PERCEPTUAL PROCESSING OF AUDITORY FEEDBACK DURING SPEECH PRODUCTION AND ITS NEURAL SUBSTRATES

by

Zane Zheng

A thesis submitted to the Centre for Neuroscience Studies in conformity with the requirements for the degree of Doctor of Philosophy

Queen’s University
Kingston, Ontario, Canada
August, 2012

Copyright ©Zane Zheng, 2012
Abstract

One enduring question in the study of speech concerns the nature of the link between speech perception and production. Although accumulating evidence suggests that these two facets of spoken language are tightly coupled, the cognitive structure and neural organization underlying the interactions between the two processes are not well understood.

In this thesis, I focus on questions that arise from observations related to when individuals are both talking and listening, and assess the sensitivity of talkers and listeners to the same change in the acoustics of speech. First, I aim to elucidate the neural substrates of auditory feedback control during vocalization by examining the brain response to acoustic perturbations towards auditory concomitants of speech using functional magnetic resonance imaging (fMRI) (Chapters 2 and 3). I demonstrate, for the first time, an extensive network of brain regions involved in the detection and correction of auditory feedback errors during speech production, for which three functionally differentiated neural systems can be delineated. Then I set out to address the online perception of own voice identity as individuals are talking. Chapters 4 and 5 measure the perceptual sensitivity of individuals to the auditory concomitants of their own speech by presenting temporally gated auditory feedback in stranger’s voices during talking. The results show that people perceive stranger’s voices as a modified version of their own voice and adjust their vocal production accordingly, when their utterances and heard feedback are phonetically congruent. Chapter 6 further examines this perceptual effect by using experimental paradigms in the domain of body ownership and shows that the misattribution of the stranger’s voice, is not predicted by individual differences in suggestibility; rather it is related to the integration of multimodal cues.

In summary, by focusing on how the acoustics of speech are simultaneously processed for both the perception and production sides of spoken language, the series of studies add significantly to our understanding of the psychophysical, cognitive and anatomical relationships between speech
perception and production, and are relevant to a wide range of clinical pathologies (e.g., stuttering, schizophrenia).
Co-Authorship

I have been actively involved in all aspects (study conception, experimental design, data collection and analysis, and composition) of the presented work under the supervision of Dr. Ingrid Johnsrude. The Chapter 3 has been submitted for publication and the Chapter 5 has already been published:


Acknowledgements

When I first came to Kingston to study for my Master’s degree 7 years ago, Ingrid invited me to her house. While preparing for the dinner, she says, over the next few years, I am going to help you with your Ph.D; in return, you will need to teach me how to cook Chinese food. Now 7 years have passed, Ingrid has fulfilled her promise and yet she hasn’t even had a chance to taste any Chinese food made by me. Needless to say, Ingrid is all that I can ever ask for as a supervisor, and without her, I wouldn’t be nearly as close to where I am today. Thank you, Ingrid – you are and will be a very important person in my life!

Another academic hero on my journey is Kevin Munhall. Kevin, you are one of the most knowledgeable, friendly, (and funny 😊) professors I have ever met. You have made every project that we have done together that much more enjoyable, memorable and fulfilling.

I would particularly like to thank my parents for their trust and support all these years - I am so blessed to have you as my family and I love you always!

I am also grateful to the tremendous help from the Conchlab members and I really enjoy all the time we have spent together in the past few years!
Table of Contents

Abstract ....................................................................................................................... ii
Co-Authorship ........................................................................................................... iv
Acknowledgements ................................................................................................... v
Table of Contents .................................................................................................... vi
List of Figures ......................................................................................................... vii
List of Tables ......................................................................................................... ix

**Chapter 1 - Introduction** .................................................................................... 1
  1.1 - Preamble ....................................................................................................... 1
  1.2 - Sensorimotor control and interactions in speech production and perception .... 4
  1.3 - Sensorimotor interactions for perception of non-speech aspects of auditory feedback during vocalization ................................................................. 18
  1.4 - Multimodal integration and individual differences in supporting perception of the voice identity of auditory concomitant of speech ........................................................................ 22

**Chapter 2 - Cortical areas involved in auditory feedback control of speech** ........ 24
  2.1 - Introduction .................................................................................................. 24
  2.2 - Methods and Materials ................................................................................ 26
  2.3 - Results ........................................................................................................ 31
  2.4 - Discussion ................................................................................................... 39

**Chapter 3 - Multivoxel patterns reveal functionally differentiated networks underlying auditory feedback control of speech** ................................................. 46
  3.1 - Introduction .................................................................................................. 46
  3.2 - Methods and Materials ................................................................................ 47
  3.3 - Results ........................................................................................................ 54
  3.4 - Discussion ................................................................................................... 64

**Chapter 4 - Perception of voice identity from the auditory concomitants of speech** 70
  4.1 - Introduction .................................................................................................. 70
  4.2 - Methods and Materials ................................................................................ 71
  4.3 - Results ........................................................................................................ 78
  4.4 - Discussion ................................................................................................... 86

**Chapter 5 - Perceiving a stranger's voice as being one's own: a 'rubber voice' illusion?** 91
  5.1 - Introduction .................................................................................................. 91
  5.2 - Methods and Materials ................................................................................ 93
List of Figures

Figure 1.1. Functional overlap between brain areas that are speech sensitive and areas supporting auditory feedback control of speech.................................................................18

Figure 2.1. Schematic diagram of the first 10 seconds from a functional run in our fMRI study.................................................................29

Figure 2.2. Brain areas in which the difference in activity between production with noise and production with clear speech feedback is significantly greater than the difference in activity between listening to noise and listening to clear speech.............................................33

Figure 2.3. Brain areas that are more strongly activated for production than for listening .............36

Figure 2.4. Brain areas that are more strongly activated for speech than for noise....................38

Figure 3.1. Similarity structure contrast matrices for the experimental contrasts in our fMRI study........................................................................................................................................56

Figure 3.2. Brain areas that are speech-sensitive........................................................................57

Figure 3.3. Brain areas that are implicated in speech error coding............................................59

Figure 3.4. Brain areas in which neural patterns exhibited Feedback Type by Production/Listening interaction........................................................................................................................................61

Figure 3.5. A bar graph showing neural pattern correlations between experimental conditions...64

Figure 4.1. Schematic of the experimental paradigm for the rubber voice study.........................75

Figure 4.2. Subjective ratings as percentage of response at each of the seven response categories for the four questionnaire statements.................................................................................................................................79

Figure 4.3. Participants’ fundamental frequency (F0) for each key word presented and rubber voice F0........................................................................................................................................82

Figure 4.4. The average of the difference between talker’s F0 and that of the rubber voice........83

Figure 4.5. A bar graph showing an interaction between Block and Group..................................86

Figure 5.1. Schematic diagram of the four stages of the redesigned rubber voice study.............95

Figure 5.2. The box plots for subjective ratings on the two questionnaire statements...............97

Figure 5.3. The box plots for ratings on the two questionnaire statements for the two different experimental groups........................................................................................................................................100

Figure 5.4. The F0 time course for the word ‘day’ from one representative participant.............104

Figure 6.1. The significant correlation between the subjective ratings of the two illusions........116

Figure 6.2. The averages of participants’ F0s relative to the rubber voice F0..........................118

Figure 6.3. The relationship between rubber voice ratings and F0 shift.....................................119
List of Tables

Table 2.1. The factorial design used in the fMRI study ............................................................. 26
Table 2.2. Brain areas in which the difference in activity between production with noise feedback and production with speech feedback is significantly greater than the difference in activity between listening to noise and listening to speech ......................................................... 32
Table 2.3. Brain areas that show enhanced activation for production with noise feedback relative to production with speech feedback ................................................................. 34
Table 2.4. Brain areas that are more strongly activated for listening to speech than for listening to noise .......................................................................................................................... 34
Table 2.5. Brain areas that are more strongly activated for production than for listening .......... 35
Table 2.6. Brain areas that are more strongly activated for speech than for noise ..................... 37
Table 2.7. Brain areas that are more responsive to noise than to speech ................................ 37
Table 3.1. Speech-sensitive areas revealed by analysis of multivoxel patterns of brain activity . 57
Table 3.2. Brain areas in which neural patterns of activity encode speech errors ................... 59
Table 4.1. Statistics for the magnitude of F0 over time .............................................................. 85
Table 5.1. The number of participants who shifted their F0 up, down, or did not shift their F0, when hearing rubber voices ....................................................................................................... 104
Table 6.1. The number of participants associated with each order in which the three illusions were presented .......................................................................................................................... 111
Table 6.2. The questionnaire items used in the three illusions .................................................. 113
Chapter 1

1.1 – Preamble

This thesis focuses on an enduring question in the study of speech: how are perception and production linked? One of the oldest theories of speech posits that speech perception must be closely linked to speech production (Liberman et al., 1967; Liberman and Mattingly, 1985). When we speak, the sound of our own voice is a powerful influence on the accuracy of fluent speech and is a major force in learning to speak (e.g., Lane and Tranel, 1971; Stuart and Kalinowski, 1996; Callan et al., 2000; Dhanjal et al., 2008; Simmonds et al., 2011; MacDonald et al., 2012). Despite the theoretical and empirical importance of such a link for understanding the cognitive and neural organization of the language system, a great deal of mystery still surrounds the relationship between perception and production.

In this thesis, I describe a set of behavioral and neuroimaging studies that address several questions related to when individuals are both talking and listening. All the studies are based on an auditory feedback perturbation paradigm that tests the sensitivity of talkers and listeners to the same changes in speech acoustics. The key to this work is a real-time speech tracking system that permits perturbations of auditory feedback of speech with a very short time delay (less than 10 msec). Thus, talkers in these studies produce utterances but hear modified versions of them through headphones as they speak, while measures of their perceptual and/or neural response are taken. By focusing on the impact of auditory feedback on both perception and production of speech, these studies add significantly to our understanding of the psychophysical, cognitive and anatomical relationships between speech perception and production. The findings of the studies presented in this thesis are also relevant to a range of clinical populations (e.g., stuttering,
schizophrenia) in which the relation between perception and production is an ongoing research and intervention focal point.

What follows in this opening chapter is a comprehensive review of the relevant literature that lays the foundation for all the studies. The primary goal of the review is to introduce the theories of, and empirical data related to, auditory feedback processing and elaborate how this knowledge motivates the experiments conducted in this thesis. Evidence from a wide range of domains (behavior, neurophysiology, neuroanatomy, and computational models) is presented to show the tight coupling between speech perception and production processes, as well as how the complex interaction between the two processes is mediated by auditory feedback of vocalization.

Chapter 2 describes a functional magnetic resonance imaging (fMRI) study that was motivated by, and extended, a previous fMRI study (Zheng et al., 2010) focused on the functional overlap between the neural systems involved in auditory feedback monitoring and in speech perception. A clear understanding of the extent to which the two systems interact would be the first step to resolving the question of how we process the acoustics of auditory feedback for articulatory control and for perception of speech. To this end, auditory feedback was modified during talking in two different ways: it was either masked by signal-correlated noise, or the speech formants (F1 and F2) were shifted in frequency. Using fMRI, the brain response to these distorted feedback types, compared to unaltered speech feedback, was compared, both when these stimuli were presented during talking and during passive listening. Chapter 2 presents the conventional univariate analysis of these data, whereas in Chapter 3 they are analyzed using a more sensitive data analytic framework (multi-voxel pattern analysis, or MVPA). This framework is based on a recently developed MVPA prototype (Linke et al., 2011) but greatly extended to be better tailored to the needs of our study. One unique feature of this framework is hypothesis testing by
multivariate contrasts, which allows predicted/hypothesized neural pattern similarity among experimental conditions to be explicitly tested within a general linear model (GLM). The unique methodological framework permits me to explore the brain circuits implicated in auditory feedback control of speech from a new perspective, and reveal, for the first time, three functionally differentiated networks subserving distinct aspects of auditory feedback processing.

Based on the results showing the processing of auditory feedback for perception of speech, Chapters 4 and 5 address how people perceive non-speech aspects of auditory feedback during vocalization. In two related studies, I focus on voice-identity information and ask how normal individuals respond to a feedback voice that is heard as the auditory concomitant of their own vocalization (i.e., temporally gated with their own speech production). The idea is that, by manipulating the identity (and/or the speech content) of the voice used for auditory feedback while participants are talking, I can assess how people perceive voice-identity information when speech content is congruent with their own vocal production and when it is not, as well as how subjective perception of voice identity is related to the acoustics of vocal production. This line of inquiry elucidates the possible roles of auditory feedback in the recognition of self-generated vocalization and in speech motor control, and contributes to the understanding of the cognitive principles underlying the distinction between self and other.

Since own-voice identification belongs to the general problem of self-attribution, Chapter 6 presents a study in which established paradigms in the field of self-recognition are used to assess the factors contributing to the recognition of voice identity in auditory feedback. By comparing the subjective perception of an alien voice with that of a prosthetic body part, as well as with self-reported suggestibility ratings, this study reveals that self-voice recognition during talking appears to be due to convergence of multimodal cues and not due to individual differences in
susceptibility to suggestions. This exploration sheds new light on how vocal production may contribute to the perception of self-produced voices, and provides a general framework for future investigations related to body ownership during motor actions.

Finally, Chapter 7 includes a brief discussion of the main findings of my dissertation research and their significance, as well as ideas for future work.

1.2 – Sensorimotor control and interactions in speech production and perception

Speech production and perception seem as natural and effortless as breathing to us, likely because communicating with each other via spoken language is such a key part of our everyday existence. However, the apparent ease of speaking and listening belies the complexity and sophistication of the motor, sensory, and linguistic processes underlying the production of speech and of the resulting acoustic signal that the listener perceives. For example, speech production relies on the coordinated movements of the laryngeal, supraglottal, and respiratory articulators to transform a discretely specified linguistic message into a continuous stream of concrete sounds (Perkell, 2012). On the listener’s side, perception of speech is complicated by the considerable variability of the acoustic environments and phonetic contexts in which speech sounds are produced (Liberman et al., 1967; Diehl et al., 2004). Thus, speech perception is not merely an acoustic event attributable to the extraction and analysis of phonological sounds; rather there are complex perceptual, cognitive, and physiological processes involved.

1.2.1 – Speech production and perception sharing cognitive and neural processes

Evidence suggests that the two sides of spoken language share substantial cognitive resources (Kempen, 2000; Pickering and Garrod, 2004, 2007) and neural organizations (Heim et al., 2003; Awad et al., 2007; Menenti et al., 2011). Most theoretical approaches hypothesize a relationship between production and perception (Liberman et al., 1967; Fowler, 1986; Diehl et al., 2004) and
findings from clinical populations demonstrate correlations between the degree of impairment in talking and listening (e.g., Forrest and Rockman, 1988; Rvachew, 1994). One recent, compelling study exploring speaker-listener neural coupling demonstrated alignment of spatial and temporal structures of neural dynamics in a speaker’s and his or her listener’s brains during natural verbal communication; further, the extent of such neural coupling between the pair of conversation partners predicts the success of communication (Stephens et al., 2010).

Although there is an extensive literature on the perception of speech produced by others (e.g., McGurk and MacDonald, 1976; Miller and Eimas, 1995; Watkins and Paus, 2004; Skipper et al., 2005; McQueen et al., 2006; Obleser and Eisner, 2009; Kotz and Schwartz, 2010; McGettigan et al., 2010; Pulvermuller and Fadiga, 2010; Friederici, 2012), the cognitive structure and neural organization subserving perception of one’s own speech during vocalization are less well understood. Vocal production provides important sensorimotor feedback that plays a vital role in language development and the maintenance of fluent speech articulation. An in-depth understanding of perception of one’s own speech during vocalization would, therefore, offer insight into the psychophysical, cognitive, and anatomical basis of the complex sensorimotor interactions supporting speech communication.

Mounting evidence has shown that being able to hear one’s own voice is crucial for learning and maintaining fluent and stable speech. Congenital or early deafness in life without cochlear implantation at an early age severely hinders the development of normal speech in children (Geers et al., 2003; Schauwers et al., 2004). In addition, 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). Conversely, articulatory dysfunctions such as stuttering can be significantly ameliorated by altering auditory feedback (AAF treatment) (Kalinowski et al., 1993;
Sparks et al., 2002; Stuart et al., 2004; see Lincoln et al., 2006 for review). Thus, it is uncontroversial that perception of one’s own speech is linked to vocalization in some way. However, there is considerable debate over how, and at what stage(s) of processing, perception and production are linked. Second-language learning programs and therapeutic interventions are typically directed towards either perception or production, but, by necessity, usually involve both. Knowing precisely how perception and production interact would allow therapists and educators to strategically use both modalities for more effective intervention.

A significant step towards bridging this gap in our knowledge would be to specify how the acoustics of speech are processed for both the input (perception) and output (production) sides of spoken language. Empirical findings lend support to the view that the input and output of spoken language are coupled in a sensorimotor control process in which auditory feedback provides a key control signal. Behavioral studies have consistently shown that vocal production is altered in response to a change of parameters in acoustics of auditory feedback of speech. For example, delayed auditory feedback (DAF) during vocalization produces disruptive effects on running speech such as 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; Finney and Warren, 2002; Stuart et al., 2002). Experimental paradigms geared towards exploring the effects of auditory feedback on vocal production have reported compensatory vocal responses following intensity-regulated (Bauer et al., 2006; Liu et al., 2007) and frequency-shifted feedback (Elman, 1981; Kalinowski et al., 1993; Burnett et al., 1997; Burnett et al., 1998; Hain et al., 2000; Jones and Munhall, 2000; Natke and Kalveram, 2001; Donath et al., 2002; Houde and Jordan, 2002; Natke et al., 2003; Bauer and Larson, 2003; Xu et al., 2004; Jones and Munhall, 2005; Sivasankar et al., 2005;
Three main findings have emerged from auditory feedback perturbation studies. First, pitch or formant frequency perturbations in auditory feedback generally lead to adjustment of vocal production of the shifted pitch or formants (i.e., compensation) in the opposite direction (but note that this pattern is not uniform across subjects - there are some subjects that tend to change their acoustics in the same direction as the perturbations, for reasons that are not well understood; see Burnett et al., 1998). When the auditory feedback returns to normal, the shifted acoustics in vocal production only gradually return towards but do not reach the baseline level immediately. Second, 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”) (Houde and Jordan 1998), apparently reflecting a systematic speech-motor remapping in response to perceptual alterations (Houde and Jordan 1998; Jones and Munhall, 2005). Third, compensatory behavior in the face of feedback perturbation appears automatic and obligatory, and the vocal response is not easily modified by conscious strategy or intention (e.g., explicit instruction to ignore the feedback changes) (Munhall et al., 2009). Taken together, these findings provide empirical support for the idea that interactions between the processes supporting perception and production of one’s own speech might be coupled in a sensorimotor control system, and that such interactions are sensitive to (and mediated by) the auditory concomitants of speech.

1.2.2 – Sensorimotor control mechanisms linking speech production and perception

Talking, like all motor control, requires sensory feedback to maintain the accuracy and stability of movement. Strong evidence suggests that a motor plan is formulated prior to the onset of
motor movement, and that this plan is continuously updated by an internal loop that relies on both
motor outflow and sensory inflow to assess the consequence of the motor plan (e.g., Desmurget
and Grafton, 2000). The unique advantage of such an internal loop in sensorimotor control is that
it is predictive in nature and thus counteracts some of the effects related to the physiological delay
required to process the sensory feedback. This motor prediction comes in the form of an efference
copy (von Holst and Mittelstaedt, 1950) or corollary discharge (Sperry, 1950), and together with
the sensory feedback, comprises an error-correction mechanism used for control of rapid motor
movement (i.e., in this case, speech). The concepts and processes related to the sensorimotor
control mechanism are explained in more detail below.

Over a hundred years ago, von Helmholtz (Helmholtz, 1878) 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 consequence 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 launches the motor command as motor efferent and the pathway that conveys
information to the central nervous system as sensory afferent. The input through the sensory
afferent caused by the activation of the motor efferent was referred to as reaference. 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. The efference copy is sent internally
to the intended sensory modality and used to cancel the reaference, since the reaference can be
predicted from the motor command and thus contributes little information to our perceptual
experience (Miall, 1998; Blakemore et al., 2002). The cancellation, however, is a non-trivial process that presumably involves the transformation of the motor command into a neural representation that is directly comparable to the sensory input (Webb, 2004).

Over the past several decades, numerous studies of the neural basis of sensorimotor control (e.g., Blakemore et al., 1998; Shadmehr et al., 2010; Wurtz et al., 2011) have demonstrated that the cancellation effect has important functional implications (even at the neuronal level). The efference copy, issued from the motor control area to the sensory area, can serve as a target for incoming sensory feedback; if the incoming feedback does not match this target, sensory cortical neurons will be activated, counteracting the inhibitory/cancellation effect of the efference copy. In the event of such a mismatch, the motor control system can compensate for and then possibly minimize the error by updating the subsequent motor command.

Two fundamental issues (e.g., Shadmehr et al., 2010) with this sensory feedback control scheme are that 1) the context in which motor movement occurs can be highly variable, rendering the sensory feedback information inherently noisy; and 2) the transmission of sensory feedback may be delayed due to neurophysiological constraints imposed by our biological system (e.g., synaptic and/or axonal transmission times; Kandel et al., 2000). However, the two issues can be ameliorated by integrating a feedforward control component into the feedback control system. The main feature of feedforward control is that it issues control signals in a pre-defined way, so it does not use feedback and correct for errors. Under most circumstances, feedforward and feedback control are combined to realize important functions. For example, consider the process of learning to produce a new sound. At the beginning, the vocal motor system may not have a properly tuned feedforward command for the sound; thus it must rely heavily on the feedback of the auditory and somatosensory systems to adjust the motor commands. On each attempt to
produce that sound, the feedforward command is updated to incorporate the information fed back from the sensory system, which leads to a more accurate motor command for the subsequent attempt. This iterative process continues until the feedforward command is sufficiently accurate so that it causes very few auditory errors during production of that sound. At this point, feedforward control of the process becomes dominant to facilitate rapid (speech) motor output, while feedback information is continuously monitored for possible error correction.

1.2.3 – Neurophysiological and neuroimaging evidence supporting sensorimotor interactions during speech

During speech movements, auditory systems are important for the sensorimotor interactions between motor predictions (i.e., efference copy) and auditory feedback. Indeed, inhibition of auditory neurons during vocalization has been shown in birds, bats, monkeys and humans. For example, Poulet and Hedwig (2002) have demonstrated, using intracellular recordings of auditory neurons in the singing cricket, 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 nucleus of the lateral lemniscus 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 neural 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. Furthermore, the auditory cortical neurons that exhibited suppression during (unaltered)
vocalization exhibited increased activity in response to frequency-shifted feedback during vocalization, suggesting that the heightened sensitivity to vocal feedback may play an important role in auditory self-monitoring (Eliades and Wang, 2008), consistent with an error-detection process.

In humans, reduced responsiveness of temporal-lobe regions during speech production has been demonstrated 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 conducted 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). The attenuation of neural suppression (or increased neural activity) in response to altered auditory feedback also appears to be modulated by the extent of deviation of the altered feedback: the more altered the feedback, the less suppressed the neurons (Behroozmand and Larson, 2011). This modulation is time-sensitive as well, such that temporal delay of the altered feedback also enhances neural responsiveness compared with the same feedback occurring at vocal onset (Behroozmand et al., 2011). All of these results are consistent with the idea that neural suppression in auditory cortical areas during vocalization reflects 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., an efference copy) as part of the sensorimotor control process (e.g., Flinker et al., 2010).

The attenuation of neural suppression in response to altered feedback can also be studied using functional imaging methods such as positron emission tomography (PET) or functional magnetic resonance imaging (fMRI). Using PET imaging, McGuire et al. (1996) found that when their subjects’ speech was distorted while they read aloud, by elevating the fundamental (pitch), there was 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 white masking noise that prevented perception of their own speech. As syllables rate increased, cerebral blood flow increased in several areas associated with speech production as well as in the left planum temporale and posterior peri-Sylvian cortex. The authors interpreted the activation of the temporal regions as evidence for motor-to-sensory discharge (i.e., efference copy) during vocal motor output.

Using fMRI, Hashimoto and Sakai (2003) examined the delayed auditory feedback effect (i.e., auditory feedback is fed back to the ears with a temporal delay during talking) and observed significant bilateral superior temporal gyrus (STG) activation when delayed feedback was compared to normal feedback during 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) explored the neural correlates of auditory self-monitoring using fMRI, and found increased superior temporal activation for distorted voice feedback relative to normal voice feedback or alien voice feedback, during reading aloud. The greater temporal activation associated with the mismatch between
expected and actual auditory feedback may reflect auditory self-monitoring that is involved in the
detection and correction of auditory feedback errors during speech production.

1.2.4 – *Computational models outlining sensorimotor control of speech*

Computational models have been proposed to account for a wide range of behavioral,
neurophysiological and neuroimaging data, and to provide a framework for investigating
sensorimotor control of speech with putative anatomical substrates. One influential neural model
of speech motor control, called ‘Directions Into Velocities of Articulators’ (DIVA), integrates
both feedforward and feedback components into its control system, and provides a quantitative
account of the interactions among motor, somatosensory, and auditory cortical areas during
speech production (Guenther et al., 1998; Guenther, 2006; Golfinopoulos et al., 2011). The DIVA
model posits that, for well-learned speech sounds, the feedforward subsystem is responsible for
generating output controls; the feedback control subsystem only responds to disturbances
(random control variations or external sensory feedback perturbations) that cause sensory
feedback to stray outside the range specified by speech category. Therefore, the central nervous
system actively monitors the sensory feedback (both auditory and somatosensory) during
speaking, and uses that feedback for adjustments to the articulatory trajectory being generated,
particularly when there are disturbances present in sensory feedback.

However, two important limitations of the DIVA model are that 1) the sensory consequences of
motor commands to the vocal articulators may not be precisely predicted (due to motor variance)
and 2) the dynamics of the articulators (i.e., their current velocity or their momentum) are not
taken into account in formulating vocal tract controls (Houde and Nagarajan, 2011). The implicit
assumption of the DIVA model is that the dynamics of the articulators are regulated by the lower
motor system such that desired articulatory trajectories are faithfully executed (Guenther et al.,
1998; Guenther, 2006; Houde and Nagarajan, 2011). However, when correcting for disturbances from the desired output, the DIVA model is not able to anticipate dynamic responses when outputting these corrective control signals (Hickok et al., 2011; Houde and Nagarajan, 2011). This is problematic, especially for fast movements like speech, when articulator dynamics are particularly important for a rapid, smooth execution of motor trajectories (e.g., Gomi and Kawato, 1996).

These issues are very well addressed by a more recent model, named the State Feedback Control (SFC) model of speech (Ventura et al., 2009; Hickok et al., 2011; Houde and Nagarajan, 2011). A fundamental principle of SFC is that control must be based on a running internal estimate of the dynamic state of the system (Jacobs, 1993), which, in the case of speech, requires that the central nervous system maintain an internal model of the dynamics of the vocal tract. This principle is important, because it is only through their effects on the dynamics of the articulators that motor commands affect articulator positions and velocities, and thus auditory and somatosensory feedback from the vocal tract (Houde and Nagarajan, 2011). According to the SFC model, online articulatory control is achieved primarily through an internal model that predicts the effects of motor output based on the current system dynamics, and actual sensory feedback is then used to train and update such an internal model. In other words, the dynamical state relates exerted controls to their sensory outcome, and by monitoring the dynamical state of the vocal articulators, their dynamics can be incorporated into the formulation of control commands. Thus, compared to the DIVA model, the SFC model represents a more fine-grained feedback control structure that endows the CNS with the ability to respond to sensory feedback with much better sensitivity.
Both DIVA and SFC models provide a well-defined architecture that emphasizes the sensorimotor interactions during speech motor planning, execution, and monitoring. In addition, components of both DIVA and SFC have been hypothesized to be associated with specific anatomical regions in the brain (DIVA: Guenther, 2006; SFC: Hickok et al., 2011); this can be examined using neurophysiological methods, including functional neuroimaging. For example, a number of fMRI studies have been conducted to validate the DIVA model (Ghosh et al., 2008; Tourville et al., 2008; Bohland et al., 2010; Golfinopoulos et al. 2011), and the findings of these studies have significantly contributed to the expansion and refinement of the model components. In the meantime, although neuroimaging studies that directly test the SFC model are lacking, converging results in the speech motor control literature support the neural plausibility of the SFC model in speech production. Neuroimaging studies of speech motor control can test the specific anatomical predictions of these computational models and lead to a unified account of sensorimotor control of speech integrating several lines of evidence.

1.2.5 – Experimental and methodological issues with previous neuroimaging studies addressing auditory feedback control of speech

Previous neuroimaging studies of auditory feedback control of speech production based on auditory feedback perturbation paradigms (e.g., McGuire et al., 1996; Paus et al., 1996; Hashimoto and Sakai, 2003; Fu et al., 2006; Christoffels et al., 2007; Toyomura et al., 2007; Tourville et al., 2008; Zheng et al., 2010; Christoffels et al., 2011) suffer from two main limitations. First, altered and normal auditory feedback were not necessarily matched acoustically on duration and length (e.g., McGuire et al., 1996; Paus et al., 1996; Hashimoto and Sakai, 2003; Toyomura et al., 2007), which confounds results with low-level acoustic and/or linguistic processes. Second (and more importantly), previous studies typically compared altered feedback
with normal feedback during production, without explicitly examining the interaction of these feedback conditions under production and listening (i.e., identifying brain areas that are sensitive to the distortion of auditory input during production but not during passive listening). Observing a change in auditory cortical activity during altered, compared with normal, feedback is not sufficient evidence for an online monitoring mechanism. A speech control signal would be generated uniquely during speech production but not during passive listening. Accordingly, it is best identified as an interaction, where the same acoustic difference between altered and normal feedback in the context of talking provokes a different pattern of activity than when identical stimuli are presented in the context of listening.

In order to address the first issue, more recent fMRI studies have adopted real-time feedback perturbation systems that match the timing of ongoing articulation well (Tourville et al., 2008; Zheng et al., 2010; Parkinson et al., 2012). By comparing pitch-shifted vocalization against non-shifted, normal vocalization, Parkinson et al. (2012) observed bilateral superior temporal gyrus and middle temporal gyrus activation. By shifting the first formant frequency (F1) of the subject’s utterances of monosyllabic words, Tourville et al. (2008) found increased activation in posterior superior temporal cortex in response to formant-shifted feedback compared to non-shifted feedback during articulation (note that this effect was not significant at the population level; i.e., it was evident in group data, but only when subjects were treated as a fixed effect). For the second issue (i.e., interaction), Christoffels et al. (2011) used a design in which both overt speaking and passive listening were paired with four levels of externally generated acoustic noise as input. The noise level was parametrically varied across four intensity levels, with the first level set to zero intensity (i.e., normal auditory feedback) and the fourth level to maximum intensity (i.e., noise feedback). The authors found that the auditory cortical response was specific, such that the
The magnitude of the neural activation in the superior temporal regions varied in proportion to the parametric level of the noise mask during overt speech (but not during passive listening). This pattern of neural activation appeared to represent an interaction between tasks and levels of acoustic mismatch, even though such an interaction was not directly tested.

Zheng et al. (2010) addressed both issues, and was the first to employ a 2x2 factorial design to explicitly test an interaction between Task (production or listening) and Feedback type (masking noise or clear speech). In the production conditions, subjects whispered a consonant-vowel-consonant word (‘ted’), and heard either normal auditory feedback, or masking, signal-correlated noise that was temporally gated with their production (with the aid of a real-time signal processing system; see Purcell et al., 2006a,b). Whispered speech was used to minimize the audible bone-conducted sound during vocalization (Barany, 1938). To ensure that the production components were properly controlled for, two other listening-only conditions were yoked to the production conditions, with identical auditory feedback (clear ‘ted’, or noise) as the corresponding production trials. As expected, a significant interaction between Feedback Type (clear speech or noise) and Task (production or listening) was observed in bilateral superior temporal regions (i.e., areas in which the difference in activity between production with noise and production with normal speech is significantly greater than the difference in activity between listening to noise and to speech) (see Figure 1.1). This pattern of interaction indicated that, consistent with the findings of previous studies of auditory self-monitoring, the superior temporal regions appear to be involved in the generation and processing of an error signal caused by the mismatch between the predicted and actual sensory consequences of vocalization. In addition, a portion of the temporal regions were also found to be more strongly activated by speech than by noise during passive listening (i.e., speech-sensitive), suggesting that brain regions involved in
online articulatory control appear to overlap with the regions supporting speech processing (see Figure 1.1) – a finding that highlights the close link between speech production and perception.

Figure 1.1. Functional overlap between brain areas that are speech sensitive and areas supporting auditory feedback control of speech is shown. Areas in which the difference in activation between noise and clear speech is significantly greater in production than in listening conditions (i.e., interaction) are shown in blue. Areas that are activated more by listening to speech than by listening to noise are shown in green. The activation map is thresholded at $p < 0.001$ uncorrected and overlaid on the group average structural image.

1.3 – Sensorimotor interactions for perception of non-speech aspects of auditory feedback during vocalization

The discussion so far has been restricted to the perception of speech during production. However, the study of the relationship between perception and production can also be extended to the non-speech aspects of articulation (e.g., speaker identity). This extension is important for ongoing debates about the modularity of speech as well as for the burgeoning field of social neuroscience. After all, articulation is not just about conveying the content of messages; it is also
about conveying socially relevant cues such as the identity of the talker and his or her emotional state (Lass et al., 1976; Nygaard and Pisoni, 1998; Belin et al., 2004; Grandjean et al., 2005; Smith et al., 2005).

1.3.1 – Theoretical and empirical distinction between speech and non-speech aspects of articulation

Traditionally, a clear distinction has been made between the “linguistic” features of an utterance (i.e., features that are used to map phonetic structure onto higher-level words or sentences) and its “paralinguistic/indexical” features (i.e., non-verbal features that encode talker identity and emotion) (Abercrombie, 1967). Support for this view comes from neuropathological evidence that demonstrates “pure word deafness” in clinical patients with preserved speaker identification ability (Gazzaniga et al., 1973), indicating that processing of the linguistic content of speech can be dissociated from the processing of voice-specific information for the purpose of talker identification (Laver and Trudgill, 1979).

Indeed, in most voice perception studies, speech is simply treated as the carrier of talker-specific information (e.g., Legge et al., 1984; van Lancker et al., 1985; Schweinberger et al., 1997). These studies typically assess the listener’s ability to discriminate familiar and unfamiliar voices without taking into account the role of the speech content in the perceptual process. On the other hand, speech perception studies are mostly focused on the linguistic content of speech, and explore the extraction and representation of the acoustic signals in the context of speech intelligibility (e.g., Davis and Johnsrude, 2007; Scott et al., 2009; Pulvermüller and Fadiga, 2010). This theoretical and empirical dissociation between the encoding of content- and talker-relevant information seems to be supported by neuroimaging studies revealing distinct, or at least partially dissociable, neural substrates for the processing of different aspects of voices (von
Kriegstein et al., 2003; von Kriegstein and Giraud, 2006; Formisano et al., 2008). For example, by using a sophisticated machine-learning framework, Formisano et al. (2008) demonstrated the feasibility of decoding (and discriminating) speech content (“what” is being said) and speaker identity (“who” is saying it) based on the neural patterns of brain activity in the auditory cortex. This finding suggests that “what” and “who” aspects of speech may be represented as segregated processes in the brain.

1.3.2 – Interactions between speech and non-speech processes during perception

However, during verbal communication, we not only extract the underlying messages; but additionally (and simultaneously) pay attention to an extensive amount of information about the individual who is speaking (e.g., Scharinger et al., 2011). This raises the question of whether the linguistic and paralinguistic processes may be interacting. Empirical evidence appears to suggest that this indeed is the case. A growing body of work on the effects of talker characteristics has shown that identity information contained in the speech may have a strong impact on both word recognition (Mullennix et al., 1989; Mullennix and Pisoni, 1990) and word recall later (Martin et al., 1989; Goldinger et al., 1991). By familiarizing participants with a set of novel voices, Nygaard et al. (1994) showed that the ability to identify a talker’s voice improved the intelligibility of novel words produced by that talker, suggesting that talker-contingent process is not ‘stripped away’ during speech perception; instead it may become part of a rich representation of the talker’s utterance (Nygaard et al., 1994; Nygaard and Pisoni, 1998; Bradlow et al., 1999; Nygaard and Queen, 2008). This account is also consistent with recent neuroimaging findings showing that speaker identity information may form an integral part of the neural representation of words (e.g., Chandrasekaran et al., 2011). Given the extensive literature on interactions between speaker-specific information and speech content, I intend to explore whether speaker
identity is an aspect of what is perceived from auditory concomitants of speech (i.e., assessing how who is saying it is modulated by what is being said).

1.3.3 – *Perceiving the voice identity of the auditory concomitant of speech (self-voice identification)*

Most previous studies exploring the cognitive and neural mechanisms of speech content v.s. talker-specific information have focused on the perception of someone else’s voice (Belin et al., 2004; Belin et al., 2011; Latinus and Belin, 2011). By comparison, relatively less emphasis has been placed on perception and recognition of one’s own voice. Furthermore, in studies that have investigated own-voice processing (e.g., Kaplan et al., 2008; Rosa et al., 2008; Aruffo and Shore, 2012), the target voice was typically presented for perception while participants were passively listening; in contrast, much less attention has been paid to the perception of voice through the auditory feedback resulting from a participant’s own vocalization.

Vocal production provides rich sensory (i.e., auditory and somatosensory) feedback information which, when combined with the motor representations arising from articulation, can significantly contribute to the recognition of one’s own voice through the auditory self-monitoring mechanism (Frith et al., 2000; Blakemore et al., 2002). The idea is that during vocal production, a sensory prediction (i.e., an efference copy) of the outcome of the motor output is produced along with the actual motor command. The results of the comparison between the sensory prediction and the sensory consequences of the vocal production can then be used to determine the ‘origin’ of the voice being produced – the congruency and temporal alignment between what is produced and what is heard would give rise to the perception that ‘it is my own voice’). Therefore, it is both theoretically interesting and ecologically valid to investigate the
perception and identification of voices presented as an auditory concomitant of speech production.

1.4 – Multimodal integration and individual differences in supporting perception of the voice identity of auditory concomitant of speech

With the addition of motor production, the perceptual process of voice identification during vocalization becomes inherently multimodal (i.e., integration of auditory, somatosensory and motor cues for perception). This begs the question of what is the underlying mechanism supporting self-voice recognition while talking: is it multimodal integration? Or is it simply individual differences in suggestibility (Benton and Bandura, 1953; Braffman and Kirsch, 1999; Hergovich, 2003)? In other words, does the perception of the identity of the voice giving auditory feedback depend on how well an individual integrates multimodal cues or on how suggestible he/she is to hearing someone else’s voice as their own?

Given the fact that self-voice identification is not only a problem of voice perception, but also a problem of self-attribution (e.g., Tsakiris, 2008; Longo and Haggard, 2012), we may use established paradigms in the general field of self-recognition to help provide an answer to the aforementioned question. The rubber voice illusion (RHI: Botvinick and Cohen, 1998; Ehrsson et al., 2005) - a phenomenon where temporally coincident visual and somatosensory inputs (i.e., the feeling of someone stroking the fingers of your own hidden hand, and the sight of a prosthetic hand being stroked the same way) combine to create the perception of self-body ownership – is one of such paradigms. It has been suggested that the induction of the RHI strongly depends on the integration of the multisensory (i.e., visual, tactile, and proprioceptive) cues – a conclusion given credence by the observations that asynchronous stroking of the rubber hand and the hidden hand markedly attenuates the illusion. It has later been shown that multisensory integration
appears to be the general mechanism for the perception of self-body ownership (e.g., Makin et al., 2008; Slater et al., 2009; Petkova et al., 2011). If this is the case, then would self-voice identification during talking also rely on the integration of multimodal (i.e., motoric, auditory, and somatosensory) information?

One possible way to address this question would be to compare, within subjects, the perception of an alien voice presented as an auditory concomitant of speech with the perception of a prosthetic hand under multisensory inputs. If the perceived strengths of the two otherwise very different illusions are strongly correlated, it might suggest that people perceive the identity of the alien voice as a result of multimodal convergence, just like how they incorporate the prosthetic hand into their own bodily system based on multisensory inputs. To control for general task complexity, I can also employ an illusion (Lackner, 1988) to evaluate how strongly one’s body ownership might be influenced by external sensory/motor perturbation. Furthermore, I can use an established subjective suggestibility scale (Kotov et al., 2004) to assess whether and how perception of self-voice is related to individual differences in suggestibility. This exploration would offer insight into the factors contributing to the identification of the voice as auditory concomitants of speech, and provides a general framework for future studies on body ownership during motor movements.
2.1 – Introduction

Using a real-time speech tracking system and a 2x2 factorial design, Zheng et al. (2010) assessed the cortical response when the same acoustic difference between masking noise and normal speech feedback was compared during whispered vocalization and during passive listening. Strong bilateral superior temporal activation was observed in response to masking noise feedback, uniquely during vocalization but not during passive listening, presumably reflecting neural processes implicated in speech error detection and correction. The focus of the design was an interaction between Feedback Type (masking noise and normal speech) and Task (speech production and passive listening) that captured the characteristics of a speech error signal - this interaction provided a way of assessing the auditory feedback error response while controlling for both motor production processes and acoustics.

Although Zheng et al. (2010) improved over previous studies methodologically (with the aid of a real-time speech signal processing system) and experimentally (by utilizing a design centered on the test of an interaction), there were a few limitations. For example, whispered speech was used to control the bone-conducted acoustic signals in the production conditions. However, the extent to which this may have affected the results is not clear – previous studies have shown that the level of neural activation in auditory regions evoked by whispered speech and by vocalized (overt) speech can be somewhat different (e.g., Haslinger et al., 2005). Furthermore, signal-correlated masking noise was used to generate an auditory feedback error, compared to clear speech feedback. However, envelope-matched noise and speech stimuli differ on multiple levels, including acoustic properties, lexico-semantic value, and even familiarity. A condition that more
closely matches the real speech would allow us to better isolate processes specific to auditory feedback errors. In addition, the inclusion of another erroneous feedback condition would also afford us the opportunity to examine whether, and how, the magnitude of neural response varied according to the level of distortion in the auditory feedback during articulation, thereby extending the results of Zheng et al. (2010) in which only a single level of distortion was used.

Therefore, the main goal of the present study was to explore, in more detail, the speech error response system in the brain, by 1) using normal, instead of whispered, speech during production, and 2) including another distorted feedback condition – formant-shifted speech (Houde and Jordan, 1998). Here, the formant-shifted speech refers to the CVC words for which certain frequency components of the vowel sounds are experimentally shifted. The formant-shifted speech is perceived as speech, and matches the normal speech condition well in terms of temporal and spectral envelopes. The addition of this feedback type would allow for 1) a better assessment of the neural basis of speech error response, with a better acoustic control; and 2) an examination of multiple interaction patterns related to distinct auditory feedback error types.

The present study employed a factorial design plus a low-level rest control condition. In addition to the 2x2 conditions presented in Zheng et al. (2010), formant-shifted speech was delivered both during production and during passive listening, giving rise to a 2x3 factorial design (production and listening crossed with three types of acoustic input: clear speech, signal-correlated masking noise, and formant-shifted speech) (see Table 2.1). Following the logic of Zheng et al. (2010), the focus of the current study was the three sets of interactions between Feedback Type and Task within the 2x3 grid, in order to specifically explore the processes related to different feedback processing during vocalization. However, I will first attempt to replicate the findings of Zheng et al. (2010), where the interaction contrasting masking noise feedback and
clear speech feedback was assessed, before moving on to the other two sets of interactions involving the newly-added formant-shifted speech condition.

Table 2.1. The 2x3 factorial design used in the present study is shown in the table. Three types of acoustic input (clear speech, formant-shifted speech, and signal-correlated masking noise) are crossed with two tasks, speech production and passive listening.

<table>
<thead>
<tr>
<th></th>
<th>Production</th>
<th>Listening</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clear Speech</strong></td>
<td>Production with clear speech feedback ($production-clear$)</td>
<td>Listening to clear speech ($Listen-clear$)</td>
</tr>
<tr>
<td><strong>Formant-shifted speech</strong></td>
<td>Production with shifted speech feedback ($production-shift$)</td>
<td>Listening to shifted speech ($Listen-shift$)</td>
</tr>
<tr>
<td><strong>Signal-correlated Masking Noise</strong></td>
<td>Production with shifted speech feedback ($production-noise$)</td>
<td>Listening to noise ($Listen-noise$)</td>
</tr>
</tbody>
</table>

### 2.2 – Methods and Materials

#### Participants

Written informed consent was obtained from twenty participants (mean age 21 years, range 19-27 years; 8 females). All were right handed, without any reported history of neurological or hearing disorder, and all spoke English as their first language. Each participant received $15 to compensate them for their time. Procedures were cleared by the Queen’s Health Sciences Research Ethics Board.

#### Experimental Design and Stimuli

---

1 Since this and the next chapter are based on the same dataset, the Experimental Design and Stimuli sections of the two chapters overlap extensively. Some of the content presented in this section is also presented in Chapter 3 because Chapter 3 is a submitted manuscript and therefore stands on its own.
I employed a 2x3 within-subject factorial design with 6 experimental conditions, plus a low-level silence/rest control condition. Two tasks (speech production and passive listening) were crossed with three types of auditory signal (formant-frequency-shifted speech, signal-correlated noise, and normal speech), which served as auditory stimuli during listening conditions and as temporally gated auditory feedback during production conditions.

For the production with normal speech condition (production-clear), participants vocalized “had” and heard unaltered auditory feedback. For the production with formant-shifted speech condition (production-shift), participants vocalized “head” but heard processed feedback such that the first (F1) and second (F2) formants of the vowel were shifted by -200 Hz and +250 Hz respectively (i.e., ‘head’ shifted towards ‘hid’). The direction and magnitude of the shifts were chosen based on empirical studies of formant perturbations (MacDonald et al., 2010, 2011). Previous studies have shown that participants would compensate for the formant shift by producing in an opposite direction of the shift, i.e., when they hear ‘head’ shifted towards ‘hid’, they would likely compensate by producing ‘had’. For the production with signal-correlated noise condition (production-noise), participants produced either “had” or “head”, and heard masking noise temporally gated with the onset and offset of their vocalizations. Having participants produce ‘had’ half the time and ‘head’ half the time in this condition ensured that any difference in response patterns between distorted- and normal-feedback production conditions could not simply be due to a difference in what the participant was asked to say. For yoked listening conditions (listen-clear, listen-shift and listen-noise, respectively), participants heard auditory feedback that they had previously heard during matched production conditions.

In the three production conditions, participants spoke into an optical microphone (Phone-Or, MA) and their utterances were digitized at 10 kHz with 16-bit precision using a National
Instruments PXI-6052E data acquisition board (National Instruments, TX). Real-time analysis was carried out using a National Instruments PXI-8176 embedded controller with which processed acoustic signals were converted back to analog by the data acquisition board and played over high-fidelity magnet-compatible headphones (NordicNeuroLab, Bergen, Norway) in real time. The processing delays were too short (iteration delay less than 10 msec) to be noticeable to the participants. The processed signals were also recorded and stored on an IBM ThinkPad X32 laptop (IBM, USA) to be used as yoked stimuli during listening conditions.

For the production-shift condition, vowel formants were estimated with an iterative Burg algorithm (Orfanidis, 1988) and formant shifting was implemented using an infinite impulse response (IIR) filter in real-time (see Purcell and Munhall, 2006a,b). For the production-noise condition, masking noise was generated by applying the amplitude envelope of the utterance to uniform Gaussian white noise. This ensures that the noise level was utterance-specific and exactly intense enough at every moment to mask the energy of the spoken words. The resulting ‘signal-correlated noise’ (Schroeder, 1968) has the same spectral profile and amplitude envelope as the original speech signal, but is completely unintelligible.

**Image Acquisition and Experimental Procedures**

Functional magnetic resonance imaging data were collected on a 3-T Siemens Trio MRI system, using a rapid sparse-imaging paradigm (Orfanidou et al., 2006). This paradigm allowed us to present our auditory stimuli and record responses during a 1600-msec silent period between successive 1600-msec acquisitions in order to minimize acoustic interference (GE-EPI sequence; TR 3200 msec; TA 1600 msec; 211 x 211 matrix; in plane resolution 3.3 x 3.3 mm²; 26 transverse slices with interleaved acquisition). A high-resolution T1-weighted magnetization-
prepared rapid gradient echo (MPRAGE) structural scan was also acquired in each participant (TR 1760 msec, TE 2.6 msec, voxel resolution 1.0 x 1.0 x 1.0 mm³, flip angle 9°).

Trials were presented in the silent period between successive scans. Participants were instructed to pay attention to a rectangular, grey prompt in the middle of a computer screen that appeared 100 msec prior to the offset of each scan, which signaled the onset of the next 1600-msec silent period/trial (see Figure 2.1). Depending upon the condition type, the prompt either contained a word, or was blank. Participants were asked to produce the word when there was one, and remain silent when there was not one presented. The words “had” and “head” were shown in the production-clear and production-shift conditions respectively. The words “had” or “head”, presented with equal probability, were shown in the production-noise condition. The prompt was blank during listen-clear, listen-shift, listen-noise, and rest conditions.

Figure 2.1. Schematic diagram of the first 10 seconds from a functional run with predicted hemodynamic response function (HRF) is shown. The prompts (indicated by an arrow) appears on a computer screen 100 msec before scan offset and cues the participant to either produce a word (word prompt) or remain silent (blank prompt). The 1600-msec long scans are separated by the 1600-msec long silent periods permitting speaking and listening.
Participants were scanned in three functional runs, each lasting 9 min and comprising 24 blocks of 7 trials. Within each block, conditions were presented in pseudorandom order, such that transitional probabilities from one condition to another were approximately equal and one trial from each of 7 conditions were presented (conditions could repeat at the transition from one block to the next). The stimuli for listening conditions were taken from the recordings of the matched production trials in the preceding block, except for the first block, where the listen stimuli were taken from the production trials in the same block, i.e., production trials always preceded listening trials in this first block.

Each participant practiced three blocks (21 trials) of the experiment in the scanner before scanning commenced. During scanning, their behavior on each trial was monitored in real-time by the experimenter (ZZZ) for possible errors. Participant’s vocal production and auditory feedback signals were segregated into two different channels and therefore monitored and recorded simultaneously.

**Behavioral Data Analysis**

Recordings from vocal production and auditory feedback on each trial were reviewed to ensure that incorrect trials were identified and properly accounted for. Trials were considered incorrect if 1) participants made production errors (i.e., failed to produce, produced wrong words, or vocalized during listening/rest trials); or 2) auditory feedback was not triggered due to the very low magnitude of the vocal production (no-trigger trials). In fact, no participant made any errors of the former type, production errors (see Results).

**Functional Data Analysis**

SPM2 ([www.fil.ion.ucl.ac.uk/spm/spm2.html](http://www.fil.ion.ucl.ac.uk/spm/spm2.html)) was used for functional data analysis and visualization. Data were first realigned, within subjects, to the first true functional scan of the
session (after discarding 2 dummy scans), and individual’s structural image was coregistered to the mean fMRI image. The coregistered structural was spatially normalized to a standard template (ICBM152), and the realigned functional data were normalized using the same deformation parameters. The fMRI data were then smoothed using a Gaussian kernel of 10 mm (full-width at half maximum, or FWHM).

Data from each participant were entered into a fixed-effects single-subject general linear model using an event-related analysis procedure (Josephs and Henson, 1999). Six event types were modeled for each run. Incorrect trials, if present, were also modeled as a covariate of no interest for each run to reduce the error variance. In addition, six parameters from the motion correction (realignment) stage were included as regressors in the model to ensure that variability due to head motion was properly accounted for. This model was convolved with the hemodynamic response function (HRF) and fitted to the MR time series in each voxel, resulting in parameter estimates indexing the magnitude of response to the experimental conditions. A high-pass filter (cut-off 128 sec) and AR(1) correction for serial autocorrelation were applied. Contrast images were created for each participant and were entered into random-effects analysis (one-sample t-tests) comparing the mean parameter-estimate difference over subjects to zero. Clusters were deemed significant if they exceeded a statistical threshold of p < .05 after correction for multiple comparisons (Family-wise error correction, or FWE) at the cluster level (Worsley, 1996).

2.3 – Results

Behavioral Data

Both real-time monitoring during scanning and inspection of recorded behavioral data indicated that participants followed the instructions in all trials. In production trials, they did not produce the incorrect word, or fail to produce any word, and in listening trials they did not produce
speech. There were, however, some production trials in which the vocalization failed to trigger the delivery of auditory feedback, i.e., no-trigger trials. There was no significant difference in the distribution of no-trigger trials across the three runs, $F(2, 18) = 2.419, p = .117$.

**Functional Data**

*Interaction: (production-noise – production-clear) vs (listen-noise – listen-clear)*

This was the main interaction tested in Zheng et al. (2010), in which bilateral superior temporal gyrus (STG) exhibited a greater increase in activity for masking noise compared to normal speech, when these were heard as the auditory concomitants of one’s own utterances compared to when they were presented during passive listening.

Here, analysis of the same interaction effect revealed significant clusters in right STG, and bilateral insula (see Table 2.2 and Figure 2.2). There was a cluster in the left STG/STS that did not survive the cluster-level statistical threshold (FWE), but the peak of the cluster survived a statistical threshold of $p < .05$ at the voxel level, FDR corrected for multiple comparisons across the whole brain (Genovese et al., 2002).

Table 2.2. Areas in which the difference in activity between *production-noise* and *production-clear* is significantly greater than the difference in activity between *listen-noise* and *listen-clear*. STG: superior temporal gyrus.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>t-statistic</th>
<th>No. of voxels</th>
<th>Anatomical area</th>
</tr>
</thead>
<tbody>
<tr>
<td>-33</td>
<td>21</td>
<td>3</td>
<td>6.26</td>
<td>337</td>
<td>Left Insular Cortex</td>
</tr>
<tr>
<td>48</td>
<td>-33</td>
<td>3</td>
<td>5.96</td>
<td>202</td>
<td>Right STG</td>
</tr>
<tr>
<td>39</td>
<td>21</td>
<td>0</td>
<td>5.83</td>
<td>216</td>
<td>Right Insular Cortex</td>
</tr>
</tbody>
</table>
Figure 2.2. Areas in which the difference in activity between production-noise and production-clear is significantly greater than the difference in activity between listen-noise and listen-clear. The activation map is thresholded at $p < .001$ uncorrected for multiple comparisons across the whole brain, overlaid on the group average structural image. The color scale indicates the p-value corrected for multiple comparisons using the family-wise error (FWE) correction (Worsley, 1996), such that any color in the red-orange-yellow range on the color scale is statistically significant ($p < .05$) at the whole-brain level.

In order to better understand how differences among conditions produced this significant interaction, I explored the simple effects that constitute this interaction. I observed that, within the
regions of left supplementary motor area, left cerebellum, right inferior frontal gyrus (IFG), and right middle temporal gyrus, activation for production-noise was significantly greater than for production-clear (see Table 2.3). As expected, the contrast of (listen-clear – listen-noise) yielded strong activation in the speech-sensitive areas in bilateral STG, extending into the middle temporal gyrus (see Table 2.4).

Table 2.3. Areas that show enhanced activation for production-noise relative to production-clear. MTG: middle temporal gyrus; SMA: supplementary motor area; IFG: inferior frontal gyrus.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>t-statistic</th>
<th>No. of voxels</th>
<th>Anatomical area</th>
</tr>
</thead>
<tbody>
<tr>
<td>45</td>
<td>-48</td>
<td>18</td>
<td>5.98</td>
<td>108</td>
<td>Right MTG</td>
</tr>
<tr>
<td>0</td>
<td>12</td>
<td>60</td>
<td>5.33</td>
<td>220</td>
<td>Left SMA</td>
</tr>
<tr>
<td>33</td>
<td>33</td>
<td>0</td>
<td>5.23</td>
<td>121</td>
<td>Right IFG</td>
</tr>
<tr>
<td>-18</td>
<td>-66</td>
<td>-51</td>
<td>4.91</td>
<td>163</td>
<td>Left Cerebellum</td>
</tr>
</tbody>
</table>

Table 2.4. Areas that are more strongly activated for listen-clear than for listen-noise. STG: superior temporal gyrus; MTG: middle temporal gyrus.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>t-statistic</th>
<th>No. of voxels</th>
<th>Anatomical area</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>-33</td>
<td>0</td>
<td>10.53</td>
<td>438</td>
<td>Right STG/MTG</td>
</tr>
<tr>
<td>-63</td>
<td>-21</td>
<td>0</td>
<td>9.82</td>
<td>362</td>
<td>Left STG/MTG</td>
</tr>
</tbody>
</table>

Interaction: (production-clear – production-noise) vs (listen-clear – listen-noise)

The opposite interaction contrast did not yield any significant clusters in the brain.

Production vs Listening: (production-noise + production-clear + production-shift) vs (listen-noise + listen-clear + listen-shift)

The main effect of (production – listening) activated an extremely large region involving both hemispheres, consistent with the previous finding (Zheng et al, 2010). As in Zheng et al. (2010), I
increased the threshold to $p < 10^{-11}$ for this contrast – this revealed significant clusters in a number of areas including bilateral insula, bilateral thalamus, bilateral postcentral gyrus, left supplementary motor area, left superior temporal gyrus, and right rolandic operculum (see Table 2.5 and Figure 2.3). The reverse contrast, in which activity during production conditions was subtracted from that during listening conditions, did not reveal any significant activation.

Table 2.5. Areas that are more strongly activated for production than for listening. Cluster peaks are reported if they exceeded a statistical threshold of $p < 10^{-11}$ threshold after correction for multiple comparisons at the cluster level. SMA: supplementary motor area; STG: superior temporal gyrus.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>t-statistic</th>
<th>No. of voxels</th>
<th>Anatomical area</th>
</tr>
</thead>
<tbody>
<tr>
<td>-36</td>
<td>0</td>
<td>3</td>
<td>21.22</td>
<td>97</td>
<td>Left Insular Cortex</td>
</tr>
<tr>
<td>39</td>
<td>6</td>
<td>3</td>
<td>19.77</td>
<td>113</td>
<td>Right Insular Cortex</td>
</tr>
<tr>
<td>12</td>
<td>-15</td>
<td>3</td>
<td>19.17</td>
<td>115</td>
<td>Right Thalamus</td>
</tr>
<tr>
<td>-18</td>
<td>-9</td>
<td>-3</td>
<td>17.88</td>
<td>85</td>
<td>Left Thalamus</td>
</tr>
<tr>
<td>-48</td>
<td>-18</td>
<td>30</td>
<td>17.86</td>
<td>84</td>
<td>Left Postcentral Gyrus</td>
</tr>
<tr>
<td>57</td>
<td>-6</td>
<td>12</td>
<td>17.01</td>
<td>39</td>
<td>Right Rolandic Operculum</td>
</tr>
<tr>
<td>57</td>
<td>-6</td>
<td>30</td>
<td>15.45</td>
<td>19</td>
<td>Right Postcentral Gyrus</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>63</td>
<td>15.32</td>
<td>15</td>
<td>Left SMA</td>
</tr>
<tr>
<td>-57</td>
<td>-12</td>
<td>12</td>
<td>14.50</td>
<td>7</td>
<td>Left STG</td>
</tr>
</tbody>
</table>
Figure 2.3. Areas that are more strongly activated for production than for listening. The activation map is thresholded at $p < 10^{-11}$ uncorrected for multiple comparisons and overlaid on the group average structural image. The color scale indicates the p-value corrected family-wise error (FWE), such that any color brighter than dark blue survived a statistical threshold of $p < .001$ (FWE) at the whole-brain level.

Speech vs Noise: $(production-clear + listen-clear) vs (production-noise + listen-noise)$

This contrast yielded strong bilateral STG clusters extending into middle temporal gyrus (MTG) (Table 2.6 and Figure 2.4). In the left hemisphere, the peak of the STG cluster was anatomically close to the peak of the STG cluster revealed in the contrast $(listen-clear – listen-noise)$: (-63 -30 3) vs (-63 -21 0), suggesting that both these two speech-related contrasts might reveal putative brain regions recruited for speech processing. On the other hand, brain regions more responsive to noise than to speech included left middle occipital cortex, left superior parietal cortex, right
angular gyrus, right superior frontal gyrus, right supramarginal gyrus, and right putamen (see Table 2.7). Significant activation in auditory regions for this contrast was not observed.

Table 2.6. Areas that are more strongly activated for *speech* (*production-clear + listen-clear*) than for *noise* (*production-noise + listen-noise*). STG: superior temporal gyrus; MTG: middle temporal gyrus.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>t-statistic</th>
<th>No. of voxels</th>
<th>Anatomical area</th>
</tr>
</thead>
<tbody>
<tr>
<td>-63</td>
<td>-30</td>
<td>3</td>
<td>11.33</td>
<td>392</td>
<td>Left STG/MTG</td>
</tr>
<tr>
<td>63</td>
<td>-12</td>
<td>0</td>
<td>8.88</td>
<td>398</td>
<td>Right STG/MTG</td>
</tr>
</tbody>
</table>

Table 2.7. Areas that are more responsive to *noise* (*production-noise + listen-noise*) than to *speech* (*production-clear + listen-clear*). MOC: middle occipital cortex; SFG: superior frontal gyrus; SPC: superior parietal cortex.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>t-statistic</th>
<th>No. of voxels</th>
<th>Anatomical area</th>
</tr>
</thead>
<tbody>
<tr>
<td>-42</td>
<td>-81</td>
<td>18</td>
<td>9.60</td>
<td>1717</td>
<td>Left MOC</td>
</tr>
<tr>
<td>24</td>
<td>-60</td>
<td>45</td>
<td>8.74</td>
<td>922</td>
<td>Right Angular Gyrus</td>
</tr>
<tr>
<td>18</td>
<td>9</td>
<td>66</td>
<td>7.30</td>
<td>558</td>
<td>Right SFG</td>
</tr>
<tr>
<td>60</td>
<td>-27</td>
<td>30</td>
<td>6.76</td>
<td>282</td>
<td>Right Supramarginal Gyrus</td>
</tr>
<tr>
<td>-18</td>
<td>-69</td>
<td>42</td>
<td>5.92</td>
<td>346</td>
<td>Left SPC</td>
</tr>
<tr>
<td>33</td>
<td>-15</td>
<td>-9</td>
<td>5.34</td>
<td>317</td>
<td>Right Putamen</td>
</tr>
</tbody>
</table>
Figure 2.4. Areas that are more strongly activated for speech than for noise. The activation map is thresholded at $p < .001$ uncorrected and overlaid on the group average structural image. The color scale indicates the p-value corrected family-wise error (FWE), such that any color brighter than dark blue survived a statistical threshold of $p < .001$ (FWE) at the whole-brain level.

Now, I will explore the other two sets of interactions involving the formant-shifted speech (which is the focus of the present study), to see whether the same brain regions involved in processing auditory feedback error (in the form of masking noise) are also involved in processing the feedback error that is acoustically speech (i.e., formant-shifted speech).

Interaction: (production-shift – production-clear) vs (listen-shift – listen-clear)
This contrast was expected to reveal brain regions that were more strongly activated for shifted speech than for normal speech feedback, only during production but not during listening – these regions would presumably be recruited for the processing of an auditory feedback error that is acoustically speech. However, I only observed a marginally significant cluster (p = .068) located in the left caudate nucleus. The opposite interaction contrast \((production-clear – production-shift) vs (listen-clear – listen-shift)\) failed to yield any significant clusters.

*Interaction: \((production-noise – production-shift) vs (listen-noise – listen-shift)\)*

This contrast was used to assess whether the two types of auditory feedback error would differentially engage distinct brain regions during production. However, neither this nor the opposite interaction revealed any significant clusters.

### 2.4 – Discussion

Motivated by methodological and experimental limitations of Zheng et al. (2010), the present study was designed to both replicate the findings of the previous study and explore the brain regions sensitive to auditory feedback errors in more detail. Although the observed patterns of brain activation are in general agreement with those shown in Zheng et al. (2010), the current study does not yield substantially more information regarding whether and how distinct feedback error types might be represented in the brain.

The bilateral STG activation observed in the main interaction contrast in Zheng et al. (2010) was also observed in the same contrast in the present study – this activation has been interpreted as the cortical response to auditory feedback error during articulation and has been replicated in a number of studies of auditory feedback monitoring (e.g., Fu et al., 2006; Christoffels et al., 2007; Tourville et al., 2008; Christoffels et al., 2011; Parkinson et al., 2012). It is worth noting that, in the present study, I also observed bilateral insula activation that was not evident in Zheng et al.
(2010). This enhanced response in the insular cortex is not surprising, as this region has been repeatedly documented in studies of sensorimotor control (e.g., Anderson et al., 1994; Fink et al., 1997; Riecker et al., 2000; Ackermann and Riecker, 2004). For example, Riecker et al. (2000) found the activation of the right anterior insula when participants were reproducing a non-lyrical tune whereas the left insula was activated by overt production of speech. The authors argue that the right insula might mediate temporo-spatial control of vocal tract musculature during overt singing and the left insula appears to support the coordination of speech articulation. Given the fact that both speech production and singing require sophisticated sensorimotor planning and control, it is plausible that the bilateral insula was also activated in a situation that required a particularly high level of articulatory control, when auditory feedback was masked by noise during speaking.

On the other hand, the two interaction contrasts involving the formant-shifted speech did not yield meaningful and interpretable patterns of brain activation. The formant-shifted speech, albeit acoustically similar to normal speech, was erroneous and thus was expected to activate the brain regions recruited for speech error processing, such as bilateral STG. Furthermore, the interaction contrast involving both formant-shifted speech and masking noise feedback was also expected to engage the brain regions implicated in the differential processing of distinct feedback error types, which might or might not include bilateral STG. The failure to identify significant activation for these two interaction contrasts at the whole-brain level might be due to the lack of statistical power and/or the loss of voxel spatial pattern information through the averaging of neural activation across multiple voxels in conventional imaging analysis – which can be addressed by adopting a more sensitive analytic framework. Multi-voxel pattern analysis (MVPA) is such a
technique. Below, I will provide a brief overview of the general MVPA methodology – this overview is also background to the analysis described in Chapter 3.

Conventional imaging studies use an analytic approach in which brain activation from a large patch of neural tissue is averaged, and then levels of activation between different conditions are compared. This type of analysis may obscure distinct patterns of functional specificity within a given brain region and lead to non-significant results that could have been identified by a more sensitive analysis approach such as MVPA. Unlike conventional methods of fMRI analysis where statistical tests are used to determine whether a voxel is more responsive to one experimental condition versus another, MVPA is concerned with many voxels all at once and considers the spatial pattern of the neural activation within a given brain region. Conventional analyses are univariate such that the analysis of one voxel has no effect on that of any other voxel; whereas MVPA is multivariate in nature and is based on extracting the information contained in the patterns of neural activity (i.e., the relative differences in activity among voxels contain relevant information) (Haynes and Rees, 2006; Norman et al., 2006; Kriegeskorte and Bandettini, 2007; Mur et al., 2009; Rissman et al., 2012; Tong and Pratte, 2012). Therefore, MVPA can be used to distinguish the spatial neural patterns elicited by different experimental conditions, even if the average/aggregate activity does not differ among these conditions (e.g., Serences et al., 2009; Staeren et al., 2009; Abrams et al., 2011; Gallivan et al., 2011; van den Hurk et al., 2011; Tamber-Rosenan et al., 2011).

Haxby et al. (2001) was among the first to use distributed pattern information in the object-sensitive regions in the ventral temporal cortex to discriminate whether subjects were viewing pictures of faces, houses, chairs, cats, bottles, shoes, scissors, or nonsense objects. Based on comparison of within-category and between-category correlations of neural patterns, the authors
revealed that different categories of objects are distinctly represented in the object-selective cortex. This finding has later been replicated and extended by a number of other studies using MVPA (Carlson et al., 2003; Cox and Savoy, 2003; O’Toole et al., 2005; Downing et al., 2007; Kriegeskorte et al., 2007; Williams et al., 2007; Peelen et al., 2009; Reddy et al., 2009). Subsequently, a few studies (Harrison and Tong, 2009; Kamitani and Tong, 2005; Mannion et al., 2009) found that orientation and direction-sensitive neural responses can be decoded with remarkable precision in the early visual cortex, even though such feature-selective information is organized at sub-millimeter spatial scales. Both Kamitani and Tong (2005) and Swisher et al. (2010) suggest that MVPA is sensitive to the signals arising from variability in the spatial arrangement of cortical columns, which can lead to locally biased signals on the order of millimeters encoded in the neural spatial patterns – this fine-grained information would otherwise be impossible or difficult to detect with conventional methods of analysis that emphasize averaged magnitude of signals over many voxels. Given its sensitivity and decoding power, MVPA has been widely used in recent years to explore high-level cognitive functions and mental states, including visual scene perception (MacEvoy and Epstein, 2011; Epstein and Morgan, 2012), biological motion processing (Thompson and Baccus, 2012; Connolly et al., 2012), imagery and memory (Polyn et al., 2005; Stokes et al., 2009; Rissman et al., 2010), phonological representations (Formisano et al., 2008; Staeren et al., 2009), speech perception (Abrams et al., 2011), and decision-making (Hampton and O’Doherty, 2007; Haynes et al., 2007).

The key component of MVPA involves discriminating patterns of brain activity across different conditions, which is typically done in one of the two ways. The first one is based on classifiers, where a classification algorithm is trained on one subset of data and then the classification performance is tested on a separate, independent subset of data (Mur et al., 2009;
Pereira et al., 2009). The purpose of training is to establish a decision boundary that separates neural responses to two different conditions based on pattern information (e.g., a decision line that separates two clusters of sampled data points in a 2-D space representing two voxels). The most frequently used classifiers are linear classifiers which are straightforward to use and easy to interpret, such as support vector machine (SVM) (Cox and Savoy, 2003; Kamitani and Tong, 2005; Kamitani and Tong, 2006; Eger et al., 2009; Ester et al., 2009; Harrison and Tong, 2009; Parkes et al, 2009; Serences et al., 2009), and Fisher’s linear discriminant analysis (LDA) (Haynes and Rees, 2005a, b; Serences et al., 2009; Shmuel et al., 2010); but non-linear classifiers have also been used (e.g., Hanson and Schmidt, 2011). In the testing phase, the classifier that was determined in the training phase is applied to an independent dataset to assess whether this classifier is still able to reliably separate the two conditions; if so, the multi-voxel patterns evoked by the two different conditions are said to be discriminable, indicating that the underlying brain region of interest may play a role in the cognitive function associated with the conditions.

The second way involves contrasting within- and between-condition neural patterns, which leads to measures of condition-relevant neural pattern similarity. The simplest way to achieve this is to correlate two neural patterns, which yields a scale- and mean-level invariant measure (Haxby et al., 2001; O’Toole et al., 2005; Downing et al., 2007; Williams et al., 2007; Kay et al., 2008; Peelen et al., 2009; Stokes et al., 2009). A more complex way, however, is to consider the level of activation of each voxel as one dimension in a multi-dimensional vector (representing multiple voxels), and compute distance metrics of the two vectors associated with two different conditions based on the Euclidean distance (which is sensitive to differences in both pattern and mean level of neural activity; Mur et al., 2009), or the Mahalanobis distance (which is a scale-invariant measure; see Haynes and Rees, 2005a, Kriegeskorte et al., 2006; Kriegeskorte et al., 2007).
Patterns of brain activity are identified as being similar to, or different from each other based on the similarity measures, by contrasting the within-condition similarity measures against the between-condition similarity measures (e.g., as in the case of Haxby et al., 2001). If the within-condition similarity is greater, meaning that a specific condition can consistently and reliably evoke a specific neural pattern, then some feature represented in that condition might be encoded by the brain region of interest (e.g., feature of objects in Haxby et al., 2001).

The choice of classifier-based or similarity-based pattern analysis is not clear-cut and there are a few factors that should be taken into account. The first one is the number of different conditions that need to be compared simultaneously. Most classifiers are designed for pairwise comparisons and therefore not suitable (or optimal) for coping with more than two conditions at a time. Similarity measures, on the other hand, have the advantage of being possibly high-dimensional (in a square matrix, where each entry is a correlation between two conditions) and handling any number of conditions simultaneously: 2 conditions (Stokes et al., 2009), 3 conditions (Williams et al., 2007), 4 conditions (Downing et al., 2007), 7 conditions (Haxby et al., 2001), 42 conditions (Kriegeskorte et al., 2008), and 120 conditions (Kay et al., 2008).

The second important consideration is performance. The performance of classifiers can be task-dependent and sensitive to many factors (e.g., what measure is taken from each voxel: t-value or beta-value?). Generally speaking, similarity-based measures perform as well as classifier-based methods (Cox and Savoy, 2003; Mitchell et al., 2004; Ku et al., 2008; Misaki et al., 2010). However, due to the fact that classifiers such as SVM or LDA are not scale/mean-level invariant, and therefore take differences in mean activation into account, they may yield slightly better performance in some cases.
Finally, classifier-based methods require separate training and testing datasets to avoid circularity in analysis; but preserving datasets for training naturally reduces the statistical power of the remaining dataset for testing. In addition, there is always a decision to make in terms of the actual algorithms to be used for classification; sometimes an algorithm that appears to be the best for one problem may not necessarily be the best for another problem. In contrast, the beauty of similarity-based measures is computational simplicity and conceptual clarity, as correlation is easy to implement and straightforward to understand.

Therefore, after taking all the pros and cons of different MVPA methods into account, I extended a similarity-based MVPA framework (Linke et al., 2011), and applied this novel analytic approach to the same dataset presented in this chapter (see Chapter 3).
Chapter 3

3.1 – Introduction

The articulatory movements comprising speech must be produced very quickly and must be precise in their execution and timing. Auditory feedback is essential for such accurate speech production (e.g., Guenther et al., 2006; Hickok et al., 2011; Houde and Nagarajan, 2011), implying a complex control system. Here we demonstrate that the complex system controlling speech is supported by a complex neural network that is involved in linguistic, motoric and sensory processing. With the aid of novel real-time acoustic analyses and representational similarity analyses of fMRI signals, our data show functionally differentiated networks underlying a single aspect of articulation – auditory feedback control.

Theoretical accounts of speech monitoring posit multiple functional components required for detection of errors in speech planning (Levelt, 1983; Levelt et al., 1999; Postma, 2000). However, neuroimaging studies generally indicate either single brain regions sensitive to speech production errors, or small, discrete networks (McGuire et al., 1996; Paus et al., 1996; Hashimoto and Sakai, 2003; Fu et al., 2006; Christoffels et al., 2007; Tourville et al., 2008; Zheng et al., 2010; Christoffels et al., 2011). The discrepancy between the complexity of theoretical accounts and the neuroimaging data may be attributable to the univariate analyses that are typically conducted (but see Tourville et al, 2008 for an ROI-based analysis): these analyses are not well suited to the characterization of distributed brain networks.

Here, we use pattern-information analysis of fMRI data (Haxby et al., 2001; Haynes and Rees, 2006; Norman et al., 2006; Kriegeskorte et al., 2008; Kriegeskorte, 2011) to explore auditory

---

2 This Chapter is currently in revision for publication in Journal of Neuroscience (see Zheng et al., in revision).
feedback control networks. Within a general linear model (GLM), we examine multi-voxel neural patterns during speech production and listening with different types of auditory input, and probe for commonalities/differences in neural response profiles across different conditions. The GLM framework provides a conceptually and computationally straightforward way to test hypotheses based on brain pattern similarity, rendering our method a simple and flexible variant within the pattern-information analysis family.

We searched the whole brain (Searchlight method, Kriegeskorte et al., 2006) for patterns of neural activity that are consistent when processing load is specifically placed on putative systems subserving articulatory control. We use a real-time speech-tracking system to deliver normal feedback and two different types of distorted auditory feedback (formant-shifted speech (Houde and Jordan, 1998) and signal-correlated noise (Schroeder, 1968)) in response to spoken words. These same auditory stimuli are also presented in passive-listening conditions. We look for brain regions exhibiting a distinctive neural response profile in which the two acoustically different, distorted feedback conditions elicit similar patterns of activity, but in which the patterns elicited by either of these conditions and normal feedback are dissimilar. Any such profile, present during speech production but not passive listening, can be assumed to reflect processes engaged by auditory feedback that does not match the intended motor gesture, regardless of the nature of the mismatch. Based on the neural pattern signatures, we characterize three functional networks that appear sensitive to distinct aspects of the auditory feedback control process.

3.2 – Methods and Materials

Participants

Written informed consent was obtained from twenty participants (mean age 21 years, range 19-27 years; 8 females). All were right handed, without any reported history of neurological or hearing
disorder, and spoke English as their first language. Each participant received $15 to compensate them for their time. Procedures were cleared by the Queen’s Health Sciences Research Ethics Board.

**Experimental Design and Stimuli**

Standard behavioral paradigms to investigate auditory feedback control, from our lab and others (e.g., Burnett et al., 1998; Donath et al., 2002; Houde and Jordan, 1998; Purcell and Munhall, 2006a, 2006b; Villacorta et al., 2007; Munhall et al., 2009; MacDonald et al., 2010, 2011), typically involve the introduction of a feedback perturbation that changes the acoustics that participants hear, and then maintain these conditions until the talker’s vocalizations have measurably altered. This design is inappropriate for functional MRI since imposing a speech perturbation consistently over many trials (i.e., a low-frequency effect) confounds this manipulation with the slow signal fluctuations (noise) characteristic of fMRI. Furthermore, we are not specifically interested in the process of behavioral compensation (e.g., see Tourville et al. (2008) for evidence of within-utterance formant compensation using long duration trials) but rather are focusing on the neural correlates of trial-specific error response.

We employed a 2x3 within-subject factorial design with 6 experimental conditions, plus a low-level silence/rest control condition. Two tasks (speech production and passive listening) were crossed with three types of auditory signal (formant-frequency-shifted speech, signal-correlated noise, and normal speech), which served as auditory stimuli during listening conditions and as temporally gated auditory feedback during production conditions. To avoid confounding the frequency of our manipulation with the low-frequency noise characteristic of fMRI, we adopted a paradigm in which condition types changed from trial to trial. Although we cannot measure behavioral compensation with this design, single-cell recording studies in marmosets (e.g.,
Eliades and Wang, 2008) and electrophysiological studies in humans (Houde et al., 2002; Heinks-Maldonado et al., 2005; Behroozmand and Larson, 2011; Behroozmand et al., 2012) show that neural responses to auditory feedback perturbations occur rapidly (with a latency of approximately 50 msec), and this would be reflected in the BOLD signal.

For the production with normal speech condition (production-clear), participants vocalized “had” and heard unaltered auditory feedback. For the production with formant-shifted speech condition (production-shift), they vocalized “head” but heard processed feedback such that the first (F1) and second (F2) formants of the vowel were shifted by -200 Hz and +250 Hz respectively (i.e., ‘head’ shifted towards ‘hid’). The direction and magnitude of the shifts were chosen based on empirical studies on formant perturbations in our lab (MacDonald et al., 2010, 2011). For the production with signal-correlated noise condition (production-noise), participants produced either “had” or “head”, and heard masking noise temporally gated with the onset and offset of their vocalizations. Having participants produce ‘had’ half the time and ‘head’ half the time in this condition ensured that any difference in response patterns between distorted- and normal-feedback production conditions could not simply be due to a difference in what the participant was asked to say. For yoked listening conditions (listen-clear, listen-shift and listen-noise, respectively), participants heard auditory feedback that they had previously heard during matched production conditions.

In the three production conditions, participants spoke into an optical microphone (Phone-Or, MA) and their utterances were digitized at 10 kHz with 16-bit precision using a National Instruments PXI-6052E data acquisition board (National Instruments, TX). Real-time analysis was carried out using a National Instruments PXI-8176 embedded controller with which processed acoustic signals were converted back to analog by the data acquisition board and
played over high-fidelity magnet-compatible headphones (NordicNeuroLab, Bergen, Norway) in real time. The processing delays were too short (iteration delay less than 10 msec) to be noticeable to the participants. The processed signals were also recorded and stored on an IBM ThinkPad X32 laptop (IBM, USA) to be used as yoked stimuli during listening conditions.

For the production-shift condition, vowel formants were estimated with an iterative Burg algorithm (Orfanidis, 1988) and formant shifting was implemented using an infinite impulse response (IIR) filter in real-time (see Purcell and Munhall, 2006a, 2006b). For the production-noise condition, masking noise was generated by applying the amplitude envelope of the utterance to uniform Gaussian white noise. This ensures that the noise level was utterance-specific and exactly intense enough at every moment to mask the energy of the spoken words. The resulting ‘signal-correlated noise’ (Schroeder, 1968) has the same spectral profile and amplitude envelope as the original speech signal, but is completely unintelligible.

Image Acquisition and Experimental Procedures

Functional magnetic resonance imaging data were collected on a 3-T Siemens Trio MRI system, using a rapid sparse-imaging paradigm (Orfanidou et al., 2006). This paradigm allowed us to present our auditory stimuli and record responses during a 1600-msec silent period between successive 1600-msec acquisitions in order to minimize acoustic interference (GE-EPI sequence; TR 3200 msec; TA 1600 msec; 211 x 211 matrix; in plane resolution 3.3 x 3.3 mm$^2$; 26 transverse slices with interleaved acquisition). A high-resolution T1-weighted magnetization-prepared rapid gradient echo (MPRAGE) structural scan was also acquired in each participant (TR 1760 msec, TE 2.6 msec, voxel resolution 1.0 x 1.0 x 1.0 mm$^3$, flip angle 90°).

Trials were presented in the silent period between successive scans. Participants were instructed to pay attention to a rectangular, grey prompt in the middle of a computer screen that appeared
100 msec prior to the offset of each scan, which signaled the onset of the next 1600-msec silent period/trial (see Figure 2.1). Depending upon the condition type, the prompt either contained a word, or was blank. Participants were asked to produce the word when there was one, and remain silent when there was not one presented. Hence, the words “had” and “head” were shown in the production-clear and production-shift conditions respectively. The words “had” or “head”, presented with equal probability, were shown in the production-noise condition. The prompt was blank during listen-speech, listen-shift, listen-noise, and rest conditions.

Participants were scanned in three functional runs, each lasting 9 min and comprising 24 blocks of 7 trials. Within each block, conditions were presented in pseudorandom order, such that transitional probabilities were approximately equal and all 7 conditions were presented exactly once (conditions could repeat at the transition from one block to the next). The stimuli for listening conditions were taken from the recordings of the matched production trials in the preceding block, except for the first block, where the listen stimuli were taken from the production trials in the same block, i.e., production trials always preceded listening trials in this first block.

Each participant practiced three blocks (21 trials) of the experiment in the scanner before scanning commenced. During scanning, their behavior on each trial was monitored in real-time by the experimenter (ZZZ) for possible errors. Participant’s vocal production and auditory feedback signals were segregated into two different channels and therefore monitored and recorded simultaneously.

**Behavioral Data Analysis**

Recordings from vocal production and auditory feedback on each trial were reviewed to ensure that incorrect trials were identified and properly accounted for. Trials were considered incorrect if
1) participants made production errors (i.e., failed to produce, produced wrong words, or vocalized during listening/rest trials); or 2) auditory feedback was not triggered due to the very low magnitude of the vocal production (no-trigger trials). In fact, no participant made any errors of the former type, production errors (see Results). No-trigger trials were excluded from further analysis, as were runs in which more than 25% of trials within the same condition (i.e., 6 trials) were no-trigger trials. When, across all three runs, more than 25% of trials within the same condition (i.e., 18 trials or more) were no-trigger trials, data from that participant were eliminated. This ensures that we had a sufficient number of samples from each condition for each participant for the pattern analysis.

Multi-voxel Pattern Analysis

Functional images were analyzed using SPM5 (www.fil.ion.ucl.ac.uk/spm/) and a custom-made, modular toolbox (MVPaa) implemented in an automatic analysis (aa) pipeline system (http://cambridgeneuroimaging.com/aawiki). To preserve spatial pattern information in the native subject space, data were only realigned to the first true functional scan of each run (after two dummy scans were discarded) without further preprocessing. Realigned data were then entered into a general linear model (GLM) using an event-related analysis procedure (Josephs and Henson, 1999). For the GLM, we modeled three successive trials of the same condition as one regressor in each run. This amounts to temporal smoothing of the data, leading to a better signal-to-noise ratio and increased sensitivity of parameter estimates for the time series (Meyer et al., 2011). The modeling resulted in eight regressors (i.e., 24 trials per condition/3 successive trials per regressor) associated with each experimental condition in each run. In addition, incorrect trials, if present, were modeled as a covariate of no-interest for each run to reduce the error variance. We also included six motion parameters from the realignment stage as regressors to
ensure that variability due to head motion was accounted for. This model was convolved with the hemodynamic response function and then fitted to the MR time series in each voxel, resulting in parameter estimates (β) indexing the magnitude of response to the experimental conditions.

A searchlight analysis (Kriegeskorte et al., 2006), constrained to grey-matter voxels, was performed on the β images. The grey matter mask applied to the original whole brain β images was segmented from each participant’s native-space structural image and then coregistered with the individual’s EPI space. We extracted, for each condition-specific regressor, the multi-voxel pattern of voxelwise β values within each spherical searchlight of 4 mm radius (for all searchlights containing at least 30 grey matter voxels). Therefore, each condition was associated with 24 multi-voxel patterns (8 regressors per run x 3 runs) in each sphere. The regressor-specific patterns were then compared with each other using Spearman correlation (Haxby et al., 2001; Kriegeskorte et al., 2008). By taking the mean across all comparisons/correlations, the resulting data were condensed into a 6 x 6 (three feedback types for both production and listening) correlation matrix representing the similarity measures within and between each pair of experimental conditions.

A second, repeated-measures GLM was then applied to the similarity measures derived from within- and between-condition correlations. Here, the predictors were multivariate contrasts that contained hypothesis-driven predictions regarding the relative magnitude of pattern correlations within and between conditions. The resulting images of parameter estimates (βs), each corresponding to one of the multivariate contrasts performed, were spatially transformed into MNI space (Mazziotta et al., 1995) using a nonlinear stereotaxic normalization procedure (Friston et al., 1995) and smoothed with an 8 mm FWHM Gaussian kernel, in order to compensate for anatomical variability across participants. These images were then entered into random-effects
analyses (one-sample t-tests) comparing the mean parameter-estimate difference over subjects to zero. Clusters that survived the statistical threshold of $p < .05$ corrected for multiple comparisons over the whole brain using Gaussian random field theory, i.e., familywise error correction (FWE) at the cluster level (Worsley, 1996), were deemed significant.

3.3 – Results

Behavioral Data

Both real-time monitoring during scanning and inspection of recorded behavioral data indicated that participants followed the instructions in all trials. In production trials, they did not produce the incorrect word, or fail to produce any word, and in listening trials they did not produce speech. There were, however, some production trials in which the vocalization failed to trigger the delivery of auditory feedback, i.e., no-trigger trials. Based on the pre-defined elimination criteria, we identified four participants with elevated rates of no-trigger trials (i.e., more than 25% of trials in the same production condition across three runs), and these participants were excluded from further analysis. For the remaining 16 participants, no significant difference in the distribution of no-trigger trials was observed either across gender, $\chi^2 = 1.182$, $p = .277$, or runs, $\chi^2 = 5.788$, $p = .055$, or production conditions, $\chi^2 = 4.692$, $p = .096$. These no-trigger trials were subsequently modeled as a covariate of no interest in our imaging analysis.

Multi-voxel Pattern Analysis

Our MVPA analysis was based on an assessment of the model fit between predicted and measured multi-voxel pattern similarities within and between conditions. The resulting brain maps reflect the localization of regions in which multi-voxel patterns were consistent with the expected pattern similarity for a given effect of interest, or contrast. Such individually generated maps were used to create a whole-brain ‘similarity map’ at the group level using a random-effects
Anatomical structures were identified based on the LPBA40 probabilistic brain atlas (Shattuck et al., 2008).

**Highlighting speech-sensitive regions**

We reason that brain regions in which multi-voxel patterns are consistently similar for speech stimuli are involved in processing spectrotemporal features characteristic of speech. Such regions will be expected to reveal multi-voxel patterns that are 1) highly correlated between listening to clear speech (*listen-clear*) and listening to shifted speech (*listen-shifted*) conditions; and 2) weakly correlated between listening to clear speech (*listen-clear*) and listening to signal-correlated noise (*listen-noise*), AND between listening to shifted speech (*listen-shifted*) and listening to noise (*listen-noise*). These between-condition similarity predictions ensure that the multi-voxel patterns contain information that is generalizable across different stimulus types that share a crucial ‘feature’ (e.g., the presence of speech information), not just within a stimulus type. The predicted similarity structure among conditions is schematically presented in Figure 3.1a. We call this a “similarity structure contrast matrix”.
Figure 3.1. Similarity structure contrast matrices for the tested contrasts in our study are shown. 
a) For speech-sensitive regions, we predicted high similarity in multi-voxel patterns between the
two speech stimuli AND low similarity between speech stimuli and noise during passive
listening. b) For the regions sensitive to distorted auditory feedback, we predicted an interaction
pattern between Feedback Types and Production/Listening. In other words, a high similarity
between the two types of distorted feedback, AND low similarity between either type of distorted
feedback and normal feedback, during production but not during listening. This was tested by
conducting a paired t-test on the individually generated similarity maps from the minuend and
subtrahend of the formulation shown here.

We observed clusters localized to bilateral superior temporal gyrus (STG), extending anteriorly
and ventrally into the superior temporal sulcus (STS) and middle temporal gyrus (MTG) (see
Table 3.1 and Figure 3.2). The observation of strong bilateral STG/MTG clusters in the
neighborhood of putative speech-sensitive areas is consistent with a large body of literature on the
perception of speech using a variety of stimuli/paradigms (see Table 1 from Zheng et al., 2010 for
review). Our MVPA analysis yielded results that are highly compatible with the literature on this benchmark contrast compared to conventional univariate voxelwise analysis, attesting to the validity of our analytic framework.

Table 3.1. Speech-sensitive areas revealed by analysis of multi-voxel patterns of brain activity.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>t Value</th>
<th>Cluster Size (# of Voxels)</th>
<th>Anatomical Areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>-62</td>
<td>-32</td>
<td>-8</td>
<td>12.75</td>
<td>9605</td>
<td>Left STG/MTG</td>
</tr>
<tr>
<td>70</td>
<td>-22</td>
<td>-10</td>
<td>11.04</td>
<td>8431</td>
<td>Right STG/MTG</td>
</tr>
</tbody>
</table>

Figure 3.2. Speech-sensitive areas in which multi-voxel patterns are more similar between listening to clear speech and listening to shifted speech, than between listening to either speech stimulus and listening to noise. Results are shown at p < .001, FWE corrected for multiple comparisons at the cluster level.
Highlighting networks that are differentially sensitive to auditory feedback error during vocalization, compared to during passive listening

The brain regions involved in auditory feedback error processing during vocalization should be sensitive to the discrepancy between articulation and its auditory concomitant. This requires that multi-voxel patterns of activity in such regions be similar whether they are evoked during production with shifted speech feedback (production-shifted) or during production with masking noise feedback (production-noise). Furthermore, pattern similarity in error-sensitive areas should be greater between these two distorted feedback conditions than between either of these conditions and production with clear speech feedback (production-clear). This similarity structure contrast matrix is depicted on the left side (minuend) of the symbolic formulation in Figure 3.1b. Our MVPA analysis based on this production-related contrast matrix revealed significant clusters in the right posterior STG/inferior angular gyrus at the temporoparietal junction and in the right supplementary motor area (SMA). Additionally, there were marginally significant clusters located in the right cerebellum (p = .053) and in the left cerebellum (p = .067). The peaks of these cerebellum clusters survived a statistical threshold p < .05 at the voxel level, FDR corrected for multiple comparisons across the whole brain (Genovese et al., 2002), see Table 3.2 and Figure 3.3.
Table 3.2. Areas in which multi-voxel patterns of activity were more similar for production-shifted and production-noise, than for production-shifted and production-clear, or for production-noise and production-clear.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>t Value</th>
<th>Cluster Size (# of Voxels)</th>
<th>Anatomical Areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>-34</td>
<td>-44</td>
<td>-24</td>
<td>5.86</td>
<td>902</td>
<td>Left Cerebellum</td>
</tr>
<tr>
<td>42</td>
<td>-54</td>
<td>34</td>
<td>4.99</td>
<td>1118</td>
<td>Right AG</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>46</td>
<td>4.73</td>
<td>3180</td>
<td>Right SMA</td>
</tr>
<tr>
<td>30</td>
<td>-58</td>
<td>-30</td>
<td>4.59</td>
<td>1017</td>
<td>Right Cerebellum</td>
</tr>
</tbody>
</table>

Figure 3.3. Error-coding areas in which multi-voxel patterns are more similar between the two types of distorted feedback than between either distorted feedback and normal feedback, uniquely during production but not during passive listening, are shown.
To explore whether the greater similarity between shifted speech and masking noise conditions was specific to production, we conducted the same MVPA analysis on the matched listening conditions, as depicted in the listening-related contrast matrix on the right side (subtrahend) of the formulation in Figure 2b. However, no brain regions exhibited similarity patterns consistent with the prediction that listening to shifted speech \((\text{listen-shifted})\) and listening to noise \((\text{listen-noise})\) were significantly more similar to each other than they were to listening to clear speech \((\text{listen-clear})\).

The apparent difference between the similarity patterns elicited during production and those during listening (i.e., the Feedback Type by Production/Listening interaction) was formally tested by conducting a paired t-test on the individually generated similarity maps (see the comparison of production- and listening-related contrast matrices in Figure 3.1b). The analysis revealed a distributed network of brain areas that partially overlapped with the network sensitive to distorted feedback during production. A significant interaction was observed in right inferior angular gyrus, right SMA, and bilateral cerebellum, as well as in a number of additional areas including bilateral STG/MTG, extending from the most posterior regions in the STG to the anterior part of the MTG, and bilateral IFG extending dorsally towards the precentral gyrus (see Figure 3.4).
Figure 3.4. Brain areas in which multi-voxel patterns exhibited a Feedback Type by Production/Listening interaction are shown. This interaction was inclusively masked by production simple effect and by inverse listening simple effect in order to demonstrate their relative contributions: predominantly driven by production simple effect (cyan), predominantly driven by inverse listening simple effect (magenta), and a combination of the two (blue).

We refer to the individual similarity structure contrast matrices depicted in Figure 3.1b as production and listening “simple effects”, by analogy with conventional ANOVA. We also refer to the inverse of the listening similarity structure contrast matrix as the “inverse listening simple
effect”. The inverse listening simple effect highlights regions in which multi-voxel patterns of activity elicited by listen-shifted or listen-noise are more similar to the pattern elicited by listen-clear than they are to each other (i.e., in Figure 3.1b, the labels in the similarity structure contrast matrix on the right hand side would flip such that ‘high’ becomes ‘low’ and vice versa). Significant pattern response in the regions identified by the interaction may, therefore, reflect the production simple effect, the inverse listening simple effect, or a combination of the two.

To determine the relative contributions of the production and inverse listening simple effects to the observed interaction, we show the interaction inclusively masked by both effects in Figure 3.4. The interaction masked by the production simple effect yielded regions including right angular gyrus, right SMA, and bilateral cerebellum, as shown in cyan in Figure 3.4. The interaction masked by the inverse listening simple effect yielded regions including bilateral STG/MTG and a small portion of the left precentral gyrus, as shown in magenta in Figure 3.4. A breakdown of the inverse listening simple effect, comparing the similarity of listen-clear to either listen-shifted or listen-noise separately, indicates that similarity between listen-clear and listen-shifted was driving the patterns observed in bilateral STG/MTG and left precentral gyrus (in magenta in Figure 3.4). This is not surprising, since both of these conditions are speech-like, unlike the auditory stimulus in the listen-noise condition, and the observed brain regions are known to be speech sensitive (e.g., Binder et al., 2000; Wilson et al., 2004; Obleser et al., 2007).

The bilateral IFG regions that were observed in the interaction did not seem to arise either because of the production simple effect or because of the inverse listening simple effect alone, but by a combination of the two.

*Assessing pattern similarity among conditions reveals functionally differentiated networks*
We assessed functional specificity of the brain networks observed in the Feedback Type by Production/Listening interaction (as shown in magenta, cyan, and blue color in Figure 3.4), by exploring whether these networks generated differentiable response profiles across conditions. This was done by creating a mask for the volume of significant voxels within each network, and then extracting the mean correlation coefficients for the six between-condition correlation pairs (i.e., three pairs - clear/shift, shift/noise, and clear/noise - for both production and listening). A repeated-measures ANOVA with networks (3 levels) and condition pairs (6 levels) as within-subject factors indicated that there was a significant interaction between networks and condition pairs, F(6, 10) = 9.67, p = .006 (see Figure 3.5). Three separate ANOVAs, each testing one pair of networks, confirmed that the patterns of correlation across the 6 condition pairs differed significantly among networks: (network by condition pair interaction: cyan v.s. magenta: F(5, 11) = 26.34, p < .001; cyan v.s. blue: F(5, 11) = 4.58, p = .017; magenta v.s. blue: F(5, 11) = 18.90, p < .001). Post-hoc comparisons revealed that for the cyan network (i.e., production simple effect), production-shift/noise pattern correlations were significantly stronger than either production-shift/clear correlations, p = .001, or production-clear/noise correlations, p = .003. For the magenta network (i.e., inverse listening simple effect), listen-clear/shift pattern correlations were significantly stronger than either listen-shift/noise correlations, p < .001, or listen-clear/noise correlations, p < .001. For the blue network (i.e., showing neither simple effect clearly but sharing features of both), three out of the four pair-wise comparisons from the two simple effects demonstrated significant or nearly significant differences: production-shift/noise correlations being stronger than production-clear/shift correlations, p = .052, or production-clear/noise correlations, p = .007; listen-clear/shift correlations being stronger than listen-shift/noise correlations, p = .054. These results suggest that the Feedback Type by Production/Listening
interaction arises from three distinct profiles of between-condition pattern similarity, indicating that the brain networks exhibiting these distinct profiles are functionally dissociable.

![Graph showing pattern similarity](image)

Figure 3.5. The magnitudes of the six between-condition pattern correlations are plotted for three functional networks. For each network, three pairs of comparisons are shown for production (shades of blue color on the left) and listening (shades of red color on the right). The vertical bars indicate standard errors and asterisks represent significant difference between the means at $p < .05$.

### 3.4 – Discussion

In the present study, we combined fMRI and an MVPA framework to explore the functional architecture subserving auditory feedback control of speech. Whole-brain analysis of multi-voxel neural pattern similarity revealed three functionally differentiable networks that exhibited different patterns of sensitivity to auditory input across production and listening conditions. The distinct patterns of sensitivity, presumably reflecting the operation of functionally specialized
networks, suggest a distributed neural architecture supporting sensorimotor control of speech production.

One specialized network, including bilateral cerebellum, right angular gyrus, and right supplementary motor area (SMA), yielded patterns of activity that were similar for the two acoustically different, distorted-feedback conditions, but were different for clear speech feedback compared to either distorted feedback condition. Furthermore, this profile was observed uniquely during articulation but not during listening, a pattern that would be consistent with encoding an error signal during talking. That the cerebellum is part of this network is not surprising: other studies have suggested that it integrates sensory and motor inputs to control the precision and timing of movement (Jacobson et al., 2008; Dean et al., 2010). Clinical patients with cerebellar lesions consistently show uncoordinated, disruptive or impaired motion (Diener and Dichgans, 1992; Baier et al., 2009), largely due to poor utilization of sensory-feedback information. The cerebellum maintains an internal model that is capable of adaptively updating motor commands in order to achieve an intended motor output (Wolpert et al., 1998; Blakemore et al., 2001; Ito, 2008). This function is highly relevant for speech motor control (Riecker et al., 2006; Callan et al., 2007; Ackermann, 2008), particularly when, as in our case, the expected and actual sensory consequences of the motor output do not match.

The right angular gyrus has been implicated in monitoring one’s own visuomotor movements, acting as a high-level motor control centre in detecting mismatch between predicted and perceived sensory outcome (Sirigu et al., 2004). Farrer et al. (2008) observed significant neural activation in the right angular gyrus when the correspondence between the intended and perceived sensory consequences of a motor movement was manipulated by way of a sensory
feedback delay. The magnitude of brain activation in this area also correlated with the degree of the discrepancy (Farrer et al., 2003), indicating a neural representation of sensory feedback error.

The final region observed to be part of this network was the supplementary motor area (SMA) bordering on dorsal anterior cingulate cortex (ACC). Given the relatively rostral location of the SMA peak in the current study, it primarily encompasses the pre-SMA subdivision (Picard and Strick, 1996; Tanji, 1996). The pre-SMA interconnects with prefrontal cortex (Bates and Goldman-Rakic, 1993; Luppino et al., 1993), and is known to be involved in motor preparation and planning (Hikosaka et al., 1996), general sensorimotor integration (Kurata et al., 2000), and updating of motor commands and plans (Matsuzaka and Tanji, 1996; Shima et al., 1996). The dorsal ACC appears to be involved in error detection and conflict monitoring (Garavan et al., 2003) and is critical to the processing of error-related responses (Hester et al., 2004). In sum, all of the areas highlighted in this network are highly compatible with a function of encoding an articulatory error signal and adjustment of subsequent motor commands.

The second specialized network includes bilateral anterolateral STG and left precentral gyrus. In these regions, multi-voxel patterns were more similar between clear speech feedback and formant-shifted speech feedback than between either of these conditions and the noise-feedback condition. Furthermore, this speech-sensitive pattern was present only during listening and not during production. The observed network has consistently been implicated in speech processing (e.g., Binder et al., 2000; Scott et al., 2000; Narain et al., 2003; Wilson et al., 2004; Liebenthal et al., 2005; Obleser et al., 2007), but here we demonstrate that speech-sensitivity in this network is not evident during production. Our results show that these regions are not functionally restricted to speech perception – they appear to change their function during production, such that sensitivity to heard speech is attenuated. This is consistent with previous literature demonstrating
a role of bilateral STG in auditory feedback control of speech (Fu et al., 2006; Christoffels et al., 2007; Tourville et al., 2008; Zheng et al., 2010). The high similarity observed across stimulus conditions during production is also compatible with the postulated role of temporal-lobe regions in neural suppression during self-vocalization (Eliades and Wang, 2003, 2005), since such suppression might attenuate any stimulus-driven differences, resulting in highly similar neural patterns across different stimuli.

A final network, localized to bilateral IFG, shares characteristics with both of the other two networks, and may serve as an ‘intermediate stage’ during auditory feedback control. The multi-voxel patterns in these regions appeared more similar for the two distorted-feedback conditions during production and for the two speech conditions during listening. Previous work implicates IFG regions in visuomotor interactions (Johnson-Frey et al., 2003), motor execution based on observation (Iacoboni et al., 1999), and sensorimotor integration in motor action (Parsons et al., 2005). Here these frontal regions may be involved in using incoming auditory information to modify how articulatory gestures are being programmed.

We have extended previous correlation-based MVPA approaches (Haxby et al., 2001; O’Toole et al., 2005; Downing et al., 2007; Williams et al., 2007; Kay et al., 2008; Kriegeskorte et al., 2008; Peelen et al., 2009; Stokes et al., 2009), by incorporating a distinctive feature of conventional activation-based analysis: hypothesis-driven predictions (Linke et al., 2011). These predictions regarding the relative strength of within- and between-condition correlations are formulated as individual contrast-like matrices and tested against the measured correlations of multi-voxel patterns through GLMs, leading to model-based estimates of within- and between-condition similarity. The use of GLMs circumvents the choice of complex algorithms (e.g., Carlson et al., 2003; Cox and Savoy, 2003; Mitchell et al., 2004; Haynes and Rees, 2005;
Kamitani and Tong, 2005; Kriegeskorte et al., 2006; Pessoa and Padmala, 2006; Serences and Boynton, 2007; Friston et al., 2008) for pattern classification, and renders the evaluation of experimentally relevant hypotheses intuitive. The methodological simplicity and conceptual clarity ensure that our method can be readily adapted to other studies to draw inferences about the functional architecture of brain networks based on multi-voxel pattern information.

It is important to acknowledge that in the context of auditory feedback control, experimental designs that are optimal to observe behavioral effects are not necessarily optimal for fMRI data acquisition. In the current design, the randomization required for optimal imaging eliminates the measurable behavioral compensation that would normally build up over many successive distorted-feedback trials. However, we did observe statistically reliable, condition-dependent brain responses even when the overt behavioral changes were not evident. We argue that such brain responses are interpretable, since one of the strengths of neuroimaging as an experimental technique is that overt behavior is not always necessary (see, for example, work in which neuroimaging is used to probe for awareness in behaviorally unresponsive patients diagnosed in vegetative state; Owen and Coleman, 2008; Coleman et al., 2009; Monti et al., 2010; Cruse et al., 2011).

In our study, we elicited an error response by utilizing two acoustically different error types and using a converging-operations approach to extract the commonality (i.e., error response) that is independent of acoustic features. These two types of error are very different: in one case production is masked and inaudible, and in the other case, it is clearly audible and speech-like, but altered. Any differences we might have observed in brain activity between these two conditions could result either from these being different types of error, or trivially from them being different types of sound. Therefore, we concentrate on what they have in common, not how
they are different. This rather conservative approach reveals multiple, extensive brain networks involved in auditory feedback control of speech.

Computational models of speech motor control, including the DIVA (Guenther et al., 2006; Golfinopoulos et al., 2010) and State Feedback Control (SFC) (Hickok et al., 2011; Houde and Nagarajan, 2011), are embodied within a large network of anatomical regions supporting feedforward and feedback control of vocal motor production. The feedback control systems in both of these models are assumed to be anatomically constrained and functionally unitary (e.g., Tourville et al., 2008; Golfinopoulos et al., 2010). In contrast, we demonstrate here that such control systems cover widely distributed brain areas and include functionally differentiated components. One distributed network is directly involved in capturing a speech error signal resulting from the disparity between what is spoken and what is heard. This network might be particularly important for detection of errors during speech movements. Another network appears to alter its function in the context of listening compared to the context of production, such that sensitivity to distorted feedback as an auditory concomitant of speech is accentuated. A third network captures aspects of the previous two networks and may be involved in programming motor output on the basis of sensory information. These functionally heterogeneous networks operate in concert to support auditory feedback control of speech, reflecting the complexity and intricacy of the neural processes recruited for speech movements.
Chapter 4

4.1 – Introduction

How we distinguish ourselves from others is something we easily take for granted. Theories of self-recognition posit that the cognitive processing of self may involve monitoring, evaluation and integration of information from different sources (e.g., see Northoff & Bermpohl, 2004). Evidence from developmental psychology, neurophysiology, and neuroimaging studies suggests that we may represent both self and others in a shared framework to facilitate social interactions (Meltzoff, 2007 a,b; Keenan, 2001; Keenan et al., 2001; Gallagher & Frith, 2003; Sebanz et al., 2003; see Decety & Sommerville, 2003 for review). Yet the cognitive factors supporting our ability to perceive and recognize self-generated actions is not well understood.

Wegner (Wegner, 2002) proposed that the illusory feeling of an action being generated by oneself arises based on priority (thought precedes action), consistency (thought congruent with action), and exclusivity (no alternative causes). This notion is in accord with behavioral studies showing that when cues from different modalities are manipulated to be coherent and congruent, participants perceive externally generated events as being self-produced (Daprati et al., 1997; Knoblich & Prinz, 2001; Sato and Yasuda, 2005; Sirigu et al., 1999; Van den Bos & Jeannerod, 2002). For example, Van den Bos & Jeannerod (2002) demonstrated that when an experimenter’s hand was the only visual information presented to participants, and when the experimenter performed the same finger movements as the participants, participants were prone to identify the experimenter’s hand as their own hand. Sato and Yasuda (2005) developed an illusion in which participants misattributed a tonal sound evoked by a button press by an experimenter as self-generated, when the button press by the experimenter was synchronized with that of the subject’s. All these examples indicate that the illusory perception of an action as self-generated can be
experimentally elicited, provided that the constraints proposed by Wegner (2002) are properly satisfied.

Here, I designed an experiment to explore whether an alien voice, when presented as the auditory concomitant of a participant’s own speech, would be perceived as the participant’s own voice. Specifically, participants heard someone else’s pre-recorded voice (what we call the “rubber voice” below) through headphones during vocalization. The voice they heard matched their own vocalization in terms of both time of onset and content, i.e., temporally gated, congruent, auditory feedback. I was interested in examining how participants would perceive the rubber-voice feedback both when it was congruent with their own production and when it was not (as measured by subjective questionnaires), as well as how participants would change the acoustics of their vocal production in response to the rubber-voice feedback. According to the criteria of Wegner (2002), I predicted that participants would perceive the rubber voice as being from themselves when the rubber voice feedback was consistent/congruent with what they said, but not when the rubber voice feedback was incongruent (i.e., when the property of consistency is violated). In addition, I predicted that participants might change their vocal production when they heard the rubber-voice feedback, based on the literature showing altered vocalization following perturbations of auditory feedback (e.g., Burnett et al., 1998; Jones and Munhall, 2000; Larson et al., 2000; Sivasankar et al., 2005; Chen et al., 2007; Liu et al., 2007).

4.2 – Methods and Materials

Participants

Participants (pilot participants and experiment participants as detailed below) were right-handed female adults (mean age: 20.7; SD: 2.2) from the Queen’s community, without any history of neurological or hearing impairment. Only female participants were recruited, because the pre-
recorded voice was a female voice. Having sex-matched auditory feedback would eliminate the possibility that participants could judge voice identity simply based on the sex cues (e.g., fundamental frequency). Participants had lived most of or all of their lives in southern Ontario and spoke English as their first language with no discernible accent. Written informed consent was obtained from all participants. Each participant was compensated at the rate of $10.00 per hour (rounded to the next quarter hour) for their time, and debriefed at the end. All procedures were cleared by the Queen’s General Research Ethics Board.

**Materials**

**Selection of word stimuli.**

The purpose of this procedure was to ensure that the list of words selected for stimulus delivery would be minimally variable in production duration and difficulty across the tested participants. This is important because the illusion might be critically dependent upon the timing and content of what is produced (participants’ vocalization) matching with what is heard (auditory feedback of another voice). Should there be apparent inconsistency or mismatch between the two, it would be obvious to the participant that the voice she heard is not her own voice.

The initial list of 100 words was selected from the MRC Psycholinguistic Database (Wilson, 1988) based on several criteria, including number of letters (3-4), number of syllables (1), Kucera-Francis written frequency (100) (Kucera & Francis, 1967), concreteness rating (400), and phonetic transcription (C-V-C). In the end, the words selected for further screening were monosyllabic, easy-to-pronounce, and frequently used words with a consonant-vowel-consonant phonetic structure.

Pilot participants (N = 10) were seated in a soundproof booth with a microphone, a set of headphones and a computer screen. A custom made MATLAB (MathWorks, Natick, MA) script
was used to present words on the computer screen. Each word was presented for 1200 ms, and participants were instructed to read each word into the microphone in a clear and consistent manner, once per item. The acoustic signal was recorded online using Adobe Audition (Adobe Inc. USA). The participants were specifically told not to make any noise (i.e., sniffing, coughing, etc) during the reading that would interfere with the quality of the recording. The mean and standard deviation of the production duration of each word were calculated across pilot individuals, and the 5 least-variable of the set of 25 words were selected to be used in the experiments reported below. These chosen words were **dear, pole, dare, kid, and day**.

**Stimulus voice recording.**

We recruited one female participant to assume the role of the ‘rubber voice’ (a term adapted from the ‘rubber hand’ illusion, see Botvinick and Cohen, 1998). In addition to consenting to the study, we asked this participant to consent to have her voice heard by individuals participating in the behavioral test (next section). She received the same instructions and recorded the stimuli using the same experimental procedure outlined above. To ensure the quality and intelligibility of the auditory stimuli, noise reduction procedures were applied to the recorded speech using Adobe Audition (Adobe Inc. USA). In addition, the silent periods before the onset and after the offset of each trial were trimmed.

**Procedure**

**Induction of “Rubber Voice Illusion”**

Fifty participants were randomly assigned to one of two groups: the *no break* group (N = 25) and the *break* group (N = 25). The two groups only differed in the auditory feedback given on two trials near the end of the procedure. Testing took place in a soundproof booth where a microphone and a set of headphones were connected by a real-time processor on which a speech-
tracking system (LabView, National Instruments, TX) developed in our lab was implemented. This system is capable of delivering auditory stimuli through the headphones without noticeable delay from the onset of speech production (iteration delay less than 10 ms). It can deliver either pre-recorded stimuli, or stimuli relayed directly from the microphone.

The experiment consisted of 82 trials. On each trial, participants were prompted to read a stimulus cue word presented on the screen into the microphone and heard concomitant auditory feedback through the headphones. Low-level white noise was present in the headphones to minimize the contribution of bone-conducted speech during vocalization to the auditory feedback (Barany, 1938). Participants were instructed to keep up with the pacing of the prompts (each prompt remained on screen for 1500 ms) and speak as clearly and consistently as possible. They were also asked to remain still and keep a constant distance from the microphone (which we adjusted at the beginning for each participant) to avoid sound fluctuations. Performance was monitored by the experimenter (ZZZ) in real-time from outside the booth, using another pair of headphones.

The experiment consisted of four blocks (see Figure 4.1). In the first block, both groups produced five repetitions of each of the five key words in random order, and heard their own unaltered voice. This was in order to obtain baseline measures of vocal production from each participant. In the second block, both groups produced the 20 non-key words from the set, followed by the 5 key words in random order (the acoustics of these productions were measured). On each trial, the auditory concomitant of their vocalization was the same word as they had uttered, but produced by the rubber voice. The rubber-voice feedback was temporally gated with the vocalization of the same word produced by the participant. The third block was similar to the second, except that the order in which words were presented was different. The last block,
containing 7 trials, was where the two groups diverged. Participants in the break group were cued to produce, on two consecutive trials, a word from the set of 20 non-key items, and heard temporally gated but incongruent rubber voice feedback. For example, she might be cued to say ‘dot’, but heard the rubber voice producing ‘kill’; then cued to say ‘hot’, and heard the rubber voice producing ‘job’. These two ‘break’ trials were followed by five key words with congruent rubber-voice feedback. For the no break group, the same two words as used in the break group were presented, followed by the presentation of the five key words. However, participants always heard congruent rubber voice feedback, as in blocks 2 and 3.

Figure 4.1. Schematic of the experimental paradigm is shown. B1: Block 1; B2: Block 2; B3: Block 3; B4: Block 4. NF: normal feedback; RVF: rubber-voice feedback; MRVF: mismatched rubber-voice feedback; Q: questionnaire.

At the end of blocks 2, 3, and 4, participants were required to answer four questions adapted from the ‘rubber hand illusion’ questionnaire (Botvinick and Cohen, 1998), each on a seven point Likert scale (Likert, 1932) ranging from “Strongly Disagree” to “Strongly Agree”. “Disagree” and “Agree” are points 2 and 6 on this scale; 3 and 5 are “Inclined to disagree/agree”; whereas 4 is “Neutral”. The four questions are:

“It seemed as if the voice I heard was caused by my production of the word”.
“It felt as if the voice I heard was my voice”.

“It seemed as if the voice I heard was a modified version of my voice”.

“It seemed as if I might have more than one voice”.

Questionnaires were facing down on the table in front of participants so that they wouldn’t see them beforehand. Participants were instructed to turn over and complete the top sheet when they saw a prompt on the computer screen stating “please fill out the questionnaire now”, and to set the questionnaire aside when done. The questionnaires were administered three times, at the end of blocks 2, 3, and 4.

Control test

To ensure that participants were able to distinguish their own voice from the rubber voice so that any perceptual effect can not be attributed to participants’ inability to distinguish the two voices, we administered a 2-interval 2-alternative force choice test to all experimental participants right after the behavioral testing. Recordings of the participant’s voice and the matching rubber voice items, randomly selected from 20 trials of the behavioral testing were presented. On each of the 20 trials, participants were instructed to listen to the two utterances and indicate whether their own voice was in the first or second interval, by pressing one of two keys. Results showed that the tested participants were all able to distinguish their own voices from the rubber voice with 100% accuracy.

Data analysis

The data analysis included the analysis of subjective questionnaire ratings as well as analysis of vocal data.
Analysis of subjective ratings.

I collected ratings on the questionnaires from each participant, and examined how the ratings of the same question changed across blocks within groups. In addition, I was also interested in the interaction between groups and blocks, in terms of whether and how the ratings from the two groups differed before and after the break point (and the homologous point in the no-break group; i.e., before and after block 4).

The Likert scales used for the questionnaires generated ordered categorical data (Jamieson, 2004). The response categories for each question have a rank order, but the interval between two adjacent categories may not be equal (non-interval). It is debatable whether parametric statistics based on the general linear model (GLM) are appropriate for Likert-type items, which are arguably non-interval (Blaikie, 2003; Knapp, 1990; Kuzon et al., 1996). Therefore, I adopted nonparametric statistical methods to analyze the questionnaire data (Pett, 1997; Clegg, 1998; Reinard, 2006). In addition, parametric statistics (repeated-measures ANOVA) were also used to explicitly test for interactions.

We employed Friedman two-way ANOVA to evaluate the within-group changes of ratings on each question across the three blocks. If a significant difference was found, a Wilcoxon Signed Rank test was conducted to identify which pair(s) of blocks was significantly different. Furthermore, we used Mann-Whitney U test to compare the distributions of ratings between the two groups in the different blocks. All statistical analyses were done in SPSS (SPSS Inc. Chicago, IL).

Vocal adaptation data (indicated by change of fundamental frequency F0).

Since the ‘rubber voice’ had a different fundamental frequency to that of all of our participants, I expected a change in F0 over trials. This prediction was based on substantial research that
suggests a vocal pitch shift in response to a shift of pitch in auditory feedback (e.g., Hain et al., 2000; Jones and Munhall, 2000; Natke et al., 2003; Liu and Larson, 2007; Burnett et al, 2008). I analyzed F0 across the 82 trials for each participant, in order to track the acoustics of produced vowels. More specifically, I was interested in examining how vocal production was affected by hearing the rubber-voice (RV) feedback (blocks 2 and 3) compared to hearing normal feedback (block 1, or baseline), as well as whether and how the two groups differed in vocal production before and after the break point (for the break group)/homologous trials (for the no-break group). Segmentation boundaries for the vowel in each trial were first calculated using an automated process that examined the harmonicity of the power spectrum. These boundaries were then reviewed visually and corrected.

4.3 – Results

Error trials. The procedure of the experiment requires that participants follow the trials and produce the prompted words clearly and accurately. Although I carefully selected the word stimuli to avoid potential errors, there were some participants who unexpectedly paused, laughed, coughed, or misread words during the trials, rendering the experiment invalid. The data from these participants (4 from the break group, attrition rate 16%; 8 from no-break group, attrition rate 32%) were therefore removed from further analysis, leaving a total of 38 participants (21 in the break group and 17 in the no-break group).

Subjective ratings of questionnaires. The results from the questionnaires are summarized in Figures 4.2.
Figure 4.2. Subjective ratings are plotted as percentage of response at each of the seven response categories for the four questions across blocks. B: break group; NB: no-break group.

“It seemed as if the voice I heard was caused by my production of the word”:

A mixed ANOVA on the ratings for this item, with Block (Blocks 2, 3, and 4) as a within-subjects factor and Group (break and no-break) as a between-subjects factor indicated that there was a significant Block by Group interaction $F(2, 35) = 8.819$, $p = .001$. In Blocks 2 and 3 with congruent rubber voice feedback, participants generally “Agree” (Median = 6) with the statement “it seemed as if the voice I heard was caused by my production of the word”. However, at the end of Block 4, ratings from the break group, where participants had heard two trials of incongruent rubber-voice feedback, dropped to “inclined to disagree” (Median = 3); whereas ratings from the
no-break group, where participants had heard congruent rubber voice feedback all along, remained high (Median = 6). Subjective ratings for this question were significantly different across the 3 blocks for the break group (Friedman: $\chi^2(2) = 14.63$, $p = .001$), whereas this difference was not present for the no-break group (Friedman: $\chi^2(2) = 1.41$, $p = .495$). A follow-up test revealed that ratings at Blocks 2 and 3 were both significantly higher than that at Block 4 for the break group (Wilcoxon: $p = .013$ and $p = .001$, respectively).

Furthermore, Mann-Whitney U test at each block indicated that only at block 4 was the rating from the no-break group significantly higher than that from the break group (M-W: $p = .001$), but not before. This pattern suggests that the incongruent (content-mismatched) auditory feedback appeared to be the source for the between-group difference in ratings at Block 4, in that participants from the break group were exposed to the apparent inconsistency in content between their production and auditory feedback which led them to believe that the voice they heard could not have been caused by their own production.

There was a significant main effect of block, $F(2, 35) = 4.871$, $p = .014$, such that the ratings from blocks 2 and 3 were not different whereas ratings from block 4 were lower than those from block 3 ($p = .012$). In addition, there was also a main effect of group, $F(1, 36) = 4.688$, $p = .037$, such that overall ratings from the no-break group were higher than those from the break group.

“I felt as if the voice I heard was my voice”:

Most participants from both groups responded negatively to this question. Friedman two-way ANOVA indicated that no significant changes in ratings across blocks were observed for either group.

“It seemed as if the voice I heard was a modified version of my voice”: 
A mixed ANOVA revealed a strong interaction between ratings across the three different blocks and two groups, $F(2, 35) = 5.582, p = .008$. For the break group, subjective ratings dropped from “inclined to agree” (Median = 5) before Block 4 to “inclined to disagree” (Median = 3) after Block 4, whereas for the no-break group, ratings remained positive across blocks (Block 2: Median = 4; Block 3: Median = 6; Block 4: Median = 6). Indeed, ratings from the break group changed significantly across blocks (Friedman: $\chi^2(2) = 10.18, p = .006$). Multiple comparisons revealed that for the break group, ratings at Blocks 2 and 3 were both significantly higher than that at Block 4 (Wilcoxon: $p = .003$ and $p = .049$, respectively). The no-break group, on the other hand, did not show significant changes in ratings across blocks (Friedman: $\chi^2(2) = 3.86, p = .138$).

Between-group comparisons revealed that the break group yielded significantly lower ratings than did the no-break group at Block 4 (M-W: $p = .008$), whereas prior to block 4, ratings were not significantly different between the two groups. The data suggests that both groups developed an illusory perception of the rubber voice being a modified version of their own voice when the rubber voice feedback was congruent with their own production; this perception, however, was disrupted by incongruent rubber voice feedback at Block 4 for the break group, but maintained for the no-break group where congruent rubber voice feedback was presented. There was no main effect of either block or group.

“It seemed as if I might have more than one voice”:

In general, participants from both groups assigned low ratings to this question (Median $\leq 3$). There were no significant changes in ratings across blocks for either the break group (Friedman: $\chi^2(2) = 0.33, p = .846$), or the no-break group (Friedman: $\chi^2(2) = 0.78, p = .677$). Neither was there any significant difference in ratings between groups.

*Vocal motor adaptation (indicated by change of F0)*
To determine whether and how participants changed their production when the rubber voice feedback was present, the fundamental frequency (F0) shift of participants’ vocal production, relative to the rubber voice F0, was examined over time and between groups. The shift of F0 was examined in two different ways – the first one simply measured the magnitude of the shift whereas the second one also took the direction of the shift into account (see below).

**Difference of talker’s baseline F0 and rubber voice F0**

The rubber voice had a higher F0 than most of the participants (see Figure 4.3). The difference between each talker’s F0 and that of the rubber voice varied across individuals. The average of the difference between talker’s F0 in Block 1 (baseline) and that of the rubber voice can be seen in Figure 4.4. An unpaired t-test was conducted to verify that these individual differences were similar for talkers in the break and no break groups (t(36) = -0.881, p = .38).

![Figure 4.3](image-url)

**Figure 4.3.** Participants’ F0 (range and mean values) for each key word averaged over five repetitions in Block 1 (baseline) and rubber voice F0 are shown (for the two groups together). The green bar indicates the rubber voice F0.
Figure 4.4. The average of the difference between talker’s F0 in Block 1 and that of the rubber voice is shown for the two groups. Error bars represent standard deviation.

**Magnitude of F0 shift**

A MANOVA was conducted, with the difference (ΔF0) between each talker’s F0 and rubber-voice F0 as the dependent variable (here ΔF0 = |talker F0 - rubber voice F0|; absolute value used to eliminate directional bias), Word and Block were included as two within-subjects factors, and Group was included as a between-subjects factor. The Word factor had five levels, corresponding to the five key words: dear, pole, dare, kid, and day. The Block factor had four levels, corresponding to the four time points at which ΔF0 was measured: for Block 1, ΔF0 was calculated as an average of the five repetitions for each key word, and for the remaining three blocks, ΔF0 was calculated for each of the five key words at the end of each block.
No three-way interaction was found among Word, Block, and Group, \( p = .299 \). Since there was no interaction with Word, I collapsed across this factor. The dependent variable entered into the new analysis was the \( \Delta F_0 \) averaged over the five key words at each block (i.e., removing the factor of Word). In addition, since I was more interested in how F0 changed in the two groups before and after the break point, I averaged the \( \Delta F_0 \) at Blocks 2 and 3 (as Block 2_3), to obtain an estimate of \( \Delta F_0 \) before the break point. A mixed ANOVA was conducted, with Block (three levels: Block 1, Block 2_3 and Block 4) as a within-subjects factor, and Group as a between-subjects factor.

This analysis yielded a significant interaction between Block and Group, \( F(2, 35) = 3.31, p = .048 \). A follow-up analysis revealed that, for the break group, \( \Delta F_0 \) was significantly smaller at Block 2_3 (with rubber voice feedback) than at Block 1 (with their own voice feedback), \( p = .008 \). \( \Delta F_0 \) at Block 4 did not differ significantly from that at either Block 1 or Block 2_3. For the no-break group, \( \Delta F_0 \) at Block 2_3 was significantly smaller than that at Block 1, \( p = .049 \), just like the break group. However, \( \Delta F_0 \) at Block 4 was significantly smaller than that at Block 2_3 and Block 1; \( p = .027 \) and \( p = .01 \), respectively. Thus it appears that when participants heard congruent rubber voice feedback during vocalization, they shifted F0 of their production towards that of the rubber voice, leading to a smaller \( \Delta F_0 \) at Block 2_3, compared to that at Block 1.

**Magnitude of F0 shift with the relationship between talker’s F0 and rubber voice F0 taken into account**

In calculating a measure of the change in F0 (\( \Delta F_0 \)), I realized that I could have been more careful in taking into account the relationship between each talker’s F0 and rubber voice F0. Equation 1 takes the relationship between each talker’s Baseline F0 (i.e., average F0 from Block 1) and that of the rubber voice into account. Using this equation, a positive \( \Delta F_0 \) is interpreted as a change in
F0 from Baseline towards that of the rubber voice (i.e., the talker is following the rubber voice). Similarly, a negative $\Delta F0$ is interpreted as a change in F0 from Baseline away from that of the rubber voice.

$$\Delta F0 = (\text{Block } F0 - \text{Baseline } F0) \times \text{sign(Rubber Voice } F0 - \text{Baseline } F0) \quad \text{Eq 1}$$

For Blocks 2, 3, and 4, a positive change in $\Delta F0$ was observed for both groups. This was confirmed by a series of one-sample t-tests (See Table 4.1).

Table 4.1. Change in F0 and one-sample t-test differences from zero.

<table>
<thead>
<tr>
<th>Group</th>
<th>Block</th>
<th>Mean $\Delta F0$</th>
<th>Std. Dev</th>
<th>t</th>
<th>Df</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-Break</td>
<td>2</td>
<td>5.6</td>
<td>9.7</td>
<td>2.383</td>
<td>16</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>7.1</td>
<td>9.3</td>
<td>3.124</td>
<td>16</td>
<td>&lt; .01</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>9.2</td>
<td>10.7</td>
<td>3.540</td>
<td>16</td>
<td>&lt; .01</td>
</tr>
<tr>
<td>Break</td>
<td>2</td>
<td>7.4</td>
<td>8.4</td>
<td>4.035</td>
<td>20</td>
<td>.001</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6.8</td>
<td>9.0</td>
<td>3.457</td>
<td>20</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5.1</td>
<td>10.1</td>
<td>2.319</td>
<td>20</td>
<td>.031</td>
</tr>
</tbody>
</table>

Although the average $\Delta F0$ was similar for both groups in Blocks 2 and 3, a difference between groups was observed in Block 4. The similarity for Blocks 2 and 3 was supported by the results of a mixed ANOVA, with Block as a within-subjects variable and Group as a between-subjects variable. No significant main effects were found for either Block, $F(1,36) = 0.18$, $p = .67$, or Group, $F(1,36) = 0.07$, $p = .79$. As well, no significant Group by Block interaction was found, $F(1,36) = 1.02$, $p = .32$. Given this lack of difference, each individual’s Block 2 and Block 3 $\Delta F0$ were averaged (as Block 2_3) and then compared with $\Delta F0$ from Block 4 (See Figure 4.5). A second mixed ANOVA with Block as a within- and Group as a between-subjects factor was
conducted on these data and a significant Block by Group interaction was found, $F(1,36) = 8.89$, $p = .005$, such that in Block 2-3 (in which both groups received rubber voice feedback), $\Delta F0$ did not differ between the groups; however, in Block 4, the magnitude of shift (compared to the baseline) for the no-break group was significantly larger than that for the break group, presumably due to the no-break group maintaining the shift of F0 after the trials on which the break group received mismatched rubber voice feedback but they received congruent feedback.

Figure 4.5. An interaction between Block and Group is shown. Error bars represent standard deviation.

4.4 – Discussion

In the present study, I investigated a perceptual effect related to voice perception. While they spoke single words, participants heard the same words produced in another voice. There were three main observations: 1) The synchrony of onset time and the congruence of contents between the participant’s own utterance and the presented rubber (alien) voice feedback caused participants to misattribute the rubber (alien) voice to themselves; 2) Participants shifted the
fundamental frequency (F0) of their production in response to the rubber voice feedback, such that the produced F0 appeared to be following that of the rubber voice over time; 3) The difference in F0 depended both on Block and Group, such that the group who received congruent rubber voice feedback across blocks maintained the shift of F0 towards that of the rubber voice, whereas the group who received incongruent rubber voice feedback in the last block did not continue to shift F0 afterwards. Taken together, both subjective and objective measures show behavioral evidence for a perceptual illusion, induced by an alien voice temporally and phonetically matched with participants’ own vocalization, suggesting that the perception of the feedback voice identity may depend on the sensory cues arising from the dynamics of articulation.

Our data fit well with the idea of a “forward model” as the central mechanism underlying a possible distinction between self and others (Blakemore et al., 2002; Wolpert and Miall, 1996; Wolpert, 1997; Wolpert and Ghahramani, 2000; Wolpert et al., 1995). According to this forward-model idea, a copy of the motor command (i.e., efference copy) predicts its sensory outcome, which will be compared to the actual sensory consequences when the movement occurs. The prediction is used to anticipate the sensory effects of movement, attenuating the component that is due to self-movement from that due to changes in the outside world and thereby leading to online self-recognition (Blakemore et al., 1998; Blakemore et al., 2002). In the context of our study, temporal and content matching of the ‘rubber’ auditory feedback with participant’s vocal production ensured that participant’s intention to produce a prompted word, participant’s production of that word, and the actual feedback (in terms of content) as a result of producing that word were all consistent. Therefore a plausible interpretation regarding the identity of the feedback voice would be that it was self-produced, even if it was an alien voice.
Our results may also give insights about what is happening for clinical patients with deficits in self-attribution. Previous studies have shown that individuals with schizophrenia who are experiencing the positive symptoms of auditory hallucinations and delusions of control are more likely to misidentify self-generated speech as being from another person, compared to normal controls (Allen et al., 2004; Cahill et al., 1996; Johns et al., 2001; Johns et al., 2006; Johns and McGuire, 1999). It has been speculated that these patients may have a defective forward-modeling mechanism. Although the current studies do not include clinical participants, the results are consistent with the idea that the process of self-attribution is closely related to the mechanism of a forward model, thereby providing support that deficits of self-attribution may be inherently tied to the dysfunction of a forward model mechanism.

This ‘rubber voice’ perceptual effect was accompanied by vocal adaptation. One line of studies using an online pitch-shift paradigm has shown that people compensate for a change of F0 in their auditory feedback during vocal production (Burnett et al., 1998; Chen et al., 2007; Jones & Munhall, 2000; Larson et al., 2000; Liu et al., 2007; Sivasankar et al., 2006). This compensation presumably reflects the ability of our audio-vocal system to correct for possible errors by aligning production with the sensory consequences. The implicit assumption for this compensatory strategy, however, is that the participant must acknowledge that the feedback they are hearing is generated by themselves. One recent study done by our group (unpublished results) revealed that when the magnitude of shift was increased from 50 cents up to 200 cents, participants compensated for this shift. However, when the magnitude of shift was further increased from 200 cents to 500 cents, the compensation halted. One possible interpretation is that beyond a certain shift magnitude, participants fail to recognize the voice as their own voice, and this results in the termination of the compensatory behavior. In the present study, participants followed, instead of
compensated for, the rubber voice in the auditory feedback. Although the reasons for this following pattern are not clear, I have the following speculations: since the rubber voice is qualitatively very different from the participant’s own F0, this big difference may be too big to prompt compensation (as participants somehow ‘know’ that no matter how they alter their vocal tract, they can’t compensate for the rubber voice). Nevertheless, following, like compensation, is a change in vocal production consequent upon feedback. Such changes in vocalization indicate that a) the listener perceives the auditory feedback as the result of their own vocalization; and b) they perceive this feedback as erroneous in some way. Another possibility that is worth noting is phonetic convergence that is commonly observed during conversational interaction (Garrod and Pickering, 2004; Pardo, 2006). The convergence is based on an interactive processing mechanism that leads to the alignment of linguistic representations between the interlocutors, causing a following pattern of acoustics that is similar to what we have observed here. If this is the case, then our results appear to suggest that individuals might show phonetic convergence with a voice that they perceptually categorize as being their own. This is also consistent with the view that the phonetic convergence is rather automatic (Garrod and Pickering, 2004).

Issues in the experimental design and methodology limited the interpretation of the data. First, there were only five trials after the break point for the break group. Auditory feedback perturbation studies typically impose sustained feedback manipulation over an extended period of time (at least 30 trials in a row) and examine the acoustics of vocal response. Thus, a longer post-break period would potentially allow for a better assessment of the differences between the two groups in both subjective perception and vocal response. Second, I realized that the selection of the key words was still not optimal; e.g., it was very difficult to match the onset time of production of the consonant at the end of the key word ‘kid’ for a lot of participants. This
possible mismatch would act effectively as a ‘break’ point (that should not have been) and would add variance to the data (making the no-break and break groups more similar than they should be). Third, this study was limited by having only one rubber voice with a higher F0 than that of the majority of participants. Hence it is not clear whether the observed patterns of subjective ratings and vocal response were specific to the acoustic features of this particular rubber voice. Finally, the questionnaires could have been more automatically administered, without having participants fill out questionnaires on paper. All these issues provided a direction for improvement and motivated me to conduct a refined version of this experiment, in order to more precisely examine subjective perception and vocal motor adaptation with rubber-voice feedback. This refined study is described in the next Chapter.
5.1 – Introduction

The rubber hand illusion describes a phenomenon where temporally coincident visual and somatosensory inputs (i.e., the feeling of someone stroking the fingers of your own hidden hand, and the sight of a prosthetic hand being stroked the same way) combine to create the perception of body ownership, a critical component of self-awareness (Botvinick and Cohen, 1998; Ehrsson et al., 2005). Similar perceptual illusions have also been demonstrated with respect to the face (Tsakiris, 2008), and the whole body (Petkova and Ehrsson, 2008). It has been argued that self-attribution of body parts is mediated by multisensory perceptual correlations (Bahrick and Watson, 1985; Mitchell, 1997; Botvinick and Cohen, 1998). This means that the illusion itself is not modality specific and similar phenomena might be observed across other modalities, as long as multi-modal cues are converging.

Behavioral studies have documented illusory self-attribution during voluntary action when action-related sensory cues were manipulated to be coherent and congruent (Daprati et al., 1997; Sirigu et al., 1999; van den Bos and Jeannerod, 2002). For example, Van den Bos and Jeannerod (2002) demonstrated that when an experimenter’s hand was the only visual information presented to the participants and when that experimenter performed, in synchrony, the same finger movements as the participants did, participants tended to identify the experimenter’s hand as their own hand. This series of studies can be taken as evidence that motor information provided during action can modulate the perception of body ownership when the motor movement and its sensory consequences are consistent (Tsakiris et al., 2007).

3 This Chapter has been published in PLoS ONE (see Zheng et al., 2011)
Although there is a growing literature on the perception of body ownership, empirical evidence regarding the perception of voice ownership as a result of vocal motor output is lacking. Vocal production provides rich sensory feedback signals (i.e., auditory and somatosensory) which, together with representations generated during articulation, can contribute to the recognition of one’s own voice through a self-monitoring system (Frith et al., 2000; Blakemore et al., 2002). Previous studies have shown that psychotic patients with positive symptoms of auditory hallucinations and delusions of control have difficulty identifying self-produced sounds, and this appears to be due to an impaired self-monitoring system (Cahill et al., 1996; Blakemore et al., 2000; Johns et al., 2006). However, it remains unclear how normal individuals, in whom the self-monitoring system is intact, would perceive the identity of an external voice that is heard as concomitant auditory feedback of their own vocalizations.

An important aspect of auditory feedback during vocal production is that it is used for vocal motor control of ongoing speech (e.g., Guenther, 2006). Based on online feedback perturbation paradigms, a number of studies have either provided behavioral (Burnett et al., 1998; Houde and Jordan, 1998; Jones and Munhall, 2000) or neuroimaging (Fu et al., 2006; Christoffels et al., 2007; Tourville et al., 2008; Zheng et al., 2010) evidence for the role of auditory feedback in articulatory control, as part of an error correction mechanism. The question that remains to be addressed here, however, is whether auditory feedback is used the same way for vocal motor control as for the recognition of one’s own voice.

To address these issues, we examined, both subjectively and objectively, how normal participants responded to a feedback voice that was heard as the auditory concomitant of their own vocalizations. Specifically, participants produced one of two target words on each trial, and heard auditory feedback temporally gated with their own utterances using a real-time signal
processing system. We assessed participants’ subjective perception of, and vocal-motor adaptation to, online auditory feedback of a) their own voice, and b) one of two stranger voices, during vocal production. We were interested in exploring how participants would perceive the stranger voice feedback both when it was congruent with their own production and when it was not, as well as how participants’ subjective perception of voice identity was related to the acoustics of their vocal production.

5.2 – Methods and Materials

Participants

Ninety-three right-handed female participants (age range: 18-28 years, mean: 22 years) recruited from the Queen’s community participated in this study. Participants were without any history of neurological or hearing impairment, and spoke English as their first language. Written informed consent was obtained from all human participants. All procedures were cleared by the Queen’s General Research Ethics Board.

Materials

Two female speakers of southern Ontario English were recruited as stimulus voices (V1 and V2). Prior to the experiment, utterances of the target words, ‘day’ and ‘too’, from these two stimulus voices were recorded in a soundproof booth. The two target words were selected because they are one-syllable English words of consonant-vowel (CV) form. A pilot study demonstrated that temporal aspects of production are similar across talkers. Utterances of two additional words, ‘page’ and ‘test’, were also recorded. The individuals recruited as stimulus voices were chosen because the pitch of their voices was either higher (V1) or lower (V2) than that of an average female talker of southern Ontario English. MacDonald et al. (2010) reported that for the vowels /e/ and /u/, the F0 of an average female talker was 204 and 213 Hz, respectively. For the
utterances used in this study, V1 had an F0 of 226 and 241 Hz for /e/ and /u/ respectively; V2 had an F0 of 187 and 200 Hz for /e/ and /u/.

Procedures
Testing took place in a soundproof booth with a microphone (Sennheiser E845S, Sennheiser Electronic, Germany) and a set of headphones (Sennheiser HD265 Linear, Sennheiser Electronic, Germany) connected through a Fireface 400 audio interface (RME, Germany) to a real-time signal processing computer (National Instruments, TX), on which a deterministic signal processing program was implemented (Purcell and Munhall, 2006a,b). This real-time system is capable of delivering auditory stimuli, either pre-recorded or relayed directly from the microphone, through the headphones without noticeable delay (iteration delay less than 10 ms) from the onset of speech production. In the cases when participants’ utterances were shorter than the recorded stimulus voice utterances, our system would match the offset of their vocalizations by truncating the recorded utterances after production ceased.

The experiment consisted of 155 trials. On each trial, participants were prompted to speak either ‘day’ or ‘too’ into the microphone and heard concomitant auditory feedback through the headphones. Low-level white noise was present in the headphones to minimize the bone-conducted speech feedback while they vocalized (Barany, 1938). Performance on each trial was monitored by the experimenter (ZZZ) from outside the booth and also recorded by our system. Over the 155 trials, 80 trials of ‘day’ and 75 trials of ‘too’ were presented in the same pseudorandom order for each participant, such that no more than three consecutive trials of the same word occurred.

a) Experimental paradigm and groups (Early Mismatch and Late Mismatch)
The experiment consisted of 4 stages: Baseline, Stimulus Voice Match, Stimulus Voice Mismatch, and Post Mismatch (see Figure 5.1). During the Baseline stage, participants produced 20 utterances of the target words while receiving their own unaltered feedback. During the three other stages, when participants produced a target word, they heard one of the utterances produced by a stimulus voice (V1 or V2). This feedback matched the produced word in the Match and Post Mismatch stages, but differed in the Mismatch stage where participants produced ‘day’ or ‘too’ but heard the stimulus voice saying ‘page’ or ‘test’.

**Early Mismatch Group**

<table>
<thead>
<tr>
<th>Stage</th>
<th>Baseline</th>
<th>Stimulus Voice Match</th>
<th>Stimulus Voice Mismatch</th>
<th>Post Mismatch</th>
</tr>
</thead>
<tbody>
<tr>
<td>own unaltered feedback</td>
<td>(20 trials)</td>
<td>congruent stimulus voice feedback (45 trials)</td>
<td>incongruent stimulus voice feedback (5 trials)</td>
<td>congruent stimulus voice feedback (85 trials)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 65</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 130</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 135</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Late Mismatch Group**

<table>
<thead>
<tr>
<th>Stage</th>
<th>Baseline</th>
<th>Stimulus Voice Match</th>
<th>Stimulus Voice Mismatch</th>
<th>Post Mismatch</th>
</tr>
</thead>
<tbody>
<tr>
<td>own unaltered feedback</td>
<td>(20 trials)</td>
<td>congruent stimulus voice feedback (110 trials)</td>
<td>incongruent stimulus voice feedback (5 trials)</td>
<td>congruent stimulus voice feedback (20 trials)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 110</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 130</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 135</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.1. Schematic diagram of the four stages of the experiment. See text for details.

Participants were randomly assigned into one of two experimental groups (Early Mismatch and Late Mismatch) that differed in when the Mismatch stage occurred. The Match stage consisted of 45 trials (Early Mismatch) or 110 trials (Late Mismatch). The Mismatch stage occurred at trials 66-70 (Early Mismatch) or trials 131-135 (Late Mismatch). The final stage, Post Mismatch, was similar to the Match stage, but consisted of 20 and 85 trials for the Late- and Early Mismatch groups respectively.

*b) Subjective report*
At five time points over the course of the experiment (see Figure 5.1), participants responded to two questions concerning the perceived identity of the feedback voice using a 7-point Likert scale in the form of a sliding pointer on the computer screen (with 1 indicating “strongly disagree” and 7 indicating “strongly agree”). The two questions were adapted from the rubber voice illusion questionnaire (Botvinick and Cohen, 1998): 1) “It felt as if the voice I heard was my voice”, and 2) “It felt as if the voice I heard was a modified version of my voice”. For the Early Mismatch group, these questions were asked once before and four times after the Mismatch stage, whereas for the Late Mismatch group, they were asked four times before and once after the Mismatch stage. Responses were measured to the nearest 0.1 units.

To test the validity of our questions, five participants had their own voice utterances recorded during the Baseline stage. These own utterances were then used as stimulus voice feedback during the remaining 135 trials of the experiment (i.e., feedback always temporally gated and congruent with vocalizations). As expected, these participants rated both Q1 and Q2 high across the five time points (see Figure 5.2). This suggests that these two question items are not mutually exclusive.
Figure 5.2. The box plots for ratings on the two questions (Q1 and Q2) across five time points are shown for a small control group (N=5).

However, a trivial reason for people reporting that the stimulus voice was like their own voice would be that they could not distinguish between the two. We examined this in a separate group of 10 participants tested in a pilot version of the procedure reported here. After the procedure, their ability to distinguish their own recorded productions of five monosyllabic words (including ‘day’ and ‘too’) from those of the stimulus voice was tested on 20 trials. In each trial, one monosyllabic word was presented twice; once in their voice and once in the stimulus voice. They were asked to report the interval in which their own voice was presented. Participants in this study were all able to distinguish their own voices from the stimulus voice with 100% accuracy. This result indicates that any perceptual effects in our study cannot be due simply to the inability of our participants to distinguish between their own and the stimulus voice.

c) Vocal production data analysis
We extracted the fundamental frequency (F0) across the 155 trials for each participant in order to track the acoustics of produced vowels (see Purcell and Munhall, 2006a,b). The F0 track for every trial from each participant was individually reviewed for discontinuities and/or gaps caused by glottal fry using Praat (Boersma and Weenink, 2005). Trials with glottal fry were excluded from analysis, as were participants who 1) exhibited glottal fry on more than 30% of baseline trials (> 6 trials) or more than 20% of the entire trials (> 30 trials); 2) produced sounds (e.g., coughs) that resulted in unplanned mismatches in feedback; or 3) had an F0 that was higher than the stimulus voice for one target word but lower for the other. As a result of these exclusion criteria, 29 and 33 participants were included in the Early- and Late Mismatch groups respectively.

The F0 data for each participant were then examined to determine whether vocal production shifted up, or down, or not at all during stimulus voice feedback, relative to baseline. For each word, the mean and standard error of the last five trials of the Baseline (excluding the first 5 trials of each word to allow the participant to adapt to the microphone) were used to determine the 95% confidence interval (CI). If the mean F0 of the 10 trials immediately preceding the Mismatch stage fell within the baseline 95% CI, the participant was classified as ‘no shift’; if the F0 was outside the 95% CI, the participant was classified as ‘up’ if it was higher and ‘down’ if it was lower. In addition, for every trial after the Baseline stage, for each participant, a normalized F0 shift (relative to baseline) was calculated by subtracting the mean baseline F0 for each word separately.

5.3 – Results

a) Subjective ratings
In general, ratings were low across all time points for Q1 (i.e., “I felt as if the voice I heard was my own voice”), but high before and low after the Mismatch stage for Q2 (i.e., “I felt as if the voice I heard was a modified version of my own voice”) (see Figure 5.3).
Figure 5.3. The box plots for ratings on the two questions (Q1 and Q2) are shown for the Early Mismatch and Late Mismatch group. a) Question 1: I felt as if the voice I heard was my own voice, and b) Question 2: I felt as if the voice I heard was a modified version of my own voice, were rated on a 7-point Likert scale across five time points. The Mismatch stage occurs after the first time point for the Early Mismatch group and after the fourth time point for the Late Mismatch group, as indicated by a red vertical dashed line.
These observations were confirmed with MANOVAs on the ratings on each of the two questions across the five time points, with stimulus voice (V1 and V2) and group (Early Mismatch and Late Mismatch) as between-subjects factors. For Q1, a pattern of results consistent with the mismatch events having a marked effect on ratings was rather weakly observed. There was a marginally significant interaction between time and group $F(4, 55) = 2.52, p = .052, \eta^2_p = .16$, with ratings being higher in the Early Mismatch group than in the Late Mismatch group at both first and last time points, $p = .001$ and $p = .002$, respectively. However, ratings dropped significantly after the Mismatch only in the Early Mismatch group, $p = .004$.

In addition, we observed a higher overall rating in the Early Mismatch compared to the Late Mismatch group, $F(1, 58) = 8.56, p = .005, \eta^2_p = .13$, and a marginally significant effect of time, $F(4, 55) = 2.47, p = .055, \eta^2_p = .15$. A trend analysis indicated that there was a cubic trend of ratings across the time points, $F(1, 58) = 8.04, p = .006, \eta^2_p = .12$, such that ratings decreased after the first time point, and then started to gradually increase, before decreasing again at the last time point.

Sign tests on Q1 ratings at each of the time points indicated that for the Early Mismatch group, the ratings at the first and last time points were not different from ‘neutral’ (i.e., a rating of ’4’), $p \geq .061$, but below ‘neutral’ at time points 2, 3, and 4, $p \leq .008$. For the Last Mismatch group, the ratings were below ‘neutral’ at all time points, $p \leq .001$.

The three-factor MANOVA on Q2 ratings revealed a strong interaction between time and group $F(4, 55) = 25.26, p < .001, \eta^2_p = .65$. Follow-up pairwise comparisons with Bonferroni correction revealed that, for the Early Mismatch group, Q2 ratings at time point 1 (pre-Mismatch) were significantly higher than those at all the later time points (post-Mismatch), $p < .001$. In this group, Q2 rating at time point 2 was also lower than that at time point 4, $p = .003$. For the Late
Mismatch group, Q2 ratings at the first four time points (pre-Mismatch) were all significantly higher than that at the last time point (post-Mismatch), p < .001. The ratings of the two groups did not differ at either the first or last time point (i.e., before the mismatch event for both groups or after it, p ≥ .100) but the ratings from the Late Mismatch group were significantly higher than those of the Early Mismatch group at time points 2, 3, and 4, p ≤ .001, which are pre-Mismatch stage for the Late Mismatch group but post-Mismatch for the Early group (see Figure 5.3b).

In addition to this expected interaction, participants who heard V1 gave higher ratings that those who heard V2, F(1, 58) = 5.63, p = .021, \( \eta^2_p = .09 \), and participants in the Late Mismatch group gave higher ratings than the Early Mismatch group, F(1, 58) = 9.81, p = .003, \( \eta^2_p = .15 \). Finally, ratings varied across the time points, F(4, 55) = 34.39, p < .001, \( \eta^2_p = .71 \). A trend analysis revealed a combination of linear, F(1, 58) = 33.43, p < .001, \( \eta^2_p = .37 \), and cubic, F(1, 58) = 88.65, p < .001, \( \eta^2_p = .60 \), components for ratings across the time points, such that ratings dropped after the first time point, and then slowly increased from time point 2 to 4, before dropping again at the last time point.

Sign tests on Q2 ratings indicated that, for the Early Mismatch group, Q2 ratings at the first time point (pre-Mismatch) were reliably greater than ‘neutral’, p < .001, but ratings dropped to well below ‘neutral’ at time point 2, p = .001 (post-Mismatch) and then were not different from ‘neutral’ for time points 3, 4, and 5, p ≥ .458. For the Late Mismatch group, Q2 ratings at the four pre-Mismatch time points were all reliably greater than ‘neutral’, p ≤ .001, whereas the rating at the post-Mismatch time point dropped to well below ‘neutral’, p = .014.

The results, particularly from Q2, suggest that the Mismatch stage, characterized by incongruent stimulus voice feedback, appeared to disrupt the illusion of the stimulus voice being attributed to the ‘self’, as evidenced by altered ratings. Higher Q2 ratings at the later time points
for the Early Mismatch group (see Figure 5.3b) may indicate that after many further trials of congruent feedback, the illusory percept appeared to build again. Overall, it seems that the perceptual illusion regarding the perceived identity of the feedback voice is elicited by congruent feedback, matched in timing and content to the participant’s own vocalization.

b) Vocal motor adaptation

To determine whether and how participants altered production in response to the stimulus voice feedback, we examined the F0 of participants’ vocal production over the course of the experiment (see Table 1). Chi-square tests revealed that significantly more participants altered their F0 than did not change when hearing either V1, $\chi^2(1, N = 32) \geq 10.13, p \leq .001$, or V2, $\chi^2(1, N = 30) = 6.53, p = .011$. The direction of change depended on the relation between the participant’s baseline F0 and the stimulus voice F0, such that participants were more likely to shift their F0 towards (i.e., ‘follow’), than away from (i.e., ‘compensate’), that of the stimulus voice for both V1, $\chi^2(1, N = 32) \geq 6.76, p \leq .009$, and V2, $\chi^2(1, N = 30) \geq 4.55, p \leq .033$ (see Table 5.1). Thus participants tended to shift their F0 upward if their F0 was lower than that of the stimulus voice, and downward if their F0 was higher (e.g., see Figure 5.4).
Table 5.1. The number of participants who shifted their F0 up (Up), down (Down), or did not shift their F0 (No-shift) when hearing either V1 or V2 are shown for ‘day’ and ‘too’. Each participant was assessed based on whether the average F0 of 10 trials immediately preceding the Mismatch stage was higher than (Up), lower than (Down), or inside (No-shift) the range defined by the 95% CI for the mean baseline F0 (see Procedure c for a more detailed description). For those participants who significantly shifted their F0, the direction of the shift was also determined as to whether the shift was towards (Follow) or away from (Compensate) the stimulus voice heard.

<table>
<thead>
<tr>
<th></th>
<th>Up</th>
<th>Down</th>
<th>No-shift</th>
<th>Follow</th>
<th>Compensate</th>
<th>No-shift</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>day</td>
<td>V1</td>
<td>22</td>
<td>4</td>
<td>6</td>
<td>22</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>V2</td>
<td>16</td>
<td>6</td>
<td>8</td>
<td>17</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>too</td>
<td>V1</td>
<td>19</td>
<td>6</td>
<td>7</td>
<td>19</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>V2</td>
<td>13</td>
<td>9</td>
<td>8</td>
<td>16</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

Figure 5.4. The F0 (Hz) time course for ‘day’ from one representative participant is shown. This participant was from the Last Mismatch group and assigned V1 as the stimulus voice. The solid purple vertical line at trial 20 indicates the end of the Baseline stage. The two solid red vertical lines indicate the beginning and end of the Stimulus Voice Mismatch stage. The black dashed horizontal line indicates the F0 of the stimulus voice V1.
To understand the effect of the Mismatch stage on the pattern of changes in F0, we compared the magnitude of normalized F0 shifts before and after the Stimulus Voice Mismatch stage for the two groups. Here we used, for each word, the mean magnitudes of F0 shifts during the last 10 trials in the pre-Mismatch and the first 10 trials in the post-Mismatch stages as estimates of before and after F0 shifts respectively. We conducted a MANOVA with stage (before and after) and word (‘day’ and ‘too’) as repeated measures, and stimulus voice (V1 and V2) and group (Early Mismatch and Late Mismatch) as between-subjects factors. We observed a higher magnitude of F0 shifts in the after stage than in the before stage, F(1, 58) = 8.00, p = .006, \eta_p^2 = .12. We also observed a three-way interaction between stage, stimulus voice, and group, F(1, 58) = 6.34, p = .015, \eta_p^2 = .10. Post-hoc comparisons with Bonferroni correction indicated that the magnitude of F0 shifts was the same before and after the Mismatch stage for V1 in both groups, p \geq .288, but greater after than before the Mismatch stage for V2, p \leq .042. The observation that F0 shift was either unaltered or even greater after the Mismatch stage, is in sharp contrast to the pattern of subjective ratings. If production data mirrored the subjective ratings, F0 shifts should have been diminished when the illusion was broken.

In addition, the magnitude of F0 shift was higher in the Late Mismatch group than in the Early Mismatch group, F(1, 58) = 4.43, p = .040, \eta_p^2 = .07. There was also an interaction between group and voice, F(1, 58) = 4.32, p = .042, \eta_p^2 = .07, such that the magnitude was higher in the Late Mismatch group only for V1, p = .004.

5.4 – Discussion

We believe that the emergence of this auditory illusion results from the convergence of sensory cues in the context of voluntary action. Previous studies have shown that motor action significantly contributes to the self-recognition process by structuring the perception of bodily
multisensory signals (van den Bos and Jeannerod, 2002; Tsakiris et al., 2005; Tsakiris et al., 2007). The coherence between motor movement and its sensory consequences plays a critical role in modulating the perceptual experience of both body ownership and movement agency (Daprati et al., 1997; Sato and Yasuda, 2005). In the present study, auditory feedback was temporally and phonetically congruent with motor and somatosensory feedback from the articulators. The alignment between vocal motor movement and the resulting sensory events allowed the participants to perceptually categorize a stranger’s voice as being from themselves, suggesting that the coherent action-related cues are integrated into a unified sense of voice ownership.

It has been suggested that the induction of ownership in the rubber hand illusion depends critically on a top-down process of evaluating the rubber hand against a pre-existing cognitive representation of the body, based on whether the rubber hand is a plausible substitute for the body part (Tsakiris and Haggard, 2005). This explains why using a piece of wood (Tsakiris and Haggard, 2005), a wooden hand (Tsakiris et al., 2010), or even a rubber hand covered with non-natural skin texture (Haans et al., 2008) either reduces or abolishes the illusion, depending on the degree of implausibility. Similarly, the stranger’s voice in our study, although gender-matched to the participant’s own voice, was otherwise very different, and may therefore have been perceived as a somewhat implausible substitute. This may be why participants endorsed Q2 rather than Q1. The large variances in the Q1 ratings may reflect variability in the degree of judged implausibility across individuals.

In addition to the subjective ratings indicating that participants were experiencing illusory ownership of the feedback voice, a trial-by-trial assessment of vocal production revealed that participants shifted their F0 to follow that of the feedback voice. A line of studies using online F0 shift paradigms have shown that, in general, people compensate for a change of F0 in their
auditory feedback during vocal production, i.e., by shifting their production in the direction opposite to the shifted feedback signal (Burnett et al., 1998; Jones and Munhall, 2000; Larson et al., 2000). Although the central mechanism underlying the direction of vocal motor adaptation is not well understood, the shift must reflect the operation of a sensorimotor control process involved in regulation of ongoing speech production. That our participants tended to follow, rather than compensate, might be due to a number of factors. One factor might be related to the large magnitude of effective ‘shift’ between participants’ own voice F0 and the stimulus voice F0, which is within the range over which Burnett et al. (1998) observed the greatest proportion of following responses as a result of F0 perturbation.

The observation that F0 remains shifted despite a change in perception of the feedback voice identity suggests a divergence between conscious perception and sensorimotor control. This is consistent with a two-level model of self-action recognition, which posits that an automatic level of action control and a conscious level involving the perception of action agency can be separated (Jeannerod, 2009). Empirical studies involving online sensory perturbation during motor movements have revealed both mismatch (Fourneret and Jeannerod, 1998) and temporal lags (Castiello et al., 1991) between objective motor responses and subjective awareness of the perturbation. Data supporting such a divergence also comes from clinical studies of patients with visual form agnosia who demonstrate striking precision of hand movements towards a visual target that they fail to perceive (Goodale et al., 1991), and of patients with schizophrenia who are capable of initiating an action but are impaired in attributing the action to its correct source (Mellors, 1970). Our data further add to this literature in demonstrating that the cognitive systems that process auditory feedback for the differentiation between ‘self’ and ‘other’ and for control of ongoing vocal production appear to be at least partially dissociable.
In summary, our study provides a new framework to explore the sense of ownership of voice by examining both perceptual judgment of voice identity, and acoustics of vocal production, in the same context. Our findings shed new light on how identity and acoustic information of voice are processed during talking, and are relevant to the understanding of clinical conditions involving impaired voice ownership attribution such as schizophrenia.
Chapter 6

6.1 – Introduction

The ability of people to perceive and interact with environments requires integration of input from multiple sensory sources. Recent insights gained from studies of multisensory perception suggest that different people integrate multimodal information with varying degrees of efficiency (Stone et al., 2001; Grant, 2002), and these individual differences might predict individual differences in multimodal perception, such as the strength of illusions based on multimodal integration. Here, I examine this possibility in the context of sensory and sensorimotor illusions that are contingent upon multimodal integration. By assessing illusion susceptibility across individuals, I predict that the subjectively reported strength of the illusion will correlate between two multimodal illusions, despite the fact that the induction of the illusions relies upon distinct modalities. This will support the idea that multimodal integration is a reliable construct within individuals, and does not depend on the specific modalities tested.

Multimodal illusions provide us with an ideal framework to explore cross-modal integration performance since the perceived strength of an illusory percept can serve as an index of the magnitude of multimodal perception. I examine two multisensory illusions: the rubber-hand illusion (Botvinick and Cohen, 1998), and the rubber-voice illusion (Zheng et al., 2011). Although both illusions arise due to correlated input in different modalities, the rubber hand illusion reflects multisensory processing in which information from vision and touch converge to produce the illusory impression that a fake hand is actually one’s own, whereas in the rubber voice illusion, the motoric and/or somatosensory concomitants of speaking are integrated with correlated auditory feedback to generate the illusory impression that the feedback voice is one’s own. The involvement of different modalities in the two illusions allows us to ask whether
perceptual integration of multimodal cues generalizes across tasks, as assessed by examining whether the strength of illusions correlates within subjects.

As a control for subject-specific and other confounding effects, I also employ a third, proprioceptive, illusion (Goodwin et al., 1972; Lackner, 1988). Here, stimulation of the muscle spindles produces the illusory sense that the arm is moving. Unlike the other illusions, the perception of this illusion does not require that correlated information across modalities be integrated. Hence, individual perception of this illusion, compared to the two multimodal illusions, will help us exclude potential confounds that might have contributed to the multimodal illusion susceptibility.

Previous studies have implicated a role for suggestibility in susceptibility to illusions (Eysenck, 1943; Van Ittersum and Wansink, 2011). To further examine whether the perceived strength of illusions may be related to individual suggestibility, I adopt a multidimensional suggestibility scale (Kotov et al., 2004) as a measure of general tendency to accept suggestion.

6.2 – Methods and Materials

Participants

Right-handed female participants (N = 37) were recruited from the Queen’s community. Only female participants were recruited here to accommodate the rubber-voice illusion (as in the rubber-voice illusion, the pre-recorded voice was a female voice; see Chapters 4 and 5). One participant was excluded for making errors in one of the experimental procedures (production errors in the Rubber Voice Illusion; see below), leaving 36 (mean age ± standard deviation [SD]: 20.53 ± 1.81). All participants spoke English as their first language and had no history of neurological or hearing impairment. Informed written consent was obtained from each participant
before the study and each participant received $10 as compensation for their time. All procedures were cleared by the Queen’s General Research Ethics Board.

**Experimental Procedures**

Three experimental procedures (Rubber Hand Illusion, Rubber Voice Illusion, and Somatosensory Illusion; see below) were administered for each participant. The order in which the three procedures were presented was randomly determined for each participant (see Table 6.1 for the number of participants in each permutation of the three procedures). The Multidimensional Suggestibility Scale (Kotov et al., 2004) was administered at the end, as a self-report measure of suggestibility. The rubber voice procedure took place in a sound-attenuating booth, and the rest of the study occurred in a quiet testing room at the Queen’s Biological Communication Centre.

Table 6.1. The number of participants associated with each order in which the three procedures were presented. RHI: rubber hand illusion; RVI: rubber voice illusion; SI: somatosensory illusion.

<table>
<thead>
<tr>
<th>Order</th>
<th>Number of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>RHI-RVI-SI</td>
<td>3</td>
</tr>
<tr>
<td>RHI-SI-RVI</td>
<td>7</td>
</tr>
<tr>
<td>RVI-RHI-SI</td>
<td>11</td>
</tr>
<tr>
<td>RVI-SI-RHI</td>
<td>5</td>
</tr>
<tr>
<td>SI-RHI-RVI</td>
<td>5</td>
</tr>
<tr>
<td>SI-RVI-RHI</td>
<td>6</td>
</tr>
</tbody>
</table>

**Rubber Hand Illusion**

The rubber hand illusion was induced using a method similar to that of Botvinick & Cohen (1998). Each participant was seated with her left arm resting under a mini-table (53.5cm x 30cm x 12cm), while a life-sized rubber model of a left hand was placed on the mini-table directly in front of her. During the experiment, a black blanket was used to cover the left forearm of the participant as well as a large portion of the mini-table, so that only the fingers of the rubber hand
were visible to the participant. The participant was instructed to fixate the rubber hand, while the experimenter used two small paintbrushes to stroke the fingers of the rubber hand and the participant’s hidden hand, synchronizing the timing of the stroking as closely as possible. The strokes were restricted to the middle and ring finger. Which finger was stroked each time was pseudorandom, with the constraint that no more than three consecutive strokes were made on the same finger. This procedure lasted for 2.5 minutes, leading to 150 strokes in total with a frequency of stroking at 1Hz.

Immediately before and after the stroking procedure, participants were required to indicate, with their right index finger, where they felt their left index finger (hidden under the mini-table) was located. With eyes closed, they drew their right index finger along the front edge of the mini-table until they judged it to be vertically aligned with the left index finger, which was resting directly beneath the mini-table during the stroking procedure. The horizontal displacement, if any, between the veridical position of the left index finger and its perceived position (i.e., indicated by where the right index finger was) was marked and measured using a standard ruler.

At the end, participants were asked to complete a questionnaire (adapted from Botvinick & Cohen, 1998) assessing several perceptual effects (e.g., ‘It felt as if the rubber hand were my hand’, see Table 6.2) on a 7-point Likert scale.
Table 6.2. The questionnaire items used in the three illusions are shown.

| **Rubber Hand Illusion**                  | 1. It seemed as if I were feeling the touch of the brush in the location where I saw the rubber hand touched. |
|                                          | 2. It seemed as if the touch I felt was caused by the brush touching the rubber hand. |
|                                          | 3. It felt as if the rubber hand were my hand. |
| **Rubber Voice Illusion**                | 1. It seemed as if the voice I heard was caused by my production of the word. |
|                                          | 2. It felt as if the voice I heard was my own voice. |
| **Somatosensory Illusion**               | 1. It seemed as if I were experiencing forearm flexing at the time of the vibration. |
|                                          | 2. It seemed as if the flexing of my forearm was caused by the vibrator. |
|                                          | 3. It felt as if the forearm being vibrated were not my forearm. |

**Rubber Voice Illusion**

The rubber voice procedure is described in a previous paper (Zheng et al., 2011). Participants were seated in a sound-proof booth in front of a desktop computer (Dell, Inc. U.S.A.). During the experiment, they spoke one word at a time (either ‘day’ or ‘too’), as prompted on the computer screen. They spoke into a microphone and heard concomitant auditory feedback through a set of headphones. Low-level white noise was present in the headphones to minimize the bone-conducted speech feedback during vocalization (Barany, 1938). There were 10 trials of practice at
the beginning during which participants alternated saying ‘day’ and ‘too’, and heard veridical auditory feedback. This was followed by a block of 150 trials, during which participants spoke the two words in a pseudorandom order (such that no more than three consecutive trials of the same word occurred) while hearing congruent, pre-recorded tokens of the two words in another female voice. Both the rate of word presentation and the duration of the experimental session matched those of the rubber hand procedure, with 150 trials of two different stimuli delivered at 1Hz. A real-time speech tracking system developed in our lab ensured that the delivery of the auditory feedback was temporally and phonetically gated with participants’ own vocal production, and that the two streams of acoustic signal (i.e., production and feedback) were separately monitored and recorded for further analysis. The acoustics of the participants’ vocal production and vocal motor drifts towards the heard rubber voice (if any) were measured following the same procedure of the rubber voice illusion (Zheng et al., 2011). At the end, participants were asked to complete a questionnaire assessing the generation and strength of the illusion (e.g., ‘It felt as if the voice I heard was my own voice’, see Table 6.2).

**Proprioceptive Illusion**

This procedure was modified from a proprioceptive illusion first reported by Goodwin et al. (1972) and later by Lackner (1988). Participants rested their forearms on the table and kept their eyes closed. During the experiment, the participants were instructed to slowly move their forearms up off from the table and towards their chest, while keeping the two forearms aligned with each other. A hand-held electromagnetic physiotherapy vibrator (Touch ’n Tone, Conair Consumer Products Inc., Canada) was applied to the biceps tendon at the elbow on their left arm by the experimenter during the movement. The vibration stimulated muscle spindles in the biceps that would normally be stimulated by the muscle’s stretching, thereby creating a kinesthetic
illusion that the forearm was moving away from the face. Consequently, participants would have moved the stimulated forearm faster than the other forearm without awareness (i.e., to compensate for the illusion that it’s actually moving the other way), leading to a displacement of the distance travelled by the two forearms. When the right forearm (i.e., the forearm without stimulation) reached a preset location (∼65° angle between the forearm and the table) the participants were asked to stop the movement, and the displacement of the two forearms, if any, was measured as the horizontally projected distance between the knuckles of the middle fingers of the two hands. This procedure was repeated 3 times for each participant to obtain a stable estimate of the displacement.

After the procedure, participants completed a questionnaire assessing the process of the illusion (e.g., ‘It felt as if the forearm being vibrated were not my forearm’, see Table 6.2).

**Short Suggestibility Scale**

In order to evaluate whether self-reported suggestibility had an impact on the strength of the illusions, the short suggestibility scale (Kotov et al., 2004) was administered at the very end for each participant. This scale contained 21 self-report items on a 5-point Likert scale. The benefit of using this scale is that it has been validated on a large number of human subjects (N = 1,308).

**6.3 – Results**

I measured subjective ratings for all three illusions and questionnaire data for the suggestibility scale. In addition, I measured finger displacement in the rubber hand illusion (Bovinick and Cohen, 1998), vocal production change in the rubber voice illusion (Zheng et al., 2010) and forearm displacement in the proprioceptive illusion (Lackner, 1988) as a way to measure the magnitude of illusory perception more objectively. Data were then analyzed using SPSS (SPSS, Inc., U.S.A.).
Relationships among illusion ratings

Ratings of the key statement designed to measure the illusion strength from both rubber hand illusion (i.e., ‘It felt as if the rubber hand were my hand’) and rubber voice illusion (i.e., ‘It felt as if the voice I heard were my own voice’) were correlated (Spearman $\rho = .510$, $p = .001$), such that participants who were more strongly (or weakly) susceptible to one illusion were also more strongly (or weakly) susceptible to the other illusion (see Figure 6.1). In contrast, the strength of neither the rubber hand nor the rubber voice illusion correlated with susceptibility to the proprioceptive illusion.

Figure 6.1. The significant correlation between the ratings of the two illusions is shown.

Objective measures of illusions

Consistent with the results of previous studies (Botvinick and Cohen, 1998; Ehrsson et al., 2004; Tsakiris and Haggard, 2005), I observed proprioceptive drift in the rubber hand illusion, such that
the perceived position of the participants’ left index finger shifted reliably towards that of the rubber hand (mean displacement ± SD: 2.95 cm ± 2.83; p < .001). This indicates that participants appeared to incorporate the rubber hand into their own body system as a direct consequence of illusion. However, there was no significant correlation observed between the subjective ratings and the proprioceptive drift in the rubber hand illusion.

Similarly, I found vocal motor drift of participants’ production in the rubber voice illusion, such that participants were more likely to shift their F0 towards, rather than away from, that of the heard rubber voice, for the words ‘day’ and ‘too’ at least, $\chi^2(1, N = 36) \geq 5.444, p \leq .02$ (see Figure 6.2), in accord with our previous findings (Zheng et al., 2011). More importantly, I observed a significant correlation between the magnitude of the participants’ F0 shifts towards the heard rubber voice (mean shift ± SD: for ‘day’, 14.96 Hz ± 12.60; for ‘too’, 14.99 Hz ± 13.89) and their subjective ratings: for ‘day’: $R^2 = .213, p = .016$; for ‘too’: $R^2 = .033, p = .010$ (see Figure 6.3). Therefore, it appears that the degree to which participants perceive the rubber voice as being their own voice may be reflected in (and predicted by) how much they change the acoustics of their vocal production. We note, however, that this change of F0 was not due to the acoustic discrepancy between the participants’ voice and the rubber voice, as there was no correlation between the magnitude of the F0 shift and the difference between the participants’ F0 and the rubber voice F0.
Figure 6.2. The averages of participants’ F0s that were higher and lower than the RV F0 are separately shown for ‘day’ (upper panel) and ‘too’ (lower panel). The red lines indicate the end of the practice trials.
Figure 6.3. The relationship between rubber voice ratings and F0 shift is shown for ‘day’ (upper panel) and ‘too’ (lower panel). The ratings were based on the question: ‘It felt as if the voice I heard were my own voice’ on a 7-point Likert scale. It appears that participants who more strongly perceived the rubber voice as being their own voice were also those who more strongly shifted their F0 during production.
Proprioceptive Illusion

Participants generally disagreed with the statement ‘It felt as if the forearm being vibrated were not my forearm’ in the proprioceptive illusion (median ratings = 3, p = .803), since, unlike the rubber hand or rubber voice illusion, there were no multimodal cues for misperception while the arm was vibrated. Consistent with this view, I failed to find a significant correlation between the ratings for this statement and those for the statement ‘It felt as if the voice I heard were my own voice’ in the rubber voice illusion, although there was a trend towards a positive correlation with the ratings for the statement ‘It felt as if the rubber hand were my hand’ in the rubber hand illusion, p = .057. However, participants did exhibit displacement between the two arms as a result of somatosensory perturbation (mean displacement ± SD: 2.66 cm ± 1.36, p < .001), with the vibrated arm moving more slowly (probably due to an illusory flexing feeling) than the other arm.

Subjective Illusion Ratings and Suggestibility Scale

To assess how the ratings of the three illusions might be related to individual suggestibility, I averaged ratings across the 21 items from the multidimensional suggestibility scale (Kotov et al., 2004), and then correlated these averaged scores with the illusion ratings. Neither the rubber hand illusion (‘It felt as if the rubber hand were my hand’) nor the rubber voice illusion (‘It felt as if the voice I heard were my own voice’) ratings were predicted by the suggestibility scale (rubber hand/suggestibility: ρ = -.030, p = .864; rubber voice/suggestibility: ρ = .011, p = .948), suggesting that individual differences in susceptibility of these two illusions are not simply attributable to suggestibility. However, there was a significant correlation between the somatosensory illusion ratings and the averaged scores of the suggestibility scale (ρ = .370, p =
indicating that participants who more strongly endorsed the statement ‘It felt as if the forearm being vibrated were not my forearm’ were also more suggestible.

6.4 – Discussion

By using illusion-inducing paradigms and self-report questionnaires, the present study examined the possible involvement of multimodal integration in the self-attribution process. Three main findings emerged from the results: 1) the two self-related illusions (the rubber hand illusion, RHI and the rubber voice illusion, RVI) were predictive of each other, despite the fact that distinct sets of modalities were required for the induction of the two illusions; 2) neither of the two illusions pertaining to sense of self was predicted by the subject’s self-reported suggestibility, as a significant correlation between the perceived illusion strength and suggestibility ratings was not observed; and 3) for the RVI, the subjective and objective measures of the illusion were correlated, such that the individuals who more strongly perceived the rubber voice as being their own voice were also those who shifted their F0 more towards the rubber voice. Taken together, the findings appear to suggest that the ability to integrate multimodal cues, rather than individual suggestibility, contributes significantly to the induction of the self-related illusions, and during the course of illusion development, subjective perception of the illusion and objective behavioral response were tightly coupled.

It has been argued that the induction of the RHI critically depends upon the multimodal integration of visual, tactile, and proprioceptive inputs (Botvinick and Cohen, 1998; Ehrsson et al., 2004; Tsakiris and Haggard, 2005; Longo and Haggard, 2012), as asynchronous visual and tactile stimulation abolishes the illusion (Ehrsson et al., 2004; Makin et al., 2008; Shimada et al., 2009). However, whether the perception of the illusion is related to personal suggestibility, and whether the malleability of the sense of self based on multimodal sensory evidence is a general
phenomenon (i.e., observable in another, different self-related multimodal illusion dependent on a different set of modalities) have never been tested. The current study addressed these questions by showing that the perceived strength of the RHI was correlated with that of a seemingly unrelated illusion, the RVI, that also pertains to the sense of self. Since the induction of the two illusions (RHI and RVI) rely on different modalities (i.e., visual, tactile, and proprioceptive for the RHI; motoric, somatosensory, and auditory for the RVI), our findings appear to suggest that the development of multimodal illusions related to the sense of self is unlikely to be the result of strong influence from one single modality; rather, it may reflect a high-level, cognitive process that somehow integrates multimodal inputs giving rise to the coherent percept – a cross-modal mechanism that operates on many different modalities.

For the RHI, previous work (Botvinick and Cohen, 1998; Ehrsson et al., 2004; Longo et al., 2008) demonstrated a significant correlation between subjective perception and proprioceptive drifts associated with the illusion. However, the present study and other studies (e.g., Morgan et al., 2011) failed to reveal such a correlation. Although it is not entirely clear what factors contribute to the mixed results on the relationship between subjective and objective measures (Ehrsson, in press), lack of power might play a role in the current study. Note, however, that there was a significant drift towards the rubber hand during the illusion development, indicating that incorporating the rubber hand into the subject’s own bodily system did have a strong impact on their behavioral response.

The proprioceptive illusion was adapted from the original illusion developed by Goodwin et al. (1972) and Lackner (1988). Previously this illusion was implemented by having people hold their arms in midair and then move them down during the stimulation of biceps tendon by a vibrator. Vibrating the biceps tendon stimulates the muscle spindles in the biceps, which are also activated
when the biceps are stretched, giving the illusion that the elbow is extending. Thus people would ‘compensate’ for the illusory extension by moving the stimulated arm more slowly than the unstimulated arm. Here, what the subjects were asked to do (in terms of moving direction) was the opposite of what was previously done – subjects were instructed to move their forearms up off a table and towards their chest. Since stimulation of muscle spindles simulates extension, participants would have moved their stimulated arm even faster/further than their unstimulated arm, in order to compensate for the illusory extension. However, what I observed was that the stimulated arm was actually lagging, instead of leading, at the end of the vibration period. One possible explanation is that the duration of the vibration in the current study, ranging from 45 – 75 secs, was longer than that used in previous studies. This longer duration would have allowed for windows of movement reversals near the end of the vibration so that, moving towards (as in compensation) might have developed into moving away, from the face while the other arm was catching up (e.g., see Seizova-Cajic et al., 2007). An alternative explanation is that perhaps the stimulation caused proprioceptive uncertainty, so that it’s difficult for participants to tell whether they were extending or flexing. This uncertainty was reflected in a slower, less immediate movement. Note that these views are highly speculative, and more studies need to be conducted in order to clarify the issue.

Individual ratings from the RHI and RVI did not correlate with individual scores on the suggestibility instrument, indicating that integrating an alien voice or a prosthetic hand into the ‘self’ system does not simply reflect a general tendency to accept messages/suggestions. The multidimensional suggestibility scale (Kotov et al., 2004) was chosen to measure the suggestibility in the present study because the structure is replicable and the scales are reliable (i.e., validated on a large number of subjects).
As noted by Kotov et al. (2004), a high rating on the suggestibility scale is expected to be associated with yielding to messages and to increased responding on persuasion tasks. This is presumably why the subjects who scored high (low) on the suggestibility scale also gave a high (low) rating to the statement ‘It felt as if the forearm being vibrated were not my forearm’ in the proprioceptive illusion. Unlike the matched alien voice in the RVI or the prosthetic hand in the RHI, there was little conflicting information that could have misled the subjects to claim that the vibrated arm was NOT their forearm. Thus, it seems reasonable to argue that the subjects who were likely to agree with the statement are also potentially more suggestible, resulting in high ratings in the self-report suggestibility scale.

To summarize, the present study reveals, for the first time, that the degree to which perception of self is malleable based on coherent multisensory input. The self-attribution process in two multimodal self-related illusions, the rubber hand illusion and rubber voice illusion, does not appear to be associated with suggestibility. Instead, it seems to depend on a mechanism governing cross-modal integration, where the information being integrated can come from any of several sensory domains. The findings advance our understanding regarding how people use multimodal cues for the perception of human attributes and shed new light on the underlying mechanisms of self-recognition.
Chapter 7

7.1 – General Discussion

Converging evidence supports a close link between speech perception and production, yet the cognitive mechanisms and neural architecture of such a link remains elusive. In this thesis, I have reported a series of studies directed towards understanding the cognitive and cortical organizations subserving the coupling between speech production and perception of the auditory concomitants of speech. This was made possible by using a real-time speech tracking system that allowed for the assessment of both perceptual and cortical response to acoustic changes during talking. The methodology, combined with novel experimental designs, enabled me to make valuable contributions to the field, which I summarize below.

The reported neuroimaging studies further clarify the functional overlap between brain regions involved in speech motor control and in speech perception, and reveal three functionally differentiated neural systems underlying auditory feedback control of speech. The results suggest, for the first time, that there is an extensive network of regions sensitive to speech production errors, and within the network, there are functionally specialized systems with somewhat specific properties associated with distinct aspects of articulatory control. The extent and complexity of the revealed network go substantially beyond the predictions of the computational models of auditory feedback control (e.g., Guenther, 2006; Hickok et al., 2011) and significantly advance our fundamental understanding of the neural architecture of speech motor control. In particular, the MVPA study presented in Chapter 3 provides a concrete example of how a novel brain imaging analytic framework can be applied to offer insights into the functional organization of cognitive processing that would not be possible with conventional image analysis.
The link between production and perception is not restricted to speech motor control. The human voice is the direct consequence of speech production, and as such, perception of own voice identity may be best studied online as individuals are vocalizing. However, there is a striking omission in the literature regarding how people perceive the identity of a voice that is heard as a consequence of their own production. Chapters 4 and 5 explored voice identity perception when an alien voice is presented as an auditory concomitant of speech, and demonstrate, for the first time, that talkers perceive the alien voice as being their own voice modified, despite the salient acoustic disparity. The presented findings represent both theoretical and empirical advancement in our understanding of human voice identity perception. Just as the rubber hand illusion (Botvinick and Cohen, 1998) offers a valuable framework to study the cognitive mechanisms of body-ownership, the studies described in Chapters 4 and 5 open a new window to explore the sense of ownership of voice, and are relevant to our understanding of clinical conditions involving impaired voice ownership attribution, as well as of the cognitive mechanism underlying the distinction between self and other.

Capitalizing on a range of multimodal illusions and questionnaires, Chapter 6 explored the cognitive factors contributing to the perception of the voice presented as an auditory concomitant of speech, and identifies multimodal integration, rather than individual differences in suggestibility, as a potentially important mechanism mediating the recognition of a self-generated voice. This study represents a first attempt to link voice-ownership with body ownership. When placed in the context of self-ownership literature, may point to future directions of research related to the remarkable malleability of the perceived self (e.g., cognitive factors implicated in the self-attribution process).

**7.2 – Future work**
This thesis also leaves unsolved questions. For example, if there are multiple, distributed, and functionally differentiated systems supporting speech motor control, would a similar neural architecture also be observed in other types of sensorimotor control? This is a potentially fruitful area to explore as sensory feedback has been shown to be critical for motor movement and actions (e.g., Grafton, 1994; Shadmehr et al., 2010). Furthermore, perception of self-generated voices can also be studied with the addition of facial information in order to clarify how people make perceptual judgments when seeing their own faces v.s. others’ faces making either congruent or incongruent sounds. This line of exploration would undoubtedly shed new light on human voice processing and is relevant to the literature on audiovisual speech perception as well.

The presented studies in the thesis also have clinical implications. For example, do patients with speech motor control pathologies (e.g., stuttering) have altered functional coupling and anatomical connectivity among the extensive network of regions revealed in our MVPA study? In addition, it has been suggested that patients with psychotic disorders such as schizophrenia are more likely to perceive an alien voice as being from themselves (see Hugdahl et al., 2007) and they are also more susceptible to the rubber hand illusion (Thakkar et al., 2011). Would such patients also demonstrate a correlation between the perceived strengths of the rubber voice and rubber hand illusions as shown in our behavioral studies? Future work directed towards addressing these questions would further elucidate the cognitive and neuroanatomical relationships between speech production and perception, and enhance our understanding of the underlying mechanisms of a range of clinical conditions.
References


