Auditory Information in Speech Production
Sensorimotor control of speech production

• Motor

• Speech-relevant areas of motor cortex (M1) make direct connections with the motor neurons of the lips, tongue, and other speech articulators (Jürgens et al., 1982; Jürgens, 2002; Ludlow, 2004).

• Damage to these M1 areas causes mutism and dysarthria (Jürgens, 2002; Duffy, 2005)

• Sensory Feedback

• types
  • auditory
  • somatosensory

• Important role in children (particularly auditory)

• Unclear in adults (post-lingually deafened adults remain intelligible for years)

• only fricatives, f0 affected
Evidence for role of feedback

- Delayed auditory feedback (DAF) disrupts speech (~200ms)

- Compensation to altered auditory feedback
  - vowel formant frequencies
  - f0
  - centroid of acoustic energy in fricatives
Figure 1. Apparatus and behavior. (A) Diagram of the pitch perturbation apparatus. A DSP shifted the pitch of subjects’ vocalizations (red line) and delivered this auditory feedback (blue line) to subjects’ earphones. (B) Spectrogram (Upper) and pitch track (Lower) of an example trial with pitch perturbation applied. (C) Histogram of compensatory responses as a percentage of pitch shift. The green arrow denotes the trial shown in B.

Human cortical sensorimotor network underlying feedback control of vocal pitch

Edward F. Chang,a,1,2 Caroline A. Niziolek,b,1 Robert T. Knight,a Srikantan S. Nagarajan,c and John F. Houde,b,2

Compensatory Responses

- Compensatory response begins ~150 ms after perturbation onset.

- Compensation is partial
  - Magnitude varies across subjects (e.g. -25 to 60%)
  - Magnitude varies across trials

- Why not complete compensation?
  - Conflict between somatosensory and auditory feedback.
Difficulties with feedback control in biological systems

- Sensory feedback is noisy (real-world noise)
- Feedback is delayed
  - synaptic times
- Relevant measures (formants, f0) need to be computed in windows of 30-100ms
- ~100 ms response latency of neurons in a recently discovered area of pitch-sensitive neurons in auditory cortex (Bendor and Wang, 2005)
Dynamical Feedback Model

**FIGURE 2 | The control problem in speech motor control.** The figure shows a snapshot at time $t$, when the vocal tract has produced output $y_t$ in response to the previously applied control $u_{t-1}$.

**Speech production as state feedback control**

*John F. Houde* and *Srikantan S. Nagarajan*
FIGURE 3 | Ideal state feedback control. If the controller in the CNS had access to the full internal state $x_t$ of the vocal tract system (red path), it could ignore feedback $y_{t-N}$ and formulate a state feedback control law $U_t(x_t)$ that would optimally guide the vocal tract articulators to produce the desired speech output $y_t$. However, as discussed in the text, the internal vocal tract state $x_t$ is, by definition, not directly available.
Internal (““Forward”)” model

The SFC model can also be viewed as a type of Kalman filter, with a prediction/correction process in which, in the prediction (green) direction, efference copy of vocal motor commands are passed through a Smith predictor that will be applied at the next timestep to the vocal tract. There are a number of lines of evidence supporting the idea that the Kalman gain function attenuates the influence of feedback prediction errors on correcting the current state estimate. When prediction error is so optimally determined, the combination of feedforward controls (e.g., reaching, head, and eye movement control; Do rov and Jordan, 2002; Do rov and Jordan, 2002) and the observer is leading down toward the temporal pole (Rauschecker and Tian, 2011), which in turn sensory processing. First, even in other primates, there appear to be ventral stream serving speech comprehension and a dorsal stream serving speech production as state feedback control. In particular, a variety of studies have implicated the ventral parietal temporal area (Vpta; Hickok et al., 2007; Rauschecker and Scott, 2009; Hickok et al., 2011) and the superior parietal temporal area (Spt; Hickok and Poeppel, 2009) as areas that are involved in object recognition (Hickok et al., 2011). Subsequently, studies of models of reaching motor control have been advanced to explain how people optimize their movements (K. Roweis and Ghahramani, 1999; Houde and Nagarajan, 2012), while, in humans, the dorsal stream lesions, respectively; Smith, 1959; Miall et al., 2008; Shadmehr and Krakauer, 2008). This is accomplished by activating neurons that are sensitive to auditory source type were found in a ventral pathway (Hickok and Poeppel, 2009), while the dorsal “where” stream leading down toward the temporal pole (Rauschecker and Tian, 2011) is concerned with object location, and a position-sensitive area (Hickok et al., 2007) that can be advanced for the visual system, with a dorsal “where” stream leading to parietal cortex that is concerned with object location, and a position-sensitive area (Hickok et al., 2007). Since then, the problem of state feedback control has been addressed in a number of lines of evidence supporting the idea that the Kalman gain function attenuates the influence of feedback prediction errors on correcting the current state estimate.
Solution: forward model

- Efference copy (aka “corollary discharge”)
  - encodes the predicted sensory consequences of motor commands
  - suppression of neural response to the resulting sensory stimulation through a comparison with, or subtraction of, the predicted feedback (Von Holst & Mittelstaedt, 1950).

- Originally meant to explain:
  - When the eye moves, the image on the retina shifts. Why doesn’t the world look like it shifts.

- Also: Why you can’t tickle yourself (Blakemore, 1998).
Evidence from speech

Speech-induced suppression:
Response from some areas in auditory cortex (posterior superior temporal cortex) is reduced while speaking, compared to just listening to the same audio signal.

Sensorimotor adaptation:
Sustained alteration of feedback can result in adaptation of speech that is maintained after the alteration ends. (Retuning of forward model)
Dual-stream model (Hickok & Poeppel, 2007)

• Two streams of speech processing
  • ventral - comprehension
  • dorsal - sensorimotor integration and feedback
• Evidence
  • conduction aphasia (spt)
  • SIS in the superior posterior temporal area
Could also be active during passive listening to speech, and indeed, studies with sensory areas suggest the possibility that premotor cortex may play an intermediary role in speech production? First, reciprocal connections (blue dashed arrows in Figure 6) between premotor cortex and other premotor, frontal cortices, respectively. In this way, the key parts of the SFC model are well connected to motor cortex (green) and correction (red) processes running between motor and sensory cortices. The premotor cortices are ideally placed for such an intermediary role: premotor cortex is both bidirectionally well connected to sensory cortices, and Rilling et al., 2008; Upadhyay et al., 2008, has been found to be active during passive listening to speech, and indeed, Kandel et al., 2000, has postulated that premotor cortex was functionally connected to these auditory areas during listening (Watkins and Paus, 2004). On the auditory end, auditory areas were also activated more for speech sounds rated least producible, and that svPMC was functionally connected to these auditory areas during listening to and speaking meaningless syllables, but not listening to altered feedback versus non-altered trials, To urville et al. (2004), found the superior ventral premotor area (svPMC), bilaterally was activated by both auditory and somatosensory feedback by auditory and somatosensory Kalman gain functions \( K_t(y) \). The feedback prediction errors \( \hat{y}_{t-N} \) and \( \hat{y}_{t-N} \) generated in auditory and somatosensory cortex are converted into separate state corrections \( \hat{e}_t \), based on auditory and somatosensory feedback by auditory and somatosensory Kalman gain functions \( K_t(y) \), in high order auditory and somatosensory cortices, respectively. The auditory- and somatosensory-based state corrections are then added to \( \hat{x}_{t-1} \) in premotor cortex to make next state estimate \( \hat{x}_t \). Finally, the key operations depicted in blue are all postulated to be modulated by the current speech task goals (e.g., what speech sound is currently meant to be produced) that are expressed in other areas of frontal cortex.

FIGURE 6 | State feedback control (SFC) model of speech motor control with putative neural substrate. The figure depicts the same operations as those shown in Figure 5, but with suggested cortical locations of the operations (motor areas are in yellow, while sensory areas are in pink). The current model is largely agnostic regarding hemispheric specialization for these operations. Also, for diagrammatic simplicity, the operations in the auditory and somatosensory cortices are depicted in the single area marked “sensory cortex,” with the understanding that it represents analogous operations occurring in both of these sensory cortices: i.e., the delayed state estimate \( \hat{x}_{t-1} \) is sent to both high order somatosensory and auditory cortex, each with separate feedback prediction modules (\( \text{vout}(\hat{x}) \)) for predicting auditory feedback in high order auditory cortex and \( \text{vout}(\hat{x}) \) for predicting somatosensory feedback in high order somatosensory cortex. The feedback prediction errors \( \hat{y}_{t-N} \) generated in auditory and somatosensory cortex are converted into separate state corrections \( \hat{e}_t \), based on auditory and somatosensory feedback by auditory and somatosensory Kalman gain functions \( K_t(y) \), in high order auditory and somatosensory cortices, respectively. The auditory- and somatosensory-based state corrections are then added to \( \hat{x}_{t-1} \) in premotor cortex to make next state estimate \( \hat{x}_t \). Finally, the key operations depicted in blue are all postulated to be modulated by the current speech task goals (e.g., what speech sound is currently meant to be produced) that are expressed in other areas of frontal cortex.
Neural responses and compensation

- What is neural response to altered feedback?
- Failure of efference copy when feedback is altered
- Does magnitude of SIS predict compensation?
- speech perturbation-response enhancement (SPRE).

Human cortical sensorimotor network underlying feedback control of vocal pitch
Edward F. Chang,a,1,2 Caroline A. Niziolek,b,1 Robert T. Knight,a Srikantan S. Nagarajan,c and John F. Houde,b,2

Locations exhibit:

- SIS
- SPRE
Chang et al. close to auditory sites exhibiting SIS and SPRE. Fig. 3 compensation and activity was weaker in other frequency bands, activity for the same four electrodes. (Correlations of neural subject (GP35). Asterisks denote statistical signi marks the time of peak compensation. (***(P < 0.001). (**(P < 0.01); **(P < 0.05); *(P < 0.2). Across subjects, the same pattern holds across temporal grids: SPRE electrodes showed stronger behavioral correlations electrodes in four left-hemisphere and three right-hemisphere corrected motor signal. suggesting that SPRE, and not SIS, is a marker for in more, the degree of enhancement (SPRE) for an electrode was significantly correlated electrodes (circled) and SPRE and used in the analysis in (s). Per-trial correlations for the same electrode (red; opacity denotes degree of SPRE). The white box contains related electrodes (white circles) and that of the SPRE electrodes parietal junction, with additional SPRE responses found along the anterior extent of the superior temporal cortex, the posterior superior temporal cortex, and premotor cortex and in the left temporal and ventral premotor areas in these subjects. For this reason, we have focused subsequent analyses on the four subjects perturbation during speaking, the more that electrode correlated enhanced response in the speak condition (SPRE; red). The SPRE enhancement from speaking (blue), and electrodes with an en- tempsales for the left (e21, e22, e23) and right (e45) hemispheres are shown on the brain map. The solid black line is the best-
appears to be positioned in the sulci are the result of surface coregistration (Pearson SIS and SPRE in any given electrode were not significant; in left-hemisphere grids: directional modulation of auditory cortex: suppression during listening to recordings of the same auditory stimuli. Consistent with comparing neural responses during speech with those evoked by the comparison between efference copy and external feedback also may distinguish between sensations that come from the speaker and theorized that such suppression affords a mechanism to dis- or subtraction of, the predicted feedback (34, 35). It has been suggested separable mechanisms for suppression of predicted and mismatches between intended and observed sensory outcomes. In many current models of motor control, a forward model underlies the corrective adjustment of output pitch and con-
What does efferent copy represent?

1. Predicted sensory consequences of actual motor commands

2. Predicted sensory consequences of selection of a particular motor act (whose motor commands can vary stochastically)

If (2) magnitude of SIS should vary as a function of prototypicality of particular produced token.

What Does Motor Efference Copy Represent? Evidence from Speech Production
Caroline A. Niziolek,¹ Srikantan S. Nagarajan,² and John F. Houde¹
What Does Motor Efference Copy Represent? Evidence from Speech Production

Caroline A. Niziolek,¹ Srikantan S. Nagarajan,² and John F. Houde¹

The Journal of Neuroscience, October 9, 2013 • 33(41):16110–16116

stimuli: “eat”, “Ed”, “add”

MEG: weights were extracted for the peak voxel in each hemisphere, determined by activity strength in a window around the M100 response (50–150 ms after stimulus onset)
A spread of formant values is apparent across multiple productions, with those closest to each vowel’s median colored in green (“center”) and those farthest colored in dark red (“periphery”).

Centrally versus peripherally produced vowel trials in a sample subject. Each point represents an individual vowel production.
Peripheral tokens become more centered later in the recording (first 50 ms vs. middle 50%).

Degree of centering correlates inversely with the magnitude of the SIS.

Prediction error correlates with centering.
Conclusions

• Sensory predictions do not accurately track feedback variability across repetitions of the same motor task.

• Efference copy may reflect a motor plan, not outgoing motor commands, and that is generated up-stream of primary motor cortex.

• Predictions reflect a desired sensory target.

• Peripheral tokens are like “errors”

• Suppression may reflect an error-detection/correction process.

• Reduced suppression causes more correction
On sensory targets as efferent copies

- What is “sensory target” for pitch (in a non-tone language)?

- In an experiment with altered formant frequencies, would SI-S regions predict compensation (as opposed to pitch, where it does not)?
On correction

• Niozelek et al. view the peripheral tokens as like “errors” that get corrected (centered)

• But the variation could just be temporal.
  • Articulators form constriction over time.
  • On some tokens, phonation begins before the vowel has arrived at its target.
  • It doesn’t get corrected, just gets to its target in a feedforward way.
Evidence for coupling and entrainment in speech: Speech Errors

- Speech errors have been traditionally analyzed as ‘rearrangement’ of symbol sequence in planning.

- Examine gestures during repetitive tasks that elicit speech errors.
  - EMA (Pouplier)
  - real-time MRI (Byrd, Narayanan, Proctor, Lammert)

- Most common errors are gestural intrusion errors.

“top cop top cop...”
Gesture Intrusions

**Overlaid frames**

- “Extra” copy of a gesture appears at an inappropriate temporal location.
- Usually **co-produced** with the gesture that is appropriate at that time (reductions occur but much less frequently).
- Vary continuously in magnitude and differ in perceptibility (Pouplier & Goldstein, 2006)
  - small perceived as “normal
  - large perceived as substitutions (t->k)
Why do intrusions occur?

- Under the conditions of repetition, individual constriction task variables (LL, TT, TB) function as coupled oscillators.
- The oscillators entrain.
- Under control of rate (and/or other parameters), they can exhibit a shift to a different mode.
- What are the relevant modes?
Frequency-locking modes

- When performing oscillatory motions of multiple body parts
  - Simple rhythms (e.g., 1:1, 2:1) can be performed without learning or practice.
  - Complex multi-frequency rhythms (e.g., 4:3, 5:2) can be learned.
- Spontaneous transitions are observed from

Skilled drummers begin by bimanual tapping with a 5:2 frequency-locking and gradually increase movement frequency.

Frequency-locking mode

- In “cop top”, TT and Lips are frequency-locked in 1:2 relation.
- TT intrusion errors can be viewed as a fluctuation enroute to transition to a simpler mode of frequency-locking (1:1) between TT and Lip constrictors.
Alternative hypothesis

• During an intrusion, the intended gesture is actually a correction.

• Test by looking at timing of the intended and the error gesture.

• If the intended is a correction, it should follow intrusion.
Measurement of lag (Pouplier & Goldstein, 2010)

Intrusions

Upper Lip (UL)
Tongue Tip (TT)
Lower Lip (LL)
Tongue Dorsum (TD)
Tongue Body (TB)
the point of minimum velocity before the constriction movement begins. The other kinematic events of interest were then determined using a 20% threshold of the relevant local velocity as follows. GONS was defined as the time-point at which the velocity exceeded 20% of the velocity range from the minimum velocity preceding the movement to the maximum velocity during constriction formation. TONS was defined as the point in time at which the velocity fell below 20% of the range between the maximum formation velocity and the minimum velocity during the target plateau (MAX). At TOFFS, the velocity exceeded the 20% of the velocity range between the velocity at MAX and the maximum release velocity. These kinematic events are exemplified in Figure 1 for the tongue tip constriction of an intended /t/. For identifying the labial coda consonant /p/, the variable lip aperture was calculated as the Euclidean distance between the upper and
Timing of Intrusions

- At onsets, there is zero lag, no preference for one or another to precede.
- Intended can’t be a correction.
- Synchrony predicted by coupling analysis.
- But intrusion are end sooner, so that at release, there is no audible error (or little).
- Possible suppression of intrusion based on somatosensory feedback.
Any potential role of consonant type in contributing to the change in lag values over the kinematic events will be considered in a later section.

Figure 4 shows a histogram of the lag values across all subjects and tokens. In order to evaluate whether there is an asymmetry in lag values for GONS as predicted by a monitoring account, we evaluated the number of positive and negative lag values for GONS. There are, across subjects, 541 tokens which show a positive lag value, 29 tokens with a lag value of exactly zero and 525 tokens with a negative lag value. A sign test was not significant \( (p = .646) \). Negative lag values would be expected to predominate on the monitoring account, yet there is no evidence for such an asymmetry in our data. Looking at the number of positive and negative GONS lag values on a by-subject basis, no coherent picture emerges: For each subject, the sign test is significant at the .01 level, but, as can be seen from Table 2, for two subjects there are significantly more positive lag values, while for the other two subjects there are significantly more negative lag values (see also Figure A1 in the Appendix). This between-subject variability in the direction of the

Figure 4. Histograms of GONS, TONS, MAX, and TOFFS lag values (ms) across all tokens and subjects.