The Role of Linguistic Contrasts in the Auditory Feedback Control of Speech

by

Caroline A. Niziolek

B.S. Brain and Cognitive Sciences
Massachusetts Institute of Technology, 2005

SUBMITTED TO THE DIVISION OF HEALTH SCIENCES AND TECHNOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN SPEECH AND HEARING BIOSCIENCE AND TECHNOLOGY
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
and
HARVARD UNIVERSITY

September 2010

Signature of Author: _____
Harvard-MIT Division of Health Sciences and Technology
September 1, 2010

Certified by: _____
Frank H. Guenther, Ph.D.
Professor of Cognitive and Neural Systems, Boston University
Thesis Supervisor

Accepted by: _____
Ram Sasisekharan, Ph.D.
Director, Harvard-MIT Division of Health Sciences & Technology
Edward Hood Taplin Professor of Health Sciences & Technology and Biological Engineering

The Role of Linguistic Contrasts in the Auditory Feedback Control of Speech

by

Caroline A. Niziolek

Submitted to the Harvard-MIT Division of Health Sciences and Technology
on September 1, 2010 in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy in
Speech and Hearing Bioscience and Technology

ABSTRACT

Speakers use auditory feedback to monitor their own speech, ensuring that the intended output matches the observed output. By altering the acoustic feedback signal before it reaches the speaker's ear, we can induce auditory errors: differences between what is expected and what is heard. This dissertation investigates the neural mechanisms responsible for the detection and consequent correction of these auditory errors.

Linguistic influences on feedback control were assessed in two experiments employing auditory perturbation. In a behavioral experiment, subjects spoke four-word sentences while the fundamental frequency (F0) of the stressed word was perturbed either upwards or downwards, causing the word to sound more or less stressed. Subjects adapted by altering both the F0 and the intensity contrast between stressed and unstressed words, even though intensity remained unperturbed. An integrated model of prosodic control is proposed in which F0 and intensity are modulated together to achieve a stress target.

In a second experiment, functional magnetic resonance imaging was used to measure neural responses to speech with and without auditory perturbation. Subjects were found to compensate more for formant shifts that resulted in a phonetic category change than for formant shifts that did not, despite the identical magnitudes of the shifts. Furthermore, the extent of neural activation in superior temporal and inferior frontal regions was greater for cross-category than for within-category shifts, evidence that a stronger cortical error signal accompanies a linguistically-relevant acoustic change. Taken together, these results demonstrate that auditory feedback control is sensitive to linguistic contrasts learned through auditory experience.

Thesis Supervisor: Frank H. Guenther
Title: Professor of Cognitive and Neural Systems

Caroline Niziolek
77 Massachusetts Ave.
MIT Room 36-569
Cambridge, MA 02139
carrien@mit.edu

EDUCATION

2005–2010	Ph.D., Speech and Hearing Bioscience and Technology Massachusetts Institute of Technology, Cambridge, MA.
2001–2010	B.S., Brain and Cognitive Sciences Massachusetts Institute of Technology, Cambridge, MA.

EXPERIENCE

2005–2010	Graduate Research Assistant MIT Speech Communication Group, Cambridge MA Boston University Speech Lab, Boston, MA
2007–2008	Teaching Fellow, Department of Neurobiology NB101. Auditory Neurobiology of Language and Music Harvard University, Cambridge, MA
2003–2005	Undergraduate Research Assistant MIT Computational Cognitive Science Lab, Cambridge, MA
2004	Research Assistant Institut de Neurosciences Cognitives de la Méditerranée, Marseille, France

AWARDS

2005–2009	NIH Training Grant
2008	Martha Gray Prize for Excellence in Research Harvard-MIT Health Sciences and Technology Forum Best Poster: Imaging and Optics
2004	Kelly-Douglas Traveling Fellowship

PUBLICATIONS

Patel, R., **Niziolek, C.**, Reilly, K.J., and Guenther, F.H. (submitted). Prosodic compensations to pitch perturbation during running speech. *Journal of Speech, Language, and Hearing Research.*

Niziolek, C. and Guenther, F.H. (in preparation). Phonetic categories influence auditory feedback control of speech.

POSTERS & PRESENTATIONS

Patel, R., **Niziolek, C.**, Reilly, K.J., and Guenther, F.H. (2010) Prosodic compensations to pitch perturbations in running speech. Oral presentation given at the *Fifteenth Biennial Conference on Motor Speech*, Savannah, Georgia, 2010.

Niziolek, C. and Guenther, F.H. (2009). The influence of perceptual categories on auditory feedback control during speech. Poster presented at the 15th Meeting of the Organization for Human Brain Mapping, San Francisco, California.

Patel, R., Campellone, P., Reilly, K.J., **Niziolek, C.**, and Guenther, F.H. (2008). Prosodic compensations to pitch perturbation during running speech. Poster presented at the Fourteenth Biennial Conference on Motor Speech, Monterey, California.

ACKNOWLEDGEMENTS

The research described in this dissertation was supported by the National Institute on Deafness and other Communication Disorders (R01 DC002852; PI: Frank H. Guenther). The neuroimaging experiment was carried out at the Athinoula A. Martinos Center at the McGovern Institute for Brain Research, MIT.

First and foremost, my deepest gratitude goes to my thesis advisor, Frank Guenther, who has been an exceptional mentor and teacher. I'm indebted to Frank for his encouragement, patience, friendship, generosity, and marvelous barbecuing. I'm also especially grateful to Joseph Perkell, who served as my thesis chair with a fine attention to detail and a sharp ear for grammar. Joe has given me a warm welcome to his lab, his home, and his sailboat, taking a keen interest in my research and acting as a second advisor to me.

I would also like to thank the rest of my committee members for their interest in and backing of my work. Edward Flemming is the best kind of linguist—the kind who knows math (and the kind who follows his students' research with great thoughtfulness). Jennifer Melcher has seen me through all my years in the Speech and Hearing Bioscience and Technology program with support and encouragement, and I count her as a true role model. I'm extremely grateful to my entire committee for putting in the time and energy to guide me through the end of my graduate training.

The Speech Lab at Boston University was a great community I'm sorry to be leaving. I'm indebted to Jon Brumberg, Simon Overduin, Maya Peeva, and Jenn Segawa for their patient help and company through late-night fMRI sessions. Jason Tourville and Elisa Golfinopoulos deserve a special thank-you for their valuable guidance and expertise (in addition to doing all of the above). At MIT, I would like to acknowledge Shanqing Cai, who provided me with office company, homegrown ROI analysis code, and soldering expertise (I'm a little surprised he is not graduating before me.) Warm thanks to Satra Ghosh for his generous help in—well, *everything*—and for hooking me up with the Python pipelines that become instrumental to my analysis. At Northeastern, Rupal Patel dedicated many hours of time to discussion, analysis, and writing, and was like a second advisor to me.

I would also like to acknowledge the scientists responsible for the fifteen pages of research cited at the end of this document, without whose contributions this work would never be possible.

I am sincerely grateful to my friends for letting me act the hermit during the writing of this thesis while occasionally providing much-appreciated distraction. Thanks especially to Marissa Cheng for lending me her Adobe-laden laptop; to Keith Winstein for arguing about *p*-values and for printing this dissertation after I'd fled to the west coast; and to Josh Mandel for encouragement, entertainment, and countless late-night talks on confidence intervals and perceptual theories that ultimately motivated me to follow through. Finally, I'd like to thank my parents for giving me the freedom and support to do the work that I love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Abbreviations

A1	primary auditory cortex
AAL	automated anatomical labeling
ACC	anterior cingulate cortex
BA	Brodmann area
BOLD	blood-oxygenation-level dependent
DIVA	directions into velocities of articulators
DTI	diffusion tensor imaging
EEG	electroencephalography
EPI	echo-planar imaging
F0	fundamental frequency
F1	first formant
F2	second formant
fMRI	functional magnetic resonance imaging
FMRIB	Oxford Centre for Functional MRI of the Brain
FSL	FMRIB software library
HRF	hemodynamic response function
IFg	inferior frontal gyrus
MEG	magnetoencephalography
MIS	motor-induced suppression
MMF	mismatch field
MMN	mismatch negativity
MNI	Montreal Neurological Institute
MTg	middle temporal gyrus
pSTg	posterior superior temporal gyrus
SIS	speaking-induced suppression
SMg	supramarginal gyrus
SPM	statistical parametric mapping
STg	superior temporal gyrus
TA	acquisition time
TE	echo time
TR	repetition time
VOT	voice onset time

"And who in time knows whither we may vent
The treasure of our tongue, to what strange shores
This gain of our best glory shall be sent
T'enrich unknowing nations with our stores?
— Samuel Daniel, *Musophilus*, 1599

"All language, at some level, is body language."
— Roy Blount Jr.

CHAPTER I

INTRODUCTION

The primary goal of speech is communication. Speakers use articulatory movements to produce an acoustic signal, while listeners use their auditory systems to decode a linguistic message from this signal. This process of speech production, transmission, and perception is often referred to as the *speech chain*: a chain of events linking a speaker's brain with a listener's brain, along whose links messages are sent and received (Fig. 1-1).

The speech chain has an important side branch that is sometimes overlooked: that linking the speaker's brain with itself. Every time we speak, we hear our own voices, both the air-conducted sound waves that we direct toward listeners and the bone-conducted vibrations of our own vocal folds. Thus speakers double as a kind of listener, serving not as a passive audience but as active self-monitors of vocal output. This on-line vocal monitoring is in essence a comparison of the speech sounds being produced with an internal representation of target speech sounds. Through feedback-based control, speakers can make the adjustments necessary to match their productions with their intentions.

The principal aim of the research described in this dissertation was to assess linguistic influences, both phonetic and prosodic, on auditory feedback-based control. A functional magnetic resonance imaging (fMRI) experiment was designed to characterize neural responses to unexpected changes in auditory feedback. The experiment tested the hypothesis that phonetically-relevant changes would evoke a greater response than acoustically-salient but phonetically-irrelevant changes. Additionally, a behavioral experiment was designed to induce speaker adaptation to fundamental frequency (F0) changes in auditory feedback. An emphatic stress task was used to test whether the response to F0-shifted feedback would extend to other acoustic features that are also cues to stress.

The thesis of this work can be divided into two parts: (1) auditory feedback influences speech production, and (2) language experience influences speech perception, which in turn affects the feedback-based control of production.

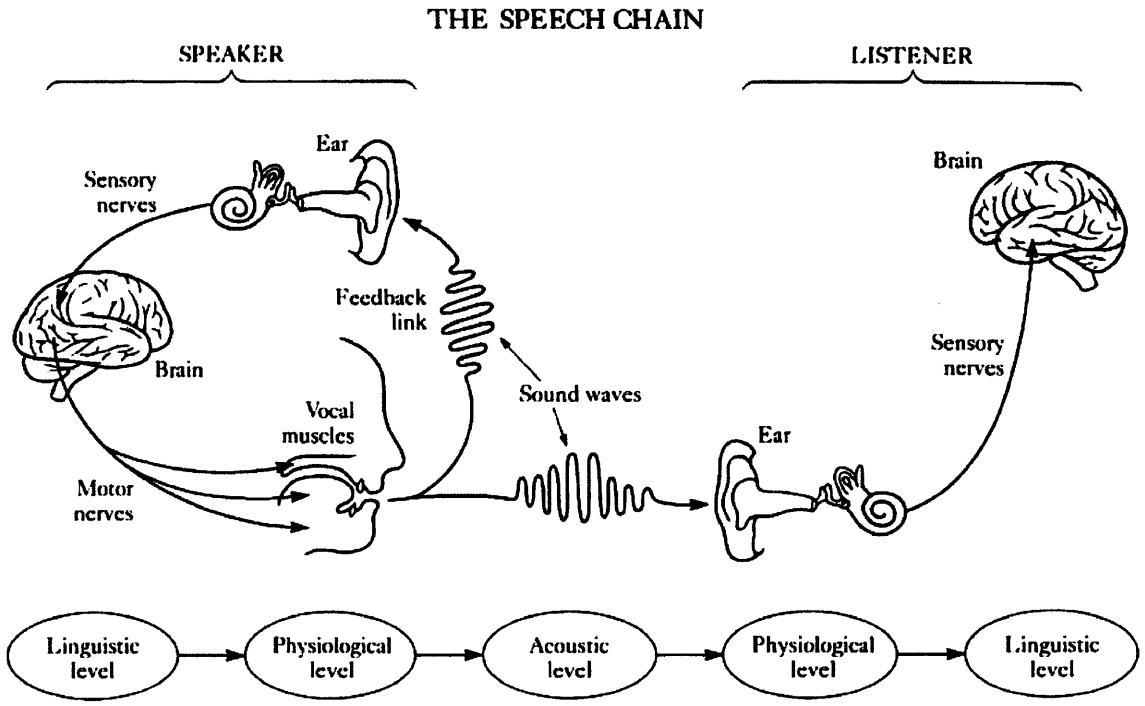

Figure 1-1. The speech chain: the different forms of a spoken message in its progress from the brain of the speaker to the brain of the listener (and, through auditory feedback, back to the brain of the speaker). From Denes and Pinson, 1993.

1.1 Auditory feedback shapes production

Although the end goal in speech is an acoustic signal, the act of speaking is very much a motor act. As Roy Blount Jr. said in his book, *Alphabet Juice*, "All language, at some level, is body language." The configuration of the vocal tract gives rise to the acoustic properties that are perceived as speech sounds. Thus, to produce speech, motor commands are sent from motor cortex to labial, glossal, palatal, mandibular, velopharyngeal, and laryngeal muscles, as well as the muscles of respiration. It takes precisely orchestrated articulator movements to achieve the dazzling consecutions of coordinated gestures that characterize speech. To master a spoken language, we must become adept at producing well-formed speech with phonetic components that can be easily categorized by listeners.

The feedback link in the speech chain is essential for the development of proper motor speech output. When we first learn to speak as infants, we must learn a mapping between the motor commands for speech gestures and the sounds these gestures produce. Hearing our own speech enables us to become versed in the articulatory-acoustic relations that define this mapping. Our own voices act as auditory feedback, enabling a precise tuning, over time, of our knowledge of the correspondences between gesture and sound. Adults with years of speech experience have well-tuned mappings, and need not rely on the sound of their own speech to guide their pronunciation. Nevertheless, even in mature speakers, auditory feedback continues to affect speech production. When auditory feedback indicates a disconnect between the expected and observed acoustic consequences of an articulatory gesture, neural feedback control allows for the consequent correction of the perceived error.

By experimentally manipulating subjects' perceptions of their own speech, it is possible to induce such an acoustic discrepancy between the expected and observed speech output. Both the magnitude of the discrepancy and the perceptual judgment of a particular speaker determine the corrective response to this type of speech perturbation. Under these conditions of perceived error, the studies described in this dissertation aim to observe the compensatory responses to both unexpected and sustained acoustic perturbations, as well as the neural activations of the cortical circuits that underlie such compensation.

1.2 Language experience shapes perception (and production)

When we learn a first language, our auditory experience shapes how we will segment the acoustic space into phonetic units—where we will draw the boundaries that differentiate speech sounds. These phonetic boundaries correspond to the edges of perceptual categories that help listeners differentiate between, say, *degree* [dəgɹi] and *debris* [dəbɹi]. To avoid confusions in word meaning, it is critical for language learners to robustly characterize all the distinctions relevant to their language.

Robust characterization comes at the cost of universal discrimination: with language experience, there is a loss in sensitivity of contrasts that have no phonetic relevance. In

other words, in order to quickly recognize a spoken [t] as belonging to the phonetic category /t/, it is to listeners' advantage to ignore meaningless within-category variance in voice onset time, spectral composition, or other features whose variation, within limits, does not affect phonetic identity. This loss of within-category sensitivity, coupled with a hypersensitivity to phonetic contrasts that are meaningful, acts to "warp" acoustic space, making particularly discriminable the changes in the acoustic regions that mark the boundaries between speech sounds.

A central aim of this research is to characterize the responses to perturbations that alter the phonetic identify of the perceived sound, versus those that modify the acoustics only within category limits. The neuroimaging experiment in this dissertation tests the hypothesis that the perceived acoustic error arising from a feedback change near the sensitive boundary region will be greater than that arising from a feedback change that lies safely within the accepted variability for a given speech sound.

1.3 Organization of dissertation

This dissertation is divided into four parts. The following two chapters outline the two central theses of this work, providing background and supporting evidence for the studies performed. Chapter 2 is an explanation and exploration of auditory feedback and its role in speech development, everyday speech production, and speech production under artificially-perturbed conditions. Chapter 3 is an overview of linguistic influences on speech perception, including the neural evidence for phonetic representations and the effects of categorical perception.

The final two chapters describe two experiments designed to investigate the linguistic influences on auditory feedback control. Chapter 4 describes an experiment which contrasts feedback control under normal speaking conditions and under conditions of speech perturbation—modification of formant frequencies in speakers' auditory feedback—both across and within phonetic category boundaries. This perturbation changes the character of the vowel, creating a sudden, unexpected mismatch between the vowel target and its acoustic realization. A subject who says "bet" might hear herself instead saying "bit" or "bat," for

example. By altering the speech feedback signal before it reaches the ear, it is possible to induce the perceived errors that engage the feedback circuit under investigation. Two hallmarks of phonetic influence on the feedback pattern are discussed: differences in neural activation as measured by fMRI, and differences in acoustic output as measured by shifts in subject-produced formant frequencies.

In Chapter 5, a second perturbation experiment explores linguistic contrasts at a suprasegmental level by perturbing *prosodic*, not segmental, cues. In this study, an adaptive, sustained perturbation acts to decrease or increase the perceived F0 of the stressed word in a multi-word sentence. A subject who stresses the word *Bob* in the sentence "BOB bought a dog" might hear herself placing more ("***BOB*** bought a dog") or less ("Bob bought a dog") stress on the first word, a perceived increase or decrease of the contrast between that word and the lower-pitched unstressed words. Because the stressed-unstressed contrast is linguistically relevant and is cued by other features besides pitch (namely, increased intensity and duration), it is hypothesized that the adaptive response is not a pure low-level pitch correction, but instead extends to one or more of these supporting cues.

Past studies of auditory feedback perturbation have investigated responses to changes in low-level acoustic dimensions: for example, a decrease in pitch, or an increase of the first formant. In contrast with these studies, the experiments described here specifically tailor the perturbations to be perceptually relevant, capitalizing on individual speakers' phonetic and prosodic contrasts. In this way, this dissertation addresses the nature of auditory feedback control under conditions of linguistically-meaningful perturbation.

CHAPTER II

AUDITORY FEEDBACK CONTROL IN SPEECH PRODUCTION

2.1 Auditory feedback of speech

Motor control systems are classically described as using one of two control schemes: *feedback* (closed-loop) control or *feedforward* (open-loop) control (Åström & Murray, 2008). A closed-loop controller uses feedback to control the outputs of a dynamical system—for example, monitoring auditory feedback to correct deviations from a desired acoustic trajectory. An open-loop controller responds in a predefined way based on previously learned command signals—for example, executing the motor program for the well-learned sequence "hello" with no influence from incoming sensory information.

Speech production employs both feedback and feedforward control. A child learning to speak must first construct an internal neural model for feedforward speech movements. Feedback allows the brain to build up a correspondence between these articulatory movements and their acoustic consequences. The monitoring of vocal feedback is critical for achieving verbal fluency, as evidenced by production deficits in those with imperfect feedback. Speakers with congenital hearing impairments show commensurate impairments in babbling (Oller & Eilers, 1988) and in learning to speak (Smith, 1975). Even speakers who become deaf late in childhood, after learning to speak, experience a marked deterioration of speech production (Cowie & Douglas-Cowie, 1992; Waldstein, 1990), since the growth of the vocal tract alters the previously learned acoustic–articulatory relationship. Feedback allows ongoing auditory experience to retune motor gestures as physical and acoustical properties change.

As important as auditory feedback is for speech development, we need not rely entirely on feedback control to speak. Our ability to speak in the presence of feedback-masking noise (Lane & Tranel, 1971) or after post-lingual hearing loss (Cowie & Douglas-Cowie, 1983) is evidence for a feedforward speech controller. Furthermore, feedback

control is too slow to be used in moment-to-moment motor control of speech. Feedforward mechanisms are essential for fluent sequences of rapid movements in which there is no time for feedback to play a part (Rosenbaum, 2009). These fast feedforward mechanisms predominate in the execution of predictable, well-practiced movements. Spoken syllables frequent in our language are perhaps the most oft-practiced gestures we carry out. Evidence that the initiation of high-frequency syllables is faster than for low-frequency syllables (Levelt & Wheeldon, 1994) supports the idea that these highly overlearned movement patterns are stored as preprogrammed motor routines: what Levelt and colleagues call the "mental syllabary."

However, even with the existence of feedforward motor programs, auditory feedback continues to play a significant role in the maintenance and control of ongoing speech. In hearing individuals, auditory feedback is important for the generation of complex and rapid speech acoustics (Ventura et al., 2009). Adults who become deaf after achieving verbal fluency can still speak intelligibly, but they immediately begin to lose control of vocal pitch and amplitude (Cowie & Douglas-Cowie, 1992; Lane & Webster, 1991). Additionally, there is a gradual deterioration of speech sound contrasts in these post-lingually deaf individuals (Cowie & Douglas-Cowie, 1992; Lane & Webster, 1991; Plant, 1984), evidence of a corresponding deterioration of feedforward commands when deprived of auditory feedback. Laboratory experiments on songbirds, prolific vocalizers and vocal learners, have also found deterioration in stereotyped song patterns after auditory feedback was removed (Lombardino & Nottebohm, 2000; Nordeen & Nordeen, 1992). These data show that the motor circuitry underlying the production of adult speech (and birdsong) is not hard-wired but continually influenced by auditory feedback. Finally, auditory perturbation studies are incontrovertible evidence that feedback control is active even among proficient speakers of a language.

2.2 Feedback perturbation experiments

There are many ways to show experimentally that speakers are their own listeners. The simplest and one of the most striking examples is the delayed speech feedback effect.

First described academically in 1950 by Bernard Lee, the delayed speech feedback effect can be induced with audio software, a PA system, or, as Lee used in 1950, the "Presto PT-900" magnetic tape recorder. Speakers who hear their speech played back with a delay of approximately one-quarter second are unable to speak normally: they pause, repeat themselves, and even exhibit a pseudo-stutter in response to the delay, which they are unable to ignore (Lee, 1950).

Feedback perturbation studies such as this reveal the influence of feedback control by inducing a mismatch between auditory expectations and observations. Changing the timing of speech sequences as in the delayed feedback effect has a deleterious effect on speech. Other manipulations with a much smaller delay do not directly impair speech processes, but instead cause changes in one or more parameters of the vocal output. Another simple and well-known auditory feedback experiment is to observe speech in a noisy environment: in a phenomenon known as the Lombard effect, speakers will increase the volume of their speech in the presence of white noise (Lane & Tranel, 1971; Lombard, 1911). Through the use of computer algorithms that rapidly process and filter acoustic signals, more complex manipulations to the speech signal can be performed before it is heard by the speaker. For example, pitch, intensity, and formant frequencies each can be statically or dynamically altered (Burnett et al., 1998; Heinks-Maldonado & Houde, 2005; Houde & Jordan, 1998).

The feedback perturbation methodology is useful and revealing because it illustrates how speakers use real-time feedback information to control their speech. The speaker has no explicit task other than speaking and, if the perturbations are natural-sounding, there is no task difference from the subject's point of view between perturbed and unperturbed trials. Furthermore, two different types of experiments—those that involve *sensorimotor adaptation* and those that elicit brief, *rapid compensation* to perturbations—are each well-poised to answer different questions about the feedforward and feedback control of speech.

2.2.1 Perturbation paradigms

Sensorimotor adaptation refers to paradigms in which motor actions adapt to altered sensory feedback. In such a paradigm, feedforward motor commands are tuned over time as

feedback perturbations consistently alter the desired output signal. A sustained, often gradual shift in the speech output signal causes a commensurate sustained adaptation to the shift through the resetting of motor commands.

A simple visual example of an adaptation paradigm is the use of optical prisms in a reaching task. Prism glasses distort the wearer's visual input, providing a view of the world that is shifted to the left or right of the normal visual field. Exposure to the visual feedback displacement leads to sensorimotor learning: initial reaching errors in the direction of the visual shift disappear after about a dozen trials (Redding & Wallace, 2006) as the subject learns a spatial remapping.

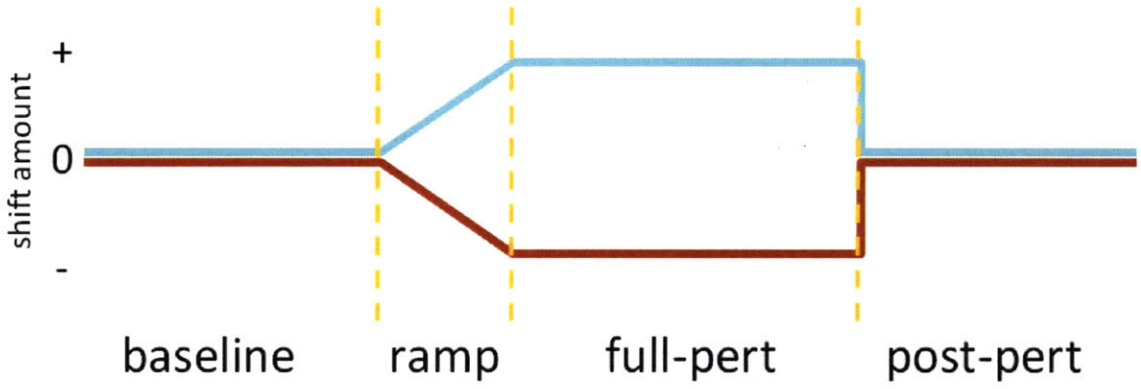

Figure 2-1. Schematic of an experiment timeline for a typical sensorimotor adaptation paradigm. The light blue line represents an upward shift of a parameter (e.g. F0, F1, or intensity); the dark brown line represents a downward shift. A single subject would be assigned to a single group (up or down).

In the same way, auditory adaptation experiments induce a consistent acoustic change to compensate for perturbed feedback. Introducing gradual shifts in formant structure (Houde & Jordan, 1998; Purcell & Munhall, 2006; Villacorta et al., 2007), pitch (Jones & Munhall, 2002, 2005), or intensity (Chang-Yit et al., 1975) causes subjects to gradually adapt to the shifts, producing speech whose formants, pitch, or intensity are modified to counteract the perturbation. The typical structure of an auditory adaptation paradigm begins with a baseline phase, continues with a gradual ramping up of the perturbation followed by a full- or sustained-perturbation phase, and concludes with a final

baseline phase, shown as a schematic in Figure 2-1. The opposing response, usually measured with respect to the baseline phase, begins soon after the onset of perturbation and is sustained over many trials. An interesting characteristic of adaptation paradigms is the typical existence of an after-effect, or an overshoot of the compensatory shift during the post-perturbation phase. In other words, subjects continue to show adaptation to the perturbation even after it has been removed. The aftereffects seen in adaptation experiments are evidence for a transient reorganization of sensorimotor neural mappings between motor commands and their corresponding acoustic targets.

The second kind of perturbation experiment uses a sudden, *unexpected perturbation* to displace speech output from its target trajectory. The perturbed trials are "unexpected" because they occur randomly and rarely throughout the experiment, typically on less than one-third of trials. The majority of trials maintain normal auditory feedback. The opposing response is measured as the difference between the average trajectories (formant or F0 traces over the course of each trial) in the perturbed trials and the average trajectories in the baseline trials.

Often, the "speech" in this paradigm is simply sustained phonation over several seconds, during which subsegments of the phonated syllable are perturbed (Burnett et al., 1998). The speaking task can also consist of whole words (Jones & Munhall, 2002; Tourville et al., 2008) or of sentences in which one or more words are perturbed (Chen et al., 2007). In all of these cases, subjects exhibit *rapid compensation* to the perturbation, altering their speech trajectories to oppose the perturbation within a few hundred milliseconds of its onset.

Because of the sparseness of perturbed trials, there is no new sensorimotor mapping as in adaptation paradigms; instead, speakers can "reset" the perturbed parameter with each normal feedback trial. The perturbations in this paradigm better resemble isolated errors in natural speech, which are corrected on the fly. Additionally, there is a smaller magnitude of compensation: 7-10% versus 25-40% in the auditory adaptation paradigm (Houde & Jordan, 1998; Tourville et al., 2008).

In summary, adaptation experiments provide evidence that a feedforward speech controller continuously monitors auditory feedback and is modified when that feedback does

not meet expectations. Brief, unexpected perturbation studies show the importance of auditory feedback in correcting speech errors or expectation mismatches very rapidly, over the course of an ongoing utterance.

Finally, much in the same way that speakers remain unaware of the articulatory gymnastics that occur in the course of normal speech, participants in both types of feedback perturbation studies are found to compensate for induced shifts even without being aware of them. As Roy Blount, Jr. put it while describing effortless feats of articulation, "It's hard to keep track of exactly what your tongue is up to."

2.2.2 Neural responses to auditory feedback perturbation

The majority of the research on auditory perturbation has taken the form of purely behavioral studies, but several landmark perturbation experiments have been performed inside the scanner, investigating neural function during vocal production. Magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) are powerful modalities for studying the brain's response to perturbation, demonstrating the neural mechanisms that underlie the detection of an auditory mismatch and the subsequent corrective motor response.

Magnetoencephalography. MEG is a technique used to measure the magnetic fields produced by the intracellular electric currents of pyramidal neurons. One hallmark of the neural response to an auditory event is the M100, an event-related potential recorded from the fronto-central region of the scalp. A reduction in the amplitude of the M100 response has been noted for self-produced—and therefore expected—speech sounds as compared with the same speech sounds presented in a passive listening condition (Curio et al., 2000; Nagarajan). This M100 reduction for speech, or *speaking-induced suppression* (SIS; also more broadly known as *motor-induced suppression*, MIS), suggests that speaking dampens the neural response to self-produced expected sounds. SIS was attenuated when participants' feedback was shifted in pitch, compared with unaltered voice feedback (Heinks-Maldonado et al., 2006); in other words, the neural response was enhanced in the presence of a feedback shift. Additionally, a response occurring 100-400 ms post-perturbation was enhanced while

subjects vocalized, compared with passive listening (Houde et al., 2007). In the view of Houde and colleagues, incoming feedback is compared with an efference-copy derived prediction of expected feedback, and this neural response is indicative of a mismatch between the two.

A weakness of MEG in speech perturbation paradigms is the presence of movement artifacts introduced during production. However, artifact-inducing articulator movements can largely be avoided if the speech produced is limited to sustained vowels that do not involve dynamic changes of the vocal tract (for example, phonating on the monophthongs /ɑ/ or /ə/) as in Houde et al. (2007).

Functional magnetic resonance imaging. Functional magnetic resonance imaging, or fMRI, affords a high-resolution spatial reconstruction of neural activity, as indirectly measured by the hemodynamic response—a pattern of oxygenated blood flow—in different regions of the brain. Speech fMRI studies have low temporal resolution, but can avoid the movement-related drawbacks of MEG using sparse temporal imaging. Because the hemodynamic response lags the stimulus by several seconds, it is possible to interleave silent intervals with periods of scanner noise, waiting until after the participant finishes speaking to acquire each image. Through the use of these sparse sampling techniques, participants can speak in relative silence and images can be acquired in relative stillness.

Using fMRI and a sparse sampling paradigm, Tourville and colleagues (2008) measured the neural response to sudden, brief perturbations of the first formant frequency (F1). In addition to a compensatory response that began 136 milliseconds after voice onset, a neural response to the perturbation was noted in bilateral posterior superior temporal cortex and right inferior frontal cortex. In the authors' view, the temporal cortical activation was indicative of the perceived mismatch between expected and observed auditory output; thus, the neurons contributing to the enhanced response were labeled auditory error cells. This study was the motivation for the fMRI experiment described in Chapter 4 of this dissertation.

Bilateral superior temporal gyrus (STg) activation has also been reported for fMRI paradigms using delayed auditory feedback (Hashimoto & Sakai, 2003). Imaging studies of pitch perturbation have revealed similar cortical regions underlying verbal self-monitoring.

Zarate and Zatorre (2005) measured the neural responses of both singers and non-musciains to pitch-shifted feedback, reporting activation in bilateral auditory cortices as well as in anterior cingulate cortex (ACC) and insula. Both singers and non-musicians were found to compensate for the perturbation when instructed to do so, but singers were more accurate at maintaining the target pitch when instructed to ignore the perturbation.

2.3 Competing models of speech motor control

Auditory perturbation experiments provide evidence for the importance of auditory feedback in guiding speech gestures. A common explanation for this sensitivity to feedback is a theory in which the goals of speech production are auditory perceptual targets. An auditory perturbation that shifts productions off-target provokes a compensatory articulatory gesture that sets the auditory output back toward the target. However, a competing theory posits articulatory gestures, or the intended motor commands that produce them, as the invariant targets of speech production. The well-known *motor theory* of speech perception developed at Haskins Laboratories holds that perceiving speech is perceiving vocal tract gestures (Liberman & Mattingly, 1985). Liberman and Mattingly proposed an "analysis by synthesis" in which the listener guesses at the speech gesture underlying an acoustic signal and internally synthesizes the acoustic consequences of this gesture to compare with the incoming acoustics. In this theory, the motor commands that produce the acoustic signal were presumed to be invariant for a given phoneme. For example, the /d/ in /di/ and the /d/ in /du/ have very different formant transitions owing to effects of coarticulation with the following vowel; however, both /d/s have an articulatory gesture in common, namely a vocal tract constriction made by contacting the tongue tip with the alveolar ridge.

Motor theory predicts that speech perception should be sensitive to visual or haptic evidence of speech gestures. One example of the influence of visual information on speech perception is the McGurk effect (McGurk & MacDonald, 1976), in which seeing the articulatory movements of a speaker can affect how a syllable is perceived auditorily. In addition, listeners do benefit from visual evidence of articulatory gestures: it is easier to

perceive speech in noise when the speaker is visible, as visual information can disambiguate syllables produced with different places of articulation (Sumby & Pollack, 1954).

However, motor theory in its strict form fails to account for the motor equivalent capabilities of the speech production system (Guenther et al., 1998). While two different acoustic patterns can both sound like a /d/ according to their context, the reverse is also true: two different vocal tract gestures can produce the same acoustic output. Speakers have the ability to use different movements to reach the same goal, and the same speaker will naturally use different movements under different conditions. In other words, speech production is motor equivalent; there is a many-to-one transformation between vocal-tract configurations and acoustic goals. Articulatory trading relations allow speakers to maintain a stable acoustic signal even when articulation method varies. For example, retroflex sounds such as /r/ can be produced either with the tongue tip raised in a "retroflex" position or with a "bunched" tongue farther back in the mouth (Ladefoged, 1993). These different articulatory gestures have the same acoustic consequence: a dip in the third formant. It is this dip, produced in either tongue position of in some combination of these extremes, that listeners use to distinguish /r/. Similarly, the low first and second formants of the vowel /u/ can be achieved either by lowering the larynx or by rounding the lips, each having the same acoustic effect (Ladefoged, 1993). Furthermore, speakers rapidly learn to reorganize vocal tract configurations to maintain steady formant patterns in the presence of articulatory impediments such as a bite block (Gay et al., 1981). Acoustic or sensory theories that propose auditory targets allow for this motor equivalent variability in articulatory gestures (Guenther et al., 1998; Perkell et al., 1997; Perkell, in press).

Another claim of the motor theory of speech perception is that the motor system is recruited for perceiving speech. *Mirror neurons*, first discovered in the premotor cortex of primates, respond both while performing a motor action, such as grasping, and while witnessing that action being performed by another (di Pellegrino et al., 1992). Mirror neurons were interpreted as a system for "action recognition" in motor cortical areas that could extend to speech motor gestures. Complementarily, there is fMRI evidence that the same motor regions active during the production of speech are also activated during passive listening (Wilson et al., 1994). However, recruitment of the motor system does not appear

to be essential for speech perception: infants and non-human animals, both unable to produce speech, can discriminate phonetic contrasts and exhibit sophisticated perceptual abilities such as categorical perception (Eimas, 1971; further discussed in Chapter 3). Additionally, lesions to speech motor cortex often show only minor effects on auditory perception of speech. Broca's aphasia, caused by damage to the brain systems subserving motor speech production, results in severaly impaired speech production but does not produce a commensurate abolishment of speech perception and comprehension ability (Naeser et al., 1989).

In summary, there is excellent support for a tight interconnection between speech perception and production. However, there is evidence both for and against the motor-based theory that perceiving speech is perceiving phonetic gestures. Another interpretation of this interconnection yields the opposite claim: producing speech is producing auditory targets. As will be discussed in the next section, successful neurocomputational models that learn speech production through auditory targets are evidence that this is the case.

2.4 The DIVA model and feedback control

Feedback is especially relevant when we have a sensory target in mind and want to track our progress toward that target. In the view of Guenther and colleagues (1994, 1995, 2006), speech targets are not motor configurations or vocal tract constructions but regions in auditory space. These targets are achieved by manipulating the velocities of the speech articulators and the vibration, abduction, and adduction of the vocal folds until the produced acoustics match the acoustic goal. Auditory feedback is used for updating and refining feedforward commands and for controlling unpredictable or novel movements.

DIVA (Directions Into Velocities of Articulators, Guenther 1994, 1995; Guenther et al., 2006) is a model of speech production that incorporates feedback and feedforward control to reach acoustic speech targets. A *speech sound map*, corresponding in function to Levelt's mental syllabary (Levelt, 2001), activates motor commands from DIVA's feedforward control map, as well as expected auditory and somatosensory targets of the production. An articulatory synthesizer (Maeda, 1990) translates DIVA's vocal tract

configuration into an acoustic signal so that the output of the motor commands can be compared with the internal sensory representations of the target, both auditory and somatosensory.

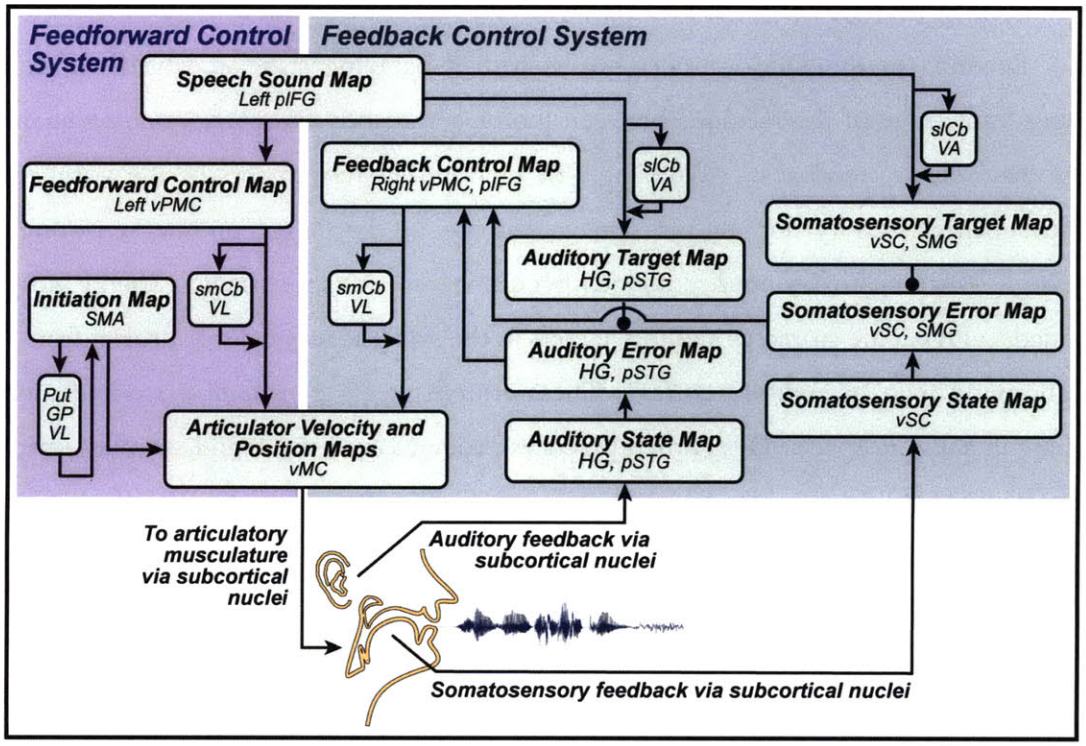

Figure 2-2. Schematic of the DIVA model. Each box corresponds to a population of neurons hypothesized to carry out processing in the specified cortical regions. The arrows correspond to synaptic projections between regions.

DIVA is a neural network model, designed to be biologically plausible; the modules in the model represent processing done in particular cortical regions. Each box in the diagram (Fig. 2-2) corresponds to a population of neurons in the brain that act as processing units. The arrows connecting the boxes correspond to synaptic projections between brain regions. As the neural signals are passed through the model, information is transformed from one type to another. For example, the auditory error map, located in auditory cortex (posterior superior temporal regions) receives input both from the speech sound map that generated the feedforward command and from the subcortical nuclei that do preliminary

processing of the incoming auditory feedback signal. In this region, acoustic representations of the two inputs can be compared and transformed into a difference signal that is used to update the subsequent motor commands. Neuroimaging studies have helped pinpoint the anatomical locations of the model components (Guenther et al., 2006; Tourville et al., 2008; Ghosh et al., 2008; Peeva et al., 2010).

Learning in the DIVA model begins with a babbling stage in which the model acquires knowledge of the relations between motor commands and corresponding auditory and somatosensory feedback. "Babbling," or randomized movements of the vocal tract, provides paired sensory and motor signals that are used to tune up the sensorimotor mapping. Learning to pronounce sounds occurs via an imitation stage. Sound samples provided to DIVA are stored as auditory targets in the synaptic weights that project from the speech sound map to auditory cortex. These stored targets constitute a kind of mental syllabary in auditory space. DIVA then practices production of the sound samples, learning a somatosensory target while it uses auditory feedback to further tune feedforward commands. With each repetition, the model relies less on feedback control and more on feedforward control.

In summary, evidence from modeling, neuroimaging, and perturbation experiments suggest that when learning to speak, we must first form internal representations or targets for speech sounds, then shape our vocal output by comparing auditory feedback with these internal targets. The DIVA model of speech production has an auditory map representation based on formant space; however, it does not take into account the perceptual boundaries dividing learned speech categories. The following chapter outlines some of the ways language learning affects the auditory perceptual space, with the goal of motivating experiments that will help extend the DIVA model.

CHAPTER III

LINGUISTIC INFLUENCES ON AUDITORY PERCEPTION

3.1 Speech sound categories

Phonetic categories are the perceptual representations of phonemes, the smallest units of sound that form meaningful contrasts in a language. As a linguistic construct, phonemes are discrete and categorical: we give names to them, and they allow us to discriminate words with different meanings. In processing acoustic input into phonetic categories, we ignore small acoustic variations around a prototype that have no phonemic consequence. In contrast, in order to effectively and efficiently discriminate speech sounds, we must pay close attention to variations around the boundary regions, where small changes matter for phonemic identity.

A robust representation of these sound categories in the brain allows us to rapidly process and understand incoming speech, and to compare our own speech productions to internal auditory schemata. As distinct entities, phonemes represent an abstract concept, but there are behavioral data and neural correlates that support their existence in the brain of the speaker.

3.1.1 Evidence of phonemes

Patterns of errors in production. Speech production errors can speak volumes about the way linguistic units are stored and sequenced. One of the most famous and oft-quoted varieties of speech error was named for Reverend William Archibald Spooner of Oxford. His "Spoonerisms," to which he was notoriously prone, involved the transposition of two words' initial segments, yielding such delightful disarrangements as "You have hissed all my mystery lectures!" (Potter, 1980) Dr. Spooner was not alone in these slips; the speech error literature shows consistent patterns of production errors across speakers that can be robustly

induced in experimental settings (Motley, 1983). Submorphemic slips of the tongue typically involve transpositions or substitutions of single phonetic segments (Meyer, 1992; Shattuck-Hufnagel, 1983, 1987), the results of which are akin to Spoonerisms: for example, "heft lemisphere" for left hemisphere. There is very little cross-pollination between sound or grammatical types: vowels exchange with vowels and consonants with consonants; nouns with nouns and verbs with verbs. This observation suggests not only that there are distinct levels of representation in the planning of speech production (Fromkin, 1980), but that phonemes are one of the lowest levels, since they generally remain intact across transpositions, anticipations, and perseverations of sound segments (Meyer, 1992). However, some studies employing electromyography (Mowrey & MacKay, 1990) and kinematic tracking of articulators (Goldstein et al., 2007) have provided evidence for lower-level representations based on subphonemic errors in articulator movement.

Reproduction conduction aphasia. Conduction aphasia is a disorder of linguistic processing related to damage in left supramarginal gyrus, left primary auditory cortices, insula, and underlying white matter (Damasio, 1992). In patients with conduction aphasia, intonation and articulation are preserved, but repetition is impaired by the presence of phonemic paraphasias, or substitutions of an incorrect phoneme for the intended one. Phonemes are deleted, transposed, or exchanged with each other even though speech production is otherwise relatively preserved (Damasio, 1992; Goodglass, 1992). For example, in one patient, the German word "Bagger" was repeated as "gabber" (Bartha & Benke, 2003). Like non-pathological slips of the tongue, these paraphasic errors maintain structure at the level of the phoneme.

Differences in cortical processing of different phoneme classes. There has been shown to be more neural activity in superior temporal cortex in response to "poor" phonemes, ambiguous sounds that lie near phonetic category boundaries, than to "good" phonemes, prototypical sounds that lie squarely in the center of the phonetic category (Guenther et al., 2004). This finding of increased brain activation to boundary stimuli implies that the brain is able to efficiently shift neural resources away from regions of acoustic space where discrimination is not behaviorally important (e.g., near the center of a sound category) and toward regions where accurate discrimination is needed. In other words, more neural

resources are devoted to processing ambiguous sounds. The formation of phonetic categories is an example of *perceptual warping* of auditory space that is contingent upon acoustic exposure.

3.1.2 Boundary effects

Phonetic boundaries, then, divide the acoustic space of speech sounds into discrete chunks within a language. Auditory perceptual experiments have demonstrated dissimilar responses to stimuli near these boundary regions and to stimuli that are safely within a particular phonetic category. For example, the *perceptual magnet effect* (Kuhl, 1991; Kuhl et al., 2008) is an account of decreased discrimination ability near category centers as opposed to near category boundaries. The canonical category center acts as a "magnet" that draws in surrounding acoustic tokens and renders them less discriminable; however, near-boundary tokens escape the magnet's pull and are perceived as more dissimilar from each other. Thus, phonetic boundaries function as discontinuities along a perceptual continuum: a continuous acoustic space is warped to yield a perceptual representation that is non-continuous. This warping also underlies the phenomenon of *categorical perception*, generally characterized by a peak in discrimination at a category boundary. Stimuli classified as belonging to different categories are easier to discriminate than stimuli classified as the same category, even when the acoustic differences are of the same magnitude (Harnad, 1990; Repp, 1984). In other words, phonemes can be relatively hard-edged, and stimuli that straddle such an edge can be more perceptually distinct than those that lie to one side of it.

3.2 Categorical perception

Categorical perception is a general phenomenon that allows us to sort the things in the world into their proper categories, "warping" perceived similarities and differences so as to compress some things into the same category and separate others into different categories (Harnad, 1990). It extends across many domains, such as color perception (Holmes et al.,

2009) and flicker-fusion (Pastore et al., 1984) in the visual domain and musical tones in the auditory domain (Siegel & Siegel, 1977).

One well-investigated area of perception research deals with how sounds from different phonemic categories are identified and discriminated. Stimuli in these experiments are generally constructed along a continuum of a single acoustic attribute, varying from one phoneme to another. Each stimulus is evenly spaced from the next in terms of the chosen acoustic attribute. (For example, voice onset time (VOT) distinguishes English judgments of /ba/ and /pa/. An experiment using these two phonemes would feature a ba–pa continuum, spaced in VOT by a set number of milliseconds.) Both identification and discrimination performance are usually evaluated.

The universal finding is that adults tested on identification show a sharp category boundary at one place along the continuum; furthermore, discrimination performance can go up dramatically when the stimuli come from different categories. Tokens judged to be within the same phonemic category are discriminated poorly, even when they are separated by the same acoustic distance as the well-discriminated tokens, one from either side of the category boundary (Liberman et al., 1957). In the words of Alvin Liberman, the speaker's discriminations have been "sharpened and dulled according to the position of the phoneme boundaries of his native language."

3.2.1 Origins of categorical perception

This phenomenon of categorical perception was initially found only in speech contexts. It was most particular to stop consonants, and somewhat harder to elicit in glides and vowels. In experiments by Liberman and colleagues (1961), continuous (not categorical) perception was obtained with non-speech stimuli generated by spectrally inverting a speech VOT continuum, suggesting that the phenomenon was specific to speech. Liberman's motor theory of speech perception, as discussed in Chapter 2, claimed that phonemes are processed by special phonetic mechanisms of hearing: a learned internal language-production model. However, later experiments succeeded in demonstrating categorical perception in non-speech sounds. Specifically, complex speech-like sounds such as noise-

buzz sequences with various lead times (analogous to VOT in plosive consonants) showed both an abrupt labeling shift and an accompanying peak in discrimination at 16 ms (Miller et al., 1976). Another particularly compelling example by Pisoni (1977) used tone onset time as a VOT analog, using two-tone complexes of differing frequencies. Pisoni found a peak in discrimination at 20 ms, no matter which tone led the other. These instances of categorical perception in non-speech contrasts argue for contrasts that are shaped by general principles of auditory perception, or, more broadly, a general property of sensory behavior.

Even more convincingly, animals and preverbal infants also show evidence of categorical perception, even in the absence of training. Furthermore, they seem to share similar category boundaries with adults. In infants, a head-turn or high-amplitude nonnutritive sucking procedure is generally used to assess discrimination performance as measured by recovery from habituation. In experiments with both English and Spanish infants, Eimas and colleagues (1971, 1987) have shown sensitivities to a VOT boundary at 25 ms, corresponding to the English boundary. Kuhl and Miller (1975), in multiple experiments with chinchillas, report voicing boundaries that correspond with adult and infant data, providing evidence for the same mechanisms of perception for all three species. Categorical perception and even perceptual compensation for coarticulation have since been demonstrated in macaques, quails, and budgerigars. These results confirm a general perceptual ability to discriminate phonetic information in CV syllables in a way that adheres to the distribution of the acoustic stimuli.

Liberman's motor theory claimed that categorical perception derives from linguistic categories. However, it is unlikely for animals and infants to have enough experience to acquire phonetic representations. This implies that it is simply the processing of the mammalian auditory system that allows for part of the observed perceptual performance. Not only do these studies show that categorical perception can arise from general auditory principles, but they suggest that the animals and infants, unbiased by previous language exposure, are responding to invariant acoustic factors in the signal that underlie phoneme categories.

Categorical perception is thus not merely the result of a specialized language-processing capability in humans. However, language learning does have a strong effect on

perceived category boundaries in adult speakers, as evidenced by the variation in boundaries from language to language. Spanish infants who show an English-like 25-ms VOT boundary grow up to become Spanish-speaking adults who show boundary at 0 ms (Eimas et al., 1987). Furthermore, there boundaries can be shifted by experimental conditions. For example, vowel identification is highly susceptible to stimulus sequence effects (Repp and Liberman, 1987). (Consonant identification, being more "categorical," is more stable.) In a phenomenon known as *selective adaptation*, hearing a sound many times can shift perceptions of ambiguous tokens away from that sound category (Eimas & Corbit, 1973). Additionally, when one possible phonetic categorization forms a word and the other does not, there is a tendency to categorize ambiguous tokens to make words (Ganong, 1990). That is, the boundary between two phonemes shifts towards the lexically-acceptable one. Categorical perception is therefore an important auditory-acoustic relation that is greatly influenced by linguistic experience.

3.2.2 Neural basis of categorical perception

It is to a speaker's advantage to reduce "the number and variety of the many sounds with which he is bombarded" (Liberman et al., 1957). The perceptual system is tasked with recoding the low-level, rapidly-changing, multi-possibility signal into something more efficiently accessed. By suppressing responses to signal variations that are not judged to be phonemic, categorical perception improves the efficiency of speech processing. Categorical perception of some non-speech stimuli show that these nonlinear effects do not require a lifetime's worth of language experience. However, language-specific category perception does require training based on acoustic input. This training serves to tune up the central auditory system to respond differently to sounds at category centers than to sounds in boundary regions.

As discussed, Guenther and colleagues (2004) found decreased BOLD activation in superior temporal areas for good exemplar than a boundary token. This observation provides a neurological explanation for perceptual behavior: sounds from the center of a

category are more difficult to discriminate from each other than sounds near category boundaries because they are represented by fewer cells in the auditory cortical areas.

Many researchers have used noninvasive electrophysiology techniques such as MEG to quantify the neural response to changes in phonetic membership. A common neural marker of acoustic change is the mismatch negativity (MMN) evoked potential, a preattentive, neurophysiologic index of auditory discrimination. The MMN, or its magnetic analogue, the mismatch field (MMF), is often elicited by an infrequent, deviant stimulus (an auditory "oddball") detected among frequent, repeated stimuli (the "standard"). Several independent studies in MEG and electroencephalography (EEG) have found an enhanced MMN to a native-language phonetic category contrast than a native within-category contrast or non-native contrast (Dehaene-Lambertz, 1997; Näätänen et al., 1997; Sharma & Dorman, 1999, 2000; Sharma et al., 2000). Furthermore, Dehaene-Lambertz and Baillet (1998) have reported similar findings in prelingual infants. Strikingly, Phillips and colleagues (2000) elicited an MMF arising from left superior temporal gyrus using stimuli that were phonetic, but not acoustic, oddballs—that is, no individual stimulus was frequent, but subjects grouped together different tokens of each phonetic category to form a standard/oddball distribution that elicited the mismatch response. This is evidence that auditory cortex responds categorically, having access to a phonological representation of category membership.

3.3 Graded sensitivity within categories

While listeners are less sensitive to within-category than between-category distinctions, they do show graded sensitivity to the "goodness" of a particular speech token. As discussed in Section 3.1, neural responses in superior temporal areas are sensitive to the goodness of fit of a sound to its phonetic category (Guenther et al., 2004). Listeners also explicitly rate near-boundary stimuli as poorer exemplars of a speech category (Miller & Volaitis, 1989), and discrimination of poorer exemplars is characterized by a longer reaction time (Pisoni & Tash, 1974). Furthermore, the goodness of speech tokens can affect the efficacy of semantic priming. At short interstimulus intervals, Andruski and colleagues

(1994) showed a decrease in the magnitude of semantic priming for words whose syllable-initial VOT was acoustically manipulated to make them worse exemplars of the phonetic category. The change in semantic priming shows the effect of acoustic fine structure on lexical access, even when the fine structure variation remains within a given category.

These results suggest that listeners do have access to information about within-category variability. Moreover, this within-category sensitivity can have downstream consequences for higher-level speech perception.

3.4 Vowel perceptual space

For vowels, which are produced with an open vocal tract, continuous articulation between two phonemes is possible. That is, speakers have the physical ability to pronounce boundary stimuli, a vocal feat that, for consonants, is difficult to perform except synthetically. Even though their production is less discrete than that of consonants, vowels have been shown to exhibit some degree of categorical perception. Vowels are well-described by their first two formant frequencies: the location of the first two peaks in their spectral envelope. It is therefore natural to think of vowels as inhabiting a two-dimensional frequency space, with each dimension representing its respective formant value. A speaker producing the sound "ah" [ɑ], for example, will by definition produce the first two formants very close together, with a very high F1 (around 700 Hz) and a very low F2 lying almost on top of F1 (around 1000 Hz). A listener, given a real or synthetic sound with these formant values, would classify it as an [ɑ] (Stevens, 2000).

As discussed in Chapter 2, a body of evidence suggests that the goals of speech gestures are regions in acoustic space (Guenther et al., 1998; Perkell et al., 1997; Perkell, in press). Put mechanistically, speaking involves reaching sequential targets corresponding to phonemes in our language, while learning to speak requires learning the appropriate motor commands to produce sound sequences across the speech target regions. Vowel targets, then, to a first approximation, are regions in two-dimensional formant space. The perceptual boundaries between vowels segment the space into frequency regions, each associated with a particular vowel category. Learning to produce vowels can be described as

associating auditory targets with motor commands that achieve those target frequency regions.

Because vowels exhibit some degree of both categorical perception and continuous or graded perception, they are interesting to study in an auditory feedback context. The experiment described in Chapter 4 examines the influence of non-continuous perception on feedback control.

3.5 Transforms of the auditory pathway

To motivate the influence of phonetic boundaries on feedback control in the DIVA model, it is necessary to deconstruct the path between acoustics and audition. The following section is an overview of the transforms that occur in the auditory system as it processes sound input into recognizable and categorizable speech.

3.5.1 Preliminary speech processing

The processing of speech begins at the peripheral auditory system. At this initial level, speech is no different from any other acoustic stimulus. The vibrations that impact our ear are first sorted out by the ear's frequency analyzer, the cochlea. The filter banks of the cochlea determine basic acoustic properties of the signal: the frequencies at which there is acoustic energy and whether that energy is periodic or aperiodic. The mechanical properties of the ear break down the acoustic input waveform and allow for analysis of duration, intensity, bandwidth, and direction of spectral changes.

By the time the acoustic signal is transduced into neural impulses in the auditory nerve, the following modifications have taken place: narrow-band filtering (by the cochlea), half-wave rectification (from the chemical response properties of hair cells), and low-pass filtering (from the loss of high frequencies due to limits on neural synchrony). Even at this early stage of processing, the input signal is different from what is represented on a spectrogram.

Auditory nerve fibers strongly phase-lock to frequencies up to 2.5 kHz. In terms of speech perception, phase-locking is a means of robust encoding of spectral information. By relying on temporal synchrony as well as place along the cochlear frequency analyzer (a "rate-place" representation), the signal is robust to noise and allows for segregation of multiple sound sources. The neural activity pattern also acts to enhance spectral peaks in the signal.

These preliminary processes provide salient dimensions—frequency, harmonicity, spectral shape—that lead to speech sound classification (de Cheveigné, 2003). If the peripheral processing in the ear worked differently to analyze sound, our percepts would not rely on the same acoustic dimensions, and we would probably have very different speech sound categories.

Some studies have attempted to rework our techniques for visualizing auditory maps so that we can more easily approach the acoustics from the point of an auditory system analyzer. The Bark scale is a psychoacoustic scale corresponding to critical bands of hearing. The mel scale was created in an attempt to equate raw Hertz values with psychological pitch distance. Both scales take into account the relative contributions of energy from different parts of the spectrum; they are filtered versions of the pure Hertz scale in an attempt to mirror auditory filtering.

3.5.2 Higher level auditory processing: extraction of features

Principles of perceptual grouping contribute to feature extraction at a relatively low level of the auditory system. Coincident events are "chunked" into united representations. Common modulation, both in frequency and amplitude, is a fundamental cue in auditory grouping. Component groups in a speech signal that are modulated in similar ways, along the same timescale, become perceptually united, making it near-impossible for a listener to hear out the individual constituents (e.g., formants). Harmonicity and spectral proximity, both in time and in frequency, are other cues that help form the percept of a united auditory event.

Still higher in the auditory pathway, neurons in primary auditory cortex (A1) maintain the tonotopic organization of the cochlea, but in a weaker sense; they respond to more complex stimulus configurations. For example, many single units in A1 respond weakly to pure tones of a certain pitch but strongly to pairs of pure tones ascending in frequency. Other units detect descending tone sequences. These neurons represent integrated successive cues from the peripheral auditory system and respond best to temporally variable sounds. These cortical cells are the rudiments of phonetic templates in the brain.

3.6 Linguistic influences on feedback control

Auditory-acoustic relations—the heterogeneous peaks and troughs in phonetic perception that underlie a warped perceptual space—are difficult to measure directly. Tasks requiring overt categorization or goodness judgments can be useful in a rough mapping of perceptual space, but they are inherently unlike normal speaking and listening conditions. Feedback perturbation studies aid in this research by using a natural speaking task and evoking an easily-measured, quantifiable response to shifts of a given acoustic magnitude. The magnitude of the neural response is a direct measure of the "auditory error" caused by the mismatch between the expected and observed signal. The magnitude of the compensatory response reflects the correction in response to that error.

Given the semi-categorical nature of vowel perception, distinctions that lie across vowel boundaries in auditory perception might be predicted to be more discriminable—more salient—and therefore to provoke a greater corrective response. The following chapter describes an experiment that examines responses, both neural and acoustic, to both perturbations that do not change the vowel category and perturbations that have phonetic relevance in the language of the speaker.

CHAPTER IV

EXPERIMENT 1: AN INVESTIGATION OF THE ROLE OF PHONETIC CATEGORIES ON AUDITORY FEEDBACK CONTROL

Even as proficient speakers, we rely on auditory feedback to monitor our speech, ensuring the observed acoustic signal matches our expectations. It is argued in this thesis that a critical factor in feedback control is the relevance of the output to the listener. Evidence presented in the previous chapter suggests that it is easier to perceive deviations from phonetic category centers when they cause a category change. For this reason, a shift in auditory feedback that crosses a linguistic boundary is predicted to be more salient and to result in a larger corrective response.

The goal of the current study is to explore the role of phonetic categories in feedback control. To this end, the study examines the neural response to a sudden disruption in the auditory feedback loop as elicited by an unexpected acoustic shift in real time. In the experiment, sudden auditory perturbations occur during subjects' speech, producing a mismatch between the auditory speech target and the realized speech. This mismatch is theorized to induce activity in the auditory error cells that detect this discrepancy. The perturbation paradigm offers insight into the error correction signal produced by the acoustic mismatch and the updated motor commands used to produce the vocal compensation.

Moreover, perturbations that caused a phonetic category boundary to be crossed were directly compared against perturbations of the same magnitude that caused only a within-category shift in acoustics. Behavioral and neural responses to these two types of perturbations were examined for differences in magnitude, and the neural activations were contrasted to test for spatially separable populations of error cells in the two perturbation conditions.

4.1 Phonetic category variation

Across different dialects, vowel production centers move around the formant frequency space. A speaker from the American southeast, for example, may produce /ɛ/ and /æ/ closer together than the average American speaker (Vaux, 2008). Another difference leading to vowel space asymmetries is simple perceptual variation: because of discrepancies in linguistic exposure, auditory acuity, or response bias, different listeners may assign a given sound to different phonetic categories. This study capitalizes on differences in dialect and in perception to counterbalance subjects with different asymmetries: inter-speaker variations in which vowel boundaries were "easier" to cross.

4.2 Methods and materials

The current experiment consists of two phases: a behavioral session, in which subjects' production and perception spaces were assessed to set experiment parameters (Section 4.2.2), and an imaging session, in which brain activity was measured using fMRI (Section 4.3.3). All study procedures, including recruitment and acquisition of informed consent, were approved by the institutional review boards of the Massachusetts Institute of Technology and Boston University.

4.2.1 Participants

Eighteen right-handed subjects between the ages of 19 and 33 (mean age = 23.5 years), nine men and nine women, participated in the study. These participants were drawn from a pool of forty subjects who completed a behavioral pretest (mean age = 23.6 years). All subjects spoke American English as a first language, had no history of hearing or speech disorders and, in order to be eligible for imaging, had no metal in the body.

4.2.2 Behavioral pretest

Vowel production. At the start of the behavioral session, vowel production data were collected between the carrier consonants /b_d/. Each subject read aloud these b_d words

as they appeared on a computer screen, producing ten tokens for each of the six vowels {/i/, /ɪ/, /ɛ/, /æ/, /ɑ/, /u/}. The words used to elicit these tokens were, respectively, "bead," "bid," "bed," "bad," "bod," and "booed." For ease of recording and subject comfort, the vowels were recorded with the subjects seated at a desk, head in an upright posture.

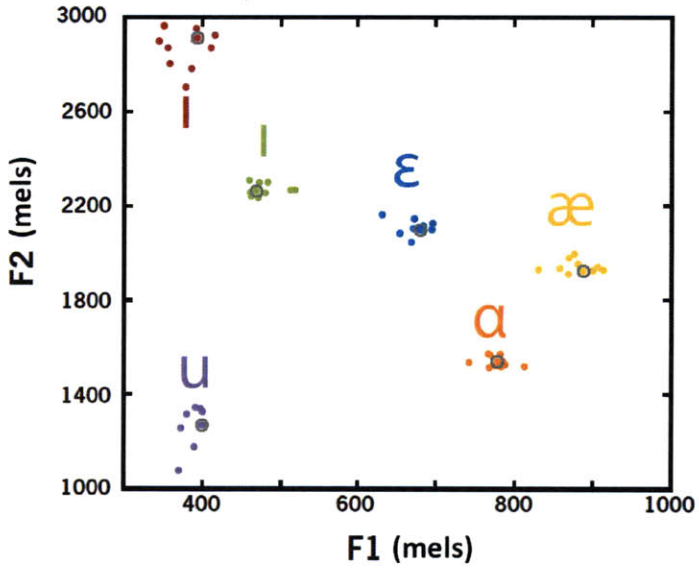

Figure 4-1. Vowel space for a sample subject. The production tokens closest to the two-dimensional median for each vowel are circled in grey; these were the inputs used in generating the continua.

Generation of formant-shifted vowel continua. From the ten productions, median values for the first two formants, F1 and F2, were determined for each vowel. All subjects showed a separation of formant values for the different vowels, whose tokens generally clustered tightly together in vowel space (Fig. 4-1). The production token closest to the two-dimensional median (F1-F2) was used as input to a formant-shifting algorithm (Boucek, 2007) that altered the first and second formant frequencies by a constant offset through the duration of the vowel but held other acoustic properties of the sound constant.

Eight vowel continua were generated across the F1-F2 spectrum for each subject, two continua for each of four pairs of the adjacent vowels {/i/–/ɪ/, /ɪ/–/ɛ/, /ɛ/–/æ/, /æ/–/ɑ/}. The median token from one vowel in the pair, called the *continuum origin*, was

shifted in formant space in ten successive increments towards the other vowel in the pair (Fig. 4-2). Thus each continuum began at the median formant values of one vowel and ended at the median formant values of a neighboring vowel, with one additional token added at each end. The step size between each continuum token was constant on the mel scale, a perceptually-derived logarithmic scale based on listener pitch comparisons (1000 mels = 1000 Hz). Furthermore, two continua were generated for each vowel pair: one starting from each end. In other words, at each step on the continuum, there were two different stimuli with the same F1-F2 values, each generated using a different endpoint vowel as the continuum origin.

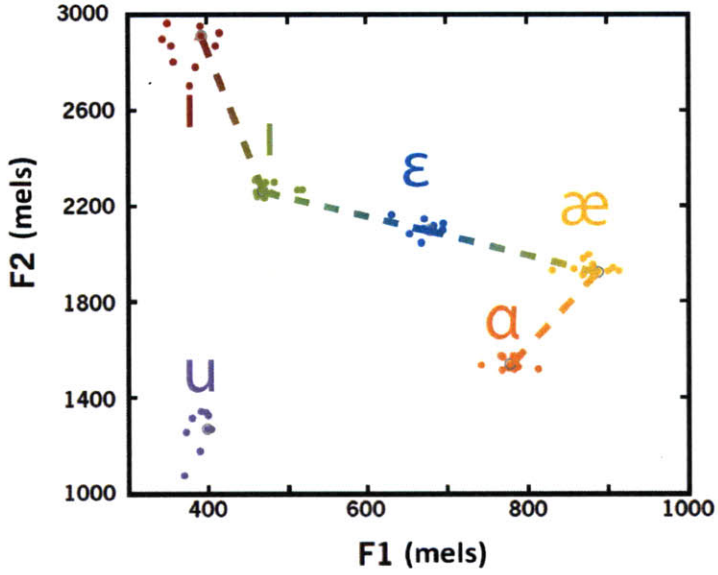

Figure 4-2. Continuum generation. Each thirteen-token continuum between pairs of vowels was generated by shifting a subject's own speech by graduated amounts in formant space. The first and second formants were shifted in the direction of each neighboring vowel. Continua formed with the vowel /u/, which was particularly far from its nearest neighbors in formant space, were judged to sound unnatural and were not used in the vowel perception test.

Vowel perception. The tokens from all eight continua were randomized and presented five times each through free-field speakers immediately following the vowel production test. Each subject heard his or her own speech and was instructed to categorize each sound as one of five possible words: bead, bid, bed, bad, or bod. The categorization data were fitted

to sigmoid curves to determine an approximate perceptual boundary between the vowels at the continuum endpoints (Fig. 4-3), defined as the point where the two sigmoid curves crossed. Furthermore, two additional points were defined: (1) *100%-within*, the token farthest from the continuum origin that was still categorized as the origin vowel 100 percent of the time, and (2) *100%-across*, the token closest to the continuum origin that was categorized as the adjacent vowel 100 percent of the time. In other words, formant values between the continuum origin and the 100%-within point were safely within the original vowel category, and formant values at the 100%-across point and beyond were safely across a category boundary, since perceptual judgments had consistently switched to a different vowel.

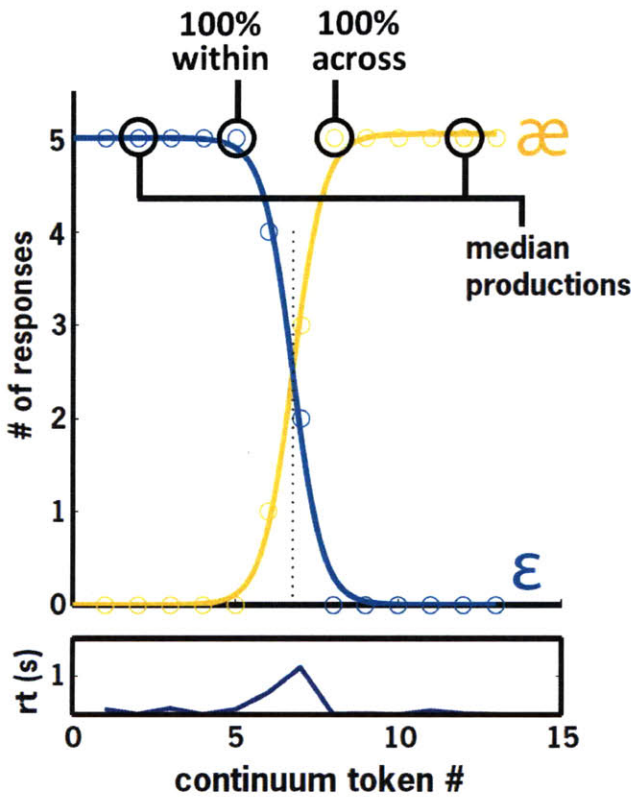

Figure 4-3. Vowel categorization responses. The identification responses for a single vowel continuum were fit to sigmoid curves. The continuum was generated by shifting the formants of the median production of /ɛ/. Shifts within a category are defined as those smaller than or equal to the shift used to generate the 100%-within token. Shifts across a category are defined are those greater than or equal to the shift used to generate the 100%-across token. As shown in the lower panel, reaction time (rt) in seconds corresponds well to the location of the category boundary.

In order to compare same-magnitude shifts within and across category boundaries, a subject's perceptual space must be "asymmetric"—the 100%-within point for one continuum must be a greater distance away from the origin than the 100%-across point for another continuum starting at the same origin (Fig. 4-4). In other words, it must take a smaller shift amount to elicit the percept of a category change in one direction than another.

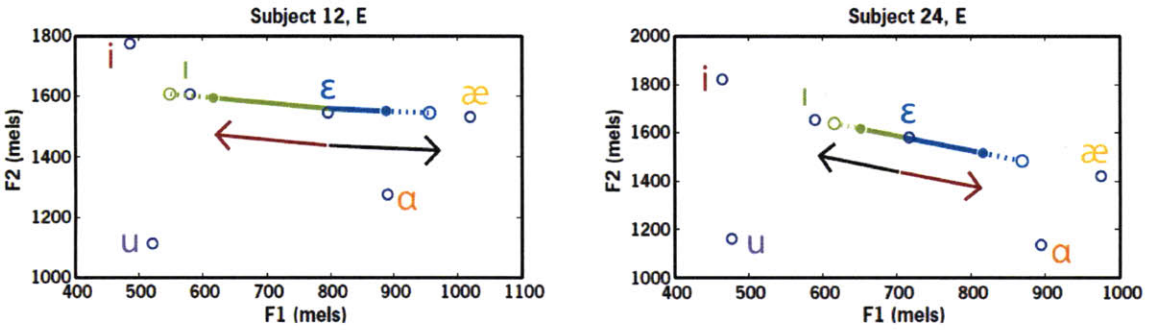

Figure 4-4. Example of counterbalanced subjects. Two sample subjects' vowel spaces overlaid with their within (red arrow) and across (black arrow) shifts. The filled circles represent the 100%-within points for the /ɛ/–/ɪ/ (green filled circle) and /ɛ/–/æ/ (blue filled circle) continua. The solid lines represent acoustic space between the continuum origin and the 100%-within point: that is, shifts that did not cause a change in vowel categorization. The open circles represent the 100%-across points for the /ɛ/–/ɪ/ (green open circle) and /ɛ/–/æ/ (blue open circle) continua. The dashed lines represent inconsistently-categorized tokens between the 100%-within and 100%-across points.

For each subject, a shift size was chosen such that it caused a category boundary to be crossed in one continuum, e.g. /ɛ/–/æ/ (see Fig. 4-4, black arrow), but not another, e.g. /ɛ/–/ɪ/ (see Fig. 4-4, red arrow). Only subjects for whom such a constant shift could be chosen—that is, whose category boundaries were asymmetric around the vowel production center—went on to complete the scanning phase. This assures that a shift of a fixed size can both effect and fail to effect a category change, depending on the direction. Counterbalancing subjects with opposite asymmetries enabled group comparisons of feedback control across and within category boundaries. Eighteen subjects qualified and went on to complete the imaging portion of the experiment. These subjects made up approximately 50% of the total subject pool.

4.2.3 Brain imaging

Functional magnetic resonance imaging (fMRI) was used to measure the BOLD response during speech, both with and without perturbation, as well as during a non-speech baseline condition. The experiment had an event-triggered design, using sparse sampling and a triggering mechanism to coordinate stimulus timing with image acquisitions.

Imaging parameters. Subjects were scanned in a 3T Siemens Tim Trio whole-body MRI machine equipped with a 32-channel volume transmit-receive birdcage head coil, located at the Athinoula A. Martinos Imaging Center at McGovern Institute for Brain Research, MIT. The subjects' speech was recorded via a custom-made MR-safe microphone, and auditory feedback was delivered via insert headphones (Stax SRS-005II electrostatic headphones). All auditory feedback had a short (~17 ms) delay owing to the processing time of the formant shift. Subjects wore supra-aural ear seals surrounded by a custom-made foam helmet, affectionately nicknamed the Head Cozy, to insulate them from the noise of the scanner.

Images were acquired in a head-first, supine position. Because of the constraints imposed by the scanner, the vowels could not be recorded with the head in an upright posture, as in the behavioral pretest. Studies examining speech acoustics under upright and supine positions have noted changes in articulation under the different gravitational loads imposed by different postures (Stone et al., 2007). However, these differences in jaw and tongue placement had little effect on acoustic output: changes were largely restricted to formant bandwidth changes, with only minor shifts in formant values (Tiede et al., 1997; Whalen, 1990). Therefore, even though postural constraints may have caused differences in vocal tract shape or jaw movements between the two experimental phases, speakers are able to use auditory feedback to tune vocal tract constrictions and maintain precise formant values.

Experimental paradigm. At the start of each trial, subjects were visually presented with a word (e.g. "bed") or a control stimulus ("***"). The words were drawn from a list of eight that depended on the vowel to be perturbed (see Appendix A). These stimuli were projected in high-contrast white-on-black and displayed on a rear projection screen, visible to the

subjects through a mirror mounted above the MRI head coil, and remained onscreen for two seconds. Stimulus delivery was controlled by custom software written in Matlab. Subjects were instructed to clearly read each word aloud when it appeared on the screen and to remain silent on the control trials. Immediately after each trial, a volume meter gave subjects feedback about the loudness of their speech. Each of five experimental runs consisted of 80 trials: 64 speech trials (eight presentations each of eight words) and 16 silent control trials.

Unbeknownst to the subjects, the speech trials were divided into three conditions: *NoShift* (normal speech feedback), *Within* (a shift was applied in the direction that did not cause a category change in the behavioral pretest), or *Across* (the same size shift was applied in the direction that did cause a category change in the behavioral pretest). The *Within* and *Across* trials each made up one-eighth of the total experimental trials, for a total of one-fourth perturbed trials. In these random 25% of trials, the formants were perturbed before being fed back to the subjects' headphones. The resultant perturbed trials sounded like mispronunciations of the trial word; the auditory output the subjects expected to hear did not correspond with the artificially-shifted output of the headphones. Trial order was randomly permuted at the start of the experiment.

In summary, the four conditions experienced by each subject were:

1. **Baseline:** a control condition in which the subject remained silent.
2. **NoShift:** speech feedback was unchanged.
3. **Within:** a within-category shift was applied to the subjects' speech.
4. **Across:** a cross-category shift was applied, the same magnitude as that of the shift-within.

Trial timeline. A sparse sampling design was used, similar to other recent studies of speech production (Birn et al., 2004; Bohland & Guenther, 2006; Ghosh et al., 2008; Tourville et al., 2008) (Fig. 4-5). After a four-second delay from the visual stimulus onset, the custom-written stimulus presentation software triggered the scanner to collect a single volume of functional data (TA = 2.75s). The delay allowed volume acquisition to occur near the peak of the hemodynamic response to speech, estimated to occur approximately 4-7 seconds post-vocalization. The functional volume was followed by a pause of 1.25 seconds

before the start of the next trial, a total trial length of eight seconds, to allow for the partial return of the BOLD signal to the steady state. Because the volume acquisition was timed to occur several seconds after the stimulus offset, subjects spoke in relative silence, an advantage of event-triggered designs. Furthermore, auditory feedback to the headphones was turned off during image acquisition to prevent the transmission of scanner noise.

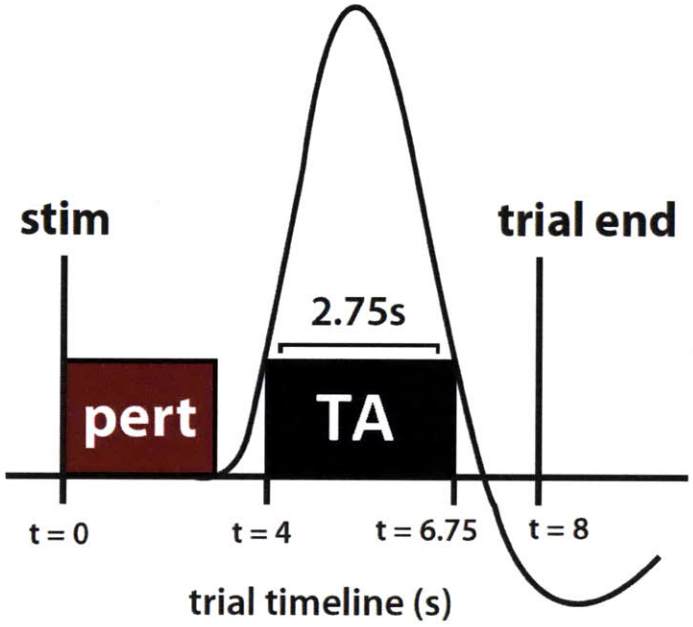

Figure 4-5. Timeline for a single trial in the fMRI experiment. The visual stimulus appeared at t=0 and lasted 2 seconds, during which perturbation was applied (if a perturbed trial). Four seconds after stimulus onset, a single volume as acquired (TA = 2.75s). The interscan interval was 8 seconds.

The sparse sampling design afforded several important advantages. First, because subjects spoke during relative silence, they could hear their own speech with no masking from the loud scanner noise. Second, the silent interval assured a relatively clean recorded signal, allowing the online speech processing and formant perturbation to be correctly applied. Finally, since the volume acquisition followed articulation by several seconds, there were no artifacts from tongue, jaw or head movement during speech.

Volume acquisition paramters. Functional volumes consisted of 45 T2*-weighted gradient echo, echo planar images aligned to the bicommissural line and covering the entire

cortex and cerebellum in the axial plane (3-mm slice thickness, 0.3 mm gap between slices, TA = 2750 ms, TE = 30 ms, flip angle = 90°, FOV = 200 mm^2). In addition to the functional data, anatomical volumes were collected in order to overlay each subject's functional data on a structural image of his or her own brain. A high resolution T1-weighted anatomical volume (128 slices in the sagittal plane, slice thickness = 1.33 mm, in-plane resolution = 1 mm^2, TR = 2000 ms, TE=3300 ms, flip angle = 7°, FOV = 256 mm^2) was collected prior to functional imaging. Diffusion tensor imaging data was also collected to track white matter tracts as they travel between connected regions of the cortex. These data were used for structural connectivity analyses between brain regions implicated in the task.

4.2.4 Auditory feedback perturbation

Subjects' recorded speech was routed through a patch panel and split into two channels using a MOTU UltraLite FireWire audio interface with on-board mixer (48 kHz sampling rate). One channel of the signal was sent to the laptop to be recorded while the other was processed on the on-board sound card. This processed signal was re-split and sent both to the laptop and back out to the subject's headphones. Because the same procedure was used for all trials, the signal underwent the same processing delay of approximately 17 ms whether or not the formants were shifted on a given trial.

Formant tracking and perturbation was carried out in the manner described by Boucek (2007) and by Cai and colleagues (2008). The speech audio signal was downsampled by a factor of four (12 kHz), then pre-emphasized to improve formant estimation by accounting for the –6 dB/octave high-frequency spectral slope typically present in the speech signal (Fant, 1960). Vowel onset and offset were detected using a root mean square (rms) threshold and rms ratio threshold. The voiced signal was then analyzed using a linear predictive coding (LPC) algorithm and the autocorrelation method to estimate the vocal tract transfer function as an all-pole model. The LPC order for each subject determined from formant-tracking performance in the behavioral pretest (9^{th}–13^{th}-order). During the vowel, formants of the incoming signal were shifted by filtering the signal through a concatenation of two digital biquad infinite impulse response (IIR) filters. These filters first

add zeros at the detected formant frequencies to neutralize the original poles, then add new poles that are shifted in frequency by the desired amount. Finally, because the formant shift changes the gain of the spectral peaks, a gain factor was applied to the filter output before the signal was upsampled and written to the sound card output buffer.

The applied two-dimensional formant shifts were constant in magnitude and direction over the duration of the vowel. During shifted speech, the vowel formants moved toward those of the neighboring vowels, either crossing or not crossing category boundaries depending on the perceptual space of each subject. To control for possible effects due to perturbation direction, each subject had a counterpart for which the Across and Within shift directions were opposite. Thus, for every subject whose across-category shift was /ɛ/–/ɪ/, there was one whose within-category shift was /ɛ/–/ɪ/. Fourteen subjects produced /ɛ/; seven of these subjects were shifted across the /ɪ/ boundary but within the /æ/ boundary, while the other seven were shifted the /æ/ boundary but within the /ɪ/ boundary. Similarly, four subjects produced /æ/; two of these were shifted across the /ɛ/ boundary but within the /ɑ/ boundary, and the other two were shifted across the /ɑ/ boundary but within the /ɛ/ boundary. A summary of subject perturbation conditions is presented in Table 4-1.

Subject ID	*Target vowel*	*Within*	*Across*
S12	ɛ	ɛ–ɪ	ɛ–æ
S13	æ	æ–ɛ	æ–ɑ
S14	ɛ	ɛ–æ	ɛ–ɪ
S15	ɛ	ɛ–ɪ	ɛ–æ
S20	ɛ	ɛ–æ	ɛ–ɪ
S21	ɛ	ɛ–æ	ɛ–ɪ
S22	ɛ	ɛ–ɪ	ɛ–æ
S24	ɛ	ɛ–æ	ɛ–ɪ
S28	ɛ	ɛ–ɪ	ɛ–æ
S29	æ	æ–ɑ	æ–ɛ
S38	ɛ	ɛ–æ	ɛ–ɪ
~~S44~~	~~ɛ~~	~~ɛ–ɪ~~	~~ɛ–æ~~
S45	ɛ	ɛ–ɪ	ɛ–æ
~~S46~~	~~ɛ~~	~~ɛ–æ~~	~~ɛ–ɪ~~
S47	æ	æ–ɑ	æ–ɛ
S52	ɛ	ɛ–ɪ	ɛ–æ
S56	æ	æ–ɛ	æ–ɑ
S58	ɛ	ɛ–æ	ɛ–ɪ

Table 4-1. Subject-specific frequency shifts in formant space. The target vowel was the vowel produced by the subject in the fMRI experiment. "Within" and "Across" show the direction of the within-category and cross-category shifts, respectively. Table rows are color-coded with like subjects matched. The blue and green rows have opposite shift directions and thus counterbalance each other; the yellow and orange rows similarly counterbalance each other. Subject 44 was later excluded owing to lack of normal activation in the Speech–Base baseline contrast; a counterbalancing subject (47) also excluded to keep the shift directions balanced in each condition.

When questioned after the scanning session, nine subjects (50%) reported no awareness of any feedback alteration, while nine (50%) did report awareness of a change. Of these, many could not articulate what had changed, but several specified the direction of perturbation, saying, e.g., that "*cab* sounded like *cob*." Of particular note, one subject reported that sometimes his "/ɛ/'s were like /æ/'s, but sometimes they were like a British accent." Another thought that "*tech* sounded like *tack*, and *bed* sounded like an Australian accent." These perceptions of vowel change and accent change corresponded to conditions of across-phoneme and within-phoneme perturbation, respectively.

4.2.5 Data analysis

Acoustic analysis. Acoustic data were compared across no-shift, shift-within, and shift-across conditions. The first and second formants tracked with LPC analysis were zero-phase filtered with an 8-point Hamming window. Formant values at each time point were averaged across all no-shift trials to yield a baseline vowel trajectory in two-dimensional formant space. Averaged F1 and F2 trajectories for the shift-within and shift-across conditions were then compared with the baseline trajectory. The greater the deviation from the baseline at each time point, the greater the measured compensatory effect. Compensation to perturbation was defined as a statistically significant ($p < 0.05$) deviation from the baseline trajectory during perturbed trials, using a "fixed effects" analysis with each trial contributing a degree of freedom.

Each subject had custom-defined shift magnitudes and directions—his own personal *shift vector* pointing towards the neighboring vowels—making a simple F1 or F2 comparison across the subject population impossible. In order to compare compensation across subjects, the two-dimensional distance between baseline and shifted conditions must be compressed into one dimension. The *2D-difference* was defined as the Euclidean distance in 2D formant space at each time point, computed by taking the square root of the sum of the squared F1 and squared F2 differences. Because this 2D-difference is positive whether subject responses counteract the perturbation or enhance it, it is not a good measure of true compensation. For a response to be considered compensatory it must mitigate the effects of

the perturbation and reset the acoustic output closer to its originally-intended values; thus, it should oppose the shift vector. The *projection* was therefore defined as the projection of the 2D-difference vector onto the inverse shift vector: the dot product of the two vectors divided by the shift magnitude. In other words, the projection is the component of the 2D-difference that is in opposition to the shift (the blue line in Fig. 4-6).

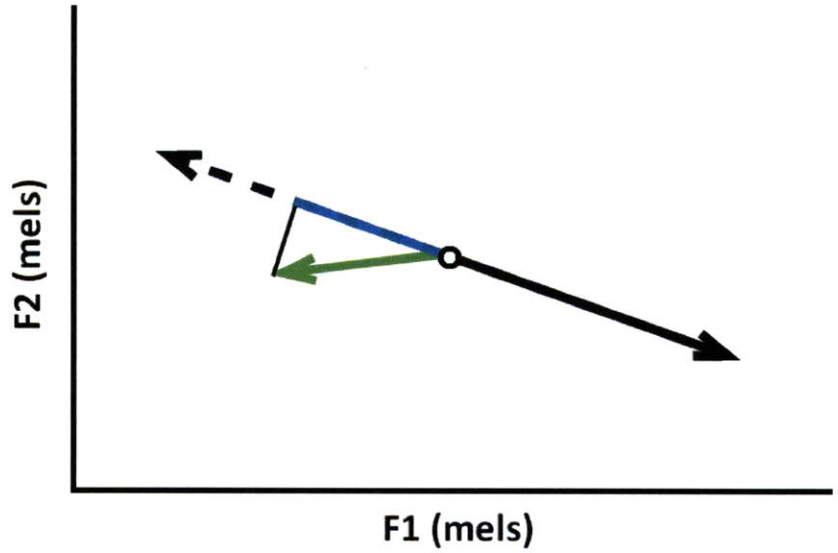

Figure 4-6. Schematic of projection and efficiency. The black shift vector (solid line) is reflected around the production (open circle) to give the inverse shift vector (dashed line), representing "perfect compensation." The actual deviation resulting from the shift (green arrow) is projected onto the inverse shift vector to yield the projection (blue line), the component directly opposing the shift. The efficiency is the ratio of the magnitudes of the blue and green lines.

Finally, the *efficiency* of compensation was defined as the projection as a percentage of the 2D-difference. This measure is equivalent to the angle between the inverse shift vector and the 2D-difference vector, scaled from –100 to 100. Responses that are perfectly aligned with the inverse of the shift vector (0°) have maximal efficiency (100%), while responses in the same direction as the shift (180°) have –100% efficiency.

Using projection traces as the primary measure of compensation, the integral of the deviation from baseline was calculated for each condition, as was the maximum excursion from the baseline (Fig. 4-7). The maximum was computed within an appropriate time

window: no earlier than 60 ms from voicing onset to allow time for the response, and no later than the time point at which one-quarter of the trials to avoid too few samples. The integral and maximum of the projection traces were also computed for individual subjects in order to compare their compensation performance.

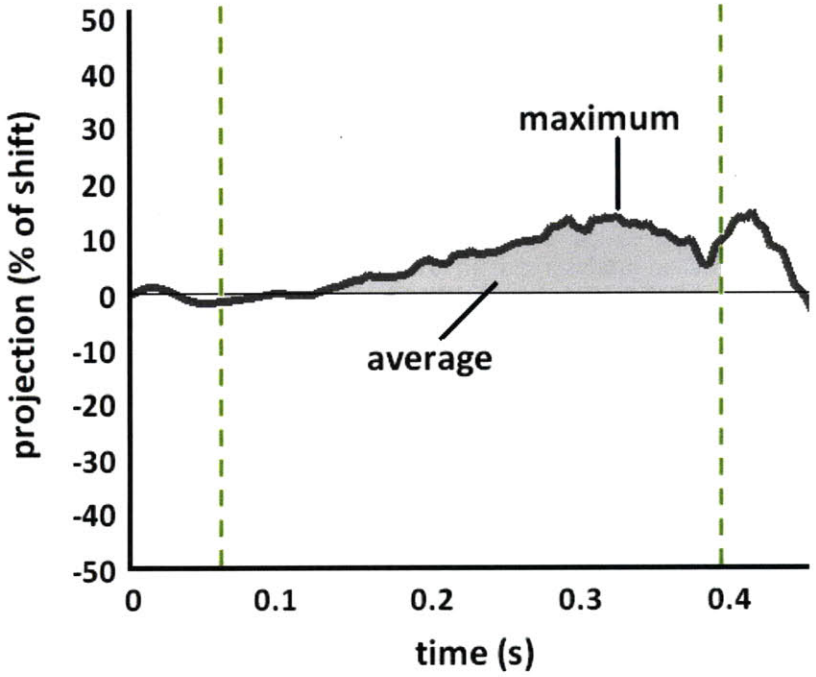

Figure 4-7. Schematic of maximum and average projection. The maximum and average were computed for the timepoints between the dashed lines, where the left line is at 60 ms to allow time for the onset of the response, and the right line is the "quarter-trial" point, the time point until which exactly one quarter of trials lasted.

Functional imaging analysis. Both voxel-based and surface-based analyses of activation were carried out to assess task-related activation. The first contrast of interest was the Shift-NoShift contrast, in which activation from both Within and Across trials was compared with activation from the NoShift speech condition. Additionally, the within-category activation patterns (Within–NoShift) were compared with the cross-category activation patterns (Across–NoShift) in terms of cortical location and extent of activation.

The functional imaging data were pre-processed and analyzed using publicly-available software packages including SPM (Friston et al., 1995), Freesurfer (Dale et al., 1999;

Fischl et al., 1999), and FSL (Smith et al., 2004; Woolrich et al., 2009). In the pre-processing stage, a rigid-body transformation was used to realign functional images to the mean EPI image, correcting for subject head movement. The realigned images were stripped of non-cortical matter via a brainmask computed with FSL's brain extraction tool, BET (Smith, 2002). Outliers with more than 2 mm of movement or with an intensity z-threshold more than 3 standard deviations from the mean were removed from the analysis using artifact detection tools (Whitfield-Gabrieli, 2009). Images were coregistered with the T1-weighted anatomical image and spatially normalized into the Montreal Neurological Institute (MNI) space (Evans et al., 1993). Finally, in the voxel-based analysis stream, the images were smoothed with a Gaussian filter (8 mm full width at half maximum).

In the surface-based analysis stream, Freesurfer was used to segment each anatomical volume into gray and white matter structures and to perform cortical surface reconstruction. The same preprocessing was applied as in the voxel-based stream except that smoothing was done on the cortical surface (6 mm full width at half maximum smoothing kernel), rather than in the volume.

For each condition of interest, a time series of finite impulses was created to represent the onsets of each event. This time series was then convolved with a canonical hemodynamic response function (HRF), generating a simulated BOLD response. The regressors for each volume were computed by sampling the height of the simulated BOLD response at the time that volume was acquired. The regressors were therefore weighted, taking into account neural responses both to the immediately preceding event and to any previous events whose resulting HRFs had not entirely decayed (Ghosh et al., 2009). These regressors were used in the general linear modeling analysis.

A standard hierarchical group model approach was used to model within-subjects and between-subjects effects (Friston et al., 2005). Contrast images were generated for each subject. Conditions were treated as fixed effects. A "summary statistics" procedure was used to model the group effects, performing one-sample t-tests across the individual contrast images. The model was applied with a p-value threshold of 0.05 and family-wise error (FWE) correction for multiple comparisons. The Automated Anatomical Labeling (AAL)

toolbox (Tzourio-Mazoyer et al., 2002) was used to identify anatomical regions for active clusters in the activation maps.

ROI analysis. Because of inter-subject variability, the alignment of functional images from multiple subjects is far from perfect. Even when normalized to a standardized stereotactic space, voxel-by-voxel comparisons are confounded by local anatomical variability. Nieto-Castanon and colleagues (2003) quantified the inter-subject overlap for various anatomically-defined brain regions by pooling different subject group sizes, demonstrating a mean overlap of 31% for two subjects. This overlap dropped to 13% for three subjects and 0% for nine subjects, a small population for imaging studies, and only half the number of participants in the current study. This problem of low overlap between subjects is typically mediated by spatially smoothing the functional data, increasing the overlap and thus the power of voxel-based analysis. However, localization of functional activation is poorer, since smoothing blurs regional boundaries, even across sulci.

A region of interest (ROI) analysis addresses this problem. This analysis compares functional responses across like anatomical regions defined from individual landmarks. By tailoring the functional analysis to the structural space of each subject prior to averaging, this method accounts for inter-subject anatomical variability and better maintains the link between structure and function.

To perform the ROI analysis, the cortex of each subject was parcellated into units using the Freesurfer cortical classifier. The classifier was trained on a set of 14 manually-parcellated brains whose speech-related ROIs were subdivided for a finer resolution (Tourville and Guenther, 2003). The BOLD response was averaged across all voxels within each ROI. The responses were fit to the same set of condition regressors used in the voxel- and surface-based analyses.

Regression analysis. The first-level voxel-based analysis yielded t-contrast maps for each subject. These t-contrast maps were then used in a simple regression analysis with the amount of compensation as a covariate measure. Amount of compensation was defined as the mean projection. F-contrast map shows the regions that have a statistically significant correlation with behavioral measures at the $p < 0.001$ level, uncorrected.

4.3 Results

4.3.1 Behavioral results

Vowel categorization data gathered during the psychophysical pretest show relatively sharp and consistent category boundaries, as evidenced by the steep slope of the sigmoid fit to the forced-choice categorization data (Fig. 4-8). For most subjects, the perceptual boundaries differed based on the continuum origin—that is, an /ɛ/–/æ/ continuum generated from an /ɛ/ token had a boundary much closer to /æ/ than an /ɛ/–/æ/ continuum generated from an /æ/ token. Even though the tokens were presented randomly, the percept from the original vowel tended to dominate each continuum.

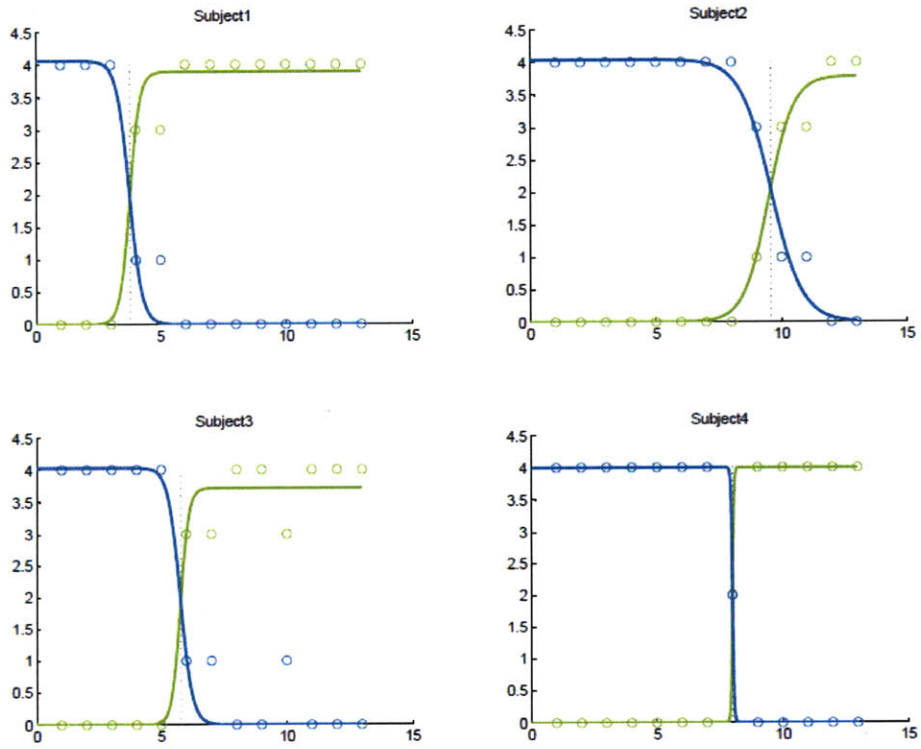

Figure 4-8. Sample vowel categorization data. As described in Figure 4-3, the identification responses for a single vowel continuum were fit to sigmoid curves. The location of category boundaries varied from subject to subject for the same vowel pair, and within a single subject across different vowel pairs.

Furthermore, there was a great deal of variability in the location of vowel boundaries, both across subjects for a given continuum, and across continua for a given subject (Fig. 4-8). That is, it takes a larger shift in formant space to elicit the perception of phonetic change for some subjects than for others, and for some vowel pairs than others. This variability enabled the direct comparison of within- and cross-category shifts of the same size within a single subject.

The production task in the scanner allowed for the comparison of the two perturbation conditions with the unshifted baseline speaking condition. Average formant trajectories for three sample subjects, chosen for their clear deviations from baseline in shifted conditions, are shown in Figures 4-9 through 4-11. Subjects responded to the unexpected shifts in formant space by altering their formant trajectories away from the baseline. Deviations from the black baseline trajectory indicated compensatory responses in the perturbed conditions.

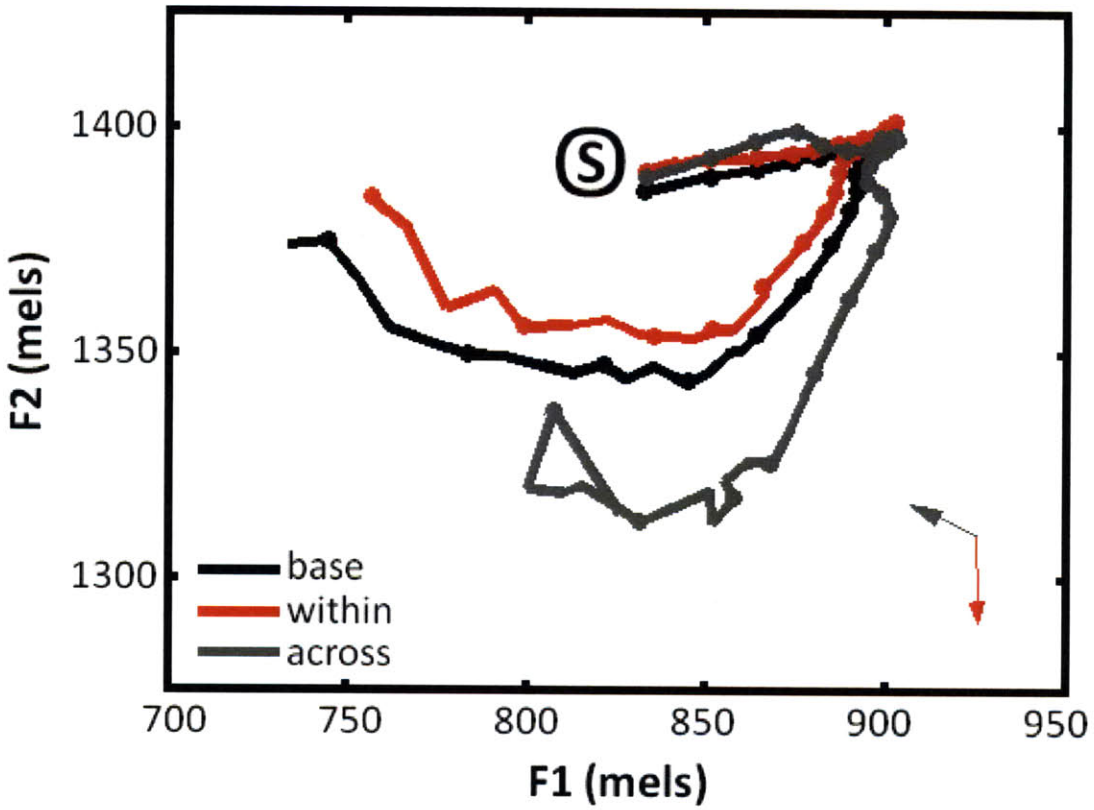

Figure 4-9. Average formant trajectories for /æ/, subject 47. The black line is the baseline production trajectory, averaged over all *NoShift* trials. The red and grey lines are the trajectories produced during the shifted conditions, averaged over all *Within* and *Across* trials, respectively. Trajectories are plotted from the onset of voicing (marked with an "S") to the "quarter-trial" point, the time point until which exactly one quarter of trials lasted. For comparison, the arrows show the direction (though not the magnitude) of the imposed shift during *Within* (red) and *Across* (grey) trials.

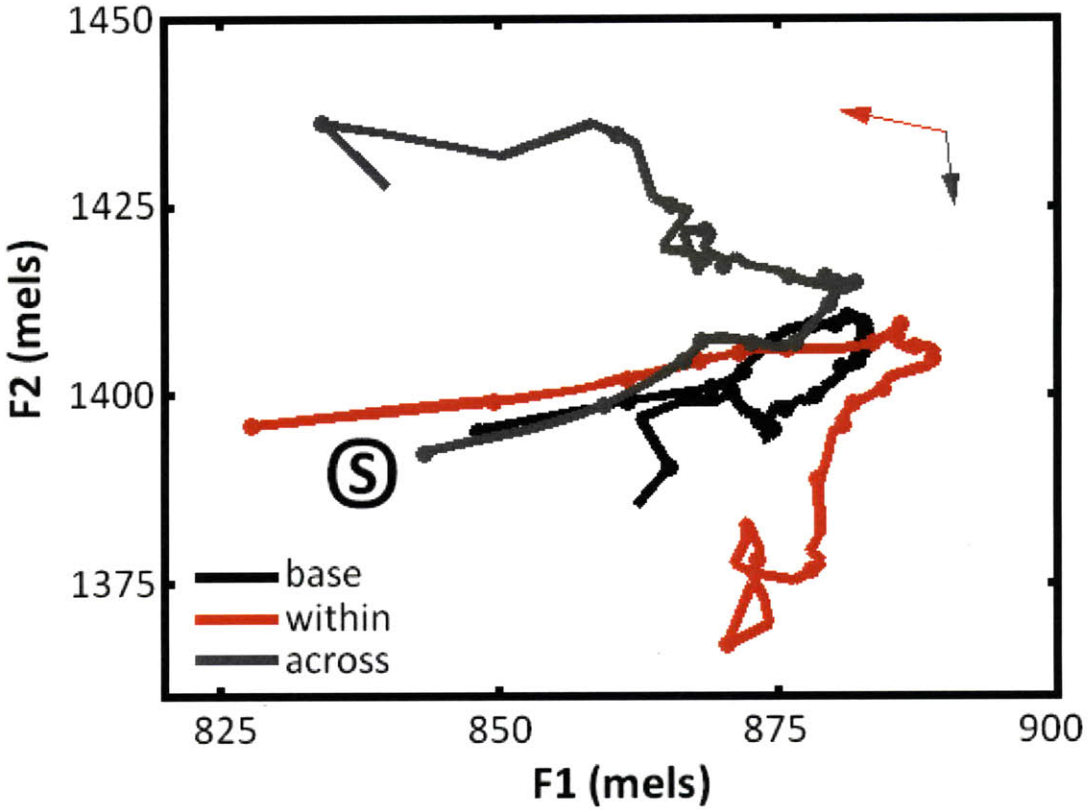

Figure 4-10. Average formant trajectories for /æ/, subject 56. (See Fig. 4-8 for figure details.) As in Figure 4-8, the arrows show the direction (though not the magnitude) of shift during *Within* (red) and *Across* (grey) trials.

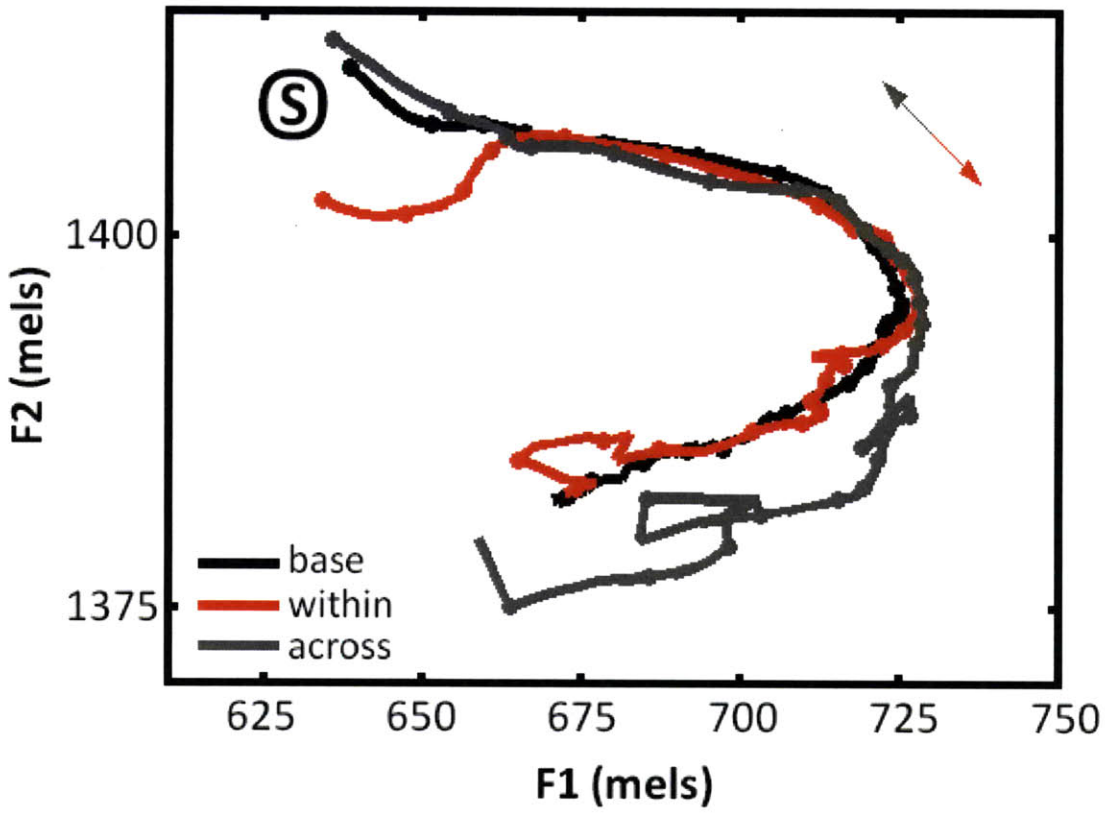

Figure 4-11. Average formant trajectories for /ɛ/, subject 21. (See Fig. 4-8 for figure details.) As in Figure 4-8, the arrows show the direction (though not the magnitude) of shift during *Within* (red) and *Across* (grey) trials.

Figure 4-12 shows the average projection: the component of the trajectory's deviation that is in opposition to each subject's custom shift. Because the projection is a magnitude, it can be averaged across all subjects, regardless of the individual's shift direction. Additionally, because a compensatory response always translates to a positive projection, the y-axis for the Within projection has been flipped with respect to that for the Across projection, to better show the separation from baseline.

Maximum projection magnitudes for each subject ranged from –20.8 to 89.4 mels (min Within: –20.8, max Within: 47.9, mean Within: 12.9; min Across: 3.4, max Across: 89.4, mean Across: 30.7). Average projection magnitudes per subject ranged from –57.2 to 24.3 mels (min Within: –107.3, max Within: 19.1, mean Within: –1.9; min Across: –25.3, max Across: 56.1, mean Across: 10.6). The projection traces deviate from the baseline at approximately 150 ms after the onset of voicing (t = 0 in Fig. 4-12). Moreover, a two-tailed unpaired *t*-test showed that the magnitude of the projection is greater for the Across condition than for the Within condition ($p < 0.05$; see Fig. 4-13).

When questioned, a subset of subjects (n = 8) reported some conscious awareness of an auditory manipulation to their speech. However, an unpaired *t*-test performed on the two sets of subjects, those aware and those unaware of the perturbation, suggests that both compensated to the same degree ($p = 0.48$).

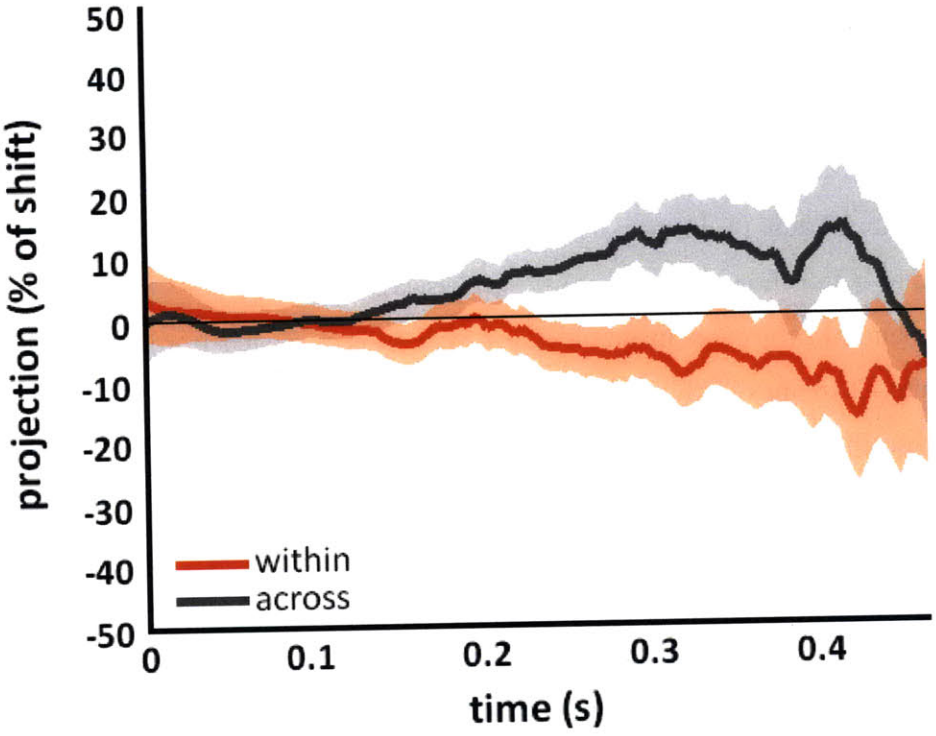

Figure 4-12. The projection of the 2D-difference vector onto the inverse shift vector, averaged across all subjects. The axis is flipped for the *Within* condition (positive projection is down) to better display the separation from baseline.

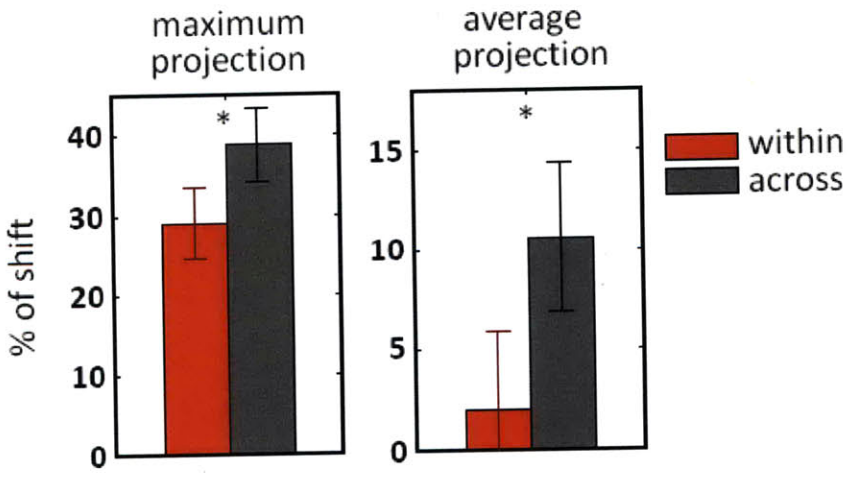

Figure 4-13. Comparison of the projection magnitude for *Within* and *Across* trials. Bars show 95% confidence intervals. The projection magnitude of the *Across* condition is significantly greater than that of the *Within* condition ($p < 0.05$), whether measured by the maximum pert-trial or the average across the trial duration.

Figure 4-14 shows the average efficiency across all subjects. As in projection, a compensatory response always translates to a positive efficiency; thus, as in Figure 4-12, the y-axis for the Within projection has been flipped with respect to that for the Across projection, to better show the separation from baseline.

The maximum efficiency for each subject ranged from –74.9% to 100.0% (min Within: –74.9%, max Within: 100.0%, mean Within: 49.3%; min Across: 54.4%, max Across: 100.0%, mean Across: 58.0%). Average efficiency per subject ranged from –96.6% to 97.2% (min Within: –96.6%, max Within: 90.4%, mean Within: 7.6%; min Across: –43.9%, max Across: 97.2%, mean Across: 17.2%). The efficiency traces deviate from baseline at approximately 150 ms after the onset of voicing. The mean efficiency for the Across condition is greater than that for the Within condition by 10%, trending toward significance (see Fig. 4-15).

While the average efficiency over all subjects is less than 15%, many subjects reached near-maximal efficiency during perturbed trials — three subjects with a mean efficiency above 90% and five more with a mean efficiency above 80% (see Figs. 4-16 and 4-17 for examples of individual subjects). In other words, these subjects altered their formants in a direction that aligned almost perfectly with the imposed shift. As expected from past studies, the magnitude of compensation was not enough to counteract 100% of the shift, but the direction of compensation approached a perfect inverse of the shift vector.

Efficiency was found to strongly correlate with projection ($r = 0.7$), but this is largely because the two measures are mathematically dependent on each other (that is, the efficiency is defined as the projection divided by raw 2D-difference).

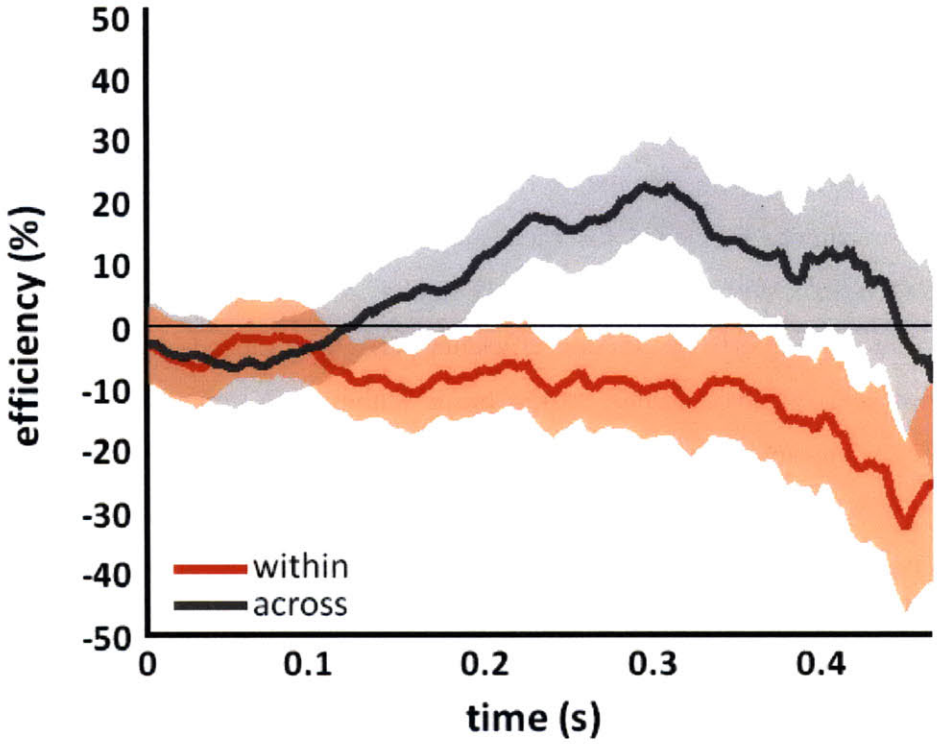

Figure 4-14. The efficiency of compensation, or the ratio of the projection and the 2D-difference, averaged across all subjects. As in Figure 4-11, the axis is flipped for the *Within* condition (positive projection is down) to better display the separation from baseline

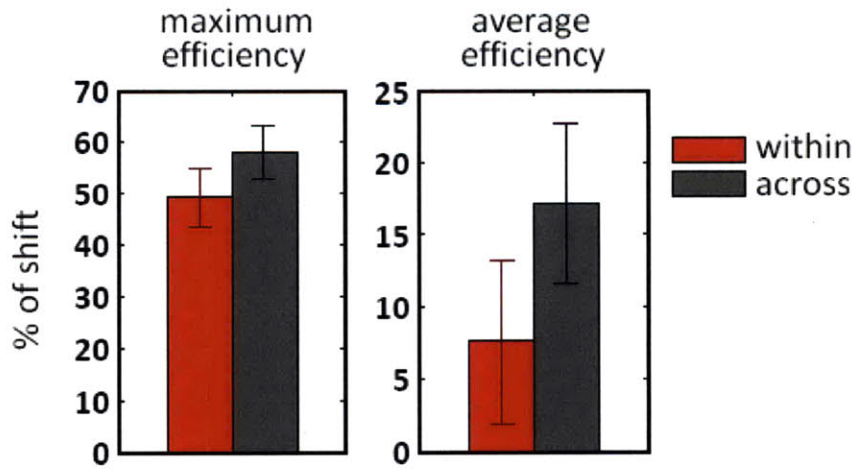

Figure 4-15. Comparison of compensation efficiency for *Within* and *Across* trials. Bars show 95% confidence intervals. The efficiency of the *Across* condition is on average greater than that of the *Within* condition, trending towards significance.

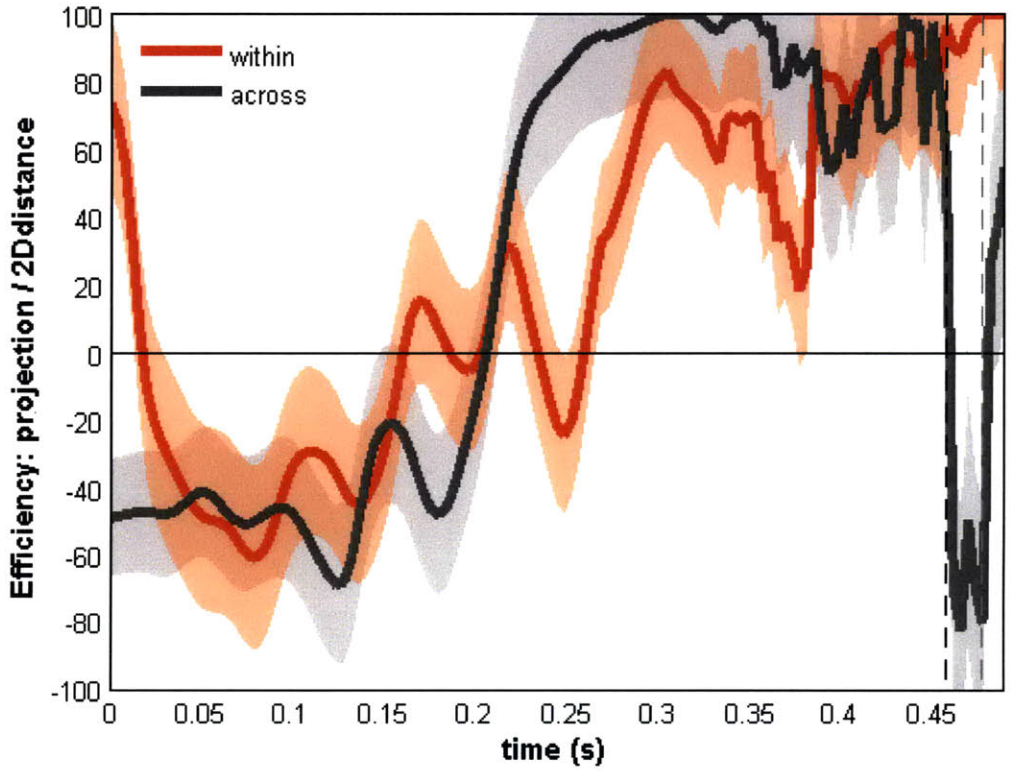

Figure 4-16. Efficiency for Subject 13. Efficiency is graphed from time of voice onset (t = 0) to the quarter-trial point (see Figure 4-8).

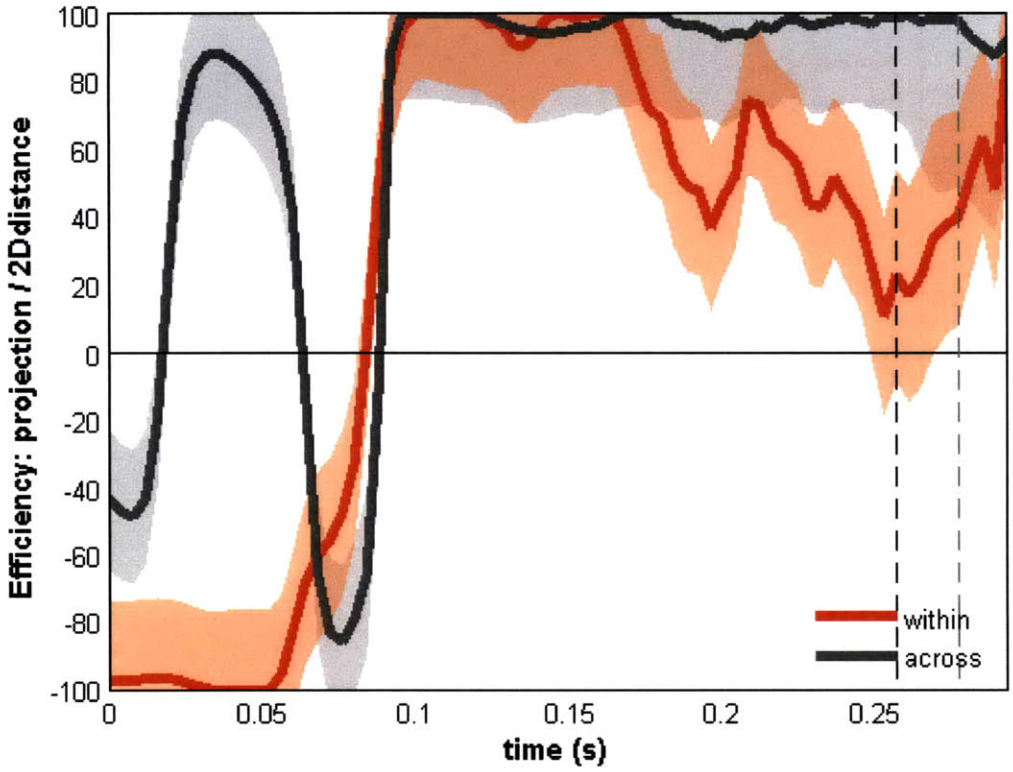

Figure 4-17. Efficiency for Subject 56. Efficiency is graphed from time of voice onset (t = 0) to the quarter-trial point (see Figure 4-8).

4.3.2 Functional imaging results

Mean activation analysis. Figures 4-18 to 4-22 show the averaged activation maps for the Speech–Baseline, Shift–NoShift, Within–NoShift, and Across–NoShift, and Across–Within conditions using a mixed-effects analysis of surface-smoothed data. Figures 4-23 to 4-42 show the averaged activation maps for these five conditions using a mixed-effects analysis of volume-smoothed data. Activations for these experimental conditions are summarized in Tables 4-2 to 4-5. Summaries of active regions are as follows:

- Speech–Baseline: activation is found in the expected "speech network" consisting of bilateral primary motor cortex, bilateral medial prefrontal cortex, and bilateral auditory cortical areas. Cortical activation in visual occipital cortex was also found, presumably owing to semantic differences in visual stimuli, which consisted of meaningful letterforms in the speech conditions and patterns of asterisks in the baseline condition ("bed" versus "***").
- Shift–NoShift: activation was seen in bilateral posterior superior temporal gyrus (pSTg: see Fig. 4-26), bilateral insula, bilateral supplementary motor area (SMA: see Fig. 4-27), bilateral inferior frontal gyrus (IFg pars opercularis and pars triangularis: see Fig. 4-25), right middle temporal gyrus (MTg), angular gyrus, and right supramarginal gyrus (SMg: see Fig. 4-28).
- Within–NoShift: activation was seen in only a subset of the Shift–NoShift areas, including left (but not right) pSTg (see Fig. 4-32), left insula, bilateral IFg (pars opercularis and pars triangularis: see Fig. 4-31), right SMA, and right SMg.
- Across–NoShift: activation was seen in a larger subset of the Shift–NoShift areas, including bilateral pSTg, bilateral SMA, left angular gyrus, and right SMg.
- Across–Within: activation was seen in pSTG (see Fig. 4-40), left IFg (see Fig. 4-39), bilateral precuneus, bilateral fusiform gyrus, and bilateral lingual gyrus.

Speech–Base surface view

left hemisphere *right hemisphere*

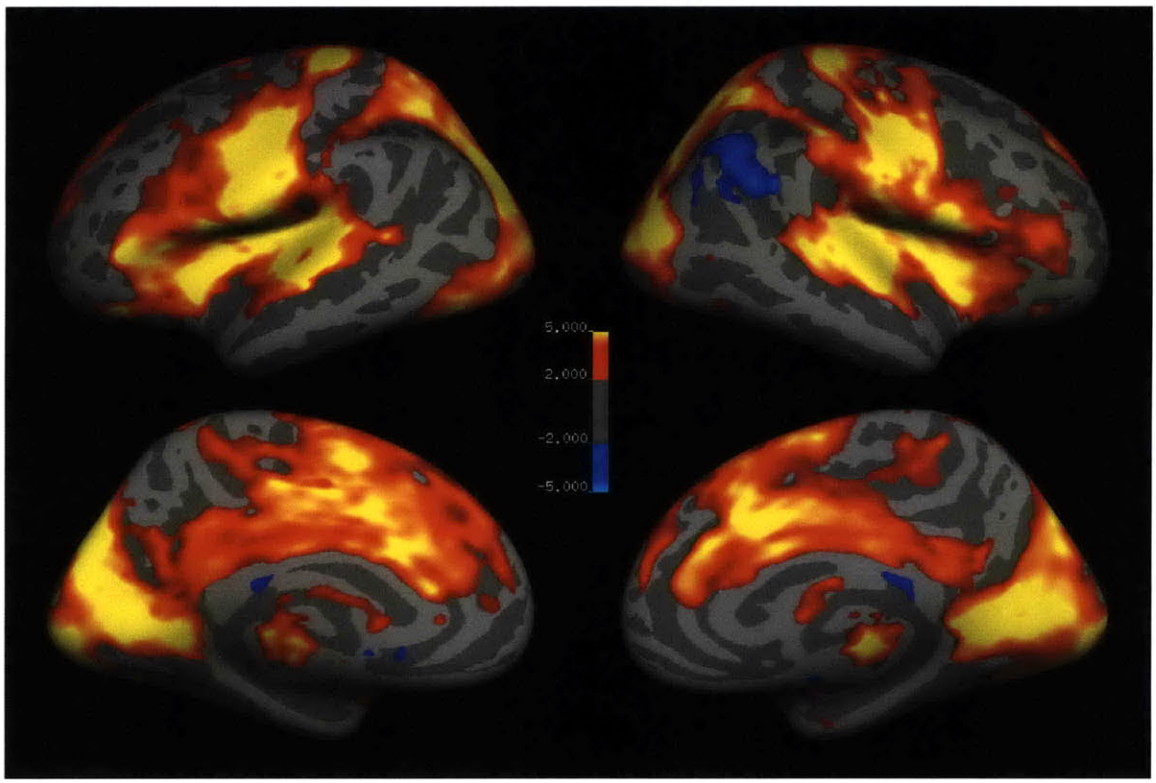

Figure 4-18. Speech–Base surface view, $p < 0.01$ (uncorrected). Mixed-effects analysis of surface-based smoothing, with lateral (top row) and medial (bottom row) views of left and right hemispheres. Light and dark gray correspond to gyri and sulci, respectively. The color bar is a T-scale, where yellow and red indicate greater activation in Speech conditions than in Baseline condition, and blue and cyan indicate greater activation in the Baseline condition than in the Speech conditions.

Shift–NoShift surface view

left hemisphere *right hemisphere*

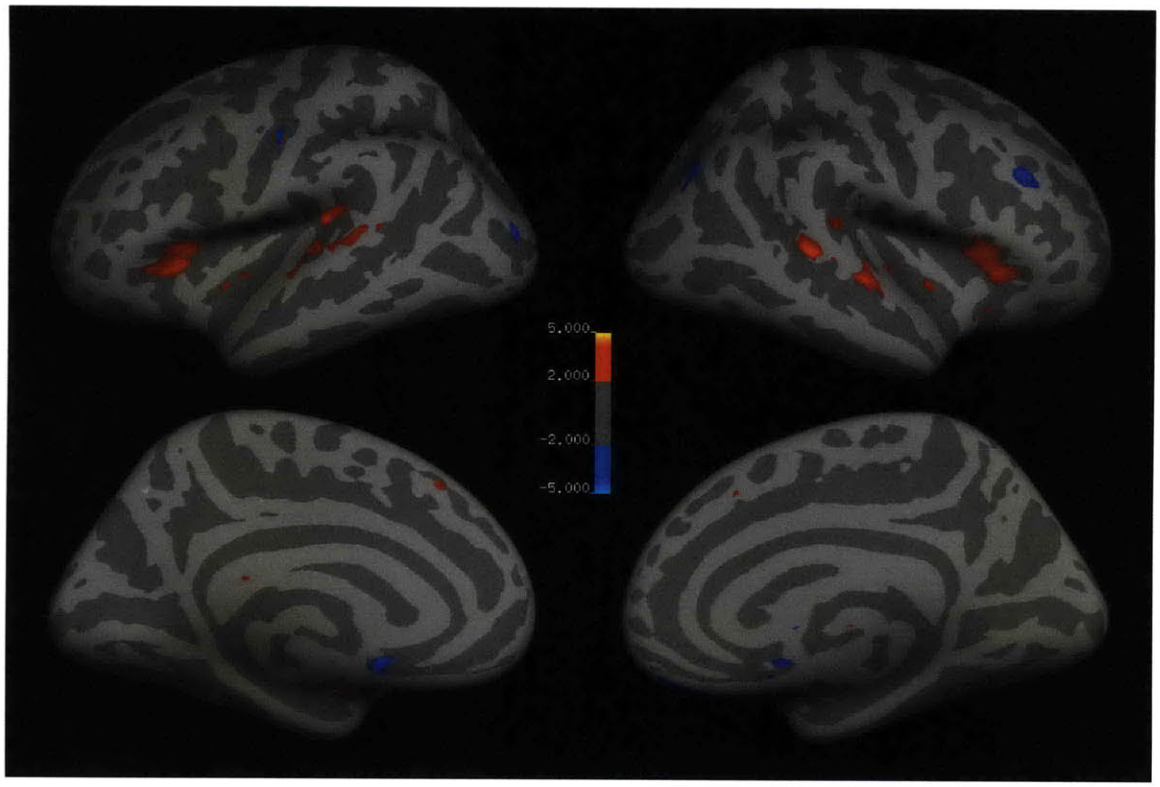

Figure 4-19. Shift–NoShift surface view. (See Fig. 4-18 for figure details.)

Within–NoShift surface view

left hemisphere *right hemisphere*

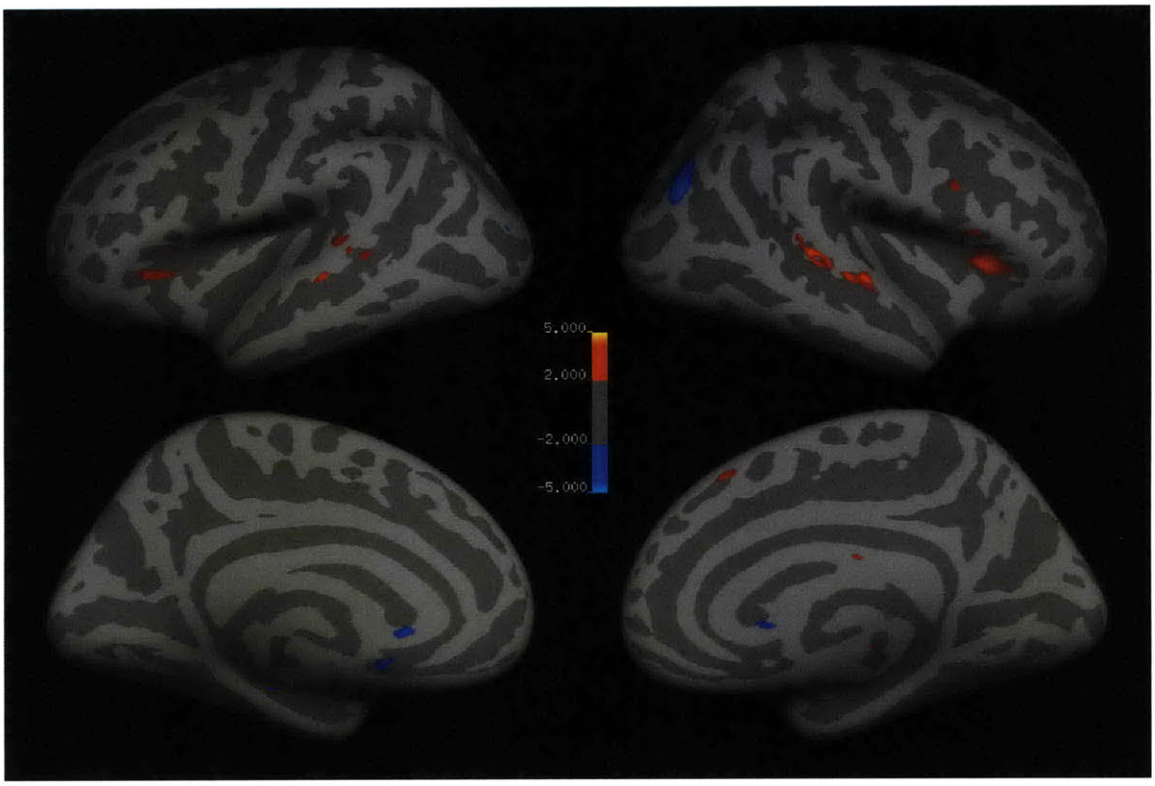

Figure 4-20. Within–NoShift surface view. (See Fig. 4-18 for figure details.)

Across–NoShift surface view

left hemisphere *right hemisphere*

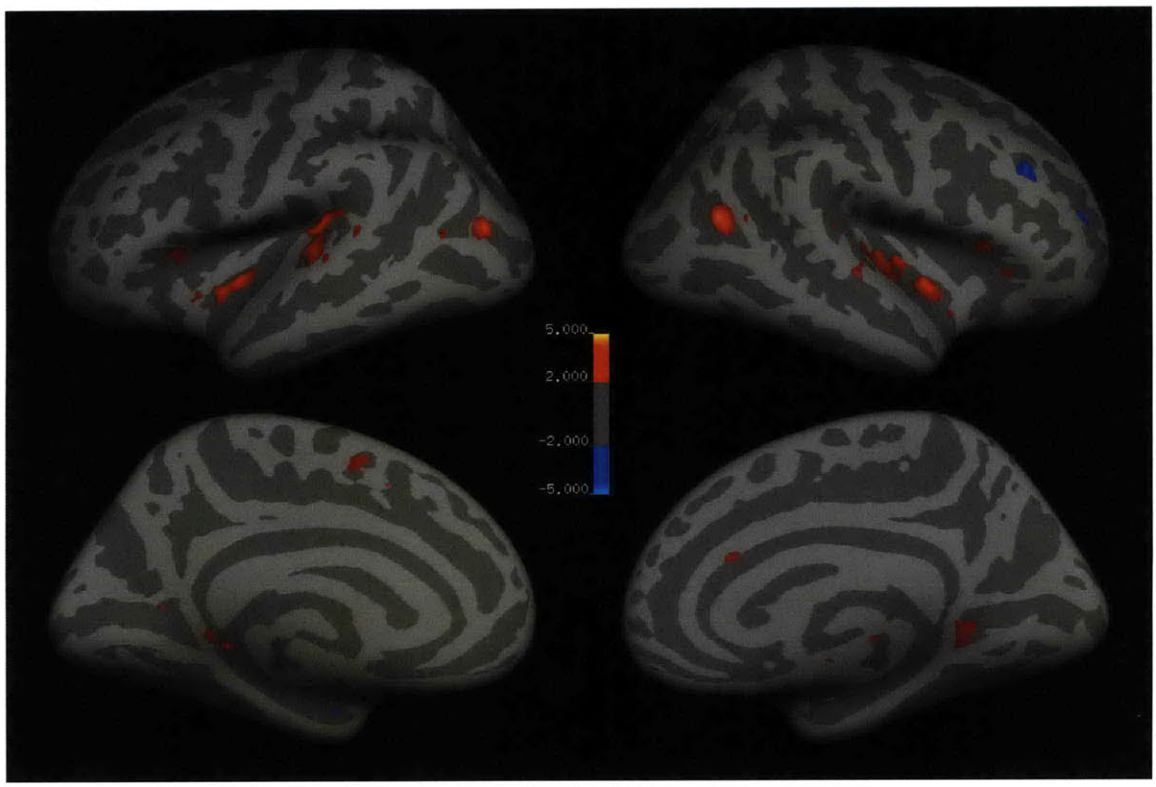

Figure 4-21. Across–NoShift surface view. (See Fig. 4-18 for figure details.)

Across–Within surface view

left hemisphere *right hemisphere*

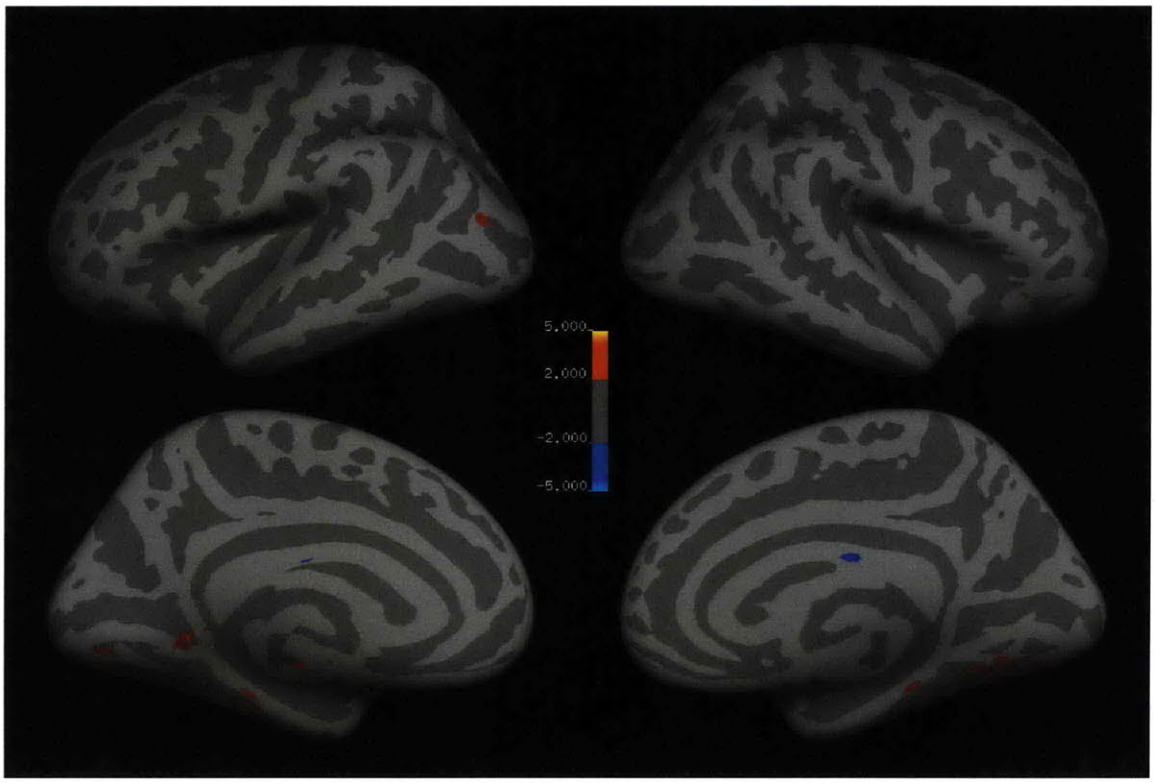

Figure 4-22. Across–Within surface view. (See Fig. 4-18 for figure details.)

Speech–Base slice view, coronal

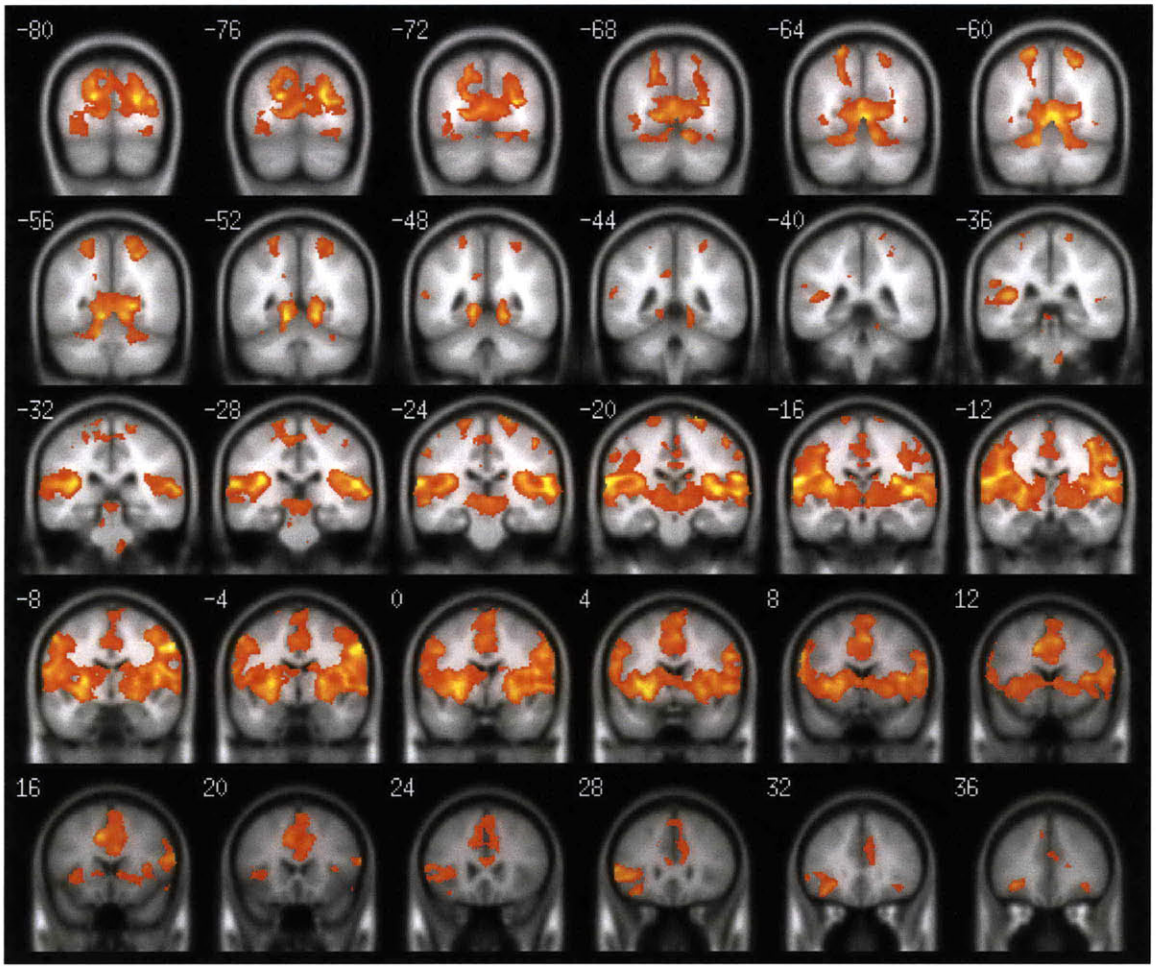

Figure 4-23. Speech–Base coronal slices, $p < 0.01$ (uncorrected). Mixed-effects analysis of voxel-based smoothing, with coronal slices through the brain (numbers are *y*-coordinates in MNI space).

Speech–Base slice view, transverse

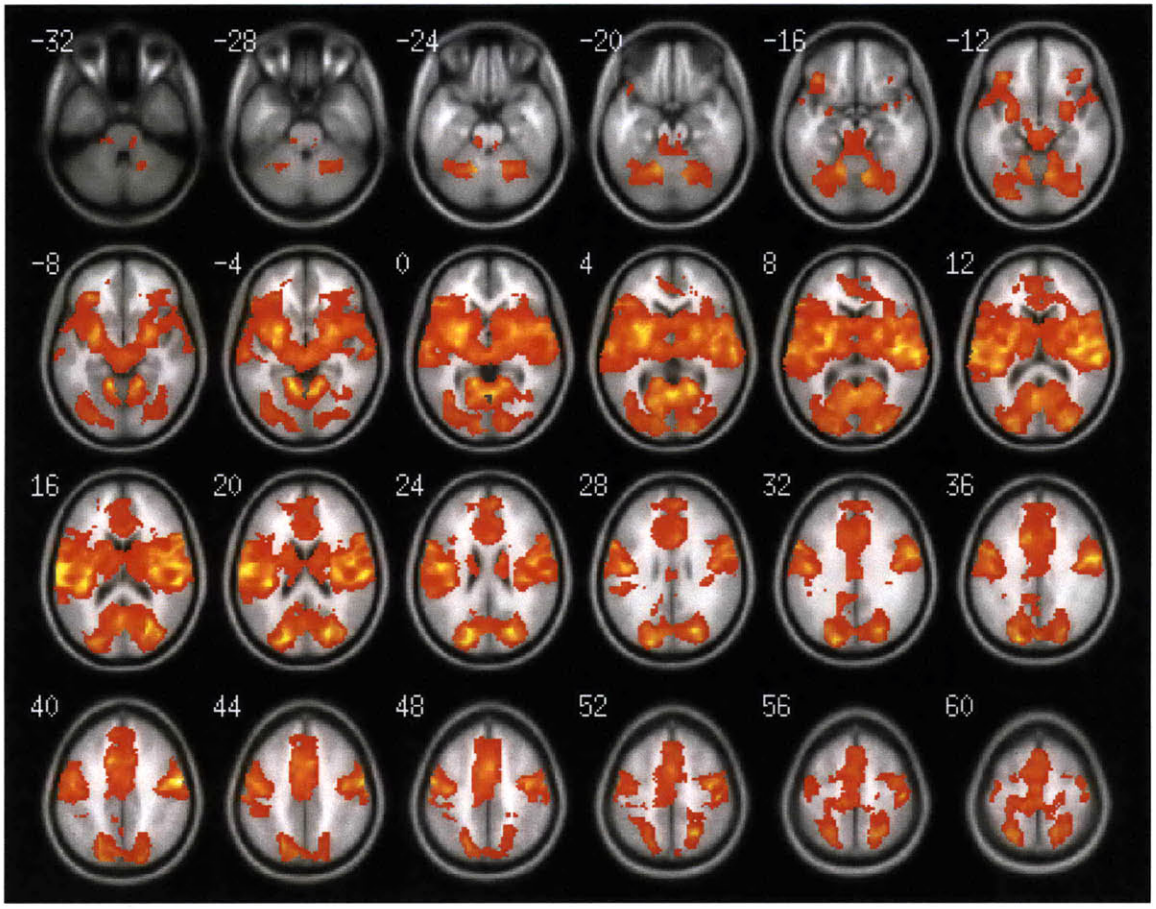

Figure 4-24. Speech–Base transverse slices, $p < 0.01$ (uncorrected). Mixed-effects analysis of voxel-based smoothing, with coronal slices through the brain (numbers are z-coordinates in MNI space).

	Shift–NoShift			
	Stereotaxic location of peak voxel (x,y,z)			
AAL label	**MNI**	**Talairach**	***T***	**norm. effect**
Frontal Cortex				
Left IFt	(–40, 36, –2)	(–38, 32, 5)	3.18	10.47
Left IFo	(–46, 10, 20)	(–44, 6, 22)	3.17	7.54
Right IFt	(38, 20, 12)	(34, 16, 17)	3.72	6.45
Right IFo	(34, 28, –12)	(31, 25, –4)	3.32	11.63
Right SMA	(–2, 24, 50)	(–3, 16, 51)	3.21	10.92
Parietal Cortex				
Right SC	(48, –24, 32)	(43, –23, 28)	3.52	4.76
Insular Cortex				
Left pINS	(–40, –44, 28)	(–38, –45, 25)	4.17	3.09
Left aINS	(–26, 28, 4)	(–25, 24, 10)	3.78	3.81
Right aINS	(32, 20, 12)	(28, 16, 17)	3.85	7.29
Temporal Cortex				
Left STg	(–56, –32, 8)	(–53, –32, 8)	3.57	11.74
Left MTg	(–40, –56, 20)	(–38, –55, 16)	3.62	13.62
Left ITg	(–54, –48, –24)	(–51, –44, –22)	3.66	3.22
Right STg	(50, –28, 4)	(45, –28, 6)	3.71	5.98
Right MTg	(70, –36, 8)	(64, –38, 9)	3.68	10.42

Table 4-2. Peak voxel responses for the Shift–NoShift contrast. Peak responses were defined as local *t*-statistic maxima ($p < 0.01$, uncorrected) separated by a minimum of 6 mm, with no more than 10 peaks reported for each cluster. Each peak voxel was mapped to a cortical region using the AAL brain atlas and is listed with the *t*-statistic and normalized effect associated with that voxel. Voxel locations are provided in both MNI and Talairach stereotaxic reference frames. See list of abbreviations on page 15.

Shift–NoShift, IFo activation

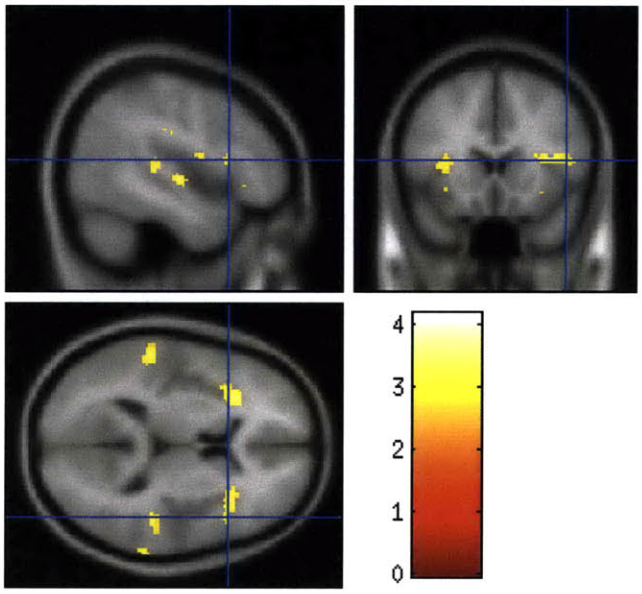

Figure 4-25. Shift–NoShift activation in right inferior frontal gyrus, pars opercularis, MNI coords (46, 20, 10), $p < 0.01$ (uncorrected). Mixed-effects analysis of voxel-based smoothing, shown in sagittal, coronal, and transverse planes (clockwise from upper left). The colorbar tracks the *t*-statistic.

Shift–NoShift, STg activation

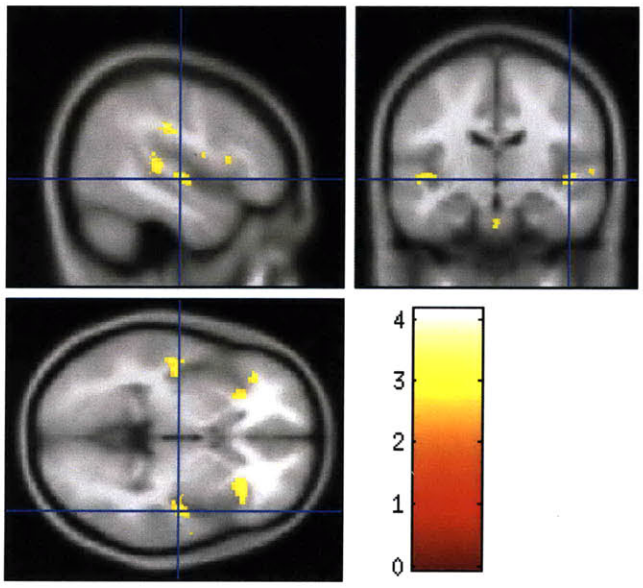

Figure 4-26. Shift–NoShift activation in right superior temporal gyrus, MNI coords (47, –13, –3). (See Fig. 4-25 for figure details.)

Shift–NoShift, SMA activation

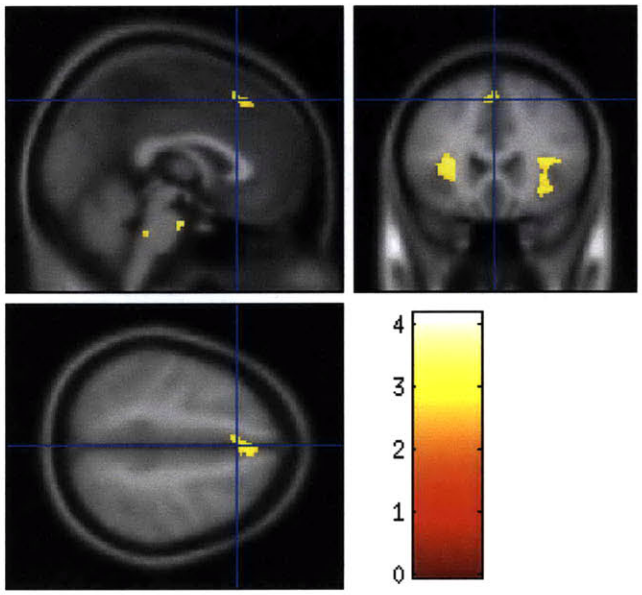

Figure 4-27. Shift–NoShift activation in supplementary motor area, MNI coords (2, 28, 48). (See Fig. 4-25 for figure details.)

Shift–NoShift, SMg activation

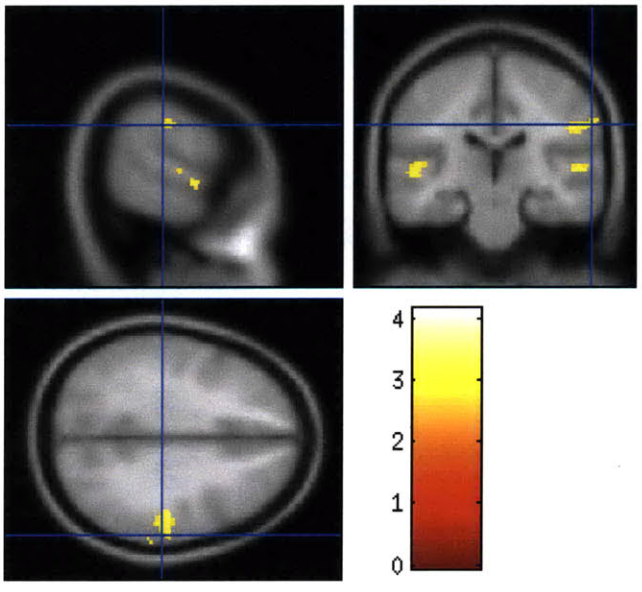

Figure 4-28. Shift–NoShift activation in right supramarginal gyrus, MNI coords (62, –22, 32). (See Fig. 4-25 for figure details.)

Shift–NoShift slice view, coronal

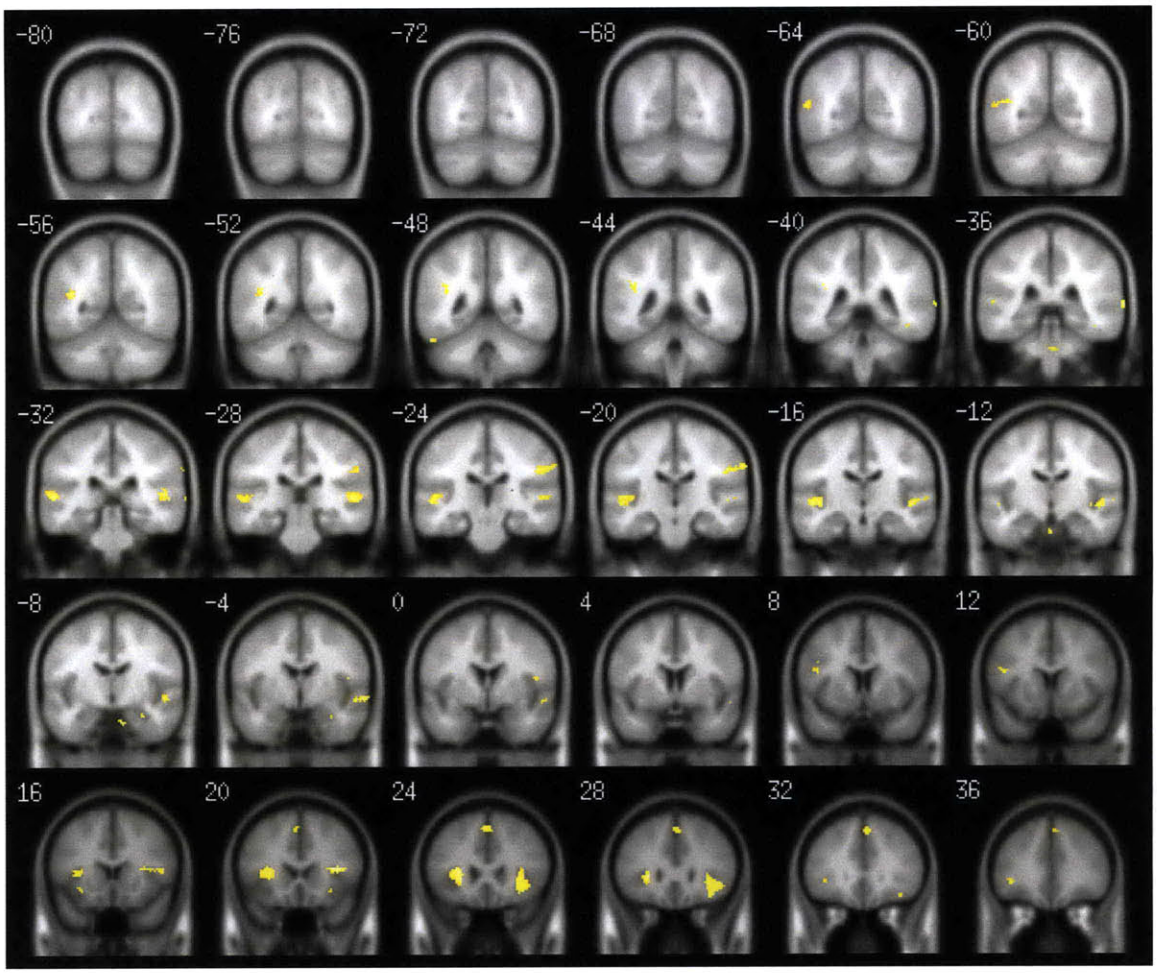

Figure 4-29. Shift–NoShift coronal slices. (See Fig. 4-23 for figure details.)

Shift–NoShift slice view, transverse

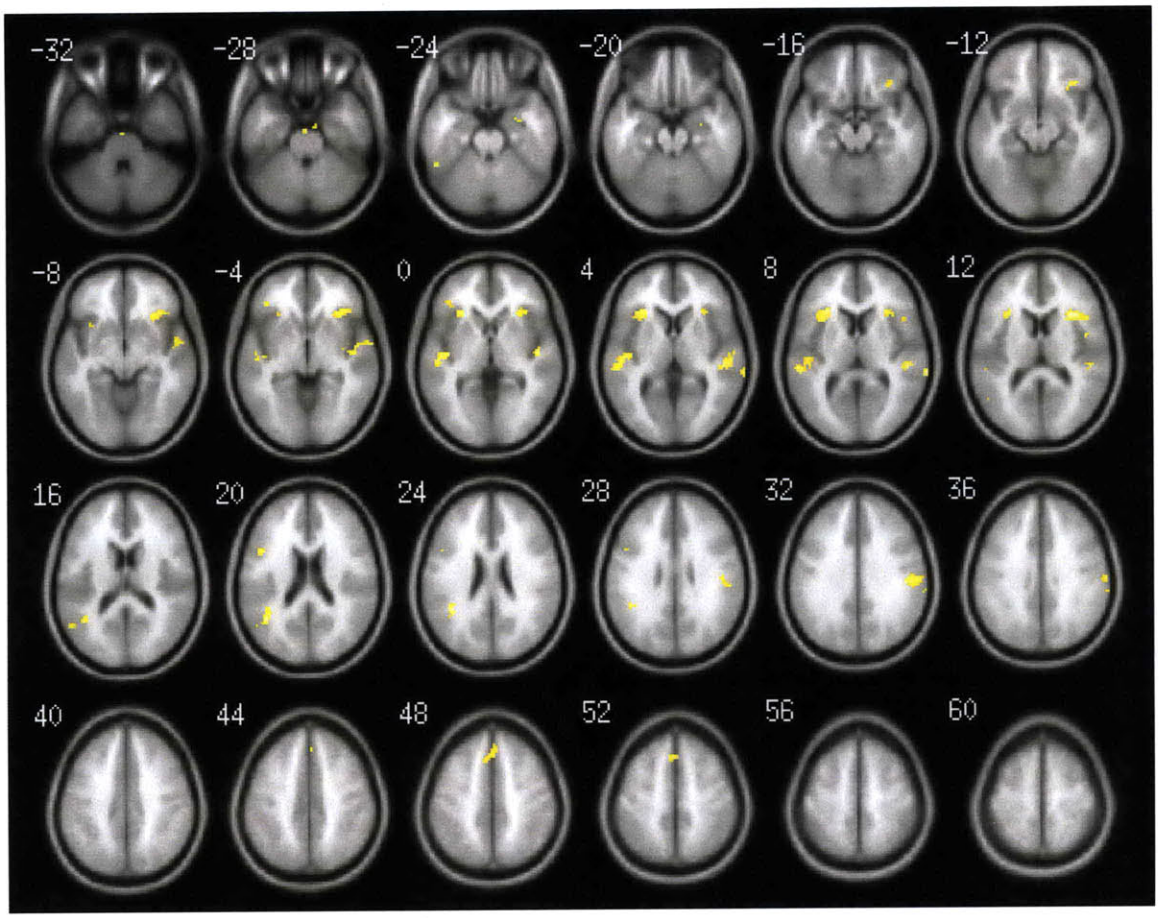

Figure 4-30. Shift–NoShift transverse slices. (See Fig. 4-24 for figure details.)

	Within–NoShift			
	Stereotaxic location of peak voxel (x,y,z)			
AAL label	**MNI**	**Talairach**	***T***	**norm. effect**
Frontal Cortex				
Left IFo	(–38, 18, 10)	(–36, 14, 14)	4.04	8.86
Right IFt	(42, 20, 10)	(38, 16, 15)	3.18	10.06
Right SMA	(8, 14, 50)	(6, 7, 50)	3.20	7.23
Parietal Cortex				
Right SMg	(66, –20, 36)	(60, –24, 35)	4.17	9.09
Insular Cortex				
Right INS	(32, 24, 0)	(29, 20, 7)	3.10	11.68
Temporal Cortex				
Left STg	(–44, –18, 0)	(–42, –18, 2)	4.61	8.62

Table 4-3. Peak voxel responses for the Within–NoShift contrast. As in Table 4-2, peak responses were defined as local t-statistic maxima ($p < 0.01$, uncorrected) separated by a minimum of 6 mm. Each peak voxel was mapped to a cortical region using the AAL brain atlas and is listed with the *t*-statistic and normalized effect associated with that voxel. See list of abbreviations on page 15.

Within–NoShift, IFt activation

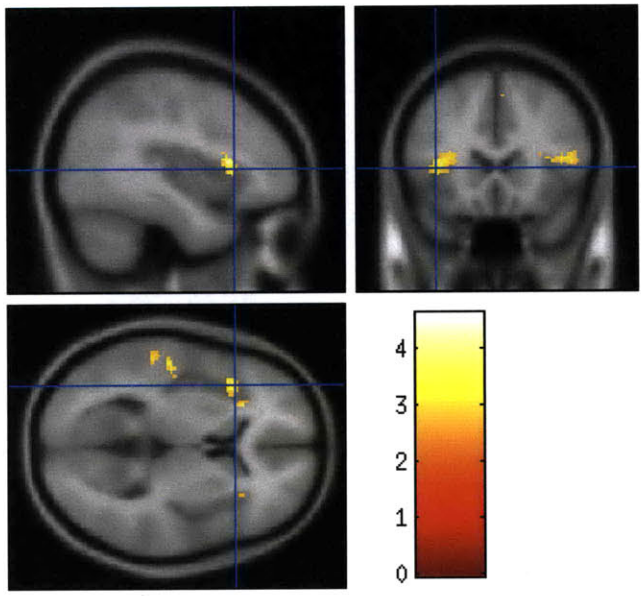

Figure 4-31. Within–NoShift activation in left inferior frontal gyrus, pars triangularis, MNI coords (–37, 19, 6). (See Fig. 4-25 for figure details.)

Within–NoShift, STg activation

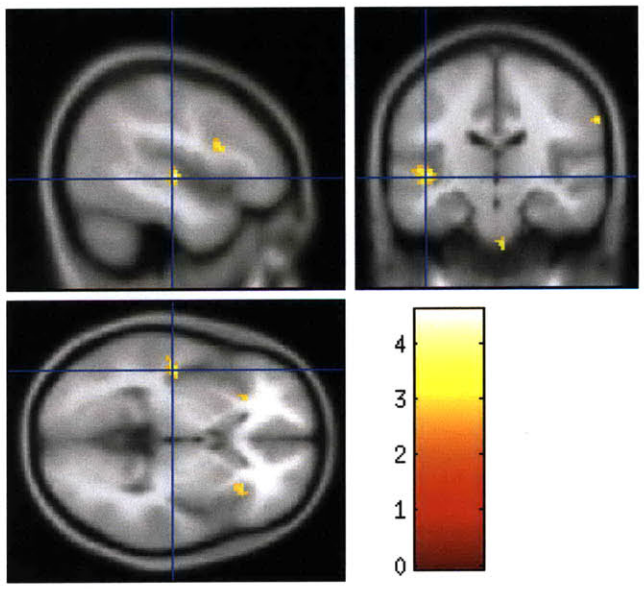

Figure 4-32. Within–NoShift activation in left superior temporal gyrus, MNI coords (–42, –18, 3). (See Fig. 4-25 for figure details.)

Within–NoShift slice view, coronal

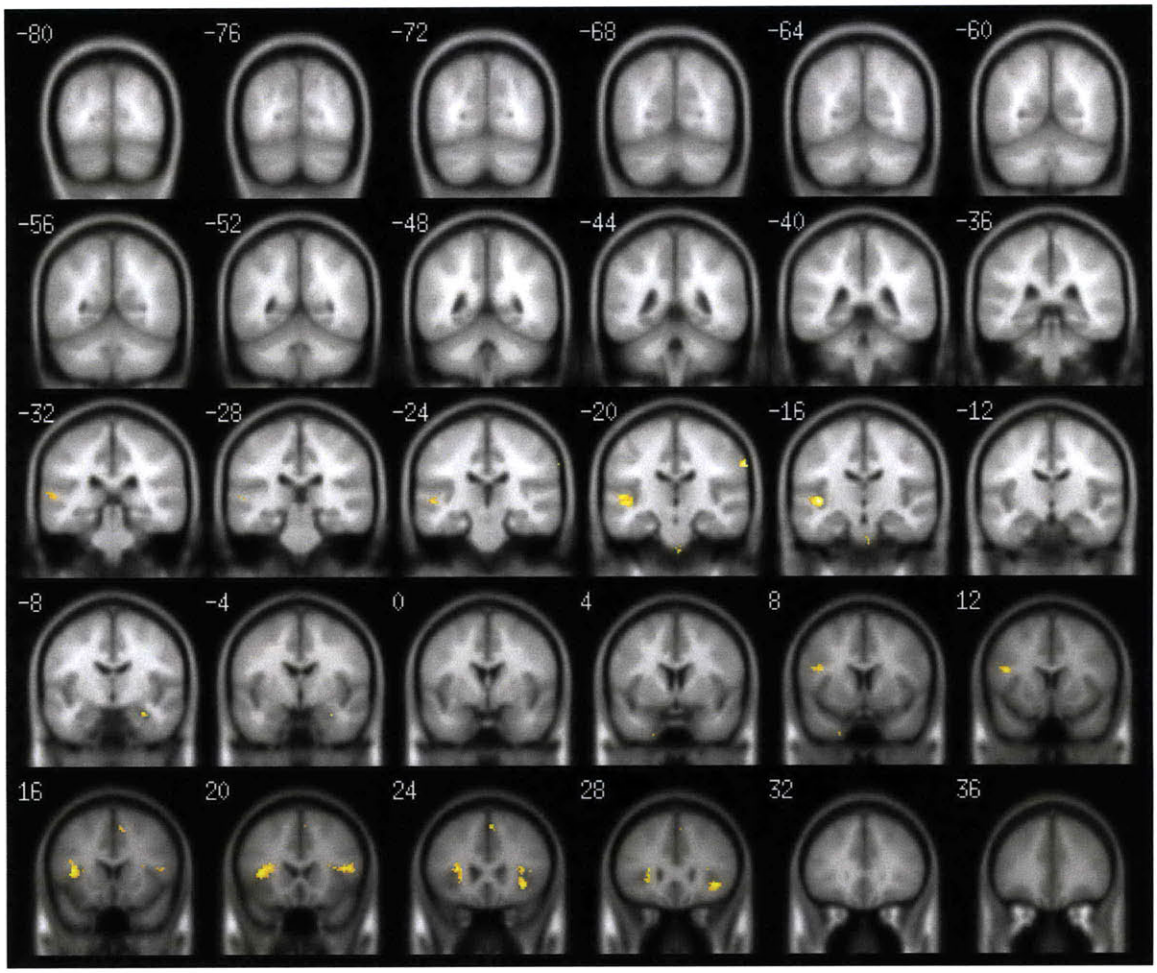

Figure 4-33. Within–NoShift coronal slices. (See Fig. 4-23 for figure details.)

Within–NoShift slice view, transverse

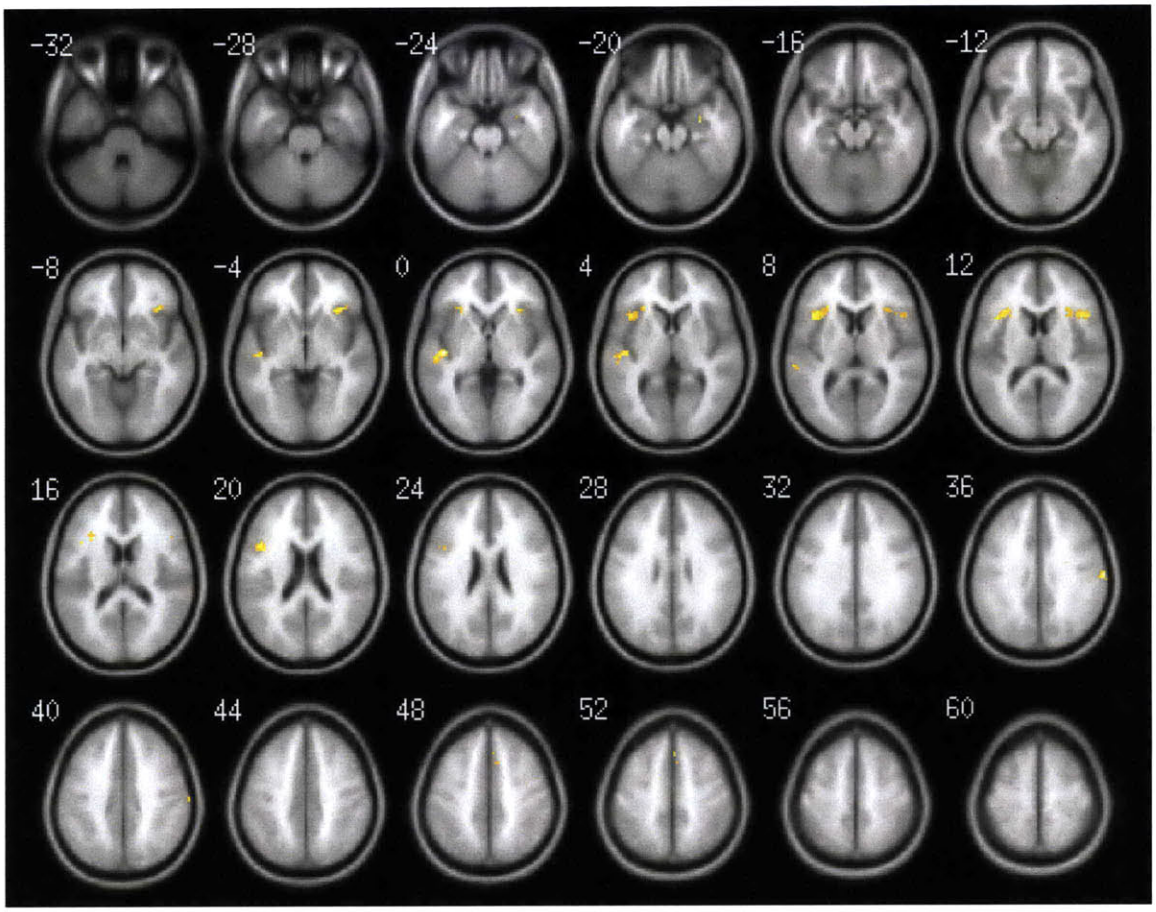

Figure 4-34. Within–NoShift transverse slices. (See Fig. 4-24 for figure details.)

	Across–NoShift			
	Stereotaxic location of peak voxel (x,y,z)			
AAL label	**MNI**	**Talairach**	***T***	**norm. effect**
Rolandic Cortex				
Right vMC	(54, 10, 46)	(48, 3, 47)	3.68	15.17
Left vPMC	(–52, 4, 50)	(–50, –2, 48)	3.62	9.75
Frontal Cortex				
Left IFt	(–52, 38, 8)	(–49, 33, 14)	3.54	16.61
Left IFo	(–46, 10, 30)	(–44, 5, 31)	2.72	13.32
Parietal Cortex				
Right PCN	(18, –48, 0)	(16, –47, 0)	3.62	8.26
Insular Cortex				
Left INS	(–40, –6, –2)	(–38, –7, 1)	2.90	9.14
Right INS	(34, 18, –10)	(31, 16, –3)	3.45	12.62
Temporal Cortex				
Left STg	(–56, –32, 8)	(–53, –32, 8)	3.46	13.12
Left MTg	(–50, –64, 18)	(–48, –63, 14)	3.88	18.64
Left ITg	(–54, –46, –24)	(–51, –42, –22)	3.72	5.20
Right STg	(50, –28, 8)	(45, –29, 10)	3.89	9.01
Right MTg	(40, –52, 20)	(36, –52, 18)	3.64	5.94
Right ITg	(56, –40, –26)	(51, –37, –22)	3.72	6.04
Right HG	(38, –20, 4)	(34, –21, 6)	3.47	8.37
Right AG	(38, –58, 24)	34, –58, 21)	3.00	4.28

Table 4-4. Peak voxel responses for the Across–NoShift contrast. As in Table 4-2, peak responses were defined as local t-statistic maxima ($p < 0.01$, uncorrected) separated by a minimum of 6 mm. Each peak voxel was mapped to a cortical region using the AAL brain atlas and is listed with the *t*-statistic and normalized effect associated with that voxel. See list of abbreviations on page 15.

Across–NoShift, IFt activation

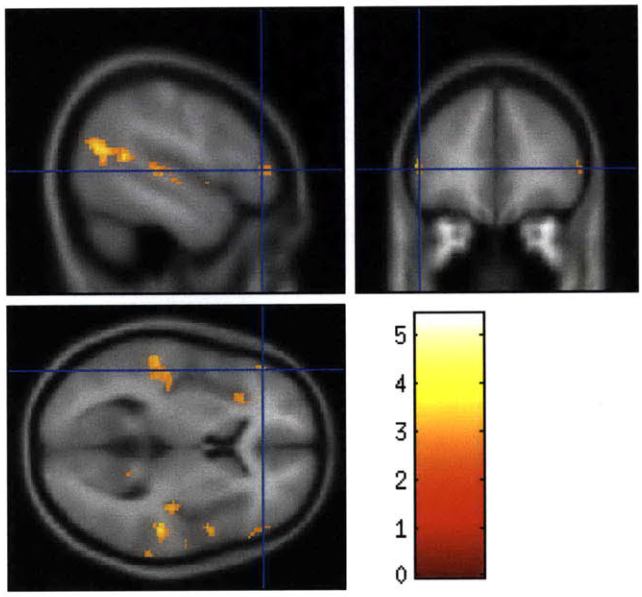

Figure 4-35. Across–NoShift activation in left inferior frontal gyrus, pars triangularis, MNI coords (–48, –39, 4). (See Fig. 4-25 for figure details.)

Across–NoShift, STg activation

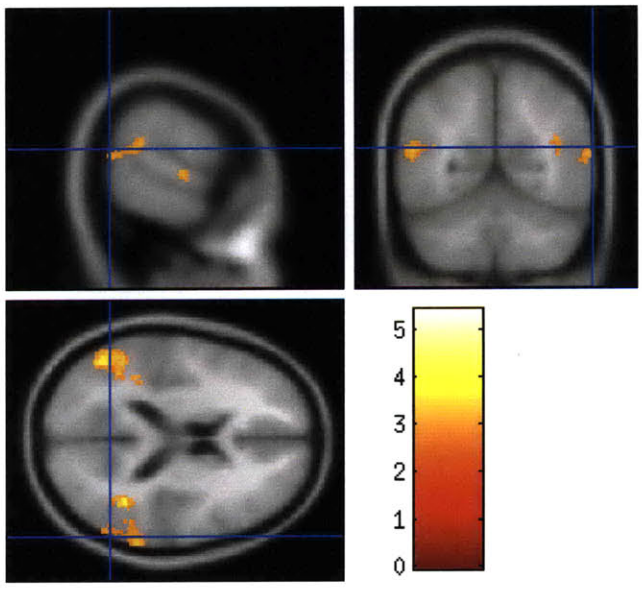

Figure 4-36. Across–NoShift activation in right superior temporal gyrus, MNI coords (66, –42, 18). (See Fig. 4-25 for figure details.)

Across–NoShift slice view, coronal

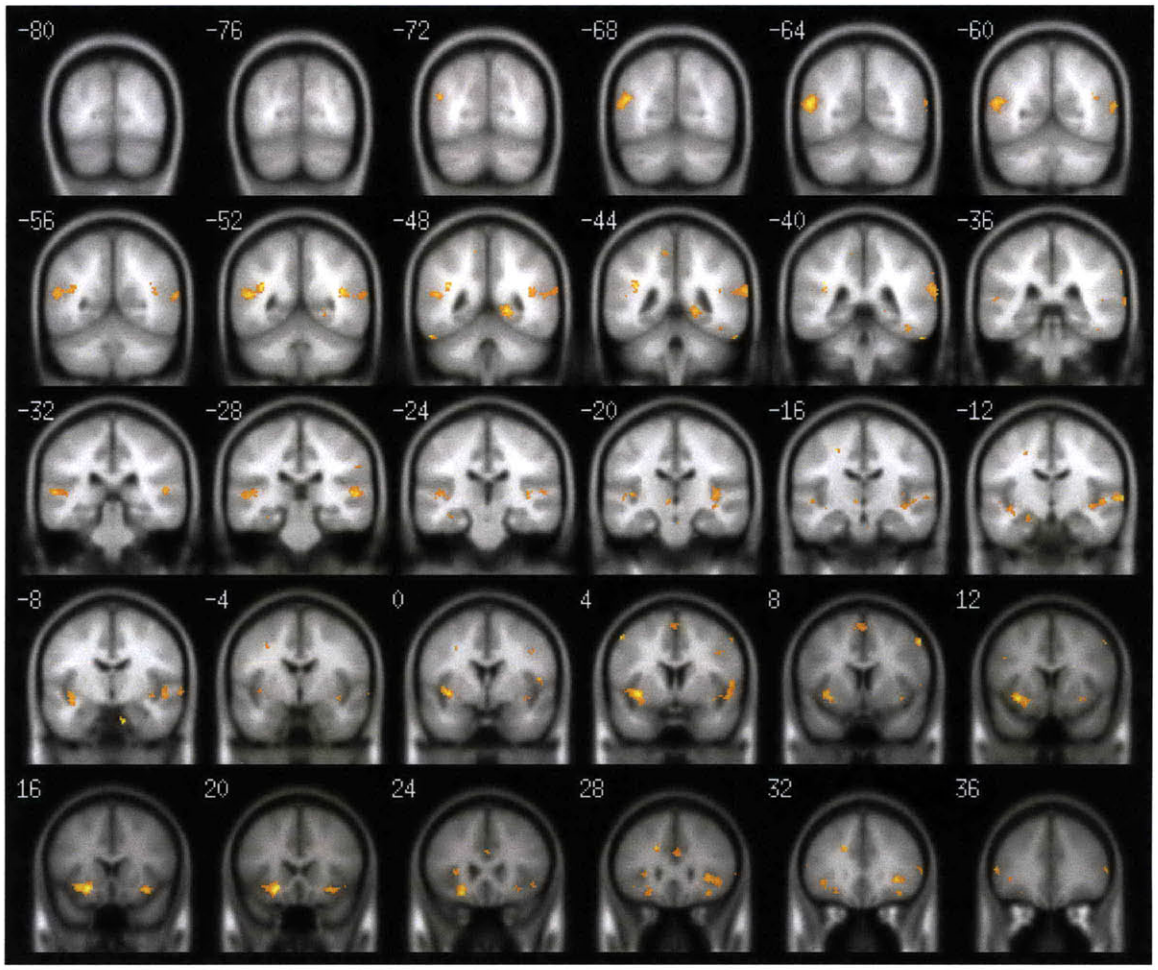

Figure 4-37. Across–NoShift coronal slices. (See Fig. 4-23 for figure details.)

Across–NoShift slice view, transverse

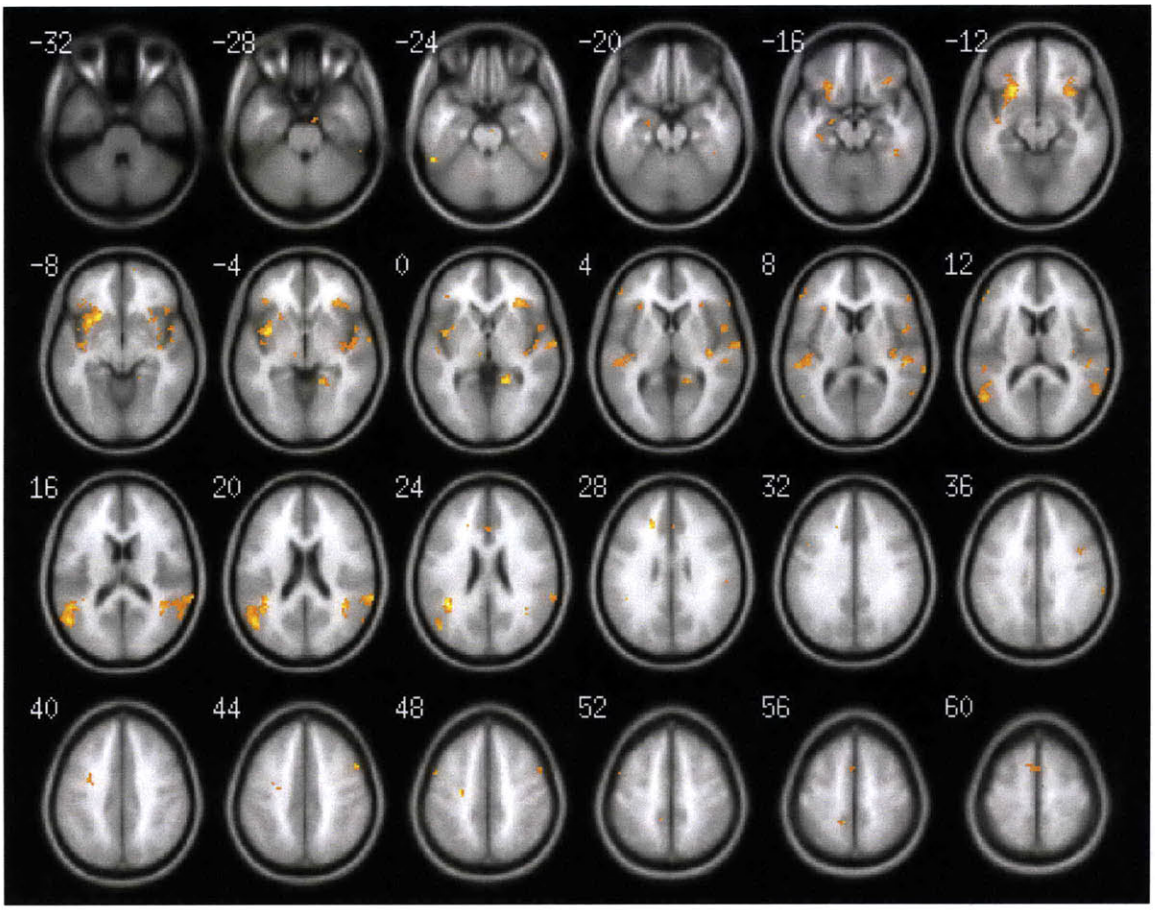

Figure 4-38. Across–NoShift transverse slices. (See Fig. 4-24 for figure details.)

Across–Within, IFo activation

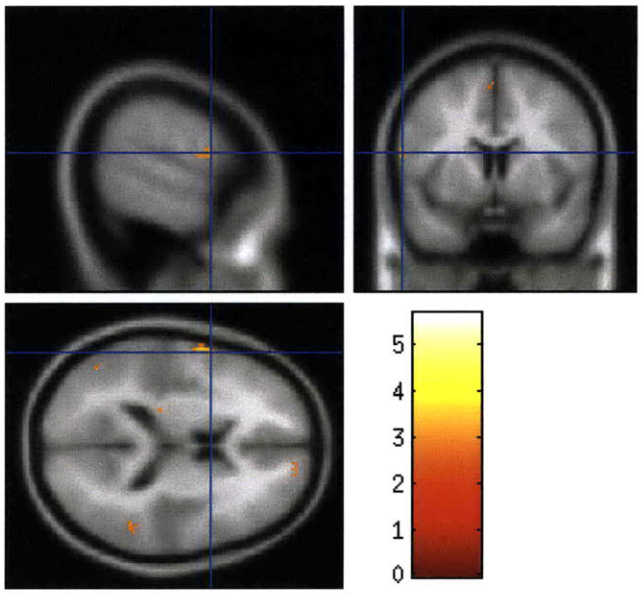

Figure 4-39. Across–Within activation in left inferior frontal gyrus, pars opercularis, MNI coords (–58, 6, 16). (See Fig. 4-25 for figure details.)

Across–Within, STg activation

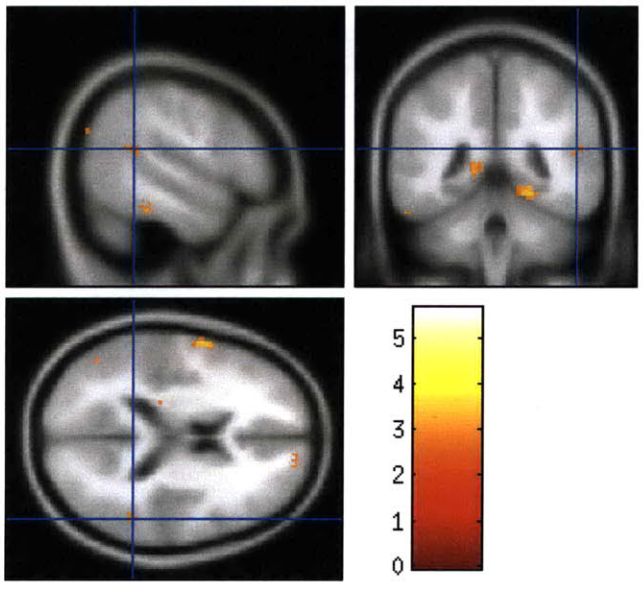

Figure 4-40. Across–Within activation in right superior temporal gyrus, MNI coords (54, –44, 18). (See Fig. 4-25 for figure details.)

Across–Within slice view, coronal

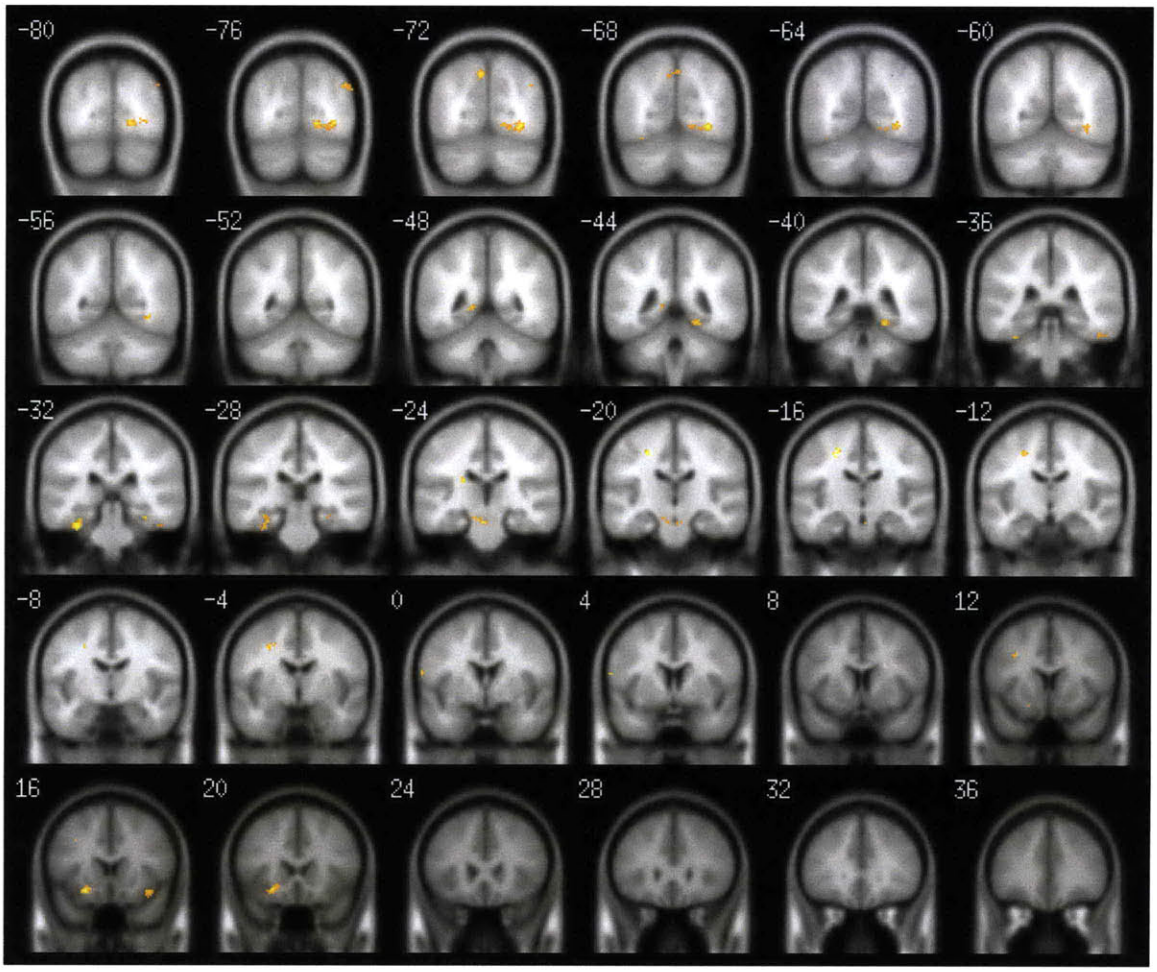

Figure 4-41. Across–Within coronal slices. (See Fig. 4-23 for figure details.)

Across–Within slice view, transverse

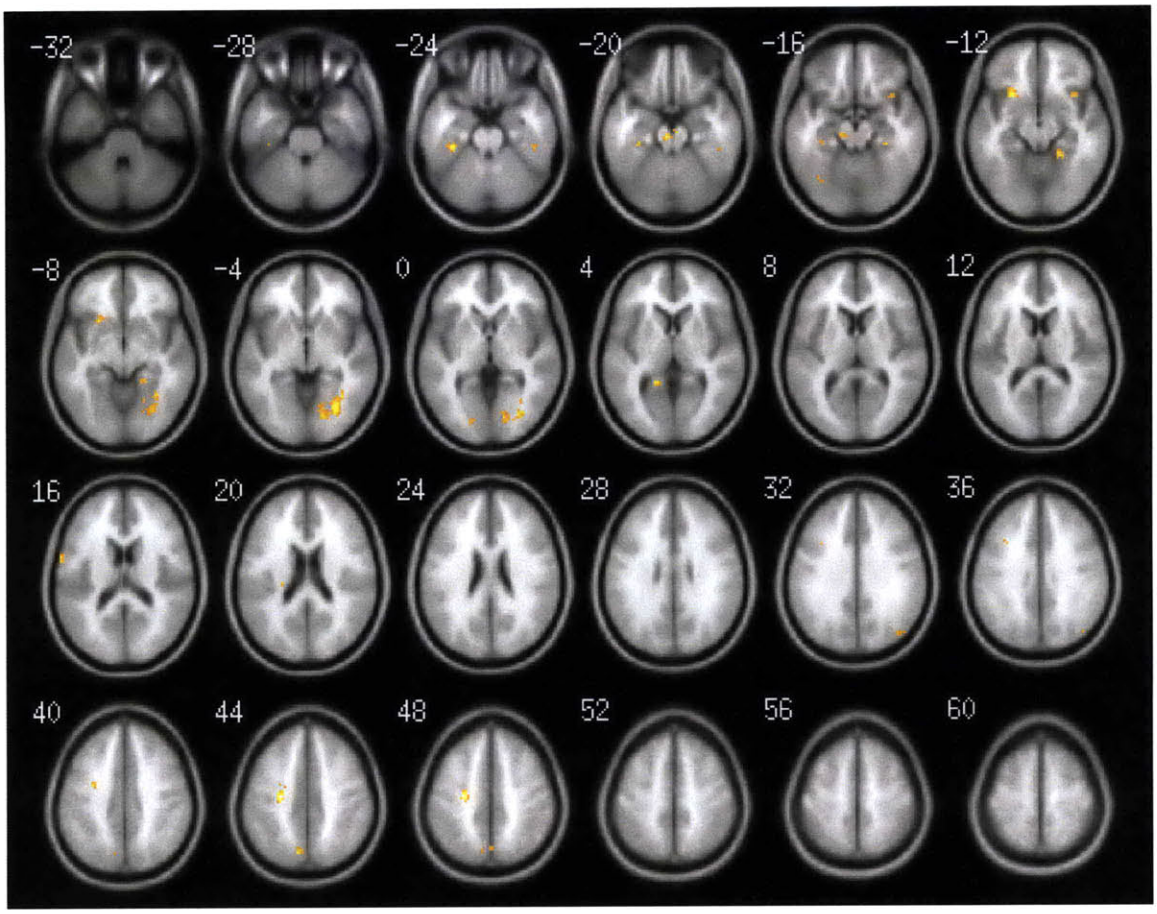

Figure 4-42. Across–Within transverse slices. (See Fig. 4-24 for figure details.)

As expected, shifted conditions showed more cortical activation in superior temporal gyrus and right inferior frontal gyrus than unshifted conditions. Results also showed increased activation of bilateral inferior frontal gyrus, superior temporal gyrus, and supplementary motor areas for across-category shifts compared with within-category shifts. In general, cortical activation was greater in extent for shifts that crossed a category boundary than for those that did not, even though these shifts were of the same magnitude.

ROI analysis. An overview of significantly active regions as determined by ROI analysis is as follows:

- Shift–NoShift: bilateral pSTg, bilateral PT, bilateral Hg, bilateral aINS, bilateral pINS, bilateral FO, right pdSTs, bilateral IFo, right IFt.
- Within–NoShift: left PT, bilateral FO, right pdSTs, left IFo, right IFt.
- Across–NoShift: bilateral pSTg, bilateral PT, bilateral Hg, bilateral PP, bilateral aINS, bilateral pINS, bilateral FO, right pdSTs, right aSTg, right aCO, right LG, left PO, left aSMA, bilateral LG.
- Across–Within: Of the regions found significant in the previous contrasts, only right PP and bilateral LG survived a direct Across–Within contrast.

Schematics of significant regions of activations are presented in Figures 4-43 to 4-45. In summary, a greater extent of neural activation is seen for the Across–NoShift contrast than for the Within–NoShift contrast. Interestingly, the Within–NoShift contrast seemed to contribute more to the inferior frontal activation seen in the Shift–NoShift contrast, while the Across–NoShift contrast contributed activation in the superior temporal areas and the intra-Sylvian region. The differences were subtle enough that most auditory cortical areas and inferior frontal areas did not survive a direct Across–Within contrast (with the exception of right PP).

Shift–NoShift ROI analysis

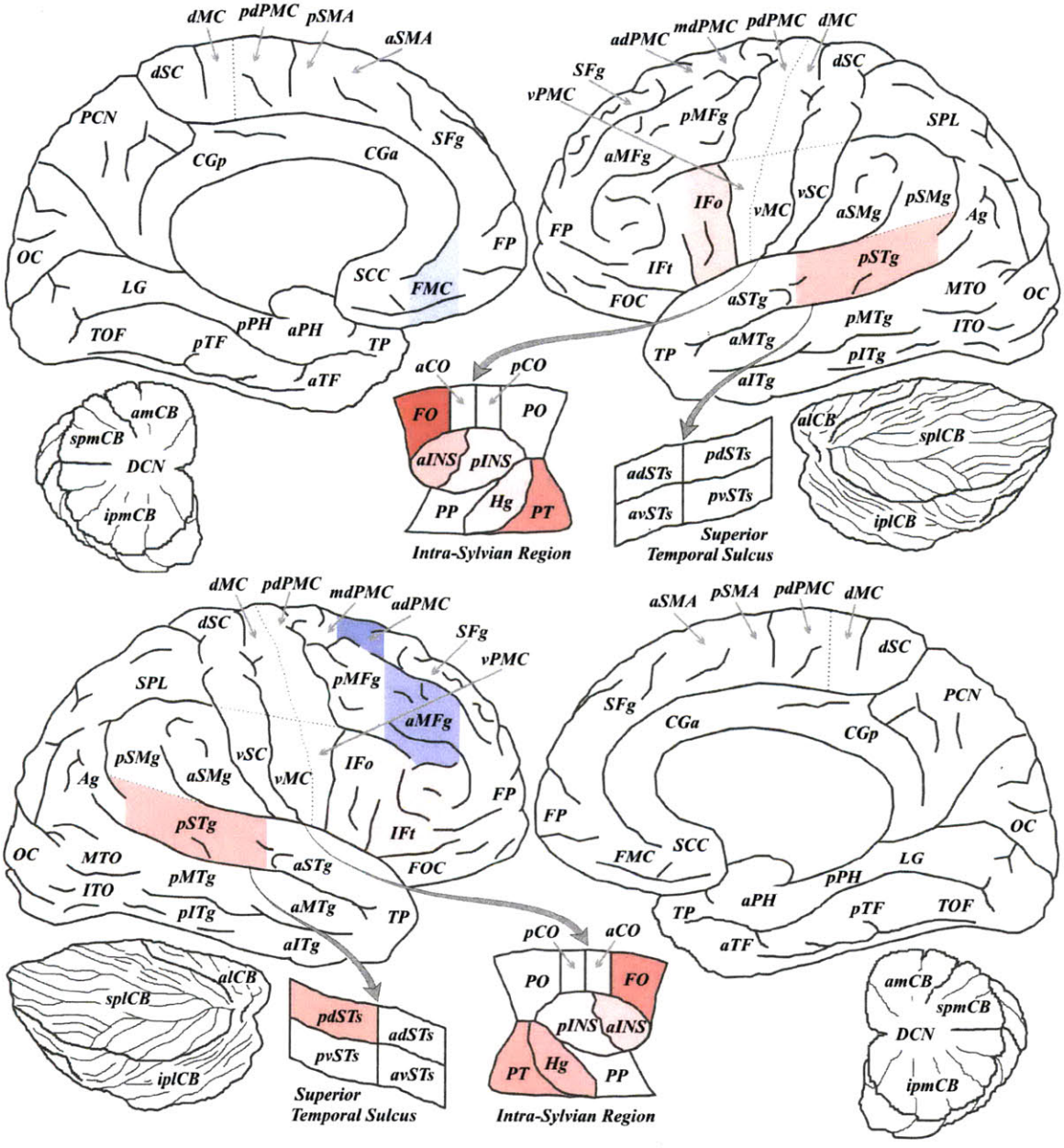

Figure 4-43. Shift–NoShift ROI analysis. Cortical regions are shaded where the *t*-statistic exceeds 1.4. Red represents regions in which Shift > NoShift, and blue represents the opposite.

Within–NoShift ROI analysis

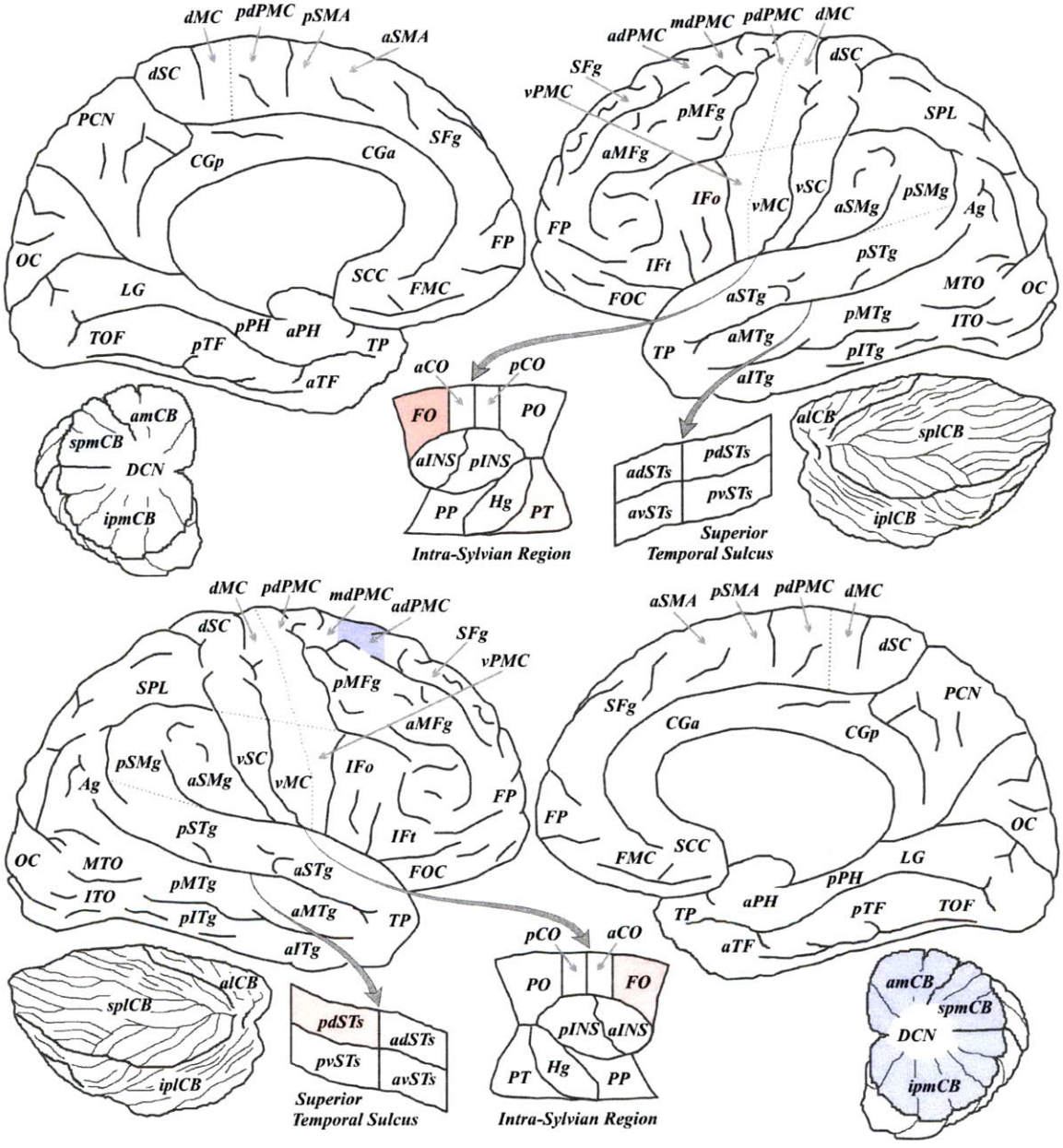

Figure 4-44. Within–NoShift ROI analysis. (See Fig. 4-43 for figure details.) Notably, differences in activation were not significant in pSTg, despite a reliable effect in pSTg in the Shift–NoShift. condition.

Across–NoShift ROI analysis

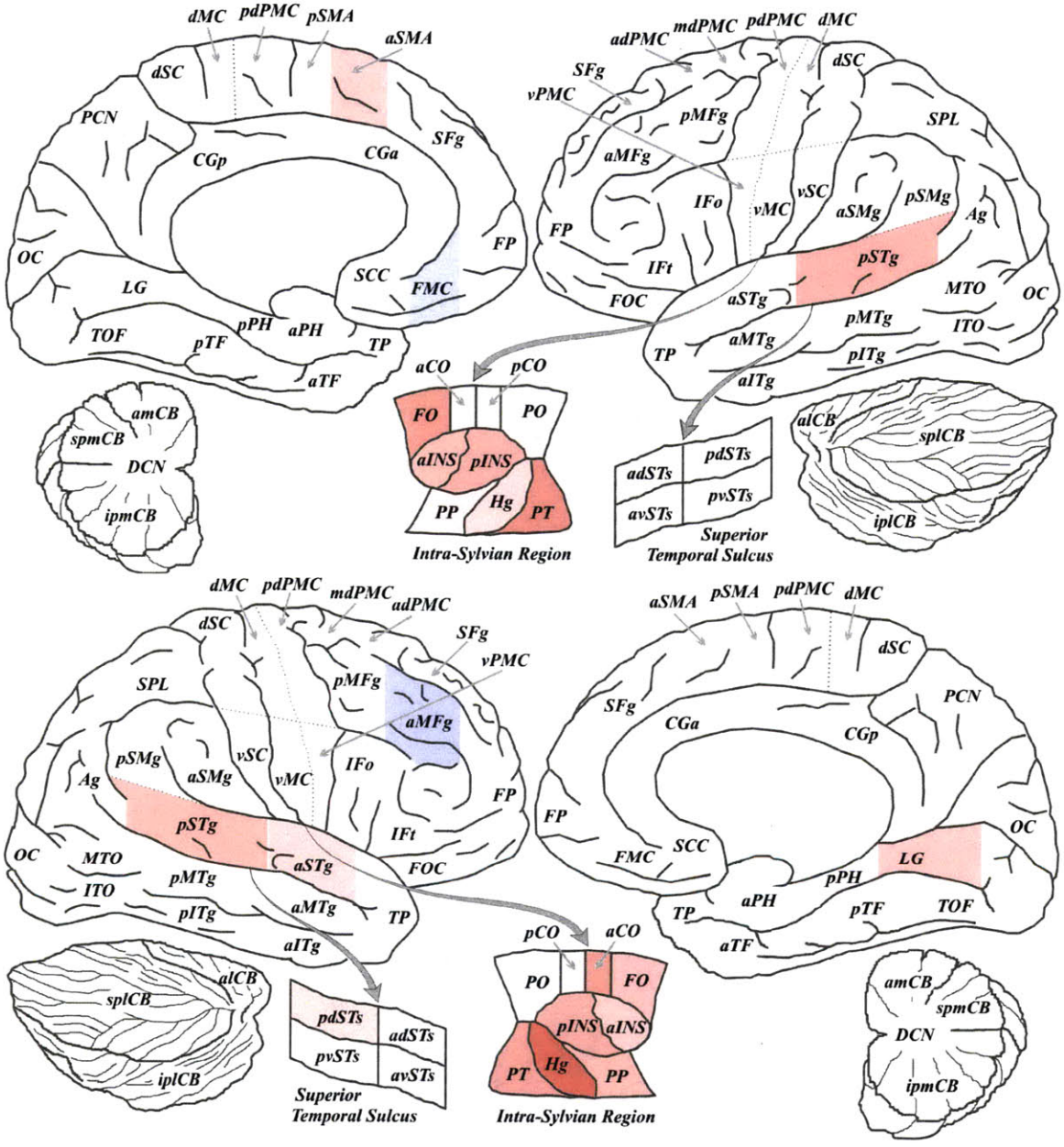

Figure 4-45. Across–NoShift ROI analysis. (See Fig. 4-43 for figure details.) Almost all posterior auditory cortical areas show significant activation, including pSTg and PT, as well as lower-order auditory cortex (Hg).

4.3.3 Correlations between functional imaging and behavior

The goal of the regression analysis was to identify cortical regions whose activity correlates with a compensatory response. Because compensatory "performance" varied across subjects, the Shift–NoShift signal may have been decreased for subjects who exhibited little compensation and who did not perceive the perturbations. Using degree of compensation as a regressor may help to refine the cortical locations that are specifically recruited in the feedback pathway.

The *t*-contrast activation maps from each single-subject analysis were used in a simple regression analysis with compensation as a covariate measure. The compensation was defined as the mean projection over all trials. When correlating with Within–NoShift and Across–NoShift activation maps, the mean projection for only the Within or the Across trials (respectively) was used as a covariate. When correlating with the Shift–NoShift and Across–Within activation maps, mean projection for Within and Across trials was averaged. The F-contrast map shows the regions that have a statistically significant correlation with behavioral measures at the $p < 0.001$ level, uncorrected (Fig. 4-45). This correlation analysis corroborates the results of the mean activation analysis in localizing activation to IFg and STg.

Brain-behavior correlation

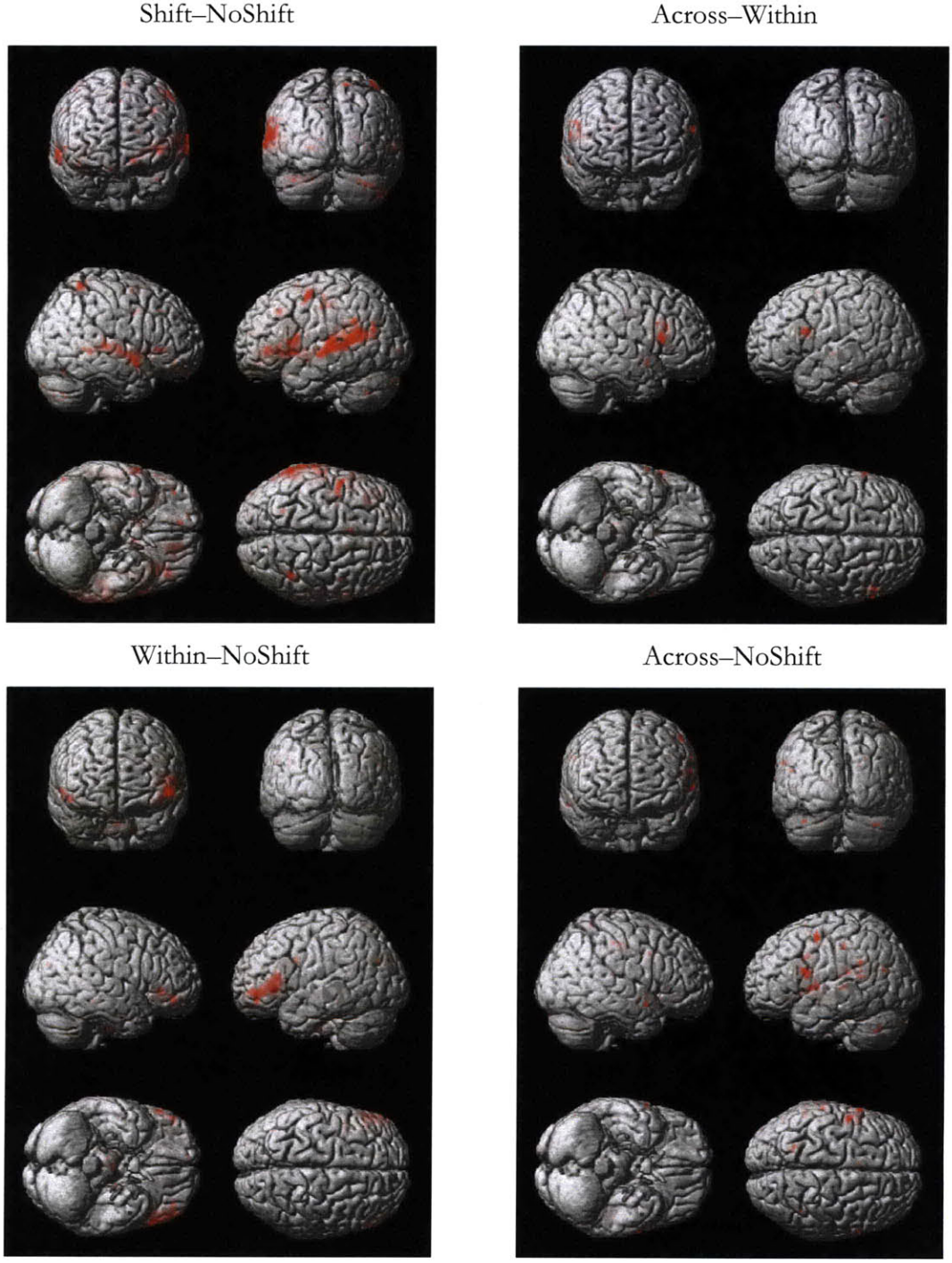

Figure 4-46. Regression analysis with compensation as a covariate, $p < 0.001$ (uncorrected). For each condition, top row: coronal view (L: anterior view, R: posterior view); middle row: sagittal view (L: left view, R: right view); bottom row: axial view (L: inferior view, R: superior view).

4.4 Discussion

Real-time formant perturbation acts as a controlled form of an articulatory error in natural speech, shifting the produced vowel away from its intended target. The auditory perturbation evokes a commensurate auditory error, triggering a compensatory corrective movement. In the current study, formant perturbation was found to activate cortical regions underlying the detection and correction of this error. Furthermore, perturbations that caused a phonetic category boundary to be crossed were found to evoke a greater behavioral and neural response than perturbations of the same magnitude that caused only a within-category shift in acoustics.

4.4.1 Compensatory responses to formant shifts

As illustrated in Figures 4-12 to 4-15, the perceived formant perturbation induced a compensatory response: a deviation from the baseline formant trajectory in opposition to the shift. Speakers altered the formants of their speech to oppose the perturbation within the first 150 ms of voicing onset. This rapid compensation is similar to that demonstrated previously (Tourville et al., 2008) and showcases the on-line vocal control mediated by the degree of mismatch between vowel target and realization. A similar response has also been reported for unexpected perturbation of F0 (Xu et al., 2004; latencies of 100-150 ms reported). As anticipated, subjects compensated for the formant perturbations within the timeframe of the utterance, despite the unpredictability of the perturbation.

Furthermore, the average magnitude of the compensatory response was greater in the Across condition than in the Within condition, confirming that the two types of perturbation evoke differential responses. In the Across condition, the formant shifts were designed to change the perceived phonetic category for each subject; given the increase in discriminability for sounds from different phonetic categories, this change was predicted to be more salient than a non-phonetic change (see Chapter 3 for background). As hypothesized, the deviation from the baseline formants was significantly greater for a cross-category shift, though both types of shifts elicited clear compensatory responses.

A similar increase in response magnitude for a linguistic contrast was reported in a study of unexpected F0 perturbation (Chen et al., 2007). The experiment involved upward or downward perturbations to the spoken pitch contour as speakers produced the phrase "you know Nina?" The audio sample imitated by the participants had a rising F0 contour at the end of the phrase, marking it as a question. Like the cross-category formant shifts, the downward perturbations in F0 had linguistic relevance, as they flattened the prosodic contour of the sentence, causing it to sound less like a question. In contrast, the upward F0 perturbations were not at odds with the target intonation contour. The authors report that compensatory response magnitudes were greater when the F0 shift was effected in the linguistically-relevant (downward) direction, suggesting that the need for a corrective response was greater than in the upward direction.

In line with this idea, Tourville and colleagues (2008) also reported a shorter latency to downward F1 shifts (shifting from /ɛ/ towards /ɪ/) than to upward F1 shifts (shifting from /ɛ/ towards/æ/). The upward and downward shift magnitudes were identical on the Hz scale (30% of the produced F1 value in Hz); when translated to a logarithmic perceptual scale, such as the mel scale used in the current study, the downward shift corresponded to a greater magnitude than the upward shift. Tourville and colleagues point out that that the downward F1 shift toward /ɪ/ was more likely to produce a phonemic or lexical categorical error than the upward shift toward /æ/, and suggest that the faster response to the downward shift may therefore reflect greater phonemic saliency. This interpretation is also supported by an experiment employing unexpected perturbations in F0 to Mandarin bi-tonal disyllables (Xu et al., 2004). Shifts in the direction opposite the intended inter-syllabic tonal transition resulted in shorter latencies and larger compensations than did shifts in the same direction.

As in previous studies of auditory perturbation, the response magnitudes did not achieve parity with the shift magnitude. Compensation, as defined by the projection (see Fig. 4-6), averaged less than 10% of the shift magnitude (see Fig. 4-13). The partial compensation can be explained by several factors. First, because the induced error was artificial and purely auditory, the unaltered somatosensory feedback would not indicate a mismatch; a large compensation would necessitate a change in somatosensory output and

would be hypothesized to produce a commensurate somatosensory error. However, with repeated exposure to perturbation, a new somatosensory mapping can be learned, allowing for greater compensation. This hypothesis is supported by a comparison with sensorimotor adaptation studies, which achieve a larger response magnitude when using a sustained perturbation (Purcell and Munhall, 2006). Secondly, a low gain prevents unstable behavior, such as oscillations, in feedback control systems with significant delays. Thirdly, by responding only partially to auditory feedback perturbations, the system allows for feedforward mechanisms to continue to mediate vocal control.

4.4.2 Brain regions implicated in feedback control

Posterior superior temporal gyrus and intra-Sylvian regions. Superior temporal gyrus contains both primary auditory and auditory association cortex (Rivier and Clarke, 1997), as well as areas associated with phonetic processing in the more posterior region of left STg (Scott et al., 2000). The activation found in this region replicates that found by Tourville (2008) in a similar fMRI study of unexpected formant perturbation, and by several other studies involving pitch shifts (Zarate and Zatorre, 2005) and auditory feedback delay (Hirano et al., 1997; Hashimoto and Sakai, 2003). Taken together, these results are strong evidence for the existence for an auditory comparison of speech targets and observed feedback in the posterior temporal regions.

Inferior frontal gyrus and anterior insula. The posterior regions of the inferior frontal gyrus, including the posterior portion of Broca's area, the frontal operculum, and adjacent anterior insula, are part of the "minimal speech network" described by Bohland and Guenther (2006). In the DIVA model (Guenther et al., 2006), the left IFg is hypothesized to contain the speech sound map responsible for the generation of feedforward motor commands. This region has been described by Indefrey and Levelt (2004) as an interface between phonological encoding and articulation.

Dronkers (1996) identified the anterior insula as a region subserving articulatory functions of speech motor control, with left anterior insular lesions associated with apraxia of speech. In a passive listening study (Mutschler et al., 2007), fMRI responses in left

anterior insula were increased for actively-learned melodies, practiced on the piano with the right hand, as compared with passively-learned melodies, familiarized through passive listening. The authors conclude that insular cortex may play a role in action-perception associations, specifically short-term auditory-motor learning. Extended to a speech context, an insula might play a role in short-term feedback-based corrective articulatory movements.

Motor areas. No motor, pre-motor, or cerebellar activation was found in response to perturbed conditions, in contrast with the areas of activation described by Tourville and colleages (2008). However, anterior supplementary motor areas were found to be active in the Across–Within condition. The SMA has been implicated in motor sequencing (Wildgruber et al., 1999) and articulatory planning (Indefrey and Levelt, 2004). It is probable that differences in vowel duration underlie these differences in activation: in the prior study, subjects were directed to vocalize until a visual prompt cued the end of the word, while in the current study, subjects were merely instructed to articulate each word clearly, so that a naïve listener could identify the word being spoken. This instructional modification led to more natural-sounding speech but a much shorter duration of vocalization—less time for a corrective response.

Precuneus. The precuneus was activated only in the Across–Within contrast. Although it is outside of the core speech network, it has been implicated in speech perception and verbal memory. Along with the left angular gyrus, the precuneus was found to be more active in response to normal than to backward speech (Dehaene-Lambertz et al., 2002), suggesting that it maybe sensitive to segmental properties. It has also been implicated in the retrieval of verbal episodic memory (Shallice et al., 1994). The anterior precuneus projects to sensorimotor areas such as the supplementary motor area, premotor cortex, and insula, found to be active in other contrasts such as Shift–NoShift.

Fusiform and lingual gyri. The fusiform and lingual gyri are visual association areas implicated in visual object recognition. The fusiform gyrus is thought to contain a number of category-specific modules, including a word area that responds preferentially to letters and letter strings (Dehaene et al., 2002; Howard et al., 1992; Rumsey et al., 1997). The fusiform gyrus especially has been implicated in face recognition, but has been postulated to underlie

expert discrimination of within-category visual objects other than faces (Price et al., 2003), and thus may be active during a perceptual task involving letterforms.

Phonetic vs. auditory error cells. In the Across condition, the formant shifts were selected to change the perceived phonetic category. This phonetic shift was hypothesized to activate a population of *phonetic* error cells, separate anatomically from the *auditory* error cells that were active in response to a lower-level auditory shift that did not affect the perceived linguistic message. While the Across–NoShift condition resulted in a greater extent and magnitude of neural activation than the Within–NoShift condition, and some ROIs were found to differ between the two conditions, no robust spatial separation of cortical activation was found. This is not conclusive evidence that a separate population of neurons is not implicated in processing a cross-category shift than a within-category shift. As described in Chapter 3, the stability of phonetic category boundaries can be affected by time, stimulus distribution, and lexical bias. Furthermore, variability in a speaker's production of a given vowel causes some productions to be farther from or closer to the category boundary, changing the likelihood that it is crossed. Therefore, despite careful individually-based selection of shift magnitudes, some of the shifts in the Within condition may have been perceived as crossing a category boundary; likewise, some of the shifts in the Across condition may have been perceived as the intended phoneme. This overlap would muddy the distinction between the two conditions and could obscure regional differences in activation caused by the phonetic and non-phonetic shifts. Further research is necessary to either establish separable neural populations or definitively illustrate the overlap between the regions active in response to each condition.

Implications for the DIVA model. According to the DIVA model, formant perturbation causes a mismatch between the auditory expectation for the current vowel and the observed auditory signal (Guenther et al., 2006). This mismatch leads to activation of auditory error cells, located in the posterior superior temporal gyrus. This prediction is strongly supported by the bilateral peri-Sylvian activation noted in the Shift–NoShift contrast. Currently lacking in the DIVA model is a representation of phonetic category boundaries that affect feedback control. Both the increased compensation magnitude and the increase in neural response to cross-category shifts should be reflected in the DIVA model's calculation of auditory error

for a given perturbation. One implementation of this result would simply check whether the perceived speech was closer to another learned category than to the intended category, and, if so, would activate additional phonetic error cells that contributed to the compensatory response. Another implementation would involve a representation of auditory perceptual space, taking into account the "warping" around the category boundaries reflected in the categorization results. In other words, auditory space would be inflated at the boundary region and shrunk at the category center, such that all speech tokens a given distance apart would be equally discriminable. With this perceptual system in place, a shift of a given magnitude would simply be perceived as larger if it crossed the inflated space of a phonetic boundary, triggering a greater response in the auditory error cells that detect the shift. This latter interpretation is better-supported by the current results, as it is yet unclear if distinct populations of phonetic and auditory error cells exist in temporal cortex.

4.4.3 Linguistic influences on the feedback control network

Comparisons to perceptual studies. In ERP studies of speech perception, MMNs were found to be larger for native language contrasts than non-native contrasts (Shamma and Dorman 1999, 2000; Shamma, Marsh, and Dorman, 2000), a result that aligns well with the increase in fMRI activation found in the current study. Similarly, an MEG study reported an MMF response arising from left superior temporal gyrus only to oddballs which crossed phonological boundaries (Phillips et al., 2000). The superior temporal activation seen here could be similar to this neural hallmark of acoustic or phonological change.

Conscious vs. pre-conscious effects. Interestingly, MMFs are often evoked with no conscious attention from the subject. In the current study, half of subjects reported some conscious awareness of the formant shift, some of whom could describe the cross-category shift as a change in vowel identity (the within-category shift was variously compared to a British or Australian accent). These perceptions of vowel change and accent change corresponded to the Across condition and the Within condition, respectively, anecdotal evidence that the cross-phoneme and within-phoneme shifts were accurately perceived as such.

An unpaired *t*-test performed on the two sets of subjects, those conscious and those unconscious of the perturbation, was not statistically significant ($p = 0.48$). In other words, both aware and unaware subjects compensated to the same degree. This lack of difference between the two groups implies the vocal correction is automatic, and that conscious awareness did not help or hinder the feedback correction response. Given that compensation is pre-conscious, it is also possible that the "better" compensators could counteract the perturbation before consciously hearing it.

4.4.4 Alternative approaches and future directions

Potential limitations. The success of these experiments is dependent on the accuracy of the psychophysical test that determines category boundaries. Additionally, as described in the previous subsection, categories may be unstable across time. A subset of subjects were retested on the categorization paradigm and showed reasonable consistency across sessions; however, to ensure stability between the behavioral and imaging tasks, all subjects should be behaviorally tested on multiple occasions, and only those whose category boundaries are stable should go on to be part of the fMRI experiment.

Another limitation of the study is its exclusion of subjects who do not show an "asymmetric" category boundary pattern. It is possible that by selecting for this pattern, the results are not universally applicable to all speakers. However, approximately 50% of participants in the behavioral pretest matched this pattern, a relatively large subset of the tested population.

Alternative approaches. One shortcoming of this paradigm is the inability to directly compare identical acoustic shifts: within a subject, one shift is always in an opposing direction to the other. We have controlled for this by counterbalancing subjects for whom the category-crossing shift is in one direction (e.g., /ɛ/–/ɪ/) and subjects for whom it is in the opposite direction (e.g., /ɛ/–/æ/). However, an alternative experiment design contrasts shifts of the same magnitude and direction *across* subjects. This approach was abandoned in favor of the method described here because of the practical issues with subject populations: a subject who was excluded or who voluntarily dropped out of the study would break a link

of the chain and cause two subjects' worth of data to be excluded. Additionally, the method used in the current study has the advantage of comparing within- and cross-category shifts *in the same subject*, allowing single-subject results to be shown on each individual's brain anatomy and avoiding normalization issues for the first-level analysis.

Future directions. Given the results of the current study, there is good evidence that an across-category or "phonetic" shift causes a greater neural response than a within-category or "non-phonetic" shift of the same magnitude. However, because of the low temporal resolution of fMRI, the neural dynamics of this response are still unknown. To remedy this, a future planned project combines psychophysical and magnetoencephalography (MEG) experiments to investigate the neural dynamics elicited by sudden modification of speakers' auditory feedback. As in the current study, the planned experiments are designed to differentiate perturbations that cause a phoneme change from perturbations of the same magnitude that do not. The goal is to use MEG to examine the time-varying neural response to unexpected feedback perturbation, contrasting that response under conditions of phonetic and non-phonetic change.

There is a natural dispersion of formant values across repeated productions of a given vowel. The goal of the first planned experiment is to probe the center and periphery of vowel production clusters for differential sensitivity to brief auditory perturbation. This experiment explores responses to sudden formant change at varying points in acoustic space. Using MEG, I will measure the neural response to such perturbations to determine whether responses are greater when the productions lie closer to a category boundary.

The second future aim is to evaluate the modulating effects of vowel category training on auditory perturbation responses. Learning a novel vowel target in formant space has the effect of adding new category boundaries between the novel vowel and the well-learned native vowels in neighboring acoustic space. The goal is to assess the degree to which these newly-learned categories affect the responses to perturbation of an existing vowel. A vowel *production training* regimen is designed to introduce a novel production target in an unused region of English formant space. An auditory *perception training* regimen, performed on a separate group of subjects, is designed to train listeners to make a new vowel categorical distinction. I aim to investigate the neural dynamics in effect when auditory

perturbations cause subjects' productions to cross newly-learned boundaries, formed either through vocal learning or perceptual learning.

The current experiment, as well as the future studies proposed here, will improve our knowledge of the neural computations that underlie speech processing. Because speech production deficits are often linked with deficits in auditory perception, this knowledge will lead to an improved diagnosis and treatment of speech disorders with manifestations in speech impairments, such as stuttering, spasmodic dysphonia, and Parkinson's disease. Auditory perturbation paradigms such as the one developed in this research may be beneficial to speakers whose pathologies prevent them from achieving intended speech targets.

4.5 Conclusions

In summary, speakers who experience an unexpected shift of their spoken formants toward another vowel will compensate whether or not the shift causes a category boundary to be crossed. However, cross-category shifts elicit a greater compensatory response than within-category shifts, even when the shift magnitudes are identical. Furthermore, the neural response to cross-category shifts is greater in extent than that to within-category shifts when compared with an unshifted speech baseline, although the effect is subtle enough that a direct contrast of the two conditions was not significant.

Taken together, these results suggest that learned phonetic categories influence the on-line feedback-based control of speech. The warping of perceptual space around category boundaries causes a cross-category shift to evoke a larger auditory error than a similarly-sized within-category shift. Although the compensatory response to perturbations occurs at a pre-conscious level, phonetic knowledge plays a role in determining the size of the compensation necessary to be considered corrective.

The research described in this chapter adds to the existing feedback control literature by introducing the distinction between meaningful linguistic changes and mere acoustic variations introduced in the speech feedback. The categorical nature of speech *perception* has been well-studied, but the influence of perceptual categories on the *motor* act of speaking is

nigh unknown. The current experiment provides evidence that the speech feedback network is differentially sensitive to changes in phonetic category membership. However, further research is necessary to better characterize the neural basis of this special sensitivity to linguistic change.

CHAPTER V

EXPERIMENT 2: AN INVESTIGATION OF PROSODIC ADAPTATION TO PITCH PERTURBATIONS DURING RUNNING SPEECH

5.1 Introduction to prosodic control of speech

Prosody is called the music of speech, encompassing the features of rhythm, pitch, and loudness that fall outside the realm of phonetic representations. Unlike the features of phonetics, prosodic features are suprasegmental; they are not confined to any one speech segment or phone but instead occur at the syllable- or word-level. Prosody plays numerous functional roles, including expressing emotional and attitudinal states, delineating phrase boundaries, and signaling linguistic contrasts such as questions versus statements. Speaker modifications to prosodic cues aid listener comprehension by biasing attention toward informative aspects of the signal (Christiansen & Dale, 2001; Cutler & Darwin, 1981; Cutler & Foss, 1977; Shields et al. 1974). The current study focuses on the linguistic function of prosody, specifically its role in marking stress within an utterance.

Despite the importance of prosody in conveying numerous linguistic and attitudinal contrasts, models of speech production largely focus on segmental and not prosodic control (Guenther et al., 2006; Saltzman & Munhall, 1989). One such model of speech acquisition and production is known as DIVA (Directions Into Velocities of Articulators; Guenther, 1994, 1995; Guenther et al., 2006). DIVA is a biologically plausible adaptive neural network in which acoustic feedback is used to acquire sensory and motor targets for speech sounds. Currently, DIVA lacks a representation of prosodic control, limiting its scope as a comprehensive model of spoken communication. Furthermore, modeling prosody may lead to improved assessment and intervention of neuromotor speech disorders that are characterized by prosodic deficits (Darley et al., 1969, 1975; Duffy, 2005).

The current study is designed to extend the DIVA model to include the control of speech prosody. Minimally, this requires representations of the acoustic cues associated with

prosody: fundamental frequency (F0), intensity, and syllable duration, perceived by listeners as pitch, loudness, and length, respectively (Bolinger, 1989; Lehiste, 1970, 1976; Shattuck-Hufnagel & Turk, 1996). It is unclear, however, whether these cues should be represented in an independent or integrated fashion. An *Independent Channel Model* would posit that F0, intensity, and duration are controlled separately, while in an *Integrated Model*, two or more acoustic cues would be jointly controlled. The current study aims to distinguish between these opposing models as a first step toward representing the complex phenomenon of prosody.

To study prosody without the influence of segmental variables, experimental stimuli were constructed to differ only in the location of emphatic stress within an utterance. While many researchers agree that F0 is the primary cue for signaling stress (Atkinson, 1978; Morton & Jassem, 1965; O'Shaughnessy, 1979), some have argued that duration and intensity cues are also important and may be "traded" for F0 cues (cf. Cooper et al., 1985; Eady & Cooper, 1986; Fry, 1955, 1958; Huss, 1978; Kochanski et al., 2005; Sluijter & van Heuven, 1996a, b; Weismer & Ingrisano, 1979). This transfer of informational cues among prosodic features has been referred to as cue trading (Howell, 1993; Lieberman, 1960). Listeners appear to be able to leverage the cue trading phenomenon to perceive stress even when the speaker's cue patterns differ from their own (see Howell, 1993; Peppé, et al., 2000 in healthy speakers; see Patel, 2002, 2003, 2004; Patel & Watkins, 2007; Patel & Campellone, 2009; Wang et al., 2005; Yorkston et al., 1984 in speakers with dysarthria).

Such cross-speaker cue trading is consistent with both an Integrated Model and an Independent Channel Model of prosodic feedback control. The two models can be differentiated by examining the effects of auditory perturbations during speech production. Perturbation paradigms show the importance of auditory feedback for online vocal control during speaking tasks. Numerous studies have investigated gradual or sudden perturbations to F0 (Burnett et al., 1998; Chen et al., 2007; Jones & Munhall, 2002, 2005; Larson et al., 2000; Xu et al., 2004), as well as to intensity (Bauer et al., 2006; Chang-Yit et al., 1975; Heinks-Maldonado & Houde, 2005) and to vowel formant frequencies (Houde & Jordan, 1998; Tourville et al., 2008; Villacorta et al., 2007). A consistent finding in perturbation studies is a compensatory response: speakers alter their production of the perturbed feature

in the direction opposite to the perturbation. This opposing response is noted both for adaptation paradigms and for paradigms that use brief, unexpected perturbations to auditory feedback. Adaptation paradigms involve persistent exposure to the same perturbation, allowing subjects to adapt their feedforward commands ("adaptation") such that they continue to respond to the perturbation even after it has been removed. In contrast, unexpected perturbation studies use one or more brief, unpredictable perturbations to elicit a compensatory response within a given trial ("rapid compensation").

Most F0 perturbation studies have examined rapid compensations during sustained vowel phonation rather than in linguistic contexts (Burnett et al., 1998; Larson et al., 2000; Xu et al., 2004). While recent work has examined linguistically-relevant perturbations to tones and tone sequences in Mandarin (Jones & Munhall, 2002, 2005; Xu et al., 2004), meaningful prosodic contrasts remain largely unexplored in English. A notable exception is the work of Chen et al. (2007) which examined brief, unexpected upward and downward F0 perturbations as speakers produced the question "you know Nina?" The authors note that upward perturbations, which were not at odds with the rising intonation contour of the target question, resulted in a smaller compensatory response than downward perturbations. Although the perturbation had linguistic relevance, the use of an imitation paradigm may have influenced speaker responses. Further work on eliciting a range of prosodic contrasts in linguistically-motivated communicative contexts is warranted. Additionally, speakers tend to use multiple acoustic cues to signal prosodic contrasts, yet compensatory responses have only been examined within the perturbed parameter, e.g., measuring compensations in F0 for pitch-shifted feedback.

The present study extends the F0 auditory perturbation literature in two main directions. First, meaningful prosodic contrasts in English are elicited by providing contextual scenarios that cue the location of stress within each utterance. Thus, during perturbed trials, speakers must compensate for F0 shifts of the stressed word to preserve the intended prosodic contrast. This linguistically-motivated task may better resemble auditory feedback control during running speech. Second, compensatory responses to F0 perturbation are examined across multiple cues. In light of cue trading relations, changes in intensity and duration may also contribute to the compensatory response, which would be

consistent with the Integrated Model. Alternatively, compensatory responses limited to F0 alone would be evidence for an Independent Channel Model.

In summary, the present study aimed to investigate the prosodic cues used to convey emphatic stress under conditions of near real-time pitch perturbation. Specifically, the following research questions were addressed:

1. Do speakers adapt to targeted F0 perturbations of stressed words within an utterance?
2. Does this adaptation response occur in other features besides F0 (e.g. intensity, duration)?

5.2 Methods and materials

5.2.1 Participants

Twenty-five monolingual speakers of American English with normal hearing and no known speech, language, and neurological disorders between the ages of 20-28 (12 M, 13 F; mean age = 22.0 years) were recruited. Participants were assigned to either the upward shift (Up, hereafter) protocol (6 M, 6 F; mean age = 22.2 years) or the downward shift (Down, hereafter) protocol (6 M, 7 F; mean age = 21.9 years). All participants passed a hearing screening with thresholds at or below 25 dB in at least one ear for 250, 500, 1000, 2000, 4000, and 8000 Hz tones, and reported having vision within correctable limits.

5.2.2 Procedures

Participants were seated in a sound-treated booth and wore a head-mounted cardioid microphone (AKG C420) and over-the-ear headphones (AKG K240), which were used to record productions and present auditory feedback, respectively. A customized graphical interface presented stimuli that participants read aloud. Four sentences were used, each consisting of four monosyllabic words. To control for vowel-dependent differences in F0, vowel nuclei were kept relatively constant across the sentence (Lehiste & Peterson, 1961;

Peterson & Barney, 1952). In each trial, participants produced the four-word sentence with stress on either the first or the second word. The stressed word was cued visually (i.e. using a capitalized, red font) and by providing a contextual scenario. For example, the context sentence "Who caught a dog?" would prompt the target sentence "BOB caught a dog" on the screen. Conversely, "What did Bob do to a dog?" prompted the sentence "Bob CAUGHT a dog." (The remaining three sentences were *Dick bit a kid*, *Doug cut a bud*, and *Dad pat a cat*.) Participants were instructed to produce emphatic stress such that a naive listener could identify the intended stress location.

Given that stressed words tend to have a higher F0 than unstressed words (e.g., Cooper, et al., 1985; Eady & Cooper, 1986; Morton & Jassem, 1965; O'Shaughnessy, 1979), participant-specific F0 thresholds allowed for selective F0 perturbation of stressed words alone. For each participant, a brief pre-test consisting of 16 stimuli was used to determine the perturbation threshold. The threshold was operationally defined as the F0 value that optimally separated stressed words from unstressed words across all 16 trials. F0 values below the threshold value were never perturbed.

In the experimental protocol, each participant produced a total of 480 sentences across four phases: a *baseline* phase with no perturbation; a *ramp* phase during which the perturbation was applied to the auditory feedback in increments; a *perturbation* phase involving full feedback perturbation on the stressed word; and a *post* phase with no perturbation. In the ramp and perturbation phases, F0 of the stressed word was scaled in proportion to the amount it exceeded the threshold. The formulae used to calculate the scaling factors that transformed input F0 to output F0 were:

```
Up:      pitchscale = 1 + ((F0/threshold - 1) * pertval);
Down:    pitchscale = 1 - ((F0/threshold - 1) * pertval);
```

The coefficient `pertval` was set to 0 during the baseline phase, gradually increased to .5 during the ramp phase, held constant at .5 during the perturbation phase, and reset to 0 during the post phase.

For example, if a subject were assigned to the Down group and her threshold was 200 Hz, a 220 Hz production during the perturbation phase would result in a scaling factor of 1 – ((220/200 – 1) * .5), or 0.95. Scaling the input F0 of 220 Hz by 0.95 would result in an output F0 of 209 Hz, an apparent decrease in F0 which would cause the stressed word to sound less stressed. On the other hand, if the same subject were assigned to the Up group, the scaling factor for the same utterance would be 1.05 and would increase the perceived F0 to 231 Hz, thereby increasing the apparent F0 contrast between the stressed word and the unstressed words (see Fig. 5-1).

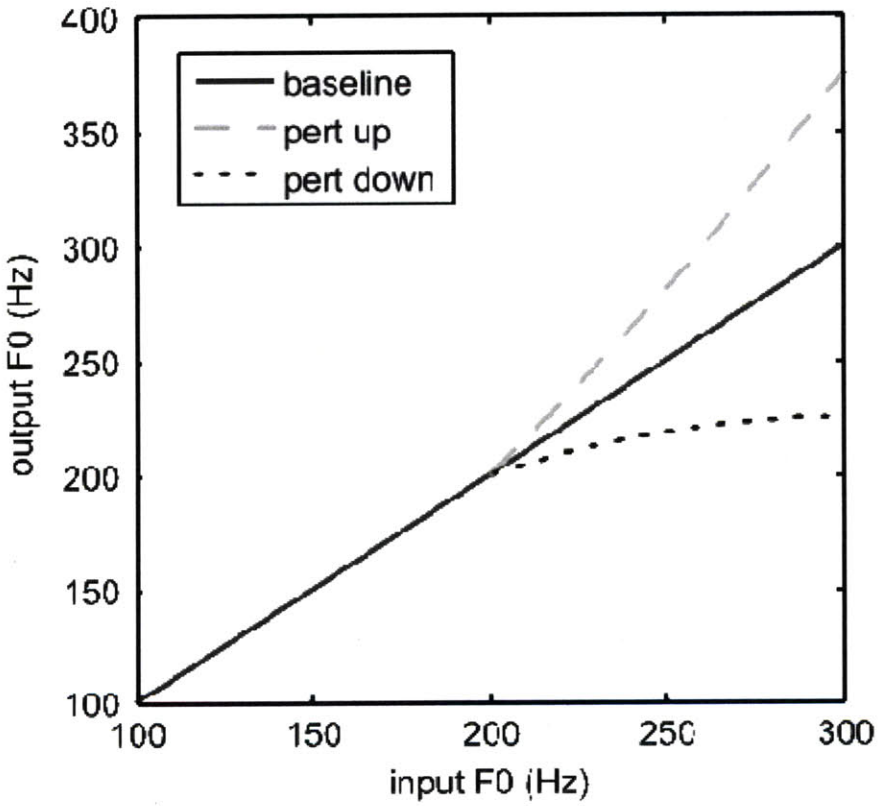

Figure 5-1. Input-output curves for perturbed conditions in a sample subject with a threshold of 200 Hz. No productions below the threshold are ever perturbed; F0 values above the threshold are scaled based on the amount the threshold is exceeded.

Perturbation was implemented using a Texas Instruments (TI DSK 6713) digital signal processing (DSP) board with only minimal processing delay (~26ms). An audio mixer split the subjects' speech signal into two channels, one sent to a computer for recording and one sent to the DSP board. The DSP board used a near-real-time autocorrelation algorithm to track and shift the F0 of each participant. This F0-shifted output was further split and sent both to the subjects' headphones and to the recording computer. Thus, each experimental session produced a stereo waveform consisting of one channel of microphone-recorded data (i.e. what the participant produced) and one channel of feedback-perturbed data (i.e. what the participant heard). The two channels were compared with and without perturbation to ensure that the F0 shift had no effect on intensity.

5.2.3 Acoustic analysis

Customized software implemented in Matlab (CadLab acoustic analysis suite (CLAAS)) was used to derive estimates of F0, relative intensity, and duration for each word across all utterances. Each utterance was manually annotated to demarcate word boundaries ($r = 0.984$ interlabeler reliability for 10% of the data). CLAAS used the Praat autocorrelation algorithm to estimate time-stamped F0 values (Boersma & Weenink, 2009). Similarly, time-stamped intensity values were derived via a root-mean-square calculation of the acoustic waveform. The software operated on the annotations and the time-stamped pitch and intensity values to calculate word duration, average F0, and average intensity across stressed and unstressed words. All analyses were performed on the original spoken utterance, not on the F0-perturbed feedback. The perturbed signal was compared with the microphone-recorded signal to ensure perturbation occurred on the intended trials.

A total of 12,000 utterances were acoustically analyzed (480 trials x 25 participants). A subset of the utterances was examined by hand to ensure correct pitch tracking of all words. Pitch tracking errors, when found, were manually corrected. Errors in pitch tracking were especially problematic for females, particularly for the third and fourth words, which were often in the glottal fry register. Manual correction of automatically generated F0 values was required on 8.3% of the total dataset; 2.7% were excluded. Two female subjects had

greater than 100 mistracked trials (>20%) and were excluded from further analysis. Furthermore, one male subject was excluded due to corrupted acoustic data, and one female subject was excluded because she produced incorrect stress on greater than 40% of trials. The resultant dataset after exclusions was 9752 utterances from 21 participants (Up: 6 M, 5 F, mean age = 22.0 years; Down: 5 M, 5 F, mean age = 22.2 years).

Although acoustic measures were obtained for all four words within an utterance, analyses were restricted to the first and second word (W1 and W2) for two main reasons. First, W1 and W2 were the only two word locations that were counterbalanced in both stressed and unstressed conditions. Second, the word length and sentence position of W3 and W4, respectively, led to variable and imprecise acoustic measurements. Specifically, W3, which was the word "a" in all stimuli, was often reduced or even omitted, while W4 was often glottalized or excessively lengthened owing to phrase-final boundary effects.

5.3 Results

Speaker responses to F0 perturbation were examined in three acoustic variables: mean F0, mean intensity, and word duration. For each trial, the dependent measure was the difference in a given acoustic variable between the stressed word (W1 or W2) and the unstressed word (W2 or W1). This difference was normalized by the mean stressed–unstressed difference in baseline. For simplicity, this normalized value will be referred to as the *contrast distance*, since it represents the degree to which speakers contrasted the stressed and unstressed words within an utterance.

Broadly, the Up and Down groups responded differently to the F0 perturbation, altering their contrast distances in opposite directions with only a short delay from perturbation onset. However, individual speakers were noted to use differing strategies to compensate for perturbations and had individualized time courses for adaptation. Additionally, there was a group-wise trend of a slow increase in F0 and intensity of stressed words across the experiment. To ensure that variations in intensity were not simply the result of a passive physiological correlation with F0, the correlation of these two measures was calculated on a trial-by-trial basis for each participant, and the resulting r scores were

Fisher z-transformed before averaging across the group. This analysis yielded weak correlations ($z = 0.14$ averaged across participants; back-converted to $r = 0.139$), suggesting that changes in subglottal pressure required to modulate F0 had little direct influence on intensity in this study.

To quantify the changes in contrast distance between and within subject groups, paired and independent samples t-tests were conducted on conditions of interest. Between-group (Up vs. Down) differences were compared at all four experimental phases. Because of the upward drift of both F0 and intensity over the course of the experiment, the analysis focuses on these between-group differences. Additionally, within each perturbation direction (Up or Down), differences between all phases (baseline, ramp, perturbation, and post) were compared; therefore, there were six comparisons for each perturbation direction, or twelve within-subjects comparisons. In total, sixteen t-tests were carried out for each acoustic variable. To account for multiple comparisons, the Bonferroni correction factor was used to adjust the α-level to 0.003.

5.3.1 Mean fundamental frequency (F0)

Between-group comparisons show evidence of adaptation to the upward and downward F0 perturbations (Fig. 5-2). In the baseline phase, in which no perturbation was applied, there was no significant difference between the Up and Down groups ($p = 0.45$). However, the two groups diverged in the ramp phase ($p = 0.0014$) and remained significantly different in the perturbation phase ($p < 0.0001$) before falling back below the adjusted significance level in the post phase ($p = 0.02$). Thus, the perturbation resulted in a difference in F0 contrast distance between the two groups. Specifically, speakers altered F0 to enhance or reduce emphatic stress, with the Down group increasing the F0 difference between stressed and unstressed words as compared to the Up group.

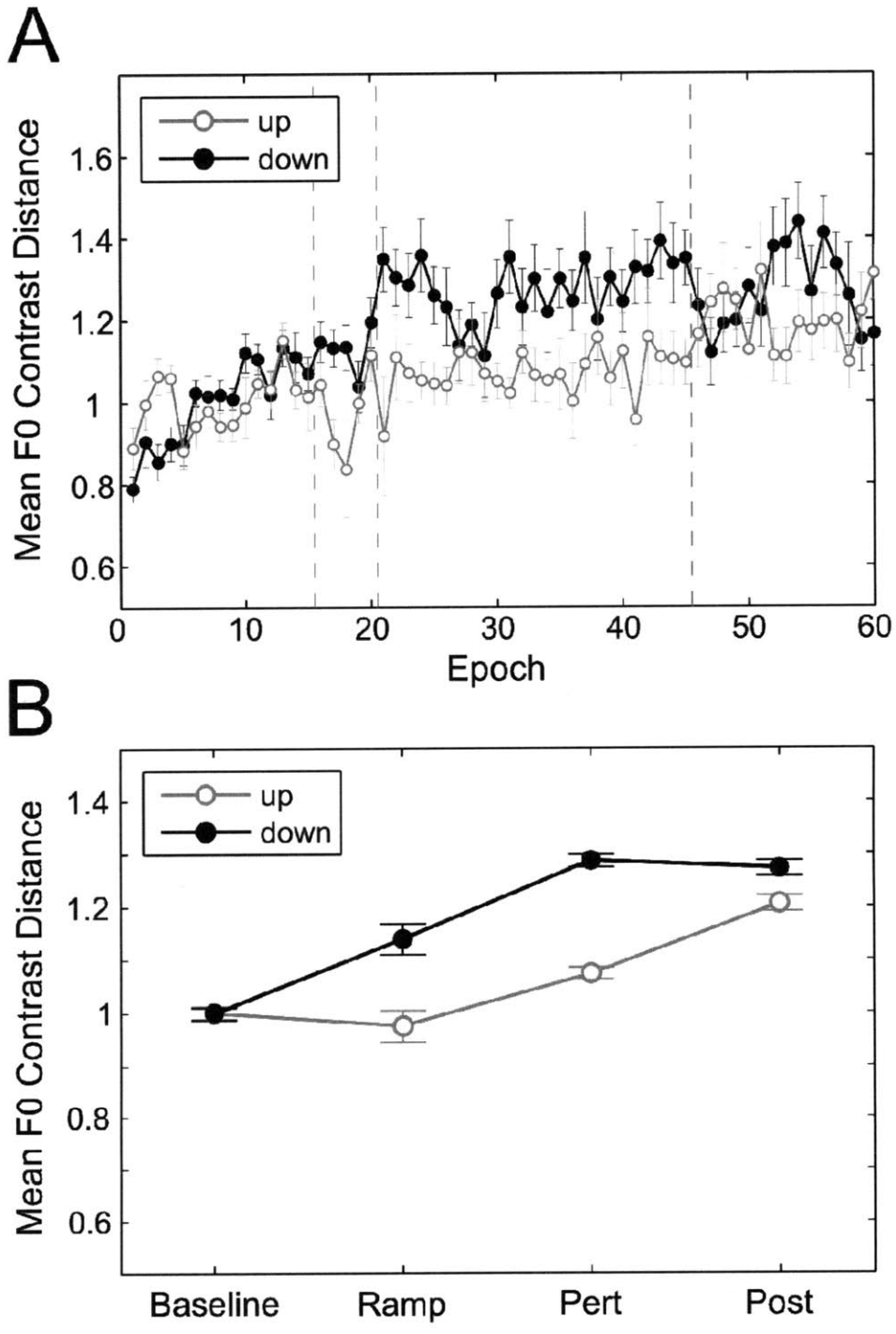

Figure 5-2. F0 contrast distance (A) by epoch and (B) by phase. Error bars show 95% confidence intervals.

Contrasts between phases were used to examine the time course of adaptation within a perturbation direction group. In the Down group, there was a difference between baseline and every other phase (ramp, perturbation, and post), as well as between ramp and every other phase (perturbation and post). There was no difference between perturbation and post phases. Thus, the F0 contrast distance increased from the baseline through to the perturbation phase and then stabilized in the post phase. In the Up group, only differences between the post phase and every other phase (baseline, ramp, and perturbation) were statistically significant. In other words, the F0 contrast distance did not change from baseline values until the upward perturbation was removed in the post phase. Patterns of adaptation within groups may be due in part to the overall upward drift of F0 during the course of the experiment.

5.3.2 Mean intensity

As with fundamental frequency, there was evidence of adaptation in intensity (see Fig. 5-3). Speakers who received a downward perturbation increased the intensity contrast between stressed and unstressed words more than speakers who received an upward perturbation, even though speakers' intensities were unaffected by the perturbation. The two perturbation direction groups significantly differed in intensity contrast distance during the perturbation phase ($p < 0.0001$); however, they were not significantly different in any other phase (baseline: $p = 0.5$; ramp: $p = 0.018$; post: $p = 0.006$).

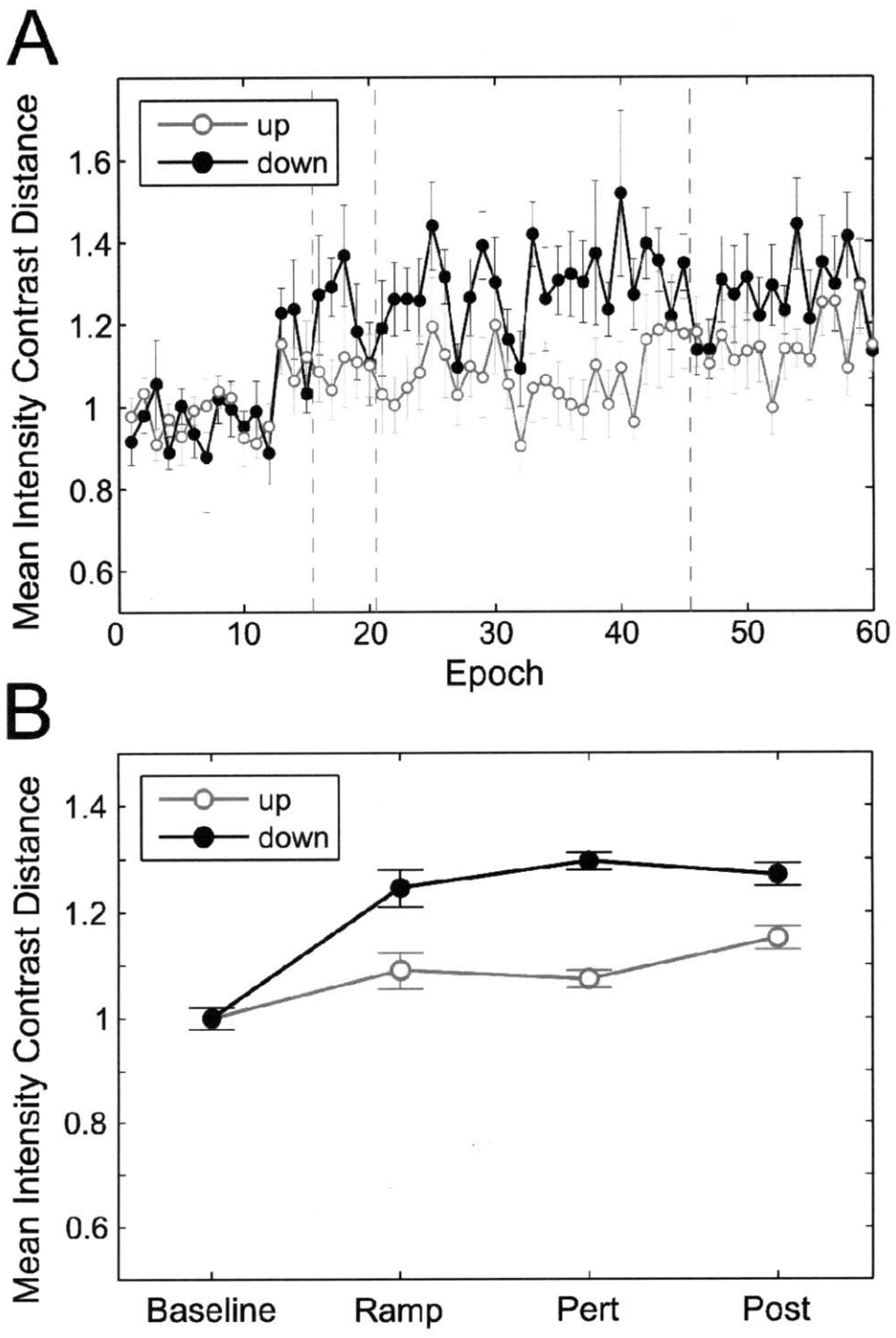

Figure 5-3. Intensity contrast distance (A) by epoch and (B) by phase. Error bars show 95% confidence intervals.

With regard to within-group contrasts, in the Down group only the baseline phase was significantly different from other phases (ramp, perturbation, and post). In other words, the intensity contrast distance increased during the ramp phase and remained increased throughout the experiment. In the Up group, the only significant phase contrast was that between baseline and post phases, again suggesting a slow drift in intensity contrast distance over the course of the experiment.

5.3.3 Word duration

Unlike in fundamental frequency and intensity, there were no significant between-subjects differences (Up vs. Down) in duration contrast distance (see Fig. 5-4). In other words, the perturbation did not effect a durational change between the stressed and unstressed words.

While there were no between-subject differences, there was a difference in the Down group between baseline and perturbation phases ($p = 0.0003$), as well as between baseline and post phases ($p < 0.0001$). In the Up group, however, experimental phase had no effect on duration contrast difference.

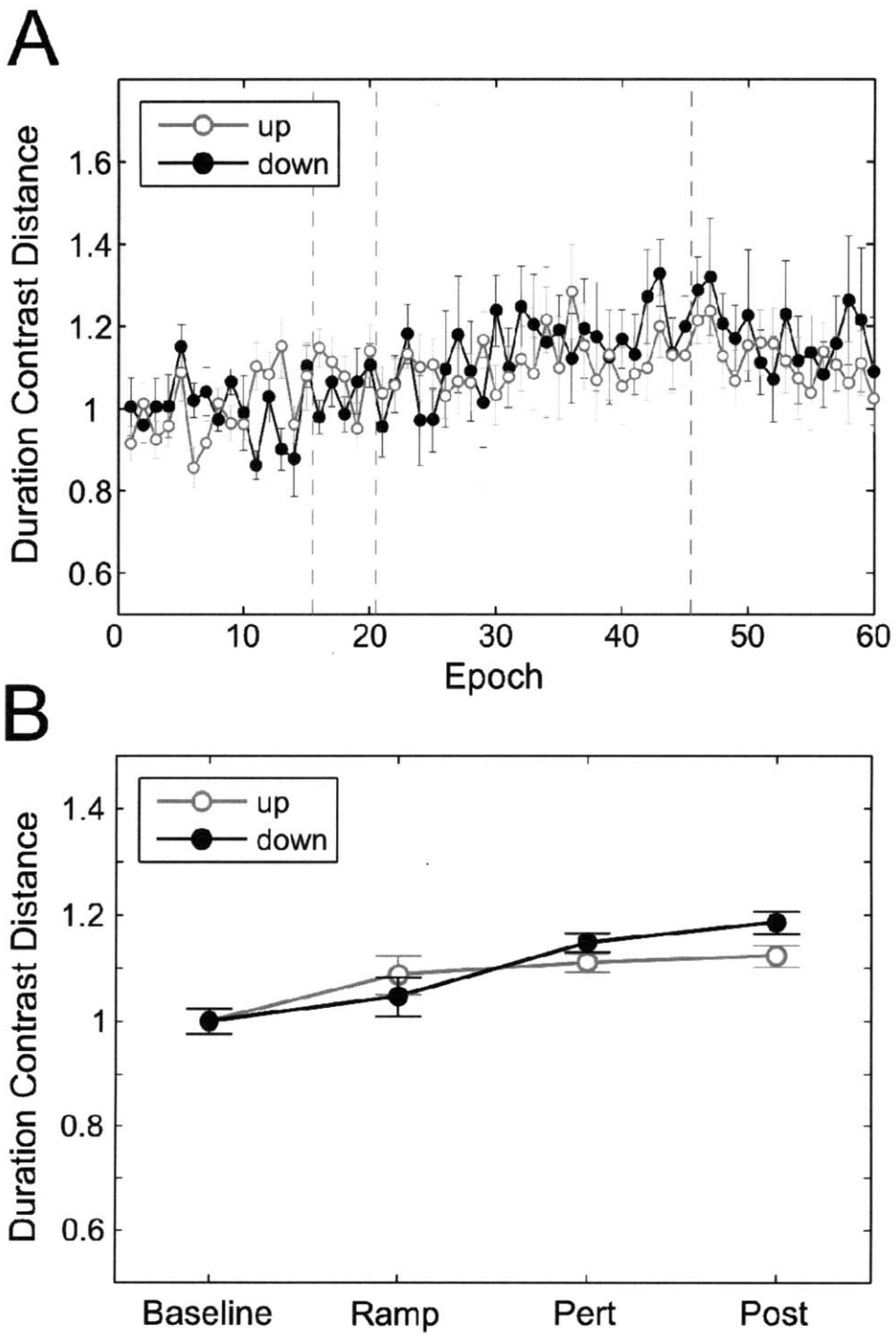

Figure 5-4. Duration contrast distance (A) by epoch and (B) by phase. Error bars show 95% confidence intervals.

5.4 Computational modeling of prosodic adaptation

Responses to F0 perturbation in the current stress task were simulated using a simple differential model with a combined pitch and intensity target. The following functions define "inputs" to the model (`f0in` and `intin`), which are internal representations of the F0 and intensity contrasts that the model attempts to produce, consisting of a baseline offset (`basef0in`; `baseintin`), a noise term (`noise`, scaled by a random factor), and an exponential drift whose terms (`a` and `b`) were derived from the upward drift of the original data.

```
f0in(t) = basef0in + a*(1-exp(b*(t/8))) + noise*rand(1);
intin(t) = baseintin + a*(1-exp(b*(t/8))) + noise*rand(1);
targ(t) = f0in(t) + intin(t);
```

The target output `targ` represents a constant stress contrast distance between stressed and unstressed words. The model adjusts its output by computing the difference between the observed stress contrast and the expected stress contrast (the internal target):

```
f0out(t) = f0in(t-1) - n*(f0perc(t-T)+intperc(t-T) - targ(t-T));
intout(t) = intin(t-1) - n*(f0perc(t-T)+intperc(t-T) - targ(t-T));
```

where `T` represents the corticocortical transmission delays, and `n` is a scaling factor.

The F0 output is then scaled by `f0scale`, representing the perturbation, to form the perceived F0 contrast, `f0perc`. The perceived intensity contrast, `intperc`, is left unchanged:

```
f0perc(t) = f0out(t)*f0scale(t);
intperc(t) = intout(t);
```

By using a combined pitch-intensity target, and combining the pitch and intensity contrasts to check against the target, the model accounts for the effect of F0 perturbation on intensity. As in the experimental results, F0 and intensity are not strongly correlated on a trial-by-trial basis.

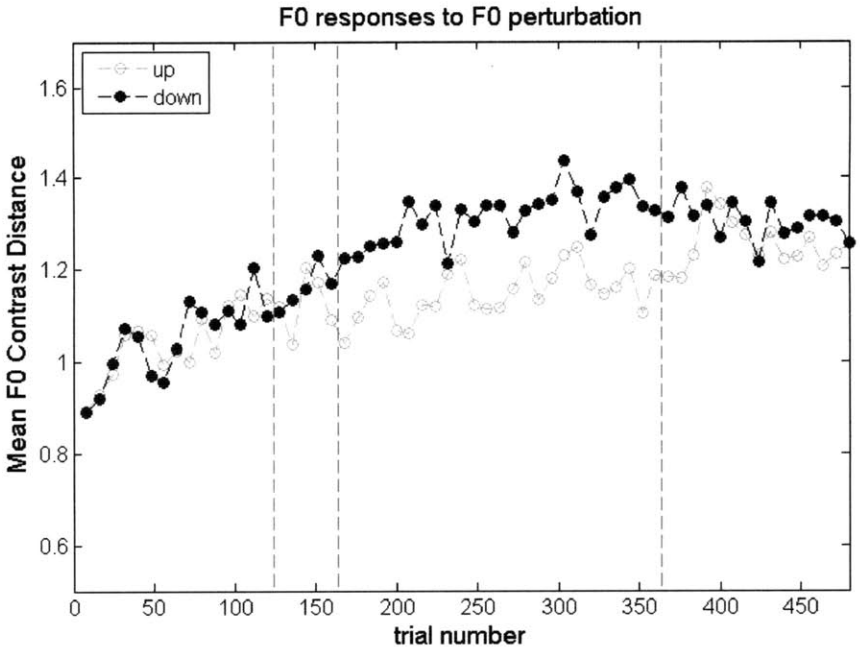

Figure 5-5. Modeling simulation of normalized stressed–unstressed intensity contrast given F0 perturbations.

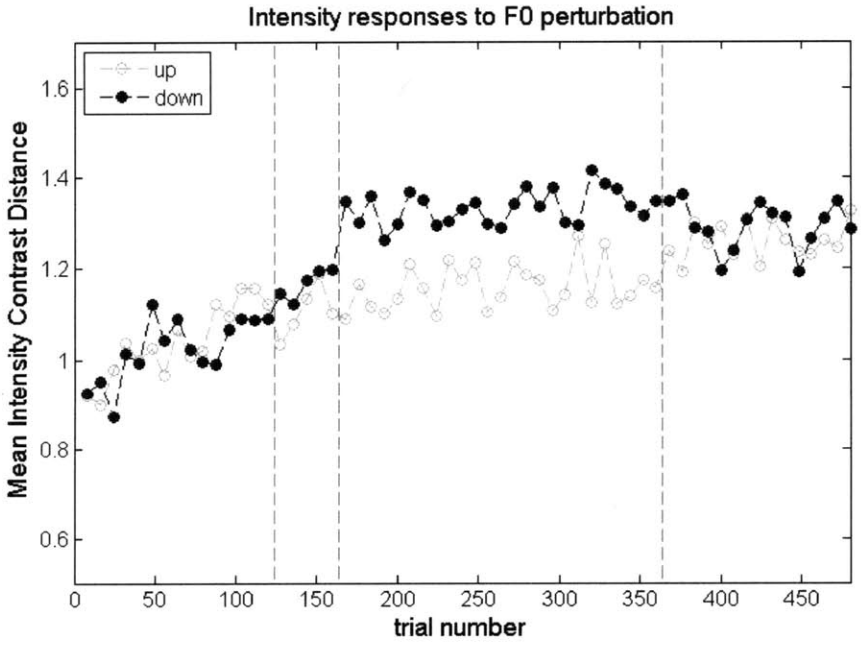

Figure 5-6. Modeling simulation of normalized stressed–unstressed intensity contrast given F0 perturbations.

5.5 Discussion

A major aim of the current study was to distinguish between two representations of speech prosody, one involving independent control of pitch, loudness, and duration, and one involving integrated control of these cues. The emphatic stress task required speakers to modulate prosody while keeping segmental units constant. Introducing a gradual F0 perturbation altered the influence of F0 as a stress-bearing cue. To maintain the appropriate degree of contrast between stressed and unstressed words, speakers might alter only F0, consistent with the Independent Channel Model, or they might alter a combination of prosodic cues to oppose the F0 shift, consistent with the Integrated Model.

Speaker responses to pitch-shifted auditory feedback were measured in three acoustic variables: F0, intensity, and duration. Results indicated that the Up and Down groups adapted to shifts in F0 by altering the contrast between stressed and unstressed words. Specifically, when participants heard their stressed F0 shifted downwards, they increased F0 contrast compared with when they heard their stressed F0 shifted upwards. The interaction between the two F0 manipulations supports the conclusion that speakers are sensitive to upward and downward shifts of F0 in a meaningful prosodic context. Furthermore, compensatory effects were not restricted to F0 but also extended to intensity: speakers altered intensity contrast distance during perturbation, making stressed words relatively louder in the Down group compared with the Up group. This change occurred even though (i) the auditory feedback preserved intensity differences between stressed and unstressed words, and (ii) intensity and F0 were only very weakly correlated in individual trials. These findings provide support for the Integrated Model in that speakers modified both F0 and intensity (although not duration) in response to F0 shifts.

In both experimental groups, Up and Down, speakers' F0 in the perturbation phase was the same or higher than their initial baseline values. This is consistent with the gradual increase in F0 noted in past sensorimotor adaptation studies (Villacorta et al., 2007; Jones & Munhall 2000, 2002). During F0 perturbations to single-word productions, subjects were found to increase F0 over many trials (Jones and Munhall 2000, 2002). In the present sentence production task, only the stressed word showed a similar drift, resulting in an

apparent increase in the contrast distance over the course of the experiment. Given that the contrast distance gradually increased, it appears that the Up group was in fact opposing the perturbation by decreasing F0 relative to an upward drift.

Unlike Jones and Munhall (2002), who found no significant differences in intensity between Up and Down groups, our present results show an increased intensity for the Down group as well as a trend towards increasing intensity over many trials. This was likely due to the prosodic nature of the emphatic stress task. Loudness can be used as a stress cue, and thus participants had reason to manipulate intensity in line with F0 to maintain emphatic stress (Fry, 1955; Kochanski et al., 2005).

While the gradual upward drift in F0 and intensity is consistent with an adaptive response from both Up and Down groups, an alternative explanation is that the two groups adapted to different degrees. The increase in F0 by the Down group may be indicative of a larger compensation than that of the Up group, whose F0 did not decrease from baseline values. The perturbation experienced by the Up group did not interfere with the planned intonation pattern of a stressed–unstressed contrast; thus, there may have been less of a need for a corrective response. That is, there may be more of an incentive to restore a stress contrast that has been attenuated (Down group) than to decrease a stress contrast that has been enhanced (Up group), since the enhancement is aligned with the speaker's goals. This interpretation is also supported by previous work in which downward perturbations elicited a larger rapid compensation response than upward perturbations during question productions (Chen et al., 2007).

While the compensatory changes in both F0 and intensity are in line with the Integrated Model, they might also be attributed to a correlation between F0 and intensity (Gramming et al., 1988; Dromey & Ramig, 1998), as F0 has been found to increase at higher than habitual loudness levels (though not at soft levels). However, our analysis of the trial-by-trial F0-intensity correlation suggests this was not the case. The correlation explains less than 2% of the variance observed in F0; thus, physiological dependencies alone are unlikely to explain the commensurate intensity increase in the group data.

A simple differential model with a combined pitch and intensity target can account for the effect of F0 perturbation on intensity. The model starts with a baseline internal

target for a stress contrast, then computes the difference between that internal target and each observed stress contrast. The internal target is a combination of pitch and intensity contrast values, allowing either cue to contribute to perceived stress. This model maintains an excellent correspondence with the changes in F0 and intensity seen in the experimental results.

In adverse listening conditions speakers enhance prosodic cues to optimize communication (Lane & Tranel, 1971; Letowski et al., 1993; Lombard, 1911; Patel & Schell, 2008; Rivers & Rastatter, 1985; Summers et al., 1988). Downward F0 perturbation in the current study served as a targeted "adverse" condition that speakers had to overcome in order to convey meaningful differences. Similar to the Lombard effect, our targeted F0 shift led speakers to alter contrast distance in multiple cues, using both F0 and intensity in an integrated fashion to signal stress contrasts.

5.5.1 Future directions

The present results support an Integrated Model of prosodic control in which the motor system modulates F0 and intensity in combination to convey stress. However, extending this protocol to perturbation of other prosodic cues will help to generalize the findings. In a planned follow-up to the current study, subjects will undergo intensity perturbations rather than F0 perturbations, allowing a direct comparison of the F0-intensity interaction in each direction. Future experiments will also help disambiguate the roles of word type and word position within an utterance. All of the stimuli in the current study followed the same pattern of word types across the sentence (*Name verbed a noun*). It will be important to vary this pattern to assess the influence of different parts of speech and sentence positions on adaptation responses. Additional experiments investigating the neural circuitry underlying the control of prosodic cues will more clearly define the roles of different brain regions in these processes. Finally, given that perturbation paradigms can induce enhanced linguistic contrasts in healthy talkers, it may be possible to leverage this paradigm for therapeutic benefit in disordered populations.

5.6 Conclusions

An F0 perturbation targeted to the stressed word of an utterance acts to increase or decrease the contrast distance between stressed and unstressed syllables. In other words, it modulates the degree of perceived stress. These upward and downward shifts cause a compensatory response in F0, the perturbed parameter, and in intensity, a feature that was untouched by the perturbation. Because both pitch and intensity modulate listener perception of stress, modulating intensity in the face of pitch perturbations is a valid strategy for overcoming these perturbations and conveying the intended message. Per-trial correlations of F0 and intensity provide evidence that this effect of intensity is not due to passive aeromechanical properties of the vocal tract but is an independent response to the perturbation. Furthermore, modeling results demonstrate the extension of the compensatory response to intensity using perturbation simulations and a simple combined pitch-intensity target for stress contrast. Intensity changes were not seen in a similar adaptation experiment in which the F0 shift had no linguistic import (Jones & Munhall, 2000). Thus, a low-level feedback control response can be broadened to a multi-featural adaptation by the linguistic relevance of the task.

CHAPTER VI

CONCLUSIONS

This dissertation described two experiments designed to advance our understanding of speech motor control through a combination of psychophysics and neuroimaging. Research on congenital deafness and hearing loss demonstrates speech motor control's strong reliance on auditory feedback for learning and maintaining intelligible productions (Cowie & Douglas-Cowie, 1992; Oller & Eilers, 1988; Smith, 1975; Waldstein, 1990). Probing the influence of auditory feedback provides insight about its use in maintaining accurate feedforward commands for speech. The feedback perturbation paradigms employed here are an elegant way to unobtrusively sever the link between speakers' acoustic signal and their reception of that signal, allowing for the investigation of the speech feedback control mechanism.

The results of these two studies illustrate the influence of linguistic experience on production and perceptual ability. The hallmarks of this influence are emphasized here in two key points.

6.1 Auditory goals are dependent on linguistic experience

Languages with different speech sound distributions produce speakers with different phonetic category boundaries. Infants raised in a given linguistic environment show language-specific perceptual ability long before they learn to produce speech; in particular, infants younger than six months a facilitation for native-language contrasts (Eimas et al., 1987; Kuhl et al., 1992, 2006). The learning of sound categories changes the distribution of the firing preferences of neurons in auditory cortical maps, thereby changing the discriminability of sounds from different parts of acoustic space (Guenther et al., 2004). This shaping of perception by auditory experience is evident in phenomena such as the

perceptual magnet effect (Kuhl, 1991) and the learned categorical perception of language-specific phonemes (Repp, 1984).

6.2 Auditory error is enhanced by a linguistic error

The perceptual warping of auditory space has low-level effects on responses to unexpected auditory perturbations of vowels, even when this perturbation is pre-conscious. In Chapter 4, a neuroimaging experiment contrasted unexpected formant shifts that crossed a category boundary with those that did not. A larger neural response magnitude in posterior superior temporal gyrus was found for the cross-category condition, evidence that the linguistic change resulted in an enhancement of auditory error. Furthermore, the cross-category condition also elicited a greater compensatory response, indicative of the corrective motor command generated by the larger perceived auditory mismatch.

A linguistically-relevant mismatch was also found to enhance auditory error in a suprasegmental context. In Chapter 5, a sensorimotor adaptation experiment gradually altered the F0 of the stressed word in a four-word phrase, making it sound more or less stressed. Compensatory responses were observed in multiple cues to stress—both F0 and intensity—even though intensity remained unchanged by the feedback manipulation. Past sensorimotor adaptation experiments that increase or decrease the F0 of spoken words have resulted in an opposing response in F0, but not intensity (Jones & Munhall, 2000), evidence that the linguistic context matters to the speaker during feedback mediation.

The DIVA model of speech production (Guenther et al., 2006), as well as other speech motor control frameworks (Eliades & Wang, 2008; Heinks-Maldonado et al., 2006), hypothesizes the existence of cortical error cells that compare planned and observed productions and generate corrective motor commands when the feedback is off-target. The responses seen in both feedback perturbation experiments are well-aligned with these models of internal auditory comparison; simultaneously, they act as a starting point for further research on incorporating linguistic representations into models of speech motor control.

REFERENCES

Ackermann, H., & Riecker, A. (2004). The contribution of the insula to motor aspects of speech production: A review and a hypothesis. *Brain and Language, 89*(2), 320-328. doi:10.1016/S0093-934X(03)00347-X

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition, 52*(3), 163-187. doi:10.1016/0010-0277(94)90042-6

Åström, K. J. & Murray, R. M. (2008). *Feedback systems: an introduction for scientists and engineers.* Princeton University Press.

Atkinson, J. E. (1978). Correlation analysis of the physiological features controlling fundamental voice frequency. *Journal of the Acoustical Society of America, 63*, 211-222.

Augustine, J. R. (1996). Circuitry and functional aspects of the insular lobe in primates including humans. *Brain Research Reviews, 22*(3), 229-244. doi:10.1016/S0165-0173(96)00011-2

Bartha, L. & Benke, T. (2003). Acute conduction aphasia: An analysis of 20 cases. *Brain and Language, 85*(1), 93-108. doi:10.1016/S0093-934X(02)00502-3

Bauer, J.J., Mittal, J., Larson, C.R., & Hain, T.C. (2006). Vocal responses to unanticipated perturbations in voice loudness feedback: An automatic mechanism for stabilizing voice amplitude. *Journal of the Acoustical Society of America, 119*(4), 2363-2371.

Birn, R. M., Cox, R. W., & Bandettini, P. A. (2004). Experimental designs and processing strategies for fMRI studies involving overt verbal responses. *NeuroImage, 23*(3), 1046-1058. doi:10.1016/j.neuroimage.2004.07.039

Blount Jr., R. (2008). *Alphabet Juice: The Energies, Gists, and Spirits of Letters, Words, and Combinations Thereof; Their Roots, Bones, Innards, Piths, Pips, and Secret Parts; With Examples of Their Usage Foul and Savory.* Farrar, Straus and Giroux.

Boersma, P. & Weenink, D. (2009). *PRAAT: doing phonetics by computer* (Version 5.1.05). [Computer software]. Amsterdam: Institute of Phonetic Sciences.

Bohland, J. W., & Guenther, F. H. (2006). An fMRI investigation of syllable sequence production. *NeuroImage, 32*(2), 821-841. doi:10.1016/j.neuroimage.2006.04.173

Bolinger, D. (1989). *Intonation And Its Uses: Melody In Grammar And Discourse.* Stanford: Stanford University Press.

Boucek, M. (2007). *The nature of planned acoustic trajectories.* Unpublished Masters Thesis, Universität Karlsruhe, Institut für Nachrichtentechnik, Karlsruhe, Germany.

Burnett, T. A., Freedland, M. B., Larson, C. R., & Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *Journal of the Acoustical Society of America, 103*(6), 3153-3161.

Cai, S., Boucek, M., Ghosh, S.S., Guenther, F.H., & Perkell, J.S. (2008). A system for online dynamic perturbation of formant trajectories and results from perturbations of the Mandarin triphthong /iau/. In *Proceedings of the 8th Intl. Seminar on Speech Production*, Strassbourg, France, December 2008, 65-68.

Chang-Yit, R., Pick, H. L., & Siegel, G. M. (1975). Reliability of sidetone amplification effect in vocal intensity. *Journal of Communication Disorders, 8*(4), 317-324.

Chen, S. H., Liu, H., Xu, Y., & Larson, C. R. (2007). Voice F_0 responses to pitch-shifted voice feedback during English speech. *Journal of the Acoustical Society of America, 121*(2), 1157-1163.

Christiansen, M.H. & Dale, R.A.C. (2001). Integrating distributional, prosodic and phonological information in a connectionist model of language acquisition. In *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society* (pp. 220-225). Mahwah, NJ: Lawrence Erlbaum.

Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America, 77*(6), 2142-2156.

Cowie, R.I. and Douglas-Cowie, E. (1983). Speech production in profound post-lingual deafness. In: *Hearing Science and Hearing Disorders* (Lutman, M. E. and Haggard, M. P., eds), pp 183-231, New York: Academic Press.

Cowie, R. I. and Douglas-Cowie, E. (1992). *Postlingually acquired deafness: speech deterioration and the wider consequences.* Berlin: Mouton de Gruyter.

Cowie, R., Douglas-Cowie, E., & Kerr, A. G. (1982). A study of speech deterioration in post-lingually deafened adults. *Journal of Laryngology and Otology, 96*(2), 101-112.

Curio, G., Neuloh, G., Numminen, J., Jousmäki, V., & Hari, R. (2000). Speaking modifies voice-evoked activity in the human auditory cortex. *Human Brain Mapping, 9*(4), 183-191. doi:10.1002/(SICI)1097-0193(200004)9:4<183::AID-HBM1>3.0.CO;2-Z

Cutler, A., & Darwin, C. J. (1981). Phoneme-monitoring reaction time and preceding prosody: effects of stop closure duration and of fundamental frequency. *Perception & Psychophysics, 29*(3), 217-224.

Cutler, A., & Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and Speech, 20*(1), 1-10.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage, 9*(2), 179-194. doi:10.1006/nimg.1998.0395

Damasio, A. R. (1992). Aphasia. *The New England Journal of Medicine, 326*(8), 531-539. doi:10.1056/NEJM199202203260806

Darley, F., Aronson, A., & Brown, J. (1969). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research, 12*, 246–269.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1975). *Motor speech disorders.* Philadelphia, PA: W. B. Saunders.

de Cheveigné, A. (2003). Time-domain auditory processing of speech. *Journal of Phonetics, 31*(3-4), 547-561. doi:10.1016/S0095-4470(03)00041-X

Dehaene, S., Le Clec'H, G., Poline, J., Le Bihan, D., & Cohen, L. (2002). The visual word form area: a prelexical representation of visual words in the fusiform gyrus. *Neuroreport, 13*(3), 321-325.

Dehaene-Lambertz, G. (1997). Electrophysiological correlates of categorical phoneme perception in adults. *NeuroReport, 8*(4), 919-924.

Dehaene-Lambertz, G., & Baillet, S. (1998). A phonological representation in the infant brain. *Neuroreport, 9*(8), 1885-1888.

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science, 298*(5600), 2013-2015. doi:10.1126/science.1077066

Denes, P. B., & Pinson, E. N. (1993). *The speech chain: the physics and biology of spoken language.* Macmillan.

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research*, 91(1). doi:10.1007/BF00230027

Dromey, C. & Ramig, L. O. (1998). Intentional Changes in Sound Pressure Level and Rate: Their Impact on Measures of Respiration, Phonation, and Articulation. *Journal of Speech, Language and Hearing Research, 41*(5), 1003-1018.

Dronkers, N. F. (1996). A new brain region for coordinating speech articulation. *Nature, 384*(6605), 159-161. doi:10.1038/384159a0

Duffy, J.R. (2005). *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management.* St Louis, MO: Elsevier-Mosby.

Eady, S. J. & Cooper, W. E. (1986). Speech intonation and focus location in matched statements and questions. *Journal of the Acoustical Society of America, 80*(2), 402-415.

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology, 4*(1), 99-109. doi:10.1016/0010-0285(73)90006-6

Eimas, P. D., Jusczyk, P. W., & Miller, J. L. (1987). On infant speech perception and the acquisition of language. In S. R. Harnad (Ed.), *Categorical Perception: The Groundwork of Cognition* (pp. 89-112). New York: Cambridge University Press.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171*(3968), 303-306. doi:10.1126/science.171.3968.303

Eliades, S. J., & Wang, X. (2008). Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature, 453*(7198), 1102-1106. doi:10.1038/nature06910

Evans, A. C., Collins, D. L., Mills, S. R., Brown, E. D., Kelly, R. L., & Peters, T. M. (1993). 3D statistical neuroanatomical models from 305 MRI volumes. *Proceedings of the IEEE Nucleur Science Symposium on Medical Imaging, 3*(1-3), 1813-1817.

Fant, G. (1960). *Acoustic Theory of Speech Production.* The Hague: de Gruyter Mouton.

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage, 9*(2), 195-207. doi:10.1006/nimg.1998.0396

Friston, K. J., Holmes, A. P., Poline, J., Grasby, P. J., Williams, S. C. R., Frackowiak, R. S. J., & Turner, R. (1995). Analysis of fMRI Time-Series Revisited. *NeuroImage, 2*(1), 45-53. doi:10.1006/nimg.1995.1007

Friston, K., Stephan, K., Lund, T., Morcom, A., & Kiebel, S. (2005). Mixed-effects and fMRI studies. *NeuroImage, 24*(1), 244-252. doi:10.1016/j.neuroimage.2004.08.055

Fry, D. B. (1955). Duration and Intensity as Physical Correlates of Linguistic Stress. *Journal of the Acoustical Society of America, 27*(4), 765-768.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech, 1*(2), 126-152.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*(1), 110-125.

Gay, T., Lindblom, B., & Lubker, J. (1981). Production of bite-block vowels: Acoustic equivalence by selective compensation. *Journal of the Acoustical Society of America, 69*(3), 802-810. doi:10.1121/1.385591

Ghosh, S., Kovelman, I., Lymberis, J., & Gabrieli, J. (2009). Incorporating hemodynamic response functions to improve analysis models for sparse-acquisition experiments. *NeuroImage, 47*(Supplement 1), S125. doi:10.1016/S1053-8119(09)71199-1

Ghosh, S. S., Tourville, J. A., & Guenther, F. H. (2008). A Neuroimaging Study of Premotor Lateralization and Cerebellar Involvement in the Production of Phonemes and Syllables. *Journal of Speech, Language, and Hearing Research, 51*(5), 1183-1202. doi:10.1044/1092-4388(2008/07-0119)

Ghosh, S. S., Whitfield-Gabrieli, S., & Nieto-Castanon, A. (2009). A Python-based software package for pipelined, batch analysis of fMRI data. *NeuroImage, 47*, Supplement 1, Organization for Human Brain Mapping 2009 Annual Meeting, San Francisco, CA.

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition, 103*(3), 386-412. doi:10.1016/j.cognition.2006.05.010

Goodglass, H. (1992). Diagnosis of conduction aphasia In: S.E. Kohn, Editors, *Conduction aphasia*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Gramming, P., Sundberg, J., Ternström, S., Leanderson, R., & Perkins, W. H. (1988). Relationship between changes in voice pitch and loudness. *Journal of Voice, 2*(2), 118-126.

Guenther, F.H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics, 72*, 43-53.

Guenther, F.H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review, 102*, 594-621.

Guenther, F.H., Ghosh, S.S., & Tourville, J.A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language, 96*, 280-301.

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review, 105*(4), 611-633.

Guenther, F. H., Nieto-Castanon, A., Ghosh, S. S., & Tourville, J. A. (2004). Representation of sound categories in auditory cortical maps. *Journal of Speech, Language and Hearing Research, 47*(1), 46-57. doi:10.1044/1092-4388(2004/005)

Habib, M., Daquin, G., Milandre, L., Royere, M. L., Rey, M., Lanteri, A., Salamon, G., et al. (1995). Mutism and auditory agnosia due to bilateral insular damage—Role of the insula in human communication. *Neuropsychologia, 33*(3), 327-339. doi:10.1016/0028-3932(94)00108-2

Harnad, S. R. (1990). *Categorical perception: the groundwork of cognition.* Cambridge University Press.

Hashimoto, Y., & Sakai, K. L. (2003). Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: An fMRI study. *Human Brain Mapping, 20*(1), 22-28. doi:10.1002/hbm.10119

Heinks-Maldonado, T. H. & Houde, J. F. (2005). Compensatory responses to brief perturbations of speech amplitude. *Acoustics Research Letters Online, 6*(3), 131-137.

Heinks-Maldonado, T. H., Nagarajan, S. S., & Houde, J. F. (2006). Magnetoencephalographic evidence for a precise forward model in speech production. *NeuroReport, 17*(13), 1375-1379. doi:10.1097/01.wnr.0000233102.43526.e9

Holmes, A., Franklin, A., Clifford, A., & Davies, I. (2009). Neurophysiological evidence for categorical perception of color. *Brain and Cognition, 69*(2), 426-434. doi:10.1016/j.bandc.2008.09.003

Houde, J. F. & Jordan, M. I. (1998). Sensorimotor Adaptation in Speech Production. *Science, 279*(5354), 1213-1216.

Houde, J. F., Nagarajan, S. S., and Heinks-Maldonado, T. H. (2007). Dynamic cortical imaging of speech compensation for auditory feedback perturbations. *Proceedings of the 153rd Meeting of the Acoustical Society of America*: Salt Lake City, UT.

Howard, D., Patterson, K., Wise, R., Brown, W. D., Friston, K., Weiller, C., & Frackowiak, R. (1992). The cortical localization of the lexicons: Positron emission tomography evidence. *Brain, 115*(6), 1769 -1782. doi:10.1093/brain/115.6.1769

Howell, P. (1993). Cue trading in the production and perception of vowel stress. *Journal of the Acoustical Society of America, 94*(4), 2063-2073.

Huss, V. (1978). English word stress in the post-nuclear position. *Phonetica*, 35, 86-105.

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition, 92*(1-2), 101-144. doi:10.1016/j.cognition.2002.06.001

Jones, J. A. & Munhall, K. G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. *Journal of the Acoustical Society of America, 108*(3), 1246-1251.

Jones, J. A. & Munhall, K. G. (2002). The role of auditory feedback during phonation: studies of Mandarin tone production. *Journal of Phonetics, 30*(3), 303-320.

Jones, J. A. & Munhall, K. G. (2005). Remapping auditory-motor representations in voice production. *Current Biology, 15*(19), 1768-1772.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America, 118*(2), 1038-1054.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1493), 979-1000. doi:10.1098/rstb.2007.2154

Kuhl, P., & Miller, J. (1975). Speech perception by the chinchilla: voiced-voiceless distinction in alveolar plosive consonants. *Science, 190*(4209), 69-72. doi:10.1126/science.1166301

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science, 9*(2), F13-F21. doi:10.1111/j.1467-7687.2006.00468.x

Kuhl, P., Williams, K., Lacerda, F., Stevens, K., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*(5044), 606-608. doi:10.1126/science.1736364

Ladefoged, P. (1993). *A Course in Phonetics* (3rd ed.), pp 272-275, Harcourt College Pub.

Lane, H. & Tranel, B. (1971). The Lombard Sign and the Role of Hearing in Speech. *Journal of Speech, Language and Hearing Research, 14*(4), 677-709.

Lane, H. & Webster, J. W. (1991). Speech deterioration in postlingually deafened adults. *Journal of the Acoustical Society of America, 89*(2), 859-866. doi:10.1121/1.1894647

Larson, C. R., Burnett, T. A., Kiran, S., & Hain, T. C. (2000). Effects of pitch-shift velocity on voice Fo responses. *Journal of the Acoustical Society of America, 107*(1), 559-564.

Lehiste, I. (1970). *Suprasegmentals.* Cambridge: MIT Press.

Lehiste, I. (1976). Suprasegmental features of speech. In N.J. Lass (Ed.), *Contemporary Issues in Experimental Phonetics* (pp.225-239). New York, NY: Academic Press.

Lehiste, I. & Peterson, G.E. (1961). Transitions, glides, and diphthongs. *Journal of the Acoustical Society of America*, 33, 268-277.

Letowski, T., Frank, T., & Caravella, J. (1993). Acoustical properties of speech produced in noise presented through supra-aural earphones. *Ear and Hearing, 14*(5), 332-338.

Levelt, W. J. M. (2001). Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences of the United States of America, 98*(23), 13464 -13471. doi:10.1073/pnas.231459498

Levelt, W. J. M., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition, 50*(1-3), 239-269. doi:10.1016/0010-0277(94)90030-2

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54*(5), 358-368. doi:10.1037/h0044417

Liberman, A. M., Harris, K. S., Kinney, J. A., & Lane, H. (1961). The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology*, 61(5), 379-388.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1-36. doi:10.1016/0010-0277(85)90021-6

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America, 32*, 451-454.

Lombard, E. (1911). Le signe de l'élévation de la voix," *Annales Maladies Oreilles Larynx Nez Pharynx, 37*, 101-119.

Lombardino, A. J., & Nottebohm, F. (2000). Age at Deafening Affects the Stability of Learned Song in Adult Male Zebra Finches. *Journal of Neuroscience, 20*(13), 5054-5064.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In W. J. Hardcastle and A. Marchel (Eds.), *Speech Production and Speech Modeling.* Boston: Kluwer Academic Publishers.

Margulies, D. S., Vincent, J. L., Kelly, C., Lohmann, G., Uddin, L. Q., Biswal, B. B., Villringer, A., et al. (2009). Precuneus shares intrinsic functional architecture in humans and monkeys. *Proceedings of the National Academy of Sciences, 106*(47), 20069 -20074. doi:10.1073/pnas.0905314106

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748. doi:10.1038/264746a0

Meyer, A. S. (1992). Investigation of phonological encoding through speech error analyses: Achievements, limitations, and alternatives. *Cognition, 42*(1-3), 181-211. doi:10.1016/0010-0277(92)90043-H

Miller, J. D., Wier, C. C., Pastore, R. E., Kelly, W. J., & Dooling, R. J. (1976). Discrimination and labeling of noise–buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America, 60*(2), 410-417. doi:10.1121/1.381097

Morton, J. & Jassem, W. (1965). Acoustic correlates of stress. *Language and Speech, 8*, 159-181.

Motley, M. T., Baars, B. J., & Camden, C. T. (1983). Experimental verbal slip studies: A review and an editing model of language encoding. *Communication Monographs, 50*(2), 79. doi:10.1080/03637758309390156

Mowrey, R. A., & MacKay, I. R. A. (1990). Phonological primitives: Electromyographic speech error evidence. *Journal of the Acoustical Society of America, 88*(3), 1299-1312. doi:10.1121/1.399706

Mutschler, I., Schulze-Bonhage, A., Glauche, V., Demandt, E., Speck, O., & Ball, T. (2007). A rapid sound-action association effect in human insular cortex. *PLoS ONE, 2*(2), e259. doi:10.1371/journal.pone.0000259

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., et al. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature, 385*(6615), 432-434. doi:10.1038/385432a0

Naeser, M. A., Palumbo, C. L., Helm-Estabrooks, N., Stiassny-Eder, D., & Albert, M. L. (1989). Severe nonfluency in aphasia: role of the medial subcallosal fasciculus and other white matter pathways in recovery of spontaneous speech. *Brain, 112*(1), 1-38. doi:10.1093/brain/112.1.1

Neilson, M. D. & Neilson, P. D. (1987). Speech motor control and stuttering: A computational model of adaptive sensory-motor processing. *Speech Communication, 6*(4), 325-333. doi:10.1016/0167-6393(87)90007-0

Nieto-Castanon, A., Ghosh, S. S., Tourville, J. A., & Guenther, F. H. (2003). Region of interest based analysis of functional imaging data. *NeuroImage, 19*(4), 1303-1316. doi:10.1016/S1053-8119(03)00188-5

Nordeen, K. W. & Nordeen, E. J. (1992). Auditory feedback is necessary for the maintenance of stereotyped song in adult zebra finches. *Behavioral and Neural Biology, 57*(1), 58-66. doi:10.1016/0163-1047(92)90757-U

Oller, D. K., & Eilers, R. E. (1988). The role of audition in infant babbling. *Child Development, 59*(2), 441-449.

Ojemann, G. A., & Whitaker, H. A. (1978). Language localization and variability. *Brain and Language, 6*(2), 239-260. doi:10.1016/0093-934X(78)90061-5

O'Shaughnessy, D. (1979). Linguistic features in fundamental frequency patterns. *Journal of Phonetics, 7*, 119-145

Patel, R. (2002). Prosodic control in severe dysarthria: Preserved ability to mark the question-statement contrast. *Journal of Speech, Language and Hearing Research, 45*(5), 858-870.

Patel, R. (2003). Acoustic differences in the yes-no question-statement contrast between speakers with and without dysarthria. *Journal of Speech, Language and Hearing Research, 46*(6), 1401-1415.

Patel, R. (2004). Contrastive prosody in adults with cerebral palsy. *Journal of Medical Speech Pathology, 12*(4), 189-193.

Patel, R. & Campellone, P. (2009). Acoustic and Perceptual Cues to Contrastive Stress in Dysarthria. *Journal of Speech, Language and Hearing Research, 52*(1), 206-222.

Patel, R. & Schell, K. W. (2008). The Influence of Linguistic Content on the Lombard Effect. *Journal of Speech, Language and Hearing Research, 51*(1), 209-220.

Patel, R. & Watkins, C. (2007). Stress identification in speakers with dysarthria due to cerebral palsy: An initial report. *Journal of Medical Speech Language Pathology, 15*(2), 149-159.

Peeva, M. G., Guenther, F. H., Tourville, J. A., Nieto-Castanon, A., Anton, J., Nazarian, B., & Alario, F. (2010). Distinct representations of phonemes, syllables, and supra-syllabic sequences in the speech production network. *NeuroImage, 50*(2), 626-638. doi:10.1016/j.neuroimage.2009.12.065

Peppé, S., Maxim, J. & Wells, B. (2000). Prosodic variation in southern British English. *Language and Speech, 43*(3), 309-334.

Perkell, J. S. (in press). Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics*, In Press, Corrected Proof. doi:10.1016/j.jneuroling.2010.02.011

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *Journal of the Acoustical Society of America, 116*(4 Pt 1), 2338-2344.

Perkell, J., Matthies, M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J., & Guiod, P. (1997). Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. *Speech Communication, 22*(2-3), 227-250. doi:10.1016/S0167-6393(97)00026-5

Peterson, G.E. & Barney, H.E. (1952). Control methods used in a study of vowels. *Journal of the Acoustical Society of America, 24*, 175-184.

Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., McGinnis, M., et al. (2000). Auditory cortex accesses phonological categories: an MEG mismatch study. *Journal of Cognitive Neuroscience, 12*(6), 1038-1055. doi:10.1162/08989290051137567

Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America, 61*(5), 1352-1361. doi:10.1121/1.381409

Pisoni, D.B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics, 15*, 289–290.

Plant, G. (1984). The effects of an acquired profound hearing loss on speech production: A case study. *British Journal of Audiology, 18*(1), 39. doi:10.3109/03005368409078927

Potter, J. M. (1980). What was the matter with Dr. Spooner? In V. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (pp. 13-34). New York: Academic Press.

Purcell, D. W., & Munhall, K. G. (2006). Adaptive control of vowel formant frequency: evidence from real-time formant manipulation. *Journal of the Acoustical Society of America, 120*(2), 966-977.

Price, C. J., U. Noppeney, J. A. Phillips and J. T. Devlin (2003). How is the fusiform gyrus related to category-specificity? *Cognitive Neuropsychology 20*(3-6), 561-574.

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In N.J. Lass, editor, *Speech and language: Advances in basic research and practice*, pp. 243–335. Academic Press, San Diego (CA), USA.

Repp, B. H., & Liberman, A. M. (1987). Phonetic categories are flexible. In S. R. Harnad (Ed.), *Categorical Perception* (pp. 89-112). Cambridge University Press.

Rivers, C. & Rastatter, M.P. (1985). The effects of multitalker and masker noise on fundamental frequency variability during spontaneous speech for children and adults. *Journal of Auditory Research, 25*(1):37-45.

Rivier, F., & Clarke, S. (1997). Cytochrome oxidase, acetylcholinesterase, and NADPH-diaphorase staining in human supratemporal and insular cortex: evidence for multiple auditory areas. *NeuroImage, 6*(4), 288-304. doi:10.1006/nimg.1997.0304

Rosenbaum, D. A. (2009). *Human Motor Control.* Academic Press.

Rumsey, J. M., Horwitz, B., Donohue, B. C., Nace, K., Maisog, J. M., & Andreason, P. (1997). Phonological and orthographic components of word recognition. A PET-rCBF study. *Brain, 120*(5), 739 -759. doi:10.1093/brain/120.5.739

Saltzman, E. L. & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*(4), 333-382.

Shallice, T., Fletcher, P., Frith, C. D., Grasby, P., Frackowiak, R. S. J., & Dolan, R. J. (1994). Brain regions associated with acquisition and retrieval of verbal episodic memory. *Nature, 368*(6472), 633-635. doi:10.1038/368633a0

Sharma, A., & Dorman, M. F. (1999). Cortical auditory evoked potential correlates of categorical perception of voice-onset time. *Journal of the Acoustical Society of America, 106*(2), 1078-1083. doi:10.1121/1.428048

Sharma, A., & Dorman, M. F. (2000). Neurophysiologic correlates of cross-language phonetic perception. *Journal of the Acoustical Society of America, 107*(5), 2697-2703. doi:10.1121/1.428655

Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In P. MacNeilage (Ed.), *The Production of Speech* (pp. 109-136). New York: Springer-Verlag.

Shattuck-Hufnagel, S. (1987). The role of word-onset consonants in speech production planning: New evidence from speech error patterns. In Keller, E. and Gopnik, M. (Eds.), *Motor and Sensory Processes in Language.* Englewood Cliffs, NJ: Erlbaum.

Shattuck-Hufnagel, S. & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*(2), 193-247.

Shields, J. L., McHugh, A., & Martin, J. G. (1974). Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology, 102*(2), 250-255. doi:10.1037/h0035855

Sluijter, A. & van Heuven, V. (1996a). Acoustic correlates of linguistic stress and accent in Dutch and American English. In *Proceedings ICSLP 96, Philadelphia: Applied Science and Engineering Laboratories, Alfred I. duPont Institute*, 630-633.

Sluijter, A. & van Heuven, V. (1996b). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America, 100*(4), 2471-2485.

Smith, C. R. (1975). Residual hearing and speech production in deaf children. *Journal of Speech, Language and Hearing Research, 18*(4), 795-811.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping, 17*(3), 143-155. doi:10.1002/hbm.10062

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage, 23*(Supplement 1), S208-S219. doi:10.1016/j.neuroimage.2004.07.051

Stevens, K. N. (2000). *Acoustic Phonetics.* Cambridge, MA: MIT Press.

Stone, M., Stock, G., Bunin, K., Kumar, K., Epstein, M., Kambhamettu, C., Li, M., et al. (2007). Comparison of speech production in upright and supine position. *Journal of the Acoustical Society of America, 122*(1), 532-541. doi:10.1121/1.2715659

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26(2), 212-215. doi:10.1121/1.1907309

Summers, W. V., Pisoni, D. B., Bernacki, R. H. Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America, 84*(3), 917-928.

Tees, R. C., & Werker, J. F. (1984). Perceptual flexibility: Maintenance or recovery of the ability to discriminate non-native speech sounds. *Canadian Journal of Psychology, 38*(4), 579-590. doi:10.1037/h0080868

Tiede, M. K., Masaki, S., Wakumoto, M., & Vatikiotis-Bateson, E. (1997). Magnetometer observation of articulation in sitting and supine conditions. *Journal of the Acoustical Society of America, 102*(5), 3166.

Tourville, J. A. & Guenther, F. H. (2003). A cortical and cerebellar parcellation system for speech studies. Technical Report CAS/CNS-03-022. Boston, MA: Boston University.

Tourville, J. A., Reilly, K. J., & Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *NeuroImage, 39*(3), 1429-1443.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage, 15*(1), 273-289. doi:10.1006/nimg.2001.0978

Vaux, B., 2008. *Harvard Survey of North American Dialects.* Available online at http://www4.uwm.edu/FLL/linguistics/dialect/. Accessed on 2/15/2010.

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *Journal of the Acoustical Society of America, 122*(4), 2306-2319.

Waldstein, R. S. (1990). Effects of postlingual deafness on speech production: Implications for the role of auditory feedback. *Journal of the Acoustical Society of America, 88*(5), 2099-2114. doi:10.1121/1.400107

Wang, Y-T, Kent, R.D., Duffy, J.R., & Thomas, J.E. (2005). Dysarthria associated with traumatic brain injury: speaking rate and emphatic stress. *Journal of Communication Disorders, 38*, 231-260.

Weismer, G. & Ingrisano, D. (1979). Phrase-Level Timing Patterns in English: Effects of Emphatic Stress Location and Speaking Rate. *Journal of Speech, Language and Hearing Research, 22*(3), 516-533.

Werker, J. F., & Tees, R. C. (1983). Developmental changes across childhood in the perception of non-native speech sounds. *Canadian Journal of Psychology, 37*(2), 278-286. doi:10.1037/h0080725

Whalen, D. H. (1990). Intrinsic velar height in supine vowels. *Journal of the Acoustical Society of America, 88*(S1), S54. doi:10.1121/1.2029052

Whitfield-Gabrieli, S. (2009). *Artifact detection tools.* [Computer software]. Cambridge: Massachusetts Institute of Technology. http://web.mit.edu/swg/software.htm.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7*(7), 701-702. doi:10.1038/nn1263

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., Beckmann, C., et al. (2009). Bayesian analysis of neuroimaging data in FSL. *NeuroImage, 45*(1, Supplement 1), S173-S186. doi:10.1016/j.neuroimage.2008.10.055

Xu, Y., Larson, C. R., Bauer, J. J., & Hain, T. C. (2004). Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *Journal of the Acoustical Society of America, 116*(2), 1168-1178.

Yorkston, K. M., Beukelman, D. R., Minifie, F., & Sapir, S. (1984). Assessment of stress patterning. In M. McNeil, J. Rosenbek, & A. Aronson (Eds.), *The Dysarthria: Physiology, Acoustics, Perception, Management* (pp. 131-162). Austin, TX: ProEd

Zarate, J. M., & Zatorre, R. J. (2005). Neural Substrates Governing Audiovocal Integration for Vocal Pitch Regulation in Singing. *Annals of the New York Academy of Sciences, 1060*(1), 404-408. doi:10.1196/annals.1360.058

This thesis was typeset in Monotype Garamond, an "old-style" or humanist typeface named for the punch-cutter Claude Garamond (c. 1480–1561) and based on the work of Jean Jannon. It was developed in 1922 by Fritz Max Steltzeris at the Monotype foundry. Garamond is considered timeless for its elegance and exceptional readability.