A New Model of Speech Motor Control based on Task Dynamics and State Feedback

Vikram Ramanarayanan[1†], *Benjamin Parrell*[2†], *Louis Goldstein*[3], *Srikantan Nagarajan*[4] *and John Houde*[5]

[1]Educational Testing Service R&D, San Francisco, CA
[2]Department of Linguistics & Cognitive Science, University of Delaware, Newark, DE
[3]Department of Linguistics, University of Southern California, Los Angeles, CA
[4]Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA
[5]Department of Otolaryngology, University of California, San Francisco, CA
[†]Both authors contributed equally to this work.
vramanarayanan@ets.org, parrell@udel.edu

Abstract

We present a new model of speech motor control (TD-SFC) based on articulatory goals that explicitly incorporates acoustic sensory feedback using a framework for state-based control. We do this by combining two existing, complementary models of speech motor control – the Task Dynamics model [1] and the State Feedback Control model of speech [2]. We demonstrate the effectiveness of the combined model by simulating a simple formant perturbation study, and show that the model qualitatively reproduces the behavior of online compensation for unexpected perturbations reported in human subjects.

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

1. Introduction

One of the outstanding issues in speech research is how the central nervous system controls the rapid, complex articulatory movements produced while speaking. A number of models have been proposed to account for this extraordinarily facile motor behavior, including (among others) the DIVA model [3], Task Dynamics [1], and State Feedback Control [2].

DIVA combines articulatory and acoustic speech synthesis with a neurologically-specified production model. DIVA, however, is only a trajectory-based kinematic model that does not account for dynamics of the articulatory process. In contrast, TADA is a dynamic model of the vocal articulatory process. However, TADA does not include a model of how auditory feedback is used in speech. State Feedback control (SFC) is an observer based model that postulates that sensory feedback is used to update estimates of the current states of the vocal tract using a state correction process. However, SFC model does not clearly specify the dynamic process underlying vocal articulation.

We propose a novel speech production model, the Task Dynamics-State Feedback Control (TD-SFC) model, combining a neuro-biologically inspired short-latency feedback control scheme (derived from State Feedback Control) coupled with a well-developed method for deriving utterance specific control laws and generating the resulting articulatory and acoustic outcomes (derived from Task Dynamics). We first describe the model, then show its utility in qualitatively modeling human behavior in response to unexpected acoustic perturbations.

2. Modeling Background

Both Task Dynamics and State Feedback control are elaborations of a basic type of control system widely used in modeling a variety of motor control domains: feedback control. In an ideal feedback control system (Figure 1), a **controller** that generates a motor command u_t based on a control law U and the current state of the system x_t. This motor command is passed to the physical **plant** (for speech, the vocal tract), which generates a change in the state of the system based on the motor command, the current state of the system, and the internal dynamics of the plant itself. This new state x_{t+1} is used by the controller to generate the next motor command x_{t+1}.

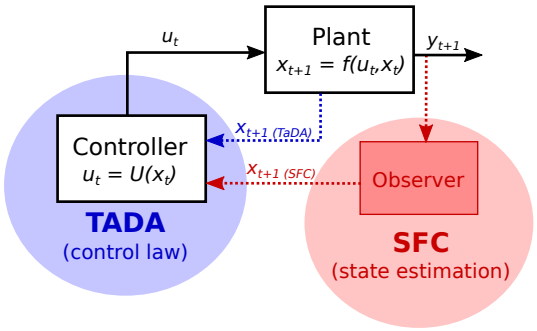

Figure 1: A basic state feedback control system, showing the shared architecture behind both Task Dynamics and SFC. A controller generates a state-dependent motor command u, which is sent to the plant to generate both changes in the state of the system x and sensory output y. Both Task Dynamics and SFC are versions of such a system for speech motor control. Task Dynamics has focused on developing the appropriate control law while assuming direct knowledge of the system state (ideal state feedback, blue line). SFC has focused on how the central nervous system can estimate the current state of the production system from sensory feedback, which is then passed to the controller in lieu of direct state knowledge (the observer, shown in red).

Such an ideal feedback controller has two principal problems. First, the human central nervous system (CNS) cannot know the current state of the production system–the only information available to the CNS is sensory information. Sensory information is problematic as it 1) reflects the consequences (auditory, somatosensory, etc) of the state of the production system rather than the state of the system itself 2) is corrupted by varying amounts of neural noise and 3) is delayed in time relative

to the true state. SFC was developed to address these issues by modeling a way the CNS could optimally estimate the state of the production system from noisy, delayed sensory signals [2].

The second problem with the ideal feedback control system described above is that the control law governing how the system changes is unknown. At least for speech, any such controller must be quite complex. Task Dynamics, and particularly the Task Dynamics Application (TaDA), provides a model of this control law, and is able to generate state-dependent motor commands that drive changes in the speech articulators [1, 4].

While both SFC and Task Dynamics have evolved out of a general feedback controller, they have focused on refining entirely different parts of the model. Task Dynamics has developed the controller, while assuming that the instantaneous state of the speech production system can be known without error. SFC has modeled the how the CNS can estimate the state of the system, but put off understanding how that state could be used by the controller to generate motor commands.

2.1. Task Dynamics Application (TaDA)

The Task Dynamics Application (or TaDA) software [5, 6, 4] implements the Task Dynamic model of inter-articulator speech coordination with the framework of Articulatory Phonology [7]. Based on any arbitrary orthographic (ARPABET) input, TaDA uses a feedback control schema to control a configurable articulatory speech synthesizer [8, 9], 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) [7].

From a particular utterance, the relevant tasks are selected by first by converting the orthographic input to the model to a phonetic string using a version of the Carnegie Mellon pronouncing dictionary that also provides syllabification. The syllabified string is then parsed into articulatory gestures and a langauge-specific model of coordination between gestures is used to generate the temporal activations of each gesture in an utterance, known as a gestural score.

The gestural score represents the Task Dynamic control law that governs the behavior of the the model. 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. Changes in the vocal tract model are further mapped to the vocal tract area function.

The Task Dynamics model of speech articulation is as follows (after [1]):

$$M\ddot{x} + B\dot{x} + Kx = u \tag{1}$$

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

$$\dot{x} = J(a)\dot{a} \tag{3}$$

$$\ddot{x} = J(a)\ddot{a} + \dot{J}(a, \dot{a})\dot{a} \tag{4}$$

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, and K is the stiffness coefficient matrix of the second-order dynamical system model. u is a control input, while J is the Jacobian matrix of the Forward Model.

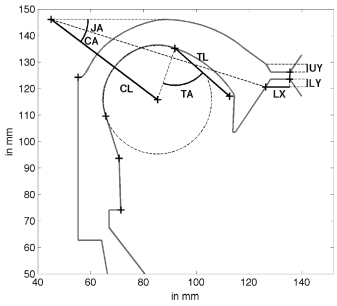

Figure 2: A visualization of the Configurable Articulatory Synthesizer (CASY) in a neutral position, showing the outline of the vocal tract model. Overlain are the key points (black crosses) and geometric reference lines (dashed lines) used to define the model articulator parameters (black lines and angles), which are: lip protrusion (LX), vertical displacements of the upper lip (UY) and lower lip (LY) relative to the teeth, jaw angle (JA), tongue body angle (CA), tongue body length (CL), tongue tip length (TL), and tongue angle (TA).

However, 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 of the positions of the organs of the model vocal tract (*articulatory* variables, a) which can be nonlinearly related to the task variables using the so-called 'direct kinematics' relationship.

2.2. The State Feedback Model of Motor Control

State feedback control (SFC) is the combination of a control law acting on a state estimate provided by an observer [2]. It is so named because if the state x_t of the vocal tract was available to the CNS via immediate feedback, then the CNS could control x_t directly via feedback control. In other words, a fundamental principle of SFC is that control must be based on a internal estimate of the state x_t because x_t is not directly observable from any type of sensory feedback, and furthermore, the sensory feedback that comes to the higher CNS is both noisy and delayed [10].

The speech production SFC model is a neurally plausible formalization of this concept, which includes an observer that ideally estimates the state of the vocal tract. Based on a copy of the motor command, the observer estimates the future state of the vocal tract, generates the predicted sensory consequences of that predicted state, compares the predicted sensory expectations with actual sensory feedback to generate a sensory error, and corrects the state prediction if any error is found.

3. Proposed Model: TD-SFC

In order to address the shortcomings of the current implementations of both Task Dynamics and State Feedback Control, we have developed a hybrid model that combines the insights from each model into a larger feedback control framework. A schematic of the current model, TD-SFC (Task Dynamics-State Feedback Control), is shown in Figure 3.

The dashed blue box replicates the current Task Dynamics model. 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 both articulatory and acoustic state, and 3) a model to generate acoustic output from time-varying articulatory trajectories. Note that the gestural score, part of the controller, essentially defines an utterance-specific, time-varying control law that determines how the state of the system will change in a state-dependent manner. Here we rep-

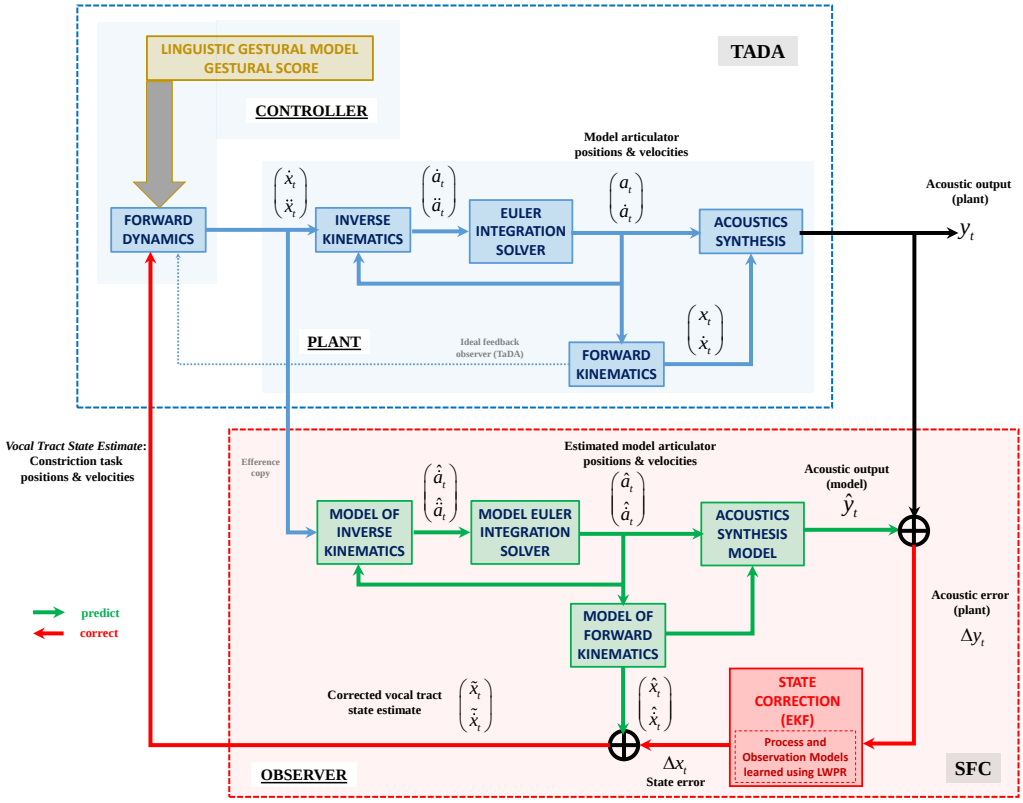

Figure 3: The proposed hybrid TD-SFC model. This model, based on the shared feedback control architecture of Task Dynamics and SFC, includes the controller and vocal tract model from TaDA (in blue) and an implementation of SFC-style state estimation (in red). The observer includes predictive components (green arrows) and mechanisms to correct predictions based on sensory feedback (red arrows).

resent the state of the vocal tract $\mathbf{x_t} = [x_t \dot{x}_t]^T$ 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:

$$\begin{bmatrix} \dot{x}_t \\ \ddot{x}_t \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \frac{-K}{M} & \frac{-B}{M} \end{bmatrix} \begin{bmatrix} x_t \\ \dot{x}_t \end{bmatrix} + \begin{bmatrix} 0 \\ u \end{bmatrix} \tag{5}$$

Next we use Equation 2 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 state of the plant (output from forward kinematics). 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. We then arrive at the corresponding state of the vocal tract at the task level by running the forward kinematics model f (see Equation 2). Finally, an appropriate synthesis model converts the model articulator and constriction task values into output acoustic parameters y_t.

Note that the current version of TaDA 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 3), which is unrealistic for the human CNS given the variety of reasons discussed above. Incorporating state feedback as prescribed by the SFC model (red box in Figure 3) allows us to overcome this limitation.

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 vocal tract and 2) an estimate of the sensory consequences of the estimated state. In our model, the output of the forward dynamics model ($\ddot{x}_t$, which is equivalent to the motor command) is passed to the internal model. The observer then estimates how this motor command would effect the speech articulators by replicating the inverse kinematic model (generating $\hat{\ddot{a}}_t$) and the Euler integration model (generating $\hat{a}_t, \hat{\dot{a}}_t$). The expected acoustic state ($\hat{y}_t$) is then derived based on the predicted articulatory state.

Crucially, the SFC model also evaluates the acoustic sensory error Δy_t between the estimated model acoustics $\hat{y}_t$ and the actual acoustics y_t and passes this value to a state estimator that computes a state error correction estimate $e_t = \Delta x_t$. In our current implementation of the model, we use an Extended Kalman Filter (EKF) [11] to perform the state estimation. The updated state estimate $\Delta x_{t|t}$ in this case is given by:

$$\Delta x_{t|t} = \Delta x_{t|t-1} + K_t \Delta \tilde{y}_t \tag{6}$$

where $\Delta \tilde{y}_t$ is the innovation or measurement residual and K_t is the Kalman Gain, which is computed in the following manner:

$$K_t = P_{t|t-1} J_{h_t}^T (J_{h_t} P_{t|t-1} J_{h_t}^T + R_t)^{-1} \tag{7}$$

$$P_{t|t-1} = J_{f_t} P_{t-1|t-1} J_{f_t}^T + Q_t \tag{8}$$

where J_{f_t} represents the Jacobian of the process model $\mathcal{F}$ at time t, J_{h_t} is the Jacobian of the observation model $\mathcal{H}$ at time t, and $P_{t|t-1}$ refers to the predicted covariance estimate. Q_t and R_t indicate process and observation noise, respectively.
Note that one of the challenges in implementing such an EKF is that both the process model $\mathcal{F}$ (that provides a functional mapping from Δx_{t-1} to Δx_t) as well as the observation model $\mathcal{H}$ (that maps from Δx_t to Δy_t) are unknown. In order to solve this problem, we *learn* the process model and observation model functional mappings required for Extended Kalman Filtering using Locally Weighted Projection Regression, or LWPR, a computationally efficient machine learning technique [12]. 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 [3].

We used the current version of TaDA (without SFC-style state estimation) to generate 972 vowel-consonant-vowel (or VCV) sequences corresponding to all combinations of 9 English monophthongs and 12 consonants (including stops, fricatives, nasals and approximants). We then extracted 10-dimensional constriction task variable trajectories and 4-dimensional acoustic variable trajectories (corresponding to the first four formants, F1 - F4) from this dataset, and used this to train the mappings for the process and observation model using LWPR.

Once the process and observation models are trained, the Extended Kalman Filter produces estimates of the state error $e_t = \Delta x_t$, which is subtracted from the current state estimate $\mathbf{x_t}$ to produce a corrected vocal tract state estimate, $\tilde{\mathbf{x}}_\mathbf{t}$, which is in turn fed back to the controller. All motor commands in this system are generated based on this *estimated state*, rather than the actual state of the system.

4. Simulation Experiments

4.1. Response of model to altered feedback

One of the strongest pieces of evidence that the speech-production system uses acoustic feedback to control ongoing speech comes from studies which perturb the spectral components of speech in real time [13, 14, 15]. In these studies, subjects repeatedly produce either a single word with an extended vowel or a sustained vowel while listening to feedback of their own voice played back in real time via headphones. On a random subset of trials, their speech formants are perturbed (either F1 alone or F1 and F2). Subjects compensate somewhat, though not completely, for this unexpected perturbation 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). These compensations generally begin roughly 200 ms after the onset of the perturbation or, for experiments with continuous perturbation throughout the production of a word, the vowel onset.

These results are generally taken as evidence that vowels have an explicit acoustic target [16]. We hypothesized that this compensatory behavior could, alternatively, be produced by a system with articulatory, rather than explicitly acoustic, goals. Although the targets in such a system might be in articulatory space, the actual articulatory state of the system cannot be directly known; rather, the current state must be estimated from 1) the expected outcome of produced motor command and 2) sensory feedback. In such a system, a given motor command issued to achieve a particular articulatory task would generate a *sensory feedback expectation*, which could then be compared with incoming sensory feedback. Any discrepancy between the expected and actual sensory feedback would generate a sensory error, which could be used to update the estimate of the current articulator state. Thus, introducing an error in auditory space could cause the speaker to (incorrectly) estimate the state of their articulatory system, leading to corrective movements.

4.2. Simulation and results

In order to study the behavior of the TD-SFC model, we simulated a simple altered auditory feedback experiment. We perturbed F1 of the vocal tract acoustic output y_t by 100Hz. Since the model acoustic output $\hat{y}_t$ is unperturbed, this should result in a large feedback error. We would then expect the model to compensate for this large feedback error by lowering F1.

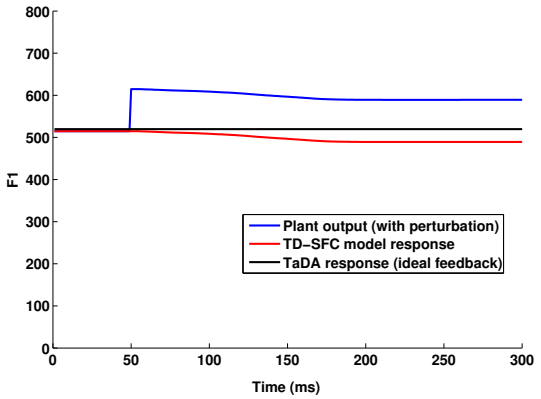

Figure 4: An example simulation run of the TD-SFC model. A perturbation of +100Hz is applied to F1 at 50ms.

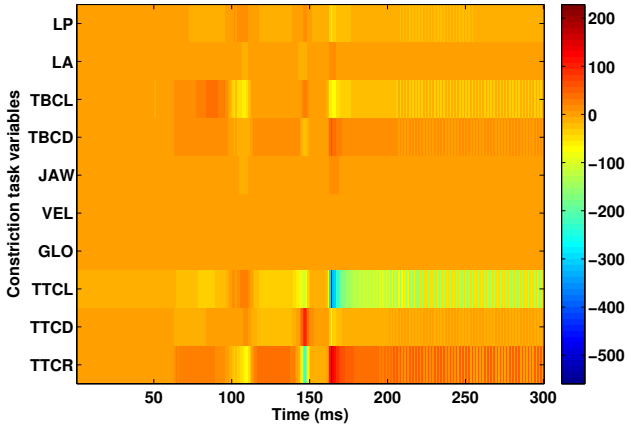

Figure 5: A visualization of how different constriction task variables x_t affect F1 over the course of the simulation. Colors represent the values of the Jacobian matrix J_{h_t} that corresponds to how F1 is affected by each of the 10 elements of x_t at each time step t. The task variables shown are: LP,LA-Lip Protrusion and Aperture, TBCL,TBCD-Tongue Body Constriction Location and Degree, JAW, VEL-Velic Aperture, GLO-Glottal Aperture & TTCL,TTCD,TTCR-Tongue Tip Constriction Location, Degree and Radius.

Figure 4 shows the results of one example simulation run of the TD-SFC model for this experiment. We observe that after perturbing the plant F1 (in blue) by 100Hz at $t = 50ms$, the model F1 (red) trajectory displays compensatory behavior by lowering F1 below the baseline of 500Hz, which is in line with previous studies on altered auditory feedback perturbations. Figure 5, which visualizes the entries in the observation Jacobian matrix J_{h_t} that affect F1 in particular, allows us to observe that the main constriction task variables which affect F1 include the tongue tip constriction location and degree.

5. Summary

We have proposed a new speech production model, TD-SFC, based on the shared feedback control architecture of prior versions of Task Dynamics and State Feedback Control models. We have shown that this model, though still under development, is capable of correcting for sensory perturbations, producing corrective responses that qualitatively match human behavior, despite having no explicit acoustic or auditory target.

6. Acknowledgements

Thanks to audiences at Haskins Laboratories, and in particular Elliot Saltzman, for comments on early versions of the model. We also gratefully acknowledge the support of NIH Grants R01DC013979, R01DC010145 and F32DC014211 and NSF Grant BCS1262297.

7. References

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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] P. Perrier and S. F. Fuchs, "Motor equivalence in speech production," in *The Handbook of Speech Production*, M. Redford, Ed. Hoboken, NJ: Wiley-Blackwell, 2015.