Simulating a state feedback model of speaking

Article · May 2014

CITATIONS
8

READS
456

5 authors, including:

John F Houde
University of California, San Francisco
87 PUBLICATIONS **3,569** CITATIONS
SEE PROFILE

Caroline A Niziolek
University of Wisconsin–Madison
33 PUBLICATIONS **357** CITATIONS
SEE PROFILE

Zarinah K Agnew
University of California, San Francisco
31 PUBLICATIONS **754** CITATIONS
SEE PROFILE

Srikantan S Nagarajan
University of California, San Francisco
373 PUBLICATIONS **16,832** CITATIONS
SEE PROFILE

Simulating a state feedback model of speaking

John F. Houde[1], Caroline Niziolek[1], Naomi Kort[2], Zarinah Agnew[1], and Srikantan S. Nagarajan[2]

[1]*Dept. of Otolaryngology – Head and Neck Surgery, University of California San Francisco*
[2]*Dept. of Radiology & Biomedical Imaging, University of California San Francisco*

houde@phy.ucsf.edu, cniziolek@ohns.ucsf.edu, naomi.kort@ucsf.edu, zarinah.agnew@ucsf.edu, sri@ucsf.edu

Abstract

An important part of understanding the neural control of speaking is determining how sensory feedback is processed. The role of sensory feedback in speaking suggests a paradox: it need not be present for intelligible speech production, but if it is present, it needs to be correct or speech output will be affected. For this reason, current models of speech motor control relegate sensory feedback to a more indirect role, with an inner feedback loop within the CNS that directly controls speech output, and a slower outer feedback loop where the possibly delayed and intermittent sensory feedback updates the internal feedback loop. Such models can be described as variations on a more general theory of state feedback control (SFC). Here we show, via numerical simulations, how the SFC model can account not only for what is known about the behavioral role of sensory feedback in speaking, but also many of our recent findings about neural responses to auditory feedback.

Keywords: speech motor control, sensory feedback, numerical simulation

1. Introduction

The paradoxical role of sensory feedback in speaking is that it is not necessary for intelligible production, but if it is present, it needs to be correct or speech will be affected (Houde & Nagarajan, 2011). For these reasons, current models of speech motor control relegate sensory feedback to a more indirect role, with an inner feedback loop within the CNS that directly controls speech output, and actual sensory feedback (both auditory and somatosensory feedback) forming slower, possibly delayed and intermittent, external loops that update the internal feedback loop (Guenther & Vladusich, 2012; Price, Crinion, & Macsweeney, 2011; Tian & Poeppel, 2010). Such models can be described as variations on the general theory of state feedback control (SFC), developed in the domain of modern control engineering theory (Houde & Nagarajan, 2011). SFC models have become more prevalent in many domains of motor control research, and we have previously described the hypothesized applicability of SFC to modeling speech motor control (Houde & Nagarajan, 2011).

Here, we construct a numerical simulation of an SFC model and show how it accounts not only for behavioral phenomena associated with the roles of sensory feedback in speaking, but also several of our past experimental findings concerning the neural phenomena involved in auditory feedback processing during phonation.

2. Structure of the SFC model

To facilitate comparison with our experiments, we illustrate our model by focusing on the production of pitch, although our model can easily be generalized to other aspects of speech production.

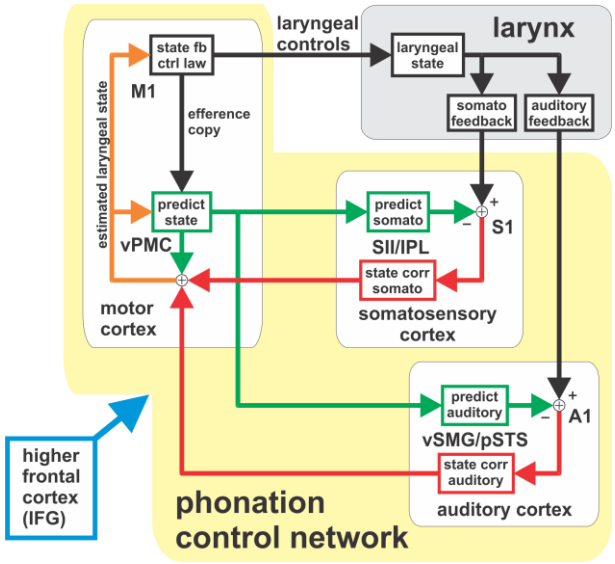

Figure 1: SFC model of how the CNS controls phonation. See text for description.

In this model, production of an utterance involving phonation begins in the CNS with higher frontal cortex (IFG) activating a *phonation control network* (**blue arrow** in Figure 1). This network controls phonation via *state feedback control* (SFC): During phonation, vPMC maintains a running estimate of the current dynamic state of the larynx (i.e., the estimated laryngeal state; **orange** in Figure 1); this state carries information about current sub-glottal pressure, vocal fold position, tension, and any other parameter the network has learned is important to monitor for achieving phonation. M1 generates laryngeal controls based on this state estimate, using a state feedback control law (**state fb ctrl law** in Figure 1) that keeps the larynx tracking a desired state (e.g., one that maintains a desired pitch). While the larynx responds to these controls, vPMC uses a copy of these controls ("efference copy") to predict the next laryngeal state. It feeds this prediction forward (**green arrows**) to the higher sensory areas (SII/IPL in somatosensory feedback and vSMG/pSTSfor auditory feedback) that use it to predict sensory feedback. At the primary sensory cortices (S1 in somatosensory cortex and A1 in auditory cortex), feedback from the larynx is compared with the feedback predictions, resulting feedback prediction errors. The higher sensory areas convert these feedback predictions into state corrections (**state corr somato**, **state corr auditory** in Figure 1) that are fed back (**red arrows**) to vPMC and added to the original state prediction, resulting in a refined estimate of the next laryngeal state (**orange**). This in turn is fed back to M1 to generate the next laryngeal controls.

3. A simulation of an SFC model of pitch control

To verify and illustrate the claims we have made about the SFC model, we have developed a simulation of SFC-based control of speech. For simplicity, the model controls a one-dimensional "speech output" which we have likened to pitch. However, it is straightforward to extend the simulation to control more realistic, multi-dimensional speech output (e.g. loudness, pitch, formants, frication). The simulation was implemented in Matlab (The Mathworks, Inc., Natick, MA), and consists of two parts: the "larynx" and the "phonation control network".

3.1. The "larynx"

The "larynx" to be controlled is modeled as a single damped spring-mass system with a variable rest length of the spring. Position of the mass of this system is taken to be the current pitch output of the vocal tract. This model is based on the idealization (admittedly incomplete) that vocal fold length (i.e., position of the mass) determines pitch, and that the muscles controlling vocal fold length (e.g. the cricothyroid muscle (Titze, Jiang, & Drucker, 1988)) can be modeled as damped spring-mass systems with variable spring rest length (Hill, 1925). This model is not intended to simulate the rich range of laryngeal behaviors captured in multidimensional models (e.g. (Story & Titze, 1995)), but rather to act as a system with dynamics that the controller (i.e., the "phonation control network") must contend with to control pitch. Rest length of the muscle controlling pitch is, in turn, controlled by "brainstem/spinal cord" lower motor system that in turn integrates descending cortical control into a constantly updated rest length of the muscle (Shalit, Zinger, Joshua, & Prut, 2012). In this way, the simulation assumes that cortical motor output codes only desired changes in the current pitch output.

Based on the findings of prior motor control studies, the descending cortical control signal is also subject to "signal dependent noise" (Harris & Wolpert, 1998). This means that the control signal actually seen by the lower motor system is the cortical motor output plus noise that scales with the magnitude of the cortical motor output.

This simulated "larynx" produces two types of sensory "feedback". First, an "auditory" output is idealized as linear conversion to Hz from muscle position to a reasonable value for a speaking pitch, and is assumed to be corrupted by additive white Gaussian "observation" noise that has a feedback delay of 150 ms (Houde & Nagarajan, 2011). Second, a "somatosensory" output is included, reflecting the current position of the mass of the muscle spring/mass system also corrupted by white Gaussian noise, with a feedback delay of 15 ms

3.2. The "phonation control network"

The simulated "phonation control network" for controlling pitch is made up of two parts: (1) an *observer*: a system that estimates the current state of the larynx and (2) a *state feedback control law* that uses the state estimate to generate controls of the vocal tract. Most of the phonation control network is engaged in implementing the observer via interaction of feedforward predictions and feedback corrections (i.e., the green and red arrows in Figure 1). Here, we focus on the observer system. Although in principle, we could use optimal control theory to implement the state feedback control law (Houde & Nagarajan, 2011), for simplicity here, we implement a more rudimentary servo control law where the forward model is first used to estimate the current, undelayed auditory output of the larynx from the current state estimate. This estimated current output is then compared with the current desired output, with the difference passed through a control gain G to generate control applied to the laryngeal simulation on the next time step.

3.2.1. The Kalman Filter Based Observer

The observer estimates state via a recurrent prediction-correction process where a prediction of next state is used to generate sensory predictions that are compared with incoming feedback. The resulting feedback prediction errors are converted by observer gains (**state corr somato**, **state corr auditory** in Figure 1) into corrections of the state prediction. When these gains are computed optimally (i.e., based on the noise characteristics of the sensory feedback), the observer is referred to as a Kalman filter (Houde & Nagarajan, 2011). We therefore implemented the observer as a Kalman filter, reflecting the assumption that the CNS would also seek optimal values for the observer gains.

The heart of the Kalman filter observer simulation is a forward predictive model of the current state of the lower motor/muscle "vocal tract". For the simulations, the forward model is simply a copy of the parameters of the state space model used to simulate the vocal tract. Our simulations here do not include the process of learning all parameters of the forward model, and instead concentrate on how the speech motor system behaves after some parameters of the forward model have been learned. In particular, the model estimates the feedback delay and the covariances of the state and observation noise that determine the Kalman gain, by cross-correlating auditory feedback with somatosensory feedback.

The predicted feedback output of the forward model is delayed by the estimated feedback delay before being compared with the incoming feedback. The resulting delayed feedback prediction error is multiplied by a Kalman gain function to compute a correction to the state prediction of the forward model. We approximated this gain by first computing the steady-state Kalman gain assuming zero feedback delay (Houde & Nagarajan, 2011), then computing the effect of this gain after it has been propagated through the forward model N time steps, where N is the estimated feedback delay. Ultimately, this way of calculating the Kalman gain quantifies the intuition that a feedback prediction error from N time steps in the past (the feedback delay) becomes less and less informative about the current laryngeal state as N increases.

4. Results

We simulated two different auditory feedback experiments we have previously conducted. Both of these experiments contrast responses to auditory feedback from the subject's ongoing speech (the speaking condition) with passive re-listening to playback of auditory feedback from the speaking condition (the listening condtion). To facilitate comparison of simulation results with experimental data, our simulation includes a simplified model of evoked response potential (ERP) generation (David, Harrison, & Friston, 2005).

4.1. Speech Onset

Figure 2 shows one trial from a simulated 100-trial speech onset experiment. The top panel shows the behavioral outputs of the simulation, showing that onset of a 120 Hz phonation target (yt) results in a pitch output (y) that initially slightly deviates from (undershoots) the target. This undershoot is due

to signal-dependent noise added to the large initial laryngeal control.

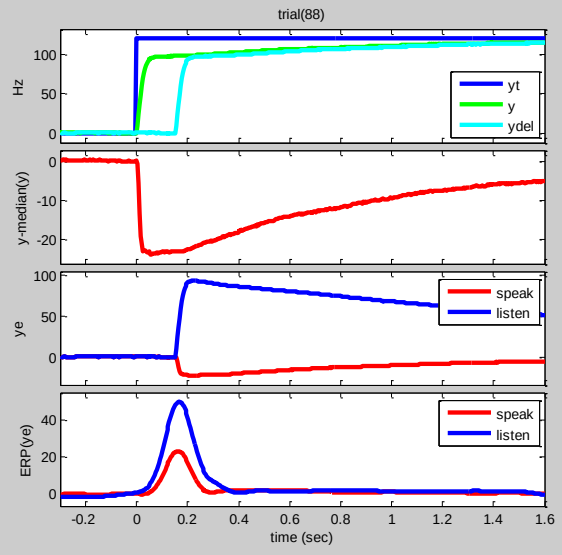

Figure 2: single trial from simulated speech onset experiment. 1st (top) panel: yt: target pitch, y: output pitch, ydel: auditory feedback received by phonation control network, with 150ms auditory processing delay. 2nd panel: deviation of this single trial from the median across trials, showing initial undershoot and subsequent "centering". 3rd panel: feedback prediction error (ye) in the speak and listen conditions of this trial. 4th panel: ERPs of the feedback prediction errors (ERP(ye)), showing SIS as the ERP difference between the speak and listen conditions.

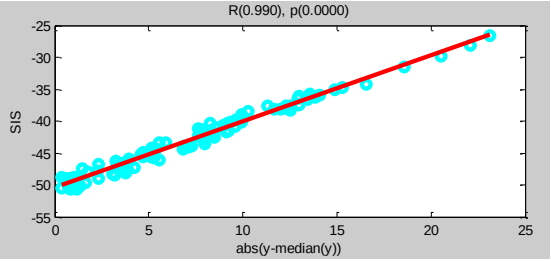

Figure 3: The SIS falloff effect. Scatterplot of the relation between SIS (y-axis) and initial deviation from median output pitch in the simulated speech onset experiment

This undershoot generates both auditory and somatosensory feedback prediction errors that, via their Kalman gains, generate state estimate corrections resulting in small, corrective laryngeal controls that counteract the initial undershoot as the utterance continues. The 2nd panel shows how this initial undershoot and subsequent correction can also be seen by measuring how much pitch output for a single trial deviates from the median pitch output over all trials. Such analysis avoids explicit reference to the pitch target and duplicates similar analysis done in our experimental studies (Niziolek, Nagarajan, & Houde, 2013), where we refer to the subsequent correction as "centering". The 3rd panel shows auditory feedback prediction errors (ye) for both the speaking condition (red), compared to the listening condition (blue), while the 4th panel shows the simulated ERPs corresponding to these prediction errors. The prediction error in the listen condition is very large because the speech onset is not predicted when passively listening to an external speech source, whereas the prediction error in the speak condition is smaller because a speaker is able to predict his/her own speech onset, via efference copy of his/her own laryngeal controls. Thus, the ERP in the speak condition is smaller than the ERP in the listen condition. This replicates the speaking-induced suppression (SIS) effect we commonly see in speech onset experiments (Kort, Nagarajan, & Houde, 2014).

In the speak condition, the size of the ERP is related to the unpredicted deviation of auditory feedback (ydel) from the target pitch. Figure 3 shows in a scatterplot across all trials that this means that initial deviation from the across-trial median (a surrogate for the target pitch) is closely related to the size of the speak – listen ERP (SIS) difference, which replicates the "SIS falloff" effect we have recently documented experimentally (Niziolek et al., 2013).

4.2. Speech Feedback Perturbation

Figure 4 shows one trial from a simulated 100-trial auditory feedback perturbation experiment. The top panel shows the behavioral outputs of the simulation, showing (in light blue) the auditory feedback, perturbed for 400 msec by a 100 cent (one semitone) shift down in pitch, and (in green) the effect this has on output pitch: in this trial, it induces 65% compensation.

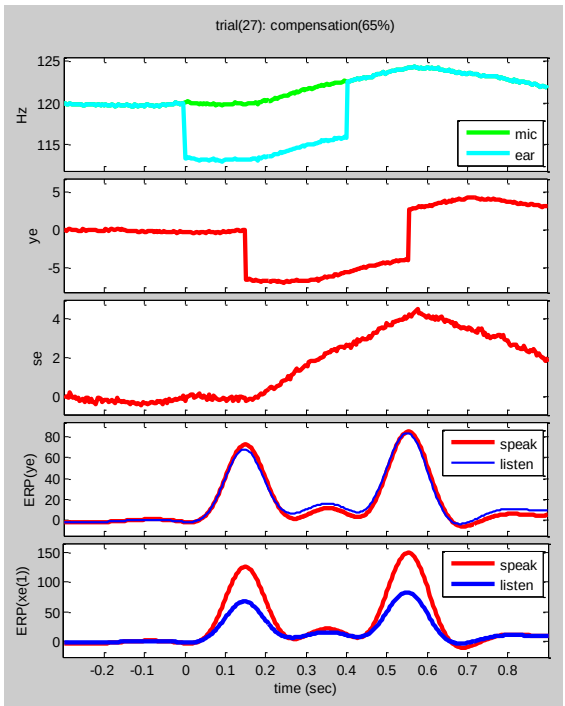

Figure 4: Single trial from the simulated auditory feedback perturbation experiment. 1st (top) panel: a -100cent pitch perturbation is introduced between the model output (mic, for "microphone") and the model's auditory input (ear for "earphone"). 2nd panel: the auditory feedback prediction error (ye) generated by the feedback perturbation. 3rd panel: the somatosensory prediction error created by the compensatory response. 4th panel: equal ERPs from feedback prediction errors (ye) in the speak and listen conditions. 5th panel: enhanced ERPs from state corrections (xe(1)) in the speak condition, compared with the listen condition

Several factors influence how much compensation will be expressed on each trial. First, compensation for the auditory feedback perturbation is moderated by conflicting information conveyed by somatosensory feedback, which remains unaltered.

The next panels of Figure 4 show these conflicting influences on compensation. The 2nd panel shows that the auditory feedback perturbation creates an auditory feedback prediction error that is then reduced by the compensatory response. But the 3rd panel shows the compensatory response itself then creates a somatosensory feedback prediction error that results

an opposing influence on the realized compensation. The strength of this opposing influence is regulated by the Kalman gain on somatosensory feedback prediction errors, which in turn is determined by the estimated somatosensory observation noise – i.e., the estimated reliability of somatosensory feedback. In the simulation shown here in the plots, mean compensation was 31.6%, but if we increase somatosensory noise from 0.005 to 0.05, mean compensation rises to 36.9% due to the decreased reliability (i.e., the "numbing") of somatosensory feedback. This is consistent with prior findings about the effect of numbing somatosensation on compensation for pitch feedback perturbations (Larson, Altman, Liu, & Hain, 2008).

Other factors cause measured compensation to vary from trial to trial around the mean. State noise and observation (sensory feedback) noise cause feedback to fluctuate over the course of the simulation, which directly affects the measurement of peak compensation used to gauge compensation on individual trials. But the fluctuating feedback also indirectly affects compensation, because we hypothesize that the Kalman gain on sensory feedback is continually re-estimated from current sensory feedback over the course of the experiment. This re-estimation causes the Kalman gain to vary slightly from trial to trial, and since the size of the Kalman gain determines magnitude of compensation on each trial, compensation therefore fluctuates because of this.

Evidence that fluctuation in the Kalman gain contributes to compensation variability in real experiments comes from consideration of additional outputs of the simulation. The 4th panel of Figure 4 shows ERPs generated from the auditory feedback prediction errors in the speak (red) and listen (blue) conditions of the experimernt. At speech onset, these two responses differ greatly, demonstrating the SIS effect, but here the two responses are identical, since unlike speech onset, the externally-applied pitch perturbation is equally unexpected in both the speak and listen conditions. The 5th panel, however, shows that these equal feedback prediction errors nevertheless result in unequal state estimate corrections. The panel shows ERPs generated from state estimate corrections in the speak (red) and listen (blue) conditions, with larger ERPs (corresponding to larger state corrections) in the speak condtion. This speech perturbation response enhancement (SPRE) matches that seen in our prior studies (Chang, Niziolek, Knight, Nagarajan, & Houde, 2013; Kort et al., 2014), and in the simulations is the result of the Kalman gain on auditory feedback being larger in the speak condition than in the listen condition (because of the inability to ascribe the total variance to anything but observation noise).

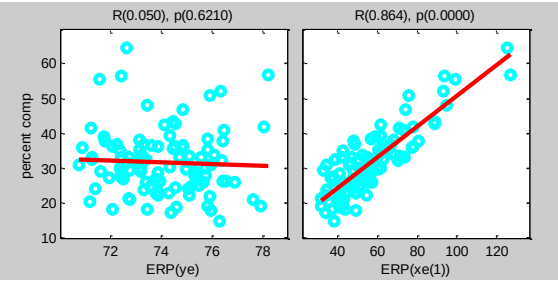

Figure 5: Correlation with compensation. Scatterplots comparing regression of percent compensation with ERPs from feedback prediction errors (ye) (left) which exhibit only SIS, and ERPs from state corrections (xe(1)) (right) which exhibit SPRE.

SPRE, therefore, is due to action of the Kalman gain on feedback prediction errors, and so activity in the parts of the model expressing SPRE will reflect trial-to-trial variability in the Kalman gain not seen in the feedback prediction errors. Thus, since Kalman gain determines compensation magnitude, ERPs from SPRE-expressing parts of the model (i.e., ERPs from state corrections) will be more correlated with trial-to-trial variation in compensation than ERPs from model components expressing only SIS (i.e., ERPs from feedback prediction errors). Figure 5 shows that this is the case in our simulations, which matches what we have found in previous pitch feedback perturbation experiments (Chang et al., 2013).

5. Conclusions

The concept of state feedback control (SFC) is a powerful and flexible model of motor control, and many current models of speech motor control can be described as examples of SFC. Here, we have considered an SFC model of speech motor control with a very general form, and found it can account for many of the known characteristics of the role of auditory feedback in the control of speech, as well as many of the phenomena observed in our previous studies of the neural processing of auditory feedback.

6. Acknowledgements

Support provided by NSF grant BCS-1262297 and NIH grant R01-DC010145.

7. References

Chang, E. F., Niziolek, C. A., Knight, R. T., Nagarajan, S. S., & Houde, J. F. (2013). Human cortical sensorimotor network underlying feedback control of vocal pitch. *Proceedings of the National Academy of Sciences, 110*(7), 2653-2658.

David, O., Harrison, L., & Friston, K. J. (2005). Modelling event-related responses in the brain. *Neuroimage, 25*(3), 756-770.

Guenther, F. H., & Vladusich, T. (2012). A Neural Theory of Speech Acquisition and Production. *J Neurolinguistics, 25*(5), 408-422.

Harris, C. M., & Wolpert, D. M. (1998). Signal-dependent noise determines motor planning. *Nature, 394*(6695), 780-784.

Hill, A. V. (1925). Length of muscle, and the heat and tension developed in an isometric contraction. *J Physiol, 60*(4), 237-263.

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience, 5*, 82.

Kort, N. S., Nagarajan, S. S., & Houde, J. F. (2014). A bilateral cortical network responds to pitch perturbations in speech feedback. *Neuroimage, 86*, 525-535.

Larson, C. R., Altman, K. W., Liu, H. J., & Hain, T. C. (2008). Interactions between auditory and somatosensory feedback for voice F-0 control. *Experimental Brain Research, 187*(4), 613-621.

Niziolek, C. A., Nagarajan, S. S., & Houde, J. F. (2013). What does motor efference copy represent? Evidence from speech production. *Journal of Neuroscience, 33*(41), 16110-16116.

Price, C. J., Crinion, J. T., & Macsweeney, M. (2011). A Generative Model of Speech Production in Broca's and Wernicke's Areas. *Frontiers in Psychology, 2*, 237.

Shalit, U., Zinger, N., Joshua, M., & Prut, Y. (2012). Descending systems translate transient cortical commands into a sustained muscle activation signal. *Cerebral Cortex, 22*(8), 1904-1914.

Story, B. H., & Titze, I. R. (1995). Voice simulation with a body-cover model of the vocal folds. *J Acoust Soc Am, 97*(2), 1249-1260.

Tian, X., & Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Front Psychol, 1*, 166.

Titze, I. R., Jiang, J., & Drucker, D. G. (1988). Preliminaries to the body-cover theory of pitch control. *Journal of Voice, 1*(4), 314-319.