13th International Seminar
of Speech Production

Proceedings

13 – 17 may 2024 Autrans FR

https://issp24.sciencesconf.org

Towards a minimal dynamics for gestures: a law relating velocity and position

Michael C. Stern[1], Jason A. Shaw[1]

[1]*Department of Linguistics, Yale University, New Haven, CT, USA*

michael.stern@yale.edu, jason.shaw@yale.edu

Abstract

Dynamical models of articulatory gestures relate the velocity of a vocal tract variable to its position via a function with one or more control parameters. In this paper we propose a minimal dynamical model of gestures. The model is empirically motivated by observations of the timecourse of the ratio of velocity to position in bilabial constriction movements by English and Mandarin speakers. We discovered that this ratio tends to follow an exponential growth curve over the course of a movement. A dynamical formalization of this empirical discovery, in combination with an assumption of point attractor dynamics, constitutes the core of our model. The model has only two parameters, T *and* r. T *corresponds to the target position of the vocal tract variable and* r *corresponds to rapidity. Simulations from the model capture key elements of gesture kinematics, performing much better than the damped mass-spring model. Our model achieves these improvements despite having fewer control parameters. Future work will extend our model to other kinds of gestures besides bilabial consonant constrictions.*

Keywords: *articulatory gesture, articulatory kinematics, dynamical system, damped mass-spring*

1. Introduction

In controlled human movement—including speech articulatory movement—peak velocity is robustly correlated with maximum spatial displacement (Ostry & Munhall, 1985). The farther an effector travels to reach its target, the faster it moves. In order to capture this empirical fact, dynamical models of articulatory movement, e.g., Task Dynamics (Saltzman & Munhall, 1989), encode a negative relationship between velocity and distance to the target, of the form in (1).

$$\dot{x} = -\lambda(x - T) \tag{1}$$

x is the state of a vocal tract variable (TV) like lip aperture (LA: the distance between the lips), T is the target state of the TV (e.g., zero or possibly negative for /b/ or /m/ [Parrell, 2011]), and λ is a control parameter modulating the relationship between velocity $\dot{x}$ and distance to the target $(x - T)$. We follow Mücke et al. (2024) in using T instead of x_0 to refer to the target position, since x_0 often refers to the initial state of x. (1) succeeds in capturing the linear correlation between peak velocity and maximum displacement. However, it fails to capture another robust fact about TV trajectories. In particular, for any fixed value of the control parameter λ, model-simulated TV trajectories achieve peak velocity instantaneously; velocity then decreases monotonically as the TV approaches its target. In real TV trajectories, peak velocity occurs later, approximately halfway through the movement (Ostry et al., 1987). In the *damped mass-spring* model of Task Dynamics, as in (2), peak velocity is delayed because velocity $\dot{x}$ is negatively related to acceleration $\ddot{x}$.

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

The timing of the velocity peak predicted by (2) is an improvement over (1). This improvement is achieved via greater model complexity: (2) is a *second order* system, referencing acceleration in addition to velocity, with four control parameters m, b, k, and T, more than the two parameters λ and T in (1). Even in (2), however, peak velocity occurs unrealistically early (Perrier et al., 1988). Thus, additional complexity has been proposed: e.g., a time-varying *activation* parameter (Byrd & Saltzman, 1998; Kröger et al., 1995), or a negative relationship between velocity and the *cube* of distance to the target (Sorensen & Gafos, 2016).

In this paper, we take a strongly empirical approach to understanding the relation between velocity and position. Rather than commit to the specific second order system in (2), we start from the minimal assumption that velocity is negatively related to distance to the target, formalized in (1). This allows us to solve for the parameter λ from measurement of data, in particular, electromagnetic articulatography (EMA) recordings of bilabial constriction movements. In this way, we address the question: what is the *empirical* relationship between velocity and position over time? The answer to this question guides further dynamical model development, which we pursue below.

2. Methods

2.1. Participants

Data was collected from 24 subjects: 12 native speakers of American English (8 female, 4 male, ages 19–28, mean = 20.75) and 12 native speakers of Mandarin Chinese (7 female, 4 male, 1 nonbinary, ages 19–33, mean = 24.00). All participants self-reported no history of speech, language, or hearing impairment.

2.2. Stimuli

Stimuli consisted of eight word-initial CV sequences in each language, where the initial consonant was bilabial—either [b] or [m]—and the vowel was either low back [ɑ] or high front [i]. Target sequences containing the vowel [i] were immediately preceded by the vowel [ɑ], and sequences containing the vowel [ɑ] were immediately preceded by the vowel [i], in order to ensure maximal vowel movement. All Mandarin target syllables bore a falling tone (T4) and were preceded immediately by a low tone (T3). Each target syllable was produced in two carrier sentences, occurring once in an informationally prominent position and once in a less prominent position. To encourage natural speech, each carrier sentence was preceded by a question, which served to provide context for the target sentences.

2.3. Procedure

Presentation of materials was controlled using E-Prime. On each trial, an audio recording of a question was played. The question was also displayed in text on the screen for 5000 ms.

Participants were instructed to listen to the question and to read aloud the answer that followed. In total, each participant produced 128 tokens (8 items × 2 carrier sentences × 8 repetitions) across four blocks of 32 items each. Within each block, stimuli were presented in a randomized order.

Articulatory kinematic data was collected with the NDI Wave Speech Research System sampling at a rate of 100 Hz. The sensors of interest for this study were attached at the vermillion border of the upper lip (UL) and lower lip (LL). Three sensors were also attached to the tongue: tongue tip (TT), tongue blade (TB), and tongue dorsum (TD), placed ~1 cm, ~3 cm, and ~5 cm from the tip of the tongue, respectively. In order to track movements of the jaw, one lower incisor (LI) sensor was attached to the hard tissue of the gum directly below the left incisor. Reference sensors were attached on the left and right mastoids and on the nasion. Measurements of the occlusal plane and a midsagittal palate trace were also collected. Acoustic data was collected using a Sennheiser shotgun microphone at a sampling rate of 22,050 Hz.

2.4. Data processing

Articulatory data was rotated to the occlusal plane and corrected for head movement computationally. Trajectories were smoothed using the robust smoothing algorithm of Garcia (2010). First and second time derivatives (velocity and acceleration) were calculated from the smoothed trajectory using central differencing, then lowpass filtered using a 5th order Butterworth filter. Consonant constriction gestures were parsed from the lip aperture (LA) signal, calculated as the Euclidean distance between the UL and LL sensors. The onset and offset of each movement were marked as the timepoints at which velocity exceeded or fell below, respectively, a 20% threshold of peak velocity, manually selected in MVIEW (Tiede, 2005). The spatial target of each gesture (i.e., T) was defined as the LA value at the timepoint of minimum velocity following gesture offset.

λ was calculated at each sample as the negative ratio of instantaneous velocity to instantaneous distance to the target: $-\dot{x}/(x-T)$ (see [1]). By demarcating gestures based on a 20% threshold of peak velocity, instead of, e.g., velocity zero-crossing, we exclude portions of the kinematics in which velocity or distance to the target are infinitesimal. This prevents λ from approaching 0 (infinitesimal velocity) or infinity (infinitesimal distance to the target). Gesture duration was calculated by subtracting the timestamp of the onset of movement from the timestamp of the offset of movement. We also calculated a measure of kinematic stiffness for each gesture by dividing peak velocity by maximum spatial displacement, i.e., onset position minus target position (Roon et al., 2021).

Out of the 3,072 tokens elicited, a total of 962 tokens (31.3%) were eliminated from analysis for the following reasons: failure of the gesture parsing tool to extract the gesture (447 tokens); a non-monotonic trajectory, i.e., instantaneous velocity changed sign for at least one sample (306 tokens); failure of the participant to produce contrastive focus on the informationally prominent syllable, as judged by the experimenters (155 tokens); disfluency (5 tokens); or data storage failure (49 tokens).

3. Results

3.1. Kinematic variables

Figure 1 displays the distributions of the kinematic variables gesture duration, peak velocity, maximum displacement, and kinematic stiffness across all 2,110 tokens from all 24 speakers.

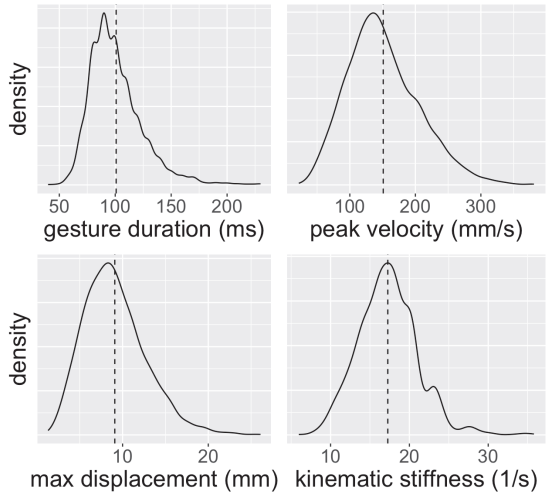

Figure 1: *Density plots of kinematic variables across all tokens (*n *= 2,110). Dashed vertical lines indicate the mean.*

3.2. λ trajectories

Next, we examine the trajectories of λ, i.e., the ratio of instantaneous velocity to instantaneous distance to the target. As seen in **Figure 2**, regardless of language and vowel context, λ generally followed an exponential growth curve from movement onset to offset.

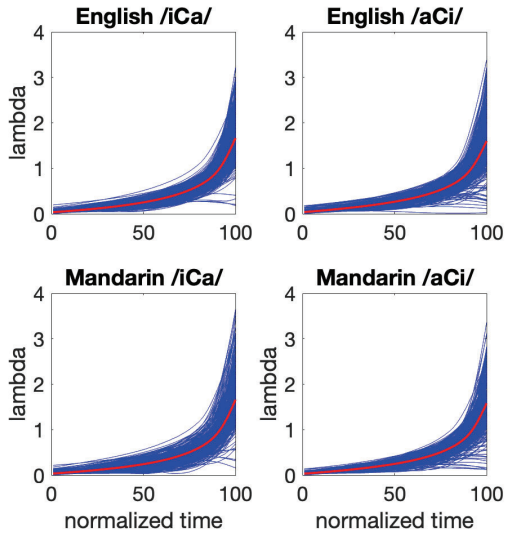

Figure 2: λ *trajectories by language and vowel context. Blue lines show individual trajectories; red lines show average trajectories. Trajectories were normalized to a 100-unit timescale using shape-preserving cubic Hermite interpolation.*

From this observation, it follows that the first time derivative of $\ln(\lambda)$ approximates a constant for each movement, which we call r. To evaluate the robustness of this generalization, a linear regression model was fit to each trajectory of $\ln(\lambda)$ over time. The fits were excellent: overall mean $R^2 = .97$. Moreover, as seen in **Figure 3**, r, the slope of each linear fit, correlates strongly with linguistically relevant measures like duration

(Spearman's $\rho = -.83$, $p < .001$) and kinematic stiffness ($\rho = .82$, $p < .001$).

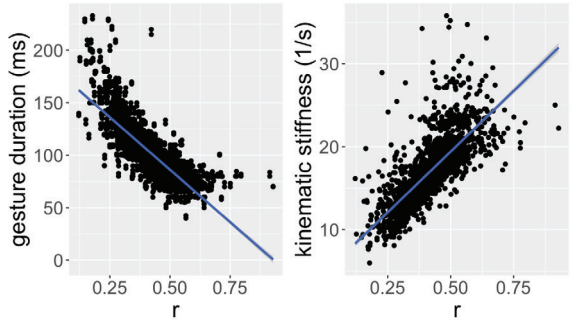

Figure 3: *Correlations between* r *(the slope of a regression line fit to* ln(λ) *and two kinematic variables: gesture duration (left) and kinematic stiffness (right).*

4. Dynamical model

The empirical observation of exponential growth in λ over time can be expressed in the differential equation in (3).

$$\dot{\lambda} = r\lambda \tag{3}$$

Together, the two first order equations in (1) and (3) express a dynamical system of two variables, x and λ. Since λ is defined in (1) as $-\dot{x}/(x - T)$, we can substitute this definition into (3) to derive a single second order equation, eliminating λ. This equation, solved for velocity $\dot{x}$, is shown in (4).

$$\dot{x} = (\ddot{x}/\dot{x} - r)(x - T) \tag{4}$$

(4) has only two parameters, r and T, which can both be inferred from data and have clear interpretations. T corresponds to the spatial target, and r corresponds to movement rapidity, similar to stiffness k in the damped mass-spring model. Moreover, the system is autonomous as it does not reference an extrinsic time variable (Fowler, 1980; Sorensen & Gafos, 2016).

In order to examine the empirical adequacy of (4), we simulated movement trajectories from (4) and compared them to observed trajectories and trajectories simulated from the damped mass-spring model (2). As seen in **Figure 4**, movement trajectories simulated from (4) correspond well with observed trajectories. For instance, peak velocity (corresponding to the zero-crossing in the acceleration curve) occurs 67% of the way through the simulated trajectory, compared to 71% on average (*SD* = 12%) in observed trajectories. For comparison, in the trajectory simulated from the damped mass-spring model, peak velocity occurs 19% of the way through the movement. In both the observed trajectories and the trajectories simulated from our model, the skew in the velocity curve is related to an asymmetry in the acceleration curve: the positive acceleration peak has a smaller magnitude than the negative acceleration peak. In particular, the ratio of the positive peak to the negative peak is 0.51 in the trajectory simulated from our model, compared to 0.83 on average (*SD* = 0.33) in the observed trajectories. In the trajectory simulated from the damped mass-spring model, on the other hand, the positive acceleration peak has a much *greater* magnitude than the negative peak (6.51 times greater).

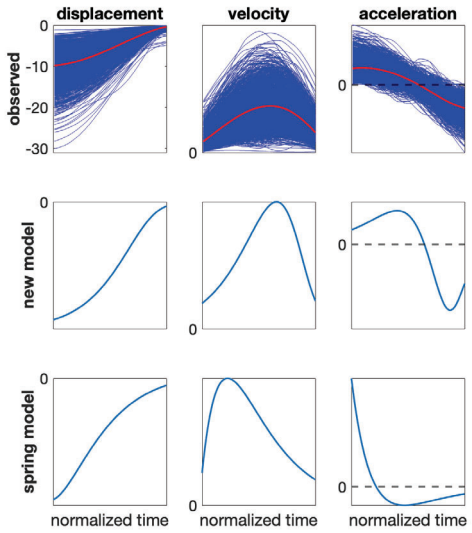

Figure 4: *Displacement (left), velocity (center), and acceleration (right) in real gestures (top), simulated by the proposed model (middle) and simulated by the damped mass-spring model (bottom). All trajectories are demarcated based on a 20% threshold of peak velocity. For both model simulations,* T *= 0 and initial* x *= 10. For the new model simulation,* r *= 10. For the damped mass-spring model simulation,* m *= 1,* b *= 10, and* k *= 25. Trajectories are reversed (multiplied by –1) in order to ease interpretation of velocity and acceleration, and vertical axes are scaled in order to focus on trajectory shapes rather than absolute magnitudes. Dashed horizontal lines indicate acceleration = 0.*

5. Discussion and conclusion

We started from the minimal assumption that articulatory gestures are defined by point attractor dynamics, i.e., a negative relationship between velocity and distance to the target. We formalized this assumption in the differential equation in (1). (1) defines the parameter λ as the negative ratio of velocity to distance to the target, a value which can be measured in articulatory kinematic data. Our investigation of λ trajectories in bilabial constriction movements from 12 English speakers and 12 Mandarin speakers revealed a robust pattern: λ generally follows an exponential growth curve over the course of a movement (**Figure 2**). We incorporated this empirical discovery into the minimal dynamics in (1), deriving (4). Our proposed dynamical system in (4) is both simpler (less parameters) and more empirically adequate than the damped mass-spring model (2). Future work will compare (4) to expanded versions of the damped mass-spring model, i.e., with time-ramped activation (Byrd & Saltzman, 1998; Kröger et al., 1995) or a cubic term (Sorensen & Gafos, 2016). While our model is simpler than those models, a direct comparison of empirical adequacy would be useful in light of the general tradeoff between model simplicity and data fitting.

It is interesting to note that, although (1) is a first order equation—only referencing the first time derivative $\dot{x}$ — formalizing the observed temporal variation in λ led to the second order equation in (4). It is not surprising that a second order description is necessary, given that the empirical shapes of velocity curves have proven difficult to capture with first

order dynamics, as described in the Introduction. Although both our model and the damped mass-spring model include an acceleration term, our model captures the shapes of acceleration curves much more closely than the damped mass-spring model, which predicts instantaneous achievement of peak acceleration (**Figure 4**). Our model likely generates more complex acceleration curves because the acceleration term is weighted by velocity, which is itself time-varying. In the damped mass-spring model, on the other hand, the acceleration term is weighted by the constant parameter m.

We have only begun to probe the empirical predictions of our model. For instance, r correlates with peak velocity. In this way, r is similar to k in the damped mass-spring model. However, in our model, the *time to achieve* peak velocity (as a percentage of gesture duration) is stable under variation in r. In the damped mass-spring model, on the other hand, k correlates with both peak velocity and time to achieve peak velocity (e.g., Z. Liu et al., 2022; Mücke et al., 2024). Thus, the damped mass-spring model predicts a negative correlation between peak velocity and time to achieve peak velocity, while our model does not. It would also be valuable to investigate the absolute magnitudes of peak velocity and acceleration, rather than just the shapes of the curves. So far, dynamical modeling work (including this work) has focused on the timing of landmarks, especially peak velocity (e.g., Sorensen & Gafos, 2016). However, the magnitude of, e.g., peak velocity, offers another kinematic dimension to constrain model building, which we have not yet explored in depth.

In future work, we plan to fit the model parameters r and T to data using least squares regression (Iskarous, 2017), rather than estimating them using heuristics. Fitting the model parameters has the potential to shed light on broader theoretical issues, such as intergestural coordination. Preliminary analysis of bilabial release and vowel constriction movements suggests that linear fits to $\ln(\lambda)$ are slightly worse, i.e., mean $R^2 = .91$ and .89, respectively. This is noteworthy because previous work suggests that the timing of target achievement for these two movements (and not consonant constriction) is coordinated (Kramer et al., 2023). It is possible that the fit is worse for these two kinds of movements because their dynamics are coupled in a way that synchronizes target achievement. Thus, model fit may be improved by the addition of a coupling term. This would constitute evidence for target-based gestural coordination (Turk & Shattuck-Hufnagel, 2020), in contrast to onset-based coordination (Nam & Saltzman, 2003).

Model fit for vowel constriction movements and other kinds of (non-labial) consonant movements may also be improved by closer consideration of the nature of targets T. A primary motivation for examining bilabial consonants is that lip aperture is a hypothesized tract variable that corresponds very closely to measurable kinematics. Movements of other articulators like the tongue body are hypothesized to unfold over two tract dimensions: constriction location and constriction degree (e.g., Browman & Goldstein, 1989; Saltzman & Munhall, 1989). In our preliminary analysis of vowel movements, we assumed a single tract variable in 3D space. This allows the target to be straightforwardly estimated from data, but represents a departure from the theoretical proposal of Articulatory Phonology/Task Dynamics. In future work, we plan to develop a method to estimate separate constriction location and constriction degree targets from data. Then, we can examine whether separating movement dynamics into two systems improves the fit of the model. In this way, our model can offer insights into the nature of the tract variables (i.e., x) governing articulatory movement.

6. Acknowledgements

We would like to thank Cherilyn Wang and Ben Kramer for collecting the data and parsing gestural landmarks, and Yuyang Liu for assistance with data processing.

7. References

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*(2), 201–251.

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

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

Garcia, D. (2010). Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational Statistics and Data Analysis*, *54*(4), 1167–1178.

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.

Kramer, B. M., Stern, M. C., Wang, Y., Liu, Y., & Shaw, J. A. (2023). Synchrony and stability of articulatory landmarks in English and Mandarin CV sequences. *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS)*, 1022–1026.

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

Liu, Z., Xu, Y., & Hsieh, F. fan. (2022). Coarticulation as synchronised CV co-onset – Parallel evidence from articulation and acoustics. *Journal of Phonetics*, *90*.

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.

Nam, H., & Saltzman, E. (2003). A Competitive, Coupled Oscillator Model of Syllable Structure. *Proceedings of the 15th International Congress of Phonetic Sciences*, 2253–2256.

Ostry, D. J., Cooke, J. D., & Munhall, K. G. (1987). Velocity curves of human arm and speech movements. *Experimental Brain Research*, *68*(1), 37–46.

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

Parrell, B. (2011). Dynamical account of how /b, d, g/ differ from /p, t, k/ in Spanish: Evidence from labials. *Laboratory Phonology*, *2*(2), 423–449.

Perrier, P., Abry, C., & Keller, E. (1988). Vers une modélisation des mouvements du dos de la langue. *Vers Une Modélisation Des Mouvements Du Dos de La Langue*, *2–1*, 45–63.

Roon, K. D., Hoole, P., Zeroual, C., Du, S., & Gafos, A. I. (2021). Stiffness and articulatory overlap in Moroccan Arabic consonant clusters. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *12*(1), 8.

Saltzman, E. L., & Munhall, K. G. (1989). A Dynamical Approach to Gestural Patterning in Speech Production. *Ecological Psychology*, *1*(4), 333–382.

Sorensen, T., & Gafos, A. (2016). The Gesture as an Autonomous Nonlinear Dynamical System. *Ecological Psychology*, *28*(4), 188–215.

Tiede, M. (2005). *MVIEW: Software for visualization and analysis of concurrently recorded movement data* [Computer software]. Haskins Laboratories.

Turk, A., & Shattuck-Hufnagel, S. (2020). *Speech Timing: Implications for Theories of Phonology, Phonetics, and Speech Motor Control*. Oxford University Press.