Interspeech 2018
2-6 September 2018, Hyderabad

FACTS: A hierarchical task-based control model of speech incorporating sensory feedback

Benjamin Parrell[‡⋆], *Vikram Ramanarayanan*[†§⋆], *Srikantan Nagarajan*[§] *and John Houde*[§]

[†]Educational Testing Service R&D, San Francisco, CA
[§]University of California, San Francisco, CA
[‡]University of Wisconsin, Madison, WI
[⋆]Joint first authors, contributed equally to this work.
vramanarayanan@ets.org, bparrell@wisc.edu

Abstract

We present a computational model of speech motor control that integrates vocal tract state prediction with sensory feedback. This hierarchical model, called FACTS, incorporates both a high-level and low-level controller. The high-level controller orchestrates linguistically-relevant speech tasks, which are represented as desired constrictions along the vocal tract (e.g., closure of the lips). The output of the high-level controller is passed to a low-level controller that can issue motor commands at the level of the speech articulators in order to accomplish the desired constrictions. In order to generate these articulatory motor commands, the low-level articulatory controller relies on an estimate of the current state of the vocal tract. This estimate combines internal predictions about the consequences of issued motor commands with auditory and somatosensory feedback from the vocal tract using an Unscented Kalman Filter based state estimation method. FACTS is able to reproduce several important aspects of human speech behavior such as: (i) stable speech behavior in the presence of noisy motor and sensory systems, (ii) partial acoustic compensation to auditory feedback perturbations, (iii) complete compensations to mechanical perturbations only when they interfere with current production goals, and (iv) the observed relationship between sensory acuity and response to sensory perturbations.

Index Terms: speech motor control, auditory feedback, task dynamics, state feedback control, feedback perturbation, speech production, speech modeling

1. Feedback and Speech Motor Control

Sensory feedback is important for speech motor control. Multiple research studies have shown that speakers compensate for perturbations to auditory and/or somatosensory feedback [1, 2], and that delayed auditory feedback disrupts the production of fluent speech [3, 4]. However, one intriguing aspect of the speech production process is that while it is responsive to auditory and somatosensory feedback, it is not critically dependent on it. We know this because post-lingually deafened adults can produce intelligible speech [5]. In addition, speech is highly intelligible during oral sensory and auditory deprivation, even though articulatory precision is affected [6]. Models of speech motor control therefore need to account for the effects of sensory feedback on the articulation process, without critically relying on it.

We gratefully acknowledge the support of NIH Grants R01DC013979, R01DC010145, R01NS100440 and F32DC014211 and NSF Grant BCS1262297.

A number of speech motor control models have been proposed in recent years, including (among others) the DIVA model [7], Task Dynamics [8], and State Feedback Control or SFC [9]. While both SFC and Task Dynamics have evolved out of a general feedback controller, Task Dynamics has at its heart a controller that generates state-dependent motor commands that drive changes in the speech articulators [8, 10], while assuming that the instantaneous state of the speech production system can be known without error. SFC models how the CNS can estimate the state of the speech production system from noisy, delayed feedback signals [11] using optimal control principles [12], but does not model how that state can be used by the controller to generate motor commands. More recently, we proposed a novel speech production model – TD-SFC [13] – that overcomes the individual disadvantages of each model by composing a neurobiologically inspired short-latency feedback control scheme (derived from State Feedback Control) with the well-developed method for deriving utterance-specific control laws and generating the resulting articulatory and acoustic outcomes (derived from Task Dynamics).

In this paper, we present a more robust and updated version of that model, which we dub FACTS (*Feedback-Aware Control of Tasks in Speech*). We show that the FACTS model can reproduce several important aspects of human speech behavior.

2. Modeling

2.1. Task Dynamics Modeling Preliminaries

The Task Dynamics Application (or TaDA) model [14, 15, 10] implements the Task Dynamic model of inter-articulator speech coordination with the framework of Articulatory Phonology [16]. Based on any arbitrary orthographic (ARPABET) input, TaDA uses a feedback control schema to control a configurable articulatory speech synthesizer [17, 18], generating both articulatory and acoustic output. In TaDA, articulatory control and functional coordination of the speech articulators is accomplished with reference to speech 'tasks' which are coordinated together in time. Speech tasks, or 'gestures', are taken to be constriction actions of the vocal tract (e.g., close the lips), with specific spatial targets and temporal extents. Each gesture controls multiple speech articulators that are used coordinatively to achieve that particular task (e.g., the upper lip, low lip, and jaw move together to close the lips) [16]. Each gesture is modeled as as a point attractor with second-order mass-spring dynamics, which when active forms part of the multi-dimensional control law that governs how the vocal tract changes through time. This time-varying control law, unique to each utterance, is known as a gestural score.

10.21437/Interspeech.2018-2087

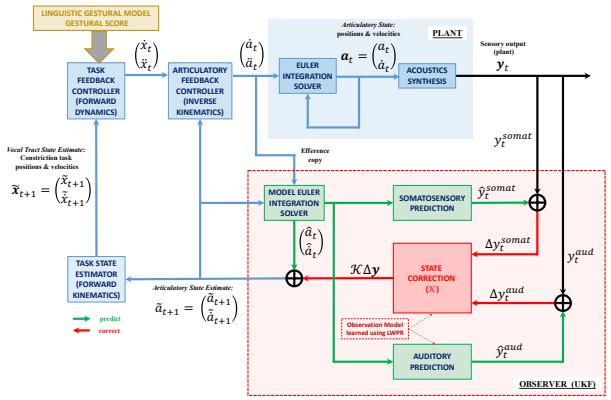

Figure 1: *The proposed FACTS model. This model includes a controller and vocal tract model from TaDA (in blue) and an implementation of SFC-style state estimation (in red). The observer, implemented as a UKF, includes predictive components (green arrows) and mechanisms to correct predictions based on sensory feedback (red arrows).*

However, as in any motor control scheme, we cannot directly control these task variables. Rather, we control the lower level articulators (or, at a level not modeled here, muscles or motor neurons). As such, TaDA generates changes in the positions of the organs of the model vocal tract (*articulatory* variables, $\mathbf{a}$) which can be nonlinearly related to the task variables using the so-called 'direct kinematics' relationship.

2.2. FACTS model

This section extends an earlier version of a task-based state feedback control model [13], TD-SFC (Task Dynamics-State Feedback Control). A schematic control diagram of our current model is shown in Figure 1. The dashed blue boxes replicate the current Task Dynamics model [8]. This model contains 1) a controller, 2) a model vocal tract or plant that receives a motor command from the controller and produces changes in the articulatory state, and 3) a model to generate acoustic output from time-varying articulatory trajectories. Here 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 and their velocities $\dot{x}_t$. Given a gestural score generated using a linguistic gestural model as described earlier, the Forward Task Dynamics model allows us to compute the state derivative $\dot{\mathbf{x}}_\mathbf{t}$ as follows:

$$\begin{bmatrix} \dot{x}_t \\ \ddot{x}_t \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \frac{-K}{M} & \frac{-B}{M} \end{bmatrix} \begin{bmatrix} x_t \\ \dot{x}_t \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{c}{M} \end{bmatrix} \tag{1}$$

where x refers to the task variable (or goal variable) vector, which is defined in TaDA as a set of constriction degrees (such as lip aperture, tongue tip constriction degree, velic aperture, etc.) or locations (such as tongue tip constriction location). M is the mass matrix, B is the damping coefficient matrix, K is the stiffness coefficient matrix of the second-order dynamical system model, and c is a constant.

Next we use Equation 2 (after [8]) 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 $\tilde{\dot{a}}_t$. J is the Jacobian matrix of the forward kinematics model relating articulatory states to task states.

$$x = f(a) \tag{2a}$$

$$\dot{x} = J(a)\dot{a} \tag{2b}$$

$$\ddot{x} = J(a)\ddot{a} + \dot{J}(a, \dot{a})\dot{a} \tag{2c}$$

Euler integration allows us to compute the model articulator positions and velocities for the next time-step, which effectively "moves" the articulatory vocal tract model. Then, an appropriate synthesis model converts the model articulator and constriction task values into output acoustic parameters y_t.

While Task Dynamics assumes perfect observability and feedback of the current vocal tract state at every iteration of the model (represented by the dotted blue arrow in Figure 1), which is unrealistic for the human CNS given the variety of reasons discussed above. We implement a state-estimation procedure (Section 2.3, below), to estimate the articulatory state from an efference copy of the motor commands issued to the plant and sensory feedback.

This articulatory state estimate is passed back to the articulatory feedback controller and used to estimate the current state of the speech tasks $\tilde{\mathbf{x}}_\mathbf{t}$. This task state estimate is calculated by running the forward kinematics model f (see Equation 2) based on the estimate of articulatory state $\tilde{\mathbf{a}}_\mathbf{t}$, and the output of this process is passed to the task feedback controller.

2.3. State estimation

The basic concept of SFC is that a copy of the motor command ("efference copy") is passed to an internal model of the vocal tract. Based on this efference copy, the internal model generates 1) an estimate of the next state of the speech articulators and 2) an estimate of the sensory consequences of the estimated state.

In FACTS, this forward modeling of articulatory state and sensory consequences is accomplished through an Unscented Kalman Filter (UKF) [19]. The UKF is an extension of the principles of the Kalman Filter to nonlinear systems that has been shown to be more stable and accurate than the method we previously emplyed [13], the Extended Kalman Filter [20]. In order to generate a posterior mean and covariance, the EKF approximates a non-linear transformation function and projects a single prior throught that linearized function. However, the approximations in this process can sometimes lead to sub-optimal performance. In a UKF, multiple prior points (called sigma points, $\mathcal{X}$) 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 true un-transformed nonlinear function, after which the posterior mean and covariance can be calculated from the transformed points. This process is called the unscented transform. This is used both to predict the future state of the system (process model) as well as the expected sensory feedback (observation model). The means and covariances calculated through these unscented transforms are then used analogously to their use in a standard Kalman Filter to estimate the optimal posterior state.

In our model, the output of the inverse kinematics model ($\ddot{a}_t$, which is equivalent to the motor command) is passed to the observer/UKF. This is combined with an estimate of the current articulatory state $\mathbf{a_{t-1}} = [a_{t-1}\ \dot{a}_{t-1}]^T$ to generate an articulatory state prediction. First, the sigma points ($\mathcal{X}$) are generated:

$$\mathcal{X}_{t-1} = [\hat{\mathbf{s}}_\mathbf{t-1} \pm \sqrt{(L+\lambda)\mathbf{P_{t-1}}}] \tag{3}$$

where $\hat{\mathbf{s}}_\mathbf{t-1} = [\mathbf{a_{t-1}^T}\,\mathbf{v_{t-1}^T}\,\mathbf{n_{t-1}^T}]^\mathbf{T}$), and $\mathbf{v}$ and $\mathbf{n}$ are the process and observation noise, 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 $\mathbf{n}$.

The observer then estimates how the motor command $\ddot{a}_t$ would effect the speech articulators by replicating using the

Euler integration model ($\mathcal{F}$) to generate the state prediction $\hat{\mathbf{a}}_\mathbf{t} = [\hat{a}_t\ \hat{\dot{a}}_t]^T$. First, all sigma points reflecting the articulatory state $\mathcal{X}^a$ and process noise $\mathcal{X}^v$ are passed through $\mathcal{F}$:

$$\mathcal{X}^a_{t|t-1} = F[\mathcal{X}^a_{t-1}, \ddot{a}_{t-1}, \mathcal{X}^v_{t-1}] \tag{4}$$

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.

$$\hat{\mathbf{a}}_\mathbf{t} = \sum_{i=0}^{2L} W_i\ \mathcal{X}^a_{i,t|t-1} \tag{5}$$

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 $\mathcal{X}^a$ and observation noise $\mathcal{X}^n$ sigma points through the articulatory-to-sensory transform $\mathcal{H}$.

$$\mathcal{Y}_{t|t-1} = \mathcal{H}\ (\mathcal{X}^a_{t|t-1}, \mathcal{X}^n_{t-1}) \tag{6}$$

$$\hat{\mathbf{y}}_\mathbf{t} = \sum_{i=0}^{2L} W_i\ \mathcal{Y}_{i,t|t-1} \tag{7}$$

In the current version of the model, the sensory state ($\mathbf{y_t}$) includes both auditory feedback ($\mathbf{y_t^{aud}}$) as well as somatosensory feedback ($\mathbf{y_t^{somat}}$), which has been show to play a role in informing the state of the system [21]. Acoustic feedback is implemented as the values, in Hz, of the first three vowel formants (F1-F3). Somatosensory feedback is implemented as the positions and the velocities of the oral articulators in the CASY synthesizer.

This estimate of the sensory state is then compared against incoming sensory feedback ($\mathbf{y}$) to adjust the predicted articulatory state. To model sensory noise, Gaussian white noise (ω) is added to the formant values and articulator positions and velocities produced by the CASY with separate standard deviations for auditory and somatosensory signals . The updated state estimate $\tilde{\mathbf{a}}_\mathbf{t}$ in this case is given by:

$$\tilde{\mathbf{a}}_\mathbf{t} = \hat{\mathbf{a}}_\mathbf{t} + \mathcal{K}_t(\mathbf{y_t} - \hat{\mathbf{y}}_\mathbf{t}) \tag{8}$$

where $\mathbf{\Delta y} = \mathbf{y_t} - \hat{\mathbf{y}}_\mathbf{t}$ is the sensory error and $\mathcal{K}_t$ is the Kalman Gain, which is computed as a function of the posterior covariance matrices $\mathbf{P_{x_t y_t}}$ and $\mathbf{P_{y_t y_t}}$ in the following manner:

$$\mathcal{K}_t = \mathbf{P_{x_t y_t}} \mathbf{P^{-1}_{y_t y_t}} \tag{9}$$

$$\mathbf{P_{x_t y_t}} = \sum_{i=0}^{2L} W_i[\mathcal{X}_{i,t|t-1} - \hat{\mathbf{a}}_\mathbf{t}][\mathcal{Y}_{i,t|t-1} - \hat{\mathbf{y}}_\mathbf{t}]^T \tag{10}$$

$$\mathbf{P_{y_t y_t}} = \sum_{i=0}^{2L} W_i[\mathcal{Y}_{i,t|t-1} - \hat{\mathbf{y}}_\mathbf{t}][\mathcal{Y}_{i,t|t-1} - \hat{\mathbf{y}}_\mathbf{t}]^T \tag{11}$$

One of the challenges in implementing such an UKF is that both the process model F (that provides a functional mapping from $[a_{t-1}\ \dot{a}_{t-1}\ \ddot{a}_{t-1}]^T$ to $\hat{\mathbf{a}}_\mathbf{t}$) as well as the observation model $\mathcal{H}$ (that maps from $\mathbf{a_t}$ to $\mathbf{y_t}$) are unknown. Currently, we implement the process model $\mathcal{F}$ by replicating the Euler integration equations used to drive changes in the CASY model. 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 ($y_t^{aud} = \mathcal{H}(\hat{\mathbf{a}}_\mathbf{t})$)required for Unscented Kalman Filtering using Locally Weighted Projection Regression, or LWPR, a computationally efficient machine learning technique [22]. 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 [7]. Currently, we learn only the auditory prediction component of $\mathcal{H}$. 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 ($y_t^{somat} = \hat{\mathbf{a}}_\mathbf{t}$).

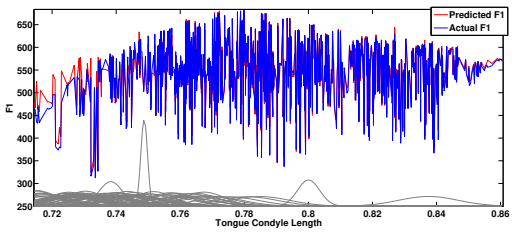

Figure 2: *Predicted F1 produced by running the LWPR model of the articulatory-to-sensory transform and visualizing it in the 2D plane of F1–Tongue Condyle Length (CL). We also plot Gaussians corresponding to the mean and variance of all LWPR receptive fields (below a certain variance threshold) to illustrate how LWPR models different regions of the CL space.*

2.4. Model training

We used the TaDA model [15] (which given our current use of many of that model's components, is essentially equivalent to the current FACTS model without the SFC component) to generate a set of 2938 vowel sweeps covering the extent possible of tongue body movements. Half of these productions started at the neutral tongue position for the vowel [ə] and moved the tongue body to some other location in the vocal tract. The range of end positions covered the full extent of allowable tongue body positions in the model. The other half of the productions reversed these starting and ending points.

We then extracted 20-dimensional articulatory variable trajectories(corresponding to the positions (a) and velocities ($\dot{a}$) of the oral articulators in the CASY model [18]), 10-dimensional articulatory control signals issued to the plant ($\ddot{a}$) and 3-dimensional acoustic variable trajectories (corresponding to the first three formants, F1 - F3) from this dataset, and used this to train the mappings for the process and observation model using LWPR. Figure 2 shows an example of predicted v.s actual F1 (in red and blue, respectively), for different values of one particular dimension of the articulatory parameter, "tongue condyle length" (CL). We also plot, in gray, Gaussians that are representative of the mean and variance of the learned LWPR receptive fields (visualized in this 2D plane) that are used to make the predictions. Notice that the predicted F1 pretty closely matches the actual F1, particularly towards the upper end of the range of CL values.

3. Simulation Experiments

3.1. Stability of the model with sensory noise

One of the key factors that contribute to the instability of systems that rely purely on feedback control is the presence of noise in the sensory processes that monitor the motor output of those systems. In humans, the auditory and somatosensory feedback pathways, like all neural signals, are corrupted by some amount of noise [23]. Given the role feedback pathways play in the FACTS model, test the effects of sensory noise on stability by simulating production of [ə], varying the standard deviation of the additive sensory noise in the auditory and somatosensory feedback signals between 1% and 50% of the ob-

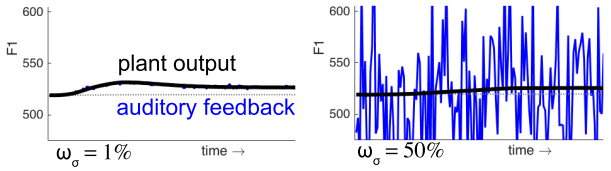

Figure 3: *Example runs of the FACTS model with different degrees of additive sensory noise. Note the feedback is barely visible in the left figure due to the low amount of noise.*

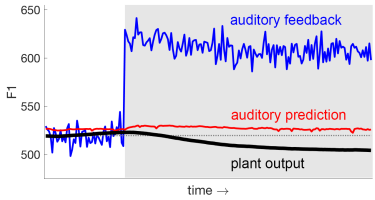

Figure 4: *An example simulation run of the FACTS model with altered auditory feedback. A perturbation of +100 Hz is applied to F1 during the shaded time period.*

served ranges in our training data. As shown in Figure 3, the model is able to produce stable output (though with some run-to-run variability) regardless of the amount of sensory noise, due to the state prediction component of the UKF (only endpoints shown).

3.2. Response of model to altered feedback

While the speech motor system can operate independently of sensory feedback, there is strong evidence that the system does use acoustic feedback, when available, to control ongoing speech production [24, 25, 26]. When subjects' speech formants are perturbed they compensate by shifting their own formants in the opposite direction (e.g. a positive shift in F1 played back to the subject induces a negative F1 shift in the subject's production). In order to evaluate the ability of the FACTS model to reproduce this compensatory response to auditory perturbations, we simulated a simple altered auditory feedback experiment: we perturbed F1 of the vocal tract acoustic output y_t^{aud} by 100 Hz while the model was producing [ə].

Figure 4 shows the results of one example simulation run of the FACTS model under F1 perturbation. The model initially starts with veridical (though noisy) feedback. At the onset of the shaded region, the perturbation is turned on. This causes a discrepancy between the perceived auditory feedback and the auditory predictions generated by the UKF. In response to this feedback error, the produced F1 lowers below the baseline value of 520 Hz, though not enough to fully compensate for the 100 Hz perturbation. This partial compensation qualitatively replicates human behavior in previous studies on altered auditory feedback perturbations. Critically, FACTS is able to replicated the compensatory behavior seen in response to auditory perturbations despite the absence of any explicit auditory goals in the model.

3.3. Consequences of sensory acuity

While the simulations in Section 3.1 showed that FACTS is stable in the presence of noise, the amount of noise in the sensory input does influence model behavior. Specifically, more noise for a given sensory channel results in a a smaller weight in the Kalman gain for that channel. This can be seen when an +100 Hz perturbation is applied to the auditory feedback in the model

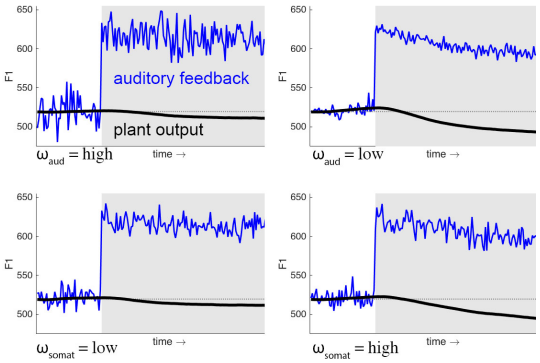

Figure 5: *Four simulation run of the FACTS model with altered auditory feedback and varying sensory noise. A perturbation of +100 Hz is applied to F1 during the shaded time period. A larger response is produced when auditory noise is low or when somatonsensory noise is high.*

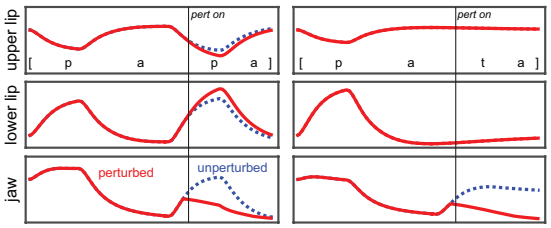

Figure 6: *Simulation run of the FACTS model with a mechanical pertubation applied to the jaw.*

(Figure 5). More noise results in a smaller compensatory response. This simulation result mirrors behavioral findings that speakers who produce smaller compensatory responses to auditory perturbations have less acute auditory systems [27]. The opposite change is observed when somatosensory noise is manipulated: a lower somatosensory noise results in a smaller compensatory response. This suggests a trading relationship between audiotory and somatosensory acuity, consistent with a similar trade-off in auditory vs somatosensory compensation in human behavior [28].

3.4. Responses to mechanical perturbations

When mechanical perturbations are applied to the jaw during a consonant closure, other articulators move to compensate for the lowered jaw position. The response of these articulators depends on the current production goal [29]. For example, the upper lip will lower in response to a jaw perturbation during /p/ but not /f/ [30]. We replicated these experiments by applying a downward force to the model jaw. FACTS qualitatively replicates human behavior, as seen in Figure 6. The upper and lower lip move to a compensate for the lower jaw only when needed to produce a bilabial /p/ (left), but not when producing a coronal /t/ (right).

4. Summary

We have elaborated the computational architecture of a new model of speech production, FACTS. We have shown that this model, though still under development, is capable of producing speech in the presence of sensory noise or even complete absence of sensory feedback, yet is able to use sensory information to correct for auditory and mechanical perturbations, producing corrective responses that qualitatively match human behavior.

5. References

[1] J. A. Jones and K. G. Munhall, "Learning to produce speech with an altered vocal tract: The role of auditory feedback," *The Journal of the Acoustical Society of America*, vol. 113, no. 1, pp. 532–543, 2003.

[2] S. Tremblay, D. M. Shiller, and D. J. Ostry, "Somatosensory basis of speech production," *Nature*, vol. 423, no. 6942, p. 866, 2003.

[3] B. S. Lee, "Effects of delayed speech feedback," *The Journal of the Acoustical Society of America*, vol. 22, no. 6, pp. 824–826, 1950.

[4] A. J. Yates, "Delayed auditory feedback." *Psychological bulletin*, vol. 60, no. 3, p. 213, 1963.

[5] R. Cowie and E. Douglas-Cowie, *Postlingually acquired deafness: speech deterioration and the wider consequences*. Walter de Gruyter, 1992, vol. 62.

[6] C. M. Scott and R. L. Ringel, "Articulation without oral sensory control," *Journal of Speech, Language, and Hearing Research*, vol. 14, no. 4, pp. 804–818, 1971.

[7] J. A. Tourville and F. H. Guenther, "The diva model: A neural theory of speech acquisition and production." *Lang Cogn Process*, vol. 26, no. 7, pp. 952–981, 1 2011.

[8] E. Saltzman and K. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, 1989.

[9] J. F. Houde and S. S. Nagarajan, "Speech production as state feedback control," *Front Hum Neurosci*, vol. 5, p. 82, 2011.

[10] H. Nam, V. Mitra, M. Tiede, M. Hasegawa-Johnson, C. Espy-Wilson, E. Saltzman, and L. Goldstein, "A procedure for estimating gestural scores from speech acoustics," *The Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3980–3989, 2012.

[11] O. L. R. Jacobs, *Introduction to control theory*. Oxford ; New York: Oxford University Press, 1993.

[12] E. Todorov and M. I. Jordan, "Optimal feedback control as a theory of motor coordination," *Nature neuroscience*, vol. 5, no. 11, p. 1226, 2002.

[13] V. Ramanarayanan, B. Parrell, L. Goldstein, S. Nagarajan, and J. Houde, "A New Model of Speech Motor Control Based on Task Dynamics and State Feedback," in *Interspeech 2016*, 2016.

[14] H. Nam, L. Goldstein, C. Browman, P. Rubin, M. Proctor, and E. Saltzman, "TADA (TAsk Dynamics Application) manual," *Haskins Laboratories Manual, Haskins Laboratories, New Haven, CT (32 pages)*, 2006.

[15] E. Saltzman, H. Nam, J. Krivokapic, and L. Goldstein, "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.

[16] C. Browman and L. Goldstein, "Dynamics and articulatory phonology," in *Mind as motion: Explorations in the dynamics of cognition*, R. Port and T. van Gelder, Eds. Boston: MIT Press, 1995, pp. 175–194.

[17] P. Rubin, E. Saltzman, L. Goldstein, R. McGowan, M. Tiede, and C. Browman, "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.

[18] K. Iskarous, L. Goldstein, D. Whalen, M. Tiede, and P. Rubin, "CASY: The Haskins configurable articulatory synthesizer," in *International Congress of Phonetic Sciences, Barcelona, Spain*, 2003, pp. 185–188.

[19] E. A. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*. Ieee, 2000, pp. 153–158.

[20] S. S. Haykin, S. S. Haykin, and S. S. Haykin, *Kalman filtering and neural networks*. Wiley Online Library, 2001.

[21] T. Ito, M. Tiede, and D. J. Ostry, "Somatosensory function in speech perception," *Proceedings of the National Academy of Sciences*, vol. 106, no. 4, pp. 1245–1248, 2009.

[22] D. Mitrovic, S. Klanke, and S. Vijayakumar, "Adaptive optimal feedback control with learned internal dynamics models," in *From Motor Learning to Interaction Learning in Robots*. Springer, 2010, pp. 65–84.

[23] A. A. Faisal, L. P. J. Selen, and D. M. Wolpert, "Noise in the nervous system," *Nature Reviews Neuroscience*, vol. 9, no. 4, pp. 292–303, Apr. 2008.

[24] D. W. Purcell and K. G. Munhall, "Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation," *The Journal of the Acoustical Society of America*, vol. 120, no. 2, pp. 966–977, 2006.

[25] J. A. Tourville, K. J. Reilly, and F. H. Guenther, "Neural mechanisms underlying auditory feedback control of speech," *Neuroimage*, vol. 39, no. 3, pp. 1429–1443, 2008.

[26] C. A. Niziolek, S. S. Nagarajan, and J. F. Houde, "What does motor efference copy represent? evidence from speech production," *The Journal of Neuroscience*, vol. 33, no. 41, pp. 16 110–16 116, 2013.

[27] V. M. Villacorta, J. S. Perkell, and F. H. Guenther, "Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception." *J Acoust Soc Am*, vol. 122, no. 4, pp. 2306–19, 2007.

[28] D. R. Lametti, S. M. Nasir, and D. J. Ostry, "Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback." *J Neurosci*, vol. 32, no. 27, pp. 9351–8, 2012.

[29] J. A. Kelso, B. Tuller, E. Vatikiotis-Bateson, and C. A. Fowler, "Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures." *J Exp Psychol Hum Percept Perform*, vol. 10, no. 6, pp. 812–32, 1984.

[30] S. Shaiman and V. L. Gracco, "Task-specific sensorimotor interactions in speech production," *Experimental Brain Research*, vol. 146, no. 4, pp. 411–418, 2002-10-01.