption ($M_6 > M_0$), reveals only one significant cluster ($P < 0.05$ corrected for cluster-extent): left middle frontal gyrus (shown to the left in red, on a standard reconstructed cortical surface). Extracting percent signal change for each testing condition in these clusters shows no release from adaptation for same-category pairs (i.e. $M_0$ and $M_{3\text{within}}$, shown to the right in blue bar graphs) and full release for different-category pairs (i.e. $M_{3\text{between}}$ and $M_6$). Asterisks indicate significant differences from the $M_0$ condition ($p < 0.05$ or better, post-hoc paired t-test). MNI coordinates for the cluster are listed above the bar graph.
For this analysis, we used the group defined aMTG and pSTG clusters, and the individually defined PMC clusters from the preceding analysis. We then extracted the measured BOLD time course in each of these ROI, and looked for correlations. We detected a significant correlation between aMTG and pSTG \((r = 0.4, p<0.04)\) and between pSTG and PMC \((r = 0.3, p<0.044)\). A strong trend between aMTG and PMC was detected, but did not reach significance \((r = 0.17, p<0.07)\). We further found the functional correlation between pSTG and PMC to be significantly stronger \((p<0.0006, \text{two tailed t-test})\) than the correlation between aMTG and PMC. These results suggest that the PMC category-selectivity signals are located at a high-level of the dorsal stream rather than the ventral stream.

**PMC signals predict behavior**

We also determined whether the category selectivity observed in left premotor cortex was correlated to subject behavior. More specifically, we predicted that subjects with larger differences between \(M3_{\text{between}}\) and \(M3_{\text{within}}\) signals would have better categorization performance (sharper category boundaries). For each individual subject, we searched the \(M6+M3_{\text{between}} > M3_{\text{within}}+M0\) contrast map and identified activity peaks within a 20mm radius of the group-defined PMC cluster. In cases where multiple peaks were detected, the one closest to the group-defined PMC cluster was chosen. We then centered 5mm spherical ROI on each of these coordinates and extracted the percent signal change for each testing
condition. **Figure 11** shows the signal difference $M_{3\text{between}} - M_{3\text{within}}$ (in percent signal change) measured in these ROI against the sharpness of the behavioral category boundary (beta value resulting from sigmoid fitting to behavioral categorization performance) for each subject. Multiple linear regression of these data confirmed a significant correlation between these measures ($r = 0.63$, $p<0.016$). This result strongly implicates left PMC as playing a direct role in perceptual category processing. This result is particularly noteworthy, as subjects were not asked to categorize the sounds but rather were actively distracted from the category nature of the sound pairs by an attention-demanding direction-matching task.

### E. Discussion

Our study used a sound-morphing method to create stimulus pairs for which acoustic differences could be disentangled from category differences, and an fMRI-RA paradigm designed to probe neural selectivity more directly than traditional methods based on BOLD-contrasts. We found no evidence of phoneme category representations in the ventral stream, and instead find only a representation of their acoustic-features in the left anterior middle temporal gyrus. Similar results have been reported by others (Frye et al., 2007; Toscano et al., 2010), supporting the idea that this region maintains a representation of the phonetic continuum with neurons tightly tuned to complex combinations of acoustic features but independent of the phonetic category.
Figure 11: Signal in the left premotor ROI. PMC ROI were defined for each subject, by detecting the peak of the M6 + M3\textsubscript{between} > M3\textsubscript{within} + M0 contrast, within 20 mm radius of the group-defined PMC cluster. We then calculated the BOLD-contrast response difference between the M3\textsubscript{within} and M3\textsubscript{between} conditions for each morph line and subject (y axis) and plotted this index against the sharpness of the category boundary measured behaviorally outside the scanner (x axis). Data are shown for all participants (n = 14) and the regression line (r = 0.63, p<0.016).
In contrast to our results in the ventral stream, in the dorsal stream we find evidence of acoustic selectivity in posterior superior temporal gyrus as well as category selectivity in left premotor cortex. This is consistent with the suspected role of the dorsal stream in mapping sounds to motor-articulations (i.e., calculating an “inverse” model). As opposed to the ventral stream where there is no categorical meaning onto which /da/ and /ga/ should be separately mapped, the motor-articulations that produce them are categorically different and are likely represented in PMC. This provides the basis for a mechanistic account of the classic effect of categorical perception, since we show that the category signal present in PMC does in fact influence subjects’ ability to categorize the stimuli, which has also been reported previously (Dufor et al., 2009), and altered phoneme discrimination resulting from TMS applied to premotor cortex has also been reported (Meister et al., 2007; Sato et al., 2009). How and where the two streams interact to produce behavior, however, remains an open question.

Furthermore, category selectivity was observed in PMC despite subjects never having been informed of, and actively being distracted from, the category nature of the sound pairs. This suggests that this category processing is highly automatic, being performed regardless of whether or not the resulting category-label information is task-relevant. A recent meta-analysis of sublexical speech perception (Turkeltaub and Coslett, 2010) reports that the only region responding more to speech than control sounds, outside of auditory cortex, is in the left ventral PMC. The location of this cluster is in good agreement with the PMC
cluster identified in the present study. The appearance of this cluster without any explicit task also supports the notion of highly automatic processing of speech in the PMC.

Finally, we also observe stronger functional connectivity between the pST and PMC than between the aMTG and PMC. When added to existing theories and anatomical evidence, we conclude that the category selectivity observed in PMC reflects the endpoint of processing in the dorsal stream as opposed to the ventral stream. A limitation is that our connectivity results do not imply direction, and fMRI does not have sufficient temporal resolution to determine the direction in which information flows between pST and PMC. A recent ERP study (De Lucia et al., 2010), however, found selectivity for conspecific vocalizations in humans 169-219 ms after sound onset, but not until 291-357 ms in a prefrontal/precentral gyrus cluster ([−53 −3 40]) very near our PMC cluster ([−54 4 44]).

Several previous studies, similar in design to our own, have reported conflicting results concerning the neural representation of phoneme categories. One of these studies (Myers et al., 2009) reports findings that are largely consistent with ours, including sensitivity to acoustic differences in left anterior superior temporal gyrus, and category selectivity in left inferior frontal sulcus. In contrast, another recent study reported category selectivity for phonemes in the left anterior superior temporal sulcus and middle temporal gyrus (Joanisse et al., 2007a). A
likely cause of this discrepancy is differences in the total number of times stimuli were presented. Whereas our study ensured that all stimuli are presented an equal number of times throughout the scans, the study from Joanisse et. al. presented the stimuli used for their same-stimulus and within-category conditions many more times than one of the stimuli used in the between-category condition. By being presented less frequently throughout the scan, the stimuli for the between-category condition would have been adapted less than the others and thus likely evoked a disproportionately large response, confounding any possible adaptation interpretation.

Another study reported “selective amplification” of different-category phoneme pairs in several regions throughout the brain. This study focused largely on left supramarginal gyrus but also included bilateral prefrontal regions and the anterior cingulate gyrus (Raizada and Poldrack, 2007). However, this study also presented stimuli a different total number of times, with those near the center of the continuum being presented more frequently than those near the ends. The effect of this over-adaptation is indeed visible in their responses to same-stimulus pairs across the continuum, with responses near the middle of the continuum being lower than near the ends (i.e., the response is “U-shaped”). Because “amplification” was determined by comparing different phoneme pairs with same-stimulus pairs from the same region of the continuum, this would have the effect of decreasing the baseline (i.e., signal for the same-stimulus conditions), causing the signal to the different-stimulus pairs to appear as though it was “amplified.”
A. Summary of goals

In Chapter I, I argued, as others have (Holt and Lotto, 2010) that speech recognition in humans fits well into the more general framework of perceptual categorization. I would propose here that the same argument should apply not only to speech but also to auditory recognition in general, including other aspects of vocal sounds (e.g., voices, affect, etc.) as well as environmental sounds. I then also discussed a theoretical and computational “two-stage” model that has been shown to solve problems of perceptual categorization (Riesenhuber and Poggio, 2000), using two separate stages of processing

3. a stimulus-selective, hierarchical neural representation
4. task representations separate from the sensory representation

While this model has been primarily developed and applied in the visual domain, the underlying computations are general in nature and thus hold the potential to generalize to other sensory domains, such as audition. And while existing models of speech processing qualitatively account for a diverse collection of evidence across species (Rauschecker and Scott, 2009b) and across normal and impaired humans (Hickok and Poeppel, 2007), our understanding of the underlying neural computations remains limited. Given that many similarities exist between current models of auditory and visual processing, it is reasonable to ask whether the two systems employ similar neural computations to solve analogous problems. The
experiments in Chapters II and III set out to test whether the auditory system exhibits evidence of performing these two types of computations proposed in the quantitative model from vision.

**B. Summary of results**

In Chapter II, we demonstrated a series of hierarchically related auditory cortical fields in the human. Early stages of this hierarchy (i.e., core) are activated by both simple and complex stimuli, whereas later stages (i.e., belt and parabelt) respond only to stimuli containing sufficiently complex combinations of features. It is in these later stages that we expected to find stimulus-selective responses. As argued in Chapter III, however, simple BOLD contrasts are a poor indicator of neuronal selectivity, preventing this aspect from being investigated in Chapter II. This question was thus taken up by the experiments in Chapter III using the fMRI-RA technique, which is thought to probe neuronal tuning more directly than traditional BOLD contrasts. In left aMTG, we observed significant adaptation only for pairs of identical stimuli (taken from a phonetic continuum from /da/ to /ga/), and similarly higher responses to all other pairing conditions, suggesting that the different stimuli evoked responses in few if any of the same underlying neurons.

The fMRI-RA experiment in Chapter III also found the same pattern of adaptation responses in the left pSTG, with no significant adaptation for any pairing in which the stimuli were not identical. However, given the suggested role of this region in mediating aspects of sensorimotor interface (Hickok et al., 2011; Rauschecker,
2011), we believe the underlying computations in this region are quite different from those in aMTG. While responses in aMTG likely reflect selectivity directly to the acoustic features present in the stimulus, we interpret the responses in pSTG to reflect the process of calculating an “inverse model” of the speech sounds (i.e., mapping individual acoustic features to specific motor-articulations most likely to have caused them). This interpretation is further supported by the observation of category specific responses in left PMC, to which the left pSTG was functionally connected. These categorical responses in left PMC likely reflect categorically different high-level (i.e., abstract) /da/ and /ga/ motor-articulation representations, independent of low-level variations that might be introduced to compensate for the present state of the articulators (e.g., saying /da/ normally and with gum in your mouth require slightly different end-articulations).

C. Interpretation of results

Given that the responses observed in Chapter III were evoked automatically, in the absence of an explicit categorization task, it is awkward to describe the responses in left PMC as consistent with a “task” representation. It is perhaps helpful to repose the two-stage computational scheme as

1. what was observed
2. what it meant

In the case of our dorsal stream results, the pSTG responses provide the sensorimotor mapping at a low-level (i.e., high detail), whereas the category responses in PMC provide the ultimate meaning (i.e., an articulation produced
with the tongue or the back of the throat). This interpretation applies to the results in the ventral stream as well, in that there is a detailed representation of the sound heard (in aMTG) but no categorical representation in the ventral stream. This makes sense in many ways, since there is no explicit meaning associated with either /da/ or /ga/ alone (other than the articulation mapped in PMC). Indeed, others have argued that categorical recognition of individual phonemes may not be necessary, and may even be detrimental, for recognizing whole words (Holt and Lotto, 2010). This is because the actual sound of a phoneme may vary depending on what comes before and/or after it in the speech stream (i.e., co-articulation). Thus, categorical representations would then only be expected in the ventral stream when the sounds being processed have categorical meanings associated with them. One example would be when subjects do explicitly categorize the phonemes, in which case activation in left IFG has often been reported (Binder et al., 2004; Blumstein et al.; Husain et al., 2006). Another would be the use of real words, such as “date” and “gate.” While the acoustic details of these utterances would still be expected near the aMTG (or possibly more anterior), categorical representations may also be elicited in prefrontal regions reflecting their categorically different semantic meanings (Martin 2007).

Taken together, these studies provide support for the idea that the auditory system does in fact utilize a two-stage computational architecture to solve sound recognition problems. These results most directly support this interpretation in the dorsal stream, which has also recently been shown to be compatible with
two-stage processing in the visual stream (Freedman and Assad, 2011). The computational appeal of this architecture is that it allows a single perceptual representation to support many different tasks. Take, for example, the case of the words “hid” and “head” spoken by two voices. One task circuit may categorize “what” was said, while another may identify “who” was speaking. Whereas biasing the perceptual representation to facilitate one task would interfere with the other, the two-stage architecture successfully avoids interference by keeping these task processing separated from the sensory representation.

**D. Hierarchical vs. ‘modular’ processing**

The stimulus representation in this model is organized in a functional hierarchy. Ascending from early to later stages in the hierarchy, simple and localized receptive fields are alternately combined and then pooled over to gradually build specificity for complex stimuli and invariance to their simple transformation (e.g., change in location or size). Importantly, the same generic operations are proposed to underlie recognition of objects in general, as opposed to a more ‘modular’ framework in which specific cortical regions perform computations specialized for recognizing specific object classes. The later case has been variously proposed to explain processing in the well-known fusiform face area (FFA) in the ventral visual pathway as being specific to faces (Sergent et al., 1992; Kanwisher et al., 1997) or object of ‘expertise’ (Gauthier et al., 1999). Even amongst common claims of face specificity, the representation in FFA has been
posed in various ways, such as being identity-based (Rotshtein et al., 2005) or based on differences from a “mean” face (Loffler et al., 2005). However, in a recent study using parametrically morphed faces and fMRI-RA, it was shown that the neural selectivity in FFA was compatible with a simple shape-based representation for faces, as predicted by a two-stage computational model (Jiang et al., 2006). Evidence of sensitivity to facial expressions has also been detected in FFA (Fox et al., 2009), further suggesting this region is selective to physical features rather than to identity.

In auditory cortex, recent work has pointed to specific regions of aST being selective for specific sound classes (Leaver and Rauschecker, 2010c), including voices (Belin and Zatorre, 2003; Formisano et al., 2008; Petkov et al., 2008a). Arguments have already been posed proposing the temporal voice area as a specialized voice-processing module. Indeed, there has been an active effort to explicitly tie the studies of the voice area to previous studies of FFA, describing the voice as an “auditory face” (Belin et al., 2004), exploring the question of expertise (Chartrand et al., 2008) and even reporting a “voice inversion” effect similar to the “face inversion” effect (Bédard and Belin, 2004). These results beg the question “is processing in this area specialized for voices or can it be accounted for via more general computations?” Given that the results of Chapters II and III support the use of two-stage recognition processing similar to that in visual recognition, it is possible that neuronal selectivity in the temporal voice area is compatible with a continuous acoustic-feature based
representation. In this case, similar populations of neurons in the voice area would represent similar voices, while neurons tuned to specific voice identities would instead likely be found separately in inferior frontal regions. Similar predictions might also apply to recognizing the emotional content of vocalizations, which has been reported to elicit activation in regions of right aST similar to those reported for voices, as well as increased signal in right prefrontal regions when attention is directed to emotional content (Belin et al., 2004).

E. What types of invariance should we expect?

Recordings in macaque inferotemporal cortex (IT) provided evidence of two response properties that are the hallmark of higher-order sensory neurons: they respond to complex, but not simple, stimuli and these responses are robust to changes in the location and size of the stimulus (Bruce et al., 1981). In contrast, these neurons also display little invariance to rotations in depth (Logothetis et al., 1995). Such neurons have been described computationally as “view-tuned units” (Riesenhuber and Poggio, 1999b). The responses of view-tuned units reflect the fact that, computationally, invariance to certain transformations of a stimulus can be estimated from only one view, whereas learning invariance to other transformations (such as rotation in depth) requires multiple views (Riesenhuber and Poggio, 2000). This illustrates why neurons at different levels of the processing hierarchy often exhibit different types or degrees of invariance: their responses depend on the input information they receive and what they are able
to do with that information.

An interesting question in the auditory domain is then, “what types of invariance do we expect auditory neurons to exhibit, and where should we find them?” A good starting point is to look at what types of invariance have been found in the acoustic signals themselves. A now classic study (Peterson, 1952) reported that the identity of vowel sounds could largely be determined by plotting the relationship between their first and second formants (i.e., spectral peaks). Since the formant relationships still provide information about the intended phoneme independent of the pitch of the sound, we could reasonably expect to find representations that are invariant to changes in pitch (even if only for a finite range). Evidence for pitch invariance of phoneme processing has indeed been reported by MEG studies using a mismatch paradigm (Aulanko et al., 1993). That pitch seems to be processed in specific locations of auditory cortex may also support the idea that it can be separated from the rest of the speech signal (Bendor and Wang, 2005; Griffiths et al., 2010).

What about higher-level invariance, then, such as speaker identity? For example, if the representation in TVA simply reflects acoustic-features, then we should expect to find representation of identity elsewhere. Computationally, identification is not a qualitatively different computation from categorization, but reflects merely a quantitative difference in the types of specificity and invariance needed (Riesenhuber and Poggio, 2000). Given the previous evidence of category
selectivity in prefrontal cortex, by extension we might also expect to find identity selective neurons in this region, though some existing evidence points to anterior IT regions (Kriegeskorte et al., 2007). A further question would be whether or not representations of identity are restricted to a single modality or whether they are multimodal. Little is presently known about modality-specific vs. general identity representations and how multiple sources of information interact, in particular in a synergistic way. Computationally speaking, if the task units are separated from sensory representations, then there is little reason why they shouldn’t also be capable of utilizing information from multiple sensory representations. Such multimodal task neurons might even provide a potential mechanism of the well-known McGurk illusion (McGurk and MacDonald, 1976), where subjects watching a video of a face articulating /ga/ where the audio is replaced with the utterance /ba/ report hearing /da/.

**F. Transfer of learning**

The primary computational strength of the two-stage model is the ability to do multiple tasks using a common sensory representation. An attractive feature of this architecture however is that it allows sensory learning to transfer across multiple tasks. The idea is that training on one type of task leads to sharpening of selectivity in neurons in the sensory representation. When this same representation is then utilized to support another task, it too benefits from the higher fidelity. For example, in a recent study (Jiang et al., 2007), subjects were trained to categorize a continuum of morphed car images. Before and after
categorization training, subjects were also asked to discriminate between similar car images from these continua. The results showed that subjects improved at discrimination after they had learned the categorization task. Similar notions have been proposed in the auditory domain, where trained musicians have been found to outperform non-musicians on tasks concerning speech prosody (Schön et al., 2004; Thompson et al., 2004; Magne et al., 2006) and detection of pitch (Marques et al., 2007).

This idea of transferring learning from one task to another hinges on the different task circuits sharing a common underlying representation. However, another interesting computational question arises in the converse situation where task units draw input from multiple sensory representations. In such a circuit, it would be reasonable to ask whether or not learning in one sensory representation can be transferred to another through shared multimodal task units. For example, is it possible that learning to identify a person’s face under increasingly diverse viewing conditions might ultimately lead to an improved ability to discriminate their voice from other similar voices? Or perhaps watching silent movies of someone speaking would lead to better recognition of their voice? Cross-modal transfer of learning for recognition has indeed been reported (Harman et al., 1999; Newell et al., 2001; Ernst and Newell, 2007). The voice is also known to carry significant information about various physical traits of the speaker such as gender and approximate age, and perhaps other physical characteristics (Belin et al., 2004). It is possible that by, say, watching silent movies of a person speaking
the multimodal representation of the individual’s physical traits could be refined, and thus lead to more accurate estimations of their likely vocalizations.
Literature Cited


Peterson G (1952) Control methods used in a study of the vowels. Journal of the Acoustical Society of ....


