ON SPEECH, MUSIC, AND AUDIOMOTOR INTERACTIONS

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

Brannon M. Green, M.A.

Washington, DC
January 15, 2016
ON SPEECH, MUSIC, AND AUDIOMOTOR INTERACTIONS

Brannon M. Green, M.A.

Thesis Advisor: Josef P. Rauschecker, Ph.D.

ABSTRACT

Speech and music production require repeated and accurate coordination between actions and sounds. This tightly regulated set of audiomotor interactions seems to persist for receptive speech and music perception as well. Various models for auditory processing exist that delineate how the brain processes different kinds of sounds, and roles for these auditory areas have been suggested, but they have yet to be well characterized. Similarly, as speech and music sounds are repeatedly presented they must be learned and stored somewhere, which must invariably draw upon some form of memory. Yet, memory systems have been largely absent from the models of auditory processing. This dissertation seeks to further validate the dual-stream model of auditory perception while proposing functional and anatomical additions to this model in the form of motor and memory systems. The first study investigates the location and timing of areas exhibiting categorical speech representations. Categorical tuning was found in areas of the inferior frontal gyrus, insula, and premotor strip, suggesting a role for motor areas in the categorical speech perception. The second study investigates how the brain processes novel and familiar sequences of sounds. The basal ganglia were shown to play an early role in novel sequence encoding, while the medial temporal lobe seemed to play an intermediate role as sequences were repeatedly presented until finally eliciting activation in prefrontal areas. This work provides insight into how the human auditory and motor systems interact with each other and the canonical memory systems to form lasting auditory memories.
Acknowledgements

I would first like to thank Josef Rauschecker for giving me the opportunity to combine two of my passions: music and science. We bonded over our mutual love of music almost immediately, and it was clear that there were some fun and interesting directions to scientifically explore the role of music in cognitive neuroscience. Having joined his lab with almost no experience in cognitive neuroimaging he helped forge collaboration with Mikko Sams and Iiro Jääskeläinen of Aalto University that would form the foundation of my neuroscience training. So to him and to both of them I am eternally grateful.

Thanks to my thesis committee: Michael Ullman, John VanMeter, Max Riesenhuber, and Barry Horwitz. The critique and insight that you’ve provided me in my quest to complete these projects has radically transformed and improved the quality of the work. We got a bit of a late start with the meetings, but I think we more than made up for it in the last two years. I thank you for your patience and your flexibility in managing the ups and downs along the way.

I would like to thank Penelope Kuhn for first putting me on the path to pursue a career in science, and I must give a very special thanks to Karen Gale for being a tireless student advocate and one of the sharpest scientific minds this world will ever know. You are missed.

Of course, I would finally like to thank my mom Kathy, my brother Justin, and my darling Leighton. You have all provided me with unconditional love and support for as long as I can remember. My life is better because of you.
I’d specifically like to thank Jussi Alho. He was lead author on the publication from which this chapter is adapted. He not only played a critical role in this project’s completion, but also taught me MEG data acquisition and analysis, as well as introduced me to some of the finer aspects of Finnish culture. Thanks to the other authors Patrick May, Mikko Sams, Hannu Tiitinen, Josef Rauschecker, and Iiro Jääskeläinen for contributions to the project and paper. Specific thanks must also go to Mark Chevillet and Max Riesenhuber for their contribution to the original fMRI study from which this MEG study was adapted. Marita Kattelus and Matt Huszagh were also instrumental in the MRI and MEG data acquisition. Funding for this study was provided by the National Science Foundation (PIRE-OISE-0730255), the Emil Aaltonen Foundations, and from the Academy of Finland (grant No. 138145 and FiDiPro No. 118370).

Special thanks to Dima Smirnov, who spent countless hours as my imaging technician, research assistant, programming expert, roommate, and friend. The novel design and analysis used for this project led to many problems that were beyond the scope of my understanding at the time, so I must thank fellow students Enrico Glerean, Juha Lahnakoski, Mikko Hakala, Kyle Shattuck, Jeremy Purcell, Iain Dewitt, and Amber Leaver for patiently listening to and answering my statistics, fMRI, programming, and experimental design questions. Funding for this study was provided by the National Science Foundation (PIRE-OISE-0730255), aivoAALTO and Academy of Finland (FiDiPRO No. 11837; CoE No. 129670).
# Table of Contents

**Introduction** ........................................................................................................................................... 1

Motor Control theory of Speech Perception ......................................................................................... 2

Dual Stream Models ................................................................................................................................. 3

Declarative and Procedural Memory ........................................................................................................ 5

Current Investigation ............................................................................................................................... 6

**Chapter 1: Early-Latency Categorical Speech Sound Representations in the Left Inferior Frontal Gyrus** ................................................................................................................................................. 9

Introduction ............................................................................................................................................... 9

Materials and Methods ............................................................................................................................ 13

Results ..................................................................................................................................................... 26

Discussion ............................................................................................................................................... 31

**Chapter 2: Distinct Networks of Neural Activation Subserve Encoding and Storage of Novel and Familiar Sound Sequences** .................................................................................................................................................. 42

Introduction ............................................................................................................................................... 42

Materials and Methods ............................................................................................................................ 45

Results ..................................................................................................................................................... 51

Discussion ............................................................................................................................................... 54

**Chapter 3: General Discussion** ........................................................................................................... 63

Objectives ............................................................................................................................................... 63

Motor Role in Speech Perception ............................................................................................................. 64

Basal Ganglia as Auditory Dorsal Stream Structures ............................................................................... 65

Hierarchical Representation of Auditory Objects ..................................................................................... 66

Medial Temporal Lobe as Auditory Ventral Stream Structures ................................................................. 66

Conclusions ............................................................................................................................................. 68

**References** ........................................................................................................................................ 69
## List of Figures

Figure 1.1: Examples of the Auditory Stimuli (Displayed as Spectrograms) and Adaptor-Probe Pairs Used in the MEG Experiment ................................................................. 14

Figure 1.2: Group Average Results from the Syllable Discrimination and Categorization Tasks .......................................................................................................................... 24

Figure 1.3: Neural Selectivity for Phoneme Category and Acoustic-Phonetic Features Determined with Spatiotemporal Cluster Analysis and Visualized on the Left-Hemisphere Inflated Surface .................................................................................................................. 28

Figure 1.4: Positive Correlation Between Behavioral Phoneme Categorization and Neural Phoneme Category Selectivity in the Left Anterior Insula and Posterior Operculum .......... 30

Figure 1.5: Stronger Responses to Speech Sounds When the Auditory Input was Attended to Versus Ignored ........................................................................................................ 32

Figure 1.6: Connectivity Between the Speech-Selective Cluster in the Posterior Temporal Cortex and the Rest of the Cortex during Processing of Speech Sounds in the ATTEND condition ......................................................................................................................... 33

Figure 2.1: Example of a Progressively Presented Musical Sequence consisting of 5 Subsequence Chunks .............................................................................................................. 47

Figure 2.2: Experimental Design and Stimulus Schematic ........................................................................................................................................................................ 48

Figure 2.3: Parametric Analysis of Novelty Across the final 30-Second Trial ....................... 52

Figure 2.4: Areas Showing a Parametrically Increasing Activation in Medial Frontal Cortex to Repeated Stimuli Across the final 30-Second Trial ................................................................... 53

Figure 2.5: Areas Showing a Parametric Increase in Activation Over Final 10, Final 13.5, and All 17.5 Repetitions of the Target Sequence ........................................................................ 55

Figure 2.6: Intermediate Chunk Familiarity in Axial and Sagittal slice Images with Overlaid Images of Parametrically Increasing Responses to Final 10, Final 13.5, and All 17.5 Repetitions ......................................................................................................................... 56
Introduction

Speech and music production both require the tight coupling of motor articulations and auditory feedback (Zatorre, 2007), but the role of these motor areas for receptive speech and music perception is less clear (Rauschecker, 2011; Hickok et al., 2011).

As a musician learns to play an instrument, say the piano, they are learning to map sounds to specific hand positions and movements (Lahav et al., 2007; Zatorre et al., 2007; Furuya and Soechting, 2010). These mappings refer to both absolute relationships, where a specific position maps to a single specific pitch, as well as relative relationships, where the successive or simultaneous movement from one position to another produces two pitches of a specific distance or tonal interval. Through trial and error, as well as explicit instruction, the aspiring musician learns these associations, commits them to memory, and builds a repertoire of musical performance skills that become fast, accurate, and automatic. The same could be said of the process involved in speech production.

When babies are first learning to speak they go through a process of babbling (Halle, 1989). They hear speech sounds produced by those around them and they attempt to imitate them. Their speech output is nonsense, irregular, and inaccurate but through repeated trial and error the auditory feedback helps reduce error in the motor code responsible for producing clear recognizable speech. In this manner, babies can learn their native language automatically and with ease, but when learning a second language or taking up music lessons later in life it becomes much more difficult due to attain the same level of fluency or expertise (Doupe and Kuhl, 1999). This is due to the developmental
critical period that maximizes the ability to learn these audiomotor skills during this young age (Doupe and Kuhl, 1999; Bolhuis et al., 2010).

In both examples, people are exposed to small, individual units that are concatenated to produce longer and longer sequences of sounds. In the case of speech this progression goes from individual phonemes, to combinations of phonemes that form words, and to combinations of words that form sentences. For music this progression would be from individual notes, to combinations of notes to form melodies, to combinations of melodies to form songs. Indeed, both speech and music rely on the combination of smaller units into longer sequences for the purpose of conveying increasingly complex information. Exactly how the brain transitions from perceiving and encoding individual units to processing and storing longer sequences remains an open question.

*Motor Control Theory of Speech Perception*

The motor theory of speech perception holds that premotor cortex, part of the sound-to-movement mapping network of the brain, is required for speech processing (Liberman & Mattingly, 1985; Liberman et al., 1967). The debate over the motor control theory of speech perception has endured for years. This reasoning goes that because the motor system is involved in speech production it is necessarily recruited for speech perception (Schwartz et al., 2012). Evidence for this comes from studies showing activations in motor areas during speech listening tasks (Schulze et al., 2012; Chevillet et al., 2013). While the role of motor control on babbling babies and budding musicians seems
self evident, the controversy comes into play when assuming that motor system activation is required for speech or music perception in normal adults later in life. Experimental evidence showing speech processing is preserved in patients with conductive aphasia (Damasio and Damasio, 1980) seems to go contrary to the necessity of motor areas in receptive speech processing. An alternative explanation is that while premotor cortex can be activated by speech it may not be required (Hickok et al., 2011; Schwartz et al., 2012). Furthermore, despite conflicting evidence on either side of the debate, the motor theory of speech perception doesn’t provide an explanation for neural activations in cortical and subcortical areas for musical paradigms in non-musicians who would not have experienced the requisite sound-to-movement mapping.

**Dual Stream Models**

To understand the role of motor areas in the processing of speech and music sounds it is necessary to understand how the brain encodes, processes, and stores auditory information. Current models and theories on auditory perception have borrowed heavily from those used to explain vision. A popular model for understanding the organization of vision is the dual-stream model (Goodale and Milner, 1992; Milner and Goodale, 2008). This model posits that there are two largely segregated streams of sensory processing that are each responsible for different types of information. Anatomically, the ventral stream originates in primary auditory cortex and proceeds ventrally and anteriorly through temporal cortex, insula, and terminating in the (ventrolateral) prefrontal cortex, while the dorsal stream originates in A1 and proceeds dorsally and posteriorly through parietal, motor and premotor cortices before terminating in the (dorsolateral) PFC (Rauschecker &
Scott, 2009). The ventral stream, dubbed the “what” stream is said to be responsible for visual object quality or identity (Kravitz et al., 2013; Rauschecker & Scott, 2009), while the dorsal stream, dubbed the “where” stream is responsible for visuospatial processing (Kravitz et al., 2011; Rauschecker, 2014).

Monkey electrophysiological data have been used as evidence to extend this model to the non-human primate auditory system (Rauschecker & Tian, 2000), which was then also applied to human auditory processing (Rauschecker & Scott, 2009; Hickok & Poeppel, 2008; Dewitt & Rauschecker, 2012; Leaver et al., 2009). Some functional aspects of the dual-stream model for vision are not universally regarded as analogous with dual stream hearing. For example, spatial processing in the visual domain may be more selectively part of the dorsal stream (Kravitz et al., 2011) than in the auditory domain. There have been studies to support a dorsal stream role in auditory spatial processing (Rauschecker and Tian, 2000; Warren et al., 2002; Renier et al., 2010), but there seems to be some evidence that object and spatial processing may interact within the streams rather than being strictly segregated (Zatorre et al., 2002). So although the dorsal stream pathway in vision can maintain the “where” moniker, for audition this doesn’t necessarily provide the complete picture.

In light of this, the role of the dorsal stream has been expanded to include additional (or more specific) functions and incorporated into similar but different versions of the dual-stream model from different groups. The dual-stream model at the root of the work in this dissertation is the functional and anatomical model of auditory processing described
above (Tian & Rauschecker, 2000; Rauschecker & Scott, 2009; Rauschecker, 2011) that identifies roles for the ventral stream in auditory object identification (Leaver & Rauschecker, 2010) and hierarchical speech processing (Dewitt & Rauschecker, 2012), while the dorsal stream subserves musical sequence processing (Leaver et al., 2009) and categorical speech perception (Chevillet et al., 2013).

Building upon the Rauschecker (2011) model and combining it with a computational model for sentence comprehension, Bornkessel-Schlesewski & Bornkessel (2013) have defined the dorsal and ventral streams as being time-dependent and time-independent, respectively. This provides a different computational basis for understanding the functional distinction between the two streams, while at the same time further supporting the former model, i.e. identifying words or other auditory objects is not dependent on a temporal component, while sequence processing is by definition the processing of sounds over time.

**Declarative and Procedural Memory**

These dual-stream auditory models often focus on processing and perception while ignoring an important aspect of cognition i.e. memory. The repositories of words and speech sounds, musical tones and rules, melodies and sound sequences, all have to be stored in memory. Like the dual-stream models of auditory processing, memory comes in various kinds that are subserved by various areas. Two classical divisions are between declarative and procedural memory systems: declarative memory is memory for facts, dates, or events, while procedural memory is rules, procedures, or actions (Cohen & Squire,
1980). Applied toward human language, the declarative/procedural model highlights the overlap of the lexicon with declarative memory and grammar with procedural memory (Ullman, 2004). The declarative/procedural model of language seems to be consistent with the dual-stream model of auditory processing, such that ventral and dorsal processing streams, respectively, and could work together as part of a greater memory-perception network (Ullman, 2004; Eichenbaum, 2007; Rauschecker, 2014). Furthermore, the basal ganglia subserve a variety of procedural memory functions (Ullman, 2004). They have also been shown to play a role in auditory dorsal stream processing (Leaver et al., 2009), although not formally identified as a dorsal stream structure. Declarative memory is more dependent on the medial temporal lobe structures—although they also subserve some procedural processes (Ullman, 2004; Eichenbaum et al., 2012). Auditory processing studies have shown limited activity in the medial temporal lobe (Watanabe et al., 2008; Leaver et al., 2009), but it has seldom been the focus of the discussion.

**Current Investigation**

Music and speech are both produced through the coordinated movement of the body to produce specific sounds, and the dorsal stream plays an important role in the coupling of auditory and motor systems in speech and music processing (Zatorre, et al., 2007), but speech and music both utilize a repository of words, sounds, and auditory representations that should require the object representation offered by the ventral stream, as well as learning and memory structures, including but not limited to the medial temporal lobe and basal ganglia (Miranda & Ullman, 2007; Ullman, 2004; Kravitz et al., 2013; Eichenbaum et al., 2012).
Previous work in our lab investigated the categorical processing of speech phonemes found categorical speech representations in left premotor cortex using a rapid-adaptation fMRI design (Chevillet et al, 2013). We seek to replicate this study using MEG in order to benefit from the increased temporal resolution afforded by this technique as compared with fMRI. This approach allows us to detect precise timing differences within and between regions of interest to identify areas sensitive to specific stimulus conditions and directional flow of information. We also seek to expand on this work by using a music paradigm to investigate the transition from low-level unit encoding and concatenation to the binding and storage of longer chunks and sequences. We use a music paradigm to avoid the pitfalls of confounding variables in lengthening speech sounds, such as their semantic content.

Detection of structure in a sound sequence requires both learning and storage of ordered sequences. The dorsal auditory stream has been implicated in sensorimotor integration and concatenation of sequential events (Rauschecker, 2011) and plays an important role in the coupling of auditory and motor systems in speech and music processing (Zatorre et al., 2007). Previous work in our lab has investigated auditory sequence learning with a study that used fMRI to image brain activation during anticipatory imagery between well-known songs or recently-learned short musical sequence pairs (Leaver et al., 2009). This study showed differences in activation based on the novelty and familiarity of sequences, but the experiments were carried out with different subjects at different times. Our study seeks to investigate the role of the dorsal
and ventral streams, including subcortical motor and medial temporal lobe structures, in the processing of musical sequences presented in such a way as to produce a continuum of novelty and familiarity within a single stimulus in order to see the exact progression from areas responsible for encoding the sequences to areas responsible for storing the sequences in a repetition-dependent manner.
Chapter 1: Early-Latency Categorical Speech Sound Representations In The Left Inferior Frontal Gyrus

Introduction

Phonemes, the elementary units of speech, greatly vary in their acoustic structure when produced by different speakers in different contexts. The brain therefore faces a fundamental challenge of mapping highly variable acoustic signals to distinct phonetic categories. Although categorical perception of speech sounds is well documented, it remains unresolved how the brain accomplishes this many-to-one mapping. Specifically, it is unclear which cortical areas exhibit categorical speech processing and at what latencies from sound onset this occurs. Current theories postulate that speech is cortically processed by parallel ventral and dorsal auditory streams (Rauschecker, 1998a, b; Wise, 2003; Hickok and Poeppel, 2007; Rauschecker and Scott, 2009). The ventral stream, involving superior-to-middle temporal areas and terminating in pars triangularis (roughly corresponding to Brodmann area 45) of the inferior frontal gyrus, has been suggested to process speech signals for comprehension, whereas the dorsal stream, projecting from auditory cortex via the temporoparietal junction to premotor cortex and pars opercularis (Broadmann area 44) of the inferior frontal gyrus, has been proposed to mediate a mapping between auditory and articulatory-motor representations (Rauschecker, 2011). Given that each human has a repertoire of potential speech gestures which is less variable than the mass of acoustic speech signals one has to categorize (Liberman et al., 1967; Liberman and Mattingly, 1985), it can be hypothesized that categorical speech representations are found

---

1 Adapted from Alho et al. NeuroImage. 2016. 129:214-23.
in the speech-motor areas (e.g. pars opercularis/premotor cortex) and that they guide speech categorization via the dorsal stream.

The sensorimotor nature of speech processing is supported by empirical findings whereby disrupting speech-motor areas with transcranial magnetic stimulation (TMS) impairs speech sound discrimination or categorization (Meister et al., 2007; Möttönen and Watkins, 2009; Sato et al., 2009; D’Ausilio et al., 2012; Grabski et al., 2013). Furthermore, stimulus-specific adaptation of fMRI signals revealed categorical speech representations in the left premotor cortex (Chevillet et al., 2013), pars opercularis (Myers et al., 2009; Lee et al., 2012) and anterior insula (Myers et al., 2009). Importantly, the categorical processing in these fMRI studies was task-independent, as subjects engaged in a listening task wherein phoneme category information was irrelevant.

Lower-level phonological processing areas in temporal and parietal lobes have also been implicated in categorical speech processing. An fMRI-adaptation experiment where subjects watched a film without the soundtrack and were under instruction to ignore the sounds found categorical speech representations in the left superior temporal sulcus and the supramarginal gyrus (Joanisse et al., 2007). The supramarginal gyrus exhibited categorical speech representations also when the auditory input was attended to (Raizada and Poldrack, 2007). However, re-analysis of these same data with multivariate rather than univariate techniques revealed categorical speech representations in the pars opercularis, rather than in supramarginal gyrus, with the discrepancy supposedly due to different spatial scales of cortical representations in different dorsal-stream areas (Lee et al., 2012).
Electrocorticographic recordings from the posterior superior temporal gyrus revealed categorical speech representations during passive listening to speech sounds with small acoustic differences (Chang et al., 2010). The left medial superior temporal sulcus was associated with phonemic perception being more strongly activated by familiar speech sounds than acoustically (i.e. spectro-temporally) matched non-phonemic sounds (Liebenthal et al., 2005). Further, a study using combined fMRI and electroencephalography (EEG) suggested that categorization of highly familiar (e.g. native) and newly acquired speech sounds rely on long-term representations in medial superior temporal sulcus and short-term representations in posterior superior temporal sulcus, respectively (Liebenthal et al., 2010). Another fMRI-study identified category-selective responses to speech sounds in anterior superior temporal regions (Leaver and Rauschecker, 2010). In line with these findings, a recent meta-analysis localized invariant phoneme representations consistently in anterior-to-mid superior temporal gyrus (DeWitt and Rauschecker, 2012).

As evidence for categorical perception of phonemes has been found both within the ventral and dorsal streams, it seems plausible that invariant representations are formed independently based on both spectro-temporal and articulatory-motor information (for review, see Rauschecker, 2012). An intriguing question is what determines the engagement of the two respective streams in speech categorization. Previous research has proposed a modulatory role for the dorsal stream in speech perception, particularly in the learning of new sound categories (Liebenthal et al., 2010), under adverse listening conditions (Osnes et al., 2011; Du et al., 2014), or during sublexical tasks, such as syllable discrimination.
(Hickok and Poeppel, 2007). However, none of these conditions were present in the above-mentioned studies reporting categorical speech representations in the dorsal stream areas, which raises the possibility that the discrepant results between the ventral and dorsal stream involvement in speech categorization could be explained by differences in allocation of auditory attention. In support of this interpretation, a recent study using TMS and MEG demonstrated that the involvement of articulatory-motor areas in the early (<100 ms) processing of acoustic-phonetic features of speech depended on attention, while the longer-latency auditory-motor interaction (>170 ms) occurred even when the subjects were under instruction to ignore the sounds and to focus on watching a silent film (Möttönen et al., 2014).

Here, we used stimulus specific adaptation and cortically-constrained MEG source estimates to infer the location, latency, and attention-dependence of categorical speech representations. Sounds from a phonetic continuum were presented to participants while they were performing a non-phonetic listening task and, in a separate passive recording condition, ignoring the sounds while watching a film without the soundtrack. The following questions were addressed: Are categorical speech representations observed in speech-motor areas regardless of auditory attention? Is the latency of categorical speech representations compatible with the proposal that phonological categories are accessed ~150 ms after sound onset (Salmelin, 2007)? Does the neural selectivity underlying categorical speech representations correlate with behavioral categorization?
Materials and Methods

Participants

All 22 subjects were right-handed and reported neither a history of hearing problems nor neurological illnesses. MEG data from four subjects were excluded from analyses due to poor a signal-to-noise ratio (SNR). This resulted in a final sample of 18 subjects (6 females; age mean ± SD = 25.3 ± 4.0, range 21–38 years). The experiment was approved by the Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa, Finland.

Stimuli

The present study utilized the same stimulus material as that in Chevillet et al. (2013), that is, a place-of-articulation continuum between the natural utterances /da/ and /ga/ (Fig. 1.1). Place-of-articulation refers to the point of maximum obstruction in the vocal tract in the articulation of a consonant. The stimuli were produced with the STRAIGHT toolbox (Kawahara and Matsui, 2003) for MATLAB (MathWorks), which allows for the parametric manipulating of the acoustic and acoustic-phonetic structure of natural voice recordings. The natural utterances were taken from recordings provided by Shannon et al. (1999). Two phonetic continua (or “morphlines”) were generated at 0.5% intervals between the /da/ and /ga/ prototypes: one for a male voice and the other for a female voice. Morphed stimuli were generated up to 25% beyond each natural utterance (i.e. from -25% /ga/ to +125% /ga/), for a total of 301 stimuli per morphline. The stimuli created...
Figure 1.1: Examples of the Auditory Stimuli (Displayed as Spectrograms) and Adaptor-Probe Pairs Used in the MEG Experiment. The anchor points for mapping from one stimulus to the other are denoted with circles (o). The arrows show how the stimuli were paired to probe selectivity to acoustic-phonetic features and to phoneme categories: ID, identical sounds; 33S, 33% acoustic-phonetic difference and same category; 33D, 33% acoustic-phonetic difference and different category; 67D, 67% acoustic-phonetic difference and different category. Modified from Chevillet et al. (2013).
beyond the natural utterances were qualitatively assessed to ensure their intelligibility and behaviorally verified in a categorization test (described below).

All stimuli were resampled to 48 kHz, trimmed to 300-ms duration, and root-mean-square normalized in amplitude. A linear amplitude ramp of 10-ms duration was applied to sound offsets to avoid auditory artefacts. Amplitude ramps were not applied to sound onsets so as to avoid interfering with the natural features of the consonant sound.

**Discrimination Behavior**

Prior to brain imaging, the subjects completed a discrimination test to identify individual category boundaries. The discrimination thresholds of the subjects were measured at 10% intervals along both male and female voice continua. The adaptive staircase algorithm QUEST (Watson and Pelli, 1983), implemented in MATLAB using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997), was used to adjust the difference between paired stimuli based on subject performance. This allowed the measurement of the just-noticeable difference at each location (for both morph directions), which is known to have its minimum value at category boundaries. To diminish the risk that the subjects would categorize the sounds during MEG, the task on each trial was to report as quickly and accurately as they could whether the two sounds were exactly the same or in any way different without assigning them to a specific phonetic category. A maximum period of 3 seconds was allowed for a response before the next trial started. In half of the trials, the paired stimuli were identical and in the other half they were different from each other. Of the pairs where the stimuli differed from each other, half represented a displacement in
one direction along the continuum, and the other half a displacement in the opposing direction. In total, 560 trials were presented, with 20 conditions (10% intervals from 0 to 90% with displacements toward 100%, and 10% intervals from 10 to 100% with displacement toward 0%) and 28 trials per condition.

Categorization Behavior

After brain imaging, the subjects were asked to categorize the auditory stimuli along both morphlines to confirm the location of their individual category boundary as well as to measure its sharpness. 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, the subject was presented with a single sound and given up to 3 seconds to indicate as quickly and accurately as possible whether s/he had heard /da/ or /ga/. Each subject completed 15 runs of 20 trials per condition, for a total of 300 trials per morphline. The resulting data were fitted with subject-specific sigmoidal functions to estimate boundary locations as well as boundary sharpness. The sigmoid was given by the generalized logistic curve:

\[ f(x) = \frac{1}{1 + e^{-(x-\alpha)/\beta}} , \]

where \( x \) is the location along the morph line, \( \alpha \) is the location of the boundary along the morph line, and \( 1/\beta \) is the steepness of the boundary (with lower values of \( \beta \) resulting in sharper boundaries).
MEG Paradigm

To infer neuronal stimulus selectivity, the stimulus specific adaptation paradigm was used. In stimulus specific adaptation, two stimuli – an adaptor and a probe – are presented in succession in each trial, and the similarity between the two stimuli is varied between trials to investigate neuronal tuning along the dimensions of interest (Butler, 1972). In this setup, the attenuation of the response to the probe reflects the overlap between the neural pars opercularisulations responding to the adaptor and the probe, respectively. By presenting stimuli across a continuum of acoustic and categorical similarity this paradigm results in one of three response profiles indicating broad acoustical selectivity, sharp acoustic selectivity, or categorical selectivity. In the present study, the silent interval between the adaptor and the probe was 500 ms and the interval between successive adaptor-probe pairs varied randomly and uniformly in the 5–7 second range. For each subject, four sounds along the morphline were selected on an individual basis according to the pre-imaging behavioral discrimination test. These sounds were combined into the following four adaptor-probe pairs defined by the acoustic-phonetic change and the phoneme category change between the adaptor and the probe: (1) identical sounds (ID), (2) 33% acoustic-phonetic difference, same category (33S), (3) 33% acoustic-phonetic difference, different category (33D), and (4) 67% acoustic-phonetic difference, different category (67D) (Fig. 1.1). Thus, any difference in adaptation between 33S and 33D, which were equalized with regard to acoustic-phonetic dissimilarity, can be attributed to an explicit representation of the phoneme categories.
Brain regions containing category-selective neurons should show larger adaptation (i.e. reductions of the probe response in relation to the adaptor response) in the 33S trials than in the 33D trials, as the stimuli in each pair in the latter condition would activate different neuronal pars opercularisulations, whereas the same neuronal pars opercularisulations would be activated in the former condition. In this way, stimulus specific adaptation enables the dissociation of phoneme category selectivity from mere tuning to acoustic-phonetic differences. On the one hand, category-selective neurons respond similarly to dissimilar stimuli from the same category but differently to similar stimuli belonging to different categories (Freedman et al., 2003; Jiang et al., 2007). On the other hand, in the case of tuning to acoustic-phonetic differences, neuronal responses gradually drop off with acoustic dissimilarity, without the sharp transition at the category boundary that is the hallmark of perceptual categorization. Morphlines were 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 were centered at the category boundary for each subject.

To observe responses independent of overt phoneme categorization, the subjects were scanned while they performed an attention-demanding distractor task for which phoneme category information was irrelevant (from here on referred to as the ATTEND condition). For this reason, each presented sound in a pair persisted 30 ms longer in one ear than in the other. In ATTEND, the subject was asked to listen out for these offsets and to report whether the two sounds in the pair persisted longer in either the same ear or in different ears. The reporting was done by pressing a button either with the left or right
hand, indicating either "same" or "different". To disentangle activation resulting from categorical decisions from that associated with categorical motor activity (i.e. to average out the motor responses), the label of the left- and right-hand side responses alternated on each run (i.e. whether left-hand side response indicated “same” or “different”). For controlling the effect of attention, the subjects were additionally scanned while they watched a film without soundtrack (a wildlife documentary) and were under instruction to ignore the sounds (from here on referred to as the IGNORE condition). To diminish the possibility that data obtained in the IGNORE condition could be partly contaminated by prior exposure of the stimuli, the interval between the discrimination test and brain imaging was at least two days and on average 21 days.

Both ATTEND andIGNORE conditions comprised 512 trials (128 for each adaptor-probe pair). The trial order was randomized and the number of presentations between all stimuli was equalized. With an average 6-s trial duration, the measurement time per condition totaled ~51 min, which was divided into eight ~6.4-min blocks to prevent fatigue. The measurements were divided over two days (one condition per day), with the order counterbalanced across subjects. The auditory stimuli were delivered through insert earphones (Etymotic Research Inc., IL, USA), comprising plastic tubing and earplugs, with the sound level set at 65 dB.

**MEG Data Acquisition And Preprocessing**

The MEG data were acquired with a whole-head 306-channel neuromagnetometer (VectorView, Elekta-Neuromag, Helsinki, Finland) of the MEG Core of Aalto NeuroImaging
infrastructure at Aalto University. The device was situated in a magnetically shielded room, with a three-layer μ-metal and aluminum cover to attenuate effects of outside magnetic fields. An additional active noise-cancelation system was used. Before each MEG recording session, the locations of five head-position indicator (HPI) coils attached to the scalp were recorded with respect to three anatomical landmark points (the nasion and two preauricular points) using a 3-D digitizer (Isotrak, Polhemus, Colchester, VT). Additional scalp surface points (~30) were digitized to facilitate the coregistration with anatomical magnetic resonance (MR) images. Vertical and horizontal electro-oculograms (EOG) were used to detect eye blinks and movements. The MEG signals were band-pass filtered at 0.03-330 Hz and digitized at 1 kHz sampling frequency. To compensate for the MEG signal changes due to movements, the head position of the subject was continuously tracked during the data acquisition by exciting the HPI coils at high frequencies (290–330 Hz). During preprocessing, the head movements were estimated continuously at 200-ms intervals and compensated using the signal-space separation method (Taulu and Simola, 2006) (Maxfilter, Elekta-Neuromag, Helsinki, Finland). MEG signals time-locked to the onset of the stimuli were averaged across trials for each condition. Amplitudes were measured with respect to a 200-ms pre-adaptor baseline. Trials where the MEG gradiometer, MEG magnetometer, or EOG channel peak-to-peak amplitude exceeded 3000 fT/cm, 4000 fT, or 150 μV, respectively, were rejected from the average. For each subject, more than 100 trials per adaptor-probe pair and per condition were included in the analyses. The averaged signals were band-pass filtered between 1 and 40 Hz.
**Structural Magnetic Resonance Imaging (MRI)**

The individual MR images were acquired with a 3T scanner (Magnetom Skyra, Siemens) of the AMI Centre of Aalto NeuroImaging infrastructure at Aalto University. Coregistration between MEG data and MRIs was done by identifying the fiducial point locations in the MRIs. FreeSurfer software was used to reconstruct the cortical mantle from the MRI data (Dale et al., 1999; Fischl et al., 1999a).

**MEG Source Estimation**

The source currents were estimated at each cortical location by computing a depth-weighted minimum-norm estimate (Hämäläinen and Ilmoniemi, 1994; Lin et al., 2006b). The minimum norm estimate solution does not represent true brain activation but rather recovers a source distribution with minimum overall power that is consistent with the measured MEG signals. The forward solutions for all source locations were computed using a single-compartment boundary-element model (BEM) based on the information from individual structural MRIs and MEG sensor locations (Hämäläinen and Sarvas, 1989). The cortical surface of each subject was decimated to ca. 7000 source locations per hemisphere with an average 5-mm spacing between adjacent locations. A noise covariance matrix was estimated from 200-ms pre-adaptor stimulus baselines of the raw MEG data. Activity at each source location was estimated for each time point of the evoked response using an inverse operator computed from the forward solution and the noise covariance matrix. The source orientations were controlled with a loose orientation constraint (Lin et al., 2006a). In addition to minimum norm estimates, dynamic statistical parametric map (dSPM) estimates were generated (Dale et al., 2000). As a measure of signal-to-noise (derived
through normalizing the minimum norm estimate by the noise sensitivity at each cortical location), dSPM indicates the locations where minimum norm estimate amplitudes are above noise level. Since individual MRI could not be obtained for one subject, a FreeSurfer average brain was used as a surrogate in this subject (by aligning the individual fiducial points to the fiducial points of the average head).

Spatiotemporal Cluster Analysis Of Evoked Responses

Source analysis of evoked responses was conducted with a nonparametric randomization test based on spatiotemporal clustering (Maris and Oostenveld, 2007). The data were downsampled to 200 Hz and the individual cortical surfaces were morphed to a FreeSurfer average brain with 10242 dipoles per hemisphere (Fischl et al., 1999b). The medial wall of the cerebral cortex, as defined by an automatic parcellation (Desikan et al., 2006), was excluded from the analysis due to low SNR. To quantify the stimulus specific adaptation effect, the adaptor-probe reduction rate R (i.e. strength of adaptation) was defined for a 500-ms time window as $R = [\text{adaptor response} - \text{probe response}] / \text{adaptor response}$.

A $t$-value was calculated for each data point (i.e. dipole/time point) for the given contrast between pairing conditions with a two-sided paired-samples $t$-test. All data points with $p$-value < 0.05 (uncorrected for multiple comparisons) were clustered on the basis of spatial and temporal adjacency. Cluster-level statistics were calculated by summing the $t$-values within every cluster and the maximum of the cluster-level statistics was used as the actual test statistic. A reference distribution of test statistics was produced by taking a
thousand random partitions of the combined data across the conditions and calculating a test statistic for each partition. A multiple-comparisons-corrected cluster $p$-value was obtained by comparing the test statistic of the contrast of interest against the reference distribution. The null hypothesis of no difference between the conditions was rejected if this $p$-value was smaller than 0.05. The analysis was performed in five consecutive 100-ms time windows starting from the probe onset and implemented with the MNE-Python toolbox (Gramfort et al., 2013; Gramfort et al., 2014).

For any observed cluster, a paired-samples $t$-test was conducted to reveal the differences in the reduction rates between the adaptor-probe pairs. The reduction rates were obtained by first morphing the observed cluster onto the individual cortical surface (Fischl et al., 1999c), extracting time courses by averaging over the source locations within the region of the cluster, averaging the time courses over the temporal extent of the cluster with respect to both adaptor and probe onsets, and finally calculating the adaptor-probe reduction rate as described above.

For any observed cluster exhibiting phoneme category selectivity, a correlation test was conducted to examine whether subjects who exhibited better behavioral categorization performance also exhibited stronger neural selectivity to phoneme categories. The behavioral categorization performance was measured after the scanning in a categorization test and quantified as the sharpness of category boundary (see Materials and Methods; Fig. 1.2). The selectivity to phoneme categories was quantified as the percent change between the reduction rates in 33S and 33D (i.e. $[R_{33S} - R_{33D}] / R_{33S}$). The reduction
Figure 1.2: Group Average Results from the Syllable Discrimination and Categorization Tasks. The JND was measured at 10 percentage intervals along each morphline (male and female voice) both toward /ga/ (red curve) and toward /da/ (blue curve). The halfway point between the minima of the two curves predicts the category boundary, measured in a separate categorization test (black curve). Error bars indicate SEM.
rates were obtained as in the paired-samples t-test (described in section 4.9). The Spearman rank correlation test was applied to examine the level of correlation.

**Connectivity Analysis**

For estimating inter-areal connectivity, the phase slope index (Nolte et al., 2008) was computed from single-trial minimum norm estimates. Phase slope index is based on the notions that the imaginary part of the coherency/cross-spectrum is insensitive to false connectivity caused by volume conduction (Nolte et al., 2004) and that the direction of information flow can be derived from the slope of the phase of the cross-spectrum. Accounting for the inevitable delay when distinct brain areas interact through a physical medium, phase slope index provides a robust measure of effective connectivity insensitive to common challenges in MEG connectivity analyses, such as low SNR and signal mixing due to volume conduction. Minimum norm estimate inverse solutions were first computed for all (max 512) artefact-free epochs (-200–500 ms with respect to adaptor onset) in the ATTEND condition. For estimating the cross-spectra, the epochs were then filtered with the continuous Morlet wavelet transform between 8–40 Hz with 4 Hz steps (with the wavelet width linearly increasing from 1.5 to 7.5 cycles for the lowest to the highest frequency). Since phase slope index is a signed quantity, indicating both the connectivity strength as well as the direction of information flow, the null hypothesis is simply that the values are drawn from a zero-mean distribution (Haufe et al., 2013). The significance of the phase slope index values was statistically tested with spatiotemporal cluster analysis (for details, see section 4.9).
Phase slope indices were calculated between any observed speech-selective cluster and all other source locations (excluding the medial wall). Individual seed regions-of-interest (ROIs) were defined by first morphing the clusters onto the individual cortical surface and then applying functional constraints by selecting the sources within the clusters where the individual dSPM values exceeded a threshold of 6 (F-statistic) at any time within 500 ms after adaptor onset. Seed time courses were obtained by averaging across source locations within the seed ROI. Only the radial components of the seed time courses were kept and, depending on source orientation, sign-flips were applied to reduce signal cancellations. The analysis was performed with the MNE-Python toolbox (Gramfort et al., 2013; Gramfort et al., 2014).

**Results**

**Behavior**

The syllable discrimination test showed clear minima in the just noticeable difference for morph differences in each direction for all subjects. The category boundary was inferred to be halfway between the two smallest just noticeable difference measurements (one in each direction). The explicit phoneme category boundary measured in the syllable categorization test conformed with the just noticeable difference minima (Fig. 1.2). During the MEG scanning, the average performance in the non-phonetic listening task of the ATTEND condition across subjects was 84.8% (ID: 84.7 %, 33S: 85.2%, 33D: 85.4%, 67D: 84.0%; with no statistically significant differences between the pairing conditions), indicating that the task was demanding and therefore minimized the chance that the subjects covertly categorized the stimuli in addition to performing the task. The
long response times in this task (mean ± SD, all: 1267 ± 293 ms, ID: 1222 ± 280 ms, 33S: 1271 ± 312 ms, 33D: 1286 ± 313 ms, 67D: 1305 ± 330 ms) in this task further validate its difficulty. The response times for ID were shorter than for 33S, 33D, or 67D (ID vs. 33S: $t_{(17)} = 3.20, p < 0.05$; ID vs. 33D: $t_{(17)} = 3.99, p < 0.001$; ID vs. 67D: $t_{(17)} = 4.28, p < 0.001$). No statistically significant differences were observed between any other pairing conditions.

**Spatiotemporal Cluster Analysis Of Evoked Responses.**

'Selectivity For Phoneme Category: In the ATTEND condition, the contrast 33S > 33D, indicative of phoneme category selectivity, yielded one cluster involving parts of the left anterior insula and pars opercularis (115–140 ms, $p < 0.05$; Fig. 1.3). Paired-samples $t$-tests revealed stronger release from adaptation in 33D compared to 33S ($t_{(17)} = 3.41, p < 0.01$) as well as in ID compared to 33D ($t_{(17)} = 2.59, p < 0.05$) and in 33S compared to 67D ($t_{(17)} = 2.17, p < 0.05$). The contrast yielded no significant clusters in the IGNORE condition.

'Selectivity For Acoustic-Phonetic Features Regardless Of Phoneme Category: Sharp acoustic-phonetic selectivity (i.e. $3*ID > 33S + 33D + 67D$) was found in the ATTEND condition in an area involving parts of the left anterior superior temporal gyrus and posterior insula (235–275 ms, $p < 0.05$; Fig. 1.3). Paired-samples $t$-test showed that the release from adaptation in 33S, 33D, and 67D were each significantly stronger than that in ID (33S vs. ID: $t_{(17)} = 2.77, p < 0.05$; 33D vs. ID: $t_{(17)} = 3.13, p < 0.01$; 67D vs. ID: $t_{(17)} = 3.87, p < 0.01$). Two clusters exhibited broad acoustic-phonetic selectivity (i.e. ID + 33S + 33D > 3*67D) in the ATTEND condition: one in the left anterior temporal cortex (300–400 ms, $p < 0.01$; $t$-test, 67D vs. ID: $t_{(17)} = 2.16, p < 0.05$; 67D vs. 33S: $t_{(17)} = 2.58, p < 0.05$; 67D vs. 33D: $t_{(17)} = 2.16, p < 0.05$).
Figure 1.3: Neural Selectivity for Phoneme Category and Acoustic-Phonetic Features Determined with Spatiotemporal Cluster Analysis and Visualized on the Left-Hemisphere Inflated Surface. The bar charts show the reduction rates for the four adaptor-probe pairs as well as the t-tests for the differences between these rates. The color coding indicates the temporal extent of the clusters, error bars indicate SEM, and asterisks indicate significant differences (*p<0.05, **p<0.01, ***p<0.001). ID, identical sounds; 33S, same category with 33% acoustic-phonetic difference; 33D, different category with 33% acoustic-phonetic difference; 67D, different category with 67% acoustic-phonetic difference.
Broad acoustic-phonetic selectivity was found also in the IGNORE condition, with one cluster located in the left middle temporal cortex (315–400 ms, \( p < 0.05 \); t-test, 67D vs. ID: \( t_{(17)} = 2.37, p < 0.05 \); 67D vs. 33S: \( t_{(17)} = 4.11, p < 0.001 \); 67D vs. 33D: \( t_{(17)} = 3.41, p < 0.01 \)) and another in the left posterior temporal cortex (300–400 ms, \( p < 0.01 \); t-test, 67D vs. ID: \( t_{(17)} = 2.25, p < 0.05 \); 67D vs. 33S: \( t_{(17)} = 2.15, p < 0.05 \); 67D vs. 33D: \( t_{(17)} = 2.64, p < 0.05 \); Fig. 1.3). Sharp acoustic-phonetic selectivity was not found in the IGNORE condition.

**Correlation Between Neural Phoneme Category Selectivity And Behavioral Phoneme Categorization**

A significant positive correlation was found between the individual degree of phoneme category selectivity in the left anterior insula/pars opercularis and the ability to categorize the phonetic stimuli (Spearman \( r = 0.61, p < 0.05 \); Fig. 1.4). The categorization ability was measured behaviorally after brain imaging in a syllable categorization test (see section 2.4). Two outliers were removed from the analysis.

**The Effect Of Attending Vs. Ignoring The Auditory Input To Speech Processing**

As differential results were obtained between the ATTEND and IGNORE conditions, an additional analysis was performed to examine the differences in the processing of speech sounds when the auditory input was attended vs. ignored. The stimulus categories
Figure 1.4: Positive Correlation Between Behavioral Phoneme Categorization and Neural Phoneme Category Selectivity in the Left Anterior Insula and Posterior Operculum. Behavioral phoneme categorization was determined as category boundary sharpness, which was measured after scanning during a syllable categorization test (see section 2.4 and Fig. 2). The measurements were averaged over the male and female voice continua, and normalized. The level of neural phoneme category selectivity was quantified by the percent change in reduction rates between 33S and 33D. The $r$ and $p$ denote the Spearman rank correlation coefficient and the corresponding $p$-value, respectively. Black line represents the regression line.
were first combined into one category and responses between ATTEND and IGNORE conditions were contrasted in five consecutive 100-ms time windows starting from the adaptor (not probe) onset (and therefore being unaffected by the repetition suppression). The contrast yielded five clusters, all indicating stronger responses in the ATTEND condition (Fig. 1.5): (1) ~ left inferior frontal gyrus (120–170 ms, \( p < 0.05 \)), (2) ~ left inferior frontal gyrus / premotor cortex (200–300 ms, \( p < 0.01 \)), (3) ~ right premotor cortex / middle frontal gyrus (200–300 ms, \( p < 0.05 \)), (4) ~ right inferior temporoparietal cortex (200–300 ms, \( p < 0.05 \)), and (5) ~ left dorsal premotor cortex / superior parietal cortex (300–400 ms, \( p < 0.05 \)).

*Effective Connectivity During Speech Processing*

Connectivity tests were conducted to determine the cortical areas to which the speech-selective areas are effectively connected. The analysis was restricted to the range 100–400 ms after adaptor sound onset and performed in 100-ms time windows. In the 120–170 ms range, the speech-selective region in the posterior temporal cortex showed significant connectivity with an area involving the left pars opercularis and premotor cortex (\( p < 0.05 \); Fig. 1.6). The direction of information flow was estimated as going from the posterior temporal cortex to pars opercularis/premotor cortex. No significant connectivity was found with respect to any other speech-selective cluster.

*Discussion*

The present study investigated which cortical areas support categorical speech processing, at what latencies from sound onset such processing occurs, and whether it
Figure 1.5: Stronger Responses To Speech Sounds When the Auditory Input Was Attended to Versus Ignored. The analysis was performed with spatiotemporal clustering and visualized on the left-hemisphere inflated surface. The color coding indicates the temporal extent of the clusters. The time ranges indicate analysis time windows after the onset of the adaptor sound.
Figure 1.6: Connectivity Between the Speech-Selective Cluster in the Posterior Temporal Cortex (Green Circle) and the Rest of the Cortex During Processing of Speech Sounds in the ATTEND Condition. The observed positive PSIs indicate that information is flowing from the seed ROI to the cluster (arrow). The color coding indicates the temporal extent of the cluster, and the time ranges indicate latencies after sound onset. MTG, middle temporal gyrus; PMC, premotor cortex; POp, pars opercularis (of the inferior frontal gyrus); STG, superior temporal gyrus; STS, superior temporal sulcus.
depends on auditory attention. MEG was recorded while participants, presented with paired sounds from a phonetic continuum, (1) engaged in an attention-demanding listening task wherein phoneme category information was irrelevant and (2) ignored the same sounds while watching a film without the soundtrack. Recent findings imply that frontal premotor structures, and more generally the dorsal auditory stream, support speech categorization (e.g. Alho et al., 2012; Chevillet et al., 2013; Alho et al., 2014). The results presented here corroborate these findings, revealing categorical speech representations in left inferior frontal areas. For the first time, our findings show that these category-specific representations are activated at early latencies (115–140 ms), compatible with the known speed of perceptual phonetic categorization (Salmelin, 2007; Bidelman et al., 2013). As further novel findings, we observed that these representations depend on auditory attention, correlate with the participants’ ability to categorize the phonetic stimuli, and show concurrent functional connectivity with left posterior auditory cortical areas.

*Neural Selectivity For Phoneme Category In Pars Opercularis*

Phoneme category selectivity was found in a left-hemisphere area involving pars opercularis and anterior insula (Fig. 1.3). The effect was present only when attention was directed to the auditory input, not when the subjects ignored the sounds while watching a silent film. Importantly, as the task in the attention condition diverted attention from the phonetic features, the observed category selectivity can still be considered task-independent. The relatively early latency of the effect (<140 ms) is in agreement with the proposal that access to phonological categories occurs at ~150 ms after stimulus onset (Salmelin, 2007) and suggests that these representations might drive the categorization.
Our findings are consistent with previous fMRI-adaptation studies showing categorical speech representations in speech-motor areas when the paradigm required attention to the auditory input (Chevillet et al., 2013; Lee et al., 2012; Myers et al., 2009), but not when the auditory input was ignored (Joanisse et al., 2007). A recent MEG study reported adaptation effects indicative of categorical speech representations in the left posterior superior temporal sulcus/posterior superior temporal gyrus between 430–500 ms after sound onset (Altmann et al., 2014). However, the participants were actively discriminating the sounds, which may explain the discrepancy with our results, considering that the left superior temporal gyrus is more strongly activated in sublexical speech perception experiments where participants engage in an active decision task compared to passive listening or non-phonetic listening tasks (Turkeltaub and Coslett, 2010). Also, due to the long latency of categorical speech representations in Altmann et al. (2014), the study does not decisively answer the question which cortical region(s) drive the categorization rather than reflect categorical speech representations as a result of projections from other regions. The latter is a concern especially in view of the fact that late evoked responses (>200 ms) seem to depend on cortical feedback connections (Garrido et al., 2007).

In interpreting the results, it has to be considered that in case any covert phonetic categorization occurred in the ATTEND condition, the discrimination of within-category sounds (33S) would presumably be more difficult than that of between-category sounds (33D), therefore making it possible that the differential activity in pars opercularis and anterior insula observed in the 33S vs. 33D contrast represents a difference in the
engagement of executive functions rather than processing of phoneme categories *per se.* However, this alternative interpretation is refuted by the apparent difficulty of the non-phonetic duration discrimination task in the ATTEND condition (as reflected in the hit rate and response times, see section 3.1) together with the fact that no differences were observed in the performance (i.e. hit rate or response times) between the 33S and 33D sound pairs in this task. Further, the early latency (115–140 ms) makes it unlikely that the difference between 33S and 33D could reflect any postperceptual processes, such as response selection or decision-making.

Supporting the attention-dependence of the early premotor contribution to speech processing, we observed stronger responses to speech sounds in the left inferior frontal gyrus between 120–170 ms in the ATTEND compared to the IGNORE condition (Fig. 1.5). Previous studies have reported similar findings; for example, a recent TMS-MEG study demonstrated that involvement of articulatory motor areas in the early processing of acoustic-phonetic features of speech depended on auditory attention, whereas longer-latency auditory-motor interaction (>170 ms) occurred even when the subjects were told to ignore the sounds and to focus on watching a silent film (Möttönen et al., 2014). Another study showed that the inferior frontal gyrus was engaged in the processing of degraded speech only when the subjects were attending to the stimulation (Wild et al., 2012).

It is worth noting that these results don’t perfectly fit into the response profiles indicative of categorical specificity. The greatest release from adaptation was seen in the 33D condition while the 67D condition seemed to show more adaptation. Categorical
selectivity would suggest a release from adaptation in both conditions, and likely increased release from adaptation in the 67D condition, not less. The response found in our data could be due to the differences in response latencies across conditions and the fact that our cluster analysis focused on the 33S and 33D conditions as they were seemingly the most interesting. That being said, despite an apparent adaptation response in the 67D condition this response was not significantly different from that of the 33D response.

*Neural Selectivity For Acoustic-Phonetic Features In Temporal Cortex*

Selectivity to acoustic-phonetic features (without category-boundary effects) was found in the left temporal cortex (Fig. 1.3). An area just anterior to the auditory core (Heschl’s gyrus) exhibited sharp acoustic-phonetic selectivity (i.e. release from adaptation for all except identical sounds). This finding is compatible with a previous fMRI-adaptation study showing acoustic-phonetic selectivity in an anterior region of auditory cortex (Chevillet et al., 2013) and with findings showing sharper tuning to sound frequency in anterior than in posterior auditory cortical areas (Jääskeläinen et al., 2004). However, similarly to the observed phoneme category selectivity, the effect was present only when the auditory input was attended to, therefore further highlighting the importance of auditory attention in the early processing of speech stimuli. Broad acoustic-phonetic selectivity (i.e. release from adaptation only in 67D) was observed, regardless of attention to the auditory input, in extensive lateral temporal lobe areas, involving parts of superior temporal gyrus, superior temporal gyrus, middle temporal gyrus, and inferior temporal sulcus. When interpreting the source localization of MEG signals, one has to take into
account the selective sensitivity of MEG to sulcal sources (although see Hillebrand and Barnes, 2002).

Together, these results are consistent with studies indicating that phonetic recognition occurs in left anterolateral superior temporal cortex (i.e. the ventral auditory stream) (Binder et al., 2000; Scott et al., 2000; Leaver and Rauschecker, 2010; DeWitt and Rauschecker, 2012). Our findings also support a functional distinction between temporal and frontal areas, in that temporal areas exhibit sensitivity to acoustic variation both within and between phonetic categories, whereas frontal areas show selectivity to phoneme category with insensitivity to within-category acoustic variation (e.g. Myers, 2007; Myers et al., 2009).

**Correlation Between Neural Phoneme Category Selectivity And Behavioral Categorization**

The phoneme category selectivity found in our MEG measurements was positively correlated with performance in a categorization test conducted after the MEG session (Fig. 1.4), thus implying that the underlying neural tuning properties predict behavioral phoneme categorization performance. Similarly, Chevillet et al. (2013) reported a correlation between behavioral categorization and neural phoneme category selectivity in the left premotor cortex (located adjacent to pars opercularis). These results are in agreement with TMS studies demonstrating that disturbing the speech-motor system impairs speech sound discrimination (Meister et al., 2007; Möttönen and Watkins, 2009; Sato et al., 2009; D'Ausilio et al., 2012; Grabski et al., 2013) as well as with our recent MEG studies showing that speech categorization during scanning correlates positively with left
premotor cortex response amplitudes (Alho et al., 2012) and with functional connectivity strength between left premotor cortex and auditory cortex (Alho et al., 2014).

Although the correlation between neural phoneme categorical selectivity and behavioural categorization has persisted in our analyses despite the removal of outliers, at face value the data are not completely convincing. It would perhaps then be beneficial to include alternative means resampling the data, e.g. jackknifing or weighted correlations, to ensure the relationship we have observed is genuine.

*Effective Connection From Posterior Temporal Lobe To Pars Opercularis*

Directed functional connectivity was observed from the left posterior temporal lobe to pars opercularis/premotor cortex between 120–170 ms after sound onset (Fig. 1.6). We interpret this to reflect the engagement of the dorsal auditory stream during speech processing and, consequently, the observed categorical speech representations in pars opercularis to reflect the result of sensorimotor integration whereby speech sounds are mapped onto the motor articulations likely to have produced them. This interpretation is supported by combined fMRI and diffusion tensor imaging (DTI) studies demonstrating a dorsal connection from posterior superior temporal cortex to pars opercularis (Brodmann area 44) and premotor cortex (Frey et al., 2008; Frey et al., 2014), which supports sensorimotor mapping of sound-to-articulation, while a ventral pathway connects the anterior superior temporal cortex with pars triangularis (Brodmann area 45) (Frey et al., 2008) and may be responsible for sound-to-meaning mapping. Also, a recent study using combined electrical microstimulation and fMRI to investigate primate frontotemporal...
effective connectivity found that while stimulating the frontal operculum (corresponding to human pars triangularis and pars orbitalis of inferior frontal gyrus) activated the anterior temporal lobe, stimulation of the more dorsal area F5 (corresponding to human pars opercularis and ventral premotor cortex) activated more posterior temporal areas (Petkov et al., 2015).

Thus, a possible scenario to account for the present findings is that the categorical speech representations in pars opercularis constrain the acoustic-phonetic interpretation of the phonetic continuum into discrete categories via feedforward and feedback connections between auditory and premotor areas (Callan et al., 2004; Davis and Johnsrude, 2007; Hickok and Poeppel, 2007; Rauschecker and Scott, 2009; Rauschecker, 2011; Schwartz et al., 2012; Bornkessel-Schlesewsky et al., 2015).

In addition to pars opercularis, we found categorical speech representations also in anterior insula, which has long been implicated in articulatory planning (Dronkers, 1996), auditory processing of vocalizations (Sander and Scheich, 2005; Remedios et al., 2009), as well as integration of auditory and motor information (Mutschler et al., 2009). Thus, the present data help to reinvigorate the view that the insular cortex represents an important, yet often overlooked, brain region involved in speech processing (Ardila et al., 2014; Oh et al., 2014).
Conclusions

Our results are very clear in identifying the left pars opercularis as the most distinct brain region for phoneme categorization. Indeed, the left pars opercularis exhibited early-latency (115–140 ms) phoneme category selectivity that was positively correlated with behavioral phonetic categorization. Taking advantage of our recording techniques, this finding demonstrates—for the first time—phoneme-category specific responses in the left pars opercularis that occurs early enough to coincide with phonetic perception. Furthermore, concurrent functional connectivity was observed between the left pars opercularis and posterior auditory cortical areas, which implies that phoneme category invariance arises from dorsal-stream-mediated integration of auditory and motor information. At the same time, due to its proximity with pars triangularis (Brodmann area 45), which constitutes the endpoint of the auditory ventral stream, pars opercularis (Brodmann area 44) is in an excellent position to enable the transformation between an articulatory and a phonological code in the inferior frontal gyrus, as previously proposed by Rauschecker and Scott (2009). Finally, as an additional novel result, we found that the category-specific representations in the left pars opercularis depend on auditory attention.
Chapter 2: Distinct Networks of Neural Activation Subserve Encoding and Storage of Novel and Familiar Sound Sequences

Introduction

Hearing a story, listening to music, or getting directions all require the accumulation of auditory information over time. As each new bit of information is processed it is assimilated with previous knowledge while the listener continues to take in new information. Representations are formed based on previous experience, and the listener can solidify this representation with each new presentation, while also predicting how the remaining sequence may progress. When a series of sounds are repeatedly presented in a specific order they become associated with each other as parts of a sequence. The perception of a meaningful series of sounds requires the learning and storage of these ordered sound sequences, but the exact mechanisms for this are largely unknown.

Both speech and music require the processing (production and perception) of long sequences of sounds that are related to movement. In the case of speech it is the movement of the tongue, lips, teeth and throat to create the sounds of speech. For music it is the coordinated movement of the fingers, limbs, mouth, or body to manipulate the instrument of choice (which includes the singing voice). This tight coupling between the auditory system and the motor system is widely documented (Zatorre et al., 2007), and the dorsal auditory stream has been implicated in sensorimotor integration and concatenation of sequential events (Rauschecker, 2011). Subcortical motor areas such as the basal ganglia and cerebellum have been linked to the learning of newly associated sound sequences but
not the storage of well-known songs (Leaver et al., 2009). Auditory sequence processing seems to depend on generalized sequencing abilities borrowed from the motor system (Janata and Grafton, 2003; Zatorre et al., 2007) but the functional and anatomical transition from areas of the brain that encode the patterns of newly presented sequences to areas that represent and store frequently repeated and well known sequences as objects has yet to be shown in humans.

The auditory ventral stream, dubbed the “what” pathway, has been shown to process auditory objects in a hierarchical fashion, proceeding anteriorly as object complexity increases (Leaver and Rauschecker, 2010). Auditory objects, similar to visual objects, are sounds or classes of sounds with specific auditory signatures making them easily identifiable and classifiable (Leaver & Rauschecker, 2010). As such, the mental repository of such objects would be in declarative memory, while the relationships between and sequences of objects would be in procedural memory (Cohen & Squire, 1980). This pattern of hierarchical processing has also been shown to reflect the timescale during which auditory stimuli are presented and to which certain areas are tuned. As the temporal receptive window increased, brain activations progress hierarchically from primary sensory areas to higher cortical areas (Lerner et al., 2012). While music and language are not strictly analogous, there are some similarities worth considering, such as how activations for phonemes and words migrate rostrally from middle superior temporal gyrus to anterior superior temporal gyrus, respectively (DeWitt and Rauschecker, 2012), and that syntactical structure has also been shown to elicit activation in both anterior superior temporal gyrus and posterior superior temporal gyrus (Bornkessel et al., 2015;
Friederici et al., 2010). This suggests that there may be both a dorsal-stream and ventral-stream component to syntax processing specifically and to sequence processing in general, since a major component of syntax is the sequence of words in language.

Statistical learning paradigms have shown that streams of visual or auditory objects can be segregated into meaningful units (e.g. words) based on the transition probabilities between stimuli (Saffran et al., 1999; Hauser et al., 2001; McNealy et al., 2006). As stimulus pairs are repeatedly presented in succession so grows the strength of their association. These experiments can be done as streams of stimuli or within the context of a task such as the serial reaction time task (Schapiro et al., 2012), and have been shown to require the medial temporal lobe for learning and storage of these learned sequences (Albouy et al., 2008, Rose, et al., 2011; Schapiro et al., 2014). So while easily identifiable or well-known sounds can be processed as discrete auditory objects as part of the ventral stream, novel sounds may require dorsal stream structures for encoding (Rauschecker, 2014; Schapiro et al., 2014; Hickok, 2007).

The current study utilized a progressively presented musical sequence to create a continuum of repetition-based novelty and familiarity within a single stimulus. Using functional magnetic resonance imaging (fMRI) we measured blood oxygen level dependent (BOLD) activity associated with novel (few repetitions) and familiar (many repetitions) segments of the musical sequence. We hypothesized that segregated networks of activity would be associated with novel and familiar sequences within this paradigm: novel sequences should activate sequence-encoding structures in the auditory dorsal stream,
which should include a role for subcortical motor areas such as the basal ganglia and the cerebellum. Familiar sequences should include activation of object-representation structures in the auditory ventral stream with potential for recruitment of medial temporal lobe structures to facilitate auditory object representation and long-term memory formation.

**Materials and Methods**

**Participants**

18 volunteers (15 male; 17 right handed; ages 21-39 years) were recruited from Aalto University and Helsinki University in Helsinki, Finland. 16 participants were native Finnish speakers, 1 participant was a native Russian speaker, and 1 participant was a native Spanish speaker. All participants spoke English as a second language with high proficiency and reported normal hearing.

**Stimuli**

Auditory stimuli consisted of four 30-second musical sequences presented at 180 beats per minute (bpm). A random number generator (www.random.org) was used to create random, atonal sequences using the notes from C₄ (Middle C) and C₅. Sequences were then constructed using GarageBand Software (Apple, Inc), divided into 30 one-second segments, and edited with Audacity audio editing software (www.audacity.sourceforge.net). Visual Stimuli consisted of 12 greyscale Gabor patches of uniformly incremental spatial frequency that were created using a custom script in
MATLAB (The Mathworks, Inc). Auditory and visual stimuli were combined into multimodal videos using VideoPad video editor (NCH Software).

Stimulus Presentation

One-second auditory sequence segments were presented progressively such that on the first trial only Segment 1 (S1) was heard; on the second trial S1 was presented followed by S2 as a continuous 2-second sequence; and on the third trial S1, S2, and S3 were presented as a continuous 3-second sequence and so on (Fig. 2.1). This progressive presentation continued through the 30th and final trial where the participants were presented with the 30 segments as one continuous 30-second sequence.

Each trial was randomly assigned between 1 and 4 musical events characterized by the presence of a note exactly 7 half steps above a coincident note from the base sequence discussed above (interval). Additionally, the visual presentation of each trial included a Gabor patch in the center of the visible field for fixation. This patch changed spatial frequency between 1 and 4 times per trial, independent of the auditory stimulus.

Behavioral Tasks

While in the scanner, all participants were presented with all four of the progressive musical sequences, but were asked to attend to perform either an auditory or visual attention task orthogonal to the sequence learning (Fig. 2.2). The auditory attention task consisted of counting the number of coincident musical events in each trial. Each trial is
Figure 2.1: Example of a Progressively Presented Musical Sequence Consisting of 5 Subsequence Chunks. A target sequence, ABCDE, is presented in a progressively increasing fashion such that on the first trial the first sub-sequence (A) is presented and on each subsequent trial an additional sub-sequence is added to the previously heard sequences. On trial 5 the target sequence is played in its entirety with segment A having been presented 5 times and segment E being presented for the first time, thus establishing the repetition based continuum of novelty/familiarity within a single sequence.
Figure 2.2: Experimental Design and Stimulus Schematic. Each run consisted of a 10 second pre-scan period where subjects were prompted to attend and complete either the auditory or visual task. During the scanning period the subjects were presented with both auditory (musical sequences) and visual (Gabor filters) stimuli during 30 trials of increasing length where the length of the nth trial was n seconds. Between trials was a 5 second response period where subjects indicated whether or not the number of target events in either the auditory or visual attention task (max = 4) was odd or even.
followed by a 5-second response period where the participants were prompted to indicate if the number of coincident intervals was odd or even. The visual attention task was identical to the auditory task except that participants were prompted to count the number of different Gabor patches presented in each trial and indicate whether that number is odd or even. There were between 1 and 4 total events for either task in any trial except those trials whose length was the limiting factor.

**MRI Protocol**

Images were obtained using a 3T MAGNETOM Skyra whole-brain body scanner (Siemens Healthcare, Erlangen, Germany), using a standard 32-channel head-neck coil. A continuous acquisition paradigm was used in which functional EPIs were collected consecutively throughout the experiment. Functional runs consisted of 325 volumes each (TR, 2s; TE, 24 ms; voxel size, 3.4 x 3.4 x 3.0 mm³).

**fMRI Analysis**

All analysis was carried out using FEAT in FSL (FMRIB). Functional images from each run were motion corrected, spatially smoothed using a three-dimensional 2.5 mm³ Gaussian filter, and high-pass filtered to remove low-frequency events. Preprocessed functional images were coregistered with their corresponding high-resolution MPRAGE anatomical images and interpolated into MNI152 standard space.

Parametrically weighted GLM analyses were conducted to assess the relationship between either sequence novelty or familiarity with the BOLD-fMRI signal. Weights of the
GLM predictors were adjusted to reflect novelty or familiarity as a function of the number of times a stimulus had been repeated. Z statistic (Gaussianized T) images were thresholded using clusters determined by $Z > 2.3$ and a (corrected) cluster significance threshold of $p = 0.05$.

**Final Trial Analysis:** The final 30-second trial of each run was analyzed with either a 2 or 6-second non-overlapping sliding window. GLM predictors were adjusted to reflect the average number of repetitions for sequences contained within the window in a linear fashion for the following analyses: 2-second temporal window analysis for novelty (-1.0, -0.818, -0.636, -0.455, -0.273, -0.091, 0.091, 0.273, 0.455, 0.636, 0.818, 1.0), 6-second temporal window analysis for novelty (-1.0, -0.333, 0.333, 1.0), 6-second temporal window analysis for familiarity (1.0, 0.333, -0.333, -1.0). Preliminary results for the 2-second temporal window analysis for familiarity did not yield significant results (data not shown) and were excluded from the current analysis.

**Intermediate Sequence Analysis:** A 6-second subsequence was traced across trials from its first to final presentations. This sequence began at 6 seconds post-stimulus onset in each trial (to account for hemodynamic onset effects), was first presented in its entirety in trial 12, and had its final presentation in trial 30 for a total of 19 presentations. BOLD fMRI signal associated with this sequence was averaged across trials into 5 groups according to the number of repetitions. Groups 1 through 5 averaged 2.5, 6.5, 10, 13.5, and 17.5 repetitions, respectively. GLM predictors were adjusted to reflect increased repetition in a linear fashion from group 1 to group 5 over 15 repetitions (-2.0, -1.0, 0.0, 1.0, 2.0), from
group 1 to group 4 over 11 repetitions (-1.0, -0.333, 0.333, 1.0) from group 2 to group 5 over 11 repetitions (-1.0, -0.333, 0.333, 1.0), and from group 3 to group 5 over 7.5 repetitions (-1.0, 0.0, 1.0). Preliminary results comparing parametric increase from groups 1 to 2, 1 to 3, and 4 to 5 were not significant (data not shown) and were excluded from the current analysis.

**Results**

We measured parametrically increasing neural responses associated with sequence novelty and familiarity as defined by the number of times a sequence had been presented. The gradient from novel to familiar was approached in two ways: 1) The final 30-second trial of each run was analyzed with either a 2 or 6-second non-overlapping sliding window; and 2) A 6-second subsequence was traced across trials from its first to final presentation.

Parametrically increasing neural responses associated with novel stimuli in the final trial were widespread at both the temporal window durations. At the 6-second timescale novel stimuli were associated with increased activations in the left inferior parietal lobule, sensorimotor areas, medial frontal gyrus, lingual gyrus, and bilateral thalamus (Fig. 2.3B). The 2-second timescale showed additional activation in bilateral inferior frontal gyrus, insula, claustrum, and left globus pallidus (Fig. 2.3A,B).

Parametrically increasing neural responses associated with familiar stimuli in the final trial was observed in bilateral medial prefrontal cortex and anterior cingulate (Fig. 2.4). Tracing a 6-second subsequence across multiple trials through repeated presentations revealed distinct networks of areas corresponding to increasing numbers of repetitions.
Figure 2.3. Parametric Analysis of Novelty Across the Final 30-Second Trial. Cortical projections of the 2 second temporal window analysis (A) and slice projections of 2 and 6 second temporal windows analysis (B), shown in red and blue, respectively were used to analyze the final 30-second trial of the stimulus. Z-statistic images were thresholded at $Z>2.3$ at $p<.05$ significance.
Figure 2.4. Areas Showing a Parametrically Increased Activation in Medial Frontal Cortex to Repeated Stimuli Across the Final 30-Second Trial. 6-second temporal window analysis revealed localized increases in activation as target sequence was repeated. Z-statistic images were thresholded at $Z>2.3$ at $p < .05$ significance.
Trials were binned into 5 groups averaging 2.5, 6.5, 10, 13.5, and 17.5 repetitions of the target sequence, respectively. Areas showing increasing activity across the full range of presentations (15 repetitions on average, from group 1 to group 5), included left middle temporal gyrus, superior temporal gyrus, fusiform gyrus, parahippocampal gyrus, superior frontal gyrus, and right anterior cingulate and medial frontal gyrus. Comparisons of parametrically increasing activation from group 1 to 2 and from group 1 to 3 were not significant, but from group 1 to group 4 showed a similar activation pattern as for 1 to 5 (data not shown). Areas showing increasing activity across 11 repetitions (group 2 to group 5) included similar medial temporal area activation as for 15 repetitions but also included activity in the left amygdala, right superior parietal lobe, and bilateral pre/post-central gyrus. Areas showing increased activity across 7.5 repetitions were found only in the bilateral pre/post-central gyrus and right superior parietal lobe.

Discussion

The current study investigated the neural mechanisms underlying auditory sequence perception, encoding, and storage. The novel paradigm used for this experiment utilized the progressive presentation of an atonal musical sequence to create a continuum of novelty and familiarity, as defined by the number of repetitions. This continuum was approached within the sequence in multiple dimensions, e.g. linearly in the final trial, and abstractly across trials. We hypothesized that auditory dorsal stream structures and basal ganglia would play a role in the encoding of novel sequences, while the perception of
Figure 2.5. Areas Showing a Parametric Increase in Activation Over (A) Final 10, (B) Final 13.5, and (C) All 17.5 Repetitions of the Target Sequence. Target sequence was traced across trials across a total of 21 presentations and analyzed in trial groups resulting in 3 distinct patterns of parametrically increasing activity across an average of 10, 13.5 and 17.5 repetitions relative to the final presentation. Z-statistic images were thresholded at $Z > 2.3$ at $p < .05$ significance.
Figure 2.6. Intermediate Chunk Familiarity in Axial (A) and Sagittal (B) Slice Images With Overlaid Images of Parametrically Increasing Responses to Final 10 (Green), Final 13.5 (Blue), and All 17.5 (Red) Repetitions. Activation maps from Figure 3 overlaid on slice images to further demonstrate the progression of activation across brain areas. Z-statistic images were thresholded at Z>2.3 at p < .05 significance.
familiar sequences as auditory objects could be mediated by the auditory ventral stream and prefrontal areas.

*Encoding Of Novel Sequences*

Listening to novel sequences revealed a broad network of active brain areas at both the 2- and 6-second timescales. This network included large clusters in superior temporal areas, inferior parietal lobule, somatosensory, motor, and premotor cortices, as well as supplementary motor and prefrontal cortex (Fig. 2.3). These are all major hubs of the dorsal auditory stream and consistent with our hypothesis that the dorsal stream would play a role in the encoding of novel sequences. Additionally, at the 2-second timescale additional clusters were seen in the left globus pallidus of the basal ganglia and the right inferior frontal gyrus.

*Basal Ganglia And Inferior Frontal Gyrus:* The basal ganglia and inferior frontal gyrus were active only during the 2-second timescale analysis of novel sequences. This suggests that these areas have shorter temporal receptive windows (Lerner et al., 2011) and are tuned to the short, novel stimuli as they are first being presented for encoding. This role for the basal ganglia and inferior frontal gyrus as processors of novel sound sequences has been shown previously in work from our lab (Leaver et al., 2009). Basal ganglia activity has been shown to arise during the early stages of sequence learning (Bornstein et al., 2012; Pasupathy and Miller, 2005). Additionally, the current study provides evidence that, when presented with novel auditory sequences, the auditory system may borrow the sequencing capabilities of the motor system for the concatenation of novel, repeated sounds into
longer sequences. These structures could serve as the building blocks for long-term storage of sequences and ultimately help to turn them into object representations in other areas.

Processing And Storage Of Familiar Sequences

Repeated presentation of sound sequences resulted in increasing activation in the medial temporal lobe, medial prefrontal cortex, somatosensory areas, and superior parietal lobule, but only at the 6-second timescale (Fig. 2.5, 2.6). This suggests that areas responsible for long term storage, recognition, and representation of sound sequences are tuned to longer sequences that have already been encoded and concatenated by the novelty-processing areas such as the basal ganglia and inferior frontal gyrus.

Medial Temporal Lobe: Signal in the medial temporal lobe, including the hippocampus, parahippocampal gyrus, and the amygdala were shown to have increased activations as a specific sequence was increasingly repeated. The medial temporal lobe structures are associated with the formation and storage of long-term memories and object recognition (Bussey & Saksida, 2007; Eichenbaum, 2012). Medial temporal lobe lesions result in temporally graded retrograde amnesia, such that older memories are less impaired than more recent memories (Squire, 2009). This suggests that while the medial temporal lobe is critical in the formation of memories, it should also facilitate the representation of these memories in other cortical targets (Ullman, 2004). Similarly, studies of implicit sequence learning has shown that patients with medial temporal lobe lesions show impaired sequence learning regarding declarative knowledge of the sequence, while procedural sequence learning (e.g. learning to play a melody on the piano) is
conserved (Schapiro et al., 2012; Valtonen et al., 2014). Although the medial temporal lobe has been shown to activate and bind information early and quickly (Schapiro, 2014), medial temporal lobe and basal ganglia have also been shown to play complementary roles in statistical and sequence learning such that the basal ganglia activate early in learning, while the medial temporal lobe activates later (Keele, 2003; Davis and Staddon, 1990; Bornstein, 2012). So while our data do seem to conflict with the conception of early medial temporal lobe and late basal ganglia facilitation of declarative and procedural memory, respectively (Ullman, 2004), we do find ample evidence to support the early basal ganglia/late medial temporal lobe activity we show. This discrepancy does not invalidate either model, nor does it make them mutually exclusive. Instead, perhaps, these contradictory findings highlight the diversity and context-specificity of function in both the medial temporal lobe and basal ganglia.

Medial Frontal Lobe: Increased activations in the medial frontal lobe were also observed for increased repetitions of target sequences. It is possible that the medial frontal lobe plays an important role in the long-term representation of sound sequences as auditory objects. The medial frontal lobe, specifically ventromedial prefrontal cortex is also an area associated with a putative cortico-striatal circuit underlying the perception or attenuation of the tinnitus percept (Rauschecker, 2010; Rauschecker et al., 2015). More generally, this area is suggested to be a location for cortical, top-down or descending inhibition of aberrant sensory signals in healthy individuals. In this capacity, this area could also facilitate the encoding of auditory sequences by acting as a signal monitor sensitive to
sequence non-conformity. The increased activations in this area could also indicate increased working memory load as target sequence length increases (Muñoz et al., 2009).

**Anterior Shift In Processing:** The current study partially replicated the findings of previous work in the lab (Leaver et al., 2009) showing an anterior shift in processing as stimuli progress from novel to familiar. While the activation maps are not direct reflections of each other, the pattern of activation progressing from sequence-processing areas, like the basal ganglia, to the memory and frontal areas, like the medial temporal lobe and medial and inferior frontal lobe. These similarities suggest that long term representations of sound sequences may reside in frontal regions, as mediated by memory areas and sequence processing structures.

**Limitations**

The novel design utilized in the current study was done so with the intention of creating a gradient of repetition-based novelty/familiarity within a musical sequence that, to the best knowledge of the authors, had not been seen previously. This design essentially flipped the location of novel and familiar stimulus sequences, familiar first and novel last, in each trial, thus reducing the effect of stimulus-specific adaptation in the analyses of highly repeated familiar sequences. The nature of this stimulus is biased heavily toward novel, low-repetition events, as all sequences are at some point novel, while only some progress to a high level of repetition-based familiarity. As such, we were not able to use the entire dataset for our analyses, but we were able to approach the analysis of novelty and familiarity in both a linear single-trial format and a more abstract across-trials format,
finding consistent results from each method. This progressively presented design adds complexity and has its limitations, but we are confident that we took appropriate measures to make sound interpretations of the results.

It is also worth noting that the behavioral tasks in our paradigm were orthogonal to the perceptual sequence variables of interest and did not reflect learning or memory of the sequences themselves. Novelty and familiarity of target sequences were operationally defined by the number of times they had been presented to subjects, low and high, respectively. Because of the implicit nature of our sequence we could not conduct such behavioral experiments prior to the image collection, as it would have revealed the nature of the experiment. While it would perhaps have been ideal to collect such data afterwards, this was not done, but this should not discount our interpretations of the observed correlations between parametrically increasing brain activations based on the number of stimulus presentations.

Conclusions

With increased repetition comes changing representation. Using a progressively presented auditory sequence our results indicate that novel sound sequences are processed by dorsal-stream auditory areas as well as the basal ganglia and inferior frontal gyrus under shorter timescales, suggesting a borrowing of sequence-encoding capabilities from motor areas attuned to that task. Alternatively, highly-repeated familiar sequences are processed across longer timescales in the hippocampus, parahippocampal gyrus, amygdala, medial prefrontal cortex, temporal pole, and superior parietal lobe, suggesting a
potential transition from dorsal- to ventral-stream processing where specific sequences are stored as entities or “auditory objects” in canonical object identification and memory areas.

Although the dual-stream theories of auditory processing have tended to focus on the cortex, the basal ganglia have been proposed to play a role in the dorsal stream as sequence-encoding areas that stitch together short, novel sequences into larger chunks for further processing by other areas. Our data are consistent with this hypothesis and provide further evidence for their inclusion into dual-stream models of auditory processing. Similarly, medial temporal lobe structures have been folded into the dual stream model from a memory perspective (Ullman, 2004; Eichenbaum, 2012; Muñoz, 2009), whereas declarative and procedural memory systems have connections with visual ventral and dorsal streams, respectively. The results in the current study parallel these models and support a potential role for the inclusion of the medial temporal lobe into the auditory ventral stream as one facilitator of auditory object perception and representation.
Chapter 3: General Discussion

Objectives

The preceding work utilized speech and music paradigms to investigate the role of motor areas in speech perception and auditory sequence processing. Speech and music are well suited for the task of investigating these audiomotor interactions as they both exemplify how complex motor movements produce sounds, which then offer sensory feedback to modify future movements. Interestingly, we’ve also shown that music elicits activity in motor regions during sequencing tasks without a trained motor component. These are by no means the only ways to elicit such audiomotor interactions, but speech and music are perhaps the two most salient and distinctly human phenomena to do so. In Chapter I we discussed an MEG experiment, adapted from fMRI, which utilized a rapid adaptation paradigm to further investigate categorical speech perception in humans. The MEG approach made it possible to not only localize brain areas displaying categorical speech representations, but to also obtain latency measures for relevant activations, and establish a flow of information through our regions of interest. Chapter II discusses an fMRI music experiment where a progressively-presented musical sequence was resulting in a continuum of repetition-based novelty and familiarity. This continuum was established to investigate the transition from areas that encode unknown sequences to those areas that store known sequences. This was a novel paradigm, not without its flaws, but it allowed us to observe repetition-based processing in multiple dimensions and revealed distinct cortical and subcortical networks subservient to high and low repetition sequences.
**Motor Role In Speech Perception**

The debate regarding the motor theory of speech perception is ongoing. There is ample evidence demonstrating that there are significant activations in motor areas during receptive speech perception. These areas do reflect those ostensibly responsible for speech production, and thus the basis for the claim that speech perception recruits speech production areas as a means of simulating and perceiving the physical movements responsible for the speech (Schwartz et al., 2012). Because the production of phonemes requires discrete motor movements, it stands to reason that there are categorical boundaries that distinguish these points of articulation to result in different speech sounds. Our findings that categorical speech perception takes place in the posterior operculum and insula, areas associated with speech production and part of the auditory dorsal stream, suggest that yes, these motor areas are recruited for speech perception. What these data don’t say is if these areas and the categorical processing therein are required for speech perception. The case against the motor theory of speech perception highlights the case of conduction aphasia, where lesions in speech production areas resulting in speech deficits do not affect speech comprehension (Damasio and Damasio, 1980). This may seem a compelling argument, but it doesn’t necessarily invalidate the motor theory. First, the motor system comprises multiple areas, anatomically and functionally connected in any number of configurations. Specific lesions that produce speech deficits don’t prove that all speech production processes, subvocal though they may be, are destroyed or impaired. The brain is plastic and redundant and finds novel ways to spare function e.g. a 44 year-old man presented with virtually no brain and relatively normal neurological function (Feuillet et
Secondly, for patients with conductive aphasia that occurs later in life, the groundwork for stable perception circuitry could be established enough pre-lesion to continue functioning post-lesion, similar to how late-deaf individuals show fewer speech deficits than early-deaf (Campbell et al., 2014).

**Basal Ganglia As Auditory Dorsal Stream Structures**

The presence of basal ganglia activation in auditory processing has been repeatedly shown (Leipzig et al., 2006; Zatorre et al., 2007; Leaver et al., 2009; Rauschecker, 2014). The results in Chapter II showing basal ganglia activation at the 2-second sliding window analysis but not the 6-second window analysis is consistent with findings that basal ganglia play an early role in learning of novel sequences (Pasupathy and Miller, 2005; Leaver et al., 2009), with a temporal receptive window tuned to short sequences. This fits with our hypothesis that the basal ganglia would serve as a sequence-encoding network that work to concatenate short novel sequences into larger chunks that can progress through the hierarchy for eventual storage as auditory objects. The basal ganglia comprise many subareas, but the two specific areas that showed activation in these data were the left globus pallidus and putamen. These structures are known to play a role in coordinated movement, and in the case of the putamen, implicit learning (Leipzig et al., 2006). The sequence learning task in our study was orthogonal to the attention demanding behavioral task, which would suggest that the putamen could be facilitating implicit sequence learning. The dual stream model of auditory processing already includes premotor structures and speech motor areas in the dorsal stream (Rauschecker and Scott, 2009; Rauschecker, 2011; Bornkessel-Schlesewsky et al., 2015). Dorsal stream inclusion has also been extended to at
least conceptually include subcortical structures like the basal ganglia and the cerebellum for various tasks related to sound sequencing and sound guided action, although they have not yet made it into the current models. Connectivity between auditory structures and the basal ganglia have been established, and it seems only a matter of time before the dorsal stream is officially reimagined with them included. The data we've presented should support this.

Hierarchical Representation Of Auditory Objects

Previous work in the lab has shown that auditory speech sounds are processed hierarchically, with a posterior-to-anterior flow of processing across the superior temporal gyrus from phonemes to words to phrases (Dewitt and Rauschecker, 2012; Chevillet et al., 2013). The current findings show a posterior to anterior spread of activation as sound sequences are presented an increasing number of times. This activation starts in primary auditory areas and both cortical and subcortical dorsal stream processing areas. It then progresses anteriorly through medial temporal lobe structures including hippocampus and amygdala, the temporal pole, and medial prefrontal cortex. This pattern of activation is also consistent with hierarchical models of visual object recognition (Bussey and Saksida, 2007; Cowell et al., 2010), and suggests overlap in the processing of objects across modalities.

Medial Temporal Lobe As Auditory Ventral Stream Structures

The results discussed in Chapter II provide some evidence to suggest that the ventral stream may include processing carried out by the structures of the medial temporal
lobe. The activation of medial temporal lobe structures, though not canonical auditory areas, has been shown to play an important role in sequence learning, specifically perceptual sequence learning, and statistical learning (Watanabe et al., 2008; Turk-Browne et al., 2010; Rose et al., 2011; Schapiro et al., 2012), while patients with medial temporal lobe lesions show deficits in these same tasks (Schapiro, 2014; Valtonen et al., 2014).

Models of medial temporal lobe based memory processing have been extended to include interconnections with dorsal and ventral stream models of vision, where both dorsal and ventral streams make their way through the network of medial temporal lobe areas.

Consistent with our findings, this places the medial temporal lobe in a convenient location to subserve the transfer of information from early, dorsal stream areas to later and slower ventral stream areas.

Why don’t contemporary dual stream models incorporate medial temporal lobe structures? One possibility would be that much of the work to establish these models was done in non-human primates, namely the macaque. While macaques do serve as a near relative to humans and thus a model species for understanding neuroanatomy and function, there may still be some crucial differences. For instance, dual stream models were first proposed and validated in vision (Ishai et al., 2000; Milner and Goodale, 2008). Monkeys are extremely adept at visual tasks have a decreased capacity for auditory tasks (Scott et al., 2012, 2014). One explanation for this could be that they don’t have the same level of access to long-term memory in the auditory domain and rely on auditory working memory to complete tasks (Fritz, 2005), perhaps due to decreased connectivity between auditory areas and the medial temporal lobe or frontal lobe structures (Munoz-Lopez et al.,
Macaques are, after all, not a vocal learning species (Bolhuis et al., 2010). Vocal learners need to reproduce or improvise heard vocalizations (Doupe and Kuhl, 1999). This is a rare trait and seen in only seen in 8 unrelated species. Without the ability to perform the auditory/vocal functions typical of vocal learning species, it is perhaps not surprising that the auditory function of the macaque is limited to short term working memory where it pales in comparison to the advanced visual system used for most purposes.

**Conclusions**

The work presented here argues for the inclusion of motor and memory areas in the current dual-stream models of auditory processing. We have shown that memory systems such as the medial temporal lobe and medial frontal lobe play a critical role in the transfer of newly learned sequences into well known sequences. We have shown that the basal ganglia and multimodal parietal areas are recruited during the sequencing of novel sounds. We have also shown that inferior frontal gyrus and premotor areas show selectivity for phoneme categories. While the necessity of motor areas in receptive speech processing remains an open debate, these data taken together suggest a highly interconnected relationship between auditory, motor, and memory systems for the continued perception of speech and music.
References


Leaver AM, Van Lare J, Zielinski B, Halpern AR, Rauschecker JP (2009) Brain activation


76


