Beyond Direct Geometry: Spring-Mass Control of Tongue Articulation for Vocal Synthesis

Debasish Ray Mohapatra*
University of British Columbia
Vancouver, British Columbia
Canada
debasishray@ece.ubc.ca

Ziyi Xia*
University of British Columbia
Vancouver, British Columbia
Canada
zxia0101@cs.ubc.ca

Sidney Fels
University of British Columbia
Vancouver, British Columbia
Canada
ssfels@ece.ubc.ca

Abstract

Human speech production relies on tightly coupled neuromuscular control of articulators and the aeroacoustic properties of the vocal tract. Vocal synthesizers employing direct geometric control of articulatory positions often struggle to generate smooth nonlinear trajectories between target vowels, as required for diphthong synthesis. We propose a biomechanically inspired control approach using a lightweight spring–mass–damper framework coupled to an acoustic wave solver, in which spring forces are parameterized to generate target tongue shapes. This physics-based interface enables synthesis through an input modality analogous to natural muscle activation. We conducted a pilot study comparing the proposed physics-based controller with a conventional geometry-driven controller on identical trajectory-generation tasks, subsequently coupling both to a vocal synthesizer. The pilot study served to refine the experimental design and verify that the system captures meaningful differences between the two controllers. Results revealed large, observable differences in the ability of each controller to generate nonlinear articulatory trajectories, both quantitatively and qualitatively. These findings support a planned controlled user study with a larger and more diverse participant pool, aimed at providing statistically valid assessments of the proposed controller's effectiveness for smooth trajectory generation.

Keywords

Articulatory Synthesis, Biomechanical Controller, Oral Cavity, Tongue

1 Introduction

The human voice—the most nuanced and expressive sound producing system— is capable of seamlessly transitioning between phonemes, pitches, and timbres through continuous articulatory movements [2, 8]. Unlike discrete-control instruments (e.g., keyboards and fretted strings), the voice operates within a high-dimensional parameter space—shaped by tongue position, lip aperture, and vocal fold tension—enabling fluid gestural control. Central to this is tongue positioning, which dynamically controls the oral cavity geometry and its acoustic resonances that define vocal timbre [9]. This articulatory-acoustic mapping motivates two complementary vocal instrument design strategies: geometric control [13] and physics-based control [16] of articulatory movements.

*These authors contributed equally to this work.

NIME '26, June 23–26, 2026, London, UK

In geometric control, speech is synthesized by altering articulatory positions or tract shape, without accounting for the underlying physiology and muscle dynamics of the vocal apparatus. This requires independent control of multiple articulatory parameters, rather than natural coupled movements, increasing cognitive load and limiting expressiveness, especially for nonlinear articulations. Alternatively, articulatory gestures can be altered by a physics-based system approximating the muscle dynamics of the tongue [3, 16, 17]. Such a framework may yield more naturalistic vocal gestures by leveraging performers' inherent motor experience. However, the coupled nature of simulated muscle dynamics may simplify nonlinear trajectories at the expense of precise static or linear control. Characterizing this tradeoff requires a dedicated experimental apparatus. This paper therefore makes two contributions: (1) the design of a comparative apparatus implementing both control approaches, and (2) pilot study results demonstrating quantitative differences between them—establishing both the existence of this difference and a controlled environment for a subsequent full user study.

2 Related Work

This section reviews control paradigms of vocal synthesizers that incorporate tongue physiology and articulatory modeling.

2.1 Mapping Input to Tongue Shape

Ogata et al. [13] developed a data-glove interface that maps finger gestures to oral constrictions and lip aperture, and feeds the resulting vocal tract area functions to an acoustic wave solver [11]. Similar geometric controllers were proposed for desktop and web-based vocal synthesizers, such as Cook's SPASM [4] and Pink Trombone [15] models, which provide real-time articulation control. While these systems demonstrate the viability of articulatory synthesis for musical performance, they rely on arbitrary gesture-to-geometry mappings, lacking biomechanical constraints and natural tongue dynamics.

2.2 Mapping Input to Tongue Biomechanics

Recent advances in physiological computing have enabled artist-researchers to incorporate biomechanics and muscle activation into musical interface design. Examples include Lionetti et al. [10], who used sEMG to extend guitar expression beyond traditional pedalboards, and Díaz-Durán et al. [5], who combined sEMG and force sensors to sonify muscle tension. Similarly, Reed and McPherson [14] applied sEMG to measure vocal musculature activity for synthesizer control. Although such physiological sensors enable direct sonification of articulatory gestures, they are invasive, interfere with natural vocal production, and introduce motion artifacts.

Wang et al. [16, 17] proposed a voice synthesizer that maps force-sensor input to muscle activations of a biomechanical tongue

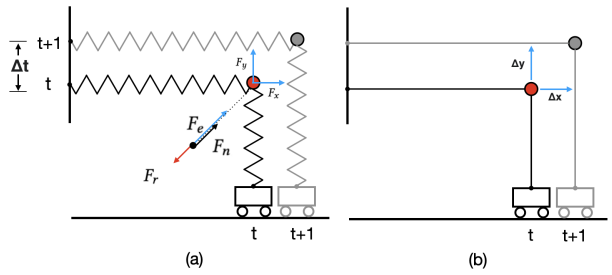

Figure 1: Schematic of a physics-based model vs. geometric model. (a) physics-based model: F_e shows the user applied force (the vector sum of F_x and F_y), F_r shows the restoring force due to spring, resulting in the net force F_n used for computing the acceleration for the next frame t+1. (b) geometric model: Δx and Δy shows the user applied displacement in x and y direction to move the mass directly from t to t+1.

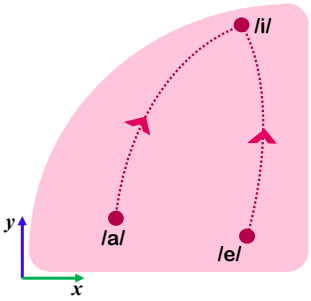

Figure 2: Illustration of diphthong trajectories within 2D Pink Trombone triangular vowel space.

model, demonstrating that force-driven control yields more intelligible and natural-sounding speech than direct geometric manipulation. This improvement stems from the intuitive mapping between force sensors and muscle activations, reducing control complexity. However, the FE-based biomechanical tongue model limits real-time voice synthesis.

We therefore adopt an intermediate approach balancing biomechanical interpretability and computational efficiency: a lightweight spring–mass–damper framework coupled to a 1D vocal tract acoustic model, where spring forces are parameterized to generate target tongue shapes. This enables real-time control of tongue shape and voice synthesis within an intuitive, physics-based control paradigm analogous to muscle activation, retaining the expressive benefits of force-driven models without the overhead of full FE implementations.

3 System Design

This section describes the design of the proposed physics-based controller and a conventional geometric controller to generate tongue trajectories.

3.1 Physics-based Controller

Our physics-based controller employs a Hill-type muscle model, well-suited to tongue biomechanics [6], as the tongue is a muscular hydrostat whose deformation arises from active contraction and passive elasticity. The classical Hill model comprises three components [7]: (1) a contractile element generating force via actin–myosin cross-bridge interactions, (2) a parallel elastic element representing passive tissue and contributing to the force–length relationship without activation, and (3) a series elastic element modeling tendon compliance and filament elasticity, enabling energy storage during deformation. Rather than implementing the full Hill-type formulation, we adopt a computationally efficient spring–mass–damper approximation, where the spring provides contraction force and the damper models energy dissipation. This yields muscle-like dynamics—contraction, return to rest, and transient responses—while remaining suitable for real-time control.

The controller is modeled as a 2D spring–mass system, where a point mass is attached to a fixed anchor by a zero–rest-length spring, as shown in Fig. 1(a). Here, we describe the system dynamics along the x axis; the same formulation also applies along the y axis. A displacement Δx of the mass from the anchor point produces a restoring force (F_r) directed toward the equilibrium according to Hooke's law (Eq. 1). With an external force F_e applied at time t, the net force is $F_n = F_e - F_r$, yielding acceleration a_t via Newton's second law (Eq. 2). We then compute velocity v_{t+1} of the mass for the next time step $t+1$ using a_t (Eq. 3). A damping term D_f is introduced (Eq. 4) to model energy dissipation, preventing sustained oscillations in response to external perturbations. Finally, the point mass position x_{t+1} is computed using Eq. 5. This simplified physics-based spring–mass model represents the interaction between active force generation, passive elasticity and damping as in muscle dynamics.

The spring constant K determines the restoring force magnitude, while D_f controls energy dissipation. Larger values of K and D_f yield faster acceleration and more reactive motion, which may reduce stability. Smaller values produce slower, more damped motion, causing the mass to settle quickly with a stickier response. We empirically tuned $K = 0.16$ and $D_f = 0.05$ to achieve a balanced control-to-display ratio, providing sufficient temporal persistence for users to generate desired tongue trajectories while maintaining stable interaction. The discrete update equations are given as follows:

$$F_y = -K\Delta u \tag{1}$$

$$a_t = F_n/m \tag{2}$$

$$v_{t+1} = v_t + a_t\Delta t \tag{3}$$

$$v_{t+1} = v_{t+1}D_f \tag{4}$$

$$u_{t+1} = u_t + v_{t+1}\Delta t \tag{5}$$

where, u represents displacement along the x/y axis

3.2 Geometric Controller

The geometric controller maps user inputs directly to the spatial position of a point mass in the control space (Fig. 1(b)), without any force-based interactions. Each incremental input produces a fixed step displacement along the corresponding Cartesian axis (x or y), updating the mass position deterministically depending on the input magnitude and direction, without inertia or damping.

3.3 Implementation

For comparative analysis, we integrated both controllers with Pink Trombone—a web-based real-time vocal synthesizer [12]. Using each controller, target tongue trajectories were produced

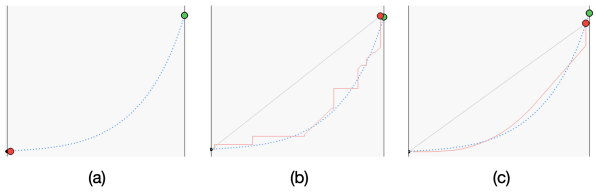

Figure 3: Interaction Window: red dot represents the mass, green dot represents the target, blue dotted line shows reference curve, red line shows user drawn curve. (a) Mass at initial position (b) Curve drawing using geometry model (c) Curve drawing using physics-based model

within the Pink Trombone control space (Fig. 2), yielding diphthong sounds. Both controllers are implemented as a 2D keyboard-based input interface, with each axis corresponding to an independent control dimension. Although the formulation naturally generalizes to higher-dimensional inputs, the control space is intentionally limited to 2D to match the degrees of freedom in Pink Trombone's vowel space.

Motion was constrained to the positive x and y directions from a fixed corner anchor to avoid symmetric redundancy. User input controlled force magnitude along each axis: when the force increased, the mass moved according to the simulated dynamics, while reducing it allowed the restoring spring force to return the mass toward equilibrium.

4 Pilot Study Development

To prepare for a larger user study, we conducted a pilot study to refine the experimental design, identify which interaction dimensions were most important to measure, and verify that our implementation could capture meaningful differences between the two control approaches. Rather than serving as a formal evaluation, this pilot functioned as a developmental step: it helped us debug the task structure, tune controller parameters, and assess whether the observed differences were strong enough to justify a full user study of musicality and expressive interaction in the Pink Trombone vowel space.

4.1 Goals and Task Design

We compared two control paradigms: a physics-based model and a geometric model. The two differ in how keyboard input is translated into motion. In the physics-based model, button presses W (up), X (down), J (left), and L (right) increment directional force applied to a point mass, resulting in motion shaped by simulated dynamics. In the geometric model, the same inputs map directly to mass position, bypassing intermediate dynamics. In both cases, interaction produces a time-varying 2D trajectory in the control space (Fig. 3).

A primary goal of the pilot was to determine whether the two controllers would differ along dimensions relevant to a later study of musicality. To probe this, we designed a trajectory-tracing task based on diphthong-like movements through the Pink Trombone vowel space. Participants followed displayed reference trajectories that included both straight and curved paths between vowel targets. These reference paths were generated using Bézier curves approximating transitions in the Pink Trombone space. The final reference set consisted of four straight trajectories, three convex curves, and three concave curves, providing a balanced range of movement demands.

We selected three quantitative measures to support later study design: Fréchet distance [1], drawing speed, and normalized bending energy. Together, these metrics allowed us to examine controllability, efficiency, and dynamic trajectory shape. Each author completed the pilot using both controllers. The protocol included a warm-up phase, a controlled trajectory-tracing phase, and a free-form exploration phase:

- **Warm-Up Phase** Participants familiarized themselves with the keyboard interface and adjusted controller-specific parameters, such as step size for the geometric model and damping for the physics-based model.
- **Controlled Trajectory-tracing Phase** Participants traced 10 predefined reference curves, each repeated three times per controller, yielding 30 trials per model per participant. At the start of each trial, the mass was reset to a fixed anchor position and a reference path was displayed as a dotted line ending at a target point.
- **Free-Form Exploration Phase** Participants interacted with a Pink Trombone vowel space visualization in which phonemes were displayed at fixed spatial locations, and the mass moved via the same keyboard interface, with audio synthesized in real time from mass position.

4.2 Pilot Study Outcomes

The pilot study revealed two main outcomes (Fig. 4). First, the two controllers were broadly comparable in basic controllability. For straight trajectories, the physics-based model yielded a mean Fréchet distance of 42.0 px, compared with 45.3 px for the geometric model. For curved trajectories, the corresponding values were 86.6 px and 83.3 px. These small differences suggest that both interfaces support similar levels of spatial accuracy.

Second, the pilot showed meaningful differences in movement dynamics. For straight trajectories, the geometric model was faster (500.3 px/s) than the physics-based model (446.4 px/s), while for curved trajectories the pattern reversed, with the physics-based model outperforming the geometric model (173.3 px/s vs. 132.6 px/s). The strongest separation appeared in normalized bending energy for curved trajectories, where the physics-based model produced substantially lower values (211.6) than the geometric model (1191.1). A lower bending energy indicates that users can generate smoother trajectories with the physics-based controller, with reduced oscillatory motion.

4.3 Summary and Discussion

Overall, this pilot study established that the implementation can reveal meaningful differences between the two control models while maintaining comparable baseline controllability. More importantly, it helped identify which aspects of performance are likely to matter in a larger study of musicality: not only whether users can reach intended targets, but how trajectory shape, temporal continuity, and motion smoothness vary across control paradigms. Although the pilot results do not establish definitive statistical claims, they demonstrate measurable and theoretically meaningful differences between the two control approaches, setting the stage for a full-scale user study. With an adequately powered sample, future work can rigorously test hypotheses regarding performance efficiency, trajectory smoothness, and multidimensional musicality generation potential across input modalities.

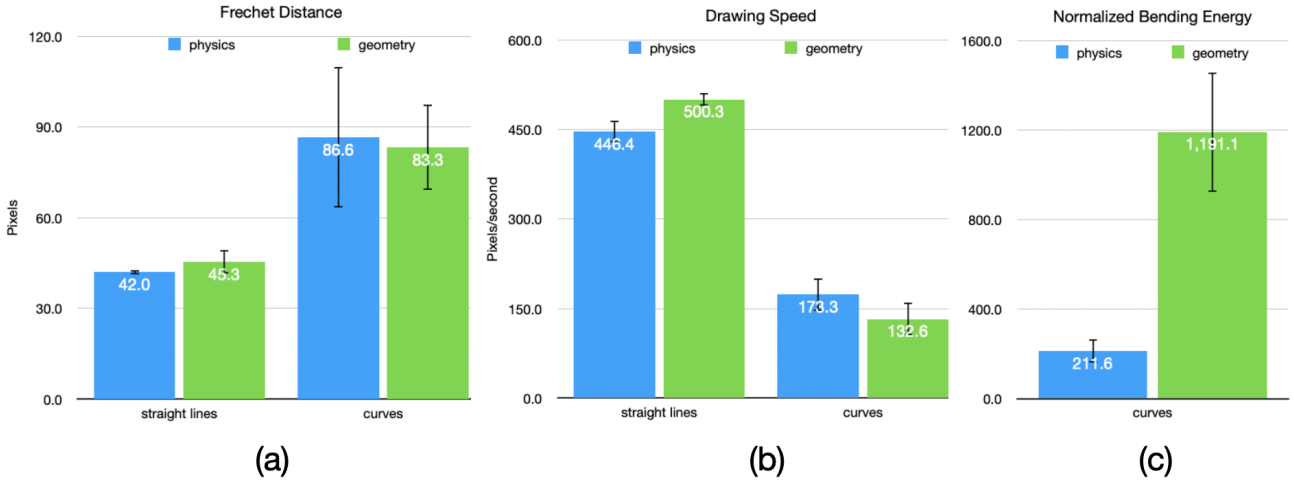

Figure 4: Quantitative Results physics-based vs. geometry model: (a) Frechet Distance to capture similarity between the reference and produced trajectory; (b) drawing speed, computed as curve length divided by completion time, to capture efficiency; and (c) normalized bending energy, to capture smoothness and oscillation. Together, these metrics allowed us to examine controllability, efficiency, and dynamic trajectory shape.

5 Ethical Standards

This work adheres to accepted principles of ethical and professional conduct, ensuring transparency and objectivity throughout. No data was collected from users; thus, informed consent and welfare statements are not applicable to this research. Additionally, there are no conflicts of interest, financial or otherwise, to disclose.

Acknowledgments

This work is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada (2025-04931).

References

[1] Helmut Alt and Michael Godau. 1995. Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications* 5, 01n02 (1995), 75–91.

[2] Anders-Petter Andersson and Birgitta Cappelen. 2013. Designing empowering vocal and tangible interaction. In *The International conference on new interfaces for musical expression*. Seoul National University, 406–412.

[3] Perry Raymond Cook. 1991. *Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing*. Stanford University.

[4] Perry R Cook. 1993. SPASM, a real-time vocal tract physical model controller; and singer, the companion software synthesis system. *Computer Music Journal* 17, 1 (1993), 30–44.

[5] Joaquín R Díaz Durán, Laia Turmo Vidal, and Ana Tajadura-Jiménez. 2023. Joakinator: An Interface for Transforming Body Movement and Perception through Machine Learning and Sonification of Muscle-Tone and Force. 94–97 pages.

[6] Nicolas Hermant, Pascal Perrier, and Yohan Payan. 2017. Human tongue biomechanical modeling. *Biomechanics of living organs* (2017), 395–411.

[7] Archibald Vivian Hill. 1938. The heat of shortening and the dynamic constants of muscle. *Proceedings of the Royal Society of London. Series B-Biological Sciences* 126, 843 (1938), 136–195.

[8] Rébecca Kleinberger, Nikhil Singh, Xiao Xiao, and Akito van Troyer. 2022. Voice at NIME: a Taxonomy of New Interfaces for Vocal Musical Expression. In *NIME 2022*. PubPub.

[9] Bernd J Kröger. 2022. Computer-implemented articulatory models for speech production: A review. *Frontiers in Robotics and AI* 9 (2022), 796739.

[10] Davide Lionetti, Paolo Belluco, Massimiliano Zanoni, Luca Turchet, et al. 2024. Muscle-Guided Guitar Pedalboard: Exploring Interaction Strategies Through Surface Electromyography and Deep Learning. In *Proceedings of the International Conference on New Interfaces for Musical Expression*.

[11] Hiroki Matsuzaki, Antoine Serrurier, Pierre Badin, and Kunitoshi Motoki. 2014. One-dimensional and three-dimensional propagation analyses of acoustic characteristics of Japanese and French vowel/a/with nasal coupling. *Acoustical Science and Technology* 35, 1 (2014), 35–41.

[12] Jack Mullen. 2006. *Physical modelling of the vocal tract with the 2D digital waveguide mesh*. Ph. D. Dissertation. University of York.

[13] Kohichi Ogata, Kohei Matsumura, and Yusuke Matsuda. 2015. Data-glove-driven vocal tract configuration methods for vowel synthesis. *Acoustical Science and Technology* 36, 6 (2015), 527–536.

[14] Courtney Reed and Andrew McPherson. 2020. Surface electromyography for direct vocal control. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 458–463.

[15] Neil Thapen. 2017. Pink Trombone: Bare-handed procedural speech synthesis. https://dood.al/pinktrombone/

[16] Johnty Wang, Nicolas d'Alessandro, Sidney Fels, and Robert Pritchard. 2012. Investigation of Gesture Controlled Articulatory Vocal Synthesizer using a Bio-Mechanical Mapping Layer. In *Proceedings of the International Conference on New Interfaces for Musical Expression*.

[17] Johnty Wang, Nicolas d'Alessandro, Sidney S Fels, and Bob Pritchard. 2011. Squeezy: Extending a multi-touch screen with force sensing objects for controlling articulatory synthesis. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 531–532.