Discovering Dynamical Laws for Speech Gestures

Sam Kirkham

Department of Linguistics and English Language, Lancaster University

Abstract

A fundamental challenge in the cognitive sciences is discovering the dynamics that govern behavior. Take the example of spoken language, which is characterized by a highly variable and complex set of physical movements that map onto the small set of cognitive units that comprise language. What are the fundamental dynamical principles behind the movements that structure speech production? In this study, we discover models in the form of symbolic equations that govern articulatory gestures during speech. A sparse symbolic regression algorithm is used to discover models from kinematic data on the tongue and lips. We explore these candidate models using analytical techniques and numerical simulations and find that a second-order linear model achieves high levels of accuracy, but a nonlinear force is required to properly model articulatory dynamics in approximately one third of cases. This supports the proposal that an autonomous, nonlinear, second-order differential equation is a viable dynamical law for articulatory gestures in speech. We conclude by identifying future opportunities and obstacles in data-driven model discovery and outline prospects for discovering the dynamical principles that govern language, brain, and behavior.

Keywords: Speech production; Dynamical systems; Model discovery; Symbolic regression; Computational modeling; Articulatory phonology; Task dynamics; Nonlinear dynamics

1. Introduction

A longstanding goal in the cognitive sciences is the development of models that capture the dynamics of mind and motion. In the seventeenth-century, Newton proposed the fundamental laws of motion and gravitation, which synthesized a diverse range of observations into a

Correspondence should be sent to Sam Kirkham, Department of Linguistics and English Language, County South, Lancaster University, Lancaster LA1 4YL, UK. E-mail: s.kirkham@lancaster.ac.uk

unified mathematical framework. This identification of common mathematical laws has come to represent a core goal of modern science (e.g., Anderson, 1972; Newton, 1687; Noether, 1918), but it has long been acknowledged that the dynamics of living systems are structured by other laws beyond fundamental physics (Schrödinger, 1944). For example, within decades of Newton's discoveries, a new paradigm emerged that was focused on discovering fundamental laws behind human behavior and cognition (Hume, 1739), thus laying the foundations for contemporary cognitive science. Two central questions in this line of inquiry are (1) what are the principles that govern behavior and cognition? and (2) can these principles be expressed as mathematical laws?

An exemplary distillation of this challenge is the case of spoken language. Spoken (and signed) languages involve mapping a set of high-dimensional continuous physical movements to a low-dimensional set of discrete tasks that comprise the combinatorial units of language. A fundamental issue concerns the relation between these two components: the qualitative aspects of phonological knowledge and their physical realization. One view proposes a translation between discrete symbolic units and continuous physical properties, such that phonetic realization is a matter of translation or an "interface" between symbolic and physical domains (Chomsky & Halle, 1968; Guenther, 2016; Keating, 1990; Turk & Shattuck-Hufnagel, 2020). An alternative view holds that the relation between discrete and continuous aspects of phonological cognition can be explained using the language of nonlinear dynamics (Browman & Goldstein, 1986). In this sense, phonetics and phonology are isomorphic, rather than separate modules requiring translation, and can be cast as intrinsically linked elements within a single dynamical system (Browman & Goldstein, 1992; Gafos, 2006; Kelso et al., 1986).

The dynamical view of phonology emerges from a broader perspective in *dynamical systems theory*, which views the world through the mathematical language of change: differential equations. Dynamical thinking has a long history in the cognitive sciences (e.g., Fowler, 1980; Kelso, 1995; Port & van Gelder, 1998; Smolensky, 1988) and is typically set in opposition to highly modular models of mind (e.g., Fodor, 1975; Turing, 1950). The overarching goal is to identify the appropriate dynamical laws that govern brain and behavior across task-specific domains, including the relations between microscopic and macroscopic scales. For example, a key analytic concept is identifying the appropriate "order parameters" (Haken, 1977) or low-dimensional variables that govern qualitative states in the dynamical system (Haken, Kelso, & Bunz, 1985). This does not necessarily mean doing away with symbolic representations; for example, Gafos (2006) outlines a cogent theory of the relation between dynamics and phonological grammar that recasts the "interface" as a "dynamic linkage." Indeed, many dynamical theories maintain representations and symbolic systems, but the primary hypothesis is that cognitive agents are dynamical systems that can be cast in terms of states, paths, and flows, rather than modules, computations, and translations (van Gelder, 1998).[1]

Given the immense complexity of human cognition, accompanied by our highly variable and noisy behavioral measurements, how do we begin to identify appropriate dynamical models of brain and behavior? In this study, we focus on developing models of articulatory trajectories—or "gestures"—in the human vocal tract. This is a well-studied and tractable problem, as it is possible to collect lots of high-quality data on the movements of speech articulators, which we can use to derive the lawful regularities that govern such movements.

We approach this problem as one of data-driven model discovery and, in doing so, leverage recent developments in physics-informed machine learning and symbolic regression, which allow us to learn simple and interpretable dynamical equations directly from large datasets. The outcome is a small set of candidate models, from which new predictions can be generated and tested. Before outlining this approach in more detail, we first review extant dynamical models of speech, before motivating the search for new models.

1.1. Dynamical models of speech

In their article "The dynamical perspective on speech production," Kelso, Saltzman, & Tuller (1986) argue that the task for a dynamical model of speech is "less one of translating a 'timeless' symbolic representation into space-time articulatory behavior, as it is one of relating dynamics that operate on different intrinsic time scales." The dynamical timescales involved in speech production span macroscopic forms of organization (syllables, words, prosodic structure) to microscopic physical activity (articulators, biomechanics, neural dynamics) and may even expand to contextualize speech production in terms of a larger agent–environment dynamical system (perception–action dynamics, interactions with other speakers, etc.). The most thoroughly worked out theory of dynamical systems in the study of spoken language is articulatory phonology/task dynamics (Browman & Goldstein, 1992; Fowler, 1980; Gafos & Benus, 2006; Iskarous, 2017; Saltzman & Munhall, 1989; Tilsen, 2016), henceforth AP/TD. In this theory, the fundamental unit is the *gesture*, an abstract force acting on the vocal tract that drives it from its current state towards a new target state (Browman & Goldstein, 1992). To this end, the gesture is both a model of articulatory motion and a model of the "cognitive control of abstract linguistic units" (Byrd & Saltzman, 2003, p. 154).

A common mathematical model of speech gestures in AP/TD is the damped mass-spring model, which is visualized in Fig. 1 and captures the dynamics of forces on the vocal tract (Saltzman and Munhall, 1989). The model approximates an idealized physical system, with a physical mass attached to a spring, plus a damper or shock absorber. Variants of this model have been well-studied since the seventeenth century (Hooke, 1678) and the model can be easily understood by way of its physical properties, such as the analytical relationship between spring damping/stiffness and the system's position, velocity, and acceleration. These insights can then be generalized to the dynamics of vocal tract tasks during speech, allowing us to advance a specific and testable model of speech dynamics based on a well-understood physical analogy.

The dynamics of the speech gesture can be explicitly formalized as a second-order, critically damped harmonic oscillator, as in (1):

$$m\ddot{x} + b\dot{x} + kx = 0 \tag{1}$$

In the above equation, x represents the state of the system, such as the current state of an articulatory variable, while $\dot{x}$ and $\ddot{x}$, respectively, represent the velocity and acceleration of x. The system's mass m is typically set to $m = 1$ and k is a stiffness constant. The damping constant is $b = 2\sqrt{mk}$ in critically damped versions of the model. Eq. 1 specifies the system's
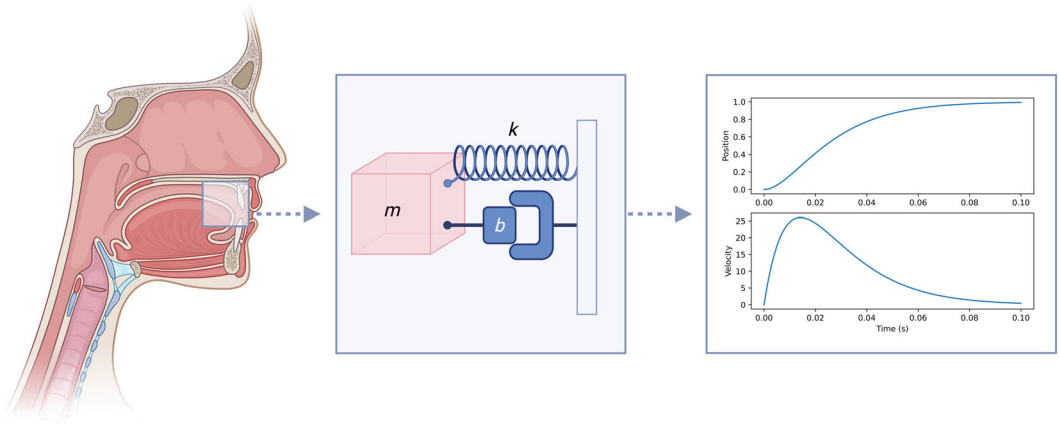
Fig. 1. The damped-mass spring model is a model of vocal tract task dynamics. The left diagram shows a midsagittal view of the vocal tract, with a box centered on the Tongue Tip task space. The middle diagram shows a physical damped mass-spring system representing the forces that act on the Tongue Tip gesture, where *m* is a mass, *k* is spring stiffness, and *b* is the strength of the damping force. The right diagram shows simulated trajectories from the damped mass-spring model, with the Tongue Tip moving from a low to a high position (arbitrary units).

equilibrium or target position as zero, but non-zero targets can be introduced by adding a target parameter (which we denote T) to the kx term as in (2). We assume that T is implicit in any subsequent formulations of these equations.

$$m\ddot{x} + b\dot{x} + k(x - T) = 0 \tag{2}$$

This equation defines the relationships between parameters as invariant over an instance of the system, where the system evolves until the value of T is reached. Research on the neural encoding of speech movements has identified signatures of critically damped oscillations in the neural populations of the ventral sensorimotor cortex, supporting a neural basis for the kinds of oscillatory models used in AP/TD (Chartier, Anumanchipalli, Johnson, & Chang, 2018).

The state of a gestural system evolves towards its equilibrium or target position, after which the target parameter changes and the system evolves towards a new state. This affords movement between different vocal tract postures, which is a fundamental characteristic of speech. Change in parameters requires a notion of gestural activation; if we cast gestural activation as a function of time $a(t)$ then we can transform (1) into (3).

$$m\ddot{x} + a(t)[b\dot{x} + kx] = 0 \tag{3}$$

In the standard model of Saltzman and Munhall (1989), $a(t)$ in Eq. 3 corresponds to the rectangular pulse function in (4). This means that gestural parameters change instantaneously at the point of gestural activation $a(t) = 1$ and remain constant until activation ends $a(t) = 0$,

where activation is bounded by the temporal interval $[t_a, t_b]$.

$$a(t) = \begin{cases} 1, & t \in [t_a, t_b], \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Eq. 3 is a critically damped harmonic oscillator with step activation. It is well-established that this model fails to capture many empirical characteristics of speech movements. This includes overly short time-to-peak velocity and highly asymmetric velocity trajectories compared with those seen in empirical data (Byrd & Saltzman, 1998; Sorensen and Gafos, 2016). An alternative approach is to introduce time-varying activation (Byrd & Saltzman, 1998). Eq. 5 is an example of a continuous activation function from Kröger, Schröder, and Opgen-Rhein (1995).

$$a(t) = \begin{cases} 0, & t < t_a \\ \sin\left(\frac{2\pi(t-t_a)}{4(t_b-t_a)}\right), & t_a \leq t < t_b \\ 1, & t_b \leq t < t_c \\ \sin\left(\frac{2\pi(t-t_d)}{4(t_c-t_d)}\right), & t_c \leq t < t_d \\ 0, & t \geq t_d \end{cases} \tag{5}$$

The activation function in (5) defines a quarter sine rise over $[t_a, t_b]$, steady activation over $[t_b, t_c)$, and a quarter sine fall over $[t_c, t_d)$. The ramped activation function corrects for the short time-to-peak velocities and velocity asymmetries of step activation (Byrd & Saltzman, 1998). In both step and continuous activation, gestures are active when they exceed a threshold of zero (with possible values of 0 or 1). An alternative view casts gestures as continuously active (Tilsen, 2020), which we discuss in Section 6.

Common to models of continuous gestural activation is that the system is explicitly time-dependent (i.e., non-autonomous) during activation. An alternative approach is to reformulate the dynamical equations that govern gestural dynamics, instead of using a more complex activation function. For example, Sorensen & Gafos (2016); Sorensen and Gafos (2023) introduce a cubic term in (6) but retain the rectangular pulse function in (4) for the activation variable $a(t)$. Note that the m term has been omitted from (6) and all subsequent related equations because it is conventional to define $m = 1$ (although see Šimko & Cummins, 2010, for a task dynamic model where different gestures are defined over different masses).

$$\ddot{x} + a(t)[b\dot{x} + kx - dx^3] = 0 \tag{6}$$

The cubic term in (6) has the effect of acting as a nonlinear restoring force on the spring in the mass-spring model, with d governing the strength of the nonlinear force. This means that the effect of the nonlinear force is greater with larger movement displacement, thereby reproducing observed nonlinear empirical relations between movement amplitude and peak velocity (Ostry & Munhall, 1985). The model also generates the symmetrical velocity profiles observed in empirical data, while retaining the step-function activation, which yields constant forcing during periods of gestural activation. The effects of this can be seen in Fig. 2, which compares the linear and nonlinear task dynamic models under step activation.
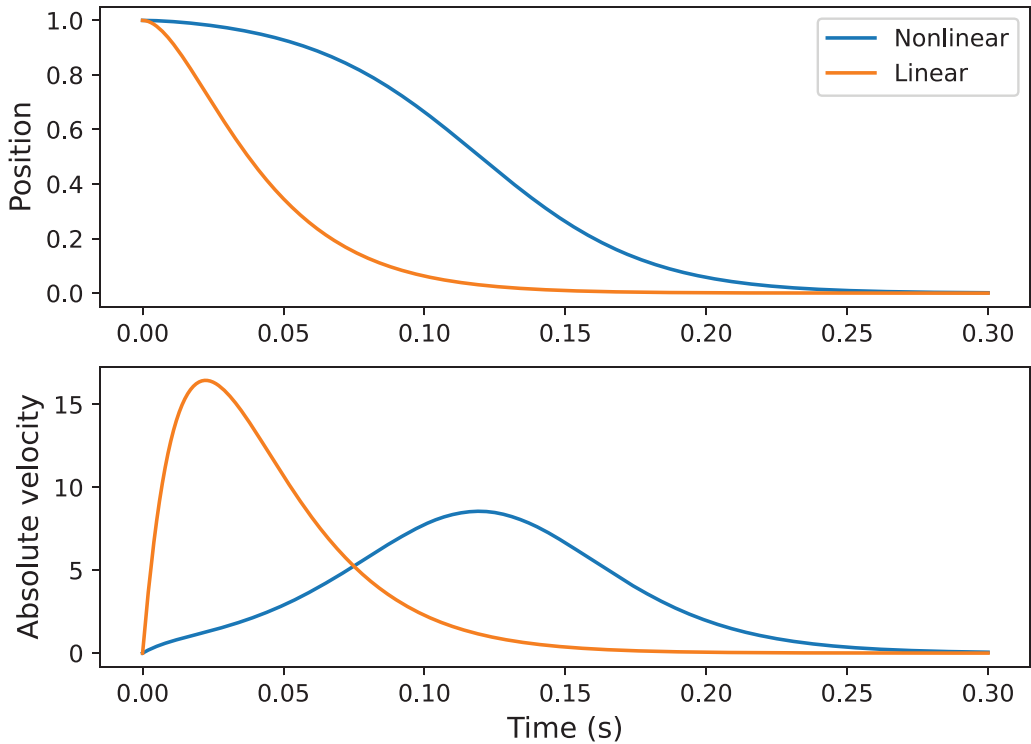
Fig. 2. Position and absolute velocity trajectories simulated using a linear damped mass-spring model and a nonlinear (cubic) damped mass-spring model. In both cases, $x_0 = 1$, $\dot{x}_0 = 0$, $T = 0$, $k = 2000$, $b = 2\sqrt{k}$. The nonlinear cubic coefficient is $d = 0.95k$.

The prevailing philosophy in the above research seeks minimal and autonomous models that capture fundamental dynamics without recourse to extrinsic timing mechanisms (Iskarous, Cole, & Steffman, 2024; Sorensen & Gafos, 2016). But we should be open to the possibility that further good models may be possible. For example, Turk and Shattuck-Hufnagel (2020) propose a General Tau model for articulatory movements, as part of a broader symbolic theory of phonology. Their model fits position and velocity trajectories better than the nonlinear dynamical model in Sorensen & Gafos (2016), although a comprehensive evaluation of how different models fit acceleration data is yet to be established (Sorensen & Gafos, 2016). There are also reasons beyond quantitative fit to search for new models, such as the trade-off between simplicity and accuracy and the value of interpretable parameters. Standard models may also not capture all kinds of speech and may require modifications or extensions to reproduce phenomena observed in disordered speech (Mücke, Roessig, Thies, Hermes, & Mefferd, 2024), speech development (Abakarova, Fuchs, & Noiray, 2022), different languages (Geissler & Nellakra, 2024), and so on. Finally, discovering new models is a core part of evaluating the success of existing models; if we can develop better models then
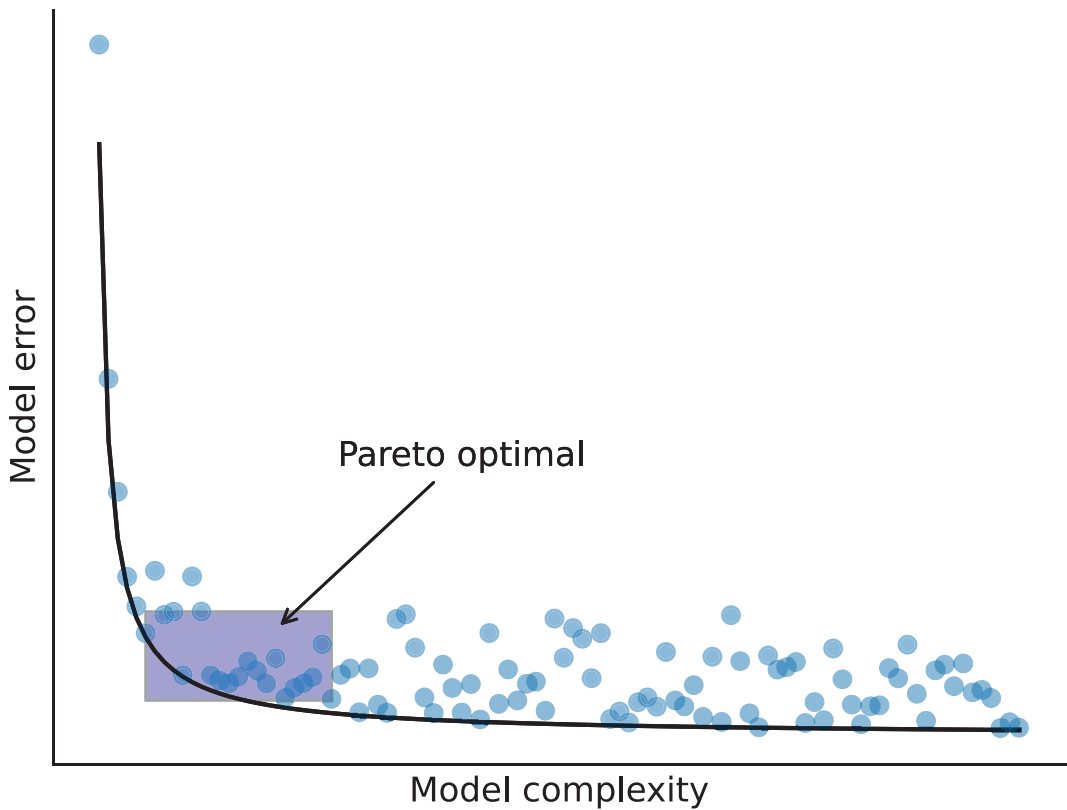
Fig. 3. A Pareto curve showing the schematized relationship between model accuracy and model complexity, after Brunton and Kutz (2022). An ideal model occupies the Pareto optimal space, which strikes a balance between accuracy and simplicity.

this would represent a major advance, but if extant models prove more successful than the discovered models then this is also an important empirical finding.

2. Discovering dynamical models from data

2.1. What does a good model look like?

There is a near-infinite number of *potential* models that could have generated a spoken utterance, but the number of *good* models is likely few in number. How do we know which models are good? A typical approach in model discovery is to strike a balance between two characteristics—parsimony and accuracy—which is schematized as complexity versus error in Fig. 3.

A maximally parsimonious model is a very simple model with as few terms as possible in the equation (top left of Fig. 3). A parsimonious model is highly interpretable because we

have a small number of model terms that make clear predictions. The downside is that those predictions are likely to be inaccurate, as they are a poor fit with the real world due to insufficient complexity. At the other extreme, a maximally accurate model would have a very high number of terms, which offer immense flexibility in fitting the model to highly variable data (bottom right of Fig. 3). An example of this would be a deep neural network, which can have upwards of many hundreds of parameters. The downside is that such a model is likely to be incredibly complex and uninterpretable, meaning we learn little about the system's fundamental dynamics. Our aim is to discover models that fit into the "Pareto optimal" space in Fig. 3, representing models that are simple but show high accuracy. The following section addresses the conceptual and technical solutions to this approach.

2.2. Sparse identification of nonlinear dynamics

A popular technique for data-driven model discovery in physics and engineering is the class of SINDy (Sparse Identification of Nonlinear Dynamics) methods that have emerged over the past decade (Brunton, Proctor, & Kutz, 2016). SINDy is based on the principles of symbolic regression (Schmidt & Lipson, 2009), which aims to approximate an unknown function from some data $X, \dot{X}$ as a combination of nonlinear functions.

$$\dot{X} = \Theta(X) \tag{7}$$

where $\Theta(X)$ is a feature library composed of an arbitrary number of mathematical functions.

$$\Theta(X) = [1 X X^2 X^3 \ldots \sin X \cos X] \tag{8}$$

The aim is to discover coefficients for the functions in $\Theta(X)$ that explain variation in $\dot{X}$. A known problem of symbolic regression is that the above procedure will produce many non-zero coefficients, leading to a complex model that contains many terms from the feature library. Such models are likely overfitted to the data and may be more complex than desirable. How can we only retain the terms that contribute substantially to the system under study? SINDy solves this by discovering the optimal *sparse* coefficient matrix Ξ corresponding to the functions in $\Theta(X)$.

$$\Xi = [\xi_1 \xi_2 \xi_3 \ldots \xi n] \tag{9}$$

The aim is to discover a coefficient matrix Ξ that provides an excellent fit to $\dot{X}$ while being as sparse as possible; that is, containing the smallest number of terms required to produce a good fit. This can be cast as Eq. 10.

$$\dot{X} = \Theta(X) \Xi \tag{10}$$

A range of sparsity-promoting algorithms exist to solve this problem, two of which are reviewed in Section 3.3. The sparsity-promoting characteristics of SINDy mean that a symbolic model derived from data often performs better at estimating parameters from new data than a higher dimensional neural network. Subsequent research has expanded the range of optimization methods for diverse problems in physics, biology, and engineering, providing a valuable toolkit for data-driven model discovery (e.g., Champion, Zheng, Aravkin, Brunton, & Kutz, 2020; Kaiser, Kutz, & Brunton, 2018).

SINDy holds great promise for learning interpretable and parsimonious models of brain and behavior. Dale and Bhat (2018) review extant approaches to equation discovery in the cognitive sciences and outline some prospects and challenges for applying SINDy to social and cognitive systems. They highlight difficulties associated with discovering models from noisy and continuously changing systems and pose a number of valuable recommendations for future research. To date, however, there have only been very limited applications of SINDy in cognitive science. In terms of related methods, Iskarous (2016) shows how least squares regression can be used to estimate dynamical models of action and perception from data, while Nalepka et al. (2019) model human multi-agent activity in games using dynamical motor primitives. Specific applications of SINDy include discovering mechanisms of human learning from experimental data (LaFolette, Yuval, Schurr, Melnikoff, & Goldenberg, 2024) and a proof-of-concept study on discovering dynamical models of speech (Kirkham, 2024).

2.3. An example: Discovering known models from simulated data

We now turn to a brief illustration of model discovery using SINDy on simulated data. This outlines the conceptual steps involved in model discovery and validates that the method can discover a known model. As an example, we generate simulated data using the damped harmonic oscillator in (11), as this is a widely used model of articulatory gestures in speech production. The equation is written with acceleration $\ddot{x}$ on the left-hand side and damping $b\dot{x}$ and stiffness $k(x - T)$ terms on the right-hand side, where T is the target or equilibrium position of the system. The damping coefficient b is typically defined as $b = 2\sqrt{k}$, which makes (11) a critically damped harmonic oscillator that will asymptotically approach the target. Note that $\ddot{x}$ is typically written as $m\ddot{x}$, but as $m = 1$ it is omitted from here onwards.

$$\ddot{x} = -b\dot{x} - k(x - T) \tag{11}$$

We solve a trajectory from this equation based on a set of initial position x_0 and velocity $\dot{x}_0$ values, and a stiffness k and target T value, all of which remain invariant over the time-course of the simulation. Numerical solutions were computed using Python's `scipy.integrate.solve_ivp` function (Virtanen et al., 2020), with a Runge–Kutta method of order 5(4) and a timestep of $\Delta t = 0.001$. Position and velocity trajectories were simulated using initial conditions $x_0 = 1$, $\dot{x} = 0$, and parameter values $k = 2,000$, $b = 2\sqrt{k} = 89.44$, and $T = 0.2$.

We pass the simulated position and velocity trajectories to a SINDy algorithm, with a first-degree polynomial library and a coefficient threshold of 0.1. Fig. 4 shows a predicted trajectory from the discovered model plotted on top of the data, with no visible differences between data and prediction, resulting in a fit of $R^2 = 1.00$. The discovered symbolic equation is $\ddot{x} = -b\dot{x} - kx + kT$, and we can rearrange terms to get $\ddot{x} = -b\dot{x} - k(x - T)$, which is the original equation that generated the simulated data. The discovered coefficients are $kT = 400.299$, $k = 2,001.31$, and $b = 89.47$. If we calculate $T = \frac{kT}{k}$ then $T = 0.2$, showing that all discovered coefficients are within 0.1% of their true values. This shows that SINDy is able to accurately discover symbolic models and coefficient values from data; Kirkham (2024) shows that the same model can be accurately discovered even when considerable noise
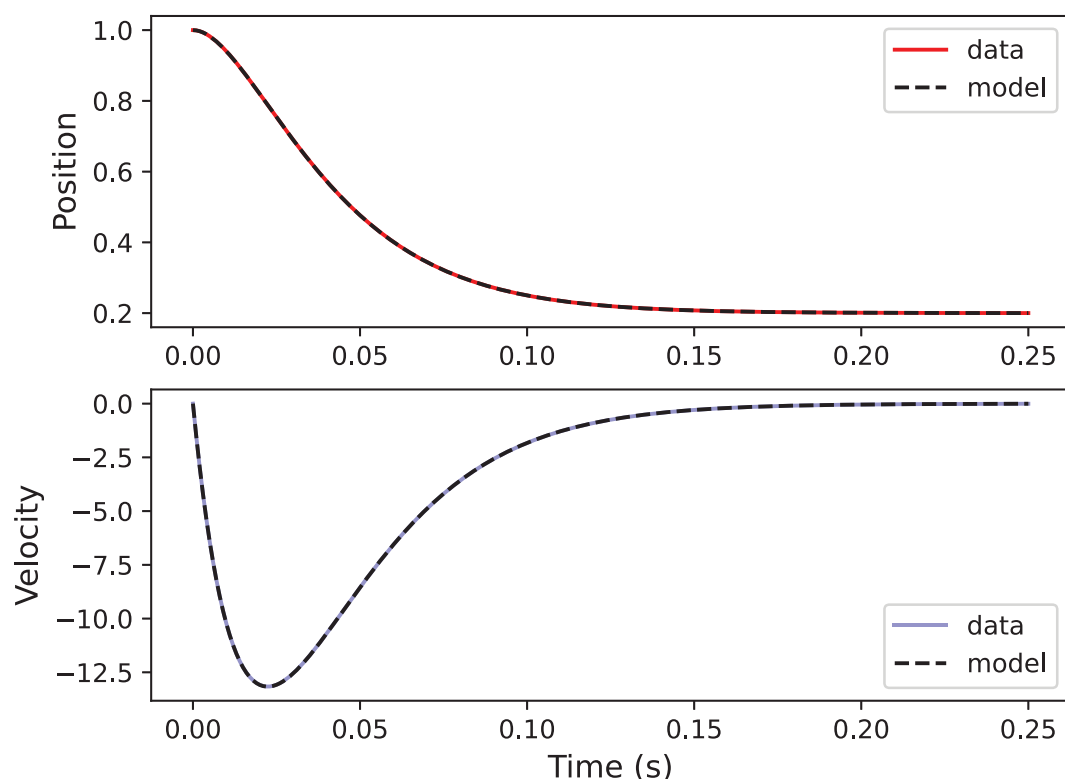
Fig. 4. Simulated position and velocity trajectories plotted against SINDy model predictions.

is added to simulated trajectories. The rest of this article applies this method to empirical data on movements of the tongue and lips during speech, with the aim of discovering accurate and interpretable models of gestural dynamics.

3. Methods

3.1. Data

We report an experiment demonstrating the use of the SINDy framework for discovering the articulatory dynamics of continuous speech. We use data from the X-Ray Microbeam (XRMB) corpus (Westbury, 1994), which contains articulatory speech data from a relatively large number of speakers. The XRMB corpus contains data from 57 speakers, most of whom speak an Upper Midwest dialect of American English (32 female, 25 male), with a median age of 21. We use a subset of 48 speakers, corresponding to those who have existing forced-aligned phone-level annotations available in the database at https://github.com/rsprouse/xray_microbeam_database. The data comprise small pellets of 2.5–3.0 mm in size attached to the upper (UL) and lower (LL) lips, surface of the tongue (T1, T2, T3, T4),

mandible, and head. The pellets were tracked using a narrow X-ray beam at 160 Hz for T1, 80 Hz for T2, T3, T4, and LL, and 40 Hz for UL, after which all pellets were resampled at 160 Hz and translated/rotated to an anatomically defined coordinate system. (See Westbury, 1994, for comprehensive technical documentation of the data and speaker sample.)

We discover models using data from XRMB tasks #11 and #101, both of which feature continuous speech of varying durations. Task #11 involved each speaker reading the following passage:

> You wish to know all about my grandfather. Well, he is nearly 93 years old, yet he still thinks as swiftly as ever. He dresses himself in an old black frock coat, usually several buttons missing. A long beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice quivers a bit. Twice each day he plays skillfully upon a small organ.

Task #101 involved each speaker producing the following three sentences in a single recording, with a short pause between each phrase:

- Elderly people are often excluded.
- When all else fails, use force.
- The dormitory is between the house and the school.

3.2. Data processing

Four articulatory variables were extracted from every recording: lip aperture (LA), tongue tip (TT, based on T1), tongue dorsum (TD, based on T3), and tongue root (TR, based on T4). We model gestural dynamics in one dimension, so we reduce all articulatory motions as follows. LA is the Euclidean distance between the vertical coordinates of the upper and lower lips, while TT, TD, and TR are the first principal component projected from the two-dimensional x/y coordinates, which was calculated separately for each speaker and each sensor (Birkholz, Kroger, & Neuschaefer-Rube, 2011; Elie, Lee, & Turk, 2023). We focus on discovering autonomous dynamical models that assume step activation of the type in Eq. 4, which has direct consequences for our approach to gestural segmentation. All signals were first divided into interpause intervals, based on the forced aligner's segmentation, and then gestural segmentation was calculated on the basis of zero-crossings in the velocity signal. A gesture is defined as a velocity peak or trough bounded by two velocity zero crossings. Some velocity trajectories have a peak that never crosses zero, meaning that such trajectories often have two or more peaks. Trajectories with two or more peaks were identified algorithmically and excluded from the analysis, which corresponds to 13.3% ($N = 2,924$) of trajectories. These trajectories were excluded because existing task dynamic models do not predict multiple velocity peaks per gesture. We also excluded trajectories longer than 200 ms in duration, because many of these represent passive speech movements during periods in which that articulatory variable was not actively involved in the production of a constriction. This corresponds to 24.4% ($N = 5,383$) of trajectories, which also includes a small number of additional trajectories that were clearly erroneous based on visual inspection.

In total, we analyze 62.3% ($N = 13{,}742$) of the segmented trajectories from the dataset. While this is a conservative approach to data filtering, it is likely that a large number of the excluded trajectories represent uncontrolled movements, anomalies at signal edges, segmentation errors, and inaccurate velocity calculations. We nonetheless retain a very large number of target gestures that are far beyond the amount possible to check manually. In total, we analyze 13,742 individual gestures across four articulatory variables (LA 3894, TT 3715, TD 3027, TR 3106). We discover models from each articulatory variable separately in order to test whether model discovery is robust across articulatory tasks.

3.3. Computational implementation

Our computational modeling is based on the Python package `pySINDy`, which is a computational framework for discovery of governing dynamical equations from data (de Silva et al., 2020). Each step of the model discovery procedure is explained below. Data and code for reproducing every analysis in this article are available at http://doi.org/10.5281/zenodo.15101639.

3.3.1. Model inputs

This study involves the discovery of two kinds of models: (1) first-order models, where the models only depend on the position and velocity of the system; and (2) second-order models, where the system includes position, velocity, and acceleration. The task dynamic models in Saltzman and Munhall (1989) and Sorensen & Gafos (2016) are both examples of second-order models, and it is well known that modeling skilled movements typically requires a second-order model (Saltzman & Kelso, 1987). To provide a point of comparison, however, we also test the hypothesis that articulatory movements can be approximated by first-order models, with no information about acceleration. First-order models can accurately model other physical dimensions of speech, such as fundamental frequency contours (Iskarous et al., 2024), although note that the models in Iskarous et al. (2024) are a series of coupled agonist-antagonist equations, which are more complex than the single-order models examined here.

The target system to be modeled is the intrinsic dynamics of the speech gesture, as defined in Section 3.2. This corresponds to an interval including the initiation of a gesture, its movement towards its target, and target achievement. The movement away from the target represents a new gesture, which is either a release gesture (e.g., returning to a rest position), or a movement towards a different target. The input data are a position signal for first-order models and a position signal with its associated velocity signal for second-order models. In no cases do we find substantially better model fits for opening versus closing gestures, which are instead reflected in simple parameter differences, so all models collapse across this distinction. The technical details of data processing are reported in Section 3.2.

3.3.2. Sequential thresholded least squares

We now review algorithms that promote sparsity in the discovered models. Recall from Section 2.2 that Eq. 12 defines the goal of model discovery, where the aim is to discover the sparse coefficient matrix Ξ corresponding to features in the library $\Theta(X)$ that optimally

model the time-derivative $\dot{X}$.

$$\dot{X} = \Theta(X)\Xi \tag{12}$$

One of the simplest methods is sequential thresholded least squares (STLSQ) (Brunton et al., 2016). STLSQ aims to (1) obtain a least squares solution for Ξ, (2) eliminate any coefficients below a pre-defined threshold, and (3) iterate until the procedure converges on an optimally sparse model. This procedure can be cast as the objective function in (13), where Ξ is the sparse coefficient matrix to be optimized and α defines the ℓ_2 regularization weight, which makes the problem a form of sequentially thresholded ridge regression. Note that the threshold parameter λ is not explicitly specified in the objective function as it is applied post hoc over repeated iterations of (13).

$$\min_{\Xi} ||\dot{X} - \Theta(X)\Xi||_2^2 + \alpha||\Xi||_2^2 \tag{13}$$

In the present study, we use STLSQ for first-order models because it provides an effective and simple technique for sparse model discovery. We set $\alpha = 0.05$, use a maximum of 20 optimization iterations to allow for convergence of the thresholding algorithm, and a coefficient threshold optimized to maximize R^2 model fit from the set $\lambda \in \{0.001, 0.01, 0.1\}$.

3.3.3. Sparse relaxed regularized regression

While STLSQ works well for first-order models that only involve a single time derivative (i.e., velocity), a second-order model introduces additional complexity that can be better constrained using alternative techniques. For example, when we numerically solve a second-order differential equation as in (14), we split it into two coupled first-order equations as in (15) and (16). This involves the introduction of a new variable y, requiring us to solve for y and $\dot{y}$.

$$\ddot{x} = -b\dot{x} - kx \tag{14}$$

$$y = \dot{x} \tag{15}$$

$$\dot{y} = -by - kx \tag{16}$$

A SINDy algorithm will simultaneously discover two equations for a second-order model. If the equations are of the form in (15) and (16) then we can simply substitute y into (16) and rearrange to obtain our second-order model for $\ddot{x}$. However, one consequence of the model discovery procedure is that SINDy will try to fit all terms in the library to both equations. In principle, this means that Eq. (15) could be something other than $y = \dot{x}$, such as (17).

$$y = 0.97\dot{x} + 0.21x - 0.13 \tag{17}$$

If we then substitute the value of y in (17) into (16), we will end up with the more complex model in (18). This represents a form of the generalized Liénard equation (Burton, 1965),

S. Kirkham / Cognitive Science 49 (2025)

where $h(x, \dot{x})$ is a damping or forcing function that depends on both position and velocity.

$$\dot{y} = -b(0.97\dot{x} + 0.21x - 0.13) - kx \tag{18}$$

While this level of complexity is not necessarily a problem, it results in a much less parsimonious model and probably contains more complexity than necessary. To solve this problem for second-order models, we use sparse relaxed regularized regression (SR3) (Champion et al., 2020). This allows to incorporate constraints on the model, such as placing bounds on coefficient values, forcing terms to be in a particular proportion to one another, or incorporating other forms of physical knowledge about the system. We use SR3-based constraints to enforce the following weak assumptions: (1) y in (15) is always equal $y \overset{!}{=} 1.00\dot{x}$, (2) all other potential terms in (15) always equal zero. This essentially constrains the damping function to $-b\dot{x}$ and prohibits additional complexity.

We use a constrained version of the SR3 algorithm that is conceptually analogous to STLSQ but with some important differences, such as the addition of a constraint matrix that makes the solution conditional on the specified constraints. This can be cast as the objective function in (19), where Ξ is the coefficient matrix to be optimized.

$$\min_{\Xi, W} \frac{1}{2}||\dot{X} - \Theta(X)\Xi||^2 + \lambda R(W) + \frac{1}{2\nu}||\Xi - W||^2 \tag{19}$$
$$\text{subject to } C\xi = d$$

The term $\frac{1}{2}||\dot{X} - \Theta(X)\Xi||^2$ measures the fit between data and model based on the sum of squared differences, where $\frac{1}{2}$ is a scaling factor that simplifies the derivatives. $R(W)$ is a regularization function that acts as a prior on sparsity promotion; we specifically use weighted ℓ_0 regularization, which is a non-convex function that can handle multiple local minima in the optimization landscape. The term W is a proxy variable for Ξ that allows us to decouple model fitting and regularization, such that Ξ can be optimized on the data and subsequently regularized to promote sparsity, with this sequence iterated during the optimization procedure. This improves numerical stability and admits greater flexibility in determining a wellfitting model. The coupling term $\frac{1}{2\nu}||\Xi - W||^2$ ensures that W and Ξ remain close and do not substantially diverge from one another. The λ parameter weights the regularization function, where $\lambda = \eta^2/2\nu$. This is sparsity promoting, where η is the threshold for the minimum coefficient magnitude in Ξ and ν determines the closeness of the match between Ξ and W.

We use a constrained variant of SR3, where optimizing the objective function in (19) is subject to a matrix of linear constraints, using the vectorized form $\xi = \text{vec}(\Xi)$. The optimization must meet the condition that $C\xi = d$, where C is a matrix specifying which terms in ξ are subject to constraints, and d is a vector defining the constraint values. In our case, $C\xi$ defines the terms in the first equation $y = \dot{x}$ that are subject to constraints, with d specifying a value of 1.0 for the term $\dot{x}$. This is what allows us to impose the specified constraints on the damping term in second-order differential equations. A proof of the convergence properties of the constrained SR3 algorithm can be found in Champion et al. (2020). In terms of hyperparameters, we use a maximum of 30 iterations to allow for convergence of the optimization

algorithm, with $\nu = 1$ and a coefficient threshold optimized to maximize R^2 model fit from the set $\eta \in \{0.001, 0.01, 0.1\}$.

3.3.4. Train-test data split

For the purposes of modeling, we randomly split the data for each articulatory variable into 80% training and 20% test sets. The training set is used for library comparison and an initial fit, where the number of discovered terms is allowed to vary between tokens. We then fit the best overall model to the test set, where the model must fit the full library (representing the best model structure) to each trajectory. We also report visualizations of model predictions on a random sample of tokens from the test set. Note that the training set is used to discover a symbolic model, rather than a statistical model, and it is this symbolic model that is fitted to the test set.

3.3.5. Library selection

We begin by sensitivity testing the candidate feature library because the number of terms in a final model can be highly sensitive to the thresholding parameter, especially with single token fits, and this allows us to compare the fits across different feature libraries. As articulatory signals are well approximated as the sum of polynomials, we use a series of polynomial libraries across first (x), second (x, x^2), third (x, x^2, x^3), and fourth (x, x^2, x^3, x^4) degrees. We fit each polynomial library to the training data using the model ensembling technique reported below and calculate summary statistics for each library. This allows us to establish the relative merits of different polynomial libraries and make transparent decisions when two libraries perform very similarly. The selected feature library is then refitted to the training data, and we report a wider range of summary statistics, which is outlined in more detail below.

The threshold hyperparameter for library comparison was optimized for each articulatory variable from the set $\{0.001, 0.01, 0.1\}$, with the final threshold value based on the highest R^2 value. In some cases, lower thresholds did not converge due to an ill-conditioned or stiff model that resulted in numerically unstable predictions. In other cases, higher polynomial libraries performed worse than lower polynomial libraries. This can appear surprising because adding an additional polynomial term to a well-performing model should not harm performance. This phenomenon arises, however, because a greater number of terms often reconfigures the model in a way that changes the magnitude of each coefficient term. In other words, a quadratic model is not necessarily the linear model with its original coefficients plus a quadratic term but can sometimes be a fundamentally different model in terms of the relationship between coefficients. As a result, even a very small threshold can eliminate key terms in such models, or result in higher polynomials fitting to noise in the signal. We address this by sensitivity testing threshold values as above and selecting feature libraries that exhibit stability across the dataset. In cases where additional complexity provides only minor performance improvements, we subsequently explore whether greater complexity simply improves the quantitative fit of models or reveals fundamentally different qualitative dynamics of the system.

3.3.6. Model ensembling

In order to discover models across large datasets, we use a model ensembling technique, whereby models are fitted to each token separately and then an ensemble model is derived

from this set of models. While a single SINDy model can be fitted using multiple trajectories, simulations show that this is generally only effective for speech gestures if the target or equilibrium position is the same or very similar across tokens. In such a case, a single model and a single set of coefficient values will be returned for the whole dataset. However, when key parameters vary—such as stiffness, damping, or target—then the resulting SINDy model will contain values for these parameters that are the average best fit to all of the tokens. As we seek good symbolic models and accurate parameters, we instead construct an ensemble model from models fitted to individual trajectories. As the model discovery procedure is repeated for each token, each trajectory could theoretically be fitted with different numbers of terms, especially in the case of larger feature libraries. We leverage this fact in order to obtain distributions on the number of terms in each model across the dataset, which allows us to then arrive at an ensemble model based on the majority model structure. Once this final model structure is determined, we then fit it to the test data, forcing the same structure on each test trajectory.

3.3.7. Generating predictions

Once we have discovered a model for each token, we then use this model to make a prediction. Predictions are generated by taking the discovered model, the discovered parameter coefficients, and a set of initial conditions from the data, which comprise the initial position and velocity value in each empirical trajectory. These initial conditions are then used to solve the discovered model forwards in time, determined by the discovered coefficient values. In an ideal scenario, this should generate position and velocity trajectories that are identical to the original data. In practice, however, the prediction is only as good as the discovered model, which allows us to use the prediction as an estimate of model fit. We quantify the fit using by-trajectory R^2 scores. All R^2 values are variance weighted, meaning that the R^2 for model fit is an average over position and velocity, weighted by the variance of each signal, which provides a more informative assessment of model performance given that position and velocity are on different numerical scales. In conventional regression analysis, R^2 values are bounded between [0,1], but the lower bound is a consequence of allowing either the intercept or the slope to vary, which is the aim of regression analysis. With a constrained intercept, however, R^2 is negative when the prediction is worse than simply fitting a horizontal line through the data (Chicco, Warrens, & Jurman, 2021). We fix the intercept for each model prediction as the initial conditions from empirical data, so R^2 will be negative when the model prediction is worse than a horizontal line through the data.

4. Discovering new models from data

4.1. First-order models

4.1.1. Library comparison

Table 1 shows the results for library comparison on first-order models. The optimal threshold for the models was $\lambda = 0.001$ (LA), 0.1 (TT), 0.01 (TD), and 0.001 (TR). The first-degree

Table 1

Comparison of different polynomial libraries in first-order models across articulatory variables (training set). The model scores are R^2 mean (standard deviations)

Number of Polynomials	LA	TT	TD	TR
1	0.40 (0.57)	0.56 (0.17)	0.53 (0.24)	0.51 (0.29)
2	0.96 (0.05)	0.96 (0.05)	0.96 (0.05)	0.96 (0.03)
3	0.96 (0.04)	0.97 (0.04)	0.96 (0.03)	0.96 (0.03)
4	0.98 (0.02)	0.98 (0.02)	0.99 (0.02)	0.99 (0.02)

Table 2

Model fit statistics for first-order models with polynomials up to third-degree. All values rounded to two decimal places

Training Set	LA	TT	TD	TR
mean(R^2)	0.96	0.97	0.96	0.96
$\sigma(R^2)$	0.04	0.02	0.03	0.03
min(R^2)	0.63	0.66	0.63	0.45
max(R^2)	1.00	1.00	1.00	1.00
Test Set	**LA**	**TT**	**TD**	**TR**
mean(R^2)	0.95	0.95	0.95	0.95
$\sigma(R^2)$	0.01	0.02	0.01	0.02
min(R^2)	0.77	0.78	0.84	0.84
max(R^2)	0.99	0.99	0.99	0.99

library performs poorly, with R^2 values across the four articulatory variables of {0.40, 0.56, 0.53, 0.51}. Libraries with two to four polynomials perform well with R^2 between 0.96 and 0.99. Overall, while the second-degree library performs well on average, it contains some negative R^2 values for LA, TT, and TD. By contrast, the lowest score for any articulatory variables in the third-degree library is $R^2 = 0.63$ (LA). For this reason, we select the third-degree library containing x, x^2, x^3, which we more thoroughly evaluate in the following section.

4.1.2. Results

Table 2 shows summary statistics for the first-order models with a third-degree polynomial library fitted to the training data. All four terms (a constant, x, x^2, x^3) were found for the majority of models (LA = 99.78%, TT = 64.67%, TD = 93.68%, TR = 99.64%), but some models omitted the cubic term (LA = 0.22%, TT = 35.23%, TD = 6.28%, TR = 0.36%) and a small percentage comprised only linear terms (TT = 0.1%, TD = 0.04%)

All models perform fairly well, with mean $R^2 = 0.96$ or above and minimum R^2 of 0.45–0.66. The test data were then fitted using the same algorithm, but all trajectories were forced to have linear, quadratic, and cubic terms. The test data show comparable performance, with mean $R^2 = 0.95$ for each articulatory variable and higher minimum R^2 values in each case,

(a) LA (Lip Aperture)

(b) TT (Tongue Tip)

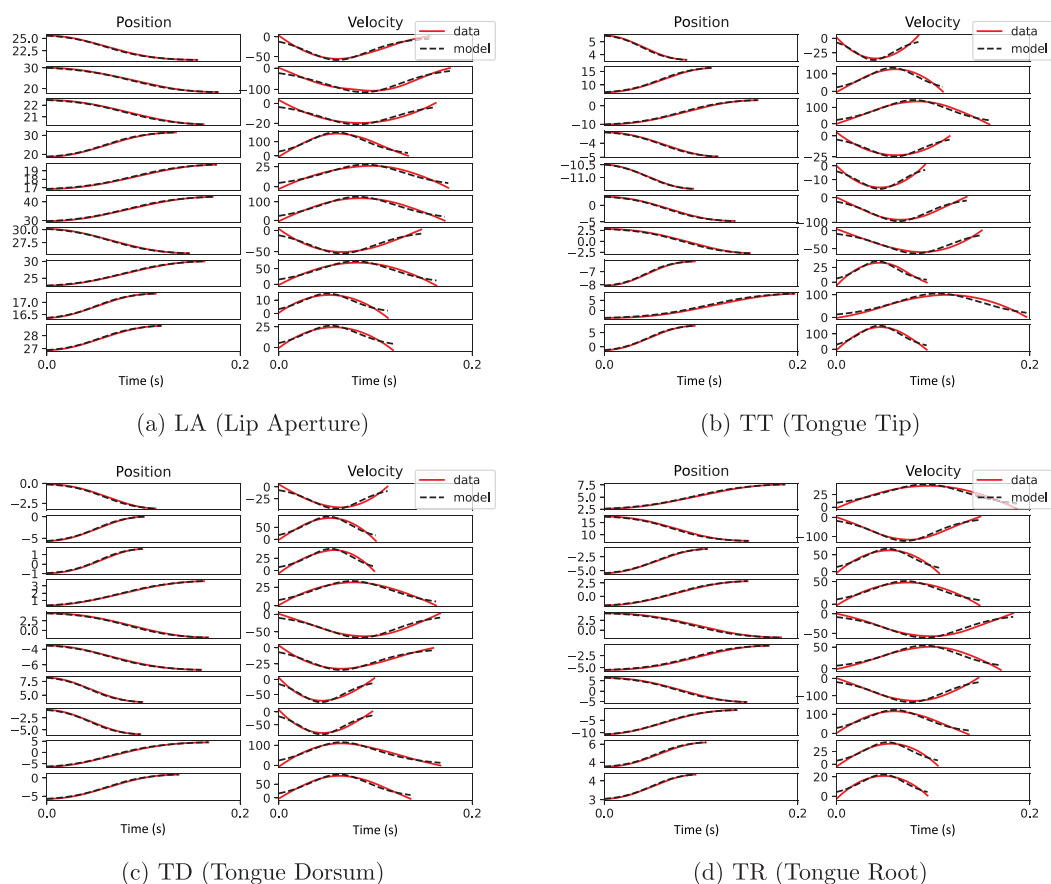(c) TD (Tongue Dorsum)

(d) TR (Tongue Root)

Fig. 5. Ten randomly sampled trajectories for each articulatory variable showing the fit between data and model predictions for first-order models with a third-degree polynomial library on the test data.

ranging from 0.77 to 0.84. The higher performance in the test data is because every token is forced to contain all terms in the library (i.e., $\lambda = 0$).

Fig. 5 shows 10 randomly sampled position and velocity trajectories for each variable, with a comparison of data and model prediction from the test data. The quantitative fit is good but with some errors in the intercept and also errors in the fit along the curves. As a result of minor errors in the position data, the resulting velocity signals show bigger errors. This is a consequence of a limited degree of flexibility in the possible shape of the velocity curve, due to the lack of an acceleration term governing change in velocity.

4.1.3. Summary

The discovered model takes the form of the symbolic equation in (20), where a is a constant and b, c, d are the coefficients of x, x^2, x^3. This makes the discovered model a cubic equation, where velocity is dependent on a constant, as well as the current position multiplied by a

Table 3
Comparison of different polynomial libraries in second-order models across articulatory variables (training set). The model scores are R^2 mean (standard deviations)

Number of Polynomials	LA	TT	TD	TR
1	0.99 (0.02)	0.98 (0.03)	0.98 (0.02)	0.99 (0.02)
2	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
3	0.99 (0.05)	0.98 (0.08)	0.99 (0.07)	0.99 (0.03)
4	−1.14 (6.26)	−3.24 (14.26)	−3.05 (19.57)	−2.00 (10.73)

coefficient, the square of the current position multiplied by a coefficient, and the cube of the current position multiplied by a coefficient. The presence of x^2, x^3 makes it a nonlinear model.

$$\dot{x} = a - bx + cx^2 - dx^3 \tag{20}$$

In summary, a first-order nonlinear model appears to be a reasonable quantitative fit across four different articulatory variables.

4.2. Second-order models

4.2.1. Library comparison

Table 3 shows the library comparison of second-order models. In this instance, the first, second, and third-degree libraries perform near-identically in terms of summary statistics, with mean $R^2 = 0.98$–0.99, although the first-degree models have slightly lower standard deviations. The optimal threshold for all models was $\eta = 0.001$. The fourth-degree library performs poorly with a very small threshold of $\eta = 0.001$; if we relax the threshold to zero then the fourth-degree library performs marginally better than all other libraries, but there is clearly no need for a model of this level of complexity. In addition to this, all models above second-degree have some negative R^2 fits, suggesting that the greater complexity forces some important coefficients to values smaller than the threshold. Overall, there appears to be little benefit in higher polynomial libraries based on this comparison. As a result, we select the first-degree library but evaluate the impacts of any additional complexity in Section 5.4.

4.2.2. Results

Table 4 shows summary statistics for the second-order models with the first-degree polynomial library. In the training data, 100% of models contain three terms, with mean $R^2 > 0.98$ in all cases. The test data show comparable performance, with minimum R^2 values of {0.77, 0.72, 0.66, 0.76}.

Fig. 6 shows 10 randomly sampled position and velocity trajectories for each variable, with a comparison of data and model prediction from the test data. The quantitative fit is excellent, with near-perfect fits for all trajectories. Notably, the fits look substantially better than those in Fig. 5, suggesting that the second-order model is superior in quantitative fitting accuracy.

S. Kirkham / Cognitive Science 49 (2025)

Table 4
Model fit statistics for second-order models with polynomials up to first-degree. All values rounded to two decimal places

Training Set	LA	TT	TD	TR
mean(R^2)	0.99	0.98	0.98	0.99
$\sigma(R^2)$	0.02	0.03	0.02	0.02
min(R^2)	0.75	0.70	0.64	0.70
max(R^2)	1.00	1.00	1.00	1.00
Test Set	LA	TT	TD	TR
mean(R^2)	0.99	0.98	0.99	0.98
$\sigma(R^2)$	0.02	0.03	0.02	0.02
min(R^2)	0.77	0.72	0.66	0.76
max(R^2)	1.00	1.00	1.00	1.00

(a) LA (Lip Aperture)

(b) TT (Tongue Tip)

(c) TD (Tongue Dorsum)
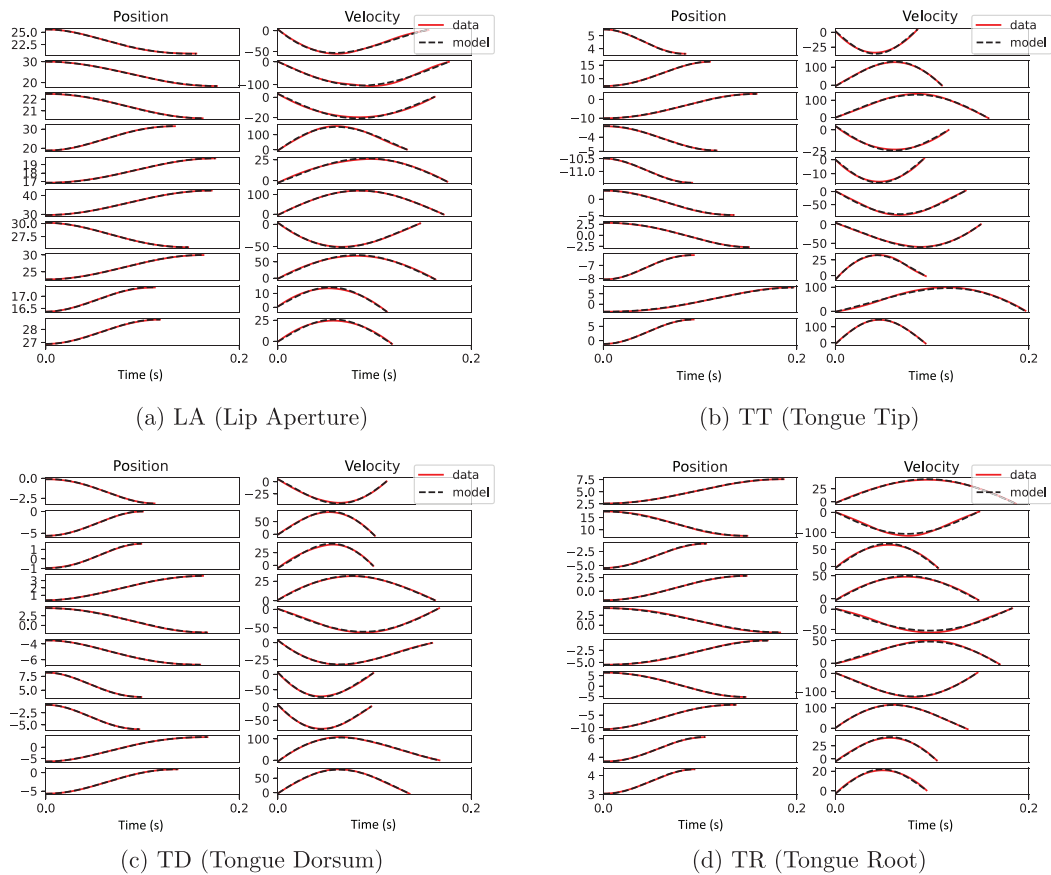
(d) TR (Tongue Root)

Fig. 6. Ten randomly sampled trajectories for each articulatory variable showing the fit between data and model predictions for second-order models with a first-degree polynomial library on the test data.

4.2.3. Summary

The discovered model takes the form of the symbolic equation in (21).

$$\ddot{x} = kT - kx - b\dot{x} \tag{21}$$

Note that SINDy models sometimes discover a constant term, but the analysis in Section 2.3 shows that this is often the term kT, such as that $kT - kx = -k(x - T)$. As such, this model takes the form of the harmonic oscillator in (22), which is a standard task dynamic model (Saltzman and Munhall, 1989).

$$\ddot{x} = -b\dot{x} - k(x - T) \tag{22}$$

In summary, a second-order linear model is a very good fit to data across four different articulatory variables. The mean accuracy of the fit is $R^2 = 0.98$ and above in all cases, with no trajectories being scored less than $R^2 = 0.64$ in either the training or test datasets. In Section 5, we conduct further interpretation of these terms and explore the implications of this model.

4.3. Interim summary

Two models fit the data very well: a first-order nonlinear model with quadratic and cubic terms, and a second-order linear model. The performance of both models is quantitatively similar, but this may be a consequence of the mean scores for both models being relatively close to ceiling. An inspection of random trajectories plotted from each model reveals that while both models fit the data well, the second-order fits are more accurate, likely a consequence of the acceleration term in the second-order models. The following sections take the two broad classes of models discovered here and explore them in greater detail. Specifically, we interpret the meaning of each model's terms in light of known systems and explore the model space via computational simulations.

5. Exploring the discovered models

5.1. Overview

So far, we have two well-fitting models, but at this point we must go beyond treating these simply as effective fits to data and understand the ways in which they govern the dynamical laws of speech. In this section, we take two models from Section 4—a first-order model and a second-order model—and explore them deeper, focusing on how to interpret the different terms in the equations as well as what predictions and assumptions they make about speech gestures. In doing so, we also explore the effects of adding complexity to the second-order model.

The analysis for each section proceeds as follows. We first plot representative examples of the data against model predictions, followed by plotting the qualitative dynamics of the system in the form of phase portraits and Hooke portraits (Beek and Beek, 1988; Mottet &
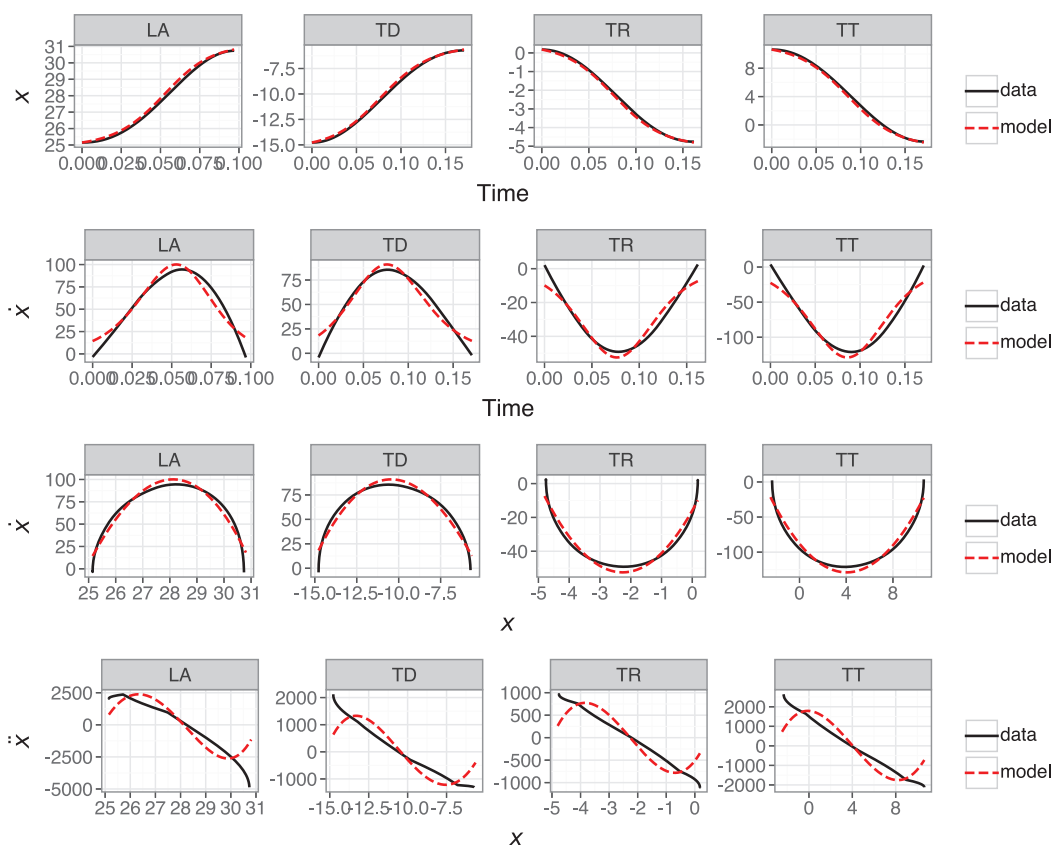
Fig. 7. First-order model predictions showing time-varying position (row 1), time-varying velocity (row 2), phase portrait (row 3), and Hooke portrait (row 4). The R^2 scores for each model are LA = 0.96, TD = 0.95, TR = 0.96, and TT = 0.95.

Bootsma, 1999). This is an essential step in moving beyond model assessment via simple data fitting because a fundamental characteristic of dynamical models of skilled movement is how they specify the relationship between position, velocity, and acceleration. We then explore the equations analytically, deriving algebraic properties that expose similarities to other well-understood systems, before exploring the dynamics of the relevant terms using computational simulations. We note that a comprehensive investigation of every aspect of each model and its numerical parameterization is beyond the scope of the current article; instead, we here focus on elucidating fundamental aspects of each system.

5.2. First-order model

Fig. 7 shows representative examples of the first-order model predictions, with position and velocity trajectories, as well as a phase portrait (position $\sim$ velocity) and Hooke portrait (position $\sim$ acceleration). While the position data are predicted with good accuracy, the higher
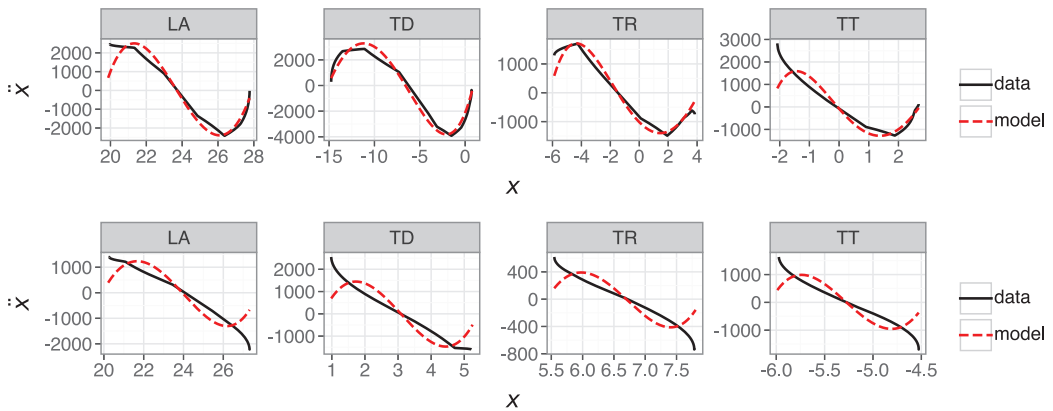
Fig. 8. Hooke portraits for the top-scoring (top row) and median-scoring (bottom row) first-order models for each articulatory variable.

derivatives and phase portraits show the inadequacy of a first-order model. The velocity trajectory and phase portrait show how the velocity estimates at signal edges are systematically incorrect as well as showing small but systematic errors in peak velocity and time-to-peak velocity. The Hooke portrait particularly highlights the poor predictions of the first-order model, where uniform nonlinearity is predicted despite quasi-linearity in the empirical data.

The above suggests that R^2 scoring of position and velocity leads to misleading conclusions about the true capacity of the first-order model. To explore this further, Fig. 8 shows Hooke portraits of data and predictions for the best scoring (top row) and median-scoring (bottom row) first-order nonlinear models. The best fitting model is a case with extensive nonlinearity between position and acceleration, especially for the TD articulatory variable. However, similarly nonlinear predictions are also made for the median-scoring models, despite the data showing a quasi-linear relation in these cases. This is a consequence of the first-order model containing no information about the system's higher derivatives, such as acceleration; this makes it impossible to accurately model change in velocity, which is particularly evident at movement onsets.

In conclusion, this analysis shows that a first-order nonlinear equation is not an appropriate model of articulatory control. While the SINDy analysis revealed a good fit to the empirical position and velocity trajectories, this did not take into account the *nature* of the mismatches between the model and data. In this instance, the phase portrait and Hooke portrait point towards a fundamental issue with the model, rather than minor errors in quantitative fit. These findings suggest that this model is insufficient for capturing the dynamical characteristics of articulatory movements. In addition, the first-order model contains a greater number of parameters (with quadratic and cubic terms), which is inevitable given the lack of terms for controlling higher derivatives. This shows that the first-order model's complexity is not warranted due to its theoretical inadequacy. As a consequence, we do not consider this model any further and move on to the second-order linear model.
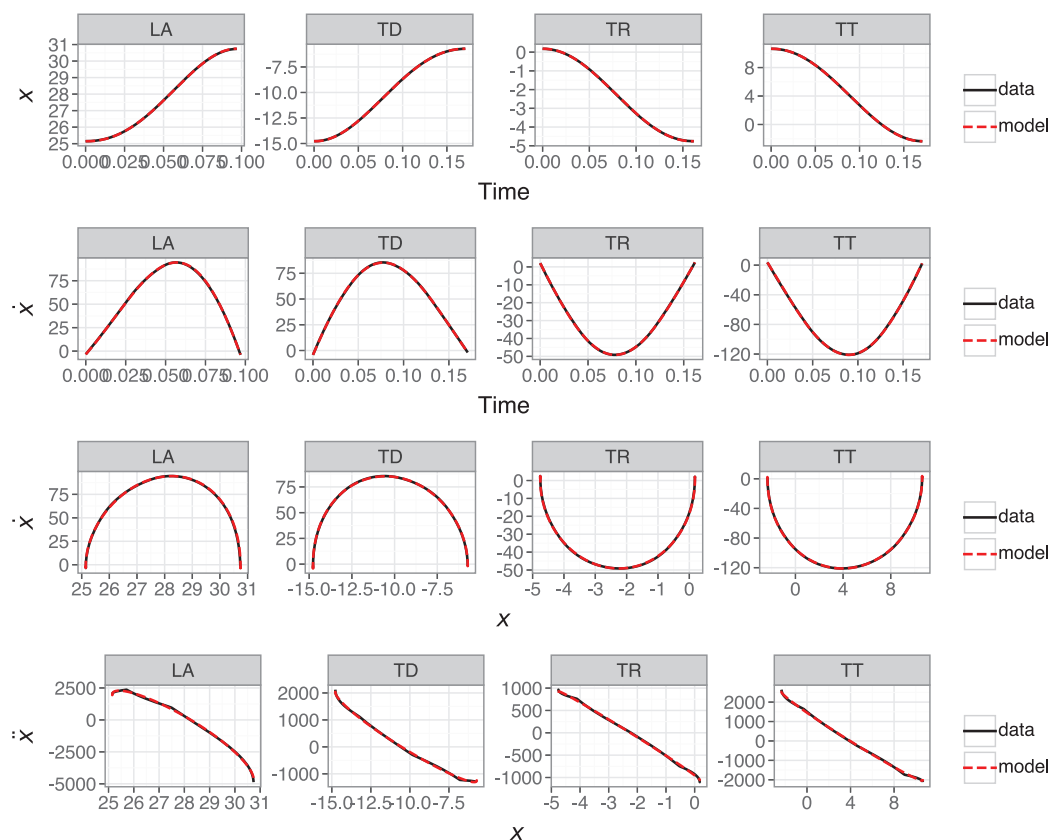
S. Kirkham / Cognitive Science 49 (2025)

Fig. 9. Linear second-order model predictions showing the time-varying position (row 1), time-varying velocity (row 2), phase portrait (row 3), and Hooke portrait (row 4). The SINDy scores for each model are LA = 0.999, TD = 0.999, TR = 0.999, and TT = 0.999.

5.3. Second-order linear model

Fig. 9 shows data and predictions for the best fitting second-order model for each articulatory variable. In contrast to the first-order model in Section 5.2, the second-order model is an excellent fit across position and velocity, the corresponding phase portrait, and the Hooke portrait. Notably, the Hooke portraits show quasi-linearity, suggesting that the dynamics are well approximated by harmonic motion (Beek & Beek, 1988). This is a key signature of a linear oscillator, in contrast to the anharmonicity that would suggest nonlinear dynamics.

We note that Fig. 9 captures cases that correspond to highly linear models. The remainder of this section focuses on these cases and their interpretation, but in Section 5.4 we explore cases where the linear model fits worse and consider whether additional complexity is warranted. For now, we turn to the second-order model that was discovered in Section 2. This
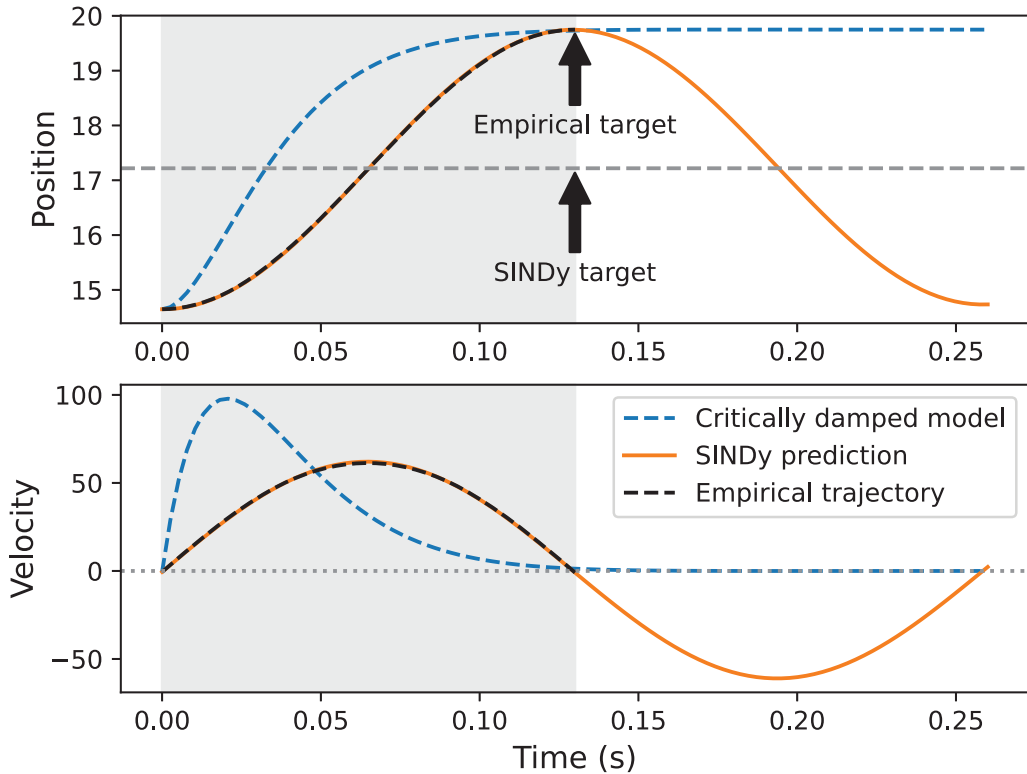
Fig. 10. Dynamics of the second-order linear model with SINDy-discovered parameters. The dashed black line is the empirical trajectory with the best fitting model, the orange line is the SINDy model prediction ($T = 17.22$, $k = 592$, $b = 0.264$, $R^2 = 0.999$), the dashed blue line represents the coefficients needed to reach the empirical target under the assumption of critical damping ($T = 19.75$, $k = 2,500$, $b = 2\sqrt{k}$). The gray-shaded area corresponds to the duration of the empirical trajectory, with the subsequent unshaded region showing the effects of continuing the simulation beyond this duration. The label "SINDy target" corresponds to the SINDy-discovered value of T, whereas the label "empirical target" is the empirical position value corresponding to the final velocity minimum.

equation took the form in (23):

$$\ddot{x} = -b\dot{x} - k(x - T) \tag{23}$$

This corresponds to the form of a linear harmonic oscillator, but it is clear that the system is not critically damped, as a critically damped version of (23) is unable to fit the symmetrical empirical velocity profiles as accurately as seen in Section 4.2 (Sorensen & Gafos, 2016). How, then, does the system achieve its equilibrium position? Fig. 10 shows the trajectory with the best fitting SINDy model ($R^2 = 0.999$), which is a Tongue Root trajectory produced by speaker JW45. The gray-shaded area shows the empirical duration of the trajectory, with the dashed black line showing the empirical data. The orange line shows SINDy predictions, which are near-perfect fits to the data, but with an important characteristic: this is only true

during the time period that corresponds to the empirical trajectory. If we continue the simulation beyond the trajectory's original duration the system begins to oscillate.

To unpack this further, the dashed blue trajectory shows the trajectory that a critically damped oscillator would need to take to reach the empirical target. This does reach the target, but with a very early time-to-peak velocity, demonstrating the poor empirical fit of a critically damped model. In contrast, the behavior that the SINDy parameters capture is as follows. It drives the system towards an equilibrium value $T = 17.22$, which is the SINDy discovered coefficient for T. This is below the empirical target $T = 19.75$ and is clearly not the true target in the sense of the gestural system (i.e., the position value at the final velocity zero-crossing). To disambiguate, we henceforth refer to the actual empirical target at the velocity zero-crossing as T and SINDy's discovered virtual target as T_v (the term "virtual target" is used entirely to refer to the SINDy discovered target and we make no claims about its theoretical status at this stage). Importantly, the SINDy discovered trajectory shows oscillatory behavior if extended beyond the empirical duration, as shown by the period after the gray-shaded area (i.e., $t > 0.13$).

How are we to relate the SINDy target $T_v = 17.22$ to the empirical target $T = 19.75$? The horizontal dashed line in Fig. 10 shows the value of T_v. We can see that the orange line reaches this target twice: once before empirical T and once after empirical T, which is only visible if we extend the simulation beyond the empirical duration of the fitted trajectory. The oscillation is due to the lack of damping $b \approx 0$, but there is a strong intrinsic relationship between T and T_v, such that T_v can be easily derived from (24), which is equivalent to saying that the virtual target is half the distance between the initial condition and T. We can then solve for T as in (25).

$$T_v = x_0 + \frac{T - x_0}{2} \tag{24}$$

$$T = 2T_v - x_0 \tag{25}$$

If we substitute (25) into our second-order harmonic oscillator model and rearrange terms then we get (26). This suggests an alternative hypothesis, where the model contains $\frac{k}{2}(T + x_0)$. This new equation still requires tuning of b, k to avoid the system producing oscillations when $b < 2\sqrt{k}$, which we address below, but it allows us to formulate a model where T captures the empirically observed target of the system. The only new parameter in the model is x_0 (initial position), which any dynamical system necessarily already has access to in calculating a trajectory, so it is still an autonomous system. The only modification here is the use of a more complex constant term $\frac{k}{2}(T + x_0)$.

$$\ddot{x} = -b\dot{x} - kx + \frac{k}{2}(T + x_0) \tag{26}$$

The new equation captures cases where the velocity trajectory is very close to a half-cycle sine wave, representing an undamped or minimally damped oscillator where $b \approx 0$. For example, if we use absolute-valued SINDy-discovered parameters for T, x_0 based on Eq. 25 then
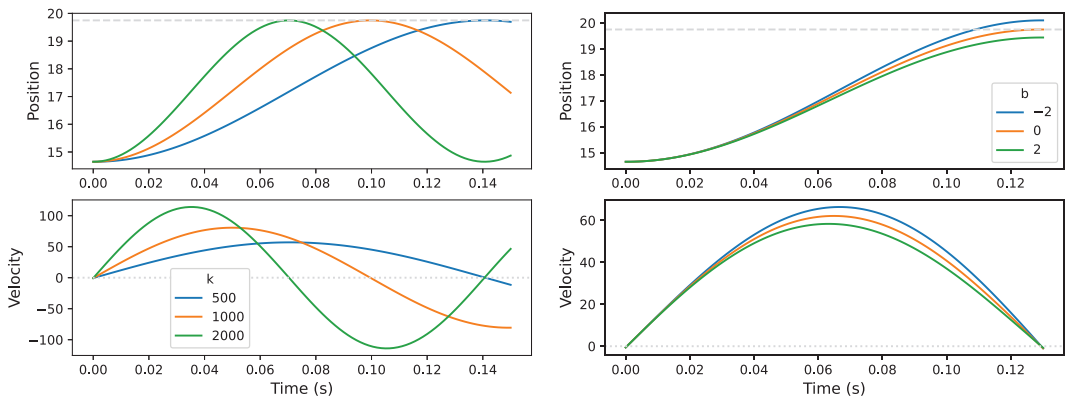
Fig. 11. The effect of varying k on time-to-target achievement when $x_0 = 14.65$, $T = 19.75$, $b = 0$. The dashed line in the upper panel represents the value of T. (left). The effect of varying b for the same trajectory, where $k = 592.74$ (right).

$|T|$ is strongly correlated with the positional value of $|x|$ at the final velocity minimum for LA ($r = .94$) and TR ($r = .91$), but only moderately correlated for TT ($r = .71$) and TD ($r = .82$). There appears to be no discernible difference between opening/closing gestures, with only minimally higher correlations for closing gestures (e.g., largest difference is LA closing $r = .94$ vs. LA opening $r = .92$). The differences between articulatory variables could be a consequence of potential differences in velocity segmentation quality. For example, correlations under the simpler formulation $\ddot{x} = -b\dot{x} - k(x - T)$ are LA $= -0.79$, TT $= -0.72$, TD $= -0.83$, and TR $= -0.91$, suggesting that only LA benefits from the formulation in (26). The other articulatory variables show greater variability in this respect, although the above example shows a case where TR velocity is highly symmetrical, suggesting that it improves model fit for some tokens but is not otherwise detrimental. Despite this, we note that the mean b values differ only slightly between articulatory variables, so this model's improved performance for LA could be a consequence of data processing (e.g., one-dimensional lip aperture vs. compression of horizontal/vertical movements into a single dimension) or other properties of lip movements (e.g., reduced coarticulation from neighboring lingual movements).

Given the above, how are the appropriate parameters determined for a given duration? Fig. 10 shows simulated trajectories based on the best fitting model, where $x_0 = 14.65$, $T = 19.75$. The left panel of Fig. 11 shows how changing $k \in \{500, 1{,}000, 2{,}000\}$ results in the target being met at shorter durations as k is increased (when $b = 0$). The system oscillates after target achievement, due to the system being undamped, so the gesture must be deactivated at target achievement. The right panel of Fig. 11 shows that, for this example, target achievement occurs when $b = 0$; when b is positive the trajectory is damped and undershoot occurs, when b is negative overshoot occurs. We note that the version of the model with $\frac{k}{2}(T + x_0)$ represents a specific case where the intended target is met, velocity trajectories are symmetrical, and the dynamics are highly linear.

In summary, the best linear model appears to be an under-damped harmonic oscillator. The model can be slightly improved by adding $\frac{k}{2}(T + x_0)$, which relates the target to the initial condition. This adds some degree of complexity, largely in the dynamics around the equilibrium position, but note that this is not substantial and it fundamentally only involves adding a constant term to the standard harmonic oscillator model. It is worth further probing the theoretical implications of such a model, but note that recent advances in nonlinear modeling of speech gestures also incorporate information on initial position and movement distance directly into the differential equation (Kirkham, 2025).

5.4. Adding complexity: Second-order nonlinear model

Section 5.3 shows that a linear model can capture movement dynamics with high accuracy, but recall that the selected examples were the best fitting models. It should not be a surprise that a linear model accurately models dynamics that show strong linear signatures. How representative is this of the second-order model's overall performance? Fig. 12 repeats the Hooke portrait for the best scoring models (top row) and compares this with median-scoring models (second row), fifth percentile scoring models (third row), and first percentile scoring models (fourth row) for each articulatory variable in the linear model. The first and fifth percentile models are those where only 1% or 5% of tokens are lower scoring than these models. Note that the median scoring models the R^2 values range between [0.989, 0.992], the fifth percentile models are in the range [0.936, 0.965], and the first percentile models are in the range [0.849, 0.915]. Recall that these scores are based on position and velocity, not acceleration, which is why they look more optimistic about the model's performance compared with the Hooke portraits. The median model is a good approximation of the data, while the fifth and first decile models struggle with the greater nonlinearity between position and acceleration. Note that this failure of the model is much clearer in the relationship between higher derivatives because the comparison of position and velocity trajectories for these tokens do not obviously reveal this behavior.

The lower scoring models in Fig. 12 correspond to cases where there is greater nonlinearity in the relation between position and acceleration. This indicates anharmonicity and is not within the scope of a linear harmonic oscillator (Sorensen & Gafos, 2016). In summary, while the first-order model can capture nonlinearity in the Hooke portraits, it overpredicts the extent of nonlinearity and fails to adequately model the empirical characteristics of most of the data. In contrast, a second-order linear model is a better empirical fit but lacks the ability to capture the nonlinearity present in some of the data. It stands to reason that a second-order nonlinear model should combine the strengths of the two approaches.

Thus far, the results suggest a role for a nonlinear term in the second-order model. How extensive is nonlinearity in the data? Fig. 13 shows the degree of linearity in the Hooke portrait for all trajectories, based on R^2 values from by-trajectory linear regression fits between position and acceleration (Mottet & Bootsma, 1999). Note that these R^2 values only correspond to the data and capture the linear fit between position and acceleration; they do not refer to any fit between the second-order model predictions and the data. To disambiguate, we subsequently refer to R^2 values for the Hooke portraits as R^2_H.
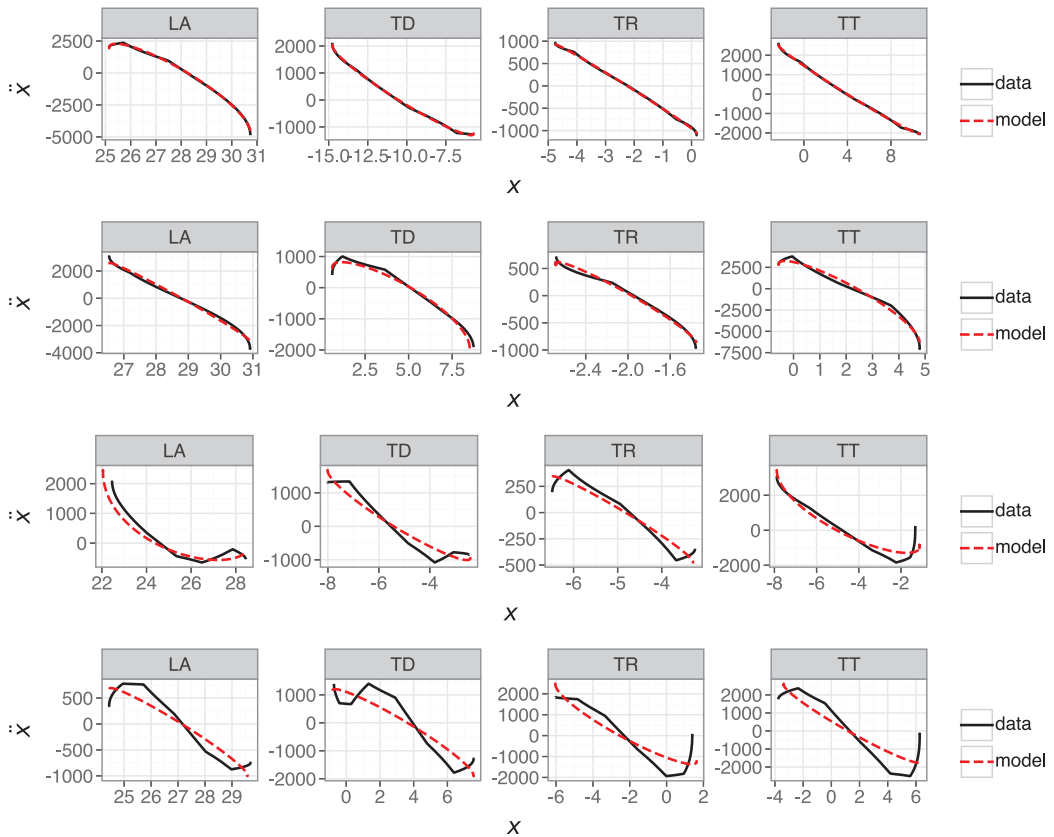
Fig. 12. Hooke portraits for second-order linear models with the highest score (top row), median score (second row), fifth percentile score (third row), and first percentile score (fourth row) for each articulatory variable.

It is obviously challenging to pose a specific R_H^2 value that indicates nonlinearity in the Hooke portrait, as opposed to an underperforming fit due to measurement or segmentation errors. The lower row of Fig. 13 (first percentile) provides sufficient context, however, because the empirical trajectories here are clearly nonlinear, with associated R_H^2 values of 0.9 (LA), 0.85 (TD), 0.82 (TR), and 0.83 (TT). This suggests that $R_H^2 \leq 0.9$ certainly indicates substantial nonlinearity that is outside of the scope of a linear model. Across all trajectories, an average of 69% of tokens across articulatory variables have $R_H^2 > 0.95$, whereas 15% of tokens have $R^2 < 0.9$. From this, we can assume that approximately 30% of tokens have a substantial degree of nonlinearity and that around half of these show extensive nonlinearity. Note that correlations between R_H^2 and spatial displacement are $r < .1$ for all articulatory variables. This suggests that nonlinearity does not straightforwardly interact with the magnitude of spatial displacement, as is predicted by a nonlinear model (Sorensen & Gafos, 2016), although see Kirkham (2025) for a nonlinear model where these dynamics are optionally under the control of the speaker.
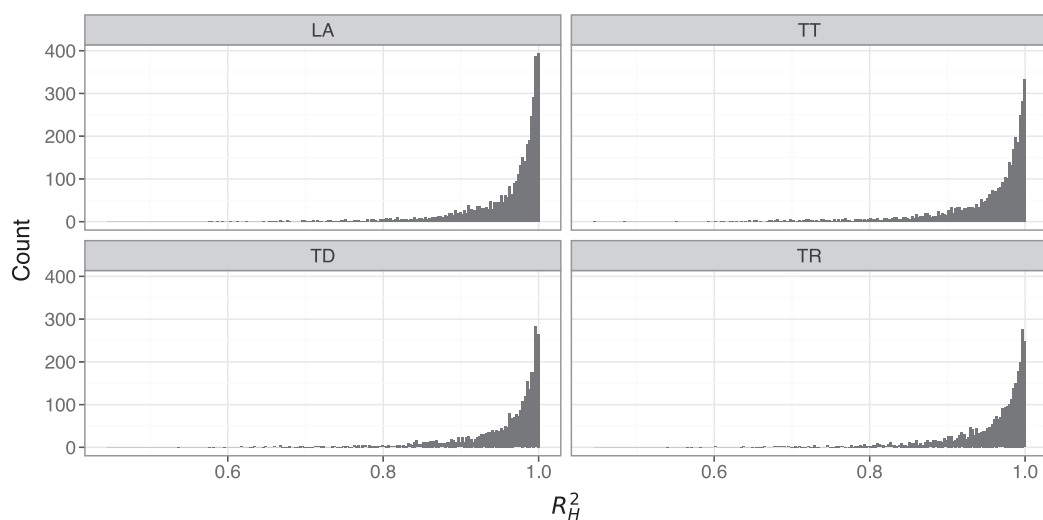
Fig. 13. Distribution of R_H^2 values for each sensor, based on a by-token linear regression of position against acceleration in the empirical trajectory data.
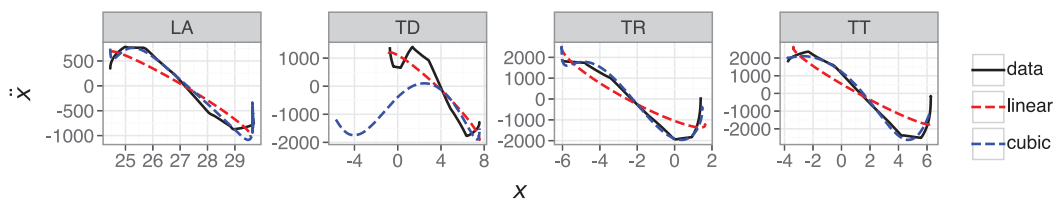

Fig. 14. Hooke portraits for linear and cubic second-order models. Note that the models for LA and TT contain x^3 and $\dot{x}^3$, whereas TD and TR only contain x^3.

   That the majority of tokens show linearity in the Hooke portrait is why the linear model in Section 5.3 shows good performance, but this analysis makes clear that a nonlinear term is required in order to account for the full dynamics of articulatory trajectories. In this case, the Hooke portraits reveal that some of the additional complexity in the higher polynomial libraries from Section 4.2 may be warranted. Fig. 14 shows the same Hooke portraits for the first percentile trajectory fits as in Fig. 12 (lower row), but with an additional cubic term in the feature library. Models were fitted using the same SINDy SR3 algorithm but with a custom library comprising linear terms and a cubic term. The cubic term was initially constrained to position-only, but the LA and TT models required both x^3 and $\dot{x}^3$ in order to improve on the linear model. This is not necessarily unusual for models of human movement (Beek & Beek, 1988; Schöner, 1990), and there may be an advantage to the inclusion of nonlinear velocity terms more generally, especially for modeling qualitatively distinct movement dynamics, such as limit cycles (Kuberski and Gafos, 2023).

The nonlinear model clearly provides a better fit than the linear model for trajectories in Fig. 14, with the exception of TD, where SINDy fails to find an optimal model. This suggests that the anharmonicity indicated by nonlinear relations between position and acceleration is well within the scope of a nonlinear model, as shown by Sorensen & Gafos (2016). It is clear that a single cubic term on position x^3 can provide an excellent nonlinear fit for the TR variable in Fig. 14, while the poor fit for TD and the presence of $\dot{x}^3$ for LA and TT are more likely to represent either (i) fitting to errors in the data or (ii) a failure of the SR3 algorithm to find an optimal solution.

Parameterization of the cubic model is clearly more challenging than the linear model and further tests showed that the SR3 algorithm failed to find an optimal fit for a cubic model in many cases, despite the empirical data clearly being within the model's scope. Indeed, a cubic model cannot be easily derived by taking the SINDy coefficients for the linear model and adding a cubic term. For example, the linear coefficient values in the discovered linear and cubic models differ substantially, such that the linear model's damping coefficient indicates very weak damping, whereas the cubic model involves more substantial (but still subcritical) damping. One potential source of difficulty in model fitting may be the complexity of the cubic coefficient, which can take on a very wide range of values depending on movement amplitude, even if movement characteristics are otherwise similar. Kirkham (2025) outlines a method for scaling the cubic coefficient by (actual or potential) movement amplitude. This constrains the parameter search to the range [0, 1], which should improve model fitting and interpretability. It may also be the case, as discussed above, that a cubic term on velocity is necessary to fully encompass the range of speech movements, which remains an open line of inquiry for future research.

5.5. Interim summary

In summary, we have explored and interpreted three candidate models. The first-order model is nonlinear, but analysis of phase and Hooke portraits revealed the insufficiency of this model for capturing the full dynamics of the trajectories. We show that the second-order linear model is a good approximation of the majority of the data, accurately capturing the dynamics for around 2/3 of trajectories, but there is significant nonlinearity in the Hooke portrait for around one third of trajectories. This is to be expected given our liberal velocity threshold, as Kuberski & Gafos (2023) show that thresholded segmentation can lead to underestimates of the extent of nonlinearity. It should be noted that we still find considerable evidence of quasi-linear relations in the Hooke portrait, but a straightforward comparison between studies is challenging. For example, Kuberski & Gafos (2023) only examine closing movements (whereas we examine opening and closing movements) and they only examine repeated syllables at different metronome rates (whereas we include a wider range of speech materials). Regardless, the presence of nonlinearity points towards the need for greater model complexity, which can be adequately captured with the addition of a cubic term.

Recall that in all cases the models are only minimally damped, which deviates from previous models that assume critical damping. Specifically, a perfectly symmetrical velocity trajectory can be achieved when the linear model is undamped, where the value of k

determines time-to-target achievement, but this requires a mechanism to deactivate the gesture upon reaching the target. Finally, a nonlinear model is clearly required to account for around one third of the data. This suggests that the movement dynamics of speech are fundamentally nonlinear, even though the nonlinear force may be minimal in some cases. In summary, the nonlinear model that best captures the full range of variation in the current data corresponds to a version of Sorensen & Gafos (2016) model without critical damping

$$\ddot{x} + a(t)[b\dot{x} + k(x - T) - d(x - T)^3] = 0. \tag{27}$$

A more complex version of this model optionally transforms the linear damping term $b\dot{x}$ into a nonlinear damping force $b\dot{x}^3$

$$\ddot{x} + a(t)[b\dot{x}^3 + k(x - T) - d(x - T)^3] = 0. \tag{28}$$

In both cases, gestural dynamics are autonomous during activation, where $a(t)$ is step activation

$$a(t) = \begin{cases} 1, & t \in [t_a, t_b], \\ 0, & \text{otherwise}. \end{cases} \tag{29}$$

The following section now discusses the theoretical implications of these models, as well as limitations and prospects for model discovery in cognitive science.

6. Discussion

6.1. An evaluation of discovered models

The first-order model initially shows good accuracy, in the sense that it provides an approximate fit to position and velocity trajectories. At first glance, any issues with the first-order model may appear to concern small deviations in fitting accuracy, but an analysis of the Hooke portraits reveals more fundamental issues. Specifically, the first-order model fails to accurately reproduce the relations between position and acceleration. In other words, what might look like minor differences in quantitative accuracy for position and velocity estimation is actually a failure to capture the fundamental dynamics of the system. This renders a first-order autonomous model insufficient for capturing articulatory dynamics and highlights the importance of exploring the scope of the discovered model's predictions, rather than only fitting to empirical trajectories.

The second-order model shows better empirical fits than the first-order model and the highest performing cases show excellent fits in the phase and Hooke portraits. Under the linear interpretation, we find that the target, stiffness, and damping terms need to be tuned in just the right way to meet the target in the specified time interval. The biggest differences between the second-order model presented here and previous task dynamic models are (1) we relax the critical damping constraint; (2) target achievement is non-asymptotic. In other words, the specific target value is achieved, rather than the system becoming asymptotically close.

A second-order linear model with minimal damping and a reformulation of the target term is capable of accurately modeling the majority of articulatory trajectories. As with the first-order model, however, cases of underperformance do not simply reflect minor issues in quantitative fit or data quality issues. Instead, the Hooke portraits show that the linear model is incapable of capturing the observable nonlinearity between position and acceleration, which occurs in approximately one third of trajectories. This motivated a second-order nonlinear model with a cubic term (Sorensen & Gafos, 2016), which is typically under-damped rather than critically damped. While the nonlinear model is significantly more difficult to parameterize, it is able to accurately model the observed nonlinearities. This confirms that the articulatory dynamics of speech are fundamentally nonlinear and adds considerable support to the Sorensen & Gafos (2016) model of articulatory control.

6.2. Autonomous dynamics and beyond

Beek & Beek (1988) outline a typology of dynamical models of rhythmic movement, including (1) simple linear models with complex external forcing, (2) nonlinear autonomous models, and (3) nonlinear models with minimal external forcing. Approach (1) is a relatively common theme in computational and modular models of mind, where movement models are simple and most of the work is offloaded to a complex forcing function. This essentially reflects a view in which physical movement is the implementation stage of highly complex cognitive processing. In contrast, the present analysis has focused mainly on the second of Beek and Beek's (1988) typology, if we also allow the inclusion of some good-performing linear models. Indeed, we confirm that the intrinsic dynamics of articulatory trajectories can be modeled without any explicit time dependence across a large database of 13,742 segmented trajectories. Our approach has been to assume very simple step-activation driving of the system, with instantaneous change in parameters at specific landmarks. However, this rather simple driving mechanism is likely insufficient, given that it does not explain how the gestural system is driven from one target to another.

A more comprehensive dynamical model would likely occupy the third category of Beek and Beek's (1988) typology, with nonlinear gestural dynamics and an external forcing function $F(t)$ that drives between system states. A clear example concerns the forces that drive the initiation and termination of gestures. This is particularly pertinent to the present study as the linear second-order model is close to undamped and the coefficients discovered for a cubic second-order model are almost never critically damped (but significantly more damped than the linear model). This means that both models require some form of gestural suppression mechanism, otherwise the system will inherently oscillate around the target. One proposal is state feedback on target achievement (Burroni & Tilsen, 2022; Tilsen, 2022), where speakers use a combination of internal and external feedback to open and close a gestural gating function. Parrrell, Ramanarayanan, Nagarajan, and Houde (2019) also outline a model of hierarchical state feedback control, which combines task dynamics with nonlinear state estimation.

An alternative proposal rejects the idea that gestures have bounded activation intervals altogether, instead casting gestures as always active but varying in their force on the vocal tract

(Tilsen, 2018). Under this view, movement preparation is represented as a dynamic neural planning field (Erlhagen & Schöner, 2002; Schöner, 2020), where articulatory control variables correspond to a parameter field. Gestures act as inputs to the planning fields, with the field's activation centroid determining the parameter value. In this view, the parameters of minimal dynamical models are recast as dynamic neural fields that generate continuous values from activation across neural populations (Kirkham & Strycharczuk, 2024; Kirov & Gafos, 2007; Roon & Gafos, 2016; Stern & Shaw, 2023; Tilsen, 2019a). The task of model discovery thus becomes uncovering the dynamics of neural fields that continuously parameterize gestural systems. This view clearly situates gestural systems as explicitly non-autonomous because parameters are no longer constant and are constantly being fed by movement planning fields.

At this juncture, it should be noted that the distinction between autonomous and non-autonomous does not necessarily have to correspond to a distinction between intrinsic versus extrinsic timing (Fowler, 1980), where the latter is associated with a central clock or time-keeper (e.g., Turk and Shattuck-Hufnagel, 2020). It is self-evident that no living system is autonomous because interaction is a signature of life (Suprunenko, Clemson, & Stefanovska, 2013), but the present study shows that the intrinsic dynamics of independent gestures can be adequately modeled as autonomous systems. That said, even if non-autonomous models turn out to be the correct direction, any time-varying parameters may still be a consequence of coupling to other dynamical mechanisms with their own intrinsic dynamics, rather than a central timekeeper. This is consistent with accounts of interoceptive rhythms in brain and body that can be coupled with the surrounding environment (e.g., Engelen, Solcà, & Tallon-Baudry, 2023). To this end, the most productive perspective may be a view of brain and behavior as a complex multiscale system with interacting, coupled, and emergent dynamics (e.g., Favela, 2024; Goheen et al., 2024; Kluger, Allen, & Gross, 2024; Senkowski & Engel, 2024; Tilsen, 2009). It is the task of future research to understand the nature of these dynamics of their coupling relations.

6.3. Prospects for data-driven model discovery in cognitive science

A major aim of the present work was to discover dynamical models from data. In doing so, we discovered some new models but also leveraged extant models in improving on these discoveries, such as the second-order nonlinear model in Sorensen & Gafos (2016). It is worth noting that articulatory trajectories are comparably easier to model than some other dynamical mechanisms in the cognitive sciences, but there remain a range of areas in which data-driven model discovery represents a promising direction.

For example, any model that is concerned with the relationship between discrete categories and their physical realization must take seriously both signed and spoken languages. While there is a very small amount of work exploring the possibility of a task dynamics of signed languages (e.g., Mertz, Pagel, Turco, & Mücke, 2024), this represents a space where model discovery would be particularly useful as a starting point, given appropriate kinematic corpora on signing. Debates also abound on the appropriate dynamical representation of various kinds of disordered speech, such as whether the addition of noise is sufficient (Mücke et al., 2024) or whether a different set of compensatory dynamics is involved.

While physical movements are obvious candidates for model discovery, this does not preclude data-driven discovery of higher level cognitive processes. This includes decision-making, perception, and working memory, all of which can be cast as relatively simple dynamical models (Schöner, Spencer, & The DFT Research Group, 2016) and are, therefore, potential areas where data-driven model discovery could be productive. More broadly, it is likely that moving beyond intrinsic speech dynamics towards a larger agent–environment system will also require a broader reconceptualization of the dynamical models we deploy, with an open system tending towards massive degrees of interactivity and complexity. This is likely to require a focus on how interacting task demands constrain the dynamics of brain and behavior, rather than attempting to model specific cognitive processes (Iskarous, 2016; Nau, Schmid, Kaplan, Baker, & Kravitz, 2024).

6.4. Limitations and future research

A productive avenue for future research would be more extensive comparisons between models (e.g., Elie et al., 2023), especially on languages other than English (Geissler & Nellakra, 2024). This is important given the emergence of new nonlinear models (Stern & Shaw, 2024) and ongoing developments in theories of articulatory control (Tilsen, 2016, 2018, 2020). It will be particularly important to compare the qualitative predictions made by different models in order to distinguish quantitative fit from qualitative adequacy (e.g., via phase and Hooke portraits). In future research, we also hope to examine bidirectional coupling between the dynamics of neural fields and the dynamics of nonlinear gestural models, which are likely mediated by dynamical feedback mechanisms (e.g., Parrrell et al., 2019; Tilsen, 2022). There is extensive scope for model development in these areas, but attention should also be directed towards developing rigorous ways of testing the predictions of different models.

A limitation of the current approach is the temporal segmentation of gestures and the method's reliance on accurate segmentation. We segmented signals at velocity zero-crossings, based on well-understood characteristics of skilled movement dynamics. If two gestures overlap then we are not able to distinguish them or estimate the parameters of each gesture separately. But this conceptualization is likely to be a gross simplification of the actual dynamics of gestures. As discussed above, Tilsen (2019a) outlines a model in which gestures are always active, but most are subthreshold at any point in time. In terms of model discovery, this becomes a significantly more complex task, but it is possible that a combination of sparse symbolic regression, predictive control algorithms, and neural networks may prove a fruitful avenue for model discovery in this area (Kaiser et al., 2018; Tilsen, 2020).

7. Conclusions

Discovering the dynamics that govern brain and behavior is a major challenge in the cognitive sciences. We have demonstrated one approach to meeting this challenge, applied to articulatory movement dynamics in spoken language. Building upon decades of research in

task dynamics and articulatory phonology, combined with recent developments in machine learning and equation discovery, we discover interpretable models directly from data. While a linear second-order equation accurately models around two thirds of trajectories in the data, a nonlinear (cubic) model is fundamentally necessary for accurately capturing the qualitative dynamics of speech movements. This supports the proposal that articulatory dynamics are well-modeled as a nonlinear autonomous system during periods of constant gestural activation. This leads us to propose that the discovered models represent the dynamical laws of motion that structure articulatory control in spoken language.

Acknowledgments

Note

1 For excellent introductions to dynamical systems and nonlinear dynamics, see Strogatz (2015) (a mathematical introduction); Abraham and Shaw (1992) (a visual introduction); van Gelder (1998) (a perspective from the cognitive sciences); Kelso (1995) (coordination), (Haken, 1977) (synergetics); Gafos (2006) (phonology); and Tilsen (2019b) (an impressive application of dynamical theory to syntactic structures). For arguments against computational metaphors of mind, see Gibson (1979), Carello, Turvey, Kugler, and Shaw (1984), Spivey (2007), Chemero (2009), and Barrett (2011).

References

Abakarova, D., Fuchs, S., & Noiray, A. (2022). Developmental changes in coarticulation degree relate to differences in articulatory patterns: An empirically grounded modeling approach. *Journal of Speech, Language, and Hearing Research*, *65*(9), 3276–3299.

Abraham, R. H., & Shaw, C. D. (1992). *dynamics: the geometry of behavior*. Redwood City, CA: Addison-Wesley.

Anderson, P. (1972). More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science*, *177*(4047), 393–396.

Barrett, L. (2011). *Beyond the brain: How body and environment shape animal and human minds*. Princeton, NJ: Princeton University Press.

Beek, P., & Beek, W. (1988). Tools for constructing dynamical models of rhythmic movement. *Human Movement Science*, *7*(2–4), 301–342.

Birkholz, P., Kroger, J., Bernd, & Neuschaefer-Rube, C. (2011). Model-based reproduction of articulatory trajectories for consonant-vowel sequences. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(5), 1422–1433.

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

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology*, *3*(1), 219–252.

Brunton, S. L., & Kutz, J. N. (2022). *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge, England: Cambridge University Press.

Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, *113*(15), 3932–3937.

Burroni, F., & Tilsen, S. (2022). The online effect of clash is durational lengthening, not prominence shift: Evidence from Italian. *Journal of Phonetics*, *91*(101124).

Burton, T. A. (1965). The generalized Liénard equation. *Journal of the Society for Industrial and Applied Mathematics Series A Control*, *3*(2), 223–230.

Byrd, D., & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics*, *26*(2), 173–199.

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

Carello, C., Turvey, M. A., Kugler, P. N., & Shaw, R. E. (1984). Inadequacies of the computer metaphor. In M. S. Gazzaniga (Ed.), *Handbook of cognitive neuroscience* (pp. 229–248). New York, NY: Plenum Press

Champion, K., Zheng, P., Aravkin, A. Y., Brunton, S. L., & Kutz, J. N. (2020). A unified sparse optimization framework to learn parsimonious physics-informed models from data. *IEEE Access*, *8*, 169259–169271.

Chartier, J., Anumanchipalli, G. K., Johnson, K., & Chang, E. (2018). Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. *Neuron*, *98*(5), 1042–1054.

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*(e623), 1–24.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York, NY: Harper and Row.

Dale, R., & Bhat, H. S. (2018). Equations of mind: Data science for inferring nonlinear dynamics of socio-cognitive systems. *Cognitive Systems Research*, *52*, 275–290.

de Silva, B. M., Champion, K., Quade, M., Loiseau, J.-C., Kutz, J. N., & Brunton, S. L. (2020). PySINDy: A Python package for the sparse identification of nonlinear dynamical systems from data. *Journal of Open Source Software*, *5*(49), 2104.

Elie, B., Lee, D. N., & Turk, A. (2023). Modeling trajectories of human speech articulators using general tau theory. *Speech Communication*, *151*, 24–38.

Engelen, T., Solcà, M., & Tallon-Baudry, C. (2023). Interoceptive rhythms in the brain. *Nature Neuroscience*, *26*, 1670–1684.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, *109*(3), 545–572.

Favela, L. H. (2024). *The ecological brain: Unifying the Sciences of brain, body, and environment*. New York, NY: Routledge.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, *8*(1), 113–133.

Gafos, A. I. (2006). Dynamics in grammar. In L. Goldstein, D. Whalen, & C. T. Best (Eds.), *Laboratory phonology 8: varieties of phonological competence* (pp. 51–79). Berlin, Germany: Mouton de Gruyter

Gafos, A. I., & Benus, S. (2006). Dynamics of phonological cognition. *Cognitive Science*, *30*(5), 905–943.

Geissler, C., & Nellakra, J. (2024). Predicting articulatory landmarks with critically-damped oscillators and general tau theory. *Proc. ISSP 2024–13th International Seminar on Speech Production*, 173–176.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Goheen, J., Wolman, A., Angeletti, L. L., Wolff, A., Anderson, J. A., & Northoff, G. (2024). Dynamic mechanisms that couple the brain and breathing to the external environment. *Communications Biology*, *7*(938), 1–11.

Guenther, F. H. (2016). *Neural control of speech*. Cambridge, MA: MIT Press.

Haken, H. (1977). *Synergetics: An introduction*. Berlin. Germany: Springer-Verlag.

Haken, H., Kelso, J. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, *51*(5), 347–356.

Hooke, R. (1678). *De Potentia Restitutiva, or of spring: Explaining the power of springing bodies*. London: John Martyn.

Hume, D. (1739). *A treatise of human nature*. London: John Noon.

Iskarous, K. (2016). Compatible dynamical models of environmental, sensory, and perceptual systems. *Ecological Psychology*, *28*(4), 295–311.

Iskarous, K. (2017). The relation between the continuous and the discrete: A note on the first principles of speech dynamics. *Journal of Phonetics*, *64*, 8–20.

Iskarous, K., Cole, J., & Steffman, J. (2024). A minimal dynamical model of intonation: Tone contrast, alignment, and scaling of American English pitch accents as emergent properties. *Journal of Phonetics*, *104*, 101309.

Kaiser, E., Kutz, J. N., & Brunton, S. L. (2018). Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *474*, 20180335.

Keating, P. (1990). The window model of coarticulation: Articulatory evidence. In J. Kingston& M. E. Beckman (Eds.), *Papers in laboratory phonology I* (pp. 3–29). Cambridge, England: Cambridge University Press

Kelso, J. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT Press.

Kelso, J. S., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics*, *14*(1), 29–59.

Kirkham, S. (2024). Discovering dynamical models of speech using physics-informed machine learning. In *Proceedings of the ISSP 2024–13th International Seminar on Speech Production* (pp. 185–188). Baixas, France: International Speech Communication Association (ISCA).

Kirkham, S. (2025). Scaling laws for nonlinear dynamical models of articulatory control. *JASA Express Letters*, *5*(2), 1–7.

Kirkham, S., & Strycharczuk, P. (2024). A dynamic neural field model of vowel diphthongisation. In *Proceedings of the ISSP 2024–13th International Seminar on Speech Production*, 193–196. Baixas, France: International Speech Communication Association (ISCA).

Kirov, C., & Gafos, A. I. (2007). Dynamic phonetic detail in lexical representations. In *Proceedings of the 16th International Congress of Phonetic Sciences*, 637–640. Saarbrücken, Germany: Universität des Saarlandes.

Kluger, D. A., Allen, M. G., & Gross, J. (2024). Brain-body states embody complex temporal dynamics. *Trends in Cognitive Sciences*, *28*(8), 695–698.

Kröger, B. J., Schröder, G., & Opgen-Rhein, C. (1995). A gesture-based dynamic model describing articulatory movement data. *Journal of the Acoustical Society of America*, *98*(4), 1878–1889.

Kuberski, S. R., & Gafos, A. I. (2023). How thresholding in segmentation affects the regression performance of the linear model. *JASA Express Letters*, *3*, 095202.

LaFolette, K., Yuval, J., Schurr, R., Melnikoff, D., & Goldenberg, A. (2024). *Data driven equation discovery reveals non-linear reinforcement learning in humans*. OSF. https://doi.org/10.31234/osf.io/65jqh.

Mertz, J., Pagel, L., Turco, G., & Mücke, D. (2024). Gradiency and categoriality in the prosodic modulation of French sign language: A kinematic approach using electromagnetic articulography. *Proceedings of Speech Prosody 2024*, 851–855.

Mottet, D., & Bootsma, R. J. (1999). The dynamics of goal-directed rhythmical aiming. *Biological Cybernetics*, *80*(4), 235–245.

Mücke, D., Roessig, S., Thies, T., Hermes, A., & Mefferd, A. (2024). Challenges with the kinematic analysis of neurotypical and impaired speech: Measures and models. *Journal of Phonetics*, *102*, 101292.

Nalepka, P., Lamb, M., Kallen, R. W., Shockley, K., Chemero, A., Saltzman, E., & Richardson, M. J. (2019). Human social motor solutions for human-machine interaction in dynamical task contexts. *Proceedings of the National Academy of Sciences*, *116*(4), 1437–1446.

Nau, M., Schmid, A. C., Kaplan, S. M., Baker, C. I., & Kravitz, D. J. (2024). Centering cognitive neuroscience on task demands and generalization. *Nature Neuroscience*, *27*, 1656–1667.

Newton, I. (1687). *Philosophiae naturalis principia mathematica*. London: Jussu Societatis Regiæ ac Typis Joseph Streater.

Noether, E. (1918). Invariante variationsprobleme. *Nachrichten der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, *1918*, 235–257.

Ostry, D. J., & Munhall, K. G. (1985). Control of rate and duration of speech movements. *Journal of the Acoustical Society of America*, *77*(2), 640–648.

Parrrell, B., Ramanarayanan, V., Nagarajan, S., & Houde, J. (2019). The FACTS model of speech motor control: Fusing state estimation and task-based control. *PLoS Computational Biology*, *15*(9), 1–26.

Port, R. F., & van Gelder, T. (Eds.). (1998). *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.

Roon, K. D., & Gafos, A. I. (2016). Perceiving while producing: Modeling the dynamics of phonological planning. *Journal of Memory and Language*, *89*(2), 222–243.

Saltzman, E., & Kelso, J. S. (1987). Skilled actions: A task-dynamic approach. *Psychological Review*, *94*(1), 84–106.

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

Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*, *324*, 81–85.

Schöner, G. (1990). A dynamic theory of coordination of discrete movements. *Biological Cybernetics*, *63*(4), 257–210.

Schöner, G. (2020). The dynamics of neural populations capture the laws of the mind. *Topics in Cognitive Science*, *12*, 1257–1271.

Schöner, G., Spencer, J. P., & The DFT Research Group (2016),. *Dynamic thinking: A primer on dynamic field theory*. Oxford, England: Oxford University Press.

Schrödinger, E. (1944). *What is life? The physical aspect of the living cell*. Cambridge, England: Cambridge University Press.

Senkowski, D., & Engel, A. K. (2024). Multi-timescale neural dynamics for multisensory integration. *Nature Reviews Neuroscience*, *25*, 625–642.

Šimko, J., & Cummins, F. (2010). Embodied task dynamics. *Psychological Review*, *117*(4), 1229–12246.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–74.

Sorensen, T., & Gafos, A. I. (2016). The gesture as an autonomous nonlinear dynamical system. *Ecological Psychology*, *28*(4), 188–215.

Sorensen, T., & Gafos, A. I. (2023). The relation between gestures and kinematics. In F. Breit, B. Botma, M. van't Veer, & M. van Oostendorp (Eds.), *Primitives of phonological structure* (pp. 251–279). Oxford, England: Oxford University Press

Spivey, M. (2007). *The continuity of mind*. Oxford, England: Oxford University Press.

Stern, M. C., & Shaw, J. A. (2023). Neural inhibition during speech planning contributes to contrastive hyperarticulation. *Journal of Memory and Language*, *132*, 104443.

Stern, M. C., & Shaw, J. A. (2024). Towards a minimal dynamics for gestures: A law relating velocity and position. In *Proceedings of the ISSP 2024–13th International Seminar on Speech Production* (pp. 262–265). Baixas, France: International Speech Communication Association (ISCA).

Strogatz, S. H. (2015). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering* (2nd ed.). Philadelphia, PA: Westview Press.

Suprunenko, Y. F., Clemson, P. T., & Stefanovska, A. (2013). Chronotaxic systems: A new class of self-sustained nonautonomous oscillators. *Physical Review Letters*, *111*(2), 024101.

Tilsen, S. (2009). Multitimescale dynamical interactions between speech rhythm and gesture. *Cognitive Science*, *33*(5), 839–879.

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

Tilsen, S. (2018). *Three mechanisms for modeling articulation: Selection, coordination, and intention* (Cornell Working Papers in Phonetics and Phonology). Ithaca, NY: Cornell University.

Tilsen, S. (2019a). Motoric mechanisms for the emergence of non-local phonological patterns. *Frontiers in Psychology*, *10*, 2143.

Tilsen, S. (2019b). *Syntax with oscillators and energy levels*. Berlin, Germany: Language Science Press.

Tilsen, S. (2020). *A different view of gestural activation: Learning gestural parameters and activations with an RNN* (Cornell Working Papers in Phonetics and Phonology). Ithaca, NY: Cornell University.

Tilsen, S. (2022). An informal logic of feedback-based temporal control. *Frontiers in Human Neuroscience*, *16*, 851991.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Turk, A., & Shattuck-Hufnagel, S. (2020). *Speech timing: Implications for theories of phonology, phonetics, and speech motor control*. Oxford, England: Oxford University Press.

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, *21*(1), 615–665.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., & SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*, 261–272.

Westbury, J. R. (1994). *X-ray microbeam speech production database user's handbook*. Madison, WI: Waisman Center.