Published in final edited form as:

Neuroimage. 2012 May 15; 61(1): 314–322. doi:10.1016/j.neuroimage.2012.02.068.

Understanding The Neural Mechanisms Involved In Sensory Control Of Voice Production

Amy L. Parkinson^a, Sabina G. Flagmeier^a, Jordan L. Manes^a, Charles R. Larson^b, Bill Rogers^a, and Donald A. Robin^a

^aResearch Imaging Institute, University of Texas Health Science Center San Antonio, San Antonio, TX, USA

^bDepartment of Communication Sciences and Disorders, Northwestern University, Evanston, IL, USA

Abstract

Auditory feedback is important for the control of voice fundamental frequency (F0). In the present study we used neuroimaging to identify regions of the brain responsible for sensory control of the voice. We used a pitch-shift paradigm where subjects respond to an alteration, or shift, of voice pitch auditory feedback with a reflexive change in F0. To determine the neural substrates involved in these audio-vocal responses, subjects underwent fMRI scanning while vocalizing with or without pitch-shifted feedback. The comparison of shifted and unshifted vocalization revealed activation bilaterally in the superior temporal gyrus (STG) in response to the pitch shifted feedback. We hypothesize that the STG activity is related to error detection by auditory error cells located in the superior temporal cortex and efference copy mechanisms whereby this region is responsible for the coding of a mismatch between actual and predicted voice F0.

Keywords

Vocalization; Auditory feedback; fMRI; Audio-vocal integration; Pitch shift

Introduction

Understanding the sensory contributions to motor control allows for a greater insight into how motor skills are acquired and performed by healthy people and those with neurological disorders. Voice production is a highly complex motor skill that provides a unique platform for understanding sensory mechanisms of motor output. Given the high incidence of voice disorders in the population, it is important to understand the neural basis of vocal control in order to assist clinical decision making relative to conditions that result from impaired phonation. Also, knowledge of the neural systems involved in the sensory control of the voice is critical given the role of auditory feedback in voice control. In particular, the use of auditory perturbations to understand voice motor control has been highly productive (Bauer et al., 2006; Burnett et al., 1998; Hain, et al., 2000), and audio-vocal integration and the

*Corresponding author: Parkinson@uthscsa.edu, Tel: +1 210-567-8169 Fax: +1 210-567-8152.

Present address: Research Imaging Institute, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX 78229

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ability to control voice fundamental frequency (F0) is essential for successful and efficient speech production.

Accurate voice F0 and amplitude control depend on vocal error detection and correction mechanisms that process sensory feedback (Burnett, et al., 1998; Hain, et al., 2000). These mechanisms have some features of reflexes (automatic and short latency) as well as voluntary responses (modification by intent) (Hain, et al., 2000). Forward models predict the next state of a process given information on the current state and motor commands via feedback from motor and sensory systems. This information is then used to correct and control output via feedback signals to the regions of the brain controlling motor output. This classic view of forward models predicting the sensory consequences of an action has been considered in human movement, vocal production and speech production (Wolpert et al, 1995; Paus et al., 1996; Houde and Jordan, 1998; Larson et al., 2008). There is also considerable evidence from studies on deafness and auditory masking that voice F0 control depends in part on auditory feedback (Angelocci et al., 1964; Binnie et al., 1982; Burns & Ward, 1978; Elliott & Niemoller, 1970; Lane, et al., 1997). The loss of auditory feedback through masking or deafness causes an increase in the absolute value of voice F0 and an increase in F0 variability (Burns & Ward, 1978; Elliott & Niemoller, 1970; Lane, et al., 1997; Murbe et al., 2002; Ringel & Steer, 1963; Svirsky et al., 1992). Moreover, many treatments for voice disorders rely heavily on auditory feedback through which patients are taught to listen to and monitor their own voice productions (e.g. Fox, et al., 2006; Ramig et al., 2001).

There have been numerous studies using pitch perturbation techniques to define these vocal responses to pitch shifted feedback and investigate the role of auditory feedback in voice F0 control (Burnett et al., 1997; Hain, et al., 2000; Larson, 1998; Tourville et al., 2008; Xu et al., 2004). The current study used an auditory feedback paradigm (Larson, 1998) in which the pitch of self-voice feedback is unexpectedly changed in real time during vocal production. The pitch shift paradigm involves the subjects listening to their own voice, and during vocalization the pitch of self-voice feedback is unexpectedly changed up or down. Subjects respond to the pitch-shift by changing their voice F0. Responses are made either in the opposite direction of the feedback, known as a compensatory response, or in the same direction as the feedback termed a “following” response. Response direction is influenced by the magnitude of the shift such that as stimulus magnitude is increased (from 25 to 300 cents) the proportion of compensatory responses decreases (Burnett, et al., 1998). The duration of the pitch shift stimulus has also been shown to affect the latency, magnitude and duration of the vocal response (Burnett et al., 1998; Hain et al., 2000). The exact neural networks underlying these vocal responses to sensory perturbations of the voice are not well understood. Thus, the current study explores functional neural networks using fMRI to define the regions involved.

It is likely that the above-described vocal responses to sensory perturbations rely upon information provided by an efference copy mechanism that predicts the motor consequences of the intended action. This idea of sensorimotor adaptation in speech production has been presented in several forward models of neural control of the voice (Borden, 1979; Houde and Jordan, 1998; Bauer, et al., 2006; Guenther et al., 2006; Hickok et al., 2011; Larson et al., 2008; Tourville & Guenther, 2010). Larson et al (2008) suggest a forward model of voice frequency control where perceived F0 from kinesthesia and audition feedback and desired F0 is compared, resulting in an error signal. This error signal is then filtered and delayed and used as feedback to motor control regions to adjust the motor drive for voice F0. Hickok et al (2011) expand this idea and suggest that sensorimotor integration control of speech production is achieved via a state feedback control mechanism where online articulatory control is achieved primarily via internal forward model predictions, and actual

feedback is used to train and update the internal model. Predicted consequences are then compared against both intended and actual auditory targets of a motor command. This view further supports the idea of online error correction and detection whereby deviation between the predicted and intended target results in an error correction signal that is fed back to the motor controllers. Hickok et al (2011) hypothesized that a specific region in the auditory cortex located in the planum temporale, at the parietal-temporal boundary (Spt), is responsible for the process of integrating auditory feedback in voice production.

It is important to understand the mechanisms controlling vocal responses to auditory feedback perturbations and to link computational models with empirically derived cortical regions. In efforts to understand the neural mechanisms and cortical regions involved in voice F0 control, there have been several studies using sparse sampling fMRI with auditory feedback shifted paradigms. Zarate and Zatorre (2010; 2008) used fMRI to examine voluntary changes in F0 in response to long duration (e.g., > 0.5 s) pitch shifts during vocalization. They found that singers compensated for pitch-shifted feedback by recruiting the putamen, superior temporal gyrus (STG) and anterior cingulate cortex (ACC), whereas non-musicians showed increased activity in the premotor cortex. Zarate and Zatorre (2005) also suggested that the left ACC and insula are important in audiovocal integration in an fMRI study with singing. Toyomura et al (2007) used a simple vocalization task with pitch shifting in a sparse sampling fMRI setup and noted increased activity in the right supramarginal gyrus, prefrontal (BA9) area, anterior insula, superior temporal area and intraparietal sulcus. Tourville et al (2008) also used a sparse sampling fMRI to examine neural responses to auditory feedback in which the first formant frequency (F1) of their speech was unexpectedly shifted in real time during production of monosyllabic words. Results of this study demonstrated that responses to unexpected auditory perturbations of F1 frequency activated bilateral STG, right prefrontal and Rolandic cortical regions.

There is another critical aspect of the previous pitch-shift studies that bears comments. It is known that pitch-shift stimuli exceeding 300ms in duration elicit voluntary responses whereas short stimuli of less than 300ms show no evidence of voluntary responses (Burnett, et al., 1998; Hain, et al., 2000). The activations in a previous study (Zarate, et al., 2010) using pitch-shifted stimuli are likely driven in part by voluntary mechanisms because the stimuli were longer than 300 ms in duration, so it is not clear whether the activations reported were a result of voluntary or reflexive responses. Tourville et al (2008) looked at long-range compensation to a formant shift resulting in a phonemic change rather than short-term (reflexive) changes associated with perturbation. It may also be expected that the networks involved in auditory feedback of voice regulation are different from those used in speech regulation through the involvement of more complex articulator control, as suggested by Tourville et al (2008). Thus, there is very little evidence relevant to the neural networks involved in pure audio-vocal pitch shift responses during vocalization. Because we are shifting voice pitch feedback and not formant frequencies (Tourville et al., 2008) we are able to examine sensory control of the voice without inference from higher level phonemic processing mechanisms. Because of this lack of information about voice control, we asked the question of whether the auditory cortex is responsive to unexpected, brief (200ms duration), small magnitude (100 cents) shifts in voice pitch feedback during vocalization. Based on the work discussed above, we hypothesized that the auditory cortex would be a primary structure that is functionally involved in the pitch-shift vocal response, specifically showing an increase in activity as a result of auditory error cells responding to pitch shifted stimuli. We also hypothesized that other regions found active in perturbation of speech (Tourville et al., 2008), such as primary motor and pre-motor regions, would also activate during perturbation of vocalization because these areas are also involved in voice control, i.e., F0 as controlled by laryngeal muscles.

Methods

Subjects

Twelve right-handed subjects (5 male, 7 Female, mean age 28 ± 11 years) with no history of neurological disorder participated in the study. All potential subjects underwent a pre-screening session to check eligibility for the study. This consisted of monitoring audio-vocal responses to the pitch shift paradigm and completing an MRI safety-screening questionnaire. All had passed a hearing screening, reported no neurological deficits, had no speech or voice disorder, had no previous musical experience in the past 10 years and were English speaking. The institutional review board of the University of Texas Health Science Centre at San Antonio approved all study procedures.

Experimental Procedure

Electrostatic headphones (Koss ESP950) were placed on the subject's head to provide acoustic feedback during each trial. The subject lay in the scanner and viewed a monitor screen reflected off a mirror directly above the eyes. Subjects were instructed to keep their jaw, lips and tongue motionless during the testing and to keep their jaw slightly open to minimize any artifact caused by movement during vocalization. Each trial began with the presentation of a speech or control visual cue stimulus projected on a screen viewable from within the scanner. The vocalization stimulus consisted of the word "ahhh" and the control stimulus consisted of the word "rest" projected onto the screen at the onset of the trial. Subjects were instructed that on the "ahhh" trials they should quickly inhale and then sustain an "ah" sound at a constant pitch and intensity until the "ahhh" cue disappeared from the screen. Subjects were also instructed to remain silent when the control stimulus cue appeared. Stimuli remained on the screen for 5 s. At the same time that the visual cue ("ahhh") appeared, a brief (500 ms) auditory cue consisting of a pre recording of the subject vocalizing the "ahhh" sound was presented over the headphones. The auditory cue was presented in order to give the subject an appropriate pitch level at which to vocalize and also to direct attention to the task.

A Matlab program provided paradigm delivery for fMRI. The Matlab program provided visual prompts, audio feedback, data acquisition and MRI triggering. By tying these aspects together in a single application, precise timing was achieved between audio capture, pitch shift timing, and MRI scan timing. The Psychtoolbox for Matlab was used for timing and delivery of visual cues. A custom C++ Matlab extension was created to generate pitch shifted audio feedback using an Eventide Eclipse Harmonizer audio processor via a MIDI connection. The C++ extension incorporated the RtAudio API (Gary P. Scavone, McGill University) for audio input and output and the RtMidi API (Gary P. Scavone, McGill University) for controlling the audio processor. The Matlab program also output TTL signals via the computer parallel port to trigger the scanner. MRI acquisition occurred after each vocalization avoiding the interference of scanner noise with the experiment. Additional TTL signals were output for the duration of each pitch shift for synchronization between audio capture and pitch shifts. A separate Powerlab data acquisition system recorded audio data and vocal perturbation timing.

Shortly after the subject began vocalizing (500–700 ms), the first of five pitch-shift stimuli was presented. With an inter-shift interval of 500–700 ms, four additional pitch-shift stimuli were presented during the same vocalization. Each pitch-shift stimulus was 100 cents (100 cents = 1 semitone) in magnitude and maintained for a duration of 200ms (Figure 1). All 5 shifts within a single trial were in the same direction (up or down). An experimental block consisted of 64 trials, 48 vocalization trials (16 shift-up, 16 shift-down, 16 no-shift) and 16 control trials. The trials were presented in a random order. Each subject performed 3

experimental blocks within the session and there was a 2-minute rest period between each block. Subject's vocalizations were recorded through a microphone (Shure SM93). Vocal signals were sampled at 10kHz, digitized using a Powerlab 8SP multi-channel recording system and recorded using Powerlab Chart Software (Version 5.2.1).

Acoustic data analysis

Data were stored to a PC workstation and analyzed off-line. The digitized voice F0 and feedback contour signals were constructed using an autocorrelation function in Praat software (Version 5.2.28, www.praat.org) and converted from Hz to cents with the formula:

$$RMag = 12 \times \frac{\log(F0/196)}{\log 2} \times 100$$

where Rmag = response magnitude, F0 = F0 waveform in Hz, 196= arbitrary reference. The voice, feedback and TTL control pulses were then displayed with a 200ms pre- and 500ms post-stimulus window on a monitor using Igor Pro (Wavemetrics, Inc.). All signals were time-aligned with the trigger pulse and averaged with like trials (e.g., 100 cent upward shift). An event related average of each individual channel was computed separately for upward and downward stimuli. Initial acceptance of averaged responses was defined by the averaged waveform exceeding a value of two standard deviations (SD) of the pre-stimulus mean (baseline F0) beginning at least 60 ms after the stimulus and lasting at least 50 ms. Measurements of response magnitude, peak latency and direction were made from the averaged waveforms. Latency of the averaged response is the time from stimulus onset until the response exceeds 2 SDs of the pre-stimulus mean. Response magnitude was measured from the averaged voice F0 contour at the point of greatest deviation from the baseline mean.

MRI data acquisition

All structural and fMRI data were acquired on a Siemens Trio 3-T Scanner. Foam padding was applied between the subject's head and the coil to help constrain head movement. Three full-resolution structural images were acquired using a T1-weighted, 3D TurboFlash sequence with an adiabatic inversion contrast pulse with a resolution of 0.8 mm isotropic. The scan parameters were TE=3.04, TR=2100, TI=78ms, flip angle=13, 256 slices, FOV=256mm, 160 transversal slices. The three structural images were combined to create an average, which was then used to register the brain of each subject to their functional data. The functional images were acquired using sparse sampling technique. T2* weighted BOLD images were acquired using the following parameters; FOV 220mm, slice acquisition voxel size = 2×2×3mm, 43 slices, matrix size = 96 × 96, flip angle = 90, TA = 3000ms, TR=11250ms and TE = 30ms. Slices were acquired in an interleaved order with a 10% slice distance factor. Each experimental run of the task consisted of 64 volumes.

Functional data were obtained using a sparse sampling technique triggered by a digital pulse sent from the stimulus computer for each event. A single volume taking 3 seconds was acquired per trial (Figure 1) beginning 5.25 seconds following the onset of the trial (visual "ahhh" cue and auditory cue). The delayed acquisition was necessary to allow collection of the bold signal near the peak of the hemodynamic response for speaking which has been estimated to occur approximately 4–7 seconds after vocalization (Tourville, et al., 2008). Auditory feedback was turned off during scanning to prevent transmission of scanner noise over the headphones. There was a three second delay following the scan before the onset of the next trial in order to allow task-related cerebral blood flow to return to baseline before the next trial.

MRI data analysis

Preprocessing and Voxel-based analysis of the data were performed using the SPM8 software package (Wellcome Department of Imaging Neuroscience, London, UK) to assess task-related effects. Functional images were realigned to the mean EPI image. The realignment parameters were included as covariates of non-interest in the study design during model estimation. Functional images were then co-registered with the T1-weighted anatomical scan, spatially normalized to a template in Montreal Neurological Institute (MNI) space and then spatially smoothed using an isotropic Gaussian kernel with full width at half-maximum (FWHM) of 6mm.

The BOLD response for each event was modeled using a single-bin finite impulse response (FIR) basis function spanning the time of acquisition on the volume (3 seconds). Three separate predictors were defined as No Shift (NS), Shift Up (SU), and Shift Down (SD). Voxel responses were fit for each of the three predictors according to the general linear model. Group statistics were assessed using a random-effects general linear model (RFX). We report the resulting group parametric maps thresholded at p -value of 0.05 (FWE corrected for multiple comparisons) unless otherwise stated.

Results

Behavioral results

Vocal response data were analyzed from 11 of the 12 subjects, as there were software problems during the recording of one subject's data (but this subject did pass the screening test showing the expected responses to the pitch shifted stimuli). This subject was still included in the fMRI analysis. Average response magnitudes for both upwards and downwards pitch shifted stimuli are displayed in figure 2. Average response magnitude was 21.36 ± 10.88 cents (mean $\pm$ SD) for downward shifted stimuli and -17.47 ± 5.48 cents for upward shifted stimuli. Average time to peak of the response was 232 ± 94 ms for the downward shifted stimuli and 202 ± 52 ms for the upward shifted stimuli. Analysis of vocal intensity changes in response to the pitch shift indicated no significant increases or decreases in loudness in response to the pitch shift stimuli.

Neural responses to the pitch shift

fMRI data were analyzed from all 12 subjects. BOLD responses during normal and shifted vocalization are displayed in Figure 3 and Table 1. These contrasts identified a number of regions that exhibited increased BOLD responses to both shifted and unshifted vocalization compared to no vocalization (rest). The results of the random-effects group analysis revealed a large network of areas including bilateral auditory cortex, pre-motor cortex, primary motor cortex, supplementary motor area (SMA) and prefrontal cortex for both shifted vocalization versus rest and unshifted vocalization versus rest (Figure 3, Table 1).

There were no significant differences in brain activity related to the direction of the shift after performing the Shift Up - Shift Down contrast, and so for contrasts of Shift - NoShift, the trials with upwards and downwards shifts were combined and compared with the NoShift trials. This analysis revealed that the auditory cortical regions of the brain in the superior temporal gyrus were active during the audio-vocal responses to the pitch shift. A random-effects analysis of Shift - NoShift revealed bilateral superior temporal gyrus and middle temporal gyrus activation (Figure 4). The activation within the superior temporal gyrus region extended from the posterior superior temporal gyrus to more anterior regions of the superior temporal gyrus. The cluster of activation in the left hemisphere also extended into the middle temporal gyrus. The contrast of NoShift - Shift did not reveal any significant activations.

Discussion

In the present study a pitch-shift perturbation paradigm was used to investigate the neural activations related to audio vocal responses to pitch shifted vocalization. We used sparse sampling fMRI that allows subjects to vocalize and to hear their digitally processed voice feedback in the absence of scanner noise (Engelien, et al., 2002; Tourville, et al., 2008; Yang, et al., 2000). We hypothesized that a primary region of increased activation would be the auditory cortex in response to unexpected perturbations in the pitch of auditory feedback during vocalization. In support of our apriori hypothesis, the main finding was increased BOLD activity bilaterally within the superior temporal gyrus during pitch-shifted vocalization compared to non-shifted vocalization, consistent with previous studies on auditory feedback disruption (Zarate, et al., 2010; Zarate & Zatorre, 2008). Indeed, this was the only region found to be significantly active using a random effects model. Because other studies used fixed effect models (sparse sampling has a relatively poor signal to noise ratio) we analyzed the contrasts with this less conservative statistical approach. When this was done we also found activations associated with pitch shifted stimuli in left lateralized premotor cortex, right middle frontal gyrus, left inferior frontal gyrus, superior frontal gyrus and bilateral cerebellum. Below we first discuss our primary finding of superior temporal gyrus activation.

In understanding the role of the STG in feedback control of the voice one must take into account the well known fact that neural control of the voice depends not only on auditory feedback but other forms of sensation as well. Previous studies have demonstrated the importance of both kinesthetic and auditory feedback for the control of voice F0 (Angelocci, et al., 1964; Binnie, et al., 1982; Burnett, et al., 1998; Burns & Ward, 1978; Elliott & Niemoller, 1970; Hain, et al., 2000; Lane, et al., 1997). Based on an extensive set of data, Larson et al (2008) proposed a model incorporating both feed-forward and feedback pathways of both auditory and kinesthetic feedback in the regulation of vocal output. In order to examine a possible interaction between auditory feedback and kinesthetic feedback for control of voice F0, local anesthetic was applied to the vocal folds in the presence of perturbations in voice pitch feedback (Larson, et al., 2008). They showed that responses to pitch shifted voice feedback were larger when the vocal fold mucosa was anesthetized than during normal kinesthesia. This model suggests that a feasible explanation for the increase in response magnitude with vocal fold anesthesia is that the motor system uses both auditory feedback and kinesthesia to stabilize voice F0 shortly after a perturbation of voice pitch feedback has been perceived (Larson, et al., 2008). Results from the present study would suggest that the superior temporal gyrus is a primary location where the predicted F0 is compared with feedback from kinesthesia and audition (actual F0) and where error signals are produced.

The concept of a feed-forward system projecting information about motor commands in order to predict the sensory consequences of an action is well supported (Wolpert et al., 1995; Blakemore et al., 1999; Jeannerod et al., 1998; Wolpert, 1997). This idea has most often been applied to visual or tactile modalities. More recently the idea of efference copy and the role of a forward model in the auditory modality is also supported by MEG (Houde & Jordan, 2002a) and ERP (Behroozmand & Larson, 2011; Heinks-Maldonado et al., 2005) studies. In this feed forward voice motor control system, corollary discharge (predicted sensory consequence of motor output) facilitates the comparison of incoming vocal feedback to predicted vocal output. Recent research suggests that the pitch-shift response relies upon information provided by an efference copy mechanism that predicts the motor consequence of an action and allows for the differentiation of self-voice and external sound (Behroozmand & Larson, 2011). There is also evidence that unexpected somatosensory feedback perturbations during speech (Golfinopoulos, et al., 2011) produce activations

similar to those produced in auditory feedback perturbations of speech. This finding suggests that the STG activations in response to perturbed feedback are not necessarily specific to auditory feedback but are more likely representative of the identification of an error between the expected and actual vocal output.

Consistent with our *a priori* hypotheses we found increased BOLD activity in auditory cortex associated with the pitch shift. This increased BOLD activity in the STG region in response to pitch-shifted feedback may be a result of reduced suppression related to efference copy. The efference copy process has been found to suppress activation (motor induced suppression) in the auditory cortex when expected and actual output match (i.e., during self-vocalization); however, when there is a mismatch (e.g. pitch shifted feedback), suppression is reduced (Aliu et al., 2009; Heinks-Maldonado & Houde, 2005; Houde & Jordan, 2002b; Liu et al., 2010a). Several ERP studies have demonstrated greatest suppression of the N1 component during vocalization with unaltered feedback and reduced suppression in conditions where the voice feedback is shifted (Behroozmand & Larson, 2011; Heinks-Maldonado et al., 2005). We recognize that Hickok (2011) has suggested a model that places the sensorimotor integration and error detection mechanism in the Spt region, a region in the auditory cortex in the planum temporale, at the parietal-temporal boundary located in the Sylvian fissure. We address this below by first providing compelling support for our hypothesis from converging data based on electrophysiological studies of the pitch shift response.

Critical to our interpretation of the role of STG in efference copy mechanisms is the fact that similar effects have also been shown in recording directly from the cortex of the STG using electrocorticography (Greenlee, et al., 2011). It is presumed that when there is a match between expected and actual output, a cancellation of sensory input results, hence the suppression seen in auditory cortex. However, when there is a mismatch, the lack of corollary discharge fails to cancel out the sensory feedback and there is an increase in sensory experience (Heinks-Maldonado & Houde, 2005; Houde & Jordan, 2002b). We hypothesize that this reduction in suppression is reflected by the increase in BOLD activity seen in the auditory cortex in response to the shifted vocalization condition, that is, without the shift, the BOLD activity in this region would be lower than found in this study. In the condition where the feedback is not shifted, there is no mismatch between expected and actual output, hence there is suppression in this region and a reduction in BOLD when compared to the pitch-shifted condition (see figure 3). As well, ERP data from our laboratory also shows increased suppression of the N1 component during vocalization with unaltered feedback (Behroozmand & Larson, 2011).

A recent fMRI study in speech also supports the idea of suppression of the auditory cortex when there is a match between expected and perceived feedback (Christoffels et al., 2011). It was shown that the BOLD signal in left and right hemisphere STG ROI's increased proportional to the parametric level of masking noise of feedback during overt speech. However no parametric brain activity was found in regions outside of the auditory cortex such as the SMA, pre-motor regions and primary motor regions. The authors suggested that the non-auditory areas are not involved in assessing the degree of mismatch between expected and perceived feedback. We also show similar findings in vocalization, where increased activity related to shifted feedback is only in bilateral auditory areas. This observation also provides support for the idea that the STG is the primary region involved in monitoring mismatches between actual and predicted feedback.

Counter to our findings, Hickok et al (2011) indicated that Spt plays an important role in integrating auditory feedback with motor commands in the control of voice production. Tourville et al (2008) found that Spt was active during a formant shift perturbation which

alters the phoneme. They interpreted this finding as evidence that Spt is part of a state feedback control system along with STG, premotor cortex and the cerebellum. However, the paradigm used by Tourville et al (2008) might be the reason for different findings. First, they looked at long term compensation relative to perturbation which may result in recruitment of more or different regions because a more long range compensatory response likely involves different processing than the pitch shift response which occurs within 250ms. Secondly, Tourville et al (2008) studied a perturbation which led to a phonemic shift (that resulted in a linguistic change at the phonological level of processing). The Spt region is known to be involved in language processing and this may therefore account for the fact that this region did not activate in our study.

In order to identify regions of the brain specifically involved in vocalization, Brown et al (2008) recently conducted a meta-analysis of phonation (syllable singing), which was then compared with Turkeltaub et al's (2002) meta-analysis of word production. Results of this meta-analysis showed significant areas of overlap in the larynx motor cortex, supplementary motor area (SMA), the rolandic operculum, posterior superior temporal gyrus and cerebellum. However, there were also areas where activation was specific to syllable singing compared to overt reading in the frontal operculum, anterior superior temporal gyrus, putamen and thalamus. These results help to identify the specific brain regions associated with models of voice control. The authors suggest a neural model of vocalization in which the principal regions for the control of phonation in speaking and singing are the posterior part of the superior temporal gyrus, the larynx motor cortex and associated premotor areas, the cerebellum (lobule VIII) and the supplementary motor area (SMA). These data show similarities with the "basic speech production network" proposed by Bohland and Guenther (2006) where it is suggested that added sequence and syllable complexity leads to increased engagement of this speech network and recruitment of additional brain areas. If this is the case then it may be expected that vocalization will involve a reduced engagement of this network when compared to more complex speech.

Further evidence of the involvement of the superior temporal gyrus in a feed-forward system to control vocal output is provided by Tourville et al (2008), who used fMRI to examine formant-shifted speech during production of monosyllabic words to test the DIVA model predictions of brain areas involved in articulatory control. The DIVA model of speech production (Guenther, et al., 2006; Guenther et al., 1998; Tourville & Guenther, 2010) is another model that incorporates auditory feedback and feed forward commands for voice and speech control (Guenther, et al., 2006). During vocalization the vocal-sensorimotor system provides both kinesthetic and auditory feedback, which is used to compare actual and intended vocal output to regulate voice F0 through error-induced corrective commands. The model specifies that auditory error cells located in the posterior superior temporal gyrus respond when a mismatch between the auditory feedback signal and the auditory target is detected. The projections from the auditory error cells transform the auditory error into a motor command to correct voice F0 to match actual vocal output with intended vocal output. Results from the current study also show increased BOLD in these STG regions when comparing the fixed effects group results of shifted vocalization versus non-shifted vocalization. Phonation is a critical component of speech, however, the control and regulation of the voice may be distinct from the articulator control involved in speech production, and the feedback system required for speech is likely to be more complex than that required for phonation. It is worth noting that the activation reported in this region by Tourville et al (2008) was from a fixed effects analysis as there were no voxels surviving random effects analysis of the *shift versus no-shift* contrast. In using a fixed effects model, findings could be driven by one or two individuals who show particularly strong responses. Using a random effects model in the current study we did not see activation in the prefrontal regions or right hemisphere pre-motor cortex as reported by Tourville et al. (2008). The

right hemisphere pre-motor region is suggested to be a region where corrective commands are sent from the auditory error cells to the motor cells driving the articulators during speech (Tourville, et al., 2008). The lack of activation in these additional regions in the current study is likely due to the random effects analysis employed and these regions may emerge with more subjects. However, differences seen could also be a result of the different tasks used (speech vs. vocalization) suggesting that the motor and premotor regions are more directly related to corrective motor commands for speech sounds. It may be the case that the activation in the motor and premotor regions is similar for the shifted and unshifted vocalization conditions. Using a similar sample size to Tourville et al (2008), we report activations from RFX analysis in the present study bilaterally in the STG where they did not find any significant areas of activation using RFX. We further reiterate that the current results point to the STG region as the primary region involved in using sensory feedback for the online control of vocalization.

As noted in the introduction, several other studies have examined vocal control mechanisms related to pitch control. However, these studies had very different methods than those used here, likely resulting in differences between the activation patterns found in those studies and the ones we report. Zarate & Zatorre (2008) showed activation bilaterally in the premotor cortex when singers were told to ignore shifts in pitch, however, when the command was to compensate for the change in pitch there was also additional activation present bilaterally in the prefrontal cortex. Neither of these previous studies using fMRI to examine perturbations in auditory feedback used comparable paradigms to the present study. Tourville et al (2008) examined the effect of shifting the first formant frequency of speech while subjects spoke monosyllabic words. Zarate and Zatorre (2008) used pitch shifted feedback during vocalization, however the feedback remained shifted for 3 seconds and subjects were asked to either voluntarily compensate for the shift or to ignore the shift. It has been shown that pitch-shift stimuli exceeding 300ms in duration elicit voluntary responses, whereas there is no evidence for voluntary responses with short stimuli of less than 300ms (Burnett, et al., 1998; Hain, et al., 2000). Therefore the responses to these 3 second shifts would likely involve voluntary control mechanisms and would not have been the same as the involuntary reflexive responses that we are measuring in the current study. Here we use short duration frequency-shifted stimuli and show average response peak latencies of around 300ms providing evidence that we were measuring reflexive responses.

We also note an important limitation to the current study associated with findings that show there are two different types of vocal responses to the pitch-shifted stimuli previously reported (Burnett, et al., 1998). One response, termed a compensatory response, is where the pitch shifted auditory feedback evokes a response in the opposite direction to the shift in order to maintain voice F0 at the desired level. The second type is a following response where subjects change their voice F0 in the same direction as the stimulus. This variability in response is still unexplained although it has been shown that with an increase in the magnitude of the stimulus shift there is a decrease in the proportion of compensatory responses (Burnett, et al., 1998). It appears that as the shifted feedback appears less like the participant's own voice, there are more following responses in the same direction as the stimulus. There may also be different neural mechanisms responsible for the different responses. Even though the responses of subjects in this study were on average compensatory (see Figure 2), one limitation in our design is in having 5 shifts per block of trials, which increases the likelihood of having mixed responses within that block (i.e. both following and compensatory responses to the same direction stimuli). That is, in some blocks there could be responses of both types, and it is not possible to distinguish whether there are differences in brain activation between these different responses. In examining the vocal responses of all subjects, $32\% \pm 16\%$ (mean $\pm$ SD) of trials consisted of compensatory responses to all 5 shifts within that particular trial and $51\% \pm 14\%$ of trials contained mixed

responses consisting of a majority being compensatory (3 or 4 compensatory and 1 or 2 following responses). In examining the blocks where all 5 responses are in the same direction, the number of trials and therefore power is greatly reduced and so activation may not reach significance when comparing compensatory responses to following responses. We may expect to see differences in neural activation to compensatory versus following responses and this may provide further insight into the role of efference copy in voice control.

There are several other methodological issues in the present study that have to be considered in the interpretation. One issue is the choice of baseline in the current study. Here we chose to compare normal and altered feedback during phonation rather than including a control condition that involved only listening to the shifted stimuli. We believe that using a condition involving listening to shifted stimuli is important but would likely not serve as a control in the current study due to the differences we would see in bold between the two conditions (upward and downward shifts) related to the act of vocalizing. The baseline BOLD activity related to vocalization would remain similar in the two conditions used here, and we would expect any differences to be related to processing the acoustic differences between the two conditions.

Another issue to address is whether or not the brain activations could be due to changes in voice amplitude. When subjects respond to the feedback shifts by changing pitch, they may also change voice amplitude even though instructed to hold their voice level steady. We examined vocal amplitude in response to the shifted vocalization and there were no significant changes suggesting that voice amplitude is not contributing to the changes seen in the BOLD signal.

A final issue of concern is the possibility that the brain activity related to the shifted feedback may also be a result of changes in attention related to the shift. There are obvious differences in the sensory experience associated with the shift and no-shift conditions, however subjects are not specifically asked to attend to the vocal feedback. There are also no changes in amplitude of the stimuli that may further enhance attention; pitch is the only acoustic difference. Thus, if attention was a factor in this study, it was probably a minor contributor to the BOLD effects that were measured.

In conclusion, using fMRI we were able to demonstrate the relevance of the STG region in audio-vocal responses to pitch shifted vocalization, visible as an increased BOLD signal in these areas in response to vocalization with pitch-shifted feedback compared to non-shifted feedback, thus supporting the involvement of this region in coding mismatches between expected and actual auditory signals. If it is the case that this region is responsible for the coding of a mismatch between actual and predicted voice F0, then the result would support previous feedback models of voice control as suggested by Larson et al. (2008). The reduction in BOLD in the unshifted feedback condition compared to the shifted feedback condition is also in agreement with response enhancement observed in ERP studies with pitch-shifted feedback (Behroozmand et al., 2009). The lack of activation in the motor and premotor regions when comparing shifted to un-shifted vocalization in random effects analysis may be due to a lack of power as a result of using the sparse sampling technique. We provide evidence of similarities in the STG region activated during audio-vocal pitch shift responses to vocalization as those found during speech (Tourville et al., 2008). However, we did not see activation in the laryngeal motor cortex and right hemisphere premotor regions in response to shifted vocalization. It may also be the case that the levels of brain activity in these regions are equal during shifted and un-shifted vocalization or that they are a result of increased engagement of the speech network due to increased complexity and involvement of the articulators in the speech tasks compared to vocalization. A further

explanation for the lack of activation in laryngeal motor cortex is that the vocal responses were quite small (e.g., 10 cents), and responses of this magnitude are associated with very small changes in laryngeal muscle activity (Liu, et al., 2010a). Such small changes imply that cortical activations involved in these responses were also quite small and did not reach threshold for the identification of a BOLD response.

Acknowledgments

This work was supported by National Institute of Health grant 1R01DC006243.

References

- Aliu SO, Houde JF, Nagarajan SS. Motor-induced suppression of the auditory cortex. *Journal of cognitive neuroscience*. 2009; 21(4):791–802. [PubMed: 18593265]
- Angelocci AA, Kopp GA, Holbrook A. The Vowel Formants of Deaf and Normal-Hearing Eleven- to Fourteen-Year-Old Boys. *J Speech Hear Disord*. 1964; 29:156–160. [PubMed: 14147499]
- Bauer JJ, Mittal J, Larson CR, Hain TC. Vocal responses to unanticipated perturbations in voice loudness feedback: an automatic mechanism for stabilizing voice amplitude. *J Acoust Soc Am*. 2006; 119(4):2363–2371. [PubMed: 16642849]
- Behroozmand R, Karvelis L, Liu H, Larson C. Vocalization-induced enhancement of the auditory cortex responsiveness during voice F0 feedback perturbation. *Clinical Neurophysiology*. 2009; 120:1303–1312. [PubMed: 19520602]
- Behroozmand R, Larson CR. Error-dependent modulation of speech-induced auditory suppression for pitch-shifted voice feedback. *BMC Neurosci*. 2011; 12:54. [PubMed: 21645406]
- Behroozmand R, Liu H, Larson CR. Time-dependent neural processing of auditory feedback during voice pitch error detection. *Journal Of Cognitive Neuroscience*. In Press.
- Binnie CA, Daniloff RG, Buckingham HW Jr. Phonetic disintegration in a five-year-old following sudden hearing loss. *J Speech Hear Disord*. 1982; 47(2):181–189. [PubMed: 7176596]
- Blakemore SJ, Frith CD, Wolpert DM. Spatio-temporal prediction modulates the perception of self-produced stimuli. *J Cogn Neurosci*. 1999; 11(5):551–559. [PubMed: 10511643]
- Bohland JW, Guenther FH. An fMRI investigation of syllable sequence production. *Neuroimage*. 2006; 32(2):821–841. [PubMed: 16730195]
- Borden GJ. An interpretation of research of feedback interruption in speech. *Brain and Language*. 1979; 7(3):307–319. [PubMed: 455050]
- Brown S, Ngan E, Liotti M. A larynx area in the human motor cortex. *Cereb Cortex*. 2008; 18(4):837–845. [PubMed: 17652461]
- Burnett TA, Freedland MB, Larson CR, Hain TC. Voice F0 responses to manipulations in pitch feedback. *J Acoust Soc Am*. 1998; 103(6):3153–3161. [PubMed: 9637026]
- Burnett TA, Senner JE, Larson CR. Voice F0 responses to pitch-shifted auditory feedback: a preliminary study. *J Voice*. 1997; 11(2):202–211. [PubMed: 9181544]
- Burns EM, Ward WD. Categorical perception--phenomenon or epiphenomenon: evidence from experiments in the perception of melodic musical intervals. *J Acoust Soc Am*. 1978; 63(2):456–468. [PubMed: 670543]
- Christoffels IK, van de Ven V, Waldorp LJ, Formisano E, Schiller NO. The sensory consequences of speaking: Parametric Neural cancellation during speech in Auditory Cortex. *PLoS ONE*. 2011; 6(5):e18307. [PubMed: 21625532]
- Elliott L, Niemoller A. The role of hearing in controlling voice fundamental frequency. *International Audiology*. 1970; 9:47–52.
- Engelien A, Yang Y, Engelien W, Zonana J, Stern E, Silbersweig DA. Physiological mapping of human auditory cortices with a silent event-related fMRI technique. *Neuroimage*. 2002; 16(4):944–953. [PubMed: 12202082]
- Fox CM, Ramig LO, Ciucci MR, Sapir S, McFarland DH, Farley BG. The science and practice of LSVT/LOUD: neural plasticity-principled approach to treating individuals with Parkinson disease and other neurological disorders. *Semin Speech Lang*. 2006; 27(4):283–299. [PubMed: 17117354]

- Golfopoulos E, Tourville JA, Bohland JW, Ghosh SS, Nieto-Castanon A, Guenther FH. fMRI investigation of unexpected somatosensory feedback perturbation during speech. *Neuroimage*. 2011; 55(3):1324–1338. [PubMed: 21195191]
- Greenlee JD, Jackson AW, Chen F, Larson CR, Oya H, Kawasaki H, et al. Human auditory cortical activation during self-vocalization. *PLoS One*. 2011; 6(3):e14744. [PubMed: 21390228]
- Guenther FH, Ghosh SS, Tourville JA. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang*. 2006; 96(3):280–301. [PubMed: 16040108]
- Guenther FH, Hampson M, Johnson D. A theoretical investigation of reference frames for the planning of speech movements. *Psychol Rev*. 1998; 105(4):611–633. [PubMed: 9830375]
- Hain TC, Burnett TA, Kiran S, Larson CR, Singh S, Kenney MK. Instructing subjects to make a voluntary response reveals the presence of two components to the audio-vocal reflex. *Exp Brain Res*. 2000; 130(2):133–141. [PubMed: 10672466]
- Heinks-Maldonado TH, Houde JF. Compensatory responses to brief perturbations of speech amplitude. *Acoustics Research Letters Online*. 2005; 6(3):131–137.
- Heinks-Maldonado TH, Mathalon DH, Gray M, Ford JM. Fine-tuning of auditory cortex during speech production. *Psychophysiology*. 2005; 42(2):180–190. [PubMed: 15787855]
- Hickok G, Houde J, Rong F. Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron*. 2011; 69(3):407–422. [PubMed: 21315253]
- Houde JF, Jordan MI. Sensorimotor adaptation in speech production. *Science*. 1998; 279:1213–1216. [PubMed: 9469813]
- Houde JF, Jordan MI. Sensorimotor adaptation of speech I: Compensation and adaptation. *J Speech Lang Hear Res*. 2002a; 45(2):295–310. [PubMed: 12003512]
- Houde JF, Jordan MI. Sensorimotor adaptation of speech I: Compensation and adaptation. *Journal of Speech, Language, and Hearing Research*. 2002b; 45(2):295–310.
- Jeannerod M, Paulignan Y, Weiss P. Grasping an object: one movement, several components. *Novartis Found Symp*. 1998; 218:5–16. discussion 16–20. [PubMed: 9949813]
- Lane H, Wozniak J, Matthies M, Svirsky M, Perkell J, O'Connell M, et al. Changes in sound pressure and fundamental frequency contours following changes in hearing status. *J Acoust Soc Am*. 1997; 101(4):2244–2252. [PubMed: 9104026]
- Larson CR. Cross-modality influences in speech motor control: the use of pitch shifting for the study of F0 control. *J Commun Disord*. 1998; 31(6):489–502. quiz 502–483;553. [PubMed: 9836138]
- Larson CR, Altman KW, Liu H, Hain TC. Interactions between auditory and somatosensory feedback for voice F0 control. *Exp Brain Res*. 2008; 187(4):613–621. [PubMed: 18340440]
- Liu H, Behroozmand R, Larson CR. Enhanced neural responses to self-triggered voice pitch feedback perturbations. *Neuroreport*. 2010a; 21(7):527–531. [PubMed: 20386347]
- Liu H, Behroozmand R, Larson CR. Enhanced neural responses to self-triggered voice pitch feedback perturbations. *NeuroReport*. 2010b; 21(7):527–531. [PubMed: 20386347]
- Murbe D, Pabst F, Hofmann G, Sundberg J. Significance of auditory and kinesthetic feedback to singers' pitch control. *J Voice*. 2002; 16(1):44–51. [PubMed: 12002886]
- Paus T, Perry DW, Zatorre RJ, Worsley KJ, Evans AC. Modulation of cerebral blood flow in the human auditory cortex during speech: role of motor-to-sensory discharges. *Eur J Neurosci*. 1996; 8:2236–2246. [PubMed: 8950088]
- Ramig LO, Sapis S, Fox C, Countryman S. Changes in vocal loudness following intensive voice treatment (LSVT) in individuals with Parkinson's disease: a comparison with untreated patients and normal age-matched controls. *Mov Disord*. 2001; 16(1):79–83. [PubMed: 11215597]
- Ringel RL, Steer MD. Some Effects of Tactile and Auditory Alterations on Speech Output. *J Speech Hear Res*. 1963; 13:369–378. [PubMed: 14071874]
- Svirsky MA, Lane H, Perkell JS, Wozniak J. Effects of short-term auditory deprivation on speech production in adult cochlear implant users. *J Acoust Soc Am*. 1992; 92(3):1284–1300. [PubMed: 1401516]
- Tourville JA, Guenther FH. The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*. 2010

- Tourville JA, Reilly KJ, Guenther FH. Neural mechanisms underlying auditory feedback control of speech. *Neuroimage*. 2008; 39(3):1429–1443. [PubMed: 18035557]
- Toyomura A, Koyama S, Miyamaoto T, Terao A, Omori T, Murohashi H, et al. Neural correlates of auditory feedback control in human. *Neuroscience*. 2007; 146(2):499–503. [PubMed: 17395381]
- Turkeltaub PE, Eden GF, Jones KM, Zeffiro TA. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *Neuroimage*. 2002; 16(3 Pt 1):765–780. [PubMed: 12169260]
- Wolpert DM, Ghahramani Z, Jordan MI. An internal model for sensorimotor integration. *Science*. 1995; 269:1880–1882. [PubMed: 7569931]
- Wolpert DM. Computational approaches to motor control. *Trends Cogn Sci*. 1997; 1(6):209–216. [PubMed: 21223909]
- Xu Y, Larson CR, Bauer JJ, Hain TC. Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *J Acoust Soc Am*. 2004; 116(2):1168–1178. [PubMed: 15376682]
- Yang Y, Engelien A, Engelien W, Xu S, Stern E, Silbersweig DA. A silent event-related functional MRI technique for brain activation studies without interference of scanner acoustic noise. *Magn Reson Med*. 2000; 43(2):185–190. [PubMed: 10680681]
- Zarate JM, Wood S, Zatorre RJ. Neural networks involved in voluntary and involuntary vocal pitch regulation in experienced singers. *Neuropsychologia*. 2010; 48(2):607–618. [PubMed: 19896958]
- Zarate JM, Zatorre RJ. Neural substrates governing audio-voal integration for vocal pitch regulation in singing. *Annals of the New York Academy of Sciences*. 2005; 1060:404–408. [PubMed: 16597793]
- Zarate JM, Zatorre RJ. Experience-dependent neural substrates involved in vocal pitch regulation during singing. *Neuroimage*. 2008; 40(4):1871–1887. [PubMed: 18343163]

Figure 1.
Experimental design for a single trial of volume acquisition and expected HRF

Figure 2. Group average vocal responses to downwards (a) and upwards (b) pitch shifted stimuli. The pitch shift occurred at time $t=0$. Figure shows mean responses in cents with standard error bars ($n=11$).

Figure 3. BOLD response for a comparison of unshifted vocalization condition compared to the baseline rest condition (NoShift-Rest). (b) Bold responses for the shifted vocalization condition compared to the baseline rest condition (Shift-Rest). Data shown represents group responses (n=12) from a random effects analysis.

Figure 4.
(a) Fixed effects ($p<0.05$, FWE) and (b) random effects ($p<0.001$) analysis of BOLD responses to the vocalization pitch shift – no shift contrast ($n=12$).

Table 1

Peak voxel responses for the contrasts of interest reported in MNI space.

Region Label	Shifted Vocalization				Unshifted Vocalization				Shift>NoShift			
	MNI Voxel location				MNI Voxel location				MNI Voxel location			
	x	y	z	T	x	y	z	T	x	y	z	T
Right Superior Temporal Gyrus	60	-12	-1	7.41	61	-5	5	5.06	63	-17	9	2.82
	66	-17	13	6.84	60	-12	-1	4.67	66	-19	13	2.81
	68	-23	3	5.79					66	-29	9	2.47
									51	-28	9	2.54
Left Superior Temporal Gyrus	-56	-2	-1	6.48	-61	-14	9	7.75	-57	-29	9	3.64
	-61	-16	6	6.05	-61	-7	6	7.02				
	-56	-24	9	5.59	-59	-5	9	7.02				
Left Middle Temporal Gyrus	-66	-31	9	5.36					-52	-28	6	4.03
Right Middle Temporal Gyrus	63	-29	3	5.57								
Supplementary Motor Area	-1	-1	65	3.75	1	0	66	3.80				
Left Precentral Gyrus	-47	-11	43	8.22	-49	-7	32	4.95				
	-21	-29	62	5.36								
Right precentral gyrus	50	-9	32	8.43	50	-10	39	4.65				
					15	-29	66	3.82				
Left Postcentral Gyrus	-47	-14	42	4.88	-52	-7	21	5.60				
Left Inferior Parietal Lobe	-35	-55	59	5.52								
Left Inferior Frontal Gyrus	-36	34	7	3.98	-33	15	-10	3.63				
Right Inferior Frontal Gyrus	43	23	11	3.54	49	14	9	4.33				
Left Insula	-32	24	9	3.45	-38	5	0	7.45				
Right Insula	37	15	19	3.66	41	19	6	7.12				
Right Putamen	22	6	6	3.59	22	3	6	5.52				
Left Putamen	-23	8	6	4.15	-23	-4	6	4.47				
Left Thalamus	-10	-18	1	3.51								