Activity in regions sensitive to auditory speech is modified during speech production:
fMRI evidence for an efference copy

by

Zhuo Zheng

A thesis submitted to the Centre for Neuroscience Studies
in conformity with the requirements for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
September, 2007

Copyright © Zhuo Zheng, 2007

Abstract

Models of speech production postulate that, in order to facilitate rapid and precise control of articulation, the predicted auditory feedback is sent to the auditory system to be compared with incoming sensory data. If this is so, an 'error' signal may be observed when the predicted auditory feedback and the sensory consequences of vocalization do not match. I used event-related fMRI to look for the neural concomitants of such an error signal. In two conditions volunteers whispered 'ted'. In one of these, voice-gated noise implemented in our real-time processing system was used to mask the auditory feedback, which should result in an error signal. Two other conditions were yoked to the production conditions (either clearly heard or masked), but were listen-only and therefore no error signal would be expected. I acquired whole-brain EPI data from 21 subjects using a fast-sparse design. Activity in the superior temporal gyrus bilaterally was significantly greater for clear than masked speech during the listen-only trials ($F(1,20) \geq 12.84$, $p < 0.002$), and significantly higher for masked than for clear speech in the production trials ($F(1,20) \geq 6.68$, $p < 0.02$). This crossover interaction indicates that speech production results in corollary discharge in the auditory system and furthermore suggests that this corollary discharge reflects expectations about the sensory concomitants of speech acts.

Acknowledgements

I would like to thank my supervisor Prof. Ingrid Johnsrude for her sophisticated guidance and generous support all long. This work would not have been possible without her trust and support. I had come into this new yet exciting field with virtually no background, and she has taken me all the way into the completion of this thesis, hand by hand. I am fortunate to have such a knowledgeable and amicable supervisor.

I truly appreciate the knowledge that I have learned from Prof. Kevin Munhall. I am grateful to both Prof. Munhall and Prof. Johnsrude, for the inspiring weekly Speech group seminars they have organized and for the great facility in the laboratory they have provided.

I would like to acknowledge Prof Stanka Fitneva (Chair), Prof Michael Dorris, Prof. Alastair Ferguson, Prof. Ingrid Johnsrude, and Prof. Kevin Munhall for their agreement to be a committee member and for their enormously helpful feedback on my thesis.

I thank all members of Johnsrude's Lab including Heather McDonald, Conor Wild, Allison Mackey, Liz Alexander, etc, for their friendly help during my graduate work and for the pleasant working environment they have established. Special thanks to Conor Wild for proof reading the final draft of my thesis.

Last but not least, my special thanks go to my parents, your love, understanding, encouragement, and support will always be my driving force to carry on. I take pride in being your son.

Table of Contents

Abstract.....	ii
Acknowledgements.....	iii
Table of Contents.....	iv
List of Figures.....	v
List of Tables.....	vi
List of Abbreviations.....	vii
Chapter 1 Introduction.....	1
Chapter 2 Methods.....	20
Chapter 3 Results.....	34
Chapter 4 Discussion.....	56
References.....	69
Appendix 1.....	89
Appendix 2.....	90
Appendix 3.....	91
Appendix 4.....	92
Appendix 5.....	93

List of Figures

Figure 1. The timeline for rapid-sparse imaging in my study.....	23
Figure 2. The experimental paradigm represented by visual cues.....	25
Figure 3. Design matrices for univariate and beta series modeling, respectively.....	32
Figure 4. Areas activated in the contrast $(PM + PN) - (LM + LN)$	35
Figure 5. Areas activated in the contrast $(PM - PN) - (LM - LN)$	39
Figure 6. Areas activated in the contrast $(PN - PM) - (LN - LM)$	41
Figure 7. Areas activated in the contrast $(PM - PN)$	43
Figure 8. Areas activated in the contrast $(LN - LM)$	46
Figure 9. Areas activated in the contrast $(LM - LN)$	47
Figure 10. The parameter estimates based on the ROI data from 21 subjects within a 10- mm cube in the left STG (-66, -45, 15).....	49
Figure 11. The parameter estimates based on the ROI data from 21 subjects within a 10- mm cube in the right STG (54, -18, -6).....	50
Figure 12. Areas functionally connected with the seed voxel (-66, -45, 15).....	55

List of Tables

Table 1. Recent fMRI studies in which perception of speech and nonspeech sounds is compared.....	37
Table 2. Activation peaks for the contrast (PM – PN) – (LM – LN).....	38
Table 3. Activation peaks for the contrast (PN – PM) – (LN – LM).....	40
Table 4. Activation peaks for the contrast (PM – PN).....	42
Table 5. Activation peaks for the contrast (LN – LM).....	44
Table 6. Activation peaks for the contrast (LM – LN).....	45
Table 7. Anatomical regions that exhibited coupling with the seed voxel.....	54
Table 8. fMRI studies discussed in the Discussion section.....	59

List of Abbreviations

ANOVA.....	analysis of variance
BOLD.....	blood oxygen level-dependent
DAF.....	delayed auditory feedback
EEG.....	electroencephalography
EPI.....	echo planar imaging
fMRI.....	functional magnetic resonance imaging
FWHM.....	full width at half maximum
HRF.....	hemodynamic response function
LM.....	listening to masking noise
LN.....	listening to normal speech
MEG.....	magnetoencephalography
MG.....	medial geniculate complex
MTG.....	middle temporal gyrus
PET.....	positron emission tomography
PM.....	production with masking noise feedback
PN.....	production with normal feedback
ROI.....	region of interest
SD.....	standard deviation
SPL.....	sound pressure level
STG.....	superior temporal gyrus
STS.....	superior temporal sulcus
TA.....	acquisition time

TE.....echo time
TMS.....transcranial magnetic stimulation
TR.....repetition time

Chapter 1

Introduction

Speech carries vital information about the identity of speakers and their emotional state, as well as the content of messages being conveyed (Davis and Johnsrude, 2007; Grandjean et al., 2005; Lass et al., 1976; Nygaard and Pisoni, 1998; Smith et al., 2005). Indeed, our ability to use our voices effectively is crucial to successful human interaction. Due to the intricacy of articulation, the acoustics of spoken language are the output of a complex time-varying system that somehow links both perception and production; intuitively, the speech signals that we perceive must reflect the kinetics and intrinsic properties of the organs involved in vocalization, and we subconsciously monitor our speech through our auditory system in order to condition the desired utterance. Even though natural speaking situations entail both vocalizing and listening, the synchrony in timing between these two processes makes it difficult to distinguish them in order to investigate the impact that one might have on the other, how they are linked, and how they interact.

The motor theory of speech perception

The idea that perception and production of speech must depend upon each other to ensure proper communication dates back at least several decades, to when Liberman and his colleagues (Liberman et al., 1967, Liberman and Mattingly, 1985; 1989) put forward the motor theory of speech perception, which states that the objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as motor commands, and speech perception requires processing the acoustic signal in order to recover the coarticulated gestures that produced it. In other words, these gestures are the

primitives that the mechanisms of speech production translate into actual articulator movements, and they are also the primitives that the mechanisms of speech perception restore from the acoustic signal. This view links the motor act of production with the perceptual analysis of speech and points to the sensorimotor nature of speech processing. Despite the theoretical significance of such a link for understanding the cognitive and neural organization of the language system, a great deal of mystery still surrounds the actual relationship between perception and production. In particular, neural mechanisms for processing one's own speech might be fundamentally different from processing speech from someone else. While there is an extensive literature on the involvement of the motor system in the perception of the speech of others (Ackermann and Riecker, 2004; Fadiga et al., 2002; Ingvar, 1983; Jardri et al., 2007; McGurk and MacDonald, 1976; Skipper et al., 2005; Watkins and Paus, 2004; Watkins et al., 2003), there is a paucity of data on how perception of one's own voice is involved in speech production.

Significance of auditory feedback in speech production

An examination of perception of our own speech in the context of speech production would enable us to better understand the psychophysical and neurophysiological basis for the complex feedback control system that underlies speech communication. Abundant evidence has shown that hearing one's own voice accurately is critical for learning and maintaining fluent and stable speech. Congenital or early deafness in life without immediate cochlear implantation severely hinders the development of normal speech in children (Geers et al., 2003; Schauwers et al., 2004). Damage to hearing capability in adults causes the deterioration of many aspects of speech such as control of fundamental frequency and intensity (Cowie et al., 1982). Articulatory dysfunctions in humans such as

stuttering can be significantly ameliorated by simply altering auditory feedback (AAF treatment) (see Lincoln et al., 2006 for review). Therefore, that speech perception and production are linked in some way is uncontroversial. The challenge for researchers and the source of controversy lies in the clarification of how and at what point they are linked. Second-language learning programs and therapeutic interventions are usually directed towards either perception or production, but, by necessity, usually involve both. Knowing precisely how the two sides of language interact would also allow therapists and educators to strategically use both modalities in intervention.

Inspiration gained from songbirds

Songbirds are one of the most extensively studied examples of vocal learners for two reasons. First, the well characterized and specialized anatomy makes songbirds an ideal system in which to examine the neural substrates of vocal motor learning. Second, phenomena related to song learning in songbirds strongly resemble those related to learning of speech by human infants (Doupe and Kuhl, 1999). Due to the similarities between song learning and speech development, the literature on songbirds may help us develop an understanding of general principles that contribute to the complexity of vocal behavior in humans.

There are generally two phases involved in song learning: a “sensory” phase (perception of the sounds that songbirds are exposed to) and a “sensorimotor” phase (song production) (Marler, 1970; Price, 1979). Early deafened songbirds before vocal practice develop abnormal song patterns that show no sign of learning (Konishi, 1965; Wang et al., 1999), indicating that auditory feedback is essential for normal vocal production. Behavioral studies have shown that when learning to sing, songbirds form an

internal song template based on previous exposure, and then shape their vocal output by comparing auditory feedback of their production with this template (Thorpe, 1958; Konishi, 1985). Leonardo and Konishi (1999) observed that perturbation of auditory feedback leads to a gradual deterioration of adult zebra finch song, and the song is recovered following the restoration of normal feedback. The authors suggested that the deterioration of song in response to the perturbation of feedback may be indicative of an actively driven process. The perturbed feedback was interpreted as “incorrect” feedback, leading to the generation of an error signal that actively imposed changes on song sequences. These behavioral findings in songbirds demonstrate a series of processes that occur during song learning: establishment of a song template, evaluation of auditory feedback against the template, and consequent modification of vocal output (Brainard and Doupe, 2000).

Behavioural evidence for linkage between speech perception and production in humans

Behavioural studies have lent support to the view that the coupling between perception and production is essentially a sensorimotor control process in which *auditory feedback* serves as a key control signal. By artificially altering the auditory feedback in laboratory settings, some studies have observed subsequent changes in production. For example, a number of experiments with normal speakers have demonstrated that delayed auditory feedback (DAF) produces disruptive effects on running speech. These negative effects include utterance errors (repetition of phonemes or syllables), changes in speech rates, and varied vocal intensity (Black, 1951; Fukawa et al., 1988; Howell, 1990; Mackay, 1968; Siegel et al., 1982; Stager and Ludlow, 1993); Zanini et al. (1999) have shown that subjects speaking at a normal rate while receiving over 200ms DAF generated

significantly more speech errors compared to those receiving normal feedback. Furthermore, using various experimental paradigms such as formant frequency shifted feedback (Elman, 1981; Howell et al., 1987; Jones and Munhall, 2000; Kalinowski et al., 1993; Purcell and Munhall, 2006a,b), or pitch-shifted feedback (Burnett et al., 1997; Burnett et al., 1998), some researchers have explored the effects of varying different auditory feedback variables on voice responses. Burnett et al. (1997) showed that when the pitch of a subject's voice was shifted either upward or downward, the subject responded by reciprocally lowering or raising the fundamental frequency (F0) of their voice. However, this pattern was not consistent; there were some subjects that changed their voice F0 in the same direction as the feedback shift. By perturbing only the formant frequencies in the production of a vowel, Purcell and Munhall (2006a,b) have shown that the manipulations led to adjustment of production in the shifted formants in the opposite direction. When the acoustic feedback was returned to normal, the formant frequencies only gradually returned towards but did not reach the baseline level immediately. This adaptive control of production can spread not only across consonant contexts (e.g., compensation for perturbation of the vowel [ɛ] in “peg” generalizes to “gep” or “teg”) but also to other vowels (e.g., to “pip” or “pap”), representing a systematic compensation for the feedback alterations (Houde and Jordan 1998). Collectively, these data can be interpreted as evidence that speech production is sensitive to modifications in parameters of perception.

The cognitive mechanism that links speech perception and production

The cognitive mechanism underlying how production is coupled with perception was not clear until the emergence of the concept of corollary discharge, which was initially developed by Helmholtz (Hermann von Helmholtz 1878). Helmholtz observed that the objects in our visual receptive field remain still at the perceptual level following a spontaneous or voluntary eye movement whereas the objects appear to be moving if the eye is passively moved by an external force. As Helmholtz argued, the stability of our visual world is achieved as a result of an “effort of will” engaged in altering the positions of the eyes which precisely matches the internally generated motor command causing the eye movement. Later on von Holst and Mittelstaedt (1950) generalized this concept by presenting a model in which they considered the pathway that conveys information to the central nervous system as the *sensory afferent* and the pathway that launches the motor command as the *motor efferent*. The input through the sensory afferent caused by the activation of the motor efferent was referred to as *reafference*. Von Holst and Mittelstaedt (1950) speculated that each motor command is accompanied at the level of the central nervous system by a signal called the *efference copy*, which is sent internally to the intended sensory modality and used to cancel the reafference since the reafference can be predicted from the motor command and contributes no meaningful information to our perceptual experience. The cancellation, however, is a non-trivial process that involves the transformation of the sensory input into a neural representation that is directly comparable to the motor command (Webb, 2004). Since the inception of this model, the efference-copy system has been extensively investigated in a wide variety of domains (Bell 2001; Blakemore et al., 1998; von Holst and Mittlestaedt, 1950; Wolf and

Heisenberg, 1990; Poulet and Hedwig, 2002; Rose 2004). One line of strong evidence for the existence of the efference copy comes from the fact that it has been shown to have inhibitory effects on sensory cortical neurons in birds, bats, monkeys and humans. For example, Poulet and Hedwig (2002) have demonstrated, using intracellular recordings in singing crickets, that presynaptic inhibition of internally generated auditory afferents and postsynaptic inhibition of identified auditory interneurons occur in phase with the song pattern. In bats, a 15-dB attenuation of the responses in the lateral lemniscus of the midbrain has been observed during self-vocalization (Suga and Shimozawa, 1974). In monkeys, Muller-Preuss et al. (1980) have found that areas activated during vocalization have an inhibitory effect on the auditory cortex, making it less responsive to external sounds; furthermore, Eliades and Wang (2003; 2005) observed a suppression of some auditory cortical neurons during voluntarily produced vocalizations in marmosets. The onset of this suppression is several hundred milliseconds prior to the vocal production; the authors conclude that this is indicative of a neurally mediated inhibition caused by an efference copy from the vocal motor area (Eliades and Wang, 2003). Dampened temporal lobe responsiveness during speech production has been demonstrated in humans using electrical recordings during neurosurgery (Creutzfeldt, et al., 1989a,b). Using MEG measurement, Houde et al. (2002) have shown that the auditory M100 activity to speech sounds is significantly attenuated during self-vocalization relative to during passive listening to recordings. A similar study done by Curio et al. (2000) indicated that speaking primes the auditory cortex by dampening reactions to self-produced sounds, but not to tape-recorded speech. The M100 response of the auditory cortex was even attenuated by 44-71% during reading aloud (overt production) compared with reading

silently (covert production). In addition, a subject's own voice feedback elicits a reduced auditory N100 response compared to that elicited by altered or alien auditory feedback (Heinks-Maldonado et al., 2005) or simply playback (Ford et al., 2001) during EEG measurements. The observed suppression in neural activity in the auditory cortical areas during vocalization presumably represents the modulation of sensory processes by motor output, indicating that the perceptual system is not solely driven by sensory input, but is also strongly influenced by predictions of sensory input generated by the motor system (i.e., by an efference copy).

Forward models and the feedback system

Based on his mathematical model (DIVA), Guenther (Guenther, 2006) hypothesizes that the efference copy, sent from the speech motor control area to auditory cortex, serves as a “target” for incoming auditory data. If the incoming auditory reafference does not match this target, auditory cortical neurons will be activated, counteracting the inhibitory effect of the efference copy. Therefore, in laboratory settings, if the reafference is experimentally manipulated or modified through the sensory afferent pathway so that it does not match the efference copy, then disparity arises which one could term an “error signal”. Mathematically, this work builds up an explicit forward model between the motor system and auditory cortical areas and presents a solution to the question of how an efference copy could possibly lead to an error signal in sensory cortical areas. An interesting feature that differentiates Guenther's DIVA model from other forward models is that this model's components have been associated with specific anatomical locations in the brain; the DIVA model suggests that the motor system including premotor cortex

and primary motor cortex serves as the vocal motor center which issues the efference copy to the superior temporal cortex during speech production.

Another model, called Modular Selection and Identification for Control (MOSAIC) approaches this problem from a more computational perspective (Wolpert and Kawato, 1998; Haruno et al., 2001; Wolpert and Flanagan, 2001). The model considers the problem of sensorimotor control as a computational process of determining the input to a specific system that one wishes to control, and the desired output that one wants to achieve. Humans are constantly selecting the input in the form of motor commands that achieve some required output that is perceived as sensory feedback. The architecture of this model consists of multiple pairs of inverse (controller) and forward (predictor) models, in which each inverse model provides a motor command that will lead to a desired change in state whereas each forward model is responsible for predicting the next state given the motor command. Each forward model is paired with a corresponding inverse model to form a predictor-controller pair. The idea being that, if the prediction of one of the forward models matches the actual sensory feedback, then its paired controller will be selected to determine subsequent motor commands (Wolpert and Flanagan, 2001). The sensory prediction error from a specific forward model is assigned a probability. The set of probabilities from an array of forward models is then used to weigh the outputs of the corresponding controllers, namely, to determine what motor command will be issued. As both forward and inverse models rely upon the dynamics of the motor system, which adapts to different situations and conditions, these models must be also adaptable.

The ability to rapidly adapt to new environments is a distinctive feature of neural systems that subserve learning. Applying both DIVA and MOSAIC models to the

auditory system, after an auditory target for a sound has been learned, the model can issue a motor command to coordinate muscles to utter that sound. However, on the first few trials, the model may not have a properly tuned forward command for the sound; thus it must rely heavily on sensory feedback to the auditory system to adjust the motor commands. On each attempt to produce that sound, the forward command is updated to incorporate the commands fed back by the auditory system on that attempt, which leads to a more accurate motor command for the subsequent trial. This iterative process continues until the forward command is sufficiently precise so that it causes very few auditory errors during production of the sound and thus no error signal will be induced. The insight that I gain from these models is that the feedback control system underlying our language capability is not a fixed entity; rather it is trained with prediction errors, and learned through experience. These models provide a good framework within which some sensorimotor processes can be interpreted and the functional regions supporting these processes can be predicted. Nevertheless, the results are preliminary and computational approaches can not always account for the physiological changes that occur within the brain.

The anatomical and functional organization of the auditory cortical system

The existing data on the mechanisms of speech processing and anatomical connectivity within the auditory cortex provide additional insight into how the efference copy may have an effect on auditory cortical activity. Before the speech signal reaches the temporal cortical region, it has already been extensively processed and recoded along the ascending auditory pathway from the cochlear nucleus to the thalamus (Eggermont, 2001). Physiological and functional data on non-human primates indicate that there are at least

three discrete processing stages in auditory cortex, comprising core, belt, and parabelt divisions (see Kaas et al., 1999; Gazzaniga et al. 2002). Each of these divisions receives extensive projections from thalamus and contains multiple parallel processing units (Morel et al., 1993; Pandya 1995; Kosaki et al., 1997; Rauschecker et al., 1997; Hackett et al., 1998). The core area comprises three cochleotopically organized fields that receive intensive inputs from the ventral division of the medial geniculate complex (MGv) (Kaas et al., 1999). These three fields, namely the most caudal primary auditory field (A1), the rostral field (R), and the rostrotemporal field (RT) act like three relatively independent entities and process auditory information from the MGv in a parallel manner. Neurons in this core region respond well to pure tones at their characteristic frequency (Rauschecker et al., 1995; 1997). Surrounding the core is the belt region, which receives its inputs primarily from the dorsal (MGd) and medial (MGm) parts of the medial geniculate complex, as well as from neighbouring core fields. The belt region is a second stage in the processing hierarchy and relays the processed acoustic signal from the core to the next level through its several subdivisions (Morel et al., 1993; Kosaki et al., 1997; Rauschecker et al., 1995). Neurons in the belt region possess more complex receptive field properties than those in the core, presumably reflecting integration over convergent inputs (Hall et al., 2003). It has been suggested that neurons in the belt are sensitive to spectrally complex stimuli such as vocalization (Rauschecker et al., 1995; Tian et al., 2001). The third well-documented division located in the auditory cortex is the most lateral region, the parabelt, which, like the belt, receives its thalamic inputs from the MGd and MGm. The parabelt begins to display selectivity in anatomical connectivity in that the caudal parabelt receives the inputs from the caudal belt and the rostral parabelt receives

inputs from the rostral belt (Tian et al., 2001). In addition, consistent with its position as a higher level processing stage along the thalamus-core-belt-parabelt pathway, the parabelt exhibits high sensitivity to certain spectral and temporal elements inherent in phonetic/acoustic signals. Even though the precise location of parabelt cortex is uncertain in humans, functional imaging studies have shown that this approximate region within STG is recruited more strongly for speech or music than for white noise (Zatorre et al., 1992; Binder et al., 1994). Thivard and his colleagues (Thivard et al., 2000) have demonstrated sensitivity of the left STG (presumably anatomically equivalent to parabelt in non-human primates) to dynamic spectral shape over time and not simply to spectral complexity; this pattern is not observed in the primary auditory cortex (A1). It has been suggested that the parabelt may serve as a specialized unit that performs high-level segregation (Griffiths and Warren, 2002).

The most notable feature of the parabelt rests on its ability to project to several regions including the dorsal bank of the superior temporal sulcus (STS), superior and middle temporal gyri, prefrontal cortex, and parietal cortex (Kaas et al., 1999; Romanski et al., 1999). Neurons in these regions have been suggested to participate in multimodal integration and exert cognitive influences on auditory percept formation, representing a higher (presumably fourth) level of processing (Hikosaka et al., 1988; Iwai et al., 1987; Kaas et al., 1999). Many authors have distinguished two functional streams involving the parabelt region, which are essentially the extremes of a broad, continuous system of connections between temporal- and frontal-lobe regions (Petrides and Pandya, 1988). The “ventral” stream involves connections between anterior parabelt regions and the anterior ventrolateral frontal regions (Hickok and Poeppel, 2004; Romanski et al., 1999; Scott and

Johnsrude, 2003; Tian et al., 2001), and the “dorsal” stream connects posterior parabelt and posterior dorsolateral frontal regions. Based on their specialization in processing, the ventral stream is sometimes referred to as “what” stream for its possible involvement in mapping sound onto meaning and accessing the mental lexicon (Scott et al., 2000; Liebenthal et al., 2005, Davis and Johnsrude, 2007), and the dorsal stream as the “where” stream for its involvement in auditory spatial perception; or the “how” stream, given its apparent involvement in auditory-echoic transformations (Kaas and Hackett, 1999; Rauschecker, 1998). It has been suggested that acoustic features of the incoming sound are first extracted in the dorsal STG region for initial spectrotemporal analysis and then in the posterior half of the superior temporal sulcus (STS) for phonological-level processing, before entering the ventral and dorsal streams (Hickok and Poeppel, 2007). Although there is a consensus regarding the role of the ventral stream in “what” processing, there are some other views concerning the dorsal processing stream in the literature. Belin and Zatorre (2000) have proposed that the dorsal stream is geared towards tracking temporal changes in the frequency spectrum of the acoustic signal, which makes it a particularly suitable candidate for speech perception. Hickok and Poeppel (2000) maintained that the dorsal stream is critical for auditory-motor integration, since the neural systems in this stream are heavily involved in a verbal working-memory task that required access not only to sound-based representations of speech, but to articulatory rehearsal component in support of articulatory-based representations of speech (Aboitiz and Garcia, 1997; Wilson, 2001). Rumiati et al. (2005) argued that the dorsal stream is implicated in the acquisition of language, as evidenced by its engagement in sound imitation. The findings from these studies can be reconciled by proposals for a dual-stream model of speech

processing (Scott and Johnsrude, 2003; Hickok and Poeppel, 2003; 2007) in which the dorsal stream links a sensorimotor interface in the left posterior STG with an articulatory network involving prefrontal and motor regions. The area Spt, located within the left posterior STG, has been previously observed to be equally activated during the reproduction of sensory stimuli that were perceived either aurally and visually (Buchsbaum et al., 2005), indicating that Spt might be more tightly coupled to a motor effector system, the vocal tract, as a part of an auditory-motor integration circuit (Hickok and Poeppel, 2007). Therefore, one important function of the dorsal processing stream might be to relay the encoded sequence of incoming sounds to vocal motor regions where it can be transformed into a sequence of articulatory gestures (Scott and Wise, 2004).

The dorsal auditory pathway has also been implicated in the preparation of constrained motor responses to auditory stimuli (Warren et al., 2005). According to a template-matching theory of auditory processing (Griffiths and Warren, 2002), the incoming spectrotemporal acoustic signal is decomposed into a series of auditory representations and then matched with stored templates based on auditory experience. These are transformed online into a sequence of motor representations with coordinated actions by the dorsal processing stream, consistent with the categorical perception of speech sounds (Hofman et al., 1998; Davis and Johnsrude, 2007). Although Warren et al. (2002) suggest that the template-matching process facilitates efficient bidirectional transfer of information between auditory and motor regions along the dorsal pathway, they do not make it clear how motor regions would have an impact on processing in the dorsal auditory pathway during speech production and what the “templates” refer to.

Haueisen and Knosche (2001) have demonstrated that pianists exhibit enhanced motor activity underlying finger movements when listening to well-trained piano music segments, implicating motor regions in the recognition, understanding and imitation of the actions in relation to the sensory stimulus. Using TMS, Watkins et al. (2003) have observed that auditory perception of speech facilitates the excitability of the motor system involved in speech production. Wilson et al. (2004) discovered that passive listening to meaningless monosyllables activates motor areas specifically involved in producing speech, indicating that speech perception involves the motor system in a process of acoustic to articulatory representations mapping to access a phonetic code with motor properties. Later, Wilson and Iacoboni (2006) argued that not only does the auditory pathway play an important role in transforming acoustic signal to phonetic codes (templates), but the motor system is actively involved in the generation of candidate phonemic categorizations which would then be compared to the sensory input in auditory areas. In their study, they presented English speakers with novel speech sounds and observed that the signal change in the bilateral superior temporal regions strongly correlated negatively with the producibility of the novel sounds, reflecting mismatch between the predicted acoustic correlates of known sounds and the incoming auditory input of the pronounced sounds. They suggested that an internal model of native sounds was formulated in the motor system and the auditory system was responsible for the comparison between the predicted and actual acoustic consequences of sounds. The mismatch represented a coding mechanism that reflected the extent of producibility of novel, heard, speech sounds (Haruno et al., 2001, Wilson and Iacoboni, 2006). Even though they did not explicitly refer to an efference copy system at work, this work

provides support for the sensorimotor integration nature of speech perception. Given the anatomical specificity and possible function of the dorsal processing stream that links posterior STG region and the dorsolateral frontal region, it seems a plausible candidate to relay an efference copy of vocal motor commands during speech production.

Previous functional imaging studies and descriptions of my study

Previous imaging studies that investigated the neural substrates of audio-vocal interactions have observed increased neural activity in the temporal regions when auditory feedback was temporally masked or altered during production. Using PET, McGuire et al (1996) found that when the subjects' speech was distorted by pitch elevation when they read aloud, there was a bilateral activation of lateral temporal cortex. In the PET study of Paus et al. (1996), subjects whispered syllables at a range of speaking rates while listening to a white masking noise that prevented perception of their own speech. As syllable rate increased, cerebral blood flow increased in several areas associated with speech production as well as in the left planum temporale and posterior perisylvian cortex. The authors interpreted this as evidence for motor-to-sensory discharge (efference copy) during speech production. Using fMRI, Hashimoto and Sakai (2003) investigated the delayed auditory feedback effect and observed significant bilateral STG activation for the contrast of delayed feedback versus normal feedback when reading aloud. They suggested that the temporal regions function as a conscious self-monitoring system to support an automatic speech production system. Adopting experimentally manipulated feedback, Fu et al. (2006) investigated the neural correlates of self-monitoring using fMRI, and found increased superior temporal activation for the condition with distorted version of their voice feedback while reading aloud relative to

the conditions of hearing their own voice undistorted and hearing another person's voice. They concluded that mismatches between expected and actual auditory feedback were associated with greater temporal activation.

However, most of the previous studies that attempted to examine the cortical response to auditory feedback alterations during vocalization utilized overt speech (read-aloud) in their production, which rendered their results somewhat difficult to interpret due to the conduction of the produced sound to the inner ear through the bones of the skull. In addition, normal feedback and altered feedback were not necessarily matched in acoustic properties (i.e., duration) and timing, which would introduce additional confounds in experiments. Most importantly, to the best of my knowledge, all previous functional imaging studies simply compared the modified feedback with normal feedback during production but did not explicitly explore the interaction between production and listening, i.e., identifying areas that are sensitive to a manipulation of auditory input during production, but not during passive listening. It is this interaction that establishes the specificity of the error signal.

To identify the neural correlates of the efference copy system that links speech perception and production, I use functional MRI which is capable of exploring the whole brain to pinpoint the regions where the interaction between the auditory processing and vocal production occurs. To this end, I implemented a factorial design in which Task (production or listening) was crossed with Feedback type (speech or noise). In the key condition, the volunteer whispers a consonant-vowel-consonant word ('ted') with their normal auditory feedback masked with a white noise burst temporally gated with their production. I hypothesize that this condition would cause a mismatch between the

effference copy and the sensory consequences of the auditory feedback, thereby generating an error signal. I use whispered speech to minimize the audible bone conducted sound during vocalization. To ensure that the production components are properly controlled for, two other listening-only conditions are yoked to the production conditions, with identical auditory feedback (noise, or clear ‘ted’) as the corresponding production trials. In order to capture the temporal coherence of activity across the whole brain, we also conduct a functional connectivity analysis using the beta-series method (Rissman et al., 2004), and examine how the coupling between different brain regions changes as a function of stimulus condition and behaviour.

Hypothesis

Based on the previous findings regarding the neurophysiology and neuroanatomy of effference copy system, we hypothesize that mismatch between the effference copy and auditory feedback would trigger an error signal which presumably involves both vocal motor regions and sensory processing regions. The relay of the effference copy from motor regions to sensory cortical regions may also involve some functional pathway that is suitable for sensorimotor integration. Given the better temporal resolution that functional connectivity analysis affords, we may observe a distributed network of regions that implicate in speech motor control, as well as regions that reflect the motor movements of articulation such as somatosensory regions. From what has been discussed, we hypothesize that:

- 1) bilateral superior temporal gyri and prefrontal regions may be activated in response to the generation and processing of an error signal;

- 2) dorsal processing pathway (pg 13, 14) may be a plausible candidate for the transmission of an efference copy;
- 3) functional connectivity analysis may reveal a large functionally coupled network including bilateral superior temporal gyri, bilateral supplementary motor areas, primary somatosensory cortex, prefrontal cortex.

Chapter 2

Methods

Participants

All subjects were recruited from a pool of potential volunteers within the Queen's community (Kingston, Ontario) who had expressed their interest in participating in psychology studies. The subjects were contacted by email with a brief description and basic criteria for participation for this study (aged 18 - 45; right-handed with English being their first language; no chronic neurological/psychological disorders; no metal in their body). Twenty-one participants (aged 20 to 32, 16 females) each received \$15 to compensate them for their time, and also received an anatomical image of their brain as an additional benefit afterwards. Written informed consent was obtained from all participants according to procedures approved by the Queen's Health Sciences Research Ethics Board.

Apparatus

My experimental paradigm entails online manipulation of auditory feedback in real-time that is temporally gated with production. A key component of my hardware is a set of National Instruments (National Instruments Corp., Austin, Texas) devices that enabled us to process and deliver my stimuli in real time. A National Instruments PXI-6052E multifunction input/output data acquisition board was mounted in a PXI-1002 chassis. This chassis also mounted a PXI-8176 signal processing computer (essentially a 1.26GHz Intel Pentium III processor) running a deterministic real-time operating system. A BNC-2110 shielded connector block with signal-labelled BNC connectors served as an I/O interface and was connected to the data acquisition board using a SH68-68-EP 68-

conductor cable for easy connectivity of my analogue input and output from microphone to headphones. I used a Phone-Or (Optoacoustic Ltd., Israel) dual-channel microphone system (Frequency response: 200-6000Hz). The NNL (NordicNeuroLab, Bergen, Norway) headphones that I used were embedded with electrostatic transducers. The results from my measurements using a sound level meter (Type 2260 by B&K, North Ryde, Australia) and headphone coupler (B&K, North Ryde, Australia) indicated that the headphones had an almost flat frequency response with no observable harmonic distortion between 100-8000Hz. Both microphone and headphones were designed specifically for fMRI experiments. The microphone and headphones were plugged into the connector block using RCA-to-BNC converters. A stimulus delivery computer (IBM ThinkPad X60; Windows XP OS; Intel Core Duo 2 Processor 2GHz) running custom software written in LabView 7.4 (National Instruments) was connected to the signal processing computer via a D-Link (D-Link System Inc., California) 10/100 fast Ethernet switch. The stimulus delivery computer and the signal processing computer were both configured into 100 full-duplex mode and communicated with each other using LabView modules that I implemented based on a previous version developed by Prof. David Purcell that was used for a different purpose. The stimulus delivery computer was considered "headquarters" and programmatically controlled the data acquisition and stimulus delivery by establishing conversations with the signal processing computer during the experiment, synchronized to MR volume acquisitions.

Experimental design

The experiment consisted of a 2x2 factorial design with four experimental conditions and a low-level "rest" control condition. The four conditions were: 1. Producing

whispered speech (the word 'ted') and hearing normal auditory feedback; 2. Producing whispered speech ('ted') and hearing voice-gated masking noise; 3. Listening to the stimuli of 1. (without production); and 4. Listening to the stimuli of 2. (without production). Whispered speech was used to minimize bone conduction (Barany, 1938) and to make sure that noise would effectively mask the speech (Houde & Jordan, 2002). All my subjects reported in the debriefing interview afterwards that they could only hear masking noise but not their own speech through the headphones when they were speaking in Condition 2. In Condition 1, the subject spoke into the microphone, and then the speech was recorded and fed back to the headphones in real time using the National Instruments deterministic signal processor mentioned above. The processing delays were short enough (iteration delay less than 1ms) that they would not be noticeable to listeners, i.e., the subject would hear exactly what was being produced. The recorded speech was sent back to the stimulus delivery computer for storage to be used for yoked trials in Condition 3. The masking noise in Condition 2 was Gaussian white noise, and temporally gated with the onset of speaking. This was realized by implementing a voice trigger as part of my program running on the real-time signal processor. The voice trigger determined the onset and offset of the voice by constantly calculating mean values and standard deviations of the data points from the microphone, and comparing them against pre-set mean and standard deviation thresholds for silence. The sensitivity of the voice trigger was dependent upon the thresholds to some degree. The thresholds set in my experiment were based on the data from pilot studies, and it turned out that they worked well. Auditory feedback in Condition 2 was recorded to be played back during trials of Condition 4. Since the auditory signal from Conditions 1 and 2 is played back exactly the

same way in Condition 3 and 4 respectively, the auditory stimuli are identical between conditions 1 and 3, and between conditions 2 and 4.

Imaging was conducted on a 3T Siemens Trio MRI system at the Queen's Centre for Neuroscience Studies MR Facility. A high-resolution anatomical image (voxel resolution of 1.0 x 1.0 x 1.0 mm, flip angle $\alpha = 9^\circ$, TR = 1769 msec, and TE = 2.6 msec) was acquired on each volunteer to facilitate intersubject averaging. A rapid sparse-imaging design was used (Orfanidou et al., 2006) for fMRI data acquisition, in which stimuli were presented and responses recorded in the silent period between successive scans, in order to minimize acoustic interference. My stimuli were presented in a 1400 msec silent gap between scans, which was long enough to record verbal responses. These silent gaps were followed by a whole-brain acquisition (T2*-weighted gradient echo EPI, TA=1600 msec, 26 slices, slice thickness 4 mm, interslice gap 0, voxel size 3.3x3.3x4.0 mm, FOV 211 mm, and repetition time TR=3000 msec) (see Figure 1).

Figure 1: The timeline for rapid-sparse imaging in my study. Each 1.4 sec stimulus delivery interval was followed by a 1.6 sec whole-brain volume acquisition. Production trials shown with a green-cross visual cue; “Listen only” trials and rest with a red-cross cue. The three trials shown in this schematic are production with normal feedback, listening to noise, and rest, respectively.

So that subjects would respond to each condition properly, I used two different visual cues: a green cross and red cross (see Figure 2). The visual cues were centered in the middle of a screen with a black background. When the subjects saw a green cross, they were expected to whisper "Ted", and when they saw a red cross, they were expected to keep silent. The green cross appeared at the onset of production trials (Condition 1 or 2), and the red cross appeared at the onset of listening trials (Condition 3 or 4), and rest trials. In other words, the green cross stands for "go" and the red cross for "no-go" in terms of vocalization. A trigger sensor in my program detected pulses from the scanner, which indicated the beginning of each fMRI scan. The stimulus delivery computer waited for 1500 msec after scan onset (i.e., 100 ms before each scan ended) and then either a green or red cross corresponding to a specific trial type was projected onto a screen that the subjects could see through a mirror attached overhead in the scanner. I presented the visual cue for the next trial 100 msec before the previous scan ended because my pilot testing based on 5 subjects revealed that it took volunteers at least 200 ms to respond to a cue, and by presenting the visual cues earlier, I could also get around the possible warm-up delay of my equipment and facilitate the real-time processing in the stimulus-delivery interval.

Figure 2: The experimental paradigm represented by visual cues. 100 ms before the onset of each trial, either a red cross or a green cross appeared on the screen. When a green cross prompted, the subject whispered ‘ted’ and heard either normal feedback (speech) or masking noise; when a red cross prompted, the subject heard either speech feedback or masking noise from previous production trials (or heard nothing, in rest trials).

The 5 conditions (4 experimental + 1 rest) were presented in a pseudo-random order so that each condition was equally likely to be followed by any of the 5 conditions. In every block of 5 trials, each condition appeared exactly once. There were 36 blocks of trials presented in a single run, which lasted about 9 minutes. I generated three different such 36-block sequences beforehand named A, B, C. (see Appendix 1, 2, 3), and each subject was assigned a random permutation of these sequences in the experiment (i.e., 3 runs for

each subject), in order to minimize systematic confounds. The stimuli presented during listening trials (Condition 3 or 4) were identical to those from the corresponding production trials (Condition 1 or 2) in the preceding block, except that for the first block in the first run, the stimuli for listening trials were from the production trials in the same block (because there was no preceding block). For the second and third run, the stimuli used for listening trials in the first block were from production trials in the last block of the previous run.

Procedure

When the subjects arrived at the imaging centre, they were required to fill out a set of forms, which screened for their eligibility to go into the scanner and gave us their consent for participation. Then I instructed subjects to respond to the green cross and whisper "Ted" in a strong and consistent manner and simply do nothing when they saw a red cross in the scanner. Considering the fact that different people may have distinct ideas about what constitutes a "whisper", I asked each subjects to perform a few practice trials outside the scanner so that I could confirm that their whisper would trigger the voice key. After the instructions, volunteers were de-metalled and settled into the scanner. They first received a 20-second localizer scan, which was followed by a 7-minute high resolution anatomical scan. Then they performed a practice run of 20 trials (4 blocks of 5 conditions in random order) that was the same as a real run except far shorter in time. This practice run gave the subjects a chance to accustom themselves to the environment of the scanner as well as provide us with their feedback on the adequacy of the sound volume over the headphones, whether they could follow the pacing of the experiment, etc. Then they were scanned in three functional runs (each with 180 trials/9 mins), with a short break between

successive runs. To catch the errors made by the subjects such as missing trials, incorrect responses, etc, I used another set of headphones to monitor in real-time how the subjects were performing, so that whenever an error occurred, I logged the trial number and removed that trial in my image analysis later. In addition, all trials were recorded for further check-up and calibration offline. When the three sessions were over, the subject exited the scanning suite and was debriefed by the experimenter, in order to gain some feedback about my experiment from the subjects. The debriefing questions are attached in Appendix 4. The subject was then told the purpose of the experiment, paid and thanked, and dismissed.

To make sure that the auditory feedback in the production trials and the corresponding listening trials was indeed the same, I also conducted an independent sound level test afterwards. The purpose of the test was to measure whether the sound volume in the headphones across all trials (except 'rest' trials) was consistent, and more importantly, whether the sound level of the playback in the listening trials was consistent with that during the corresponding production trials. The test was divided into two stages. In the first stage, normal whispered speech was used to mimic a real experiment. During the testing, an electronic artificial ear connected to a sound-level meter (Type 2260 by B&K, Denmark) was placed on one side of the headphones. The artificial ear and the headphones were placed together in a sound-proof booth to minimize the interference caused by ambient noise. However, the microphone and live voice source were located outside the sound booth to ensure that the whispered speech would not be picked up by the artificial ear. The sound pressure level (SPL) values fluctuated somewhat from trial to trial, possibly due to the latency of the sound meter. Thus, only the peak value was

recorded. I measured 35 trials, in which the SPLs of 28 trials were recorded, with the 'rest' trials omitted. The results indicated that the level on "listen only" trials was essentially identical to that on production trials (# of entries: 28, mean SPL: 45.06 dB, SD: 2.72; mean SPL of production trials: 45.46, SD: 2.59; mean SPL of listening trials: 44.65; SD: 2.99) across all trials. The procedure in the second stage of the experiment was similar to that of the first stage, except that the whispered speech was replaced by a pure tone sound level calibrator (Type 4231 by B&K) playing sounds at 1000 Hz at 94 dB, in order to eliminate the inconsistency caused by the instability of human speech. I measured 60 trials, in which there were 48 trials with non-zero feedback (i.e., 'rest' trials excluded). The results indicated that the SPLs were fairly consistent across all trials (# of entries: 48; mean SPL: 51.3 dB; SD: 0.07), as shown in Appendix 6, with no difference between listening and production trials ($t = -0.755$, $p = 0.457$), or between speech and noise trials ($t = -0.755$, $p = 0.457$).

Imaging data processing and analysis

The fMRI data processing and analyses were performed using Statistical Parametric Mapping software (SPM2, Wellcome Department of Imaging Neuroscience, London, UK). Time series diagnostics using tsdiffana (Matthew Brett, MRC CBU: <http://imaging.mri-cbu.cam.ac.uk/imaging/DataDiagnostics>) were conducted to ensure the quality of the functional data. No evidence of variance spikes or artefacts was observed in the time series for each subject.

Functional images were first realigned to the first image in the session to correct for motion. The mean of the realigned undistorted images was then coregistered with the structural T1 volume. This T1 volume was spatially normalized to a standard T1 template

(ICBM-152; standard T1-weighted normalization template in SPM2), and the same normalization parameters were subsequently applied to all the fMRI images. The normalized fMRI images were smoothed with a full-width half-maximum Gaussian kernel of 10 mm.

I modelled the 4 experimental conditions (Condition 1,2,3,4) as separate regressors, keeping the rest conditions implicit. In addition, to further model variance due to subject movement, I also included the realignment parameters in the design as user-specified regressors. I chose the hemodynamic response function (HRF) as the basis function, and downsampled the continuous hemodynamic response at each time bin. Since I used a sparse-imaging paradigm in which successive scans were separated by an interval of 1400 msec for stimulus delivery, I chose 800 msec from the onset of each scan (which was in the middle of 1600 msec acquisition) as the reference bin for specifying the event timings. I applied a high-pass filter with a cut-off of 128 sec to remove low-frequency signal components and then corrected for serial correlations with an AR(1) autoregression model. I then analyzed the contrast images across the subjects using a mixed-effects general linear model, in which I first treated conditions as a fixed effect and then treated subjects as a random effect (one sample t-test) to account for inter-subject variability. I also used the Automatic Labelling (AAL) toolbox (Tzourio-Mazoyer et al., 2002) to identify the anatomical locations of activated voxels with reference to the gross anatomy of the 'colin27' single-subject reference brain (the canonical single-subject T1-weighted image in SPM2). Voxels were deemed significant if they reached a statistical threshold of $p < 0.05$ after correction for multiple comparisons at the cluster level.

Functional Connectivity Analysis

The conventional fMRI data analysis as described in the previous section is aimed at localizing activity to brain regions in response to a specific cognitive task. However, given the limited temporal resolution afforded by fMRI acquisition compared with other imaging modalities such as EEG/MEG, the information regarding how different brain regions communicate with each other *within* a cognitive task is not accessible. More importantly, MRI signal may exhibit low-frequency fluctuations that are not associated with any task or stimulus, which would not be accounted for in the conventional analysis. However, one can not exclude the possibility that such transient signals may reflect alterations in the BOLD level induced by increased neuronal activity. To address these problems, I conducted functional connectivity analysis. “Functional connectivity” has been formally defined by Friston et al. (1993) as “measuring temporal correlation between spatially separated neurophysiological events”. To avoid confusion (Horwitz, 2003), I use the term “functional connectivity” to mean trial-by-trial (within-condition) correlation in neural activity, which can be modulated according to condition.

Traditionally, functional connectivity analysis has been geared toward blocked designs, where the mathematical correlation between the mean signals evoked by different tasks is measured (Lowe et al., 2000; Hampson et al., 2002; Sun et al., 2004). However, it is not possible to estimate what proportion of the evoked BOLD response can be attributed to each of the stages of a task when there are multiple stages within a single block (D’Esposito et al, 1999). Rissman et al. (2005) developed the beta-series method for a cognitive task that involved multiple stages in a single block, in which each stage in each

block was modeled as a unique covariate to determine how much of measured BOLD signal could be attributed to an individual stage (Postle et al., 2000; Zarahn et al., 1997).

For my event-related design with four experimental conditions, each trial can be deemed as a single ‘stage’ and the parameter estimates for all the trials associated with a particular condition can form a beta series. The rationale being that, if I assign a covariate for each individual trial in the context of general linear model, and sort the parameter estimates corresponding to each covariate based on the condition from which they were derived, then a “beta series” can be formed by concatenating all the covariates (beta values) for that condition across all trials. If I mathematically correlate different beta series from different conditions, the pattern of synchronous neural firing between conditions can be inferred. Using statistical parametric mapping, this beta series method can be easily applied to the whole brain, and statistical inference about coupling between a target voxel and every other voxel can be made.

The difference in design matrix between the modeling approach described here and the modeling used for conventional fMRI analysis is illustrated Figure 3). In the conventional analysis model, I had four conditions, each with 36 trials in a session. I modeled realignment parameters as six user-specified regressors (covariates) in each session. Therefore, the design matrix was 36×3 by $(4+6) \times 3 + 3$ (Figure 3 left, taken from one subject). In the beta series modeling, each trial that was in a column representing the same condition needed to be modelled individually as a covariate and the design matrix became 36×3 by $(36+36+36+36+6) \times 3 + 3$ (Figure 3 right). After running this GLM in SPM, 453 beta maps (beta.img) were generated, equal to the number of the covariates in the design matrix.

Figure 3: Two design matrices taken from the same subject. Left: univariate modelling; Right: beta series modelling. Two red circles correspond to the trials with the same condition in the first session. All trials from a single condition were modelled as a single covariate on the left, whereas each trial was modelled separately on the right.

The beta series method loaded all trial-specific beta images into memory and converted each 3D beta image into a 1D vector. I filled 453 such vectors into a matrix as 453 columns. Therefore this matrix was $61 \times 73 \times 85$ by 453 in dimension, with each row representing the beta series of a voxel. Since I was interested in exploring how a particular region was functionally coupled with other parts of the brain, I labelled that region as a “seed”. I first calculated the beta series of the ‘seed’ (the target region of interest; see Results) by extracting a submatrix whose rows corresponded to the rows in the matrix that were 1s and whose columns corresponded to the columns in the matrix that represented this condition. The seed beta series was obtained by averaging the columns in this submatrix, which was then used to perform correlation analysis with each row of another submatrix whose columns also represented this condition. The result of this analysis was a column vector in which each entry represented the correlation

coefficient between the ‘seed’ and the voxel that this entry stands for. This vector reflected the temporal correlation between the seed and every single voxel in the brain modulated by a specific condition. This vector was further normalized by performing arc hyperbolic tangent transformations, which resulted in the correlation coefficients being approximately normally distributed (under the null hypothesis) so that group t-tests could be conducted. I also multiplied each entry in the vector by a scalar of square root of $(N-3)$ where N represents the trial numbers associated with the condition of interest. This 1D vector was then converted back to a 3D image to reveal inter-regional connectivity according to standard statistical methods.

Chapter 3

Results

I explored several contrasts of interest, built up from the four experimental conditions: production with normal feedback (PN), production with masking noise feedback (PM), listening to normal speech (LN), and listening to masking noise (LM). These contrasts were designed to identify brain regions:

1. that are differentially recruited during speech production and perception;
2. that are engaged in different levels of acoustic processing (i.e., Speech vs Noise);
3. in which activity during speech perception is modulated during speech production.

Contrasts among conditions

Production v.s. Listening: (PN + PM) v.s. (LN + LM)

Overt speech production entails motion in the scanner (Birn et al., 1999; Huang et al., 2001), and the head movement is likely to be the greatest within 2.5s following the stimulus (Palmer et al., 2001). Therefore, this contrast might be expected to capture any motion artefact and should be interpreted with caution. Consistent with my expectation of movement artefact (Wildgruber et al., 1996; Ackermann et al., 1998; Ackermann & Riecker 2004), the contrast Production vs Listening activated an extremely large region involving both hemispheres, centered on the rolandic operculum (centered at -54, -3, 12; t-value = 17.23; 34534 voxels in the cluster; see Figure 4). Conversely, the main effect of Listening vs Production did not show significant activation.

Figure 4: Extensive areas are shown to be active in the contrast $(PM + PN) - (LM + LN)$ (shown thresholded at $p < 0.001$, uncorrected for multiple comparisons, superimposed on the SPM2 canonical single-subject anatomical image “colin27”). The numbers at the top corners of the brains indicate x coordinates in the standard space (same as other figures).

Speech v.s. Noise: (PN + LN) v.s. (PM + LM)

No voxels demonstrated signal change that passed a whole-brain corrected significance threshold. Based on apriori information from previous studies investigating activation for speech versus nonspeech sounds (see Table 1), I conducted a small volume correction, searching for elevated activity in the superior temporal region. I first calculated the centre of mass of the peak voxels from published studies (-65, -32, 3; left MTG; see Table 1), and then tested for activation in a 10-mm cube surrounding this voxel. This analysis revealed a significant focus of activation ($t=2.56$, $p<0.01$ corrected for the small volume of interest).

Authors	Contrast	Coordinates			Area	Dist
		X	Y	Z		
Uppenkamp et al. (2006)	Speech: natural & synthetic voxels v.s. Closely matched nonspeech	-66	-20	0	MTG	16.28
Scott et al. (2006)	Noise-vocoded speech v.s. Unintelligible speech	-64	-38	2	STS	3
Ashtari et al. (2004)	Phonemes v.s. Tones	-63	-36	3	MTG	3
Jancke et al. (2002)	Consonant-vowel syllables v.s. Tones + noise	-64	-32	4	STS	4.58
Giraud & Price (2001)	Words & syllables v.s. Environmental sounds + noise	-70	-38	6	STG	5.39
Vouloumanos et al. (2001)	Nonsense speech v.s. Matched sine wave sounds (complex nonspeech)	-64	-32	4	STG	4.58
Scott et al. (2000)	Speech + vocoded speech + rotated speech v.s. rotated vocoded speech	-64	-38	0	STS	4.12

Table 1: Recent fMRI studies in which perception of speech and nonspeech sounds is compared. The listed coordinates are the peak of activity observed for the contrast of speech with nonspeech sounds. Dist represents the Euclidean distance between the coordinates from previous studies and the activation peak for the contrast Speech vs Noise in the current study, in unit of *mm*. MTG: Middle temporal gyrus. STG: Superior temporal gyrus. STS: Superior temporal sulcus.

Interaction: $(PM - PN) - (LM - LN)$

This contrast reveals areas where the activation difference between production with masking noise and without is greater than the difference between listening to noise and listening to speech, i.e., areas that are expected to be most active during production with noise feedback. This is the pattern of activity expected in regions important for generating and/or processing of an error signal when the efference copy and the sensory consequences do not match. This contrast yielded significant activity in the posterior superior temporal gyrus bilaterally (see Table 2 and Figure 5).

x	y	z	t-statistics	No. of voxels	$p_{\text{corrected}}$	Anatomical areas
-66	-45	15	6.50	293	< 0.001	Left posterior superior temporal gyrus
54	-18	-6	5.88	347	< 0.001	Right superior temporal gyrus

Table 2: Activation peaks in the contrast highlighting regions in which differences between noise feedback and speech feedback are greater during production than during listening $(PM - PN) - (LM - LN)$. $p_{\text{corrected}}$: the probability value for the activation cluster, corrected for multiple comparisons at the cluster level.

Figure 5: Activation in the interaction $(PM - PN) - (LM - LN)$ (shown thresholded at $p < 0.001$, uncorrected for multiple comparisons).

Interaction: $(PN - PM) - (LN - LM)$

This contrast reveals areas where the activation difference between speech feedback and noise feedback during production is greater than the difference between speech feedback and noise feedback during listening. This revealed significant activation in the left rolandic operculum, left superior parietal lobule, right middle occipital gyri, and right fusiform gyrus (see Table 3 and Figure 6).

x	y	z	t-statistics	No. of voxels	$\rho_{\text{corrected}}$	Anatomical areas
-39	-18	18	6.26	91	0.009	Left rolandic operculum
36	-84	18	6.02	180	< 0.001	Right middle occipital gyrus
30	-54	-9	5.02	70	0.026	Right fusiform gyrus
-18	-63	54	4.85	73	0.022	Left superior parietal lobule

Table 3: The areas where the activation difference between speech feedback and noise feedback during production is greater than the difference between speech feedback and noise feedback during listening $(PN - PM) - (LN - LM)$.

Figure 6: Activation revealed by the contrast (PN – PM) – (LN – LM) (shown thresholded at $p < 0.001$, uncorrected for multiple comparisons).

Given the significant interactions, I explored the component simple effects:

Production with masked feedback (PM) v.s. Production with normal feedback (PN):

Production with masking noise feedback compared to production with speech feedback showed increased activation in the superior temporal lobe bilaterally (see Table 4 and Figure 7). In the left hemisphere, there was a significant cluster (see Table 4) in the posterior superior temporal gyrus, extending dorsolaterally into the middle temporal gyrus and posteriorly into the supramarginal gyrus. In the right hemisphere, there was one cluster in the superior temporal gyrus. The activation peak (left STG) in this contrast (PM – PN) was very close to that in the contrast (PM – PN) – (LM – LN) (Euclidean distance: 4.2mm). This is not surprising since the contrast (PM – PN) is one component simple effect for the interaction (PM – PN) – (LM – LN). When the condition PN was evaluated against the condition PM (PN – PM), no voxel passed the cluster-wise significance threshold.

x	y	z	t-statistics	No. of voxels	$p_{\text{corrected}}$	Anatomical areas
-63	-45	18	5.97	248	< 0.001	Left superior temporal gyrus
51	-54	21	5.62	118	0.006	Right superior temporal gyrus

Table 4: The areas that show increased activation for production with masking noise feedback relative to production with speech feedback (PM - PN).

Figure 7: Areas activated more in PM than in PN (PM – PN) (shown thresholded at $p < 0.001$, uncorrected for multiple comparisons).

Listening with masked feedback (LM) v.s. Listening with normal feedback (LN):

(LN – LM) highlights areas that are more responsive to speech than masking noise during passive listening (see Table 5 and Figure 8). The activation was strongly

lateralized to the left hemisphere, with a peak cluster localized to the left middle temporal gyrus extending into mid-STS, and a cluster localized to the right anterior superior temporal gyrus. This would be expected given the results of the repeated-measures ANOVAs on signal from these regions in the contrast (PM – PN) – (LM – LN) (see Figure 10 and 11). The contrast (LM – LN) yielded extensive activation in the superior temporal gyrus and middle occipital gyrus in both hemispheres, and in the left supplementary motor area, left inferior temporal gyrus, left putamen, left cingulate gyrus, and right rolandic operculum (see Table 6 and Figure 9).

x	y	z	t-statistics	No. of voxels	$\rho_{\text{corrected}}$	Anatomical areas
60	-15	-6	4.88	138	0.009	Right superior temporal gyrus
-66	-36	3	4.73	162	0.004	Left middle temporal gyrus

Table 5: The areas that are more activated by passive listening to speech than noise (LN – LM).

x	y	z	t-statistics	No. of voxels	$\rho_{\text{corrected}}$	Anatomical areas
21	-72	54	8.36	71	0.017	Right superior parietal lobule
33	-84	3	7.88	285	< 0.001	Right middle occipital gyrus
18	6	15	7.84	1213	< 0.001	Left putamen
42	-30	21	7.22	952	< 0.001	Right rolandic operculum
48	-57	-24	6.84	232	< 0.001	Right inferior temporal gyrus
-39	-84	0	5.68	684	< 0.001	Left middle occipital gyrus
-18	-63	54	5.68	104	0.003	Left superior parietal lobule
-9	-12	72	5.21	60	0.033	Left supplementary motor area
-15	12	36	4.77	56	0.042	Left cingulate gyrus

Table 6: The areas that are more activated by passive listening to noise than speech (LM – LN).

Figure 8: Areas of activity revealed by the contrast (LN - LM) (shown thresholded at $p < 0.001$, uncorrected for multiple comparisons).

Figure 9: Areas of activity revealed by the contrast (LM - LN) (shown thresholded $p < 0.001$, uncorrected for multiple comparisons).

Analysis of the interaction (PM – PN) – (LM – LN) in the bilateral STG

In order to confirm that the significant interaction that I observed in bilateral superior temporal cortex did result from PM being significantly greater than PN and LN being significantly greater than LM (crossover interaction), I conducted the following ROI analysis around the left and right STG peaks in the interaction contrast.

In accordance with the 10-mm FWHM Gaussian smoothing kernel, I defined a region of interest (ROI) as a 10-mm wide box centered on each of these two activation peaks to better reflect the spatial correlation of adjacent voxels smoothed by the kernel. The beta parameter estimates within the defined ROI were extracted in individual subjects. Repeated measures ANOVAs using SPSS on these data were performed in order to evaluate how differences among conditions produced this significant interaction. The analysis yielded significant crossover interactions among conditions in both of these regions ($F(1,20) \geq 34.31$, $p < 0.001$) (see Figures 10 and 11). Post-hoc comparisons using the Bonferroni adjustment for multiple comparisons revealed that activation for PM was significantly greater than for PN in both areas ($F(1,20) \geq 6.68$, $p < 0.02$), whereas activity was significantly greater in LN than in LM in both regions ($F(1,20) \geq 12.84$, $p < 0.002$). This crossover interaction perhaps explains why the main effect of Speech vs Noise [(PN+LN) – (PM+LM)] was only significant after a small volume correction in the left STG.

Figure 10: The parameter estimates (betas) based on the ROI data from 21 subjects within a 10-mm cube centered on the activation peak in left STG (-66, -45, 15). The estimates indicate that signal in production with masking noise feedback (PM) is significantly greater than in production with normal feedback (PN) and that signal in listening to normal speech (LN) is significantly greater than in listening to masking noise (LM); see Results. The arbitrary units refer to the parameter estimates that compare the predicted and actual signal change. The error bars represent 95% confidence intervals for the mean. Asterisk indicates difference at the significant level of $p < 0.05$.

Figure 11: The parameter estimates (betas) based on the ROI data from 21 subjects within a 10-mm cube centered on the activation peak in right STG (54, -18, -6). The estimates indicate that signal in production with masking noise feedback (PM) is significantly greater than in normal feedback (PN) and that signal in listening to normal speech (LN) is significantly greater than in listening to masking noise (PM); see Results. The arbitrary units refer to the parameter estimates that compare the predicted and actual signal change. The error bars represent 95% confidence intervals for the mean. Asterisk indicates difference at the significant level of $p < 0.05$.

Functional Connectivity Analysis

In my functional connectivity analysis (see Methods), I attempted to examine how activity in the left STG, the activation peak in the interaction contrast (PM – PN) - (LM - LN), correlated with activity in other regions. This left STG peak “seed” (-66, -45, 15) was defined as the reference voxel and I explored how the beta series in each voxel in the brain was correlated with the beta series of this seed voxel, condition by condition. A commonly used method for identifying functionally connected regions in the brain involves defining a cluster of activated voxels around the seed voxel in a region of interest (ROI) (Cordes et al., 2000; Della-Maggiore et al., 2000; Lowe et al., 2000; Ranganath et al., 2003), thus I defined a 10-mm box ROI (to be consistent with the smoothing kernel applied to my raw functional data) centered at the location of the seed voxel as the seed ROI. Since I have spatially normalized all the functional images in my dataset, single-subject data can be expected to be roughly homologous across subjects and selecting a seed ROI based on group data is reasonable.

This analysis identifies regions in which activity (scan by scan) can be predicted from activity in the seed voxel, highlighting regions in a functionally connected network. Each experimental condition was assessed individually, leading to four correlation maps for each subject. The correlation maps for each condition were then entered into group analyses in which t-values for each voxel were computed across subjects, treating inter-subject variability as a random effect. Since the t-value that is equivalent to the significance level $p < 0.001$ uncorrected for multiple comparisons is 3.55, each beta image from the group analyses was thresholded at $t = 3.55$; this resulted in four binarized images in which a ‘1’ indicating reliable correlation across subjects with the seed ROI. The areas

that exhibited coupling with the seed ROI across all four conditions were identified using an algebraic conjunction operation performed on the four binary images (one for each condition). Autocorrelation in the seed voxel, and in regions surrounding it was observed due to spatial smoothing. The autocorrelation is not helpful to show, since it simply reflects the seed voxel's correlation with itself and adjacent areas. According to SPM, the estimated intrinsic smoothness of my processed data was 16 mm full width at half maximum (FWHM), which is given by the distance between the two extreme values of an independent variable that yields half of a function's maximum value. Since the smoothing kernel I applied was Gaussian, the FWHM of a Gaussian function equals $2.355 \times \text{sd}$ (standard deviation). I assumed, conservatively, that the autocorrelation would be present up to 3 sd away from the seed voxel, because 3 sd of a Gaussian function covers 99.7% of the area under the curve; this equals $3/2.355 = 1.27$ times FWHM. Therefore, in order to remove areas in the region of the seed voxel that exhibited connectivity probably due entirely to the spatial smoothing (autocorrelation), I masked my analysis with a spherical mask with a radius of $1.27 \times \text{FWHM} = 20$ mm centered on the seed voxel.

The connectivity procedure yields only an image of clusters of correlation, coordinates indexing the centre of mass of the clusters are not provided. Consequently, I used MarsBar (Brett et al., 2002) to derive the number of clusters in the image, the centre of mass of each cluster, and the size of each cluster in voxels. Since I have spatially normalized the data using a voxel size of $3 \times 3 \times 3$, the number of voxels in each cluster can be calculated as number of the volume/ $3 \times 3 \times 3$. The anatomical label for each cluster was also determined using SPM toolbox "Anatomy" (Eickhoff et al., 2005). The areas that passed the significance threshold in all four conditions (i.e., are present after conjunction

analysis) are shown in Figure 12, and listed in Table 7. These include: right superior temporal gyrus (i.e., homologous region to the seed voxel in the opposite hemisphere), probably due to the well-established interhemispheric connections between homologous regions; bilateral superior temporal gyri, bilateral supplementary motor areas, bilateral lingual gyri, left middle temporal gyrus, left inferior temporal gyrus, left supramarginal gyrus, left rolandic operculum, left postcentral gyrus, right inferior frontal gyrus, right temporal pole, and right precuneus.

In addition, I also attempted to examine how functional coupling between the seed ROI and other regions varied with changing stimulus and task; i.e., modulated functional connectivity. I took the four correlation maps for each subject as if they were individual “functional images”, and specified contrasts among all pairs of conditions, testing these using paired t-tests, treating subjects as a random effect. However, I failed to discover any significant differences among conditions in functional connectivity between the seed region and other brain regions.

Centre	Volume (mm)	Voxel	Anatomical Label
56 -32 9	8829	981	Right superior temporal gyrus
-54 -29 6	2808	312	Left superior temporal gyrus
4 8 53	594	66	Right supplementary motor area
-53 -62 9	378	42	Left middle temporal gyrus
53 12 25	243	27	Right inferior frontal gyrus
-56 -65 -8	162	18	Left inferior temporal gyrus
53 17 -8	108	12	Right temporal pole
-18 -63 -6	81	9	Left lingual gyrus
22 -54 -7	81	9	Right lingual gyrus
9 -56 57	54	6	Right precuneus
-6 -3 60	27	3	Left supplementary motor area
-48 -45 24	27	3	Left supramarginal gyrus
-60 3 12	27	3	Left rolandic operculum
-54 -6 39	27	3	Left postcentral gyrus

Table 7: This table shows the anatomical regions that exhibited coupling with the seed ROI centered at (-66, -45, 15) across all four conditions. The centres, volumes and number of voxels for these areas were also identified.

Figure 12: Areas functionally connected with the seed voxel (-66, -45, 15), with the region of autocorrelation around the seed voxel masked out. See Table 7 for anatomical labels for these regions.

Chapter 4

Discussion

The current study investigated the neural mechanism of the efference copy system using an event-related fMRI paradigm, by artificially masking the auditory feedback with voice-gated noise during speech production. I observed an enhanced activation in the bilateral superior temporal gyri (STG) when I compared noise feedback with normal speech feedback during production relative to when I compared listening to speech with listening to noise, representing a significant interaction between Task and Feedback type. Repeated-measures ANOVAs with post-hoc comparisons in both activated areas indicated that activity was significantly greater for noise feedback than for speech feedback during production but that these same regions were more sensitive to speech than to noise during passive listening. This result suggests that cortical activity in speech-sensitive regions of the superior temporal gyrus is significantly modified during speech production when the vocal output and the sensory consequences of the auditory feedback do not match. In contrast, comparing speech feedback with noise feedback during production alone did not yield any significant activation in the brain.

The interaction assessed in my study yields an error signal

The interaction between Task and Feedback type may indicate an efference copy system at work. I hypothesize that the elevated activity in temporal regions for gated noise during noise-masked production is due to an error signal being generated and processed in these regions. In accordance with a previous EEG study (Ford et al., 2005) that utilized distorted feedback and explicitly examined the interaction between Talk/Listening and Distortion, I observed an elevated response to manipulated feedback

in auditory cortical areas only during production, but not during passive listening. This indicates that the brain regions responsible for vocal production may have modulated the neural activity in the auditory cortex through an efference copy. However, Ford et al. (2005) employed EEG as their data acquisition paradigm, and the capability of this method is limited by the spatial distribution of electrodes and frequency bands sampled. They only assessed the interaction at four frequency bands (30 Hz, 35 Hz, 40 Hz, and 45 Hz; all in the Gamma range) and could not precisely locate where the interaction may have occurred. Nevertheless, their finding that the coherence of EEG in all four bands between the frontal lobes and temporal lobes was disrupted by distorted feedback is compatible with my results, since incoherence in EEG activity might manifest as increased signal in my fMRI experiment.

The error signal is attributable to self-monitoring

The presence of an error signal when auditory feedback was masked during production provides support for the idea of a self-monitoring mechanism (Bays et al., 2006; Blakemore et al., 2000; Matsuzawa et al., 2005; Sommer and Wurtz, 2006). This mechanism is thought to comprise a forward model that outputs an internally generated predictive signal to the auditory cortex during vocal motor action and captures the error signal to facilitate online adjustment of motor commands. The anatomical location related to generation and processing of an error signal that I report is consistent with the results of a number of previous functional imaging studies (Christoffels et al., 2007; Fu et al., 2006; Hashimoto and Sakai, 2003; McGuire et al., 1996; Toyomura et al., 2007; Wilson and Iacoboni, 2006) (see Table 8). In an experiment that involved delayed auditory feedback, Hashimoto and Sakai (2003) observed bilateral STG activation when delayed

feedback was compared with normal feedback while reading sentences aloud. The authors compared different rates of feedback to exclude the possibility that the activation patterns had been induced by different speech rates and enhanced attention. Since conscious self-monitoring is explicitly required (and critical to good performance) under the delayed auditory feedback condition, the anatomical proximity of their activation clusters to mine in the bilateral STG indicates that the self-monitoring may also be responsible for the appearance of the error signal in my study (see Table 8). Fu et al. (2006), McGuire et al. (1996), and Toyomura et al. (2007) have all reported bilateral STG activation in response to a pitch shift in auditory feedback compared with normal feedback during reading aloud. Interestingly, their activation clusters were more anterior than mine in the temporal cortex, possibly near regions that have been implicated in semantic and phonological processing of speech (Belin et al., 2000; Binder et al., 2000; Majerus et al., 2005). This may be due, in part, to the fact that shifted auditory feedback still contains speech-like acoustic properties, which, unlike the masking noise that I used, may have induced activity in areas associated with speech processing. Therefore, even though the authors of the studies using pitch-shifted feedback have speculated that the self-monitoring was the underlying mechanism for the activation patterns, the extent to which this statement is valid can not be properly assessed. Christoffels et al. (2007) adopted a method involving overt picture-naming and pink masking noise to examine performance monitoring and speech-motor control. They emphasized the importance of the motor system in regulating fluent speech production with verbal feedback. Their experiment was somewhat flawed (e.g., unmatched auditory stimuli, extensive visual input, no comparison between production and matched listen-only conditions); nevertheless, when they compared pink

masking noise feedback with speech feedback during overt picture-naming, they observed bilateral STG activation. One feature shared by my study and these previous studies is that we all manipulated auditory feedback in a way that caused it to be different from what was expected. The common results across all these studies suggest that the self-monitoring mechanism may generate an error signal during self-vocalization whenever a mismatch between the predicted and actual sensory consequences arises.

	Contrast	Left	DistA	Right	DistB
Our study	(PM – PN) – (LM – LN)	-66 -45 15		54 -18 -6	
Hashimoto&Sakai (03)	Delay > Normal	-60 -30 21	17	60 -18 3	11
Fu et al. (06)	Distorted > Undistorted	-57 -17 20	30	47 10 -2	29
McGuire et al. (96)	Distorted > Undistorted	-52 -36 16	17	46 -20 4	13
Toyomura et al. (07)	Pitch-shifted > Normal			52 -10 -2	9
Christoffels et al. (06)	Noise > Voice	-41 -28 7	31	51 -18 8	14
Wilson&Iacoboni (06)	Producibility	-52 -46 14	14	52 -34 8	21
Wong et al. (06)	Unsuccessful > Successful			54 -28 -8	10
Belin et al. (00)	Speech-sensitive regions	-62 -40 10	8	52 -19 -1	5

Table 8: This table includes functional imaging studies that are discussed in this section. The activation peaks in these studies are compared with the interaction peaks in my study. DistA (DistB) represents the Euclidean distance of the left (right) peak from ipsilateral peaks in my study.

Disruption of self-monitoring in patients with temporal lobe deficits

Patients with schizophrenia often exhibit auditory hallucinations (Barta et al., 1990; McGuire et al., 1993; Silbersweig et al., 1995). Such hallucinations may result from a lack of awareness of self-generated sounds, and misattribution of sound sources (Bentall and Slade, 1985; Baker and Morrison, 1998). Indeed Allen et al. (2004) and Johns et al. (2001) have found that when auditory verbal feedback is modified while patients with auditory hallucinations speak aloud, they were significantly ($t=3.32$ $df=26$, $p=0.003$; Allen et al., 2004) more likely to misidentify their distorted voice as someone else's than patients without hallucinations. Some authors have proposed that dysfunction in temporal-lobe regions disrupts self-monitoring and the efference copy system (Frith, 1987; Keefe et al. 1999; Ford and Mathalon, 2005); and that this is the source of a self-monitoring impairment in schizophrenic patients and a contributor to auditory hallucinations (McGuire et al., 1995; Silbersweig et al., 1995). My imaging results, showing significant temporal-lobe activity during a condition that could be expected to generate an error signal in such an efference copy system is consistent with this idea.

Modulation of auditory cortical activity due to top-down influences

Davis and Johnsrude (2007) have suggested that some top-down influences arise from the interface between audition and speech perception. Here, “top-down” refers to information that is not in the external input, but has internal neural representations, gained through knowledge and experience, as its source. The fact that auditory cortical activity was modulated during speech production may be an instance of high-level top-down influences (in this case from regions supporting speech production) on lower-level perceptual processes. Davis et al. (submitted) have shown that increased sedation caused

a significant decline in the responses of prefrontal and premotor regions to speech stimuli, which was accompanied by a moderate decline in activity in temporal lobe responses to speech. They concluded that “changes in frontal lobe responses can modulate temporal lobe responses to speech through top-down processes on lower-level auditory processes” (Davis and Johnsrude, 2007, pg 10). Wilson and Iacoboni (2006) correlated signal change in the bilateral STG with the producibility of novel, non-English speech sounds presented to English speakers. They hypothesized that auditory areas that are responsible for the transformation of the acoustic signal to a phonetic code would respond differently to novel (not producible) and native (producible) sounds, since an internal representation for the novel sound cannot be readily derived, unlike for the native sound. The authors reasoned that the strength of the internal representation, and hence the degree of mismatch between the predicted acoustic consequence based on the internal representation and the actual incoming auditory input, is reflected in the producibility of the novel sound: not very producible sounds would have minimal representation and thus a match with the incoming auditory input would be difficult to obtain, whereas producible sounds would have stronger representations and a match would be readily derived. This indicates that the role of the motor system in speech perception may be to generate “top-down” internal representations under consideration (Wilson and Iacoboni, 2006). They identified a cluster of voxels located in the left posterior STG in which signal was most strongly negatively correlated with the producibility of novel sounds; this cluster is anatomically close to the activation peak in the left STG yielding activity for the interaction contrast in my study (see Table 8). In a similar study, Wong et al. (2006) investigated the neural basis of second-language learning, by comparing the brain activity

between successful and less successful learners of a pitch language to which they had not previously been exposed. After training, they found that unsuccessful learners demonstrated increased activation in the right superior temporal region and several frontal regions including right inferior frontal gyrus, prefrontal and medial frontal areas. The underlying mechanism for this network of activation might be similar to that of Wilson and Iacoboni (2006); unsuccessful learners might be characterized by their failure to produce novel sounds. Since producibility necessarily involves the motor system, the similarity in regions for these two findings suggests that the motor system may be engaged in generating internal auditory representations, and that the mismatch between the internal templates and the incoming auditory input is coded as an error signal detectable as increased BOLD signal in fMRI. In other words, the more difficult the producibility, the greater the mismatch between the stored internal representation and the sensory input, the stronger the error signal might be.

The network involved in speech production and perception

The functional connectivity analysis presents evidence for the neural circuits underlying perception and production. The areas that exhibit reliable functional coupling with the seed region in all four conditions include speech-sensitive regions bilaterally in the temporal lobes (Belin et al., 2000; Binder et al., 1997, 2000; Davis and Johnsrude, 2003; Scott et al., 2000), and also motor areas covering bilateral supplementary motor area, left postcentral gyrus, left rolandic operculum, and right inferior frontal gyrus (Watkins et al., 2003; Wilson et al., 2004), as well as bilateral lingual gyri and left supramarginal gyrus, which have been implicated in speech motor control (Ackermann et al., 1998; Mathiak et al., 2002; Wildgruber et al., 2001). Since the beta series method was

employed to model the BOLD signal within each trial, it can be expected to be sensitive to changes in functional connectivity between different regions on a trial-by-trial basis. Consistent with the motor theory of speech perception (Lieberman et al., 1967; Liberman and Mattingly, 1985; 1989; Liberman and Whalen, 2000), not only the conditions that explicitly involve vocal production activated a large number of motor areas, but the conditions that only involve passive listening, indicating that speech perception might engage the motor system in a process of auditory-articulatory mapping to access a common phonetic code (Wilson et al., 2004). The temporal coherence between the auditory cortical areas and the motor areas during both vocalization and listening implies that these regions may be part of a large interconnected, functionally coupled network that subserves speech perception and production. However, I did not observe functional connectivity to be modulated by changing stimulus and task, presumably because the same neural circuitry was involved in all four conditions.

The error signal emerges in speech-sensitive regions

The locations that I observe to be active in the bilateral STG for the interaction contrast, and which are more sensitive to speech than noise during passive listening, have been implicated many times in studies examining auditory and speech perception. The left posterior STG has been identified to be selectively responsive to speech sounds, but not to simpler sounds with less variation in physical features such as pitch, spectrum or amplitude envelope (Lieberman et al., 1967; Belin et al., 2000; Benson et al., 2001; Crinion et al., 2003; Giraud et al., 2004; Vouloumanos et al., 2001; Scott et al., 2000). This region is also activated more strongly by speech than by amplitude-modulated noise (Mummery et al., 1999; Zatorre et al., 1992), sine-wave analogues (Dehaene-Lambertz et

al., 2005; Vouloumanos et al. 2001), or tone sequences (Demonet et al., 1992, Joanisse and Gati, 2003). The right STG has been suggested to be specialized for spectral processing (Zatorre et al., 2002; Jamison et al., 2006), and sensitive to auditory stimuli with dynamic pitch variation or prosodic features (Weintraub et al., 1981; Zatorre et al., 1992, Scott et al., 2000), as in the perception of speech prosody and intonation (Rimol et al., 2005; Scott and Wise, 2004). My observation of activation in the left middle temporal region for passive listening to speech versus noise is consistent with the findings presented in (Binder et al., 1997; 2000; Rimol et al., 2005), which have identified the mid-portion of the STG as preferentially responding to speech sounds versus simple non-speech sounds. The activation pattern may represent a transition in processing from simple acoustic features to more intelligible speech along the anterior-posterior axis in the temporal lobe, thus supporting a hierarchical processing model (Davis and Johnsrude, 2003; Scott et al., 2000; Scott and Johnsrude, 2003; Uppenkamp et al., 2006). Speech vs Noise also revealed a significant focus of activation at the left MTG, at least after small-volume correction for multiple comparisons. That this contrast was not significant at the whole-brain level was probably because of the significant interaction between Task and Feedback condition. In fact, this interaction indicates that it is exactly those STG regions that are involved in processing of speech that also participate in the error signal network.

Sound level equated across conditions

It has been suggested that adding masking noise to a speaker's auditory feedback may result in an elevated voice volume (Lane and Tranel, 1971). This also raises the question of whether subjects might have increased their volume during production in the presence of masking noise, which may give rise to systematic confounds and thereby affect the

results. This does not appear to have been the case. First of all, my stimuli were yoked in that the auditory stimuli presented during passive listening conditions were identical to those in the production conditions. An inspection of the recordings of both production and listening trials also revealed that a uniform intensity of sound level was achieved across all trials ('rest' excluded). This is not surprising since Burnett and her colleagues (Burnett et al., 1998) found that when the subjects were instructed to phonate /a/ for 5 s, neither the absolute level of feedback intensity nor the presence of pink masking noise significantly affected magnitude or latency of the voice F0 response. The reason for this is that the duration of production was not long enough to accommodate such a compensatory effect. Therefore, it is highly unlikely that the subjects in my study elevated the volume of their production of a single word "ted" in an interval less than 1 s. In addition, Belin et al. (Belin et al., 2000) found that the activation of voice-sensitive regions that were localized to the upper bank of the post superior temporal sulcus (STS) on the left and middle temporal gyrus on the right were not sensitive to the amplitude envelopes of auditory stimuli. Considering the anatomical proximity between their voice-sensitive regions and my bilateral peaks (see Table 8), I argue that the effect of elevation of the subjects' volume during masking noise trials, even if it had been present, would have been negligible.

Discussion of the results of simple effects

It is interesting to note that when speech feedback was contrasted with noise feedback during production (PN – PM), no activation was observed in the superior temporal gyrus and adjacent superior temporal sulcus, which is not consistent with previous imaging studies (Christoffels et al., 2006; Fu et al., 2006; McGuire et al., 1996). It has been

suggested that there are specific regions in STG and STS that are sensitive to voices (Belin et al., 2000). Therefore, the condition PN should have activated temporal regions more strongly than the condition PM. I speculate that the reason why there was no salient activation in the (PN – PM) contrast in these regions is because a majority of auditory neurons was inhibited during production, and therefore not responsive to the auditory input. Eliades and Wang (Eliades and Wang, 2003) have shown that vocalization-induced suppression of auditory cells may begin several hundred milliseconds prior to the onset of production and last about 1.6 seconds after the onset. Since each production trial in my study was less than 1 second in duration, the auditory cells might have still been in a state of suppression by the time the vocalization ended.

Comparing passive listening to noise to passive listening to speech (LM – LN) activated a network of regions including temporal, parietal and occipital lobe, as well as supplementary motor area. These regions have been associated with increased working memory and attentional efforts (Friederici et al., 2003; Newman et al., 2001; Wong et al., 2006). When the subjects heard masking noise that they had not expected in the condition LM, they might have launched processing mechanisms seeking a mapping from the noise to “something”, which placed an increased demand on the gestural speech processing and lexical access system. On the contrary, when the subjects heard the word “ted” played back in the condition LN, this stimulus would be clearly and straightforwardly recognizable, thereby reducing the demand on the system responsible for searching the stored representations to match the signal’s acoustics.

Conclusions

In summary, my findings provide neuroimaging evidence for the network underlying the efference copy system. I have observed that, in the bilateral superior temporal gyri, greater activity was elicited during speech production when the auditory concomitants of subjects' own speech were masked with noise, compared with when they were heard clearly. In contrast, greater activity was observed in the left MTG/STG when listening to speech compared to noise, indicating that this is a speech-sensitive region. Functional connectivity analysis reveals links between this left MTG/STG region and putative speech-sensitive regions in the bilateral temporal lobes and extensive motor areas including bilateral supplementary motor areas, right inferior frontal lobe, and left rolandic operculum in both production and listening conditions. These functional links make it possible for motor speech areas interact with auditory regions in the temporal lobe to generate and process an error signal during speech production. The existence of the efference copy system is significant in two ways. First, it subserves the online control of speech production, ensuring its flexibility and adaptability to changing conditions. This is realized through an internal model of the sensory consequences of speech acts, which guides ongoing motor planning through its comparison with incoming auditory data. When the efference copy and the auditory consequences of vocalization do not match, an error signal is generated which serves to tune ongoing production. Second, the efference copy system subserves our ability to distinguish our own speech from other peoples, ensuring normal human communication. When we speak aloud, we are aware of the intention to speak and the actual act of speaking, and we also receive the feedback of hearing our voice. However, the actual sensory consequences resulted from our own

vocalization can be predicted by the efference copy from the motor regions that issued the commands to vocalize. This report concerning the neural mechanism of the efference copy has provided evidence for the sensorimotor process that directly links speech perception and production, and I anticipate that this will be of value to both basic science and clinical communities.

References

- Aboitiz, F., & Garcia, R. (1997). The anatomy of language revisited. *Biol Res*, 30(4), 171-183.
- Ackermann, H., & Riecker, A. (2004). The contribution of the insula to motor aspects of speech production: A review and a hypothesis. *Brain Lang*, 89(2), 320-328.
- Ackermann, H., Wildgruber, D., Daum, I., & Grodd, W. (1998). Does the cerebellum contribute to cognitive aspects of speech production? A functional magnetic resonance imaging (fMRI) study in humans. *Neurosci Lett*, 247(2-3), 187-190.
- Allen, P. P., Johns, L. C., Fu, C. H., Broome, M. R., Vythelingum, G. N., & McGuire, P. K. (2004). Misattribution of external speech in patients with hallucinations and delusions. *Schizophr Res*, 69(2-3), 277 - 287
- Ashtari, M., Lencz, T., Zuffante, P., Bilder, R., Clarke, T., Diamond, A., et al. (2004). Left middle temporal gyrus activation during a phonemic discrimination task. *Neuroreport*, 15(3), 389-393.
- Baker, C. A., & Morrison, A. P. (1998). Cognitive processes in auditory hallucinations: Attributional biases and metacognition. *Psychol Med*, 28(5), 1199-1208.
- Barta, P. E., Pearlson, G. D., Powers, R. E., Richards, S. S., & Tune, L. E. (1990). Auditory hallucinations and smaller superior temporal gyral volume in schizophrenia. *Am J Psychiatry*, 147(11), 1457-1462.
- Bays, P. M., Flanagan, J. R., & Wolpert, D. M. (2006). Attenuation of self-generated tactile sensations is predictive, not postdictive. *PLoS Biol*, 4(2), e28.

- Belin, P., & Zatorre, R. J. (2000). 'What', 'where' and 'how' in auditory cortex. *Nat Neurosci*, 3(10), 965-966.
- Bell, C. C. (2001). Memory-based expectations in electrosensory systems. *Curr Opin Neurobiol*, 11(4), 481-487.
- Benson, R. R., Whalen, D. H., Richardson, M., Swainson, B., Clark, V. P., Lai, S., et al. (2001). Parametrically dissociating speech and nonspeech perception in the brain using fMRI. *Brain Lang*, 78(3), 364-396.
- Bentall, R. P., & Slade, P. D. (1985). Reality testing and auditory hallucinations: A signal detection analysis. *Br J Clin Psychol*, 24 (Pt 3), 159-169.
- Binder, J. R. (1997). Neuroanatomy of language processing studied with functional mri. *Clin Neurosci*, 4(2), 87-94.
- Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., et al. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cereb Cortex*, 10(5), 512-528.
- Binder, J. R., Rao, S. M., Hammeke, T. A., Yetkin, F. Z., Jesmanowicz, A., Bandettini, P. A., et al. (1994). Functional magnetic resonance imaging of human auditory cortex. *Ann Neurol*, 35(6), 662-672.
- Birn, R. M., Bandettini, P. A., Cox, R. W., & Shaker, R. (1999). Event-related fMRI of tasks involving brief motion. *Hum Brain Mapp*, 7(2), 106-114.
- Black, J. W. (1951). The effect of delayed side-tone upon vocal rate and intensity. *J Speech Disord*, 16(1), 56-60.
- Blakemore, S. J., Smith, J., Steel, R., Johnstone, C. E., & Frith, C. D. (2000). The perception of self-produced sensory stimuli in patients with auditory

- hallucinations and passivity experiences: Evidence for a breakdown in self-monitoring. *Psychol Med*, 30(5), 1131-1139.
- Blakemore, S. J., Wolpert, D. M., & Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nat Neurosci*, 1(7), 635-640.
- Brainard, M. S., & Doupe, A. J. (2000) Auditory feedback in learning and maintenance of vocal behaviour. *Nature*, 1, 31-40.
- Buchsbaum, B. R., Olsen, R. K., Koch, P. F., Kohn, P., Kippenhan, J. S., & Berman, K. F. (2005). Reading, hearing, and the planum temporale. *Neuroimage*, 24(2), 444-454.
- Burnett, T. A., Freedland, M. B., Larson, C. R., & Hain, T. C. (1998). Voice f0 responses to manipulations in pitch feedback. *J Acoust Soc Am*, 103(6), 3153-3161.
- Burnett, T. A., Senner, J. E., & Larson, C. R. (1997). Voice f0 responses to pitch-shifted auditory feedback: A preliminary study. *J Voice*, 11(2), 202-211.
- Christoffels, I. K., Formisano, E., & Schiller, N. O. (2007). Neural correlates of verbal feedback processing: An fMRI study employing overt speech. *Hum Brain Mapp*, 28(9), 868-879.
- Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *J Acoust Soc Am*, 116(6), 3647-3658.
- Cordes, D., Haughton, V. M., Arfanakis, K., Wendt, G. J., Turski, P. A., Moritz, C. H., et al. (2000). Mapping functionally related regions of brain with functional connectivity MR imaging. *AJNR Am J Neuroradiol*, 21(9), 1636-1644.
- Cowie, R., Douglas-Cowie, E., & Kerr, A. G. (1982). A study of speech deterioration in post-lingually deafened adults. *J Laryngol Otol*, 96(2), 101-112.

- Creutzfeldt, O., & Ojemann, G. (1989). Neuronal activity in the human lateral temporal lobe. III. Activity changes during music. *Exp Brain Res*, 77(3), 490-498.
- Creutzfeldt, O., Ojemann, G., & Lettich, E. (1989a). Neuronal activity in the human lateral temporal lobe. I. Responses to speech. *Exp Brain Res*, 77(3), 451-475.
- Creutzfeldt, O., Ojemann, G., & Lettich, E. (1989b). Neuronal activity in the human lateral temporal lobe. II. Responses to the subjects own voice. *Exp Brain Res*, 77(3), 476-489.
- Crinion, J. T., Lambon-Ralph, M. A., Warburton, E. A., Howard, D., & Wise, R. J. (2003). Temporal lobe regions engaged during normal speech comprehension. *Brain*, 126(Pt 5), 1193-1201.
- Curio, G., Neuloh, G., Numminen, J., Jousmaki, V., & Hari, R. (2000). Speaking modifies voice-evoked activity in the human auditory cortex. *Hum Brain Mapp*, 9(4), 183-191.
- D'Esposito, M., Zarahn, E., Aguirre, G. K., & Rypma, B. (1999). The effect of normal aging on the coupling of neural activity to the bold hemodynamic response. *Neuroimage*, 10(1), 6-14.
- Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hear Res*, 229(1-2), 132-147.
- Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., & Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *Neuroimage*, 24(1), 21-33.

- Della-Maggiore, V., Sekuler, A. B., Grady, C. L., Bennett, P. J., Sekuler, R., & McIntosh, A. R. (2000). Corticolimbic interactions associated with performance on a short-term memory task are modified by age. *J Neurosci*, 20(22), 8410-8416.
- Demonet, J. F., Chollet, F., Ramsay, S., Cardebat, D., Nespoulous, J. L., Wise, R., et al. (1992). The anatomy of phonological and semantic processing in normal subjects. *Brain*, 115 (Pt 6), 1753-1768.
- Doupe, A.J., & Kuhl, P. K. (1999). Birdsong and human speech: common themes and mechanisms. *Annu Rev Neurosci*, 22, 567-631.
- Eggermont, J. J. (2001). Between sound and perception: Reviewing the search for a neural code. *Hear Res*, 157(1-2), 1-42.
- Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., et al. (2005). A new spm toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage*, 25(4), 1325-1335.
- Eliades, S. J., & Wang, X. (2003). Sensory-motor interaction in the primate auditory cortex during self-initiated vocalizations. *J Neurophysiol*, 89(4), 2194-2207.
- Eliades, S. J., & Wang, X. (2005). Dynamics of auditory-vocal interaction in monkey auditory cortex. *Cereb Cortex*, 15(10), 1510-1523.
- Elman, J. L. (1981). Effects of frequency-shifted feedback on the pitch of vocal productions. *J Acoust Soc Am*, 70(1), 45-50.
- Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *Eur J Neurosci*, 15(2), 399-402.

- Ford, J. M., Gray, M., Faustman, W. O., Heinks, T. H., & Mathalon, D. H. (2005). Reduced gamma-band coherence to distorted feedback during speech when what you say is not what you hear. *Int J Psychophysiol*, 57(2), 143-150.
- Ford, J. M., & Mathalon, D. H. (2005). Corollary discharge dysfunction in schizophrenia: Can it explain auditory hallucinations? *Int J Psychophysiol*, 58(2-3), 179-189.
- Ford, J. M., Mathalon, D. H., Heinks, T., Kalba, S., Faustman, W. O., & Roth, W. T. (2001). Neurophysiological evidence of corollary discharge dysfunction in schizophrenia. *Am J Psychiatry*, 158(12), 2069-2071.
- Friederici, A. D., Ruschemeyer, S. A., Hahne, A., & Fiebach, C. J. (2003). The role of left inferior frontal and superior temporal cortex in sentence comprehension: Localizing syntactic and semantic processes. *Cereb Cortex*, 13(2), 170-177.
- Friston, K. J., Frith, C. D., Liddle, P. F., & Frackowiak, R. S. (1993). Functional connectivity: The principal-component analysis of large (PET) data sets. *J Cereb Blood Flow Metab*, 13(1), 5-14.
- Frith, C. D. (1987). The positive and negative symptoms of schizophrenia reflect impairments in the perception and initiation of action. *Psychol Med*, 17(3), 631-648.
- Fu, C. H., Vythelingum, G. N., Brammer, M. J., Williams, S. C., Amaro, E., Jr., Andrew, C. M., et al. (2006). An fMRI study of verbal self-monitoring: Neural correlates of auditory verbal feedback. *Cereb Cortex*, 16(7), 969-977.
- Fukawa, T., Yoshioka, H., Ozawa, E., & Yoshida, S. (1988). Difference of susceptibility to delayed auditory feedback between stutterers and nonstutterers. *J Speech Hear Res*, 31(3), 475-479.

- Gazzaniga, M. S, Ivry, R. B., & Mangun, G. R. (2002). *Cognitive Neuroscience, Second edition*. W.W. Norton & Company, New York.
- Geers, A., & Brenner, C. (2003). Background and educational characteristics of prelingually deaf children implanted by five years of age. *Ear Hear*, 24(1 Suppl), 2S-14S.
- Giraud, A. L., Kell, C., Thierfelder, C., Sterzer, P., Russ, M. O., Preibisch, C., et al. (2004). Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. *Cereb Cortex*, 14(3), 247-255.
- Giraud, A. L., & Price, C. J. (2001). The constraints functional neuroimaging places on classical models of auditory word processing. *J Cogn Neurosci*, 13(6), 754-765.
- Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M. L., Scherer, K. R., et al. (2005). The voices of wrath: Brain responses to angry prosody in meaningless speech. *Nat Neurosci*, 8(2), 145-146.
- Griffiths, T. D., & Warren, J. D. (2002). The planum temporale as a computational hub. *Trends Neurosci*, 25(7), 348-353.
- Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *J Commun Disord*, 39(5), 350-365.
- Hackett, T. A., Stepniewska, I., & Kaas, J. H. (1998). Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. *J Comp Neurol*, 394(4), 475-495.
- Hall, D. A., Hart, H. C., & Johnsrude, I. S. (2003). Relationships between human auditory cortical structure and function. *Audiol Neurotol*, 8(1), 1-18.

- Hampson, M., Peterson, B. S., Skudlarski, P., Gatenby, J. C., & Gore, J. C. (2002). Detection of functional connectivity using temporal correlations in MR images. *Hum Brain Mapp*, 15(4), 247-262.
- Haruno, M., Wolpert, D. M., & Kawato, M. (2001). Mosaic model for sensorimotor learning and control. *Neural Comput*, 13(10), 2201-2220.
- Hashimoto, Y., & Sakai, K. L. (2003). Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: An fMRI study. *Hum Brain Mapp*, 20(1), 22-28.
- Haueisen, J., & Knosche, T. R. (2001). Involuntary motor activity in pianists evoked by music perception. *J Cogn Neurosci*, 13(6), 786-792.
- Heinks-Maldonado, T. H., Mathalon, D. H., Gray, M., & Ford, J. M. (2005). Fine-tuning of auditory cortex during speech production. *Psychophysiology*, 42(2), 180-190.
- von Helmholtz, H. (1878). The facts in perception. 698-726, in *From Kant to Hilbert: A Source Book in the Foundations of Mathematics*, 2vols. Oxford Uni. Press.
- Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends Cogn Sci*, 4(4), 131-138.
- Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1-2), 67-99.
- Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nat Rev Neurosci*, 8(5), 393-402.

- Hikosaka, K., Iwai, E., Saito, H., Tanaka, K. (1988). Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey. *J Neurophysiol*, 60, 1615-1637.
- Hirano, S., Kojima, H., Naito, Y., Honjo, I., Kamoto, Y., Okazawa, H., et al. (1997). Cortical processing mechanism for vocalization with auditory verbal feedback. *Neuroreport*, 8(9-10), 2379-2382.
- Hofman, P. M., Van Riswick, J. G., & Van Opstal, A. J. (1998). Relearning sound localization with new ears. *Nat Neurosci*, 1(5), 417-421.
- von Holst, E., & Mittelstaedt, H. (1950). The principle of reafference. *Naturwissenschaften*, 37, 464-476.
- Horwitz, B. (2003). The elusive concept of brain connectivity. *Neuroimage*, 19(2 Pt 1), 466-470.
- Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, 279(5354), 1213-1216.
- Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: An meg study. *J Cogn Neurosci*, 14(8), 1125-1138.
- Howell, P. (1990). Changes in voice level caused by several forms of altered feedback in fluent speakers and stutterers. *Lang Speech*, 33 (Pt 4), 325-338.
- Howell, P., & Powell, D. J. (1987). Delayed auditory feedback with delayed sounds varying in duration. *Percept Psychophys*, 42(2), 166-172.
- Huang, J., Carr, T. H., & Cao, Y. (2002). Comparing cortical activations for silent and overt speech using event-related fmri. *Hum Brain Mapp*, 15(1), 39-53.

- Ingvar, D. H. (1983). Serial aspects of language and speech related to prefrontal cortical activity. A selective review. *Hum Neurobiol*, 2(3), 177-189.
- Iwai, E., Aihara, T., Hikosaka, K. (1987). Inferotemporal neurons of the monkey responsive to auditory signal. *Brain Res*, 410, 121-124.
- Jamison, H. L., Watkins, K. E., Bishop, D. V., & Matthews, P. M. (2006). Hemispheric specialization for processing auditory nonspeech stimuli. *Cereb Cortex*, 16(9), 1266-1275.
- Jancke, L., Wustenberg, T., Scheich, H., & Heinze, H. J. (2002). Phonetic perception and the temporal cortex. *Neuroimage*, 15(4), 733-746.
- Jardri, R., Pins, D., Bubrovszky, M., Desprez, P., Pruvo, J. P., Steinling, M., et al. (2007). Self awareness and speech processing: An fMRI study. *Neuroimage*, 35(4), 1645-1653.
- JL, E. L. (1981). Effects of frequency-shifted feedback on the pitch of vocal productions. *J Acoust Soc Am*, 70(1), 45-50.
- Joanisse, M. F., & Gati, J. S. (2003). Overlapping neural regions for processing rapid temporal cues in speech and nonspeech signals. *Neuroimage*, 19(1), 64-79.
- Johns, L. C., Rossell, S., Frith, C., Ahmad, F., Hemsley, D., Kuipers, E., et al. (2001). Verbal self-monitoring and auditory verbal hallucinations in patients with schizophrenia. *Psychol Med*, 31(4), 705-715.
- Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of f0 production: Evidence from feedback perturbation. *J Acoust Soc Am*, 108(3 Pt 1), 1246-1251.
- Kaas, J. H., & Hackett, T. A. (1999). 'What' and 'where' processing in auditory cortex. *Nat Neurosci*, 2(12), 1045-1047.

- Kaas, J. H., Hackett, T. A., & Tramo, M. J. (1999). Auditory processing in primate cerebral cortex. *Curr Opin Neurobiol*, 9(2), 164-170.
- Kalinowski, J., Armson, J., Roland-Mieszkowski, M., Stuart, A., & Gracco, V. L. (1993). Effects of alterations in auditory feedback and speech rate on stuttering frequency. *Lang Speech*, 36 (Pt 1), 1-16.
- Keefe, R. S., Arnold, M. C., Bayen, U. J., & Harvey, P. D. (1999). Source monitoring deficits in patients with schizophrenia; a multinomial modelling analysis. *Psychol Med*, 29(4), 903-914.
- Konishi, M. (1965). The role of auditory feedback in the control of vocalization in the white-crowned sparrow. *Z. Tierpsychol*. 22, 770-783.
- Konishi, M. (1985). Birdsong: From behavior to neuron. *Annu. Rev. Neurosci*. 8, 125-170.
- Kosaki, H., Hashikawa, T., He, J., & Jones, E. G. (1997). Tonotopic organization of auditory cortical fields delineated by parvalbumin immunoreactivity in macaque monkeys. *J Comp Neurol*, 386(2), 304-316.
- Lane, A., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *J Speech Hear Res*, 14, 677-709.
- Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., & Bourne, V. T. (1976). Speaker sex identification from voiced, whispered, and filtered isolated vowels. *J Acoust Soc Am*, 59(3), 675-678.
- Leonardo, A. & Konishi, M. (1999). Decrystallization of adult birdsong by perturbation of auditory feedback. *Nature*, 399, 466-470.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol Rev*, 74(6), 431-461.

- Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1), 1-36.
- Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science*, 243(4890), 489-494.
- Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends Cogn Sci*, 4(5), 187-196.
- Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cereb Cortex*, 15(10), 1621-1631.
- Lincoln, M., Packman, A., & Onslow, M. (2006). Altered auditory feedback and the treatment of stuttering: A review. *J Fluency Disord*, 31(2), 71-89.
- Lowe, M. J., Dziedzic, M., Lurito, J. T., Mathews, V. P., & Phillips, M. D. (2000). Correlations in low-frequency bold fluctuations reflect cortico-cortical connections. *Neuroimage*, 12(5), 582-587.
- MacKay, D. G. (1968). Metamorphosis of a critical interval: Age-linked changes in the delay in auditory feedback that produces maximal disruption of speech. *J Acoust Soc Am*, 43(4), 811-821.
- Majerus, S., Amand, P., Boniver, V., Demanez, J. P., Demanez, L., & Van der Linden, M. (2005). A quantitative and qualitative assessment of verbal short-term memory and phonological processing in 8-year-olds with a history of repetitive otitis media. *J Commun Disord*, 38(6), 473-498.
- Mathiak, K., Hertrich, I., Grodd, W., & Ackermann, H. (2002). Cerebellum and speech perception: A functional magnetic resonance imaging study. *J Cogn Neurosci*, 14(6), 902-912.

- Matsuzawa, M., Matsuo, K., Sugio, T., Kato, C., & Nakai, T. (2005). Temporal relationship between action and visual outcome modulates brain activation: An fMRI study. *Magn Reson Med Sci*, 4(3), 115-121.
- Marler, P. (1970). A comparative approach to vocal learning: song development in white-crowned sparrows. *J. Comp. Physiol. Psychol.* 71, 1-25.
- McGuire, P. K., Shah, G. M., & Murray, R. M. (1993). Increased blood flow in broca's area during auditory hallucinations in schizophrenia. *Lancet*, 342(8873), 703-706.
- McGuire, P. K., Silbersweig, D. A., Wright, I., Murray, R. M., David, A. S., & Frackowiak, R. S. J. (1995). Abnormal monitoring of inner speech: a physiological basis for auditory hallucinations. *Lancet*, 346, 596-600.
- McGuire, P. K., Silbersweig, D. A., & Frith, C. D. (1996). Functional neuroanatomy of verbal self-monitoring. *Brain*, 119(Pt 3), 907-917.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746-748.
- Morel, A., Garraghty, P. E., & Kaas, J. H. (1993). Tonotopic organization, architectonic fields, and connections of auditory cortex in macaque monkeys. *J Comp Neurol*, 335(3), 437-459.
- Muller-Preuss, P., Newman, J. D., & Jurgens, U. (1980). Anatomical and physiological evidence for a relationship between the 'cingular' vocalization area and the auditory cortex in the squirrel monkey. *Brain Res*, 202(2), 307-315.
- Mummery, C. J., Ashburner, J., Scott, S. K., & Wise, R. J. (1999). Functional neuroimaging of speech perception in six normal and two aphasic subjects. *J Acoust Soc Am*, 106(1), 449-457.

- Newman, S. D., & Twieg, D. (2001). Differences in auditory processing of words and pseudowords: An fmri study. *Hum Brain Mapp*, 14(1), 39-47.
- Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Percept Psychophys*, 60(3), 355-376.
- Palmer, E. D., Rosen, H. J., Ojemann, J. G., Buckner, R. L., Kelley, W. M., & Petersen, S. E. (2001). An event-related fmri study of overt and covert word stem completion. *Neuroimage*, 14(1 Pt 1), 182-193.
- Pandya, D. N. (1995). Anatomy of the auditory cortex. *Rev Neurol (Paris)*, 151(8-9), 486-494.
- Paus, T., Perry, D. W., Zatorre, R. J., Worsley, K. J., & Evans, A. C. (1996). Modulation of cerebral blood flow in the human auditory cortex during speech: Role of motor-to-sensory discharges. *Eur J Neurosci*, 8(11), 2236-2246.
- Petrides, M., & Pandya, D. N. (1988). Association fiber pathways to the frontal cortex from the superior temporal region in the rhesus monkey. *J Comp Neurol*, 273(1), 52-66.
- Poeppel, D., & Hickok, G. (2004). Towards a new functional anatomy of language. *Cognition*, 92(1-2), 1-12.
- Postle, B. R., Zarahn, E., & D'Esposito, M. (2000). Using event-related fmri to assess delay-period activity during performance of spatial and nonspatial working memory tasks. *Brain Res Brain Res Protoc*, 5(1), 57-66.
- Poulet, J. F., & Hedwig, B. (2002). A corollary discharge maintains auditory sensitivity during sound production. *Nature*, 418(6900), 872-876.

- Price, P. H. (1979). Developmental determinants of structure in zebra finch song. *J. Comp. Physiol. Psychol.* 93, 268-277.
- Purcell, D. W., & Munhall, K. G. (2006a). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *J Acoust Soc Am*, 120(2), 966-977.
- Purcell, D. W., & Munhall, K. G. (2006b). Compensation following real-time manipulation of formants in isolated vowels. *J Acoust Soc Am*, 119(4), 2288-2297.
- Ranganath, C., Johnson, M. K., & D'Esposito, M. (2003). Prefrontal activity associated with working memory and episodic long-term memory. *Neuropsychologia*, 41(3), 378-389.
- Rauschecker, J. P. (1998). Parallel processing in the auditory cortex of primates. *Audiol Neurotol*, 3(2-3), 86-103.
- Rauschecker, J. P., Tian, B., & Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, 268(5207), 111-114.
- Rauschecker, J. P., Tian, B., Pons, T., & Mishkin, M. (1997). Serial and parallel processing in rhesus monkey auditory cortex. *J Comp Neurol*, 382(1), 89-103.
- Rimol, L. M., Specht, K., Weis, S., Savoy, R., & Hugdahl, K. (2005). Processing of sub-syllabic speech units in the posterior temporal lobe: An fMRI study. *Neuroimage*, 26(4), 1059-1067.
- Rissman, J., Gazzaley, A., & D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage*, 23(2), 752-763.

- Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., & Rauschecker, J. P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nat Neurosci*, 2(12), 1131-1136.
- Rose, G. J. (2004). Insights into neural mechanisms and evolution of behaviour from electric fish. *Nat Rev Neurosci*, 5(12), 943-951.
- Rumiati, R. I., Weiss, P. H., Tessari, A., Assmus, A., Zilles, K., Herzog, H., et al. (2005). Common and differential neural mechanisms supporting imitation of meaningful and meaningless actions. *J Cogn Neurosci*, 17(9), 1420-1431.
- Schauwers, K., Gillis, S., Daemers, K., De Beukelaer, C., De Ceulaer, G., Yperman, M., et al. (2004). Normal hearing and language development in a deaf-born child. *Otol Neurotol*, 25(6), 924-929.
- Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, 123 Pt 12, 2400-2406.
- Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends Neurosci*, 26(2), 100-107.
- Scott, S. K., Rosen, S., Lang, H., & Wise, R. J. (2006). Neural correlates of intelligibility in speech investigated with noise vocoded speech--a positron emission tomography study. *J Acoust Soc Am*, 120(2), 1075-1083.
- Scott, S. K., & Wise, R. J. (2004). The functional neuroanatomy of prelexical processing in speech perception. *Cognition*, 92(1-2), 13-45.
- Siegel, G. M., Schork, E. J., Jr., Pick, H. L., Jr., & Garber, S. R. (1982). Parameters of auditory feedback. *J Speech Hear Res*, 25(3), 473-475.

- Silbersweig, D. A., Stern, E., Frith, C., Cahill, C., Holmes, A., Grootenboer, S., et al. (1995). A functional neuroanatomy of hallucinations in schizophrenia. *Nature*, 378(6553), 176-179.
- Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *Neuroimage*, 25(1), 76-89.
- Smith, D. R., Patterson, R. D., Turner, R., Kawahara, H., & Irino, T. (2005). The processing and perception of size information in speech sounds. *J Acoust Soc Am*, 117(1), 305-318.
- Sommer, M. A., & Wurtz, R. H. (2006). Influence of the thalamus on spatial visual processing in frontal cortex. *Nature*, 444(7117), 374-377.
- Stager, S. V., & Ludlow, C. L. (1993). Speech production changes under fluency-evoking conditions in nonstuttering speakers. *J Speech Hear Res*, 36(2), 245-253.
- Suga, N., & Shimozawa, T. (1974). Site of neural attenuation of responses to self-vocalized sounds in echolocating bats. *Science*, 183(130), 1211-1213.
- Sun, F. T., Miller, L. M., & D'Esposito, M. (2004). Measuring interregional functional connectivity using coherence and partial coherence analyses of fMRI data. *Neuroimage*, 21(2), 647-658.
- Thivard, L., Belin, P., Zilbovicius, M., Poline, J. B., & Samson, Y. (2000). A cortical region sensitive to auditory spectral motion. *Neuroreport*, 11(13), 2969-2972.
- Thorpe, W. H. (1958). The learning of song patterns by birds with especial reference to the song of the chaffinch. *Fringilla Coelebs. Ibis*, 100, 535-570.
- Tian, B., Reser, D., Durham, A., Kustov, A., & Rauschecker, J. P. (2001). Functional specialization in rhesus monkey auditory cortex. *Science*, 292(5515), 290-293.

- Toyomura, A., Koyama, S., Miyamaoto, T., Terao, A., Omori, T., Murohashi, H., et al. (2007). Neural correlates of auditory feedback control in human. *Neuroscience*, *146*(2), 499-503.
- Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., 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.
- Uppenkamp, S., Johnsrude, I. S., Norris, D., Marslen-Wilson, W., & Patterson, R. D. (2006). Locating the initial stages of speech-sound processing in human temporal cortex. *Neuroimage*, *31*(3), 1284-1296.
- Vouloumanos, A., Kiehl, K. A., Werker, J. F., & Liddle, P. F. (2001). Detection of sounds in the auditory stream: Event-related fMRI evidence for differential activation to speech and nonspeech. *J Cogn Neurosci*, *13*(7), 994-1005.
- Warren, J. D., Zielinski, B. A., Green, G. G., Rauschecker, J. P., & Griffiths, T. D. (2002). Perception of sound-source motion by the human brain. *Neuron*, *34*(1), 139-148.
- Warren, J. E., Wise, R. J., & Warren, J. D. (2005). Sounds do-able: Auditory-motor transformations and the posterior temporal plane. *Trends Neurosci*, *28*(12), 636-643.
- Watkins, K., & Paus, T. (2004). Modulation of motor excitability during speech perception: The role of Broca's area. *J Cogn Neurosci*, *16*(6), 978-987.
- Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, *41*(8), 989-994.

- Webb, B. (2004). Neural mechanisms for prediction: Do insects have forward models? *Trends Neurosci*, 27(5), 278-282.
- Weintraub, S., Mesulam, M. M., & Kramer, L. (1981). Disturbances in prosody. A right-hemisphere contribution to language. *Arch Neurol*, 38(12), 742-744.
- Wildgruber, D., Ackermann, H., & Grodd, W. (2001). Differential contributions of motor cortex, basal ganglia, and cerebellum to speech motor control: Effects of syllable repetition rate evaluated by fmri. *Neuroimage*, 13(1), 101-109.
- Wildgruber, D., Ackermann, H., Klose, U., Kardatzki, B., & Grodd, W. (1996). Functional lateralization of speech production at primary motor cortex: A fMRI study. *Neuroreport*, 7(15-17), 2791-2795.
- Wilson, S. M., & Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception. *Neuroimage*, 33(1), 316-325.
- Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat Neurosci*, 7(7), 701-702.
- Wolf, R., & Heisenberg, M. (1990). Visual control of straight flight in *Drosophila melanogaster*. *J Comp Physiol [A]*, 167(2), 269-283.
- Wolpert, D. M., & Flanagan, J. R. (2001). Motor prediction. *Curr Biol*, 11(18), R729-732.
- Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Netw*, 11(7-8), 1317-1329.

- Wong, P. C., Perrachione, T. K., & Parrish, T. B. (2006). Neural characteristics of successful and less successful speech and word learning in adults. *Hum Brain Mapp.*
- Zanini, S., Clarici, A., Fabbro, F., & Bava, A. (1999). Speaking speed effects on delayed auditory feedback disruption of speech fluency. *Percept Mot Skills*, 89(3 Pt 2), 1095-1109.
- Zarahn, E., Aguirre, G., & D'Esposito, M. (1997). A trial-based experimental design for fmri. *Neuroimage*, 6(2), 122-138.
- Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends Cogn Sci*, 6(1), 37-46.
- Zatorre, R. J., Evans, A. C., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science*, 256(5058), 846-849.

Appendix 1

Condition Sequence A							
0	30	60	90	120	150		
P:masked	L:masked	P:unmasked	L:masked	L:unmasked	P:masked		
Rest	P:masked	Rest	Rest	Rest	P:unmasked		
P:unmasked	P:unmasked	P:masked	L:unmasked	L:masked	Rest		
L:unmasked	Rest	L:masked	P:unmasked	P:unmasked	L:masked		
L:masked	L:unmasked	L:unmasked	P:masked	P:masked	L:unmasked		
P:unmasked	L:masked	P:unmasked	L:masked	Rest	L:unmasked		
Rest	Rest	P:masked	P:unmasked	L:unmasked	L:masked		
L:masked	P:masked	L:unmasked	Rest	P:masked	P:unmasked		
P:masked	L:unmasked	Rest	P:masked	P:unmasked	P:masked		
L:unmasked	P:unmasked	L:masked	L:unmasked	L:masked	Rest		
L:unmasked	Rest	Rest	L:masked	L:masked	L:unmasked		
L:masked	L:masked	P:unmasked	L:unmasked	Rest	P:masked		
P:unmasked	P:masked	L:unmasked	P:masked	P:masked	P:unmasked		
Rest	L:unmasked	P:masked	P:masked	P:unmasked	L:masked		
P:masked	P:unmasked	L:masked	P:unmasked	L:masked	P:masked		
L:masked	Rest	Rest	L:unmasked	L:masked	P:masked		
P:masked	L:unmasked	P:masked	P:masked	P:masked	Rest		
Rest	P:masked	P:unmasked	L:masked	Rest	L:unmasked		
P:unmasked	L:masked	L:unmasked	Rest	P:unmasked	P:unmasked		
L:unmasked	P:unmasked	L:masked	P:masked	L:masked	L:masked		
P:unmasked	P:unmasked	P:masked	L:unmasked	P:masked	Rest		
Rest	L:masked	L:unmasked	P:masked	L:unmasked	P:masked		
L:unmasked	L:unmasked	L:masked	Rest	Rest	P:unmasked		
L:masked	Rest	Rest	L:masked	Rest	Rest		
P:masked	P:masked	P:unmasked	L:masked	L:masked	P:masked		
P:unmasked	L:unmasked	P:unmasked	P:masked	Rest	L:masked		
Rest	P:unmasked	Rest	Rest	P:unmasked	P:masked		
P:masked	P:masked	L:masked	P:masked	P:masked	P:unmasked		
L:masked	Rest	P:masked	Rest	L:masked	L:masked		
L:unmasked	L:masked	L:unmasked	L:unmasked	L:unmasked	Rest		

Appendix 2

Condition Sequence B							
0	30	60	90	120	150		
P:unmasked	Rest	L:masked	L:unmasked	P:masked	P:unmasked		
P:masked	L:masked	P:masked	L:masked	Rest	L:unmasked		
L:masked	P:masked	L:unmasked	Rest	P:masked	P:unmasked		
L:unmasked	L:unmasked	Rest	P:unmasked	P:unmasked	L:masked		
Rest	P:unmasked	P:unmasked	L:masked	L:masked	Rest		
P:unmasked	P:masked	L:unmasked	Rest	L:unmasked	P:unmasked		
L:unmasked	L:unmasked	Rest	P:masked	P:unmasked	L:unmasked		
Rest	Rest	P:masked	P:unmasked	Rest	P:masked		
P:masked	L:masked	P:unmasked	L:masked	P:masked	L:masked		
L:masked	P:unmasked	L:masked	P:masked	L:unmasked	P:unmasked		
L:masked	L:unmasked	L:masked	L:masked	P:masked	L:unmasked		
P:unmasked	P:unmasked	Rest	Rest	L:unmasked	L:masked		
P:masked	Rest	P:masked	P:unmasked	P:unmasked	L:masked		
L:unmasked	L:masked	P:unmasked	Rest	Rest	P:masked		
Rest	P:masked	L:unmasked	L:masked	L:masked	Rest		
P:unmasked	L:masked	L:unmasked	Rest	Rest	P:unmasked		
P:masked	P:unmasked	L:masked	L:masked	L:masked	L:unmasked		
L:unmasked	Rest	P:unmasked	L:unmasked	L:masked	Rest		
L:masked	L:unmasked	P:masked	P:masked	Rest	P:masked		
Rest	P:masked	Rest	P:unmasked	P:masked	L:unmasked		
P:unmasked	Rest	P:masked	Rest	Rest	P:masked		
Rest	L:masked	P:unmasked	L:unmasked	P:masked	L:masked		
L:unmasked	P:unmasked	L:unmasked	L:masked	P:unmasked	P:unmasked		
P:masked	P:masked	L:masked	Rest	Rest	L:unmasked		
L:masked	L:unmasked	Rest	P:masked	L:unmasked	Rest		
Rest	P:unmasked	P:masked	L:unmasked	P:masked	L:masked		
P:unmasked	Rest	Rest	Rest	Rest	P:unmasked		
P:masked	L:masked	L:masked	P:masked	P:masked	L:unmasked		
L:unmasked	L:unmasked	P:unmasked	P:unmasked	P:masked	P:masked		
L:masked	P:masked	L:unmasked	L:masked	Rest	L:masked		

Appendix 3

Condition Sequence C							
0	30	60	90	120	150		
P:unmasked	L:masked	L:unmasked	L:unmasked	L:unmasked	L:unmasked	Rest	
P:masked	L:unmasked	P:unmasked	P:masked	P:masked	P:unmasked	P:unmasked	
L:unmasked	Rest	P:masked	Rest	Rest	P:masked	L:masked	
Rest	P:unmasked	Rest	P:unmasked	P:unmasked	L:masked	P:masked	
L:masked	P:masked	L:masked	L:masked	L:masked	Rest	L:unmasked	
P:masked	P:unmasked	P:unmasked	Rest	P:masked	L:unmasked	P:unmasked	
Rest	L:masked	L:masked	P:masked	L:masked	Rest	L:masked	
L:unmasked	P:masked	P:masked	L:unmasked	L:unmasked	P:masked	L:masked	
P:unmasked	Rest	Rest	Rest	P:unmasked	L:masked	Rest	
L:masked	L:unmasked	L:unmasked	P:unmasked	P:masked	P:unmasked	P:masked	
L:unmasked	L:unmasked	L:unmasked	L:masked	Rest	L:masked	L:masked	
L:masked	Rest	P:unmasked	L:masked	L:masked	Rest	L:unmasked	
Rest	P:unmasked	P:masked	L:unmasked	P:masked	P:masked	Rest	
P:unmasked	P:masked	Rest	Rest	P:unmasked	P:unmasked	P:unmasked	
P:masked	L:masked	Rest	P:masked	Rest	L:unmasked	P:masked	
P:unmasked	L:unmasked	Rest	P:masked	P:unmasked	P:unmasked	L:masked	
P:masked	P:unmasked	P:masked	P:unmasked	L:masked	L:masked	Rest	
Rest	P:masked	L:unmasked	L:masked	P:masked	Rest	P:unmasked	
L:masked	Rest	L:masked	L:masked	L:unmasked	P:masked	L:unmasked	
L:unmasked	L:masked	Rest	P:masked	Rest	L:unmasked	P:unmasked	
Rest	L:unmasked	L:masked	L:masked	L:unmasked	P:masked	P:masked	
P:masked	Rest	Rest	P:masked	P:unmasked	Rest	L:masked	
L:masked	P:masked	P:masked	P:unmasked	P:masked	L:masked	L:unmasked	
P:unmasked	L:masked	P:unmasked	L:unmasked	L:masked	P:unmasked	Rest	
L:unmasked	P:unmasked	L:masked	L:masked	L:unmasked	Rest	Rest	
Rest	Rest	Rest	P:unmasked	P:unmasked	P:masked	L:unmasked	
L:masked	L:unmasked	P:unmasked	Rest	P:masked	P:unmasked	P:unmasked	
L:unmasked	P:unmasked	L:unmasked	P:masked	L:masked	L:masked	P:masked	
P:unmasked	P:masked	P:masked	L:masked	L:unmasked	L:unmasked	L:masked	

Appendix 4

Debriefing Form

1. Did you feel comfortable lying in the scanner?
2. Were you feeling drowsy, sleepy or dizzy during the study?
3. When you were listening, did you feel the sounds were too loud? Were you able to discern whose sounds they are?
4. Did you adjust the volume of your voice during the experiment or Did you keep your volume the same all along?
5. What did you think of the pacing of the experiment?
6. Do you think the masking noise is effective? Could you hear your voice through headphones in the presence of masking noise?
7. Did you feel nervous or scared at the onset of scanning?
8. What do you think of our study? Did you feel bored during the study? Were you able to concentrate from the beginning to the end?
9. Is there anything that you think we could do to make you more comfortable or focused in the experiment?
10. What do you think the study is about?

Appendix 5

Stimulus: pure tone played at 1000HZ; Unit: dB;

P:unmasked: Production with normal speech feedback;

P:masked: Production with masking noise feedback;

L:unmasked: Listening to speech; L:masked: Listening to noise;

<u>Trial type</u>	<u>Amplitude</u>
P:unmasked	51.3
P:masked	51.4
L:masked	51.3
L:unmasked	51.4
P:unmasked	51.3
L:unmasked	51.3
P:masked	51.3
L:masked	51.4
L:masked	51.3
P:unmasked	51.4
P:masked	51.3
L:unmasked	51.3
P:unmasked	51.3
P:masked	51.3
L:unmasked	51.4
L:masked	51.3
P:unmasked	51.2
L:unmasked	51.2
P:masked	51.3
L:masked	51.2
P:unmasked	51.2
P:masked	51.4
L:unmasked	51.2
L:masked	51.3
L:masked	51.4
P:masked	51.2
L:unmasked	51.4
P:unmasked	51.2