Discovering dynamical models of speech using physics-informed machine learning

Sam Kirkham

Lancaster University, UK
s.kirkham@lancaster.ac.uk

Abstract

Spoken language is characterised by a high-dimensional and highly variable set of physical movements that unfold over time. What are the fundamental dynamical principles that underlie this signal? In this study, we demonstrate the use of physics-informed machine learning (sparse symbolic regression) for discovering new dynamical models of speech articulation. We first demonstrate the model discovery procedure on simulated data and show that the algorithm is able to discover the original model with near-perfect accuracy, even when the data contain extensive variation in duration, initial conditions and target positions, as well as in the presence of added noise. We then demonstrate a proof-of-concept applying the same technique to empirical data, which reveals a small set of candidate dynamical models with increasing levels of complexity and accuracy.

Keywords: speech production, sparse symbolic regression, articulatory phonology, task dynamics, articulatory data

1. Introduction

A fundamental aim in the study of language is the discovery of abstract invariants that underlie the variability observed in performance. For example, speech production involves a set of low-dimensional combinatorial units that are physically realised as a set of variable and high-dimensional motions. How do we best model the relationship? One solution is proposed by Articulatory Phonology/ Task Dynamics (AP/TD), in which phonetics and phonology are isomorphic, with the fundamental unit being the speech gesture: an abstract goal-driven force directing the vocal tract to a target state (Browman and Goldstein 1992; Tilsen 2016; Iskarous 2017).

Saltzman and Munhall (1989) propose a model of the gesture (hereafter abbreviated as SM89) as a critically damped harmonic oscillator (1), where k is a stiffness coefficient, m is a mass coefficient, and the damping coefficient $b = 2\sqrt{m}k$.

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

The SM89 model has long been the core gestural equation underpinning AP/TD, but it fails to capture the quasi-symmetrical velocity profiles and time-to-peak velocities typical of empirical data. Byrd and Saltzman (2003) show this can be solved via ramping functions, making gestural activation time-dependent. Sorensen and Gafos (2016) argue that this is an undesirable solution and that empirically realistic trajectories can be achieved by instead allowing the restoring force to be non-linear via a cubic term dx^3 in (2). This also eliminates the need for time dependence once the gesture is initiated.

$$m\ddot{x} + b\dot{x} + kx - dx^3 = 0 \tag{2}$$

This model reproduces many characteristics of empirical velocity profiles, but there may still be some room for improvement. For instance, Elie, Lee, and Turk (2023) advance a general Tau model that outperforms the SG16 model in fitting empirical data. Beyond conventional models of the gesture, there is also considerable scope for further developing task dynamic models of other domains, such as prosodic time-series (Iskarous, Cole, and Steffman 2024), disordered speech (Parrell et al. 2023), and signed languages. In many cases, we might have a lot of data, but lack sufficient predictions of the underlying dynamics to propose a model, or we may seek alternative models that better fit empirical data. This raises a question: how can we efficiently develop new dynamical models of speech?

We solve the problem of model discovery by leveraging recent developments in dynamical systems and machine learning that allow us to learn symbolic equations directly from data (Schmidt and Lipson 2009; Brunton, Proctor, and Kutz 2016). In such cases, we want to find a small number of model terms that expose the underlying dynamics, as opposed to a neural network that may have a very large number of parameters. Underpinning this is symbolic regression, whereby a function f can be approximated from $X, \dot{X}$ – which represent time-varying states $x(t), \dot{x}(t)$ – as a combination of non-linear functions:

$$\dot{X} = \Theta(X)\Xi \tag{3}$$

where $\Theta(X)$ is a library of non-linear functions

$$\Theta(X) = [1 X X^2 X^3 \dots sin X cos X] \tag{4}$$

and Ξ is a vector of coefficients corresponding to the functions in $\Theta(X)$.

$$\Xi = [\xi_1 \xi_2 \xi_3 \dots \xi n] \tag{5}$$

Without any constraints, the above model is likely to produce many non-zero coefficients in Ξ that do not contribute much to the underlying system, adding model complexity and increasing the risk of overfitting. In order to promote sparsity in Ξ, *sparse* symbolic regression optimises for a sparse vector of coefficients for each function in $\Theta(X)$. An example optimisation is Sequential Thresholded Least-Squares, which solves a least squares solution for Ξ, thresholds any coefficients below a value λ, and repeats this process until an optimally sparse model is determined (Brunton, Proctor, and Kutz 2016).

The sparse symbolic regression method outlined above falls into a general class of SINDy (Sparse Identification of Non-linear Dynamics) models. SINDy models can accurately discover the governing equations of known systems, such as chaotic Lorenz and fluid dynamic equations, as well as discover new models in applications such as astrophysics (Pasquato et al. 2022). For more details see Brunton, Proctor, and Kutz (2016) and Champion et al. (2020).

2. Methods

The first step in model discovery is obtaining one or more time-series that represent the output of the system under study. In our case, this is the position and velocity of the vocal tract articulators. We aim to model a single speech gesture, so each trajectory represents a single gesture, defined as the interval between a pair of successive zero crossings in the velocity signal.

The next step is to select a library of candidate functions. From AP/TD research reviewed above, we know that articulatory signals are often well-approximated by polynomial functions, such that a function $f(x)$ can be approximated as a sum of polynomials of increasing order, as in (6), where a_n is the coefficient of each term (note that a_0 is a constant). In this instance, we do not allow interactions between terms, such as $x\dot{x}^2$, but allowing this would be a trivial addition.

$$f(x) = a_0 + a_1 x + a_2 \dot{x} + a_3 x^2 + a_4 \dot{x}^2 + a_5 x^3 + a_6 \dot{x}^3 + \dots \quad (6)$$

A key aspect of SINDy is that we can incorporate physical constraints on the discovered model, such that a discovered coefficient must have a specific value, or two coefficients must be in a particular ratio. To illustrate, take the equation $\ddot{x} = -b\dot{x} - kx$. In order to discover or numerically solve a second-order differential equation, we split it into a series of first-order equations with the introduction of a new variable y, such that $y = \dot{x}$ and $\dot{y} = -by - kx$. If SINDy finds $y = 1.00\dot{x}$ then we can just substitute this value easily into the second equation. If it finds a more complex equation, however, such as $y = 43.62 - 1.55x + 0.90\dot{x}$, then it would yield a final model of $\ddot{x} = -b(43.62 - 1.55x + 0.90\dot{x}) - kx$.

To avoid this level of complexity, we place a physical constraint on y such that $y \stackrel{!}{=} 1.00\dot{x}$. We later show that relaxing this constraint results in models that better fit the data, but also add significant complexity. We implement constraints using the SR3 (sparse relaxed regularized regression) algorithm (Champion et al. 2020), which aims to minimise (7), where $R(W)$ is a regularisation function that acts as a prior on sparsity promotion and λ weights this constraint. Note that $\lambda = \eta^2/2\nu$, where ν determines the closeness of the match between Ξ and W.

$$\min_{\Xi, W} \frac{1}{2}||\dot{X} - \Theta(X)\Xi||^2 + \lambda R(W) + \frac{1}{2\nu}||\Xi - W||^2 \quad (7)$$

We use weighted ℓ_0 regularisation, with a coefficient threshold of $\eta = 0.1$ and $\nu = 1$. A model is discovered for each trajectory and we perform model ensembling over these individual models to arrive at a final model. We evaluate the accuracy of the model by generating a prediction from the discovered model for each token. We then score the accuracy of the predicted trajectory using R^2 and RMSE metrics.

3. Discovering models from simulated data

3.1. Generating simulated data

In order to test the ability of SINDy to discover models from data, we generated a simulated data set with a number of parameters varied across a set of trajectories. Specifically, we simulated data across combinations of duration = {0.05, 0.10, 0.15, 0.20} seconds, initial position = {0.0, 0.1, ..., 1.0}, target = {0.0, 0.1, ..., 1.0} and noise = {normal, noise}. In all simulations, k = 2000 and $b = 2\sqrt{k}$. The noise condition corresponds to the addition of random Gaussian noise between [0, 1], scaled by a factor of 0.01, to each position and velocity sample from the simulated solution. We removed cases from the above parameter combinations where the target was equal to the initial position, as the trajectory does not move from its initial condition in these instances. These parameters were used as inputs to the SM89 second-order differential equation $\ddot{x} + b\dot{x} + kx = 0$ which was solved numerically using the `scipy.integrate.solve_ivp` function in Python. This resulted in 880 unique simulated trajectories.

3.2. Results

We perform SINDy discovery on the SM89 model using a simple candidate library containing the terms x and $\dot{x}$, which means that the maximal equation is:

$$\ddot{x} = a_0 + a_1 x + a_2 \dot{x} \quad (8)$$

The SINDy models finds equation (9) for all trajectories. Note that SINDy reports the target as kC, but we can substitute $kx - kC$ with $k(x - C)$. As such, we correctly identify the original equation that simulated the data, even in the presence of the variable durations, targets, initial conditions and noise.

$$\ddot{x} = -b\dot{x} - k(x - C) \quad (9)$$

In the no noise condition, parameter estimation is near 100% accuracy, with the difference between real/estimated coefficients at C = 0.01% ($\sigma = 0.01$), k = 0.08% ($\sigma = 0.02$), b = 0.03% ($\sigma = 0.01$). Reconstruction of the simulated trajectories is also highly accurate, with mean $R^2 = 1.00$ ($\sigma = 0.01$) and mean RMSE = 0 ($\sigma = 0.02$). The addition of noise affects parameter estimation to a minor extent, with mean $R^2 = 0.99$ ($\sigma = 0.12$) and mean RMSE = 0.03 ($\sigma = 0.02$). The difference between real and estimated coefficients in the noisy condition is C = 0.57% ($\sigma = 1.20$), k = 3.00% ($\sigma = 3.78$), b = 3.96% ($\sigma = 4.37$). The worst performing noisy trajectory had $R^2 = 0.84$

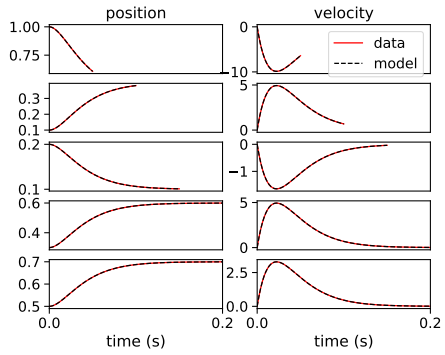

Figure 1: *Simulated trajectories and SINDy predictions for noise-free data. The y-axis varies across each plot to fit the data's range.*

Figure 1 shows 5 randomly sampled trajectories comparing simulated data and discovered model predictions. The model estimates the underlying trajectories with a very high degree of accuracy, even when the data are truncated as in the top two panels. We are unable to show a plot of the noisy data due to space constraints, but reconstruction of the underlying trajectory is also near-perfect in this condition, even in the presence of considerable random noise.

4. Discovering models from empirical data

We now move on to a proof-of-concept example, showing how we can discover parsimonious models from empirical data.

4.1. Data

We use data from the X-Ray Microbeam corpus (Westbury 1994). As a case study, we only analyse data from a single speaker (JW11), as this allows us to explore the initial interpretation of model coefficients, without having to take into account the significant added complexity introduced by between-speaker variation. Specifically, we use a task in which speakers produce a string of repetitions of the syllable /pə pə pə .../. This allows us to examine repetitions of the same gesture, which acts as a valuable test of how sensitive the model discovery procedure is to small variations within one speaker. We see this evaluation as a necessary step prior to applying the method to data with a much greater range of variation. We calculated lip aperture as the Euclidean distance between upper and lower lip sensors, and approximated velocity as the first-derivative of the position values. Gestures were segmented into separate closure and release gestures based on zero-crossings in the velocity signal. In total, we obtained 29 individual gestural trajectories from repetitions of /p/ for this speaker.

4.2. First-order models

We begin by fitting a simple model to the data: a first-order differential equation for $\dot{x}$. Note that here we are only solving for the velocity of the gesture, unlike the SM89 model which solves for acceleration $\ddot{x}$. We predict that a first-order model may be a worse fit for the data than a second-order model, but we begin with a simpler model to assess its baseline accuracy.

Table 4.2 shows a first-order model fitted with different feature libraries of polynomial degrees between one and four. Note that prediction accuracies are for the gesture's position variable only, because SINDy integrates over the velocity to return position. We comment on the model's accuracy in estimating velocity later in this section. A first-degree model performs very poorly with mean R^2 = 0.02, second/third-degree models have mean R^2 = 0.92, and the fourth-degree model has mean R^2 = 0.89. It is clear that the addition of cubic terms has only a negligible effect and the quartic term actively degrades performance, so we now explore this first-order second-degree model further.

degree	R^2 mean	$R^2\sigma$	R^2 min	R^2 max
1	0.02	0.02	0.00	0.07
2	0.92	0.01	0.89	0.94
3	0.92	0.01	0.90	0.94
4	0.89	0.17	0.01	0.94

Table 1: R^2 statistics for first-order models with different polynomial degrees fitted to lip aperture data.

The first-order second-degree model returns a simple quadratic equation:

$$\dot{x} = a - bx + cx^2 \tag{10}$$

There is a linear relationship between a, b, c, such that in these data $a \approx -14b \approx 830c$. As this is a quadratic equation, the quartic term cx^2 determines the width of the velocity peak, the linear term bx controls symmetry around the y-axis, and the constant a determines the y-intercept.

Figure 2 shows randomly sampled lip aperture trajectories and SINDy predictions. We can see very good reconstruction of the position data, but the velocity profiles are less accurate: while the qualitative shape is maintained, the onset/offset are displaced from zero and there are some noticeable mismatches. In summary, a first-order model provides a simple qualitative model that approximates the system, but clearly underperforms in predicting change in velocity. As a result, we anticipate that a second-order model should improve performance.

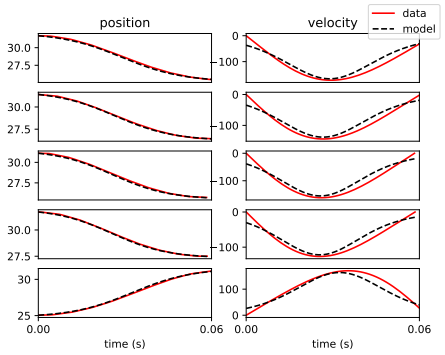

Figure 2: *Lip aperture trajectories and SINDy first-order model predictions, with polynomial terms up to quadratic. The y-axis varies across each plot to fit the data's range.*

4.3. Second-order models

We now fit a second-order model to the data, solving for the system's acceleration $\ddot{x}$. This should allow us to better capture changes in velocity. Note that we impose a physical constraint on the velocity as detailed in Section 2, which simply aims to reduce model complexity and aid interpretability. Table 4.3 shows a second-order model fitted with different feature libraries of polynomial degrees between one and four. The first- and second-degree models have mean R^2 = 0.96, which is slightly better than the higher polynomials. This suggests that a first-degree model can perform well, so we explore this further.

degree	R^2 mean	$R^2\sigma$	R^2 min	R^2 max
1	0.96	0.00	0.95	0.96
2	0.96	0.00	0.95	0.96
3	0.95	0.01	0.92	0.96
4	0.94	0.02	0.90	0.96

Table 2: R^2 statistics for second-order models with different polynomial degrees fitted to lip aperture data.

The second-order first-degree model returns (11), which is equivalent to the Saltzman and Munhall (1989) model.

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

Figure 3 shows the same 5 lip aperture trajectories as in Figure 2, with SINDy predictions from the second-order model. The discovered model fits better than the first-order model, but with some inaccuraries towards the end of the velocity trajectory. We do find, however, that this model is able to generate more symmetrical velocities than the SM89 model by relaxing the critical damping constraint. This introduces a different constraint: the model parameters must exist in a non-linear relation-

ship between b, k and duration in a way that avoids oscillation (Shaw and Chen 2019).

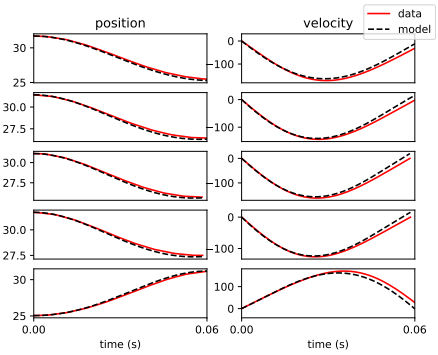

Figure 3: *Lip aperture trajectories and SINDy second-order model predictions. Model includes first-degree polynomials and physical constraints. The y-axis varies across each plot to fit the range of the data.*

If we relax the physical constraint $y \stackrel{!}{=} 1.00\dot{x}$ in $\ddot{x} = -by - kx$ then SINDy discovers the more complex model in (12):

$$\ddot{x} = -b(a - cx + d\dot{x}) - kx \tag{12}$$

Figure 4 shows example model predictions, with much improved fit between data and model. This comes at the cost, however, of adding significant complexity into the model.

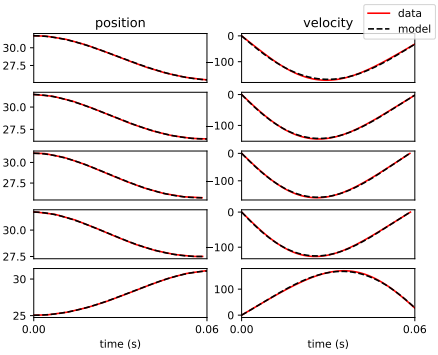

Figure 4: *Lip aperture trajectories and SINDy second-order model predictions. Model includes first-degree polynomials but no physical constraints. The y-axis varies across each plot to fit the data's range.*

5. Discussion and conclusion

This paper demonstrates how sparse symbolic regression can be used to identify dynamical principles of articulatory dynamics. The discovered models show a trade-off between simplicity and accuracy, from a simple first-order model that fits less accurately to a second-order model with no physical constraints that fits near-perfectly but is quite complex. In some cases, however, capturing the system's attractor dynamics may be more important than predicting trajectories, so the simpler models should not be immediately discounted. In future research, we will explore the discovered models via simulation to probe the dynamical principles they expose around the underlying system. In addition to this, we aim to test how well the discovered models generalise to different data sets. We note that the models should be treated with caution at this stage, as they are based on 29 trajectories of the same gesture from a single speaker, so these data may not be a good representation of all gesture types or speakers. This minimal proof-of-concept was driven by interpretability, but it clearly motivates extending this approach to a larger data set, which is the focus of ongoing research.

6. Acknowledgements

This research was supported by Arts and Humanities Research Council grant AH/Y002822/1.

7. References

Browman, Catherine P. and Louis Goldstein (1992). "Articulatory phonology: an overview". In: *Phonetica* 49.3-4, pp. 155–180.

Brunton, Steven L., Joshua L. Proctor, and J. Nathan Kutz (2016). "Discovering governing equations from data by sparse identification of nonlinear dynamical systems". In: *Proceedings of the National Academy of Sciences* 113.15, pp. 3932–3937.

Byrd, Dani and Elliot Saltzman (2003). "The elastic phrase: modeling the dynamics of boundary-adjacent lengthening". In: *Journal of Phonetics* 31.2, pp. 149–180.

Champion, Kathleen, Peng Zheng, Aleksandr Y. Aravkin, Steven L. Brunton, and J. Nathan Kutz (2020). "A unified sparse optimization framework to learn parsimonious physics-informed models from data". In: *IEEE Access* 8, pp. 169259–169271.

Elie, Benjamin, David N. Lee, and Alice Turk (2023). "Modeling trajectories of human speech articulators using general Tau theory". In: *Speech Communication* 151, pp. 24–38.

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

Iskarous, Khalil, Jennifer Cole, and Jeremy Steffman (2024). "A minimal dynamical model of Intonation: Tone contrast, alignment, and scaling of American English pitch accents as emergent properties". In: *Journal of Phonetics* 101309.1–27.

Parrell, Benjamin, Antje Mefferd, Sarah Harper, Simon Roessig, and Doris Mücke (2023). "Using computational models to characterize the role of motor noise in speech: The case of amyotrophic lateral sclerosis". In: *Proceedings of the 20th International Congress of Phonetic Sciences* 988, pp. 878–882.

Pasquato, Mario, Mohamad Abbas, Alessandro A. Trani, Matteo Nori, Kwiecinski, Piero Trevisan, Vittorio F. Braga, Giuseppe Bono, and Andrea V. Macciò (2022). "Sparse identification of variable star dynamics". In: *The Astrophysical Journal* 930.161, pp. 1–13.

Saltzman, Elliot and Kevin G. Munhall (1989). "A dynamical approach to gestural patterning in speech production". In: *Ecological Psychology* 1.4, pp. 333–382.

Schmidt, Michael and Hod Lipson (2009). "Distilling free-form natural laws from experimental data". In: *Science* 324, pp. 81–85.

Shaw, Jason A. and Wei-Rong Chen (2019). "Spatially conditioned speech timing: Evidence and implications". In: *Frontiers in Psychology* 10.2726, pp. 1–17.

Sorensen, Tanner and Adamantios I. Gafos (2016). "The gesture as an autonomous nonlinear dynamical system". In: *Ecological Psychology* 28.4, pp. 188–215.

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

Westbury, John R. (1994). *X-Ray Microbeam Speech Production Database User's Handbook*. Madison, WI: Waisman Center.