The FACTS model of speech motor control: fusing state estimation and task-based control

Benjamin Parrell[1,*,+], Vikram Ramanarayanan[2,3,+], Srikantan Nagarajan[2,4], and John Houde[2]

[1]University of Wisconsin–Madison,Department of Communication Sciences and Disorders, Madison, WI, 53706, USA
[2]University of California, San Francisco, Department of Otolaryngology - Head and Neck Surgery, San Francisco, CA, 94117, USA
[3]Educational Testing Service R&D, San Francisco, CA, 94105, USA
[4]University of California, San Francisco, Department of Radiology and Biomedical Imaging, San Francisco, CA, 94117, USA
[*]bparrell@wisc.edu
[+]these authors contributed equally to this work

ABSTRACT

We present a new computational model of speech motor control: the Feedback-Aware Control of Tasks in Speech or *FACTS* model. This model is based on a state feedback control architecture, which is widely accepted in non-speech motor domains. The FACTS model employs a hierarchical observer-based architecture, with a distinct higher-level controller of speech tasks and a lower-level controller of speech articulators. The task controller is modeled as a dynamical system governing the creation of desired constrictions in the vocal tract, based on the Task Dynamics model. Critically, both the task and articulatory controllers rely on an internal estimate of the current state of the vocal tract to generate motor commands. This internal state estimate is derived from initial predictions based on efference copy of applied controls. The resulting state estimate is then used to generate predictions of expected auditory and somatosensory feedback, and a comparison between predicted feedback and actual feedback is used to update the internal state prediction. We show that the FACTS model is able to qualitatively replicate many characteristics of the human speech system: the model is robust to noise in both the sensory and motor pathways, is relatively unaffected by a loss of auditory feedback but is more significantly impacted by the loss of somatosensory feedback, and responds appropriately to externally-imposed alterations of auditory and somatosensory feedback. The model also replicates previously hypothesized trade-offs between reliance on auditory and somatosensory feedback in speech motor control and shows for the first time how this relationship may be mediated by acuity in each sensory domain. These results have important implications for our understanding of the speech motor control system in humans.

Introduction

Producing speech is one of the most complex motor activities humans perform. To produce even a single word, the activity of over 100 muscles must be precisely coordinated in space and time. This precise spatiotemporal control is difficult to master, and is not fully adult-like until the late teenage years[1]. How the brain and central nervous system (CNS) controls this complex system remains an outstanding question in speech motor neuroscience.

Early models of speech relied on servo control[2]. In this type of *feedback control* schema, the current feedback from the *plant* (thing to be controlled–for speech, this would be the articulators of the vocal tract, as well as perhaps the phonatory and respiratory systems) is compared against desired feedback and any discrepancy between the current and desired feedback drives the generation of motor commands to move the plant towards the current production goal. A challenge for any feedback control model of speech is the short, rapid movements that characterize speech motor behavior, with durations in the range of 50-300 ms. This is potentially shorter than the delays in the sensory systems. For speech, measured latencies to respond to external perturbations of the system range from 20-50 ms for unexpected mechanical loads[3,4] to around 150 ms for auditory perturbations[5,6]. Therefore, the information about the state of the vocal tract conveyed by sensory feedback to the CNS is delayed in time. Such delays can cause huge problems for feedback control, leading to unstable movements and oscillations around goal states. Furthermore, speech production is possible even in the absence of auditory feedback, as seen in the ability of healthy speakers to produce speech when auditory and feedback is masked by loud noise[7,8]. All of the above factors strongly indicate that speech cannot be controlled purely based on feedback control.

Several alternative approaches have been suggested to address these problems with feedback control in speech production

and other motor domains. One approach, the equilibrium point hypothesis[9–11], relegates feedback control to short-latency spinal or brainstem circuits operating on proprioceptive feedback, with high-level control based on pre-planned feedforward motor commands. Speech models of this type, such as the GEPPETO model, are able to reproduce many biomechanical aspects of speech but are not sensitive to auditory feedback[12–16]. Another approach is to combine feedback and feedforward controllers operating in parallel[17,18]. This is the approach taken by the DIVA model[19–22], which combines feedforward control based on desired articulatory positions with auditory and somatosensory feedback controllers. In this way, DIVA is sensitive to sensory feedback (via the feedback controllers) but capable of producing fast movements despite delayed or absent sensory feedback (via the feedforward controller).

A third approach, widely used in motor control models outside of speech, relies on the concept of state feedback control[23–26]. In this approach, the plant is assumed to have a state that is sufficiently detailed to predict the future behavior of the plant, and a controller drives the state of the plant towards a goal state, thereby accomplishing a desired behavior. A key concept in state feedback control is that the true state of the plant is not known to the controller; instead, it is only possible to estimate this state from efference copy of applied controls and sensory feedback. The internal state estimate is computed by first predicting the next plant state based on the applied controls. This state prediction is then used to generate predictions of expected feedback from the plant, and a comparison between predicted feedback and actual feedback is then used to correct the state prediction. Thus, in this process, the actual feedback from the plant only plays an indirect role in that it is only one of the inputs used to estimate the current state, making the system robust to feedback delays and noise.

We have earlier proposed a speech-specific instantiation of a state feedback control system[27]. The primary purpose of this earlier work was to establish the plausibility of the state feedback control architecture for speech production and suggest how such an architecture may be implemented in the human central nervous system. Computationally, our previous work built on models that have been developed in non-speech motor domains[24,25]. Following this work, we implemented the state estimation process as a prototypical Kalman filter[28], which provides an optimal posterior estimate of the plant state given a prior (the efference-copy-based state prediction) and a set of observations (the sensory reafference), assuming certain conditions such as a linear system. We subsequently implemented a one-dimensional model of vocal pitch control based on this framework[29].

However, the speech production system is substantially more complex than our one-dimensional model of pitch. First, speech production requires the multi-dimensional control of redundant and interacting articulators (e.g., lips, tongue tip, tongue body, jaw, etc.). Second, speech production relies on the control of high-level task goals rather than direct control of the articulatory configuration of the plant (e.g., for speech, positions of the vocal tract articulators). For example, speakers are able to compensate immediately for a bite block which fixes the jaw in place, producing essentially normal vowels[30]. Additionally, speakers react to displacement of a speech articulator by making compensatory movements of other articulators: speakers lower the upper lip when the jaw is pulled downward during production of a bilabial [b][3,4], and raise the lower lip when the upper lip is displaced upwards during production of [p][31]. Importantly, these actions are not reflexes, but are specific to the ongoing speech task. No upper lip movement is seen when the jaw is displaced during production of [z] (which does not require the lips to be close), nor is the lower lip movement increased if the upper lip is raised during production of [f] (where the upper lip is not involved). Together, these results strongly indicate that the goal of speech is not to achieve desired positions of each individual speech articulator, but must rather be to achieve some higher-level goal. While most models of speech motor production thus implement control at a higher speech-relevant level, the precise nature of these goals (vocal tract constrictions[32–34], auditory patterns[2,12,21], or both[14,22]) remains an ongoing debate.

One prominent model that employs control of high-level speech tasks rather than direct control of articulatory variables is the Task Dynamic model[32,35]. In Task Dynamics, the state of the plant (current positions and velocities of the speech articulators) is assumed to be available through proprioception. Importantly, this information is not used to directly generate an error or motor command. Rather, the current state of the plant is used to calculate values for various constrictions in the vocal tract (e.g., the distance between the upper and lower lip, the distance between the tongue tip and palate, etc.). It is these *constrictions*, rather than the positions of the individual articulators, that constitute the goals of speech production in Task Dynamics.

The model proposed here (Feedback Aware Control of Tasks in Speech, or *FACTS*) extends the idea of articulatory state estimation from the simple linear pitch control mechanism of our previous SFC model to the highly non-linear speech articulatory system. This presents three primary challenges: first, moving from pitch control to articulatory control requires the implementation of control at a higher level of speech-relevant tasks, rather than at the simpler level of articulator positions. To address this issue, FACTS is built upon the Task Dynamics model, as described above. However, unlike the Task Dynamics model, which assumes the state of the vocal tract is directly available through proprioception, here we model the more realistic situation in which the vocal tract state must be estimated from an efference copy of applied motor commands as well as somatosensory and auditory feedback. The second challenge is that this estimation process is highly non-linear. This required that the implementation of the observer as a Kalman filter in SFC be altered, as this estimation process is only applicable to linear systems. Here, we implement state estimation as an Unscented Kalman Filter[36], which is able to account for the

nonlinearities in the speech production system and incorporates internal state prediction, auditory feedback, and somatosensory feedback. Lastly, the highly non-linear mapping between articulatory positions and speech acoustics must be approximated in order to predict the auditory feedback during speech. Here, we learn the articulatory-to-acoustic mapping using Locally Weighted Projection Regression or LWPR[37]. Thus, in the proposed FACTS model, we combine the hierarchical architecture and task-based control of Task Dynamics with a non-linear state-estimation procedure to develop a model that is capable of rapid movements with or without sensory feedback, yet is still sensitive to external perturbations of both auditory and somatosensory feedback.

In the following sections, we describe the architecture of the model and the computational implementation of each model process. We then describe the results of a number of model simulations designed to test the ability of the model to simulate human speech behavior. First, we probe the behavior of the model when sensory feedback is limited to a single modality (somatosensation or audition), removed entirely, or corrupted by varying amounts of noise. Second, we test the ability of the model to respond to external auditory or mechanical perturbations. In both of these sections, we show that the model performs similarly to human speech behavior and makes new, testable predictions about the effects of sensory deprivation. Lastly, we explore how sensory acuity in both the auditory and somatosensory pathways affects the model's response to external auditory perturbations. Here, we show that the response to auditory perturbations depends not only on auditory acuity but on somatosensory acuity as well, demonstrating for the first time a potential computational principle that may underlie the demonstrated trade-off in response magnitude to auditory and somatosensory perturbations across human speakers.

Model Overview

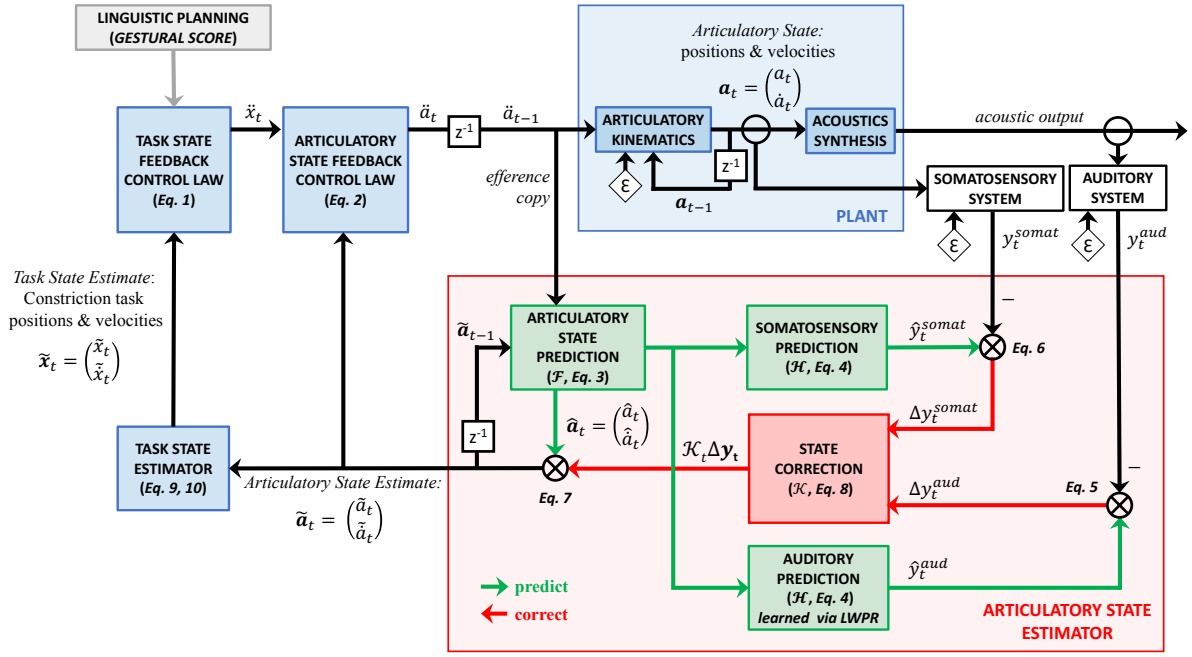

Figure 1. Architecture of the FACTS model. Boxes in blue represent processes outlined in the Task Dynamics model[32,38]. The articulatory state estimator (shown in red) is implemented as an Unscented Kalman Filter, which estimates the current articulatory state of the plant by combining the predicted state generated by a forward model with auditory and somatosensory feedback. Additive noise is represented by ε. Time-step delays are represented by z^{-1}. Equation numbers correspond to equations found in the methods.

A schematic control diagram of the FACTS model is shown in Fig. 1. Modules that build on Task Dynamics are shown in blue, and the articulatory state estimation process (or observer) is shown in red. Following Task Dynamics, speech tasks in the FACTS model are hypothesized to be desired constrictions in the vocal tract (e.g., close the lips for a [b]). Each of these speech tasks, or gestures, can be specified in terms of it's constriction location (where in the vocal tract the constriction is formed) and it's constriction degree (how narrow the constriction is). We model each gesture as a separate critically-damped second-order system[32]. Interestingly, similar dynamical behavior has been seen at a neural population level during the planning and execution of reaching movements in non-human primates[39,40], suggesting that a dynamical systems model of task-level control may

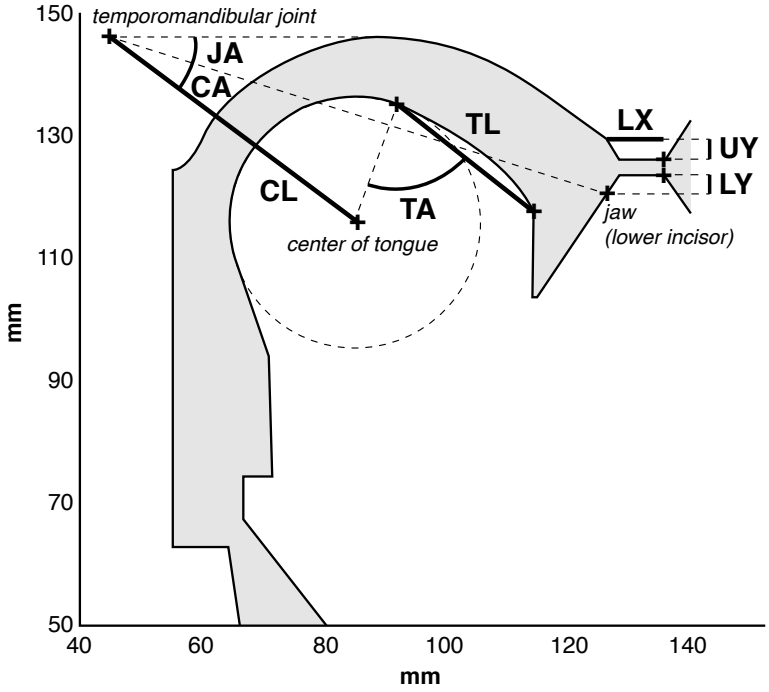

Figure 2. The CASY plant model. Articulatory variable relevant to the current paper are the Jaw Angle (JA), Condyle Angle (CA), and Condyle Length (CL). See text for a description of these variables. Diagram after[43].

be an appropriate first approximation to the neural activity that controls movement production. However, the architecture of the model would also allow for tasks in other control spaces, such as auditory targets (c.f.[13,19]), though an appropriate task feedback control law[1] for such targets would need to be developed.

FACTS uses as the Haskins Configurable Articulatory Synthesizer (or CASY)[38,41,42] as the model of the vocal tract **plant** being controlled. The relevant parameters of the CASY model required to move the tongue body to produce a vowel (and which fully describe the articulatory space for the majority of the simulations in this paper) are the Jaw Angle (JA, angle of the jaw relative to the temporomandibular joint), Condyle Angle (CA, the angle of the center of the tongue relative to the jaw, measured at the temporomandibular joint), and the Condyle Length (CL, distance of the center of the tongue from the temporomandibular joint along the Condyle Angle). The CASY model is shown in Fig. 2.

The model begins by receiving the output from a linguistic planning module. Currently, this is implemented as a *gestural score* in the framework of Articulatory Phonology[33,34]. These gestural scores list the control parameters (e.g., target constriction degree, constriction location, damping, etc.) for each gesture in a desired utterance as well as each gesture's onset and offset times. For example, the word "mod" ([mɑd]) has four gestures: simultaneous activation of a gesture driving closure at the lips for [m], a gesture driving an opening of the velum for nasalization of [m], and a gesture driving a wide opening between the tongue and hard palate for the vowel [ɑ]. These are followed by a gesture driving closure between the tongue tip and hard palate for [d] (Fig. 3).

The **task state feedback control law** takes these gestural scores as input and generates a task-level command based on the current state of the ongoing articulatory tasks. In this way, the task-level commands are dependent on the current task-level state. For example, if the lips are already closed during production of a /b/, a very different command needs to be generated than if the lips are far apart. These task-level commands are converted into motor commands that can drive changes in the positions of the speech articulators by the **articulatory state feedback control law**, using information about the current articulatory state of the vocal tract. The motor commands generate changes in the model vocal tract articulators (or **plant**), which are then used to generate an acoustic signal.

The **articulatory state estimator** (sometimes called an observer in other models) combines a copy of the outgoing motor command (or efference copy) with auditory and somatosensory feedback to generate an internal estimate of the articulatory state of the plant. First, the efference copy of the motor command is used (in combination with the previous arituclatory state estimate) to generate a prediction of the articulatory state. This is then used by a **forward model** (learned via LWPR) to

[1]Consistent with engineering control theory, we refer to the term "controller" as a "control law".

generate auditory and somatosensory predictions, which are compared to incoming sensory signals to generate sensory errors. Subsequently, these sensory errors are used to correct the state prediction to generate the final state estimate.

The final articulatory state estimate is used by the articulatory state feedback control law to generate the next motor command, as well as being passed to the **task state estimator** to estimate the current task state, or values (positions) and first derivatives (velocities) of the speech tasks (note the Task State was called the Vocal Tract State in earlier presentations of the model[44,45]). Finally, this estimated task-level state is passed to the task state feedback control law to generate the next task-level command.

A more detailed mathematical description of the model can be found in the methods.

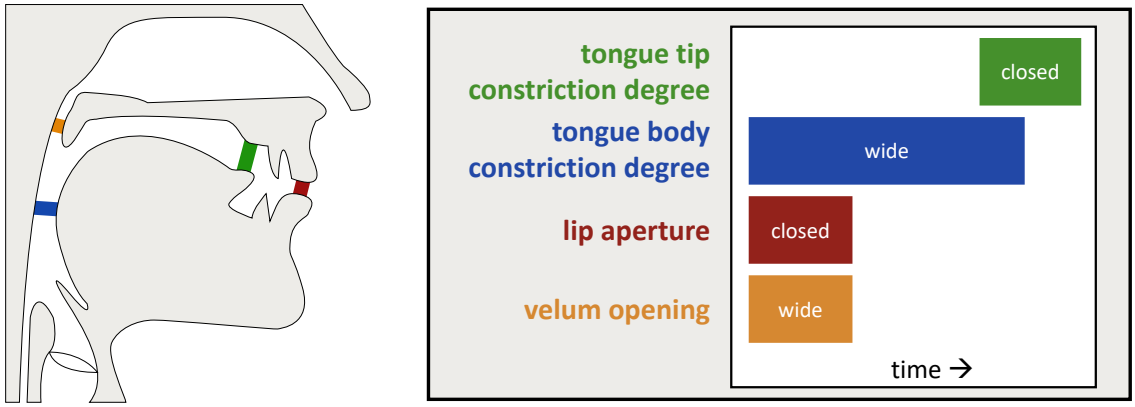

Figure 3. Example of task variables and gestural score for the word "mod". A gestural score for the word "mod" [mɑd], which consists of a bilabial closure and velum opening gestures for [m], a wide constriction in the pharynx for [ɑ], and a tongue tip closure gesture for [d]. Tasks are shown on the left, and a schematic of the gestural score on the right.

Results

Here we present results showing the accuracy of the learned forward model, and of various modeling experiments designed to test the ability of the model to qualitatively replicate human speech motor behavior under various conditions, including both normal speech as well as externally perturbed speech.

Forward model accuracy

Figure 4 visualizes a three dimensional subspace of the learned mapping from the 10-dimensional articulatory state space to the 3-dimensional space of formant frequencies (F1 – F3). Specifically, we look at the mapping from the tongue condyle length and condyle angle to the first (see Figure 4A-C) and second formants (see Figure 4D-F), projected onto each two-dimensional plane. We also plot normalized histograms of the number of receptive fields that cover each region of the space (represented as a heatmap in Figure 4A and D and with a thick blue line in the other subplots). In each figure, the size of the circles is proportional to the absolute value of the error between the actual and predicted formant values. Overall, the fit of the model is very good, with an average error of 4.2 Hz (std., 9.1 Hz) for F1 and 6.6 Hz (std., 19.1 Hz) for F2. Fit error increases in regions of the space that are relatively sparsely covered by receptive fields. In addition, the higher frequency of smaller circles at the margins of the distribution (and therefore the edges of the articulatory space) suggest that we need fewer receptive fields to cover these regions. Of course, this means that we do see some bigger circles in these regions where the functional mapping is not adequately represented by a small number of fields. Also note that we are only plotting the number of receptive fields that are employed to cover a given region of articulatory space, and this is *not* indicative of how much weight they carry in representing that region of space.

Model response to changes in sensory feedback availability and noise levels

How does the presence or absence of sensory feedback affect the speech motor control system? While there is no direct evidence to date on the effects of total loss of sensory information in human speech, some evidence comes from when sensory feedback from a single modality is attenuated or eliminated. Notably, the effects of removing auditory and somatosensory feedback differ. In terms of auditory feedback, speech production is relatively unaffected by it's absence: speech is generally unaffected by when auditory feedback is masked by loud masking noise[7,8]. However, alterations to somatosensory feedback have larger effects: blocking oral tactile sensation through afferent nerve injections or oral anesthesia leads to substantial imprecision in speech articulation[46,47].

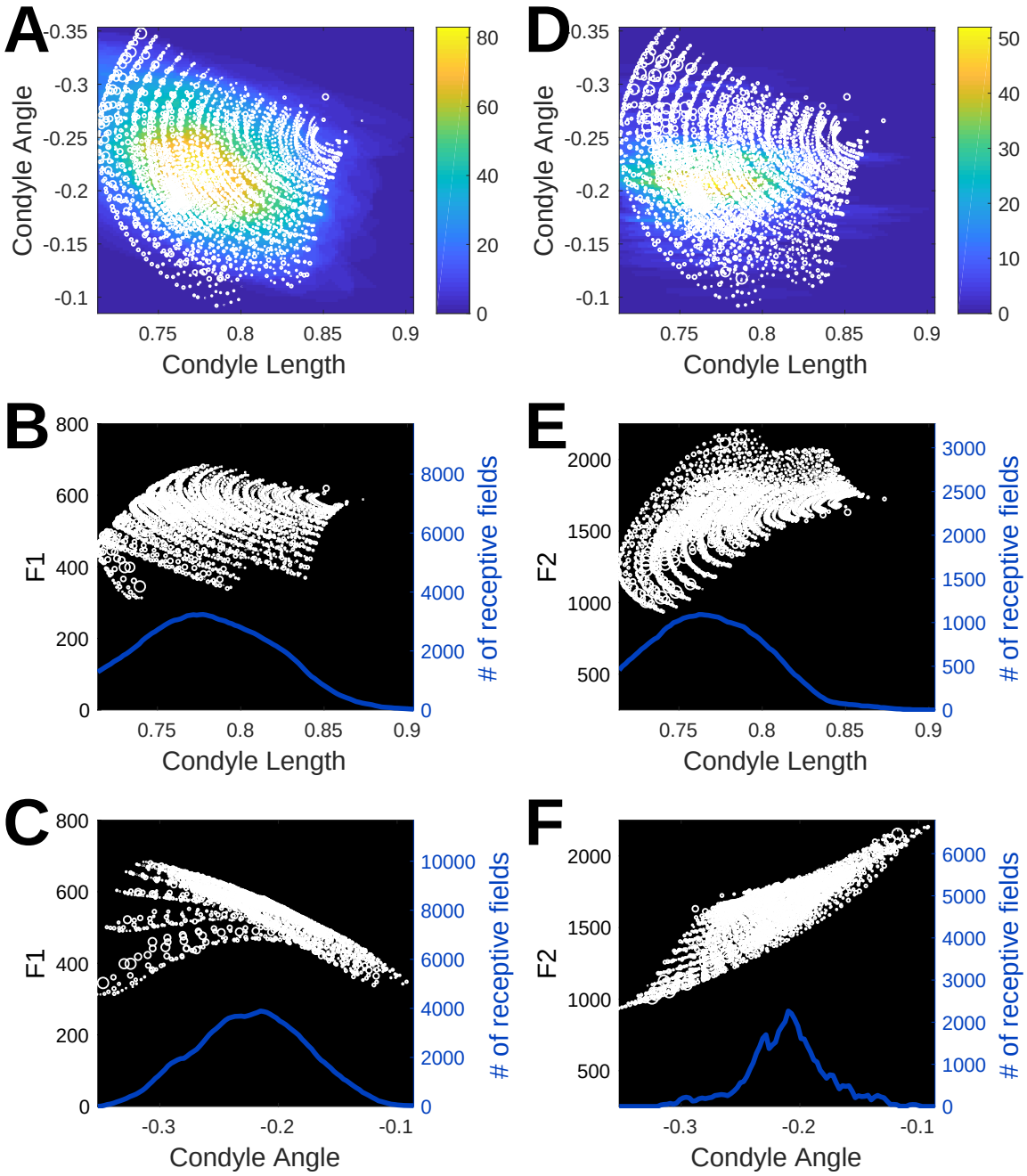

Figure 4. Learned LWPR transformations from CASY articulatory parameters to acoustics. Predicted formant values are shown as white circles. The width of each circle is proportional to the absolute value of the error between the actual formant values and the formant values predicted by the LWPR model. The density distributions reflect the number of receptive fields that cover each point (represented as colors in A, D and the height of the line in B, C, D, F). (A-C) show F1 values. (D-F) show F2 values. (A) Condyle Angle vs Condyle Length, (B) F1 vs Condyle Length, (C) F1 vs Condyle Angle. (D-F), replicate (A-C) for F2.

Fig. 5 presents simulations from the FACTS model testing the ability of the model to replicate the effects of removing sensory feedback seen in human speech. All simulations modeled the vowel sequence [ə a i]. 100 simulations were run for each of four conditions: normal feedback (5B), somatosensory feedback only (5C), auditory feedback only (5D), and no sensory feedback (5E). For clarity, only the trajectory of the tongue body in the CASY articulatory model is shown for each simulation.

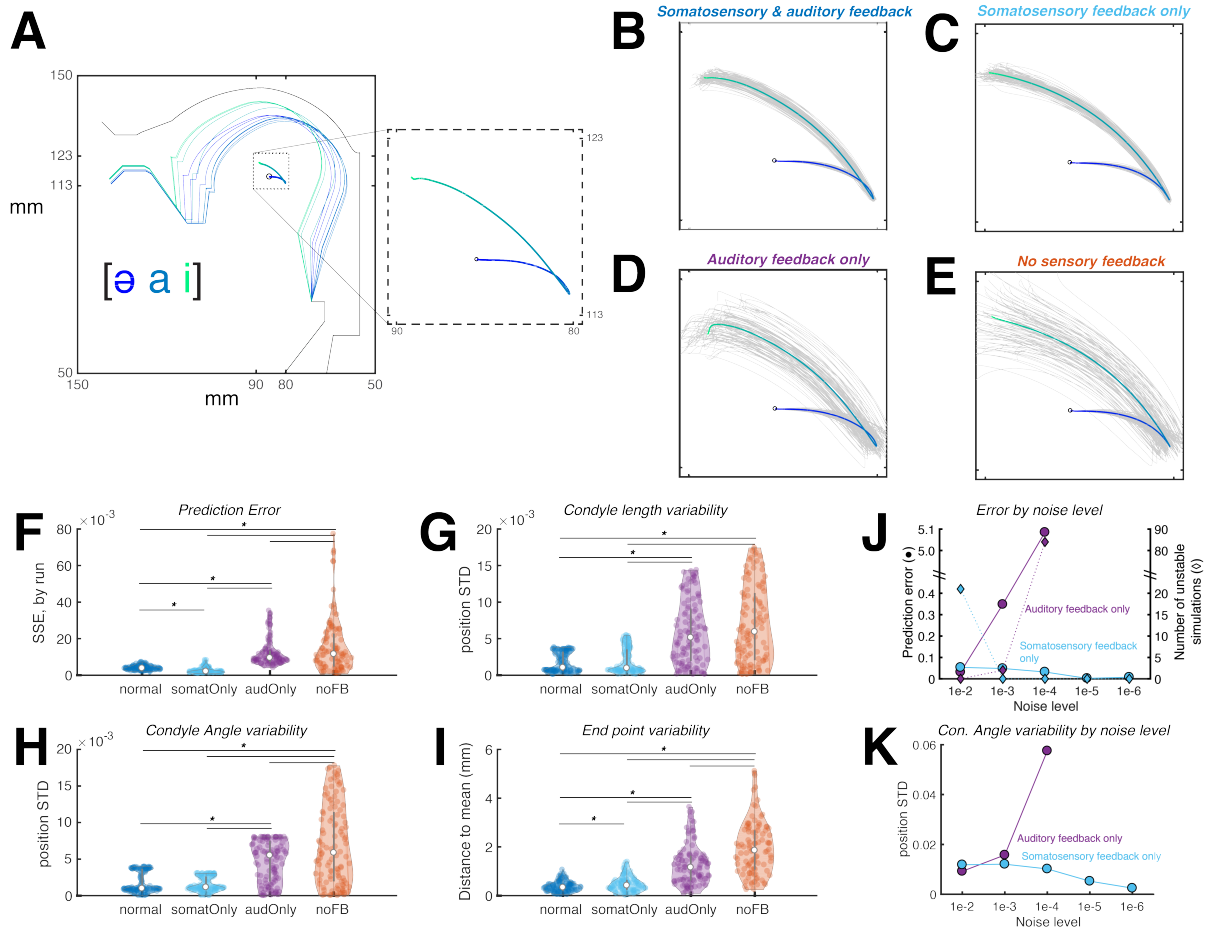

Figure 5. FACTS simulation of the vowel sequence [ə a i], varying the type of sensory feedback available to the model. (A(shows a sample simulation with movements of the CASY model as well as the trajectory of the tongue center. (B-E) each show tongue center trajectories from 100 simulations (gray) and their mean (blue-green gradient) with varying types of sensory feedback available. Variability is lower when sensory feedback is available, and lower when auditory feedback is absent compared to when somatosensory feedback is absent. (F) shows the prediction error in each condition. (G-H) show the produced variability in two articulatory parameters of the CASY plant model related to vowel production and (I) shows variability of the tongue center at the final simulation sample. (J) and (K) show prediction error and articulatory variability relative to sensory noise levels when only one feedback channel is available. Decreasing sensory noise leads to increased accuracy for somatosensation but decreased accuracy for audition. Colors in (F-K) correspond to the colors in the titles of (B-E).

In the normal feedback condition (Fig. 5B), the tongue lowers from [ə] to [a], then raises and fronts from [a] to [i]. Note that there is some variability across simulation runs due to the noise in the motor and sensory systems. This variability is also found in human behavior and the ability of the state feedback control architecture to replicate this variability is a strength of this approach[25].

The effect of removing auditory feedback (Fig. 5C) leads to a significant, though small, increase in the variability of the tongue body movement as measured by the tongue location at the movement endpoint (see Fig. 5I), though this effect was not seen in measures of Condyle Angle or Condyle Length variability (5G-H). Interestingly, while variablity increased, prediction error slightly decreased in this condition 5F. Overall, these results are consistent with experimental results that demonstrate that speech is essentially unaffected, in the short term, by the loss of auditory information (though auditory feedback is important for pitch and amplitude regulation[48] as well as to maintain articulatory accuracy in the long term[48–50]).

Removing somatosensory feedback while maintaining auditory feedback (Fig. 5B) leads to an increase in both variability across simulation runs as well as an increase in prediction error (Fig. 5F-I). This result is broadly consistent with the fact that reduction of tactile sensation via oral anaesthetic or nerve block leads to imprecise articulation for both consonants and vowels[47,51] (though the acoustic effects of this imprecision may be less perceptible for vowels[47]). However, a caveat must be made that our current model does not include tactile sensation, only proprioceptive information. Unfortunately, it is impossible to block proprioceptive information from the tongue, as that afferent information is likely carried along the same nerve (hypoglossal nerve) as the efferent motor commands[52]. It is difficult to prove, then, exactly how a complete loss of proprioception would affect speech. Nonetheless, the current model results are consistent with studies that how shown severe dyskinesia in reaching movements after elimination of proprioception in non-human primates (see[53] for a review) and in human patients with severe proprioceptive deficits[54]. In summary, although the FACTS model currently includes only proprioceptive sensory information rather than both proprioceptive and tactile signals, these simulation results are consistent with a critical role for the somatosensory system in maintaining the fine accuracy of the speech motor control system.

While removal of only auditory feedback lead to only small increases in variability (in both FACTS simulations and human speech), our simulations show speech in the complete absence of sensory feedback (Fig. 5E) shows much larger variability than the absence of either auditory or somatosensory feedback alone. This is consistent with human behvaior[51], and occurs because without sensory feedback there is no way to detect and correct for the noise inherent in the motor system (shown by the large prediction errors and increased articulatory variability in Fig. 5F.

The effects of changing the noise levels in the system can be see in Fig. 5J-K. For these simulations, only one type of feedback was used at a time: somatosensory (cyan) or auditory (purple). Noise levels (shown on the x axis) reflect both the sensory system noise and the internal estimate of that noise, which were set to be equal. Each data point reflects 100 stable simulations. Data for the acoustic-only simulations are not shown for noise levels below 1e-5 as the model became highly unstable in these conditions to due inaccurate articulatory state estimates (the number of unstable or divergent simulations is shown in Fig. 5J). For the somatosensory system, the prediction error and articulatory variability (shown here for the Condyle Angle) *decrease* as the noise decreases. However, for the auditory system, both prediction error and articulatory variability *increase* as the noise decreases. Because of the Kalman gain, decreased noise in a sensory or predictive signal leads not only to a more accurate signal, but also to a greater reliance on that signal compared to the internal state prediction. When the system relies more on the somatosensory signal, this results in a more accurate state estimate as the somatosensory signal directly reflects the state of the plant. When the system relies more on the auditory signal, however, this results in an less accurate state estimate as the auditory signal only indirectly reflects the state of the plant as a results of the nonlinear, many-to-one articulatory-to-acoustic mapping of the vocal tract.

In sum, FACTS is able to replicate the variability seen in human speech, as well as qualitatively match the effects of both auditory and somatosensory masking on speech accuracy. While the variability of human speech in the absence of proprioceptive feedback remains untested, the FACTS simulation results make a strong prediction that could be empirically tested in future work if some manner of blocking or altering proprioceptive signals could be devised.

Model response to mechanical and auditory perturbations

When a downward mechanical load is applied to the jaw during the production of a consonant, speakers respond by increasing the movements of the other speech articulators in a task-specific manner to achieve full closure of the vocal tract[3,4,31]. For example, when the jaw is perturbed during production of a bilabial stops /b/ or /p/, the upper lip moves downward to a greater extent than normal to compensate for the lower jaw position. This upper lip lowering is not found for jaw perturbations during /f/ or /z/, indicating it is specific to sounds produced using the upper lip. Conversely, tongue muscle activity is larger following jaw perturbation for /z/, which involves a constriction made with the tongue tip, but not for /b/, for which the tongue is not actively involved.

The ability to sense and compensate for mechanical perturbations relies on the somatosensory system. We tested the ability of FACTS to reproduce the task-specific compensatory responses to jaw load application seen in human speakers by applying a small downward acceleration to jaw (Jaw Angle parameter in CASY) starting midway through a consonant closure for stops produced with the lips (/p/) and tongue tip (/t/). The perturbation continued to the end of the word. As shown in Fig. 6A, the model produces greater lowering of the upper lip (as well as greater raising of the lower lip) when the jaw is fixed during production of /p/, but not during /t/, mirroring the observed response in human speech.

In addition to the task-specific response to mechanical perturbations, speakers will also adjust their speech in response to auditory perturbations[55]. For example, when the first vowel formant (F1) is artificially shifted upwards, speakers produce within-utterance compensatory responses by lowering their produced F1. The magnitude of these responses only partially compensates for the perturbation, unlike the complete responses produced for mechanical perturbations. While the exact reason for this partial compensation is not known, it has been hypothesized to relate to small feedback gains[19] or conflict with the somatosensory feedback system[14]. We explore the cause of this partial compensation below, but focus here on the ability of the

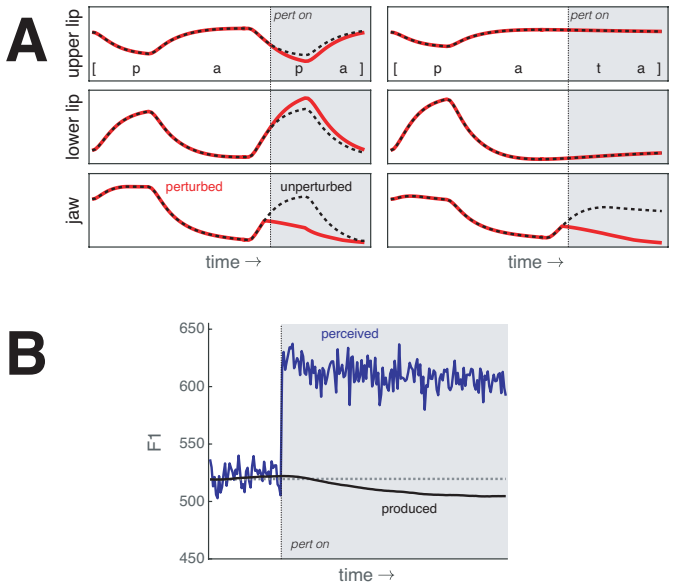

Figure 6. FACTS model simulations of mechanical and auditory perturbations. Times when the perturbations are applied are shown in gray. (A) shows the response of the model to fixing the jaw in place (simulating a downward force applied to the jaw) midway through the closure for second consonant in [papa] (left) and [pata] (right). Unperturbed trajectories are indicated with a dashed black line and perturbed trajectories with a solid red line. The upper and lower lips move more to compensate for the jaw perturbation only when the perturbation occurs on [p], mirroring the task-specific response seen in human speakers. (B) shows the response of the FACTS model to a +100 Hz auditory perturbation of F1 while producing [ə]. The produced F1 is shown in black and the perceived F1 is shown in blue. Note that the perceived F1 is corrupted by noise. The model responds to the perturbation by lowering F1 despite the lack of an explicit auditory target. The partial compensation to the perturbation produced by the model matches that observed in human speech.

model to replicate the observed behavior.

To test the ability of the FACTS model to reproduce the observed partial responses to auditory feedback perturbations, we simulated production of a steady-state [ə] vowel. After a brief stabilization period, we abruptly introduced a +100 Hz perturbation of F1 by adding 100 Hz to the perceived F1 signal in the auditory processing stage. This introduced a discrepancy between the produced F1 (shown in black in Fig. 6B) and the perceived F1 (shown in blue in Fig. 6B). Upon introduction of the perturbation, the model starting to produce a compensatory lowering of F1, eventually reaching a steady value below the unperturbed production. This compensation, like the response in human speakers, is only partial (roughly 20 Hz or 15% of the total perturbation).

Importantly, FACTS produces compensation auditory perturbations despite having no auditory targets. Previously, such compensation has been seen as evidence in favor of the existence of auditory targets for speech[14]. In FACTS, auditory perturbations cause a change in the estimated state of the vocal tract on which the task-level and articulatory-level feedback controllers operate. This causes a change in motor behavior compared to the unperturbed condition, resulting in what appears to be compensation for the auditory perturbation. Our model results thus show that this compensation is possible without explicit auditory goals. Of course, these results do not argue that auditory goals do not exist. Rather, we show that they are not necessary for this behavior.

Model trade-offs between auditory and somatosensory acuity

The amount of compensation to an auditory perturbation has been found to vary substantially between individuals[55]. One explanation for the inter-individual variability is that the degree of compensation is related to the acuity of the auditory system. Indeed, there seems to be a relationship between measurements of auditory acuity and the magnitude of the compensatory response to auditory perturbation of vowel formants[56]. If we assume that acuity is inversely related to the amount of noise in the sensory system, this explanation fits with the UKF implementation of the state estimation procedure in FACTS, where the weight assigned to the auditory error is ultimately related to the estimate of the noise in the auditory system. In Fig. 7B, we show that by varying the amount of noise in the auditory system (along with the internal estimate of that noise), we can drive differences in the amount of compensation the model produces to a +100 Hz perturbation of F1. When we double the auditory noise compared to baseline (top), the compensatory response is reduced. When we halve the auditory noise (bottom),

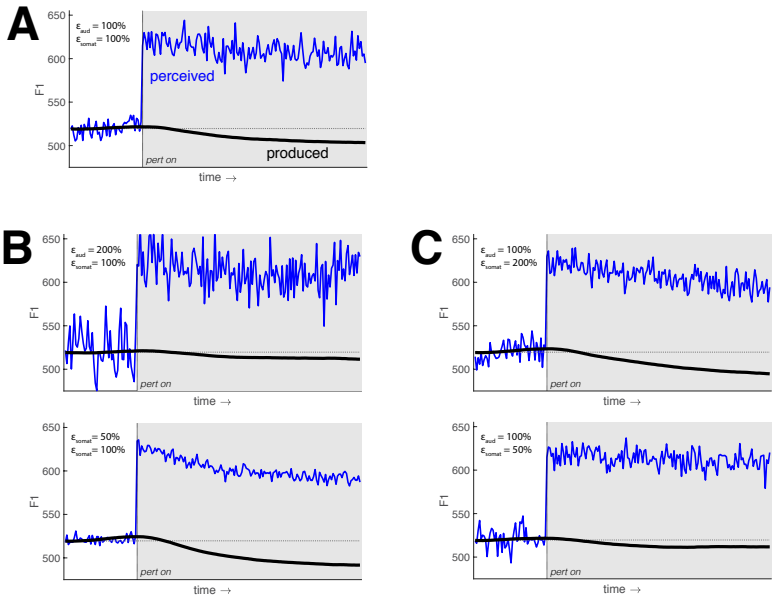

Figure 7. FACTS model simulations of the response to a +100 Hz perturbation of F1. (A) shows the baseline response. (B) shows the effects of altering the amount of noise in the auditory system in tandem with the observer's estimate of that noise. An increase in auditory noise (top) leads to a smaller perturbation response , while a decrease in auditory noise (bottom) leads to a larger response. (C) shows the effects of altering the amount of noise in the somatosensory system in tandem with the observer's estimate of that noise. The pattern is opposite of that for auditory noise. Here, an increase in somatosensory noise (top) leads to a larger perturbation response , while a decrease in somatosensory noise (bottom) leads to a smaller response.

the response increases.

Interestingly, the math underlying the UKF suggests that the magnitude of the response to an auditory error should be tied not only to the acuity of the auditory system, but to the acuity of the somatosensory system as well. This is because the weights assigned by the Kalman filter take the full noise covariance of all sensory systems into account. We verified this prediction empirically by running a second set of simulated responses to the same +100 Hz perturbation of F1, this time maintaining the level of auditory noise constant while varying only the level of somatosensory noise. The results can be seen in Fig. 7C. When the level of somatosensory noise is increased, the response to the auditory perturbation increases. Conversely, when the level of somatosensory noise is reduced, the compensatory response is reduced as well. These results suggest that the compensatory response in human speakers should be related to the acuity of the somatosensory system as well as the auditory system, a novel hypothesis which we are currently testing experimentally. Broadly, however, these results agree with, and provide a testable hypothesis about the cause of, empirical findings that show a trading relationship across speakers in their response to auditory and somatosensory perturbations[57].

Discussion

The proposed FACTS model provides a novel way to understand the speech motor system. The model is an implementation of state feedback control that combines high-level control of speech tasks with a non-linear method for estimating the current articulatory state to drive speech motor behavior. We have shown that the model replicates many important characteristics of human speech motor behavior: the model produces stable articulatory behavior, but with some trial-to-trial variability. This variability increases when somatosensory information is unavailable, but is largely unaffected by the loss of auditory feedback.

The model is also able to reproduce task-specific responses to external perturbations. For somatosensory perturbations, when a downward force is applied to the jaw during production of an oral consonant, there is an immediate compensatory response *only in those articulators needed to produce the current task*. This is seen in the increased movement of the upper and lower lips to compensate for the jaw perturbation during production of a bilabial /b/ but no alterations in lip movements when the jaw was perturbed during production of a tongue-tip consonant /d/. The ability of the model to respond to perturbations in a task-specific replicates a critical aspect of human speech behavior and is due to the inclusion of a task state feedback control law in the model[35].

For auditory perturbations, we showed that FACTS is able to produce compensatory responses to external perturbations

of F1, even though there is no explicit auditory goal in the model. Rather, the auditory signal is used to inform the observer about the current state of the vocal tract articulators. We additionally showed that FACTS is able to produce the inter-individual variability in the magnitude of this compensatory response as well as the previously observed relationship between the magnitude of this response and auditory acuity.

We have also shown that FACTS makes some predictions about the speech motor system that go beyond what has been demonstrated experimentally to date. FACTS predicts that a complete loss of sensory feedback would lead to large increases in articulatory variability beyond those seen in the absence of auditory or somatosensory feedback alone. Additionally, FACTS predicts that the magnitude of compensation for auditory perturbations should be related not only to auditory acuity, but to somatosensory acuity as well. These concrete predictions can be experimentally tested to assess the validity of the FACTS model, testing which is ongoing in our labs.

One of the major drawbacks of the current implementation of FACTS is that the model of the plant only requires kinematic control of articulatory positions. While a kinematic approach is relatively widespread in the speech motor control field–including both DIVA and Task Dynamics–there is experimental evidence that the dynamic properties of the articulators, such as gravity and tissue elasticity, need to be accounted for in its control[58–61]. Moreover, speakers will learn to compensate for perturbations of jaw protrusion that are dependent on jaw velocity[57,62–64], indicating that speakers are able to generate motor commands that anticipate and cancel out the effects of those altered articulatory dynamics. While the FACTS model in its current implementation does not replicate this dynamical control of the speech articulators, the overall architecture of the model is compatible with control of dynamics rather than just kinematics[23]. Control of articulatory dynamics would require a dynamic model of the plant and implementing a new articulatory-level feedback control law that would output motor commands as forces, rather than (or potentially in addition to) articulatory accelerations. Coupled with parallel changes to the articulatory state prediction process, this would allow for FACTS to control a dynamical plant without any changes to the overall architecture of the model.

While a detailed discussion of the neural basis of the computations in FACTS is beyond the scope of the current paper, in order to demonstrate the plausibility of FACTS as a neural model of speech motor control, we briefly touch on potential neural substrates that may underlie a state-feedback control architecture in speech[23,24,27]. The cerebellum is widely considered to play a critical role as an internal forward model to predict future articulatory and sensory states[26,65]. The process of state estimation may occur in the parietal cortex, and indeed inhibitory stimulation of the inferior parietal cortex with transcranial magnetic stimulation impairs sensorimotor learning in speech[66], consistent with a role in this process. However, state estimation for speech may also (or alternatively) reside in the ventral premotor cortex (vPMC) for speech, where the premotor cortices are well situated for integrating sensory information (received from sensory corteces via the arcuate fasiculus and the superior longitudinal fasiculus) with motor efference copy from primary motor cortex and cerebellum. Primary motor cortex (M1), with its descending control of the vocal tract musculature and bidirectional monosynaptic connections to primary sensory cortex, is the likely location of the articulatory feedback control law, converting task-level commands from vPMC to articulatory motor commands. Interestingly, recent work using electrocorticography has shown that areas in M1 code activation of task-specific muscle synergies similar to those proposed in Task Dynamics and FACTS[67]. This suggests that articulatory control may rely on synergies or primitives, rather than the control of individual articulators or muscles[68].

We have currently implemented the state estimation process in FACTS as an Unscented Kalman Filter. We intend this to be purely a mathematically tractable approximation of the actual neural computational process. Interestingly, recent work suggests that a related approach to nonlinear Bayesian estimation, the Neural Particle Filter, may provide a more neurobiologically plausible basis for the state estimation process[69]. Our future extensions of FACTS will involve exploring implementing this type of filter in FACTS.

In conclusion, the FACTS model uses a widely accepted domain-general approach to motor control, is able to replicate many important speech behaviors, and makes new predictions that can be experimentally tested. This model pushes forward our knowledge of the human speech motor control system, and we plan to further develop the model to incorporate other aspects of speech motor behavior, such as pitch control and sensorimotor learning, in future work.

Methods

Notation

We use the following mathematical notation to present the analyses described in this paper. Matrices are represented by bold uppercase letters (e.g., $\mathbf{X}$), while vectors are represented in italics without any bold case (either upper or lower case). We use the notation $\mathbf{X}^T$ to denote the matrix transpose of $\mathbf{X}$. Concatenations of vectors are represented using bold lowercase letters (e.g., $\mathbf{x} = [x\,\dot{x}]^T$). Scalar quantities are represented without bold and italics. Derivatives and estimates of vectors are represented with dot and tilde superscripts, respectively (i.e., $\dot{\mathbf{x}}$ and $\tilde{\mathbf{x}}$, respectively).

Task state feedback control law

In FACTS, we represent the state of the vocal tract tasks $\mathbf{x_t} = [x_t\, \dot{x}_t]^T$ at time t by a set of constriction task variables x_t (given the current gestural implementation of speech tasks, this is a set of constriction degrees such as lip aperture, tongue tip constriction degree, velic aperture, etc. and constriction locations, such as tongue tip constriction location) and their velocities $\dot{x}_t$. Given a gestural score generated using a linguistic gestural model[70,71], the *task state feedback control law* (equivalent to the Forward Task Dynamics model in[32]) allows us to generate the dynamical evolution of $\mathbf{x_t}$ using the following simple second-order critically-damped differential equation:

$$\ddot{x}_t = \mathbf{M}^{-1}(-\mathbf{B}\tilde{\dot{x}}_t - \mathbf{C}(\tilde{x}_t - x_0)) \tag{1}$$

where x_0 is the active task (or gestural) goal, $\mathbf{M}$, $\mathbf{B}$, and $\mathbf{C}$ are respectively the mass matrix, damping coefficient matrix, and stiffness coefficient matrix of the second-order dynamical system model. Essentially, the output of the task feedback controller, $\ddot{x}_t$, can be seen as a desired change (or command) in task space.This is passed to the articulatory state feedback control law to generate appropriate motor commands that will move the plant to achieve the desired task-level change.

Although the model does include a dynamical formulation of the evolution of speech tasks (following[32,35]), this is not intended to model the dynamics of the vocal tract plant itself. Rather, the individual speech tasks are modelled as (abstract) dynamical systems.

Articulatory state feedback control law

The desired task-level state change generated by the task feedback control law, $\ddot{x}_t$, is passed to an articulatory feedback control law. In our implementation of this control law, we use Eq. 2 (after[32]) to perform an inverse kinematics mapping from the task accelerations $\ddot{x}_t$ to the model articulator accelerations $\ddot{a}_t$, a process which is also dependent on the current estimate of the articulator positions $\tilde{a}_t$ and velocities $\tilde{\dot{a}}_t$. $\mathbf{J}(\tilde{a})$ is the Jacobian matrix of the forward kinematics model relating changes in articulatory states to changes in task states, $\dot{\mathbf{J}}(\tilde{a}, \tilde{\dot{a}})$ is the result of differentiating the elements of $\mathbf{J}(\tilde{a})$ with respect to time, and $\mathbf{J}(\tilde{a})^*$ is a weighted Jacobian pseudoinverse of $\mathbf{J}(\tilde{a})$.

$$\ddot{a}_t = \mathbf{J}(\tilde{a}_t)^*\ddot{x}_t - \mathbf{J}(\tilde{a}_t)^*\dot{\mathbf{J}}(\tilde{a}_t, \tilde{\dot{a}}_t)\tilde{\dot{a}}_t \tag{2}$$

Plant

In order to generate articulatory movements in CASY, we use Runge-Kutta integration to combine the previous articulatory state of the plant ($[a_{t-1}\, \dot{a}_{t-1}]^T$) with the output of the inverse kinematics computation ($\ddot{a}_{t-1}$, the input to the plant, which we refer to as the **motor command**). This allows us to compute the model articulator positions and velocities for the next time-step ($[a_t\, \dot{a}_t]^T$), which effectively "moves" the articulatory vocal tract model. Then, a tube-based *synthesis model* converts the model articulator and constriction task values into the output acoustics (parameterized by the vector y_t^{aud}). In order to model noise in the neural system, zero-mean Gaussian white noise ε is added to the motor command ($\ddot{a}_{t-1}$) received by the plant as well as to the somatosensory (y_t^{somat}) and auditory (y_t^{aud}) signals passed from the plant to the articulatory state estimator. Currently, noise levels (standard deviation of Gaussian noise) are tuned by hand for each of these signals (see below for details). Together, the CASY model and the acoustic synthesis process constitute the plant. The model vocal tract in the current implementation of the FACTS model is the Haskins Configurable Articulatory Synthesizer (or CASY)[38,41,42].

Articulatory state estimator

The articulatory state estimator generates an estimate of the articulatory state of the plant needed to generate state-dependent motor commands. The final state estimate ($\hat{\mathbf{a}}_\mathbf{t}$) generated by the observer is a combination of an articulatory state prediction ($\tilde{\mathbf{a}}_\mathbf{t}$) generated from an efference copy of outgoing motor commands, combined with information about the state of the plant derived from the somatosensory and auditory systems ($\mathbf{y_t}$). This combination of internal prediction and sensory information is accomplished through the use of an Unscented Kalman Filter (UKF)[36], which extends the linear Kalman Filter[28] used in most non-speech motor control models[23,25] to nonlinear systems like the speech production system.

First, the state prediction is generated using a **forward model** ($\mathscr{F}$) that predicts the evolution of the plant based on an estimate of the previous state of the plant ($\tilde{\mathbf{a}}_{\mathbf{t-1}}$) and an efference copy of the previously issued motor command ($\ddot{a}_{t-1}$). Based on this predicted state, another forward model ($\mathscr{H}$) generates the predicted sensory output $\hat{\mathbf{y}}_\mathbf{t} = [\hat{y}_t^{somat}\, \hat{y}_t^{aud}]^T$ (comprising somatosensory and auditory signals $\hat{y}_t^{somat}$ and $\hat{y}_t^{aud}$, respectively) that would be generated by the plant in the predicted state. Currently, auditory signals are modelled as the first three formant values (F1-F3; 3 dimensions), and somatosensory signals are modelled as the position and velocities of the speech articulators in the CASY model (20 dimensions).

$$\hat{\mathbf{a}}_\mathbf{t} = \mathscr{F}(\tilde{\mathbf{a}}_{\mathbf{t-1}}, \ddot{a}_{t-1}) \tag{3}$$

$$\hat{\mathbf{y}}_\mathbf{t} = \begin{bmatrix} \mathscr{H}^{somat}(\hat{a}_t^{somat}) \\ \mathscr{H}^{aud}(\hat{a}_t^{aud}) \end{bmatrix} \tag{4}$$

These predicted sensory signals are then compared with the incoming signals from the somatosensory (y_t^{somat}) and auditory (y_t^{aud}) systems, generating the sensory prediction error (comprising both somatosensory and auditory components) $\Delta\mathbf{y_t} = [\Delta y_t^{somat} \Delta y_t^{aud}]^T$:

$$\Delta y_t^{aud} = \hat{y}_t^{aud} - y_t^{aud} \tag{5}$$

$$\Delta y_t^{somat} = \hat{y}_t^{somat} - y_t^{somat} \tag{6}$$

These sensory prediction errors are used to correct the initial articulatory state prediction, giving a final articulatory state estimate $\tilde{\mathbf{a}}_\mathbf{t}$:

$$\tilde{\mathbf{a}}_\mathbf{t} = \hat{\mathbf{a}}_\mathbf{t} + \mathscr{K}_t \Delta\mathbf{y_t} \tag{7}$$

where $\mathscr{K}_t$ is the Kalman Gain, which effectively specifies the weights given to the sensory signals in informing the final state estimate. Details of how we generate $\mathscr{F}$, $\mathscr{H}$, and $\mathscr{K}$ are given in the following sections.

Forward models for state and sensory prediction

One of the challenges in estimating the state of the plant is that both the process model $\mathscr{F}$ (that provides a functional mapping from $[\tilde{a}_{t-1}\ \tilde{a}_{t-1}\ \ddot{a}_{t-1}]^T$ to $\hat{\mathbf{a}}_\mathbf{t}$) as well as the observation model $\mathscr{H}$ (that maps from $\hat{\mathbf{a}}_\mathbf{t}$ to $\hat{\mathbf{y}}_\mathbf{t}$) are unknown. Currently, we implement the process model $\mathscr{F}$ by replicating the integration of $\ddot{a}_t$ used to drive changes in the CASY model, which ignores any potential dynamical effects in the plant. However, the underlying architecture (the forward model $\mathscr{F}$) is sufficiently general that non-zero articulatory dynamics could be accounted for in predicting $\hat{\mathbf{a}}$ as well.

Implementing the observation model is more challenging due to the nonlinear relationship between articulator positions and formant values. In order to solve this problem, we *learn* the observation model functional mappings from articulatory positions to acoustics ($\hat{y}_t^{aud} = \mathscr{H}(\hat{\mathbf{a}}_\mathbf{t})$) required for Unscented Kalman Filtering using Locally Weighted Projection Regression, or LWPR, a computationally efficient machine learning technique[37]. While we do not here explicitly relate this machine learning process to human learning, such maps could theoretically be learned during early speech acquisition, such as babbling[22]. Currently, we learn only the auditory prediction component of $\mathscr{H}$. Since the dimensions of the somatosensory prediction are identical to those of the predicted articulatory state, the former are generated from the latter via an identity function ($\hat{y}_t^{somat} = \hat{\mathbf{a}}_\mathbf{t}$).

State correction using an Unscented Kalman Filter

Errors between predicted and afferent sensory signals are used to correct the initial efference-copy-based state prediction through the use of a Kalman gain ($\mathscr{K}$, Eq. 7), which effectively dictates how each dimension of the sensory error signal $\Delta\mathbf{y}_t$ affects each dimension of the state prediction $\hat{\mathbf{a}}_\mathbf{t}$. Our previous SFC model of vocal pitch control implemented a Kalman filter to estimate the weights in the Kalman gain[29], which provides the optimal state estimate under certain strict conditions, including that the system being estimated is linear[28]. However, the traditional Kalman filter approach is only applicable to linear systems, and the speech production mechanism, even in the simplified CASY model used in FACTS, is highly non-linear.

Our goal in generating a state estimate is to combine the state prediction and sensory feedback to generate an optimal or near-optimal solution. To accomplish this, we use an Unscented Kalman Filter (UKF)[36], which extends the linear Kalman Filter to non-linear systems. While the UKF has not been proven to provide optimal solutions to the state estimation problem, it consistently provides more accurate estimates than other non-linear extensions of the Kalman filter, such as the Extended Kalman Filter[36].

In FACTS, the weights of the Kalman gain are computed as a function of the estimated covariation between the articulatory state and sensory signals, given by the posterior covariance matrices $\mathbf{P}_{\mathbf{a_t y_t}}$ and $\mathbf{P}_{\mathbf{y_t y_t}}$ as follows:

$$\mathscr{K}_t = \mathbf{P}_{\mathbf{a_t y_t}} \mathbf{P}^{-\mathbf{1}}_{\mathbf{y_t y_t}} \tag{8}$$

In order to generate the posterior means (the state and sensory predictions) and covariances (used to caculate $\mathscr{K}$) in an unscented Kalman Filter (UKF)[36], multiple prior points (called sigma points, $\mathscr{X}$) are used. These prior points are chosen

carefully to capture the mean and covariance of the prior state. Each of these points is then projected through the nonlinear forward model function ($\mathscr{F}$ or $\mathscr{H}$), after which the posterior mean and covariance can be calculated from the transformed points. This process is called the unscented transform (Fig. 8). This is used both to predict the future state of the system (process model, $\mathscr{F}$) as well as the expected sensory feedback (observation model, $\mathscr{H}$).

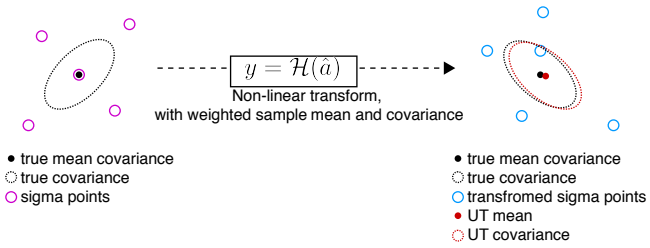

Figure 8. Cartoon showing the unscented transform (UT). The final estimate of the mean and covariance provide a better fit for the true values than would be achieved with the transformation of only a single point at the pre-transformation mean.

First, the sigma points ($\mathscr{X}$) are generated:

$$\mathscr{X}_{t-1} = [\hat{\mathbf{s}}_{\mathbf{t-1}} \pm \sqrt{(L+\lambda)\mathbf{P}_{\mathbf{t-1}}}] \tag{9}$$

where $\hat{\mathbf{s}}_{\mathbf{t-1}} = [\mathbf{a}_{\mathbf{t-1}}^{\mathbf{T}}\ \mathbf{v}_{\mathbf{t-1}}^{\mathbf{T}}\ \mathbf{n}_{\mathbf{t-1}}^{\mathbf{T}}]^{\mathbf{T}}$), and $\mathbf{v}$ and $\mathbf{n}$ are estimates of the process noise (noise in the plant articulatory system) and observation noise (noise in the sensory systems), respectively, L is the dimension of the dimension of the articulatory state $\mathbf{a}$, λ is a scaling factor, and $\mathbf{P}$ is the noise covariance of $\mathbf{a}$, $\mathbf{v}$, and $\mathbf{n}$. In our current implementation, the level of noise for $\mathbf{v}$ and $\mathbf{n}$ are set to be equal to the level of Gaussian noise added to the plant and sensory systems.

The observer then estimates how the motor command $\ddot{a}_t$ would effect the speech articulators by replicating using the Euler integration model ($\mathscr{F}$) to generate the state prediction $\hat{\mathbf{a}}_\mathbf{t} = [\hat{a}_t\ \hat{a}_t]^T$. First, all sigma points reflecting the articulatory state $\mathscr{X}^a$ and process noise $\mathscr{X}^v$ are passed through $\mathscr{F}$:

$$\mathscr{X}^a_{t|t-1} = \mathscr{F}[\mathscr{X}^a_{t-1}, \ddot{a}_{t-1}, \mathscr{X}^v_{t-1}] \tag{10}$$

and the estimated articulatory state is calculated as the weighted sum of the sigma points where the weights (W) are inversely related to the distance of the sigma point from the center of the distribution.

$$\hat{\mathbf{a}}_\mathbf{t} = \sum_{i=0}^{2L} W_i\, \mathscr{X}^a_{i,t|t-1} \tag{11}$$

The expected sensory state ($\hat{y}_t$) is then derived based on the predicted articulatory state in a similar manner, first by projecting the articulatory $\mathscr{X}^a$ and observation noise $\mathscr{X}^n$ sigma points through the articulatory-to-sensory transform $\mathscr{H}$.

$$\mathscr{Y}_{t|t-1} = \mathscr{H}\left(\mathscr{X}^a_{t|t-1}, \mathscr{X}^n_{t-1}\right) \tag{12}$$

$$\hat{\mathbf{y}}_\mathbf{t} = \sum_{i=0}^{2L} W_i\, \mathscr{Y}_{i,t|t-1} \tag{13}$$

Lastly, the posterior covariance matrices $\mathbf{P}_{\mathbf{x_t y_t}}$ and $\mathbf{P}_{\mathbf{y_t y_t}}$ necessary to generate the Kalman Gain $\mathscr{K}$ as well as the Kalman Gain itself are calculated in the following manner:

$$\mathbf{P}_{\mathbf{x_t y_t}} = \sum_{i=0}^{2L} W_i[\mathscr{X}_{i,t|t-1} - \hat{\mathbf{a}}_\mathbf{t}][\mathscr{Y}_{i,t|t-1} - \hat{\mathbf{y}}_\mathbf{t}]^T \tag{14}$$

$$\mathbf{P}_{\mathbf{y_t y_t}} = \sum_{i=0}^{2L} W_i[\mathscr{Y}_{i,t|t-1} - \hat{\mathbf{y}}_\mathbf{t}][\mathscr{Y}_{i,t|t-1} - \hat{\mathbf{y}}_\mathbf{t}]^T \tag{15}$$

$$\mathscr{K}_t = \mathbf{P}_{\mathbf{x_t y_t}} \mathbf{P}^{-\mathbf{1}}_{\mathbf{y_t y_t}} \tag{16}$$

Task state estimator

Finally, we estimate the vocal tract state estimate at the next time step by passing the articulatory state estimate into a task state estimator, which in our current implementation is a forward kinematics model (see Eq. 2)[32]. $\mathbf{J}(\tilde{a})$, the Jacobian matrix relating changes in articulatory states to changes in task states, is the same as in Eq. 2.

$$\tilde{x}_t = f(\tilde{a}_t) \tag{17}$$

$$\tilde{\dot{x}}_t = \mathbf{J}(\tilde{a}_t)\tilde{\dot{a}}_t \tag{18}$$

This task state estimate is then passed to the task feedback controller to generate the next task-level command $\ddot{x}_t$ using Eq. 1.

Model parameter settings

There are a number of tunable parameters in the FACTS model. These include: 1) the noise added to $\ddot{a}$ in the plant, y_{aud} in the auditory system, and y_{somat} in the somatosensory system; 2) the internal estimates of the process ($\ddot{a}$) and observation $\mathbf{y}$ noise; and 3) initial values for the process, observation, and state covariance matrices used in the Unscented Kalman Filter. Internal estimates of the process and observation noise were set to be equal to the true noise levels. Noise levels were tuned by hand (using a range from 1e-1 to 1e-8, scaled by the norm of each signal) to achieve the following goals: stable system behavior in the absence of external perturbations, the ability of the model to react to external auditory and somatosensory perturbations, and a partial compensation for external auditory perturbations in line with observed human behavior. The final noise values used were 1e-4 for the plant/process noise, 1e-2 for the auditory noise, and 1e-6 for the somatosensory noise. The discrepancy in the values for the noise between the two sensory domains is proportional to the difference in magnitude between the two signals (300-3100 Hz for the auditory signal, 0-1.2 mm or mm/s for the articulatory position and velocity signals). Process and observation covariance matrices were initialized as identity matrices scaled by the process and observation noise, respectively. The state covariance matrix was initialized as an identity matrix scaled by 1e-2. A relatively wide range of noise values produced similar behavior: the effects of changing the auditory and somatosensory noise levels are discussed in the results section.

References

1. Walsh, B. & Smith, A. Articulatory movements in adolescents: evidence for protracted development of speech motor control processes. *J Speech Lang Hear. Res* **45**, 1119–33 (2002).
2. Fairbanks, G. Systematic Research In Experimental Phonetics:* 1. A Theory Of The Speech Mechanism As A Servosystem. *J. Speech Hear. Disord.* **19**, 133–139 (1954). DOI 10.1044/jshd.1902.133.
3. Abbs, J. H. & Gracco, V. L. Control of complex motor gestures: orofacial muscle responses to load perturbations of lip during speech. *J Neurophysiol* **51**, 705–23 (1984).
4. Kelso, J. A., Tuller, B., Vatikiotis-Bateson, E. & Fowler, C. A. Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures. *J Exp Psychol Hum Percept Perform* **10**, 812–32 (1984).
5. Parrell, B., Agnew, Z., Nagarajan, S., Houde, J. F. & Ivry, R. B. Impaired Feedforward Control and Enhanced Feedback Control of Speech in Patients with Cerebellar Degeneration. *J Neurosci* **37**, 9249–9258 (2017). DOI 10.1523/JNEUROSCI.3363-16.2017.
6. Cai, S. *et al.* Weak responses to auditory feedback perturbation during articulation in persons who stutter: evidence for abnormal auditory-motor transformation. *PLoS One* **7**, e41830 (2012). DOI 10.1371/journal.pone.0041830.
7. Lane, H. & Tranel, B. The Lombard Sign and the Role of Hearing in Speech. *J. Speech, Lang. Hear. Res.* **14**, 677–709 (1971). DOI 10.1044/jshr.1404.677.
8. Lombard, E. Le signe de l'elevation de la voix. *Ann Maladies de L'Oreille et du Larynx* **37**, 2 (1911).
9. Feldman, A. Once more on the Equilibrium Point Hypothesis (λ) for motor control. *J. Mot. Behav.* **18**, 17–54 (1986).
10. Feldman, A., Adamovich, S., Ostry, D. & Flanagan, J. *The origin of electromyograms – Explanations based on the Equilibrium Point Hypothesis* (Springer Verlag, New York, 1990).
11. Feldman, A. G. & Levin, M. F. The Equilibrium-Point Hypothesis – Past, Present and Future. In *Progress in Motor Control*, Advances in Experimental Medicine and Biology, 699–726 (Springer, Boston, MA, 2009). DOI 10.1007/978-0-387-77064-2$_3$8.

12. Perrier, P., Ma, L. & Payan, Y. Modeling the production of VCV sequences via the inversion of a biomechanical model of the tongue. In *Proceeding of the INTERSPEECH: Interspeech'2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, 1041–1044 (2005).

13. Perrier, P., Ostry, D. & Laboissière, R. The Equilibrium Point Hypothesis and Its Application to Speech Motor Control. *J. Speech Hear. Res.* **39**, 365–378 (1996).

14. Perrier, P. & Fuchs, S. F. Motor equivalence in speech production. In Redford, M. (ed.) *The Handbook of Speech Production* (Wiley-Blackwell, Hoboken, NJ, 2015).

15. Sanguineti, V., Laboissière, R. & Ostry, D. J. A dynamic biomechanical model for neural control of speech production. *The J. Acoust. Soc. Am.* **103**, 1615–1627 (1998). DOI 10.1121/1.421296.

16. Laboissière, R., Ostry, D. J. & Feldman, A. G. The control of multi-muscle systems: human jaw and hyoid movements. *Biol. Cybern.* **74**, 373–384 (1996). DOI 10.1007/BF00194930.

17. Kawato, M. & Gomi, H. A computational model of four regions of the cerebellum based on feedback-error learning. *Biol. Cybern.* **68**, 95–103 (1992). DOI 10.1007/BF00201431.

18. Arbib, M. A. Perceptual Structures and Distributed Motor Control. In Brookhart, J. M., Mountcastle, V. B. & Brooks, V. (eds.) *Handbook of Physiology, Supplement 2: Handbook of Physiology, The Nervous System, Motor Control* (1981). DOI 10.1002/cphy.cp010233.

19. Guenther, F. H. *Neural control of speech* (The MIT Press, Cambridge, MA, 2015).

20. Guenther, F. H. Speech Sound Acquisition, Coarticulation, and Rate Effects in a Neural Network Model of Speech Production. *Psychol. Rev.* **102**, 594–621 (1995).

21. Guenther, F. H., Hampson, M. & Johnson, D. A theoretical investigation of reference frames for the planning of speech movements. *Psychol. Rev.* **105**, 611–633 (1998).

22. Tourville, J. A. & Guenther, F. H. The diva model: A neural theory of speech acquisition and production. *Lang Cogn Process.* **26**, 952–981 (2011). DOI 10.1080/01690960903498424.

23. Scott, S. H. The computational and neural basis of voluntary motor control and planning. *Trends Cogn Sci* **16**, 541–9 (2012). DOI 10.1016/j.tics.2012.09.008.

24. Shadmehr, R. & Krakauer, J. W. A computational neuroanatomy for motor control. *Exp Brain Res* **185**, 359–81 (2008). DOI 10.1007/s00221-008-1280-5.

25. Todorov, E. & Jordan, M. I. Optimal feedback control as a theory of motor coordination. *Nat Neurosci* **5**, 1226–35 (2002). DOI 10.1038/nn963.

26. Wolpert, D. M. & Miall, R. C. Forward Models for Physiological Motor Control. *Neural Netw* **9**, 1265–1279 (1996).

27. Houde, J. F. & Nagarajan, S. S. Speech production as state feedback control. *Front Hum Neurosci* **5**, 82 (2011).

28. Kalman, R. E. A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng.* **82**, 35–45 (1960). DOI 10.1115/1.3662552.

29. Houde, J. F., Niziolek, C., Kort, N., Agnew, Z. & Nagarajan, S. S. Simulating a state feedback model of speaking. In *10th International Seminar on Speech Production*, 202–205 (2014).

30. Fowler, C. A. & Turvey, M. T. Immediate compensation in bite-block speech. *Phonetica* **37**, 306–326 (1981). DOI 10.1159/000260000.

31. Shaiman, S. & Gracco, V. L. Task-specific sensorimotor interactions in speech production. *Exp. Brain Res.* **146**, 411–418 (2002-10-01). DOI 10.1007/s00221-002-1195-5.

32. Saltzman, E. & Munhall, K. A dynamical approach to gestural patterning in speech production. *Ecol. Psychol.* **1**, 333–382 (1989).

33. Browman, C. P. & Goldstein, L. Articulatory phonology: An overview. *Phonetica* **49**, 155–180 (1992).

34. Browman, C. & Goldstein, L. Dynamics and articulatory phonology. In Port, R. & van Gelder, T. (eds.) *Mind as motion: Explorations in the dynamics of cognition*, 175–194 (MIT Press, Boston, 1995).

35. Saltzman, E. Task dynamic coordination of the speech articulators: A preliminary model. *Exp. Brain Res. Ser.* 129–144 (1986).

36. Wan, E. A. & Van Der Merwe, R. The unscented Kalman filter for nonlinear estimation. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, 153–158 (Ieee, 2000).

37. Mitrovic, D., Klanke, S. & Vijayakumar, S. Adaptive optimal feedback control with learned internal dynamics models. In *From Motor Learning to Interaction Learning in Robots*, 65–84 (Springer, 2010).

38. Saltzman, E., Nam, H., Krivokapic, J. & Goldstein, L. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In *Proceedings of the 4th International Conference on Speech Prosody (Speech Prosody 2008), Campinas, Brazil* (2008).

39. Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical control of arm movements: a dynamical systems perspective. *Annu. Rev Neurosci* **36**, 337–59 (2013). DOI 10.1146/annurev-neuro-062111-150509.

40. Churchland, M. M. *et al.* Neural population dynamics during reaching. *Nature* **487**, 51–6 (2012). DOI 10.1038/nature11129.

41. Nam, H. *et al.* A procedure for estimating gestural scores from speech acoustics. *The J. Acoust. Soc. Am.* **132**, 3980–3989 (2012).

42. Rubin, P. *et al.* CASY and extensions to the task-dynamic model. In *1st ETRW on Speech Production Modeling: From Control Strategies to Acoustics; 4th Speech Production Seminar: Models and Data, Autrans, France* (1996).

43. Lammert, A., Goldstein, L., Narayanan, S. & Iskarous, K. Statistical Methods for Estimation of Direct and Differential Kinematics of the Vocal Tract. *Speech Commun* **55**, 147–161 (2013). DOI 10.1016/j.specom.2012.08.001.

44. Ramanarayanan, V., Parrell, B., Goldstein, L., Nagarajan, S. & Houde, J. A new model of speech motor control based on task dynamics and state feedback. In *INTERSPEECH*, 3564–3568 (2016).

45. Parrell, B., Ramanarayanan, V., Nagarajan, S. & Houde, J. F. FACTS: A hierarchical task-based control model of speech incorporating sensory feedback. In *Interspeech 2018* (2018).

46. Ringel, R. L. & Steer, M. D. Some Effects of Tactile and Auditory Alterations on Speech Output. *J. Speech, Lang. Hear. Res.* **6**, 369–378 (1963). DOI 10.1044/jshr.0604.369.

47. Scott, C. M. & Ringel, R. L. Articulation without oral sensory control. *J. Speech, Lang. Hear. Res.* **14**, 804–818 (1971).

48. Lane, H. & Webster, J. W. Speech deterioration in postlingually deafened adults. *The J. Acoust. Soc. Am.* **89**, 859–866 (1991). DOI 10.1121/1.1894647.

49. Cowie, R. & Douglas-Cowie, E. *Postlingually acquired deafness: speech deterioration and the wider consequences*, vol. 62 (Walter de Gruyter, 1992).

50. Perkell, J. S. Movement goals and feedback and feedforward control mechanisms in speech production. *J. Neurolinguistics* **25**, 382–407 (2012). DOI 10.1016/j.jneuroling.2010.02.011.

51. Putnam, A. H. B. & Ringel, R. L. A Cineradiographic Study of Articulation in Two Talkers with Temporarily Induced Oral Sensory Deprivation. *J. Speech, Lang. Hear. Res.* **19**, 247–266 (1976). DOI 10.1044/jshr.1902.247.

52. Borden, G. J. The Effect of Mandibular Nerve Block Upon the Speech of Four-Year-Old Boys. *Lang. Speech* **19**, 173–178 (1976). DOI 10.1177/002383097601900208.

53. Desmurget, M. & Grafton, S. Feedback or Feedforward Control: End of a dichotomy. *Tak. action: Cogn. neuroscience perspectives on intentional acts* 289–338 (2003).

54. Gordon, J., Ghilardi, M. F. & Ghez, C. Impairments of reaching movements in patients without proprioception. I. Spatial errors. *J. Neurophysiol.* **73**, 347–360 (1995). DOI 10.1152/jn.1995.73.1.347.

55. Purcell, D. W. & Munhall, K. G. Compensation following real-time manipulation of formants in isolated vowels. *J Acoust Soc Am* **119**, 2288–97 (2006).

56. Villacorta, V. M., Perkell, J. S. & Guenther, F. H. Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *J Acoust Soc Am* **122**, 2306–19 (2007). DOI 10.1121/1.2773966.

57. Lametti, D. R., Nasir, S. M. & Ostry, D. J. Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. *J Neurosci* **32**, 9351–8 (2012). DOI 10.1523/JNEUROSCI.0404-12.2012.

58. Shiller, D. M., Ostry, D. J. & Gribble, P. L. Effects of Gravitational Load on Jaw Movements in Speech. *J. Neurosci.* **19**, 9073–9080 (1999). DOI 10.1523/JNEUROSCI.19-20-09073.1999.

59. Shiller, D. M., Ostry, D. J., Gribble, P. L. & Laboissière, R. Compensation for the Effects of Head Acceleration on Jaw Movement in Speech. *J. Neurosci.* **21**, 6447–6456 (2001).

60. Ostry, D., Gribble, P. & Gracco, V. Coarticulation of jaw movements in speech production: is context sensitivity in speech kinematics centrally planned? *The Journal of Neuroscience* **16**, 1570–1579 (1996).

61. Perrier, P., Payan, Y., Zandipour, M. & Perkell, J. Influence of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *J. Acoust. Soc. Am.* **114**, 1582–1599 (2003).

62. Tremblay, S., Shiller, D. M. & Ostry, D. J. Somatosensory basis of speech production. *Nature* **423**, 866 (2003).

63. Tremblay, S. & Ostry, D. The Achievement of Somatosensory Targets as an Independent Goal of Speech Production -Special Status of Vowel-to-Vowel Transitions. In Divenyi, P., Greenberg, S. & Meyer, G. (eds.) *Dynamics of Speech Production and Perception*, 33–43 (IOS Press, Amsterdam, The Netherlands, 2006).

64. Nasir, S. M. & Ostry, D. J. Somatosensory Precision in Speech Production. *Curr. Biol.* **16**, 1918–1923 (2006). DOI 10.1016/j.cub.2006.07.069.

65. Herzfeld, D. J. & Shadmehr, R. Cerebellum estimates the sensory state of the body. *Trends Cogn Sci* **18**, 66–7 (2014). DOI 10.1016/j.tics.2013.10.015.

66. Shum, M., Shiller, D. M., Baum, S. R. & Gracco, V. L. Sensorimotor integration for speech motor learning involves the inferior parietal cortex. *Eur J Neurosci* **34**, 1817–22 (2011). DOI 10.1111/j.1460-9568.2011.07889.x.

67. Chartier, J., Anumanchipalli, G. K., Johnson, K. & Chang, E. F. Encoding of Articulatory Kinematic Trajectories in Human Speech Sensorimotor Cortex. *Neuron* **98**, 1042–1054.e4 (2018). DOI 10.1016/j.neuron.2018.04.031.

68. Ramanarayanan, V., Goldstein, L. & Narayanan, S. S. Spatio-temporal articulatory movement primitives during speech production: Extraction, interpretation, and validation. *The J. Acoust. Soc. Am.* **134**, 1378–1394 (2013). DOI 10.1121/1.4812765.

69. Kutschireiter, A., Surace, S. C., Sprekeler, H. & Pfister, J.-P. Nonlinear Bayesian filtering and learning: a neuronal dynamics for perception. *Sci. Reports* **7**, 8722 (2017). DOI 10.1038/s41598-017-06519-y.

70. Nam, H., Goldstein, L. & Saltzman, E. Self-organization of syllable structure: a coupled oscillator model. In Pellegrino, F., Marisco, E., Chitoran, I. & Coupé, C. (eds.) *Approaches to phonological complexity*, 299–328 (Mouton de Gruyter, Berlin/New York, 2009).

71. Goldstein, L., Nam, H., Saltzman, E. & Chitoran, I. Coupled oscillator planning model of speech timing and syllable structure. In Fant, G., Fujisaki, H. & Shen, J. (eds.) *Frontiers in phonetics and speech science*, 239–249 (The Commercial Press, Beijng, 2009).

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to B.P. (email: bparrell@wisc.edu).

Support This work was supported by the following grants NIH grants: R01DC013979, R01DC0176960, R01NS100440, and R01DC017091.