Current Speech Motor Control Models: An Overview of Architectures & Properties[1][2]

Benjamin Parrell[†]

Department of Linguistics & Cognitive Science, University of Delaware, Newark, Delaware, USA

Adam C. Lammert[3†]

Bioengineering Systems & Technologies, MIT Lincoln Laboratory, Lexington, Massachusetts, USA

Gregory Ciccarelli

Bioengineering Systems & Technologies, MIT Lincoln Laboratory, Lexington, Massachusetts, USA

Thomas F. Quatieri

Bioengineering Systems & Technologies, MIT Lincoln Laboratory, Lexington, Massachusetts, USA

[1]DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

[2]DISCLAIMER: This material is based upon work supported under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force, U.S. Army or the Department of Defense.

†*Both authors contributed equally to this work.*

[3]*Corresponding Author: Adam.Lammert@LL.mit.edu*

Keywords: speech production, control models, DIVA, Task Dynamics, state feedback control

Abstract

Purpose: Computational models can serve to encapsulate current understanding of the mechanisms underlying speech motor control into a compact and rigorous form. Building up from first principles, and focus on low-level control of the speech articulators, this paper provides an overview of the current state of computational modeling of speech motor control by showing how each model is constructed out of these more basic domain-general ideas and components. The authors believe that progress in developing rigorous models of speech motor control will be aided by a thorough analysis and comparison of the attributes of the more prominent models of speech motor control.

Method: The review builds an understanding of existing models from first principles, before moving into a detailed discussion of currently proposed models of speech motor control, showing how each model is constructed out of the same basic domain-general ideas and components. By showing how each model is built up on these basic elements, this approach allows for a clear comparison between the proposed models, showing where they differ as well as points of agreement.

Results: We have seen that prominent models can be cast as special cases of generalized feedforward, feedback, and model predictive components of control. Although not every model makes use of all these components, or implements these mechanisms in precisely the same way, there are abundant commonalities among all models in terms of their general architecture, reliance on a higher-level motor planning, use of feedback signals and focus on kinematics aspects of control and biomechanics.

Conclusions: The many similarities of the models discussed in this review naturally lead to questions about what is gained by the variations of the models, and what would encourage convergence to a single, standard model of speech motor control, for instance by combining hybrid feedforward/feedback pathways with forward dynamic control, as well as feedback/internal model-based state estimation. Indeed, efforts are already underway in newer models to bring together these distinct proposed modeling constructs. Future

reviews may focus on additional, important considerations, including higher-level motor planning, linguistic, communicative and even social considerations, as well as learning and developmental aspects, all of which contribute to the wide variety of speaking styles observed in real human speech.

Current Speech Motor Control Models: An Overview of Architectures & Properties[4][5]

Introduction

Why do we need computational models of speech motor control, and what can be gained from understanding the details of how they work? First, these models can serve to encapsulate current understanding of the mechanisms underlying speech motor control into a compact and rigorous form. Computational models can also deepen our understanding of those mechanisms in biological systems by defining and circumscribing the space of possible architectures and solutions to a specified biological control problem (Schaal & Schweighofer, 2005). This allows clearer connections to be drawn between specific model/engineering mechanisms and their observed or presumed neurological counterparts. The connection between engineering and biological mechanisms has been well developed in several domains of motor control, including speech motor control (Guenther et al., 1998) and oculomotor control (Robinson, 1981; Lisberger, 1988; Shibata & Schaal, 2001).

The utility of speech production models additionally extends beyonds clarifying and formalizing our understanding of speech motor control itelf. As an encapsulation of theoretical understanding, models can also guide the design of new experiments to test specific aspects of theory and modeling, inspired by the behavioral predictions of the model. Moreover, models can be potentially useful for different applications in speech synthesis and analysis. Control models, coupled with faithful mechanical models of the vocal tract, hold promise for applications in flexible and expressive speech synthesis (Shadle & Damper, 2002). Models can also inform the selection of features used for automatic speech analysis, either through an analysis-by-synthesis approach, or by direct

[4]DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

[5]DISCLAIMER: This material is based upon work supported under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force, U.S. Army or the Department of Defense.

inspection of the model itself. This can be useful for both typical (Mitra et al., 2010) and pathological speech (Ciccarelli et al., 2016).

Several formal models of speech motor control have been formulated and presented in the speech production literature. Underlying these models is the assumption, based on decades of observation, that speech motor control mechanisms are complex, and consequently require detailed description and analysis. Indeed, speech production is one of the most intricate motor activities in which humans engage. Producing speech requires fine timing and coordination of muscles that are interwoven, redundant and have complex mechanical properties, in order to move the diverse articulatory structures of the tongue, lips, jaw, velum and larynx into a wide range of configurations, all of which have a nonlinear relationship with the vocal tract's acoustic output. Control mechanisms are additionally modulated by higher-level planning processes that mediate semantic, syntactic, prosodic and phonological organization. The neurological basis for these mechanisms is widely distributed in the brain, involving, at a minimum, the laryngeal and orofacial motor cortex, supplementary motor area, inferior frontal and superior temporal gyri (Broca and Wernicke's areas), cingulate cortex, as well as the putamen, thalamus and the cerebellum (Guenther, 2016).

This paper provides an overview of the current state of computational modeling of speech motor control. Our approach is to build up an understanding of the mechanisms used in existing models of speech motor control from first principles. We start with a discussion of general motor control principles and approaches, before moving into basic domain-general models of motor control. We then move to detailed discussions of currently proposed models of speech motor control, showing how each model is constructed out of these more basic domain-general ideas and components. By showing how each model is built up on these basic elements, this approach allows for a clear comparison between the proposed models, showing where they differ as well as points of agreement.

While speech production is a complex process with developmental as well as

linguistic considerations, this review will focus specifically on control of the speech articulators in fully developed, adult speech. Other important aspects of speech production, including learning and optimization, higher-level linguistic processing, motor program generation (i.e. the "planner"), the neurological basis of hypothesized model components, and biomechanical details of the speech articulators (i.e. the "plant") will only be discussed to the extent necessary to clarify the nature and operation of the proposed control mechanisms. Figure 1 shows a schematic representation of the focus of the present paper.

The authors of this manuscript believe that progress in developing rigorous models of speech motor control will be aided by a thorough analysis and comparison of the attributes of the more prominent models of speech motor control. By providing a review of these models using the same terminology and explaining each in terms of simpler domain-general motor control principles, we provide such an analysis. We employ this approach to cover what are, to our knowledge, all current models of speech motor control: DIVA (Guenther, 1994, 2016), Task Dynamics (Saltzman & Kelso, 1987; Saltzman & Munhall, 1989), State Feedback Control (Houde & Nagarajan, 2011), ACT Kröger et al. (2009), GEPETTO (Perrier et al., 2005), FACTS (Ramanarayanan et al., 2016), Hierarchical State Feedback Control, and the Tian & Poeppel model. As our goal is to enable a basic understanding of the principles and mechanisms of each of these models for a general audience without specialized knowledge, we focus on conceptual explanations rather than mathematical details. For those readers who may desire more formal explanations, we also include some fundamental equations where appropriate, as well as an appendix with full equations for each model when possible.

Background

Motor control principles and terminology

The first step in discussing speech motor control models, it to define certain key concepts. To illustrate these terms and ideas, we will use a simple example borrowed from

the control of upper extremity reaching control, as shown in Figure 4b. This model comprises a simple two-link robotic arm moving on a planar surface. The robot's two links are connected to each other at a revolute joint that changes the angle between the links, u_2. One of the robot's links is fixed at the origin of the planar surface, defined as $(x_1, x_2) = (0, 0)$, but is free to rotate about this point, changing the angle u_1. The distal end of the second link (the "hand") is considered the end-effector of the robot, and its position can be given as coordinates in the space defined by (x_1, x_2). The space defined by the joint angles (u_1, u_2) describes the configuration of the arm. Additionally, this is the space in which the motor commands are defined, as the joint angles are directly controllable, and together will determine the final position of the hand. We will use this example to provide a more intuitive understanding of motor control principles, but will connect this simple model to the (likely more complex) speech production system throughout. This simpler example also serves as a reminder that the basic principles of speech motor control are shared with control of other effector systems.

Motor control begins with the apparatus to be controlled, commonly referred to as the plant (G). The plant transforms motor commands to movements. The plant in our example is the robotic arm, while the plant for speech production is the vocal tract (as well as nasal passage and respiratory system). The plant receives input from the controller (C), which issues motor commands to the plant. How the controller moves the plant is the primary focus of the current review. In addition to the plant and controller, most models of motor control also include a planner (P), which issues a signal, often referred to as a motor plan or motor program, to the controller that is then used to determine which motor commands are issued. These motor plans or programs are based on some higher-level or cognitive goal (e.g. reach to a particular point in space, produce the word "dad").

The desired outcome of a controlled movement is known as a *task*. *Task space* is an abstract space in which tasks can be defined, and which also functions as an abstract description of the plant's state. In our robot arm example, the task space is equal to the

space defined by (x_1, x_2). Tasks could then be defined in this space as a particular location, such as $(x_1, x_2) = (1,1)$. Alternatively, a task may be a static or time-varying trajectory through this space, such that the hand traces a specified path. In speech production task spaces might include, e.g., formant space or vocal tract constriction degree/location space.

These examples task spaces are considered abstract because the task variables that define them are not under direct control, and they typically cannot be directly used to control the system at the level of controlled variables.. Rather, task variables are distinct from, and composed of, variables that can be used to issue interpretable motor commands. In this sense, motor control can be viewed as a hierarchical arrangement of "high-level" task variables and "low-level" known as *mobility* variables (and defining a *mobility space*) [6]. Task variables and mobility variables form alternative high- and low-level descriptions of the plant's state. In our robot arm example, motor commands must be given in terms of the two joint angles (u_1, u_2), which are directly controllable), not the final hand position (which is not). This is also true for speech, where motor commands must be issued in terms of muscle lengths or the positions of the speech articulators (e.g. upper lip, lower lip, tongue tip, etc.), rather than directly in terms of desired formant values (e.g., F1 = 500 Hz) or vocal tract constrictions (e.g., lip aperture = 2 mm).

It is useful to define some formal conventions to aid in discussion of the various control architectures of interest. The *task state*, x, is a description of the plant's position in task space, whereas the *mobility state*, u, is a description of the plant's position in mobility space. A desired mobility state is also commonly used as the motor command that issued to the plant, especially in conjunction with its derivative $\dot{u}$. This derivative can be interpreted as a desired *change* in mobility state.

For many control models, it is necessary to refer to a *state estimate* (x_{est} or u_{est}),

[6] In the speech motor control literature, the term "articulatory space" is often used instead of "mobility space". We adopt the latter term here to avoid confusion with certain high-level task spaces that are defined in articulatory terms

which is a description of the controller's internal estimate of the plant's position, either in task space or in mobility space. This is often necessary, as the true state of the plant cannot be directly observed. In these cases, however, the state of the plant can be measured via its sensory output y. For our robotic arm, this may be the sensed joint angle (y_u). The sensory signal is y, an approximation to either x or u, typically corrupted by some combination of noise (e.g., neuronal noise), delays (e.g., slowed synaptic/axonal propagation) or transformations (e.g., warping). So, even though y is a observation or measurement of the state of the plant, the *sensory space* defined by y is not guaranteed to coincide with the plant itself. For speech production, sensory output is typically considered to be a combination of auditory (y_{aud}) and somatosensory signals (y_{somat}), where the somatosensory signal may include proprioceptive, tactile, and/or constriction degree information.

The *reference signal*, r_s, is a signal representing a desired configuration of the plant in the space defined by s. Reference signals may be in either task (r_x) or mobility (r_u) space. The *state error*, e_s, is a signal representing the difference between the reference and the state (or state estimate) in the space s. Again, s may be in either task (e_x) or mobility (e_u) space. These variables are used in Figure 2, and in related diagrams throughout the paper.

Formal Description of Planner, Controller & Plant: *With these conventions in mind, we can describe the function of both the planner, the controller, and the plant. The planner generates the reference signal, $P(\alpha) = r_s$ as a function of some high-level linguistic information α, and possibly as a function of time: $P(\alpha, t) = r_s$. The controller takes a reference signal as input and generates a motor command in mobility space: $C(r_s) = (u, \dot{u})$ or $C(e_s) = (u, \dot{u})$. The plant, which can also be viewed as a transfer function, converts the motor command to changes in plant state, which can be measured in both mobility and task space: $G(u, \dot{u}) = (u, \dot{u}, x, \dot{x})$.*

Types of motor control models

The purpose of this section is to lay out, in a general way, some common control architectures that are employed in various control applications, including both controling robotic systems as well as describing the functional aspects of physiological control. These general architectures are presented as a scaffold for understanding the specific architectures used in various speech motor control models, and also for the purpose of clarifying the terms used in the present paper to refer to those architectures. To illustrate these various architectures in an intuitive way, we will continue to employ the example of the planar robotic arm as in the previous section. However, these same architectures can be used to control much more complex systems, such as the speech production system.

Feedback control. Figure 2b shows an example of a *feedback* system architecture that, by definition of the term, makes use of outputs from the plant for maintaining control. These feedback signals are compared with the reference signal from the planner to generate an error signal. The error signal simply represents the difference between the current state and the desired state. The error signal is passed to the controller for determining the motor command. This type of controller is often referred to as a *closed-loop controller* in the control literature. Many types of controllers exist which make use of this error signal in a variety of ways, only a few of which will be discussed here. What all feedback controllers share is the basic idea that the error signal forms the basis for the motor command issued to the plant. The simplest feedback controller design is the proportional controller, in which the motor command is simply proportional to the error signal – e.g., $C(e) = K_p e$, where the term K_p is known as the *gain*. Larger gains lead to larger motor commands (i.e. the error has more of an effect on the system) while smaller gains result in smaller commands. Smaller gains are often preferable as large gains can lead in to instability and oscillatory behavior. Another commonly seen architecture is the proportional-integral-derivative (PID) controller, in which the motor command is a weighted combination of the error signal, its derivative and its integral with respect to time.

The second row in Figure 4b shows, across times t_1, t_2 and t_3, the progress of the robot arm as controlled by a feedback controller. At the beginning, the target is defined as desired point in task space (x_1, x_2). This type of target is sometimes referred to as a point-attractor, since the system should evolve to approach this point in task space regardless of its initial position, given sensible motor commands that reduce the error signal over time. The motor commands issued at each time step are a function of the error, e_x, between the current position of the end-effector and the point target. The error is determined by sensory feedback, which provides monitoring of the current state of the arm with respect to the position of the target.[7]. Although the error signal is in task space, the motor command issued by the controller must be given in mobility space since the only way to change the position of the end effector is to change the joint angles (u_1, u_2). The process of determining those commands requires some kind of transformation (e.g., kinematic inversion) from the desired coordinates in task space to corresponding coordinates in mobility space. Alternatively, it is also possible for the target to be a pre-specified trajectory rather than a point in task space. In this case, the error would be computed between the current position of the end-effector and the current desired position along the trajectory (typically time-locked).

Feedback control architectures have wide applicability in engineered and biological systems. Even simple designs typically lead to systems that accurately produce desired behaviors, and which can naturally handle unstable or unpredictable environments, including external perturbations to the plant. However, feedback systems can require careful calibration to ensure stability of control. Incorrectly tuned feedback systems can result in movements that grow uncontrollably or oscillate indefinitely. Feedback architectures are also heavily dependent on the quality of feedback signals. If those signals are slow to propagate, or if they require extensive processing once received, this can lead to motor commands being issued based on outdated state information, resulting in poor

[7]For this example, we make the simplifying assumption that the feedback signal is in task space, i.e. y_x

performance. This can be overcome by delaying the issuing of motor commands to match the feedback delays, but this results in slow movements. Additionally, if feedback signals are corrupted or otherwise inaccurate, this can lead to inaccurate movements. These final considerations are particularly important for biological systems, as there are substantial delays and noise inherent to neural processing of sensory feedback.

Feedforward control. One way to avoid the problems with delayed and noisy sensory information in a feedback controller is to cut out the use of feedback entirely. Figure 2a shows an example of a system architecture that makes no use of any outputs from the plant for maintaining control. Rather, the signals issued in the system are entirely *feedforward*, with the motor commands depending only on the reference signal. This architecture is commonly referred to as an *open-loop* control system, although we will use the terms *feedforward* and *open-loop* interchangeably in the present paper. The term feedforward control is sometimes used more specifically to refer to control architectures that can monitor perturbations to the plant, and adjust the motor commands to compensate without the need for explicitly monitoring outputs from the plant, usually by employing a highly accurate mathematical model of the plant (see the section on model predictive control, below). To the knowledge of the authors, this kind of architecture has not been used to model speech motor control.

The first row in Figure 4b shows the progress of the robot arm as controlled by a feedforward controller. From the beginning, the trajectory of the end-effector is defined in task space as a straight line originating at the end-effector's current position. The motor commands issued to the arm at each time step are directly determined by this pre-specified trajectory. As in a feedback controller, the reference signal is defined in task space but motor commands must be issued in mobility space. Again, this requires some kind of transformation from the desired coordinates in task space to corresponding coordinates in mobility space. Although the trajectory is most often specified in task space (for reasons which we will not elaborate on here), alternative feedforward controller could define the

plan in mobility space (for our robot example, in terms of joint angles) or even simultaneously in mobility and task space. The key aspect of feedforward control is that the plan is defined entirely beforehand, and carried out step-by-step from beginning to end. No information about the arm's actual position is used by the controller at any point throughout its movement.

Feedforward control architectures are unsuitable for unstable or unpredictable environments, where the plant can be perturbed by interference external to the system. Without the ability to detect and monitor errors in the system output, errors tend to persist, or even compound over time. Despite this obvious disadvantage, feedforward architectures are sometime attractive because they are capable of issuing motor commands quickly and without the need for complex handling of feedback signals.

Combining feedforward and feedback controllers. Each basic type of control system, feedback and feedforward control, has its own strengths and weaknesses. Feedback control is stable in the face of external perturbations, but becomes inaccurate or slow when sensory information is noisy or delayed (respectively), as in most biological systems. Feedforward control can be accomplished quickly, but is unstable when the state of the system cannot be predicted due to external perturbations.

It is possible to combine some of the strengths of feedforward and feedback systems, and mitigate the weaknesses of each, by constructing a *hybrid feedforward/feedback* controller, as shown in Figure 3a. This hybrid architecture comprises separate feedforward and feedback pathways that each issue their own motor commands, a (potentially weighted) combination of which becomes the final motor command that is issued to the plant. Such an architecture has the speed of a feedforward controller, but remains sensitive to unexpected perturbations and accumulating errors. Typically, the presence of the feedforward pathway allows for lower gains to be utilized in the feedback controller, leading to better stability. The primary disadvantage of combining feedforward and feedback pathway into a single system, in addition a more complex design, is that it is

less efficient in the use of energy expended on control itself. Relatedly, a hybrid system like this has the potential for redundancy of the control signals. For instance, if output from the plant always equals the reference (e.g., if the environment is entirely predictable), then the feedback pathway is not utilized and essentially unnecessary, since the feedforward pathway would be sufficient for control by itself.

One alternative to a hybrid control scheme is *model predictive* control. A model predictive controller, like the feedforward controller, makes no use of outputs from the plant for maintaining control. However, this architecture does make use of an *internal model* of the plant, which takes motor commands as input and transforms them into a prediction of the system's subsequent state, to predict the effects of the issued motor command. This effectively replaces feedback from the plant with a prediction of what the controller thinks that the feedback should be (Garcia et al., 1989; Miall & Wolpert, 1996). An example of this architecture is shown in Figure 2c. This state prediction acts as a kind of pseudo-feedback which can be compared against the reference, producing an error signal that is provided to the controller.

The third row in Figure 4b shows the progress of the robot arm as controlled by a model predictive controller. The functioning of such a controller is similar to the feedback controller example, above, in that the target is defined as desired point in task space, and the motor commands issued are a function of the error, e, between the current position of the end-effector and the point target. The difference is that the error is determined by comparing the desired state to the output of an internal model.

In terms of performance, the primary advantage of such an architecture is speed, since the delays associated with predicting the plant's state can often be much shorter than those associated with feedback propagation. Additionally, a model predictive controller does not need to have an entire plan (or trajectory) formulated before control is required, as is often the case with feedforward architectures. Rather, plans can be more compact, such as a single, time invariant point in task or mobility space (this is the same type of

plan often used in feedback controllers). The disadvantage of these systems is that accurate internal models can be difficult to design or learn, especially for complex, nonlinear plants such as the vocal tract. A poor internal model would mean that the predicted state may not match the true state of the plant, which can result in inaccurate control. Even small errors in the prediction will accumulate over time, since there is no way of correcting the prediction. Additionally, model predictive controllers have similar problems as feedforward control architectures in dealing with unpredictable environments and perturbations.

While pure model predictive control can be difficult to implement for these reasons, the use of an internal model is often quite useful as a component of larger architectures. For instance, internal model predictions can provide quick pseudo-feedback that can be used in conjunction with true feedback to provide fast, reliable control even in the face of long feedback propagation delays. Such methods are more stable than true model predictive control, since internal predictions do not need to be perfectly accurate, and small deviations between the predicted and actual states of the plant can be corrected via the feedback signal. An example of this concept is exemplified by the architecture known as the Smith predictor (Smith, 1959; Ghosh, 2005), as shown in Figure 3b. A Smith predictor effectively has three error comparison processes, generating state errors serially through comparing the state with a delayed version of the internal model prediction, which in turn is compared to a non-delayed internal model prediction, with this final comparison being subsequently compared against the desired state from the reference signal.

A Smith predictor is not the only controller that uses both state predictions from an internal model and feedback signals. An alternative is to combine the two signals into a single *state estimate*, which can be compared against the reference signal, rather than doing iterative comparisons. Combining the two signals provides a more accurate estimate of the plant's current state than either signal alone, due to inaccuracies in both the internal model and noise in the feedback signal. This is the approach taken by such classic control designs

as the Kalman filter (Kalman et al., 1960), which provides an optimal[8] state estimate with noisy feedback under certain strict assumptions. Importantly, the estimated state that results from combining internal predictions and feedback can then be compared with the desired state to generate a motor command (Todorov, 2004), just as in a pure feedback controller. This type of controller is sometimes referred to as *state feedback control*.

Speech models

We move now from domain-general motor control theory to models of speech motor control. Among the speech production models presented in the literature, perhaps the two most prominent are DIVA (Directions Into Velocities of Articulators) and the Task Dynamics model. The development of DIVA has been driven since the mid-1990's (Guenther, 1994) primarily by a team of researchers at Boston University, led by Frank Guenther. Task Dynamics has been developed by researchers associated with Haskins Laboratories, with Elliot Saltzman playing a key role, and with the theoretical groundwork being laid about 10 years prior to DIVA (Saltzman & Kelso, 1987; Saltzman & Munhall, 1989). More recent models include State Feedback Control (Houde & Nagarajan, 2011), the Task Dynamics State Feedback Control (FACTS) model (Ramanarayanan et al., 2016), ACT (Kröger et al., 2009), GEPETTO (Perrier et al., 2005), the Tian & Poeppel model (Tian & Poeppel, 2010), and Hierarchical State Feedback Control (Hickok, 2014).

Any model of speech production control must include, at a basic level, the ability to generate motor commands based on some motor plan. Those motor commands in turn activate a vocal tract model, possibly resulting in the generation of an acoustic signal. While complete models of speech production also need to include the formulation of motor plans, these elements are beyond the scope of the present paper, which focuses more narrowly on controlling the vocal tract for speech. An important reason for limiting the scope of the present paper is that the longstanding debate over acoustic vs. articulatory

[8]optimal here means closest to the true state of the plant, where "closest" means having the smallest mean squared error

targets of speech production tasks is often intertwined with the critical issue of how the vocal tract is controlled. For example, DIVA's tasks are formulated primarily in acoustic space, whereas applications of Task Dynamics (e.g., the Articulatory Phonology of (Browman & Goldstein, 1986) often assume tasks to be constrictions in the vocal tract. The choice of task space, however, is almost completely independent of the control formulations that are the focus of the current paper, and it is generally possible to reformulate any given control architecture using different task spaces. Therefore, the present work will discuss the task space used for each model, but will make no attempt to discuss the relative merits of the different task spaces used in different models. The concept of a task space is general enough to sit over and above the specific choice of task variables, while being well-defined enough as a concept to allowing meaningful comparisons of the control architectures underlying task space control.

Control elements that are relevant to any model of speech motor control, and which will be discussed in depth for each model in the following section, include: (a) the nature of feedforward mechanisms of control, including the formulation of the planner, (b) the nature and importance of feedback signals, (c) modeling of potentially imperfect sensory systems and/or perceptual processing of feedback, (d) the consequences of delays in feedforward and feedback pathways (e) the role of forward models in state prediction, and (f) the potential integration of both feedback and state predictions for state estimation, (g) the implementation of transformations between task space, mobility space, and sensory space, (h) the design of the controller for generating and issuing motor commands to the plant.

In order to evaluate the utility of different speech motor control models, there are a variety of key phenomena that models of speech motor control must explain in addition to simply accurately reproducing articulatory trajectories. These include, among other possibilities, (a) anticipatory and perseveratory co-articulation, (b) redundant degrees of freedom and motor equivalence, (c) compensation for acoustic and/or articulatory

perturbation that is potentially transient or persistent (d) speaking rate effects (e.g., non-linear temporal warping), (e) contextual variability (due to e.g., position in the syllable), (f) variability in production that is stochastic and/or speaker-dependent, (g) patterns of neural activity, including in the motor neurons and muscles, and (h) the causes of motor speech disorders and changes in production due to neurological and cognitive state. Where known, we will discuss how each of the current models accounts for these phenomena.

We note here that most current speech models are examples of purely kinematic controllers. That is, they do not account for dynamics or biomechanical considerations of the vocal tract. It is typically assumed that inertial parameters, centrifugal/coriolis forces and stationary external forces like gravity can all be ignored for the purposes of controller design and forward modeling. This may owe to the fact that several prominent models of the plant are purely kinematic: for instance, Maeda's model (Maeda, 1982) and the Haskins Configurable Articulatory Synthesizer (CASY) (Rubin et al., 1981, 1996; Iskarous et al., 2003). The focus on kinematics may also reflect an implicit assumption that dynamics of the plant can be ignored in the domain of speech motor control. Such an assumption is quite common in robotics and human motor control, and amounts to conceptualizing the plant as a collection of stiff articulators, akin to an industrial robotic arm. However, there is evidence that biomechanical factors play non-negligible roles in speech motor control (Derrick et al., 2015), and more recent vocal tract models such as Artisynth (Lloyd et al., 2012) incorporate dynamic and biomechanical aspects in their design.

Prominent Models of Speech Production

In the following section, we will discuss each of the current models of speech motor control in turn, explaining the architecture of the control system as it relates to the simple, domain-general systems discussed previously. Where necessary, we will touch on additional components of each model, such as motor program generation. We will also

discuss how each model accounts for the key phenomena in speech production listed above.

DIVA

In the context of classical control description, DIVA is a hybrid system combining a model-predictive controller with separate auditory and somatosensory feedback controller loops (Guenther et al., 2006; Tourville & Guenther, 2011; Golfinopoulos et al., 2011). The use of both model predictive and feedback control in DIVA is conceptually similar to a Smith Predictor. However, while a Smith Predictor uses serial error calculations to issue a single motor command, DIVA generates independent errors from each controller simultaneously. Each error is then individually transformed into a separate motor command. These three commands are then combined into a single motor command which is passed to the plant. The plant in DIVA is the Maeda model.

The basic component of the planning process in DIVA is the "speech sound", which can be a phoneme, syllable, or multisyllabic chunk. Each speech sound is linked to three distinct tasks, each a function of time: an articulatory trajectory (often called "motor" trajectory in the DIVA literature) defined in mobility space $r_u(t)$, an auditory sensory trajectory $r_{aud}(t)$, and a somatosensory trajectory $r_{somat}(t)$. The "speech sound map", which corresponds to the planner in Figure 2b, stores all three-component sets of mobility and sensory state trajectories. Each trajectory of the set serves as the reference signal to one of the controllers in DIVA: the articulatory trajectory serves as input to the model-predictive controller, and the sensory trajectories serve as input to the respective auditory and somatosensory feedback controllers. The three-component representation of speech sounds in DIVA means that each speech unit has a fully-specified articulatory trajectory and time-locked sensory expectations. Importantly, the sensory expectations are not generated online through an internal model, as in a state feedback controller.

The model predictive component of DIVA compares the predetermined desired position of the speech articulators at each point in time ($r_u(t)$) with their current predicted

position ($u_{predict}$), generating a feedforward control signal, $\dot{u}_{ff}$. In order to generate the mobility state prediction, DIVA integrates the control signal over time. This enables comparison of the estimated state of the vocal tract articulators $u_{predict}$ with the reference signal $r_u(t)$ independently of sensory feedback. Although the model-predictive controller is typically referred to as the "feedforward" controller in the DIVA literature, it is not a true feedforward controller in the sense of “open loop” control traditionally described in control systems, because it relies on a comparison between the predicted current model state and a reference. In its current implementation, the predicted state also incorporates some auditory and somatosensory feedback information, as well, since those pathways converge with the model predictive pathway. However, if the auditory and somatosensory feedback controllers in DIVA are entirely removed, the model predictive controller would function appropriately in the absence of sensory information.

Technical Details of DIVA's Model Predictive Controller: *In the model predictive controller, the control signal is generated from the following equation:* $\dot{u_{ff}} = g_{ff}G[r_u(\alpha, t) - u_{pred}]$*, where* g_{ff} *is an amplification gain applied to the motor command, and* G *is an additional gain that can be interpreted as a "go" signal, ranging between 0 (no movement) and 1 (maximal movement speed) as in Bullock & Grossberg (1988). Thus, the motor command is essentially a scaled version of an error signal, where the relevant error is between the articulatory reference signal and the predicted current position of the plant in mobility space. Note that the version of* u *that is used in computing the error signal is neither the true position of the articulators in mobility space, nor the one measured from the plant via sensory feedback, but an internal estimate of this state.*

Effectively, the quantity $r_u(\alpha, t) - u_{pred}$ *is an approximation of* $\dot{u}$ *prior to scaling. The predicted current position of the plant is used purely as a way of converting the reference signal into a velocity, because the reference signal (a set of articulatory positions) cannot be used directly as a motor command (which must specify a change in those positions). Alternative ways of computing the motor command would eliminate the need for the model-predictive component of the feedforward controller, converting it into a true "open-loop" system. For example, the planner could approximate the first derivative of the entire articulatory plan, and issue that as the reference signal. Alternatively, the planner could issue the reference signal within a window surrounding the current time point, which would allow the controller to approximate the first derivative. Further details can be found in Appendix A.*

The auditory and somatosensory feedback controllers closely follow the generic feedback control architecture. The auditory space in DIVA is defined as the first three formants (F1-F3) and the somatosensory space is defined as the positions of the individual articulators (proprioception) as well as the degree of contact between separate articulators (tactile sensation). Just as in a generic feedback controller, each feedback controller in DIVA compares the sensory target at each point in time (r_{aud}, r_{somat}) with incoming sensory feedback (y_{aud}, y_{somat}). DIVA includes delays in these feedback error correction

pathways to mimic the experimentally-determined delays in the neural system. The comparison generates an error signal (e_{aud}, e_{somat}), which is then transformed from sensory space into a control signal in mobility space ($\dot{u}_{aud}$, $\dot{u}_{somat}$).

The transformation of errors in sensory space to control signals in mobility space is difficult due to the *motor equivalence* problem – where multiple articulatory configurations could possibly achieve the same required auditory or somatosensory targets. DIVA solves the motor equivalence problem by implementing a transformation that chooses the control signal ($\dot{u_{aud}}$, $\dot{u_{somat}}$) with the smallest magnitude from all possible control signals that would eliminate the sensory error.

Technical Details of DIVA's FB Controller: *The computations performed by the sensory feedback controllers in DIVA begin with a comparison between the reference signal and the sensory output of the plant to produce an error signal in sensory space. For the sake of simplicity, only the auditory feedback computations will be presented here, but the form is the same for the somatosensory pathways. The auditory error signal is defined as:* $e_{aud} = r_{aud}(\alpha, t) - y_{aud}$. *This task-space error is then transformed into an articulatory-space error via the inverse kinematic equation:* $\Delta u_{aud} = J_{aud}^{-1} e_{aud}$. *The matrix* J *is known as the Jacobian, which provides a mapping between changes in mobility space and changes in task space. Specifically in DIVA, the Jacobian contains the rate of change for each of the dimensions of the task space for a corresponding change in mobility space. The matrix* J^{-1}, *is a pseudoinverse of the Jacobian, which allows for transforming task-space changes into mobility-space changes. The final motor command is then generated as the transformed error signal multiplied by a fixed gain:* $\dot{u}_{aud} = g_{aud} \Delta u_{aud}$. *This represents a kind of proportional control, where the motor command, ignoring transformations for the moment, is simply a scaled version of the error signal. Further details can be found in Appendix A.*

The output of the model predictive controller and sensory feedback controllers are summed to generate the final control signal, $\dot{u}$. Thus, the final control signal passed to the plant is the velocity of the articulators (or $\dot{u}$) needed to produce the desired change in the position of the articulators (termed *motor movement command*). The control signal

additionally includes the integration of $\dot{u}$ over time (u, or *motor position command*). This combined motor movement and position command is passed to the plant to drive changes in the position of the Maeda articulators. The plant also produces sensory outputs based on the position of articulators at each time point, y_{aud} and y_{somat}. In DIVA, the output of the plant is in the space of the reference signal (F1-F3 for the auditory reference, position of the articulators as well as articulator contact for the somatosensory reference). This avoids needing to model an auditory or somatosensory perceptual system.

An important detail to note is that the auditory and somatosensory reference signals are specified not as unique trajectories with a single value at each time point, but as time-varying regions. The error signal for each space (articulatory, auditory, somatosensory) is the distance from the current state to the edge of these regions. Thus, larger regions will allow greater variability in production, as no corrective error signal will be generated for any production that falls within the target region. The use of regions, rather than single points, provides a number of benefits for modeling speech behavior, as detailed below.

Explanatory power and limits. In addition to establishing the architecture of the speech motor control system, one of the primary motivations behind DIVA is establishing the neural basis of the various mechanisms that underlie speech motor control. To that end, the individual components of these model-predictive and feedback controllers have been mapped onto particular brain regions based on experimental neuroimaging results and model simulations (Guenther et al., 2006; Bohland & Guenther, 2006; Ghosh et al., 2008; Tourville et al., 2008; Golfinopoulos et al., 2011). Although other models have also hypothesized links between neural regions and control components, DIVA is the only model in which these hypotheses have been experimentally tested.

Some of the largest successes of DIVA have been in qualitatively simulating neural activity during various speech tasks. Since DIVA localizes the abstract control components (see the block diagram in Figure 5) to specific brain regions, it grounds the

block diagram process in neurobiology. Simulation studies with DIVA have provided good matches to neural activity recorded from human speakers during auditory and somatosensory perturbation experiments, including within utterance compensation (Golfinopoulos et al., 2011; Niziolek & Guenther, 2013; Tourville et al., 2008) and long-term adaptation (Villacorta et al., 2007). DIVA has also been able to simulate neural activity differences relating to syllable complexity (**?**).

DIVA simulations have been able to qualitatively match human behavioral responses to auditory perturbations, including both within-utterance compensation (Tourville et al., 2008) and long-term adaptation (Villacorta et al., 2007). Simulations also provide a fairly good match for human behavior when the lip or jaw is unexpectedly perturbed (Guenther et al., 2006). The model has also been used to derive variable productions of /r/ (Nieto-Castanon et al., 2005) based on a particular auditory target (low F3), a so-called "trading relationship" or "motor equivalence" where multiple articulatory configurations can be used for the same phoneme.

DIVA is able to model both carry-over and anticipatory coarticulation (Guenther, 1995). The key to both types of coarticulation in DIVA is that the targets for speech sounds are defined as regions, rather than points, in sensory (i.e., auditory and/or somatosensory) space. Coarticulation occurs when there is overlap between the somatosensory target regions of neighboring speech sounds.

Carry-over coarticulation (e.g. lip rounding when producing /d/ following a rounded vowel like /u/ but not an unrounded vowel like /i/) occurs when there is no need to actively change some articulatory variable from its existing value because the target region for a particular speech sound includes the current value. In the model, the target region for /d/ includes both rounded and unrounded values for the lips, so the lips are free to maintain their current degree of rounding, achieved during the preceding vowel, during production of /d/. Carry-over articulation effectively occurs as the execution level in the model, as it relies on only the starting position of the sensory parameters and the target

region of those same parameters in the current sound.

Anticipatory coarticulation arises from two sources in the DIVA model (Guenther, 2016). First, anticipatory coarticulation is acheived via a back-propogation process where speech targets are projected backwards in time until they conflict with the target for a previous speech sound. Effectively, coarticulation occurs when there is overlap between the somatosensory targets for the upcoming speech sound(s) and the current speech sound. If this is the case, the current speech sound is produced in that overlapping range. For example, when producing the sequence /adi/ the target tongue height for /d/ includes a broad region covering both high and low tongue positions. If the next segment is /i/ (high tongue position), there will be overlap between the two target regions in the high position, and /d/ will be produced with a high tongue. However, because the target tongue height region for /i/ conflicts with the region for /a/, there will not be anticipatory coarticulation between the two vowels. Secondly, the reference signals in each speech sound include coarticulated targets, because they are learned from fully coarticulated adult speech. Because it relies on either restricting the size of the target regions based on upcoming phonemes, or on the specified articulatory and sensory trajectories, anticipatory coarticulation in DIVA occurs at the planning level.

The use of target regions, rather than unique trajectories, provides a number of explanatory benefits beyond coarticulatory effects. An earlier version of the model, where articulatory targets were specified as regions (as sensory targets are in the current model), reproduced the differential effects of altering speech rate and hyperarticulation on consonant and vowel kinematics (see Guenther (1995) for details). Additionally, target regions in DIVA are hypothesized to be related to sensory acuity, where individuals with higher acuity have smaller target regions (Villacorta et al., 2007). Such differences may explain across-individual variability in the degree of adaptation to altered auditory feedback, with smaller target regions leading to a greater corrective response to feedback perturbations.

Since the current version of DIVA relies on fully time-specified articulatory trajectories (rather than time-varying regions, as for the sensory reference signals), the current implementation of DIVA cannot explain speaking rate effects other than a uniform increase or decrease in speed based on a master clock speed, though this could be accomplished through the incorporation of the approach in earlier versions of the model. The DIVA model does not currently account for any of the temporal or spatial effects of prosodic structure on speech, such as final lengthening or domain-initial strengthening.

Task Dynamics

The Task Dynamics model was formulated by Saltzman & Kelso (1987) in general motor terms, and then by Saltzman & Munhall (1989) in the particular context of speech production (see Figure 6). TD is essentially a feedback control architecture, as described in Figure 6. The controller uses a feedback comparator to relate the desired state issued by the planner (r_x) to the current state of the system (x). On the basis of this comparison (Δx), the controller computes a desired change in task space ($\dot{x}, \ddot{x}$) which is then transformed into a motor command or desired change in the state of the plant ($\dot{u}, \ddot{u}$). A crucial aspect of Task Dynamics is that both the desired state issued by the planner and the comparison performed by the controller occur in task space, not mobility space. This necessitates a transformation of the control signal from task space to mobility space before it can be utilized as a motor command. The plant in the Task Dynamics model is the CASY model (Rubin et al., 1996; Iskarous et al., 2003), which is a geometric model of the vocal tract, similar in spirit to the Maeda model used in DIVA.

A commonly-held view of Task Dynamics is that it does not incorporate a feedback process. This misconception was perhaps most recently mentioned in print by Kröger & Birkholz (2007) who stated that a serious problem with the Task Dynamics approach has been the fact that "perception [presumably feedback] as a control instance for production is not considered". Based on the discussion above, it should be clear that Task

Dynamics is, in fact, primarily a feedback-driven system. One criticism that could be made of Task Dynamics, however, is that the model assumes that the feedback process is noiseless and instantaneous, which is overly simplistic. Similarly, the model assumes that the current state of the plant can be directly known without errors from somatosensory feedback (note that this is essentially the same assumption that DIVA makes, where part of the sensory feedback signal is simply the positions of the Maeda articulators).

Technical Details of TD's Controller: *The computations performed by the controller in TD begin with a comparison between the (task-space) reference signal and the task-space position of the plant to produce an error signal:* $e_x = r_x(\alpha) - x$. *The error signal is then used, along with the task-space velocity of the plant,* $\dot{x}$, *to update the task-space acceleration of the plant via the forward dynamics equation:* $\ddot{x} = -M^{-1}B\dot{x} - M^{-1}Ke_x$, *where* M *is a diagonal matrix of inertial parameters and* B *is is a diagonal matrix of damping coefficients. Thus, the forward dynamics take the form of a second-order dynamical system that transforms the error signal into the second derivative of the task-space variable* x. *Since the task-space acceleration cannot be used directly as a motor command, it is necessary to transform this task-space acceleration into a mobility-space acceleration* ($\ddot{u}$). *This is accomplished through the use of a pseudo-inverse Jacobian function:* $\ddot{u} = J^{-1}[\ddot{x} - \dot{J}(u, \dot{u})\dot{u}]$. *This mobility space acceleration can then be integrated to produce a mobility-space velocity signal,* $\dot{u}$, *that can be used as a motor command for the plant to drive changes in the position of the speech articulators. Further details can be found in Appendix A.*

TD views speech motor control as a problem of point attractor dynamics. That is, motor tasks are conceptualized as points in task space, toward which the system is drawn by means of some governing control law which is a function of the system state. Task Dynamics describes the control law as a damped oscillator system (i.e., second-order dynamical system). Damped oscillator dynamics have a number of desirable properties in terms of defining a control law. In addition to the fact that damped oscillator dynamics are well-understood and easily characterized, the use of such dynamics to model task-directed behavior has the advantages that action patterns will be globally smooth and continuous.

Note on TD as Proportional-Derivative Control: *It is common practice in engineering control systems to take integral or derivative information of the error signal into account (e.g., the ubiquitous proportional-derivative, PD, and proportional-integral-derivative, PID, controllers – e.g., Åström & Hägglund (1995)). Integrating the feedback error, for instance, allows a controller to recognize accumulated errors, which it can then attempt to nullify. Using the derivative of the feedback error, on the other hand, can minimize undesirable future trends in the error signal, such as overshoot, oscillation and instability. TD is closely related to PD control. In PD control, the control signal* u_{PD} *is simply a weighted combination (given some weights* K_P *and* K_D*) of the error signal and its first derivative with respect to time:* $u_{PD} = K_P e_x + K_D \dot{e}_x$*. This equation looks remarkably similar to the forward dynamics equations from TD:* $\ddot{x} = -M^{-1}B\dot{x} - M^{-1}Ke_x$*, except that weights are specified, and* $\dot{x}$ *is substituted for* $\dot{e}_x$*. It can be easily shown that* $\ddot{x} = u_{PD}$*, given that* $K_P = M^{-1}K$ *and* $K_D = M^{-1}B$*, and knowing that* r_x *has a constant value, and therefore* $\dot{r}_x = 0$*. Thus, TD is equivalent to PD control up to the generation of the task variable acceleration signal, but differs in the additional transformation of the task space variables into mobility space, and integration of the mobility space variables.*

The task space in TD is defined in terms of high-level articulatory tasks (in contrast to the positions of the individual articulators themselves). For speech, the tasks are suggested to be constriction actions (i.e., gestures) of the vocal tract, such as achieving closure of the lips, rather than the positions of the individual speech articulators (for the lip closure task, these would include the upper and lower lips as well as the jaw). The point task is derived by the planner from a time-varying "gestural score" that issues the desired task state as a function of the currently active articulatory gestures. This definition allows TD to be easiiy put together with Articulatory Phonology (Browman & Goldstein, 1986). These two components form the basis for the perspective on speech production widely associated with Haskins Laboratories. Nevertheless, Task Dynamics and Articulatory Phonology are separate models that address different questions. Articulatory Phonology –

proposed roughly in parallel with Task Dynamics – asserts that articulatory gestures are the primitive units of spoken language. Gestures themselves are conceptualized with AP as discrete vocal tract constriction actions, which can be composed into gestural "scores" that function as a motor program for a given utterance. Therefore, in broad terms, Articulatory Phonology addresses the question of how speech tasks should be defined, and how they can be composed into a motor program, whereas Task Dynamics addresses the question of how those tasks can be achieved and how that motor program can be realized in a physical system.

Explanatory power and limits. Use of second-order dynamics directly connects TD to research on action planning and execution in biological systems. For instance, the VITE model is an influential neural-inspired network model for explaining kinematic trajectory formation of directed movement Bullock & Grossberg (1988). VITE comprises a network of three interacting hypothesized neural populations, each coding a distinct quantity that is needed in the generation of the motor command, given some target position. These neural populations encode quantities related to the present position of the system, the desired target position, and the difference between the target and the present position. These interacting populations are configured in such a way that there are many structural similarities to the control architecture of TD. The result of these similarities is that the present position of a population will move in a way that is consistent with a 2nd-order dynamical system, much like Task Dynamics (as pointed out by, e.g., Beamish et al. (2006)).

As an error-driven feedback system, guided by a planner following point attractor dynamics, TD is similar to other such systems that have been proposed, including the work of Hoff & Arbib (1993) and Schaal et al. (2007). Schaal et al. (2007) pointed out that this kind of formulation has several advantages, including:

- Plans are not explicitly time-dependent. Rather, they are dependent on system state/phase. This allows for more natural handling of interruptions and

perturbations during execution of the action.

- Control policies do not specify the entire trajectory of action. Rather, they only specify the parameters governing that trajectory. Therefore, they are highly flexible and globally defined.

- Control policies do not specify motor commands directly. Rather, they specify a change of task state. This generally results in policies that are (a) valid over a larger portion of the task space and (b) easy to superimpose into mixtures.

Task Dynamics can produce simple speech-rate effects by changing the dynamical parameters of the control law – e.g., by making the task-space motions more or less damped. In addition to these linear rate effects, the Task Dynamics model is able to produce a wide range of non-linear temporal effects seen in speech. Through the π-gesture model (Byrd & Saltzman, 2003), the model is able to capture the non-linear slowing found adjacent to prosodic boundaries as well as capture many of the spatial effects, such as larger movements (Fougeron & Keating, 1997), seen at those boundaries within a single framework. More recent work has extended the model to account for syllable structure and prosodic prominence (Saltzman et al., 2008).

One of the strengths of the model is accounting for coarticulatory effects. Coarticulation in this model is seen as arising from temporal overlap of independent articulatory gestures–the so-called coproduction model of coarticulation. The Task Dynamics model is naturally able to reproduce many of the coarticulatory effects seen in speech when multiple gestures are coproduced. In the case where co-active gestures control different tasks, both gestures are able to achieve their goals concurrently. This can explain vowel-influenced tongue shape during the production of consonants and variable jaw heights during bilabial closures produced near high and low vowels (Browman & Goldstein, 1992, 1995b; Fowler & Saltzman, 1993). In other cases, multiple gestures control the same tasks, in which case the relative strength of each gesture determines the final outcome.

Such gestural blending can explain variation in the place of velar articulation with front and back vowels. The model also reproduces vowel-to-vowel coarticulation in VCV sequences (Fowler & Saltzman, 1993). All these effects are produced at the execution level of the model, without changes in the speech plan. Other coarticulatory effects have been modeled at the planning level as changes in the temporal organization of gestures. These include partial vowel nasalization before nasal stops in codas, clear vs dark /l/ alternations, as well as palatalization of /s/ following high front vowels (Browman & Goldstein, 1992, 1995a; Zsiga, 1994).

Very early results from the Task Dynamics model showed that it was capable of reproducing the compensatory behavior seen in mechanical perturbation experiments, where a lowered jaw position during production of a bilabial stop is compensated for by a higher lower lip and lower upper lip (**?**). This compensation is due to the hierarchical nature of the model that controls high level tasks (like lip closure) rather than articulatory positions directly. However, the model is unable to account for auditory perturbations, as there is no auditory feedback channel.

While some recent work has started to explore neural mechanisms for some of the components of the model (Tilsen, 2016), and a connection to the VITE neural model (mentioned above) has been established, the components of TD have not been explicitly related to specific neural structures in the way that components of DIVA have been.

State Feedback Control

The State Feedback Control model for speech production (SFC) is a speech-specific instantiation of the general Kalman filter-type architecture in Figure 3c (Houde & Nagarajan, 2011; Houde & Chang, 2015). This type of model is used widely in current theories of motor control in non-speech domains (e.g., work from Todorov & Jordan (2002); Todorov (2004); Shadmehr & Krakauer (2008); Diedrichsen et al. (2010)), and is an evolution of a traditional feedback control system (Fig 2b). Recall from above

that a primary challenge of feedback control is that sensory feedback is typically noisy and delayed, making the instantaneous state of the plant impossible to know with perfect accuracy. By adopting a Kalman filter-type architecture (Fig 3c), SFC presents, in a speech motor control context, one method by which sensory feedback may be integrated with internal model predictions to produce improved estimates of the state of the plant.

In the SFC model (shown in Figure 7), estimation of the plant state is done by an "observer" (equivalent to the plant model and state estimator in the integrated model-predictive/feedback controller shown in Fig 3c). This observer receives a copy of the outgoing motor command issued by the control law (x', also known as the efference copy). Based on this signal, the observer predicts how the plant will move at the next time step ($\tilde{x}$) as well as the auditory and somatosensory feedback that will be received based on that predicted movement ($\tilde{y}$). The predicted sensory feedback is then compared with actual sensory feedback to calculate a sensory error (Δy). This error is then converted to a task state error (or task gain), via a gain function. Finally, the task state is estimated ($\hat{x}$) using the predicted state as well as the weighted sensory errors for both auditory and somatosensory predictions. As the gains associated with the sensory errors are assigned to optimize the final estimation, the observer in SFC functions is a Kalman filter (Todorov & Jordan, 2002), which provides the optimal *a posteriori* estimate of the state, under the assumption of linear processes of prediction and sensory feedback. Note that the sensory feedback the observer receives at any time point reflects the past state of the plant, while the state prediction reflects the current state. This delay is accounted for by delaying the sensory prediction before computation of sensory errors.

The model does not make explicit mention of a reference signal or a planner, and by extension does not make explicit mention of any comparison between sensory feedback and a reference. Providing a detailed description of the controller has not been a focus in the development of SFC, and therefore the controller, as presented in the literature, is represented by a generalized feedback control law which is a function $U(\hat{x})$ of only the state

estimate. This control law could take almost any form. However, the authors of this review expect any feedback control law that produces reasonable speech production behavior would need to be a function of some kind of reference, whether an explicitly planned trajectory or a gestural score. Indeed, specifying the details of this feedback control law in SFC, and the addition of a planner module, have been a primary motivation for the development of the FACTS model, described below.

By combining a state prediction with sensory feedback to estimate the current state, the SFC model is able to act quickly by operating principally on an internal prediction of the plant state. This also allows the system to operate in the absence of sensory feedback, either when that feedback is too delayed to be of use (as for very fast speech movements) or when sensory feedback is unavailable (as when speaking in loud noise or in cases of non-congenital deafness). Yet, the system is still able to respond when the internal predictions do not match the incoming sensory feedback (either due to errors in the prediction process or due to external perturbations of the plant). Thus, this system combines the major advantage of traditional feedback control systems (robustness to perturbations) with that of feedforward control (fast, accurate movement even in the absence of sensory feedback).

Explanatory power and limits. This model has been shown to be able to accurately reproduce behavior of human participants in pitch-alteration studies (Kort et al., 2014). In these studies, the vocal pitch is shifted unexpectedly while participants produce a steady-state vowel such as /a/. In the majority of cases, participants respond by shifting their pitch to oppose the perturbation (i.e. when pitch is shifted up, participants lower their pitch). This effectively cancels the sensory error.

There are two principle drawbacks to the current implementation of SFC. First, the current model only includes a basic model of pitch control, where pitch maps onto a one-dimensional mass-spring system (the position of the mass is taken to be equivalent to the produced pitch), not a model of the full vocal tract (either simple models like Maeda,

as in DIVA, or CASY, as in TADA, or more complicated models such as Lloyd et al. (2012)'s ARTISYNTH or those used by Perrier et al. (2005) or Birkholz & Kröger (2006)). This is critical for the way the state of the plant is estimated in the current model, as the Kalman filter implementation used is applicable only to linear systems, and the vocal tract is a highly non-linear system. Relatedly, there is no control law in the model that would enable control of such a complex model. The current control law only specifies vocal pitch.

The SFC model has been developed as a neuro-biologically plausible model of speech production, with various brain regions identified with the different components of the model (Houde & Nagarajan, 2011; Houde & Chang, 2015). While it has not been tested as extensively as DIVA, the neural grounding of the model could be used to make predictions of a similar sort that would be comparable with fMRI or MEG data, as in (Kort et al., 2014). Simulation results show that the model can replicate two neural effects seen in human speech, including the reduction seen in cortical electroencephalography (EEG) signals when speaking compared to listening to the one's own speech played back over headphones or speakers (speech induced suppression). The model can also replicate the enhancement of the EEG signals when seen when one's speech is externally perturbed compared to when it is unperturbed (speech perturbation).

SFC, as currently implemented, makes no distinction between task space and mobility space; they are effectively collapsed into a single space, such that commands are issued in task space. This means that the current implementation of SFC is only able to model a system where the goals of speech production are the same as the mobility space of the system. SFC has been implemented to control pitch, where the f0 maps onto a one-dimensional mass-spring system.

FACTS

Recently, a new model – the Task-Dynamics State Feedback Control model – has been proposed that combines aspects of both Task Dynamics and State Feedback Control

(FACTS, (Ramanarayanan et al., 2016)). As a merger of TD and SFC, FACTS combines elements of feedback control and model predictive controller. FACTS is an attempt to combine the strengths of each model, while addressing the major shortcomings of each. Specifically, the Task Dynamics model includes a well-developed control law that relates the movements of the speech articulators to high-level tasks, but assumes perfect knowledge of the state of the vocal tract. Conversely, State Feedback Control focuses principally on how the state of the plant can be estimated from sensory information given the noise and time delays inherent in auditory and somatosensory perception, but has to date only been used to control a very simplistic one-dimensional model of pitch. FACTS combines the state prediction and estimation framework of SFC with the planning model and vocal tract control of TD.

The architecture of FACTS is shown in Figure 8. The control component of the model is the same as that for the Task Dynamic model, with a planner generating a gestural score, which is passed to a feedback controller to generate changes at the task ($\ddot{x}$) and mobility ($\ddot{u}$) levels. This final motor command, $\dot{u}$, is passed to the plant to produce articulator movements as in Task Dynamics. However, where Task Dynamics passes the current plant and tasks states directly back to the feedback controller, FACTS uses an observer to estimate the task and plant states, as in the earlier SFC model. Here, the final motor command $\ddot{u}$ is passed to an internal model of the plant to generate predicted articulator positions ($\tilde{u}, \dot{\tilde{u}}$), as well as auditory and somatosensory feedback (y_{est}). The estimated sensory feedback is then compared with sensory feedback from the plant to generate a sensory error (e_y). The estimated mobility state is generated from the predicted mobility state and the sensory error via an unscented Kalman filter, an extension of the linear Kalman filter to nonlinear systems (Wan & Van Der Merwe, 2001). The estimated mobility state is then converted to an estimated task state, needed by the feedback controller to generate the motor command at the next time point, via the same inverse dynamics function as in Task Dynamics. Note that the architecture presented here does

not match an earlier version of FACTS presented by Ramanarayanan et al. (2016), but rather a more recent version available at (Ramanarayanan, 2017).

Explanatory power and limits. The FACTS model is relatively new, and so remains mostly untested. However, the model is able to qualitatively reproduce human behavior when vowel formants are perturbed, generating a partial compensation that opposes the perturbation. One of the features of FACTS is that it builds on the successes of the Task Dynamics model. Since the control architecture is shared between the two models, FACTS can similarly produce the successes of the Task Dynamics model, including coarticulatory effects and responses to mechanical perturbations.

The model currently lacks many of the components of the more-developed DIVA model. These include modeling neural delays and specification of the neural structures that underlie the function of the different model components. While not currently included, inclusion of delays may be relatively straightforward in this framework, as shown by their inclusion in the SFC model.

ACT

Kröger et al. (2009) introduced the ACTion-based model of speech production, speech perception, and speech acquisition (ACT model) a neurocomputational model that draws elements from both DIVA and Task Dynamics. The architecture of ACT, shown in Figure 9, is essentially a feedforward controller when viewed between the motor plan and the plant. DIVA-style dual auditory/somatosensory feedback pathways are also part of the model. However, those pathways feed indirectly to the planner, by way of high-level comparisons against abstract phoneme templates. Within the present framework, information used to modify the motor plan is considered to be part of the planner, and is therefore outside the scope of low-level control, as defined here. The plant in ACT is a three-dimensional kinematic model with articulatory control parameters similar to the Maeda and CASY models (Birkholz & Kröger, 2006).

The planner in the ACT model relates to both the speech sound map of DIVA and the gestural score in the Task Dynamics model. Like in DIVA, the basic unit of speech is assumed to be the syllable, and each syllable is represented by a model neuron in the phonemic map (cf. the speech sound map in DIVA). As in DIVA, these abstract syllable representations are linked to specific motor and sensory plans. This is accomplished in ACT through the phonetic map. Unlike in DIVA, where the motor plan is represented as a time-varying desired articulatory position signal, the motor plans in ACT are defined in terms of high-level dynamic tasks (or gestures) as in Task Dynamics. Each motor plan is, in effect, a gestural score, which defines the activation levels and temporal extent of each speech gesture, with each speech gesture being defined as a dynamical point-attractor (r_x).

The phonetic map, in addition to linking the syllable to the motor plan, also links the syllable to associated sensory (auditory and somatosensory) expectations. One conceptual difference between ACT and DIVA is that DIVA views the sensory plans as the targets of speech that have associated motor plans, while in ACT the targets are the high-level task gestures with associated sensory expectations. This conceptual difference is reflected principally in terms of how the models are trained (an issue not taken up within the scope of the present review), but the basic architecture of the models is essentially the same: a high-level syllable activates a motor plan used for feedforward or model-predictive control and a sensory plan used for feedback control.

The core control architecture in the ACT model borrows ideas from Task Dynamics, but is quite distinct. As discussed above, TD makes use of task-space comparisons between a reference, derived from the task-based gestural score, and the current (somatosensory) system state to control task-space movements given a control law that is consistent with damped oscillator dynamics. ACT, on the other hand, uses the reference, similarly derived, to directly drive motor action in a feedforward fashion. This is accomplished by the motor execution module, which uses the reference $r_x(t)$ to generate a trajectory in task space (x) that is consistent with damped oscillator dynamics. The

task-space trajectory must be transformed into a mobility-space trajectory (u) that can be used as a control signal to drive movements of the plant. This transformation is accomplished by the primary motor map. A subsequent neuromuscular processing step exists in the model, and is presently implemented as a direct, linear mapping. Plans exist for this component to eventually map control signals onto individual and/or combined muscle groups in a neuromuscular model.

Like DIVA, the ACT model also has dual somatosensory and auditory feedback pathways. The principal way these feedback pathways are used in the model is to compare the current state of the plant against pre-learned templates representing the desired somatosensory and auditory states. The comparisons may result in a high-level "cortical" sensory error signal. A crucial difference between ACT and DIVA and other models is that this error signal is used to influence the motor plan, rather than the control signals. These pathways are also used in the model is to detect sensory errors for updating the phonetic map to drive trial-to-trial adaption, a model of development and learning. The authors of ACT have also planned an additional somatosensory feedback pathway from the plant to the motor execution component. This feedback pathway, included in published figures representing ACT, would be used to "control motor execution", presumably in a fashion similar to TD. This pathway has not yet been implemented, and the details of its properties have not been fully developed.

One difference between the ACT model and others is that the mappings that relate the different signals (syllables$\mapsto r_x$, syllables$\mapsto r_y$, $r_x \mapsto r_u$, $r_u \mapsto \dot{u}$, etc.) are implemented via tunable neural networks rather than as closed-form mathematical expressions. These networks are tuned during a learning phase. Early implementations of DIVA had neural networks involved in these mechanisms (Guenther, 1994), but they were subsequently removed. The use of trained neural network models for these transformations allows for flexibility in the form of the transformations. It opens the possibility that the transformations might take forms that deviate in unexpected, and potentially even

biologically-plausible, ways when compared to mathematically-driven transformations typically adopted. The use of neural networks also makes it likely, however, that key transformations, such as the control law and the inverse kinematic transformations, cannot be easily written down analytically in closed-form.

Explanatory power and limits. The ACT model is able to produce motor equivalence in articulators linked to the same gesture due to the use of high-level tasks rather than articulatory positions as the basic unit of the motor plan. When the jaw is fixed to be low or high during a bilabial closure, this difference is compensated for by movements of the lips (Kröger et al., 2009). The model is capable of making online corrections similar to TD through low-level somatosensory feedback. The model is also capable of corrections based on high-level auditory errors or somatosensory perturbations, but only indirectly, by changing the motor plan. Such corrections may also have complicated interactions with long-term adaptations, since the feedback processing component is also used for modeling adaptation due to learning.

The ACT model includes hypotheses about the neural structures that underlies the different components (**?**) but to date has not been used to generate simulated neural activity to compare to neural data. The model also does not include neural delays.

GEPETTO

The GEPETTO model (Perrier et al., 2005; Patri et al., 2015) is a feedforward model of speech control. Its architecture is shown in Figure 10. GEPETTO shares certain characteristics with DIVA. First, speech goals are defined as regions in acoustic space (F1-F2-F3). Second, speech control is modeled as a feedforward process. Unlike the preplanned trajectories in DIVA, however, GEPETTO employs a variation of optimal control. Speech targets are defined as time-invariant regions (ellipses), and the feedforward signal for any given set of targets is calculated using a cost-optimization procedure to move the tongue between those targets in the specified order. Costs are calculated based on

changes at the level of muscle length (cf. optimal control models in other domains which usually optimize jerk or end-point accuracy, (Hogan, 1984; Flash & Hogan, 1985; Todorov & Jordan, 1998)). An additional listener-oriented cost constrains how close the final production must come to the phonemic targets. This cost effectively changes the diameter of the target ellipse in formant space.

The GEPETTO model is based on the equilibrium point hypothesis (Feldman, 1986). Unlike the rest of the models discussed in this paper, GEPETTO uses a dynamical biomechanical model of the plant. This fact has a primary influence on the way in which GEPETTO employs control in that control must be carried out at the level of individual muscles, rather at the level of geometric model parameters/articulators as in the Maeda or CASY plant models. Currently, only the tongue is included as a controllable articulator, and modeled as a finite-element model with six muscles whose length can be independently controlled. The other vocal tract surfaces and articulators are fixed.

Explanatory power and limits. GEPETTO has been able to produce articulatory trajectories that show variable vowel undershoot, with more undershoot related to lower listener-oriented and motor costs. This agrees with predictions from Lindbloms H&H theory, where less effort (a lower motor cost) is suggested to underlie hypo-articulated speech. The model has also been used to replicate patterns of consonant-vowel coarticulation (Perrier et al., 2005). The most recent version of this model (Patri et al., 2015) uses a probabilistic Bayesian model in place of the fixed optimal cost function, and is able to model the trial-to-trial variability seen in real speech. This is unique among current models, as other models generate consistent articulatory trajectories for multiple productions of any given input.

GEPETTO has two principal drawbacks. First, it does not include any sensory feedback or even any pathway for possible integration of sensory feedback into the motor control system. Second, owing to its tongue-only plant model, it presently elaborates upon control for the tongue, rather than the entire vocal tract. Therefore, it's applicability to

modeling some aspects of speech production, like inter-articulator coordination or possible motor equivalence effects, is limited.

Other models

All the above models include, at a minimum, the ability to generate motor commands based on some motor plan. These motor commands are then used to move a vocal tract model of some kind. While such complete models are the primary focus of the current review, it is important to also consider more conceptual models which have not been implemented to the same degree. As will be demonstrated, Hierarchical State Feedback Control and the Tian and Poeppel models that we will review can be described within the same general control framework as the more developed models discussed above.

Hierarchical State Feedback Control. The Hierarchical State Feedback Control model (HSFC) (Hickok, 2012a,b, 2014) is an attempt to combine speech motor and psycholinguistic approaches to speech production. It shares the same general architecture as the State Feedback Control model of speech production (Houde & Nagarajan, 2011). At it's core HSFC has a hybrid predictive model/feedback control architecture, where predicted sensory consequences of actions are compared with incoming sensory feedback to correct a system state prediction. However, there are a number of critical differences between HSFC, on the one hand, and SFC or FACTS on the other. The model itself is mainly a conceptual framework, and does not include a full computational implementation that can be used to generate speech, though a small-scale implementation of the higher-level phonological planning has been carried out (Hickok, 2012a). Also, there are two levels of prediction and correction in HSFC (hence, hierarchical). These are a lower-level phoneme system and a higher-level syllable system. The phoneme system receives input from the somatosensory perceptual system, and the syllable system receives input from the auditory system. Secondly, the sensory predictions of each system can be compared with either external sensory feedback (as in typical predictive feedback control

systems) or internally-generated sensory targets activated by higher-level planning units.

In this model, the higher level syllable system is activated by the conceptual and word system, and it selects the associated motor programs associated with each of its component phonemes. Thus, the syllable system acts as a "planner" in our terminology, and is somewhat similar to the speech sound map in DIVA. The syllable programs also activate predicted acoustic consequences, which can be compared with incoming acoustic feedback to drive "corrections" to the syllable program. Unlike the state corrections in a typical hybrid model predictive/feedback control system, however, these corrections increase or decrease the *activation* of different syllable *plans*. Thus, when the planned syllable and (auditorily) perceived syllable do not match, the system can detect this (syllable-selection) error and select a different syllable.

It is the lower-level phoneme system that directly controls the plant by monitoring and correcting for errors in somatosensory space. The details of how the phoneme system generates actual motor commands are not specified in detail, but the general premise is that learned motor programs for each phoneme are activated by the syllable system along with associated somatosensory targets. The nature of these motor programs is left unspecified, so it is unclear whether these motor programs are feedback laws that convert current somatosensory states to motor commands to achieve those targets, or pre-planned articulatory trajectories. One important difference between HSFC and typical model predictive control models is that the prediction in HSFC is based on the activation of a planning unit at the syllable or phoneme level (similar to DIVA), and is not generated on the basis of a copy of a produced motor command (or efference copy).

Overall, the HSFC model is oriented toward understanding the activation and selection of different speech production units (syllables, phonemes) rather than toward understanding the speech motor system *per se*. While the selection process is well developed, and has in fact been implemented in a basic computational model (Hickok, 2012b), the actual generation of motor commands in left unspecified. The model does

suggest which neuroanatomical structures are involved in the various proposed portions of the model.

Tian & Poeppel

Tian & Poeppel (Tian & Poeppel, 2010) propose a hybrid model predictive/feedback control model of speech motor control. As in a typical model predictive controller, an efference copy of the motor command issued by the controller is used to predict the state of the vocal tract ("motor state"). Subsequently, this vocal tract state is used to generate an auditory sensory state prediction. No separate somatosensory estimate is explicitly generated from the vocal tract state – it is implicitly assumed that the vocal tract and somatosensory states are equivalent.

To generate an estimate of the vocal tract's current state, the auditory prediction is first compared with auditory feedback to generate an auditory error, which is subsequently passed through an inverse model (mapping between auditory and vocal tract states) to generate what is referred to as a "desired motor state". This can the thought of as the predicted vocal tract state based on auditory feedback. This desired motor state is combined with the vocal tract state prediction and somatosensory feedback to generate a final vocal tract state estimate via a process called "motor state correction". This estimate is then passed back to the controller.

The Tian & Poeppel model has not developed details about what the control law governing the controller might look like. While the model does include some suggestions of the role of different neural regions, these are limited to the vocal tract state prediction (parietal cortex) and auditory prediction (auditory cortex).

The overall architecture is very similar to the State Feedback Control model (Houde & Nagarajan, 2011). However, there are a few key differences. A major differences is that, while SFC is a fully implemented computational model, the Tian & Poeppel model is only conceptual. This means that many of the details of individual proposed processes in

the model lack specification. It is possible that a concrete implementation of this model would resolve some of the other differences between the two proposals. For instance, in the Tian & Poeppel model, there is no distinction between the true vocal tract state and the somatosensory estimate of its state, unlike in SFC (as well as most of the other models reviewed here). While the vocal tract state and somatosensory states in several other models substantially overlap, it may be important to maintain this distinction in order to incorporate other types of somatosensory information (e.g. tactile information). Another difference between the Tian & Poeppel model and SFC is that the state estimation process is substantially different in the two models. In the Tian & Poeppel model, the final state estimate is a combination of the motor state prediction, somatosensory state, and "desired motor state" generated from an auditory error. The details of how these three fully-specified vocal tract states are combined is unclear. In SFC, the vocal tract state prediction is combined with weighted error signals (not full states) generated from the auditory and somatosensory feedback systems. Additionally, Tian & Poeppel posit an inverse model process to map auditory errors to motor states, but do not explicitly specify whether this is an inverse model of the type used in SFC's Kalman filter-type implementation.

Discussion & Conclusion

One of the primary goals of the current paper has been to clearly lay out the organizational structure of existing speech motor control models using a common terminology and an understanding of the range of possible control architectures. Such a comparison on equal terms is necessary to understand the commonalities between these models, as well as how they differ. We have seen that these models can be cast as special cases of generalized feedforward (Fig 2a), feedback (Fig 2b), and model predictive (Fig 2c) controllers. The models discussed here differ in which of these components are used (e.g., some are lacking either feedforward or feedback elements of control), and in the detailed

implementation of these mechanisms. These differences are summarized in Table 6. However, there are some commonalities among all models that are important to mention.

All architectures rely on a motor plan of some kind – whether an explicitly planned trajectory or a gestural score – that is formed at a higher level of motor processing, and which is issued to the controller in order to be carried out. SFC is a partial exception to this general statement in that, as mentioned above, that model not explicitly mention the incorporation of a plan, even though the generalized structure of its controller would be able to absorb a planning module if more detailed specification required it (a specification which FACTS has essentially elaborated upon). Models of speech motor planning have been discussed and elaborated upon in the literature (Bohland et al., 2010; Civier et al., 2013; Saltzman et al., 2008; Byrd et al., 2009), and display a surprising amount of variety. Although the planning level is beyond the scope of this paper, it is worth noting the variety of planning mechanisms that have been proposed in order to help narrow some of the longest-standing debates concerning speech motor control. In particular, drawing a clear distinction between control architectures and planning mechanisms, as this review has attempted to do, makes it apparent that much of the debate over the quality of competing models of speech production would appear to concentrated at the planning level, and not at the level of control. For instance, issues surrounding the nature of production goals (e.g., acoustic vs. articulatory) and the composition of those goals into utterances-size units would primarily be a concern of the planning level. Any role for muscle synergies (Ramanarayanan et al., 2014) and motor primitives would be most naturally incorporated into the planning level, and not the level discussed in this review. The nature of speech production goals has been the subject of particularly strong debate for decades, and is reflected in the nature of the feedback and reference signals in the models, which may be auditory and/or somatosensory, as in DIVA and SFC, or articulatory, as in Task Dynamics. Interestingly, the nature of the feedback signals would appear to have little bearing on the specific architectural choices of the models – the architectures being general enough to

handle a range of signals without substantial changes to their arrangement.

In almost all the models discussed here, the final execution of speech motor control is influenced by feedback signals originating from the plant. GEPETTO and ACT are a notable exception to this generalization, since it contains no explicit feedback to its control mechanisms. Despite the fact that all other models make use of feedback, there nonetheless are clear differences among models in terms of the architecture of pathways that deliver feedback information to the controller. One of the most basic differences in this regard is whether certain feedback signals are treated as idealized signals that are directly and instantaneously observable, or whether they are treated as true sensory signals that may be potentially noisy/delayed, subject to conditioning by internal models and that correspond with known neurological signals. While it seems intuitively correct that any model of biological motor control should focus on the latter, for former has been sometimes intentionally chosen in specific aspects of the models, in the interest of focusing on other aspects of control. TD provides only an idealized view of feedback concerning the position of articulators that does not model the sensory processes in any meaningful way. DIVA and TD both also make simplifying assumptions about the somatosensory feedback signal, which is assumed to be more or less equivalent to the plant's mobility variables. DIVA's auditory and somatosensory feedback pathways are less idealized in that they correspond to known, independent neurological pathways and can incorporate delays associated with sensory transduction and processing. SFC and FACTS begin with the assumption that sensory feedback will be noisy and/or inaccurate, and use that assumption to motivate the well-elaborated integration of sensory feedback with internal model predictions to provide more accurate estimates of the state of the plant. A related distinction between these models is how they function in the absence of feedback. TD, for instance, is solely a feedback architecture, and cannot function in the absence of feedback signals. Other models have explicit articulatory trajectories that are used to generate motor commands through model predictive (DIVA) or feedforward control (GEPETTO) in the absence of

sensory feedback. Still others rely on the internal state prediction component of their broader state estimation process to continue functioning through model predictive control (SFC, FACTS).

In addition to finding points of similarity and distinction between models, it is also possible to identify features that are lacking from all models. In that regard, it is worth noting the simplicity of feedback control in all feedback-driven models relative to common engineering control designs. DIVA and SFC are based on simple proportional control, meaning that the motor commands are linearly proportional to the error (for feedback) or reference (for feedforward) signals. TD, as well as FACTS by way of adopting key control elements from TD, provide slightly more complexity through a form of proportional-derivative (PD) control, albeit not in the traditional engineering sense of PD control. Traditional PD control makes use of the feedback error and its derivative, whereas these models make use of the feedback error and the derivative of the state.

Controllers in all these models are purely kinematic in how they approach control. A variety of robotic controllers perform incorporate dynamical forces (effects of inertia, centrifugal and centripetal forces, and the effects of gravity) in producing control signals, since these effects are non-negligible in many motor control situations. Systems that are rigid and slow-moving, highly damped, or systems that have specialized designs can sometimes operate purely kinematically. It is not clear whether this assumption is reasonable for speech production. A kinematic approach can be, in the opinion of the authors, partially attributed to the focus on kinematics in models of the plant. Nearly all plant models used in conjunction with the presently-reviewed models are purely kinematic, and a plant model that has no dynamics requires no understanding of dynamics by the controller. GEPETTO represents an interesting intermediate case of a non-dynamic controller that is used to control a dynamical 2-dimensional biomechanical model of the tongue. It is worth noting that other plant models are attempting to provide enhanced biomechanics (Derrick et al., 2015; Gick et al., 2011; Lloyd et al., 2012) as well, even if a

full review of biomechanical vocal tract models is beyond the scope of the present review. Combining such a model with the control frameworks summarized in this paper is far from a trivial task, as, among other complications, the control space must change from articulatory positions to muscle lengths, as in GEPETTO.

Current models are also lacking in that their control parameters time-invariant. Whereas efforts have been made to model learning and adaptation at the planning level (e.g., GODIVA), there is no time variance modeled at the level of control architectures. Models cannot account for changes in control parameters that occur, presumably, during development and aging, or as a result of injury. Time variance in control due to shorter time-scale cognitive and physiological factors – for instance, due to attention, fatigue and motivation – is currently ignored, as well. Such factors may be a natural place in these models to account for situational and interspeaker variability, as well as stochastic variability (Tilsen, 2017).

An important, but perhaps obvious, observation to make is that the models described in this review are all closely related to engineering approaches to motor control, and bear a strong resemblance to classical control-theoretic architectures when formatted in the framework of the present discussion. Given that the theory behind current understanding of biological motor control largely grew out of early advances in engineering fields (Wiener, 1948; Bellman, 1957), it is perhaps unsurprising that the same is true specifically in the area of speech motor control. Indeed, engineering approaches are a sensible place to begin investigations into the nature of speech motor control in part because our current understanding of the functional interpretation of motor control neuroanatomy follows the engineering architectures closely. Engineering approaches have also provided a detailed and rigorous description of – to borrow terminology from Marr's levels of analysis (Marr & Vision, 1982) – the *computational* challenges inherent in general problems of motor control, and have also substantially explored the space of possible solutions to those challenges at an *algorithmic* level (Schaal & Schweighofer, 2005). Despite

their initial development in and application to engineering fields, it therefore makes sense to use these models to inform the design of experiments and as a foundation for interpreting empirical results from biological systems.

The researchers and authors behind DIVA have done considerable work toward clearly drawing connections between components of the control architecture and human neuroanatomy. Related connections were subsequently adopted by SFC, and also implicitly by FACTS. Given the similarity of control architectures illustrated in the present review and the overlap of their functional components, however, it seems clear that many of the same ideas could apply to the others models discussed here, as well. The control flow described by each of the models conforms well to current theoretical understanding of the neuroanatomy underlying motor control [9]. Functional components that are included in certain models and not others (e.g., internal models) could be ideal locations for further investigation into neural correlates of speech motor control. Given the correspondence between these models and cortical neuroanatomy, a reasonable interpretation of all models is that they represent central nervous system processes, and top-down control of movement in the brain. By that same token, it is notable that there is a lack of modeling for reflexive processes. For instance, the Lombard effect has long been viewed as an involuntary, reflexive reaction in the brainstem to specific auditory inputs (Lane & Tranel, 1971; Junqua, 1993; Summers et al., 1988; Hage et al., 2006). To the extent that reflexes may exert influence on speech motor commands, there is currently no incorporation of such

[9] Care must be taken, however, because such correspondence may be partially due to the expectation of finding a functional or spatial organization in the nervous system that is consistent with the computational model architectures. The influence of engineering approaches would appear to be so strong that an important consideration in the future may be to avoid the dangers of confirmation bias, whereby control-theoretic models are forming the lens through which neural data is viewed. In addition, whereas there is a distinct possibility that computational models and neuroanatomical descriptions are converging on a consistent understanding of control flow in human speech motor control, at least at an algorithmic level, that does not necessitate that the models are converging at an *implementation* level (*sensu* Marr) in the neural substrate.

processes in any of the models described in this review.

The many similarities of the models discussed in this review naturally leads to questions about what is gained by the variations of the models, and what would encourage convergence to a single, standard model of speech motor control. For instance, there is no mathematical reason why the feedforward/feedback pathways embodied by DIVA couldn't be combined with the forward dynamic control of TD, as well as the feedback/internal model-based state estimation in SFC. Indeed, FACTS has already taken a step toward beginning these potentially useful combinations. Among all the models, DIVA and TD stand out as having a relatively long history in which they have had extensive representation in the literature. DIVA also has experimental evidence supporting its proposed neural substrates. SFC and FACTS make clear and important modeling contributions that explain specific empirical results in speech production.

This paper has focused on control of the speech articulators in fully developed, adult speech, as was the goal of this review from the outset. However, speech production is a complex process with many additional and important considerations, including higher-level motor planning, linguistic, communicative and even social considerations, as well as learning and developmental aspects, all of which contribute to the wide variety of speaking styles observed in real human speech. These aspects are beyond the scope of the present paper, but are planned as the subject of future reviews of this type.

Acknowledgments

This work is sponsored by the Assistant Secretary of Defense for Research & Engineering (ASD[R&E]) under Air Force contract #FA8721-05-C-0002. The authors would like to thank Frank Guether, Satrajit Ghosh and Bernd Kröger for their time and assistance in understanding the details of their models.

Resources

Several of the models discussed in this paper (DIVA, TD, FACTS, CASY and the Maeda model) have been implemented as software tools, and are available for download online. Their addresses on the world wide web are included in the references below.

References

Åström, K. J., & Hägglund, T. (1995). *Pid controllers: theory, design, and tuning* (Vol. 2). Isa Research Triangle Park, NC.

Beamish, D., Bhatti, S. A., MacKenzie, I. S., & Wu, J. (2006). Fifty years later: a neurodynamic explanation of Fitts' law. *Journal of The Royal Society Interface*, *3*(10), 649–654.

Bellman, R. (1957). *Dynamic programming* (1st ed.). Princeton, NJ, USA: Princeton University Press.

Birkholz, P., & Kröger, B. J. (2006). Vocal tract model adaptation using magnetic resonance imaging. In *7th international seminar on speech production (issp?06)* (pp. 493–500).

Bohland, J. W., Bullock, D., & Guenther, F. H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. *Journal of cognitive neuroscience*, *22*(7), 1504–1529.

Bohland, J. W., & Guenther, F. H. (2006). An fmri investigation of syllable sequence production. *Neuroimage*, *32*(2), 821–841.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3-4), 155–180.

Browman, C. P., & Goldstein, L. (1995a). Dynamics and articulatory phonology. *Mind as motion*, 175–193.

Browman, C. P., & Goldstein, L. (1995b). Gestural syllable position effects in american english. *Producing speech: Contemporary issues*, 19–33.

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology*, *3*(01), 219–252.

Bullock, D., & Grossberg, S. (1988). Neural dynamics of planned arm movements: Emergent invariants and speed-accuracy properties during trajectory formation. *Psychological Review*, *95*(1), 49–90.

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, *31*(2), 149–180.

Byrd, D., Tobin, S., Bresch, E., & Narayanan, S. (2009). Timing effects of syllable structure and stress on nasals: a real-time mri examination. *Journal of Phonetics*, *37*(1), 97–110.

Ciccarelli, G. A. (2017). *Characterization of phone rate as a vocal biomarker of depression.*

Ciccarelli, G. A., Quatieri, T. F., & Ghosh, S. S. (2016). Neurophysiological vocal source modeling for biomarkers of disease.

Civier, O., Bullock, D., Max, L., & Guenther, F. H. (2013). Computational modeling of stuttering caused by impairments in a basal ganglia thalamo-cortical circuit involved in syllable selection and initiation. *Brain and language*, *126*(3), 263–278.

Derrick, D., Stavness, I., & Gick, B. (2015). Three speech sounds, one motor action: Evidence for speech-motor disparity from english flap production. *The Journal of the Acoustical Society of America*, *137*(3), 1493–1502.

Diedrichsen, J., Shadmehr, R., & Ivry, R. B. (2010). The coordination of movement: optimal feedback control and beyond. *Trends in cognitive sciences*, *14*(1), 31–39.

Feldman, A. G. (1986). Once more on the equilibrium-point hypothesis (λ model) for motor control. *Journal of motor behavior*, *18*(1), 17–54.

Flash, T., & Hogan, N. (1985). The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of neuroscience*, *5*(7), 1688–1703.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The journal of the acoustical society of America*, *101*(6), 3728–3740.

Fowler, C., & Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech*, *36*(171-195).

Garcia, C. E., Prett, D. M., & Morari, M. (1989). Model predictive control: theory and practice?a survey. *Automatica*, *25*(3), 335–348.

Ghosh, S. S. (2005). *Understanding cortical and cerebellar contributions to speech production through modeling and functional imaging* (Unpublished doctoral dissertation). Boston University.

Ghosh, S. S., Tourville, J. A., & Guenther, F. H. (2008). A neuroimaging study of premotor lateralization and cerebellar involvement in the production of phonemes and syllables. *Journal of Speech, Language, and Hearing Research*, *51*(5), 1183–1202.

Gick, B., Stavness, I., Chiu, C., & Fels, S. (2011). Categorical variation in lip posture is determined by quantal biomechanical-articulatory relations. *Canadian Acoustics*, *39*(3), 178–179.

Golfinopoulos, E., Tourville, J. A., Bohland, J. W., Ghosh, S. S., Nieto-Castanon, A., & Guenther, F. H. (2011). fmri investigation of unexpected somatosensory feedback perturbation during speech. *Neuroimage*, *55*(3), 1324–1338.

Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological cybernetics*, *72*(1), 43–53.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102*, 594-621.

Guenther, F. H. (2016). *Neural control of speech*. Mit Press.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language*, *96*(3), 280–301.

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological review*, *105*(4), 611.

Hage, S. R., Jürgens, U., & Ehret, G. (2006). Audio–vocal interaction in the pontine brainstem during self-initiated vocalization in the squirrel monkey. *European Journal of Neuroscience*, *23*(12), 3297–3308.

Hickok, G. (2012a). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, *13*(2), 135–145.

Hickok, G. (2012b). The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *Journal of communication disorders*, *45*(6), 393–402.

Hickok, G. (2014). The architecture of speech production and the role of the phoneme in speech processing. *Language, Cognition and Neuroscience*, *29*(1), 2–20.

Hoff, B., & Arbib, M. A. (1993). Models of trajectory formation and temporal interaction of reach and grasp. *Journal of motor behavior*, *25*(3), 175–192.

Hogan, N. (1984). An organizing principle for a class of voluntary movements. *Journal of Neuroscience*, *4*(11), 2745–2754.

Houde, J. F., & Chang, E. F. (2015). The cortical computations underlying feedback control in vocal production. *Current opinion in neurobiology*, *33*, 174–181.

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

Iskarous, K., Goldstein, L., Whalen, D. H., Tiede, M., & Rubin, P. (2003). Casy: The haskins configurable articulatory synthesizer. In *International congress of phonetic sciences, barcelona, spain* (pp. 185–188).

Junqua, J.-C. (1993). The lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, *93*(1), 510–524.

Kalman, R. E., et al. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, *82*(1), 35–45.

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

Kröger, B. J., & Birkholz, P. (2007). A gesture-based concept for speech movement control in articulatory speech synthesis. In *Verbal and nonverbal communication behaviours* (pp. 174–189). Springer.

Kröger, B. J., Kannampuzha, J., & Neuschaefer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. *Speech Communication*, *51*(9), 793–809.

Lammert, A. C., Ramanarayanan, V., Proctor, M. I., Narayanan, S., et al. (2013). Vocal tract cross-distance estimation from real-time MRI using region-of-interest analysis. In *Interspeech* (pp. 959–962).

Lane, H., & Tranel, B. (1971). The lombard sign and the role of hearing in speech. *Journal of Speech, Language, and Hearing Research*, *14*(4), 677–709.

Lisberger, S. (1988). The neural basis for motor learning in the vestibulo-ocular reflex in monkeys. *Trends in neurosciences*, *11*(4), 147–152.

Lloyd, J. E., Stavness, I., & Fels, S. (2012). Artisynth: A fast interactive biomechanical modeling toolkit combining multibody and finite element simulation. In *Soft tissue biomechanical modeling for computer assisted surgery* (pp. 355–394). Springer.

Maeda, S. (1982). A digital simulation method of the vocal-tract system. *Speech communication*, *1*(3-4), 199–229.

Marr, D., & Vision, A. (1982). A computational investigation into the human representation and processing of visual information. *WH San Francisco: Freeman and Company*, *1*(2).

Mermelstein, P. (1973). Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America*, *53*(4), 1070–1082.

Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural networks*, *9*(8), 1265–1279.

Mitra, V., Nam, H., Espy-Wilson, C. Y., Saltzman, E., & Goldstein, L. (2010). Retrieving tract variables from acoustics: a comparison of different machine learning strategies. *IEEE journal of selected topics in signal processing*, *4*(6), 1027–1045.

Nam, H. (2012). *TADA: TAsk Dynamic Application.* Retrieved 2017-08-11, from $http://www.haskins.yale.edu/tada_download/index.php$

Nam, H., Goldstein, L., Browman, C., Rubin, P., Proctor, M., & Saltzman, E. (2006). Tada (task dynamics application) manual [Computer software manual].

Nieto-Castano, A. (2017). *VTCalcs for Matlab.* Retrieved 2017-03-01, from $http://sites.bu.edu/guentherlab/software/vtcalcs-for-matlab/$

Nieto-Castanon, A. (2016). *DIVA Source Code.* Retrieved 2017-08-11, from $http://sites.bu.edu/guentherlab/software/diva-source-code/$

Nieto-Castanon, A., Guenther, F. H., Perkell, J. S., & Curtin, H. D. (2005). A modeling investigation of articulatory variability and acoustic stability during american english/r/production. *The Journal of the Acoustical Society of America*, *117*(5), 3196–3212.

Niziolek, C. A., & Guenther, F. H. (2013). Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations. *Journal of Neuroscience*, *33*(29), 12090–12098.

Patri, J.-F., Diard, J., & Perrier, P. (2015). Optimal speech motor control and token-to-token variability: a bayesian modeling approach. *Biological cybernetics*, *109*(6), 611–626.

Perrier, P., Ma, L., & Payan, Y. (2005). Modeling the production of vcv sequences via the inversion of a biomechanical model of the tongue..

Ramanarayanan, V. (2017). *TDSFC.* Retrieved 2017-??-??, from *http://www.*

Ramanarayanan, V., Goldstein, L., & Narayanan, S. S. (2014). Motor control primitives arising from a learned dynamical systems model of speech articulation. In *Fifteenth annual conference of the international speech communication association.*

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

Robinson, D. A. (1981). Models of the mechanics of eye movements. *Models of oculomotor behavior and control*, 21–41.

Rubin, P., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *The Journal of the Acoustical Society of America*, *70*(2), 321–328.

Rubin, P., Saltzman, E., Goldstein, L., McGowan, R., Tiede, M., & Browman, C. (1996). Casy and extensions to the task-dynamic model. In *Speech production seminar* (pp. 125–128).

Saltzman, E., & Kelso, J. (1987). Skilled actions: a task-dynamic approach. *Psychological review*, *94*(1), 84.

Saltzman, E., & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological psychology*, *1*(4), 333–382.

Saltzman, E., Nam, H., Krivokapic, J., & Goldstein, L. (2008). A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In *Proceedings of the 4th international conference on speech prosody. brazil: Campinas* (pp. 175–184).

Schaal, S., Mohajerian, P., & Ijspeert, A. (2007). Dynamics systems vs. optimal control?a unifying view. *Progress in brain research*, *165*, 425–445.

Schaal, S., & Schweighofer, N. (2005). Computational motor control in humans and robots. *Current opinion in neurobiology*, *15*(6), 675–682.

Shadle, C. H., & Damper, R. I. (2002). Prospects for articulatory synthesis: A position paper.

Shadmehr, R., & Krakauer, J. W. (2008). A computational neuroanatomy for motor control. *Experimental Brain Research*, *185*(3), 359–381.

Shibata, T., & Schaal, S. (2001). Biomimetic gaze stabilization based on feedback-error-learning with nonparametric regression networks. *Neural Networks*, *14*(2), 201–216.

Smith, O. J. (1959). A controller to overcome dead time. *iSA journal*, *6*(2), 28–33.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, *84*(3), 917–928.

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

Tilsen, S. (2016). Selection and coordination: The articulatory basis for the emergence of phonological structure. *Journal of Phonetics*, *55*, 53–77.

Tilsen, S. (2017). Exertive modulation of speech and articulatory phasing. *Journal of Phonetics*.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature neuroscience*, *7*(9), 907.

Todorov, E., & Jordan, M. I. (1998). Smoothness maximization along a predefined path accurately predicts the speed profiles of complex arm movements. *Journal of Neurophysiology*, *80*(2), 696–714.

Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature neuroscience*, *5*(11), 1226.

Tourville, J. A., & Guenther, F. H. (2011). The diva model: A neural theory of speech acquisition and production. *Language and cognitive processes*, *26*(7), 952–981.

Tourville, J. A., Reilly, K. J., & Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *Neuroimage*, *39*(3), 1429–1443.

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America*, *122*(4), 2306–2319.

Wan, E. A., & Van Der Merwe, R. (2001). The unscented kalman filter. In S. Haykin (Ed.), *Kalman filtering and neural networks.* Wiley, New York.

Wiener, N. (1948). *Cybernetics: Control and communication in the animal and the machine.* Wiley New York.

Zsiga, E. (1994). An acoustic and electropalatographic study of lexical and post-lexical palatalization in american english. *Haskins Laboratories Status Report on Speech Research, SR-117/118*, 67-79.

Appendix A

To aid the speech motor control practitioner, this Appendix consolidates the key algorithmic steps of three control architectures: Task Dynamics (TD), Directions into Velocities of Articulators (DIVA), and State Feedback Control (SFC). Bold lower case letters represent vectors, and bold upper case letters represent matrices. A single overhead dot represents a time derivative, and a double dot represents a second order time derivative.

Directions Into Velocities of Articulators (DIVA)

The Directions Into Velocities of Articulators (DIVA) model is a control architecture developed by (Guenther et al., 2006) that uses a hybrid of feedback control and model predictive control. The model has been realized in software, and is available online (Nieto-Castanon, 2016).

Algorithm. In DIVA, the mobility space, $\boldsymbol{u}$, and state of the plant, $\boldsymbol{x}$, are identical, so $\boldsymbol{u}=\boldsymbol{x}$. Table 1 describes the variables in DIVA.

1. Compute a model-predictive control signal (termed *feedforward* in the published literature on DIVA).

   (a) Compute an error using the reference target in mobility space and the current predicted state of the plant.

$$\boldsymbol{e_u} = r_u(t) - \boldsymbol{u} \tag{1}$$

   (b) Compute a feedforward control update by scaling the error signal.

$$\Delta u_{ff} = g_{ff} G \boldsymbol{e_u} \tag{2}$$

2. Compute a feedback-driven control signal using the reference target and the sensed plant output to compute an error in task space. Then, use a pseudoinverse Jacobian to convert the error from task space to mobility space. Do this in both the auditory and somatosensory feedback pathways.

$$e_{aud} = r_{aud}(t) - y_{aud} \tag{3}$$

$$\Delta u_{aud} = g_{aud} J^{-1} e_{aud} \tag{4}$$

$$e_{somat} = r_{somat}(t) - y_{somat} \tag{5}$$

$$\Delta u_{somat} = g_{somat} J^{-1} e_{somat} \tag{6}$$

3. Combine the feedforward and feedback control updates to determine the new plant state.

$$u = \int (\Delta u_{ff} + \Delta u_{aud} + \Delta u_{somat})\, dt \tag{7}$$

Task Dynamics

Task Dynamics is a feedback control architecture developed by (Saltzman & Kelso, 1987; Saltzman & Munhall, 1989). The architecture has been realized in software in

the Task Dynamics Application (TADA) (Nam et al., 2006) and available online (Nam, 2012).

Algorithm. The Task Dynamics algorithm is described below, and all variables are defined in Table 2.

1. Compute error in task space. In task dynamics, the task space, $\boldsymbol{y}$, and the state of the plant are identical, $\boldsymbol{x}$, so $\boldsymbol{y} = \boldsymbol{x}$.

$$\boldsymbol{e_x} = \boldsymbol{r} - \boldsymbol{x} \tag{8}$$

2. Use a dynamical system description of the controller, a second order ordinary differential equation, to compute the new acceleration state of the plant in task space.

$$\ddot{x} = -\boldsymbol{M}^{-1}\boldsymbol{B}\dot{x} - \boldsymbol{M}^{-1}\boldsymbol{K}e_x \tag{9}$$

3. Use a pseudoinverse Jacobian to convert the task space acceleration to mobility space acceleration

$$\ddot{u} = J^{-1}\left[\ddot{x} - \dot{J}\left(u, \dot{u}\right)\left(u\right)\right] \tag{10}$$

4. Integrate mobility space acceleration to get velocity and position in mobility space.

State Feedback Control

The State Feedback Control is a hybrid feedback/model-predictive control architecture proposed by (Houde & Nagarajan, 2011).

Algorithm.

1. Create a control update using the current estimate of the plant state.

$$u_t = U_t\left(\hat{x}_t\right) \tag{11}$$

2. Create the new, true plant state using the true plant dynamics, G_{dyn}.

$$x_t = G_{dyn}\left(u_{t-1}, x_{t-1}\right) \tag{12}$$

3. Generate the subsequent plant state using the true plant transformation from plant state to plant output.

$$y_t = G_{out}\left(x_t\right) \tag{13}$$

4. Create a new, initial estimate of the plant state using the previous estimated plant state, $\hat{x}_{t-1}$, the previous control signal, u_{t-1}, and an estimate of the plant dynamics, $\hat{G}_{dyn}$.

$$\hat{x}_{t|t-1} = \hat{G}_{dyn}\left(u_{t-1}, \hat{x}_{t-1}\right) \tag{14}$$

5. Create a correction term to the estimate of the plant state using the sensed feedback from the true plant.

$$y_{t-N} = G_{out}\left(x_{t-N}\right) \tag{15}$$

$$\hat{y}_{t-\hat{N}} = \hat{G}_{out}\left(u_{t-1}, \hat{x}_{(t|t-1)-\hat{N}}\right) \tag{16}$$

$$\tilde{y}_{t-N} = y_{t-N} - \hat{y}_{t-\hat{N}} \tag{17}$$

$$\hat{e}_t = K_t\left(\tilde{y}_{t-\hat{N}}\right) \tag{18}$$

6. Combine the initial plant state estimate and the correction term to create the current estimate of the plant state.

$$\hat{x}_t = \hat{x}_{t|t-1} + \hat{e}_t \tag{19}$$

Appendix B

This appendix presents two articulatory speech synthesizers commonly referenced in the literature: the Configurable Articulatory Synthesizer (CASY), and the Maeda model. Bold lower case letters represent vectors, and bold upper case letters represent matrices. A single overhead dot represents a time derivative, and a double dot represents a second order time derivative.

Configurable Articulatory Synthesizer

The Configurable Articulatory Synthesizer (CASY) is a geometric model of the vocal tract based on the work of Mermelstein (1973) and developed by Rubin et al. (1996) and Iskarous et al. (2003). We present the governing equations below, taken from Lammert (2013). We rename the "q" variables in Lammert et al. (2013), that represent the articulators in mobility space, to "u" in this paper for consistency of notation (see Tables 4 and 5 for details about the variables/constants).

$$x_{PRO} = u_{lx} \tag{20}$$

$$x_{LA} = l_{ut} \sin(a_{ut}) + l_{lt} \cos(u_{ja}) + u_{uy} - u_{ly} \tag{21}$$

$$a = u_{cl} \sin(u_{ja} + u_{ca}) \tag{22}$$

$$b = -u_{cl} \cos(u_{ja} + u_{ca}) \tag{23}$$

$$x_{TBCL} = acos\left(\frac{a - o_x}{\sqrt{(a - o_x)^2 + (b - o_y)^2}}\right) \tag{24}$$

$$x_{TBCD} = r_{ts} - \left(\sqrt{(a - o_x)^2 + (b - o_y)^2} + r_{tb}\right) \tag{25}$$

$$c = u_{ja} + u_{ta} + s_{tb}(u_{cl} - l_{tb}) \tag{26}$$

$$d = a + r_{tb} \sin(u_{ja} + a_{tc}) + u_{tl} \sin(c) \tag{27}$$

$$e = b - r_{tb} \cos(u_{ja} + a_{tc}) - u_{tl} \cos(c) \tag{28}$$

$$x_{TTCL} = acos\left(\frac{d - o_x}{\sqrt{(d - o_x)^2 + (e - o_y)^2}}\right) \tag{29}$$

$$x_{TTCD} = r_{tb} - \sqrt{(d - o_x)^2 + (e - o_y)^2} \tag{30}$$

Maeda Articulatory Synthesizer

The Maeda articulatory speech synthesizer is a variable cross-sectional area, tube model of the vocal tract. Resonances of the tube can be computed, and these resonances are the formants. The formants can then be used to shape a vocal source (voiced or unvoiced) to create speech. A MATLAB instantiation of the Maeda synthesizer was created by Ghosh and available for download (Nieto-Castano, 2017).

Ciccarelli (Ciccarelli, 2017) created a polynomial approximation to the vocal tract component to allow fast formant computation and fast, tractable computation of the pseudoinverse of the Jacobian. The polynomial approximation was determined by running the Ghosh implementation of the Maeda model across a set of parameters, uniformly sampled from the mobility space of the model, to create a lookup table of parameters and

formant values. Formant points outside the standard vowel quadrilateral as determined by visual inspection were excluded. The remaining pairs of articulator points and formants were then fit using a least squares polynomial approximation. The order of the polynomial was a compromise between the fit to the data and the complexity of the polynomial. It was found that a second order polynomial achieved a reasonable balance between these two requirements. While the mapping from articulators to formants is preserved to within a certain error, it has not been evaluated whether the relationship between articulators encoded by the polynomial fundamental alters the trajectories of articulators in previous implementations of the Maeda model.

Figure captions

Table 1

DIVA variables.

Variable	Description
e_u	Error between reference target in mobility space and last command issued to the plant
e_y	Error between the reference target in task space and sensed task space output
$r_u(t)$	Reference target in mobility space. Defined at each point in time as a region with a center and bounds of acceptable performance.
$r_y(t)$	Reference target in task space. Defined at each point in time as a region with a center and bounds of acceptable performance.
Δu_{fb}	Change in mobility space position based on error in task space. A task space velocity update.
Δu_{ff}	Change in mobility space position based on error in mobility space. A task space velocity update.
u	Mobility space position. Computed by integrating the feedforward and feedback mobility space velocities.
y	Task space output
g_{ff}	Gain applied to feedforward velocity update
g_{fb}	Gain applied to feedback velocity update
G	Gain with a value between 0 and 1 that constrains velocities in mobility space from 0 to their maximum.
J^{-1}	Pseudoinverse of the Jacobian. The pseudoinverse converts errors in task space to

Table 2

Task dynamic variables.

Variable	Description
x, $\dot{x}$, $\ddot{x}$	Task space position, velocity, and acceleration, m by 1 vectors
u, $\dot{u}$, $\ddot{u}$	Mobility space position, velocity, and acceleration, n by 1 vectors
r	Reference target in task space, m by 1 vector
M	Inertial coefficients, m by m diagonal matrix
B	Damping coefficients, m by m diagonal matrix
J	Jacobian transformation from mobility space to task space. An m by n matrix with elements $J_{ij} = \frac{\partial x_i}{\partial u_j}$
J^{-1}	The (pseudo) inverse of the Jacobian. The Moore-Penrose pseudoinverse may be used, or other constraints can be applied to allow inversion of a non-square Jacobian.
$\dot{J}$	The time derivative of each element of the Jacobian.

Table 3

SFC variables.

Variable	Description
x_t	True plant state
$\hat{x}_t$	Estimated plant state using both the sensed plant output and the predicted plant state
$\tilde{y}_{t_N}$	Error between the sensed plant output and the predicted plant output
$\hat{y}_{t-N}$	The predicted plant output, derived from estimates of the plant state, estimates of the feedback delay, and estimate of the plant transform from state to output.
$K_t\left(\tilde{y}_{t-\hat{N}}\right)$	Transformation (e.g. a Kalman gain) applied to the error between the predicted and sensed plant output. Allows actual plant output to influence the estimate of the plant state.
$\hat{x}_{(t\|t-1)-\hat{N}}$	Estimate of the plant state using only the previous estimate of the plant state, the control signal, and the estimated plant dynamics
G_{dyn}, $\hat{G}_{dyn}$	True and estimated plant dynamics.
G_{out}, $\hat{G}_{out}$	True and estimated transformation from plant state to plant output
y_t	True plant output
u_t	Control update to the plant
$\hat{x}_{(t\|t-1)}$	Estimate of the plant state based on the control update to the plant, the estimate of the plant dynamics, and the previous estimate of the plant state

Table 4

CASY variables.

Variable	Description
x	Task space variable
u	Mobility space variable
LX	Lip protrusion
UY	Upper lip vertical displacement
UT	Upper teeth
LY	Lower lip vertical displacement
JA	Jaw angle
CA	Tongue body angle
CL	Tongue body length
TL	Tongue tip length
TA	Tongue tip angle
LA	Lip aperture
PRO	Lip protrusion
TBCD	Tongue body constriction degree
TBCL	Tongue body constriction location
TTCD	Tongue tip constriction degree
TTCL	Tongue tip constriction location

Table 5

CASY constants.

Constant	Value
l_{ut}	1.1438
a_{ut}	-0.1888
l_{lt}	1.1286
o_x	0.7339
o_y	-0.4562
r_{ts}	0.4
r_{tb}	0.02
a_{tc}	1.7279
l_{tb}	0.8482
s_{tb}	4.48

	DIVA	TD	SFC	FACTS	ACT	GEPETTO
Feedback Pathway	Y	Y*	Y	Y	Y	N
Feedforward Pathway	Y	N	N	N	Y	Y
Internal Prediction/State Estimation	N	N	Y	Y	N	N
Principle Reference	Tourville & Guenther (2011)	Saltzman & Munhall (1989)	Houde & Nagarajan (2011)	Ramanarayanan et al. (2016)	Kröger et al. (2009)	Perrier et al. (2005)

Table 6

Summary of which aspects of motor control modeling are present in each model.

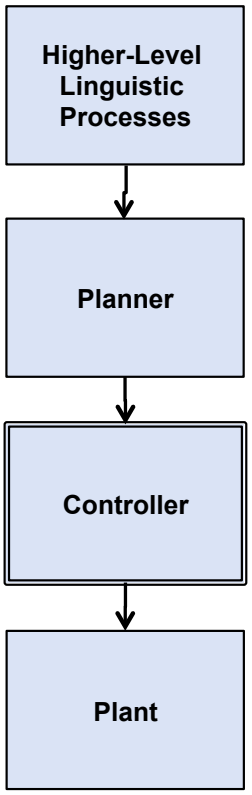
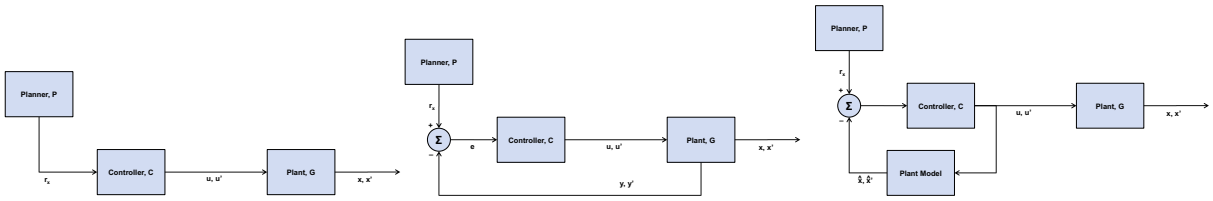
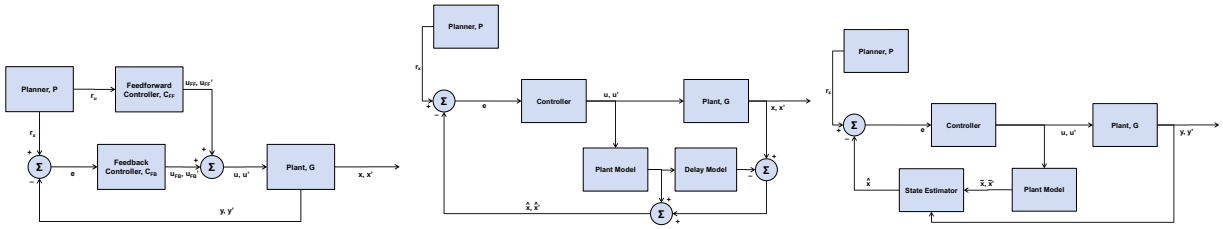


Figure 1. Representation of the distinct levels of speech production modeling. This paper focuses on modeling the speech controller, the system that takes in a speech plan and potentially feedback from the plant and issues motor commands to the plant. However, we recognize other components of the speech production hierarchy including higher level linguistic processes (prosody, semantics, syntax), the planner (low level sequencing of motor actions), and plant itself (e.g. speech synthesizers including but not limited to articulatory synthesizers such as CASY, Birkholz or Maeda).



(a) feedforward (b) feedback (c) model predictive

Figure 2. Control architecture of a generic (a) feedforward, (b) feedback, and (c) model predictive controller. The feedforward control architecture is distinguished from the other two because the controller only receives information from the planner, not information from the plant or predicted information from the plant. The feedback control and model predictive control architectures differ in the nature of the feedback received by the controller. In feedback control, the state of the plant (different than the output) is sent back to the controller. By contrast, in model predictive control, the state of the plant is sent back to the controller using an estimate of the plant and a copy of the issued control signal.

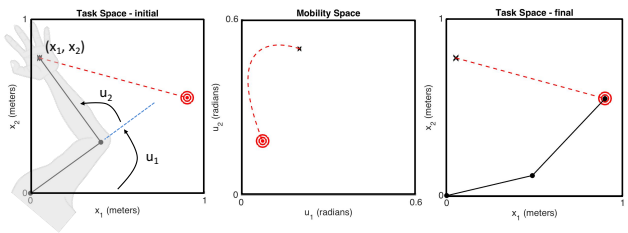


(a) hybrid feedforward/feedback

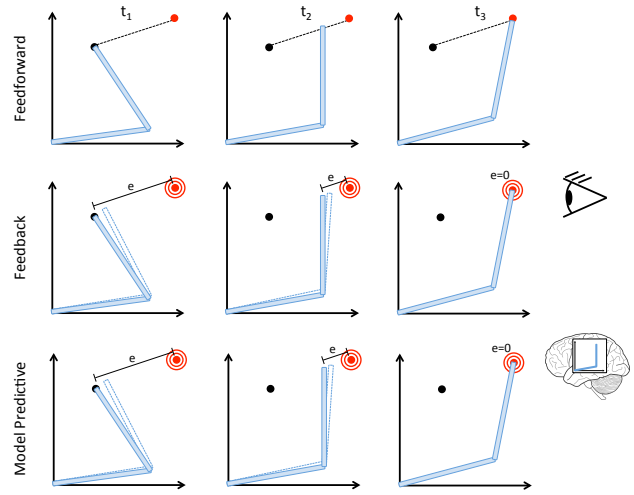
(b) hybrid model predictive/feedback

(c) integrated model predictive/feedback

Figure 3. Control architectures of generic hybrid controllers. A hybrid controller uses two of the three simple control architectures discussed in Figure 2. Diagrammed here are (a) a feedforward-feedback hybrid, (b) a model predictive-feedback hybrid with simple summation of state predictions (i.e., a Smith predictor) and (c) a model predictive-feedback hybrid with full integration of state predictions (i.e., a Kalman filter). Architectures (b) and (c) are distinguished by the particular form of information provided by the plant to the controller. In (b), the state of the plant is observed whereas in (c) the output of the plant is observed and undergoes further processing to infer the plant's state.



(a) Illustration of a simple planar robot arm in a reaching task as seen in task space and mobility space. (a) The arm in its initial configuration at (x_1, x_2) in task space, and the final goal (red circle). The arm's state variables (u_1, u_2) are defined as the angles of the shoulder and elbow. (u_1, u_2) are the parameters directly changed by the controller and therefore exist in mobility space. (b) The trajectory in mobility space. The evolution of the state variables (u_1, u_2) over time may be a non-linear trajectory despite a linear trajectory in task space. (c) The final orientation of the arm in task space at the goal.



(b) Illustration of the difference between feedforward (top row), feedback (middle row) and model predictive (bottom row) control using a simple reaching example. In feedforward control, the arm traces out a fully preplanned trajectory with no feedback about the position of the arm at any point in time. In feedback control, an error is computed between an observed state of the system (observation represented by the eye) and the target. The arm progressively works to minimize this error which drives the end effector towards the target. In model predictive control, an error is computed internally as opposed to being derived from feedback of the state of the system (represented by the brain with an internal model of the robot arm). The arm's position is updated to minimize the predicted error of the system.

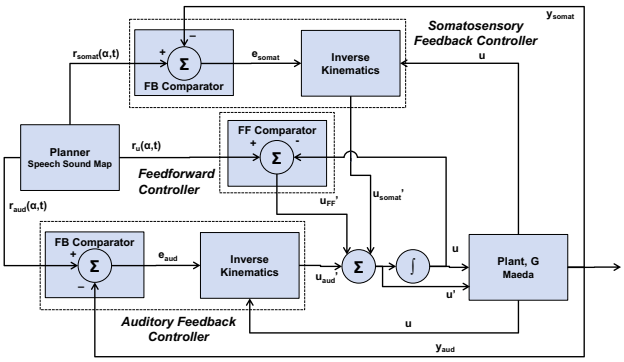


Figure 5. Control architecture of the DIVA model. The DIVA model has two feedback paths, auditory and somatosensory, that are schematically identical, and a feedforward-like pathway. The feedback pathways compute an error between the planner's signal and the output of the plant. This error is then used in conjunction with the state of the plant, u, to create a feedback control signal similar to the integrated model predictive-feedback control in Figure 3c. The feedforward-like path has more in common with model predictive control rather than feedforward control. The issued control signal is derived from a history of previously issued control signals and the planner. Therefore, DIVA is a mix of variants on both model predictive control and integrated feedback control.

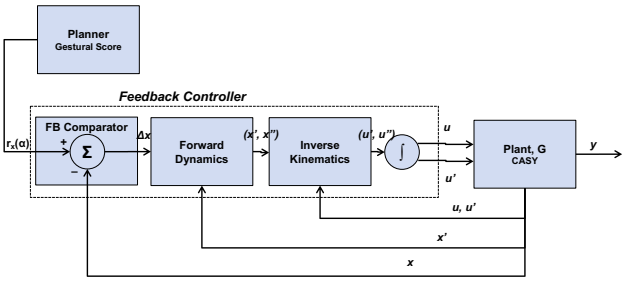


Figure 6. Control architecture of the Task Dynamics model. The system state, x, is broken out as both the state and change in state (first derivative), x'. This information is used by the controller in the rectangle. Comparing this diagram to Figure 2, we see TD is a feedback control architecture.

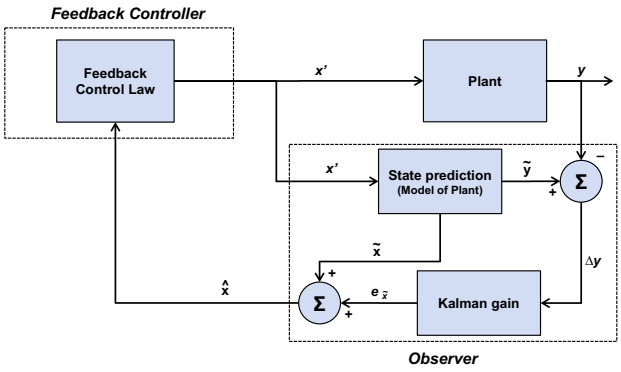


Figure 7. Control architecture of the State Feedback Control SFC model. The feedback signal, $x_{corrected}$, is the end-result of a well-elaborated and fully-integrated sensory processes in terms of state prediction and state estimation. Comparing this diagram to Figure 3, we see that SFC is an integrated model predictive feedback control architecture.

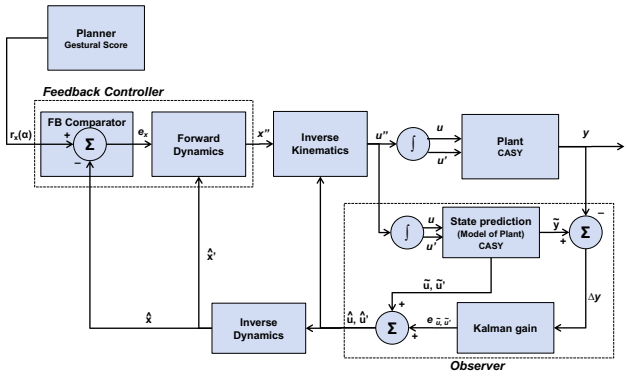


Figure 8. Control architecture of the Task Dynamics-State Feedback Control (FACTS) model. FACTS builds upon the architecture of the Task Dynamics model by substituting an estimate of the plant state for the true plant state through an observer module. The observer module does take into account previously issued state commands and includes an internal model of the plant, which gives the observer key elements of a model predictive architecture. However, the observer also has an inversion step from plant output to plant state much like integrated model predictive feedback control and more like DIVA.

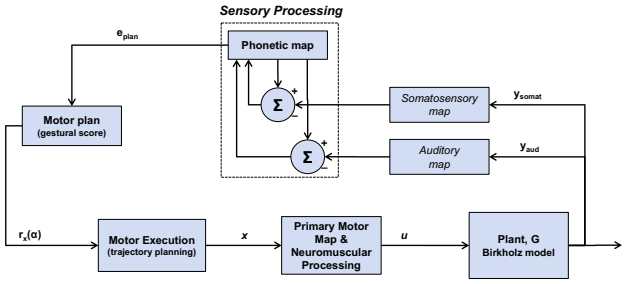


Figure 9. Control architecture of the ACTion-based model of speech production, speech perception, and speech acquisition (ACT model). ACT draws from both DIVA and Task Dynamics for its architecture, with the model comprising both feedforward and feedback pathways (both somatosensory and auditory), but relying on point-attractor dynamics for its reference signal.

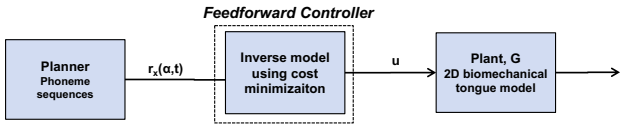


Figure 10. Control architecture of the GEPETTO model. GEPETTO is based on the equilibrium point hypothesis, employing control at the level of individual muscles, with relatively realistic biomechanics to move the articulators. Speech control is modeled as a feedforward process.