Gestural articulatory synthesizer based on real-time MRI data

Yijing Lu[1], *Justin Ly*[2], *Shrikanth Narayanan*[2], *Louis Goldstein*[1], *Asterios Toutios*[2]

[1]Department of Linguistics, University of Southern California, USA
[2]Signal Analysis & Interpretation Laboratory (SAIL), University of Southern California, USA
yijinglu@usc.edu, justinly@usc.edu, shri@sipi.usc.edu, louisgol@usc.edu, toutios@usc.edu

Abstract

This paper presents an updated implementation of an articulatory synthesis framework wherein speaker-specific articulatory models derived from real-time MRI data are animated using dynamical systems. The current implementation of these dynamical systems in Python uses discretization to solve differential equations. The efficiency and stability of the articulatory synthesizer are thereby improved compared to earlier versions, allowing multiple articulatory gestures to be simultaneously operating on the same vocal tract component. In the current implementation, we also propose an explicit specification of gestural score parameters, which controls the relative timing, targets, and natural frequencies of dynamical systems operating on vocal-tract constriction tasks, as the input of the synthesizer. We demonstrate the speaker-specificity and validity of the current implementation by comparing the synthesized vocal tract deformations in different speakers for the same word and by comparing the synthesized articulatory trajectories to the corresponding trajectories in real data.

Index Terms: articulatory synthesis, articulatory phonology, dynamical system, real-time MRI

1. Introduction

Earlier work by our group [1] proposed a computational framework for synthesizing vowel-consonant-vowel sequences from gestural specifications to dynamic vocal tract configurations and speech acoustics. The framework was directly informed by real-time magnetic resonance imaging (rtMRI) data of human speech production [2]. Speaker specific, statistical articulatory models were derived from rtMRI data by means of automatic air-tissue segmentation [3] and factor analysis of air-tissue boundaries [4]. Such individualized articulatory models were further controlled by the temporal activation of articulatory gestures to generate the time-course of a speaker's mid-sagittal vocal tract shape deformation for a target utterance. According to the theories of Articulatory Phonology [5] and Task Dynamics [6], each articulatory gesture can be represented by a dynamical system, specifically, a critically damped oscillator with two parameters: target and stiffness. These two parameters could also be estimated from rtMRI data.

Yet, the previous work did not address the issue of gestural overlap. When more than one dynamical systems were simultaneously active, the resulting vocal tract shaping became unstable. To solve this issue, in the present work we introduce a discretization of dynamical systems, which improves the stability and efficiency of the implementation. We also propose a modified gestural score specification wherein the relative timing, target, and natural frequencies of each articulatory gesture (dynamical system) are explicitly controlled. Iterative gradient descent optimization [7] is also used to refine the gestural specifications in order to better fit the observed vocal tract dynamics. With the current implementation, the utterances that can be synthesized are not limited to vowel-consonant-vowel sequences.

We demonstrate the utility of this improved articulatory synthesizer implemented in Python by simulating the mid-sagittal vocal tract dynamics of some more complex words: /spæn/, which contains a complex onset; /baɪt/ and /bjut/, which contain a diphthong. Similar hypothesized gestural scores were used to synthesize /spæn/ in one female speaker and one male speaker, and we expect to see speaker-specific vocal tract shaping to be generated. The gesture scores for /baɪt/ and /bjut/ were estimated from the rtMRI data of one female speaker. The synthesized articulatory events were compared to the recorded vocal tract dynamics from the same speaker. The gestural scores of /bjut/ were further optimized to generate better fitted vocal tract shaping dynamics.

Although our articulatory synthesizer also includes the conversion of mid-sagittal vocal tract dynamics to area function dynamics, which can further generate acoustic signals, in this paper we focus on demonstrating the synthesis results of vocal tract shaping. Both the synthesizer and the synthesis results shown in this paper are available online[1].

2. Method

2.1. Data, Articulatory Model, Forward Map

We used rtMRI data of two speakers from the USC multispeaker database [8]: a 29-year old female speaker born in Brawley, CA and a 36-year old male speaker born in Medina, OH. The segmentation of air-tissue boundaries in the mid-sagittal images (rtMRI video frames) was done by an updated version of the automatic segmentation algorithm in [3]. The derived air-tissue boundaries were further subjected to a guided factor analysis [4] to identify speaker-specific linguistically meaningful components of the vocal tract configuration (Fig. 1). The articulatory models hence represent the vocal tract shaping at any point in time as a linear combination of 8 speaker-specific components, weighted by a dynamically changing array of parameters $\mathbf{w}$, which correspond to the degree of deformation of each component.

The tasks of phonetic units in Task Dynamics [6] include constrictions produced in different vocal tract locations. To quantify the task-representation of the data, segmentation of the articulatory contours was used to measure constriction degrees (distances between passive and active articulators [9, 10]) at 6 locations (bilabial, alveolar, palatal, velar, pharyngeal, and velopharyngeal), represented by an array $\mathbf{z}$.

The forward map expresses the relation between articulatory degrees of freedom ($\mathbf{w}$) and the corresponding task representation ($\mathbf{z}$). It is in general non-linear, but we approached it as a locally linear mapping. This was achieved using a hier-

[1]https://github.com/toutios/garsy

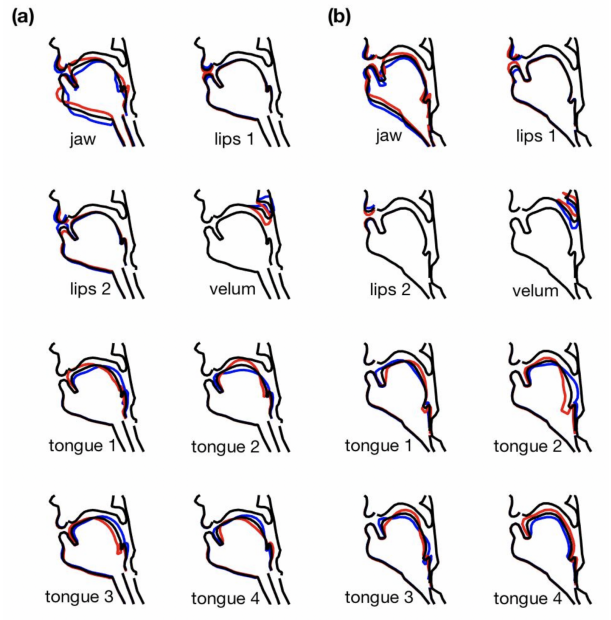

Figure 1: *Components of the articulatory models for (a) the female speaker and (b) the male speaker. Line segments denote mean and ±2 standard deviations.*

Input: Dataset of $(\mathbf{w}, \mathbf{z})$ (Root node), threshold θ
Output: Binary tree (each node is a cluster of data)
repeat
  Fit a linear regression model on $\mathbf{w} \mapsto \mathbf{z}$ for the data in the node;
  if *Error of fit below* θ *OR size of split node is less than minimum size* **then**
    Mark this node as a leaf
  else
    Split the node data in two clusters (children nodes) by applying k-means on $\mathbf{w}$
  end
until *All data have a corresponding leaf node*;
Algorithm 1: Hierarchical clustering

archical clustering procedure [9] in which the dataset (pairs of representation vectors $\mathbf{w}$ and $\mathbf{z}$) was split iteratively into two clusters until a linearity test indicated that, within a cluster, articulatory parameters could be mapped approximately linearly to constriction degrees, i.e., $\mathbf{z} = \mathbf{G}(\mathbf{w}) = F * \mathbf{w} + \mathbf{z_c}$ where F is a 6×8 matrix. The hierarchical clustering algorithm for splitting data into clusters is specified in Algorithm 1.

Each speaker ends up having multiple clusters in which the mapping from $\mathbf{w}$ to $\mathbf{z}$ is linear, representing the speaker-specific forward maps from articulatory parameters to constriction degrees.

2.2. Specification of Gestural Scores

In Articulatory Phonology [5] and Task Dynamics [6], primitive phonological units are articulatory gestures, which are goal-directed movements of vocal tract effectors that achieve local constriction tasks. The spatiotemporal unfolding of an articulatory gesture is modeled as a dynamical equation describing the change of constriction degree in a specific vocal tract region. The locally linear forward map from articulatory parameters, $\mathbf{w}$, to constriction degree tasks, $\mathbf{z}$, and its Jacobian J can also be inverted to allow the dynamical system characterizing a gesture to be expressed in terms of time evolution of the articulatory shaping parameters, $\mathbf{w}$. Given a target constriction degree $\mathbf{z}_0$, the dynamical system [1, 6] that governs the dynamics of articulatory parameters is:

$$\begin{aligned}\ddot{\mathbf{w}} =& J^*(-BJ\dot{\mathbf{w}} - K(\mathbf{G}(\mathbf{w}) - \mathbf{z}_0)) - J^*\dot{J}\dot{\mathbf{w}} \\ & - (I_8 - J^*J)B_N\mathbf{w} - G_N(-B_N\mathbf{w} - K_N\mathbf{w})\end{aligned}$$

where $\mathbf{G}(\mathbf{w})$ is the forward map from articulatory parameters to constrictions, G_N, B_N and K_N are parameters of the *neutral attractor* (see [1, 6] for details). The Jacobian J of $\mathbf{G}$, with its derivative $\dot{J}$ and pseudoinverse J^*, are readily available because of the linear mapping within each cluster. The stiffness K and damping B matrices are set dynamically as functions of the target utterance. In practice, setting an array of 6 *natural frequencies* $\omega_\mathbf{o}$, each corresponding to a place of articulation (location of constriction), fully determines K and B. The values of $\omega_\mathbf{o}$ are set on the basis of the particular gestures composing the utterance, as is an array of *target* constriction degrees, $\mathbf{z}_o$. At each point in time then, the dynamical system can be characterized by an array of 6 tuples (ω_o, z_o). These can be visualized in a way that corresponds to *gestural scores* in [5], as shown in Fig. 2. Each block represents the duration of activity of an articulatory gesture, along with the natural frequency (ω_o) and the target (z_o) of the gesture governing that dynamical system. The spaces between the blocks represent times during which no gesture actively controls that constriction, and these are internally represented with a zero value of ω_o.

2.3. Discretization of Dynamical System

In [1], the dynamical systems were implemented using MATLAB's `ode45` functions. This led to unstable solutions when, at any given time, more than one element of the natural frequency array were non-zero. Hence, overlapping gestures could not be simulated in the previous work. In the present work, we address this problem by discretization. Assuming a time-step h between two consecutive samples of (ω_o, z_o), we replace the derivatives by finite differences:

$$\begin{aligned}\dot{\mathbf{w}} &\leftarrow (\mathbf{w}[n] - \mathbf{w}[n-1])/h \\ \ddot{\mathbf{w}} &\leftarrow (\mathbf{w}[n] - 2\mathbf{w}[n-1] + \mathbf{w}[n-2])/h^2\end{aligned}$$

After some algebra, we get a linear system of the form:

$$\begin{aligned}(I_8 + hA_1 + h_2A_2)\mathbf{w}[n] = \mathbf{w}[n-2] + \\ 2\mathbf{w}[n-1] + hA_1\mathbf{w}[n-1] + h^2J^*K(\mathbf{z}_o - \mathbf{z}_c)\end{aligned}$$

with

$$\begin{aligned}A_1 &= -J^*BJ - J^*\dot{J} - B_N + J^*JB_N - G_NB_N \\ A_2 &= -G_NK_N - J^*KF\end{aligned}$$

where the matrix sum on the left side is 8×8 and in general invertible. Thus, given a (ω_o, z_o) specification and two initial samples of parameter arrays, subsequent samples can be calculated stepwise by solving the above linear system. In particular, the loop over the N samples of an utterance can be implemented, as shown in Algorithm 2.

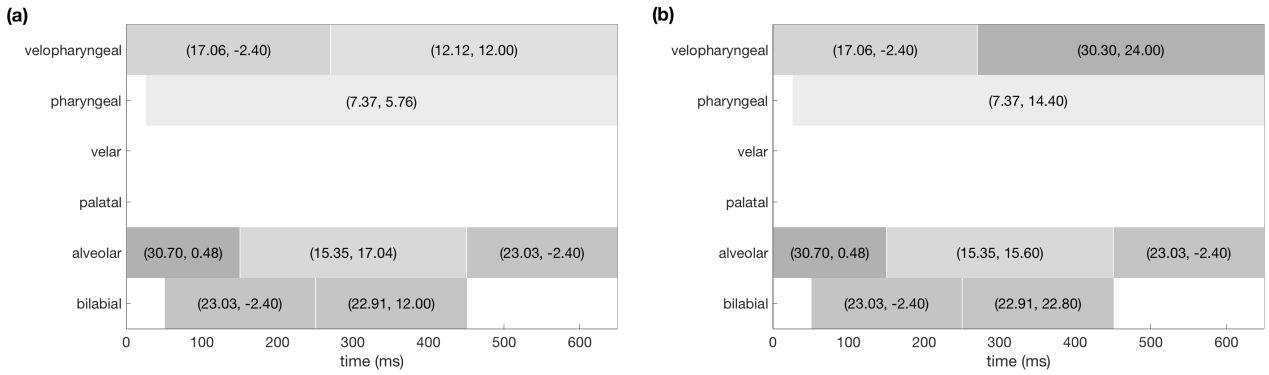

Figure 2: (ω_o, z_o) *gestural specifications of /spæn/ for (a) the female speaker and (b) the male speaker.*

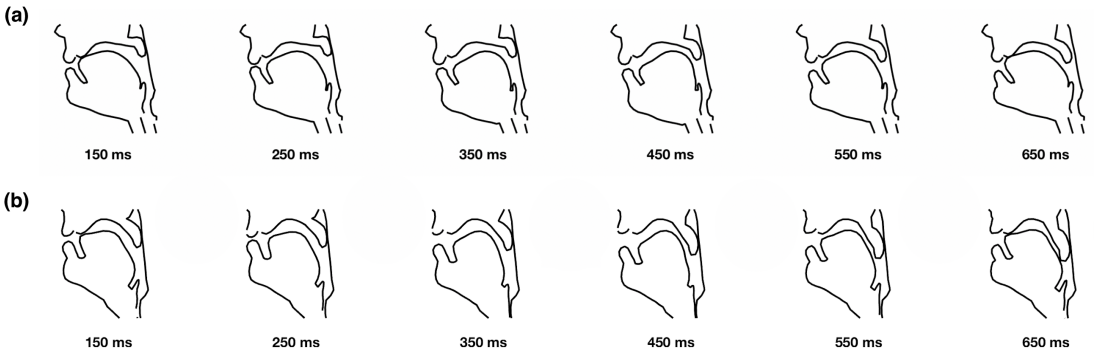

Figure 3: *Snapshots of synthesized vocal tract shapes for /spæn/ at the time points indicated, produced by (a) the female speaker and (b) the male speaker.*

Input: $\mathbf{w}[0], \mathbf{w}[1], \boldsymbol{\omega}_o[n], \mathbf{z}_o[n], n = 2...N$
Output: $\mathbf{w}[n], n = 2...N$
for *n=2...N* **do**
  Find cluster with center $\mathbf{w_c}$ closest to $\mathbf{w}[n-1]$;
  Retrieve $\mathbf{z_c}$, F, J, $\dot{J}$ for that cluster;
  Calculate $K(\boldsymbol{\omega}_o), B(\boldsymbol{\omega}_o)$;
  Solve the linear system for $\mathbf{w}[n]$
end

Algorithm 2: Stepwise solution of linear system

2.4. Optimization of Gestural Scores

With the help of an important body of experimental literature on Articulatory Phonology, combined with our own observations from rtMRI, we can infer the gestural scores of a target utterance to be synthesized. We first evaluate these hypothesized gestural scores by comparing that the synthesized articulatory events (e.g., constriction and relative timing) to their recorded counterparts when rtMRI data are available. If apparent deviations are observed, we then adjust the specific parameters of the dynamical systems comprising the gestural scores (i.e., time instants of activation, targets, natural frequencies) using an optimization loop. This optimization is achieved using a gradient descent approach [7], as shown in Algorithm 3.

Input: Recorded sequence $\mathbf{w}[n], \quad n = 1, \ldots, N$; M gestures with parameters s_j (start time), e_j (end time), ω_j (natural frequency), t_j (target); θ (threshold); γ (learning rate)
for D *iterations* **do**
  Synthesize $\hat{\mathbf{w}}[n]$ from current gestural score parameters;
  $E = \sum_i ||\hat{\mathbf{w}}[n] - \mathbf{w}[n]||_2$;
  for *j=1...M* **do**
    for x *in* $\{s_j, e_j, \omega_j, t_j\}$ **do**
      Compute $\partial E / \partial x_j$;
      Update $x_j \leftarrow x_j - \gamma \partial E / \partial x_j$;
    end
  end
end

Algorithm 3: Optimization

3. Results

We synthesized the word /spæn/ for two speakers using similar gestural scores. The input gestures specifications were hypothesized based on the example gestural scores for the same word in [5]. Fig. 2 visualizes the hypothesized, non-optimized (ω_o, z_o) specifications of /spæn/ for the female and male speakers. Each block represents an active articulatory gesture with a natural frequency ω_o and a target z_o. z_o controls the target constriction degree of each articulatory gesture. ω_o controls the how fast that target constriction can be achieved through the synergy of multiple speech articulators. The difference between the gestural specifications for the female and male speakers mainly lies in the targets of vowel gestures, which were set in proportion to the size of the speaker's oral cavity. Higher values were set as the

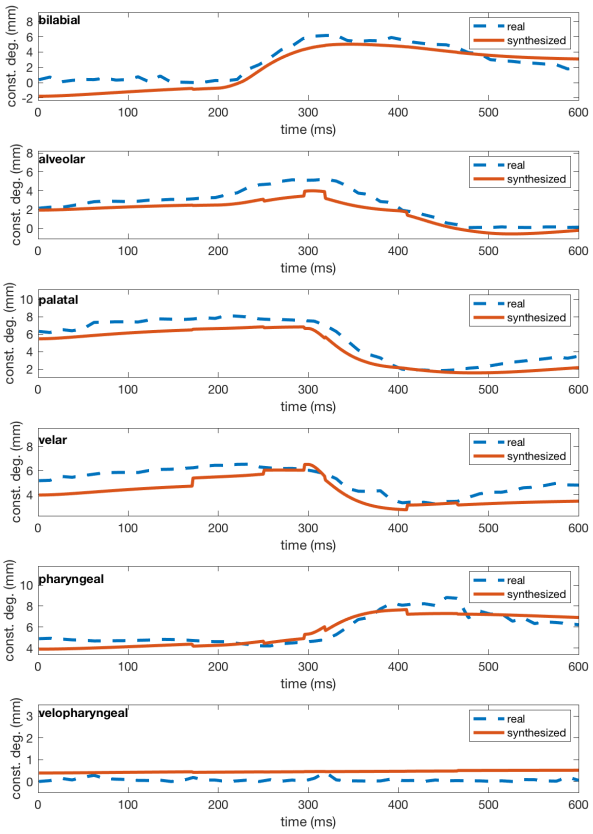

Figure 4: *Recorded (blue dashed line) and synthesized (orange solid line) degrees of constriction for /baɪt/ produced by the female speaker.*

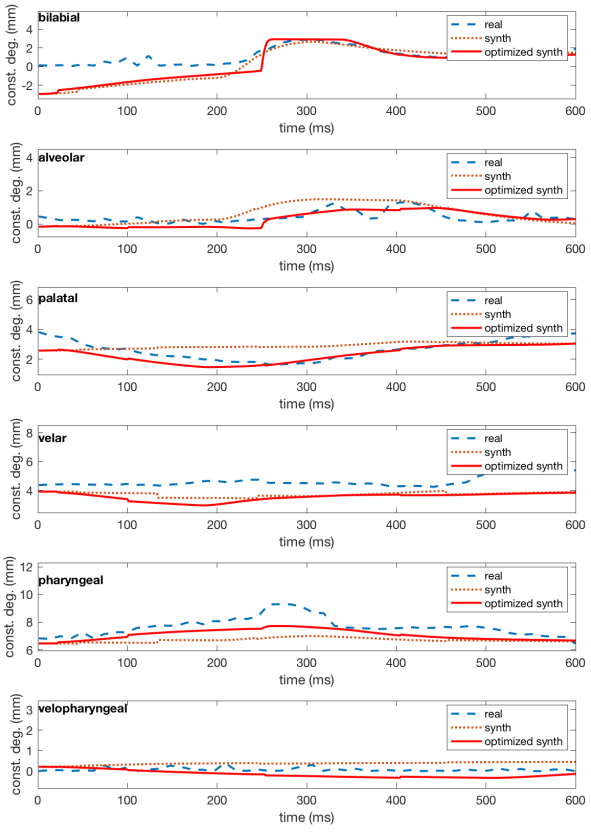

Figure 5: *Recorded (blue dashed line) and synthesized degrees of constriction for /bjut/ produced by the female speaker before (orange dotted line) and after optimization (red solid line).*

targets for the releases of the bilabial and alveolar constrictions in the male speaker due to the bigger oral cavity, which requires a higher degree of lip and tongue tip opening in the following vowels. Similarly, pharyngeal constriction during the vowel production is not as narrow in the larger vocal tract. Following the same logic, the target of the velopharyngeal constriction for /n/ also has a wider constriction target in the male speaker. We also increased the natural frequency of the velum opening gesture. Other than these, the target and the natural frequency of oral consonant gestures and the temporal coordination of all gestures were kept the same across speakers.

Fig. 3 shows the temporal evolution of synthesized mid-sagittal vocal tract shapes for /spæn/ produced by the female and male speakers. Critical articulatory events like the tongue tip constriction for word-initial /s/ (150 ms) and word-final /n/ (650 ms), the velum lowering for /n/ (650 ms), the labial constriction for /p/ (250 ms), and the overall tongue configuration for /æ/ (450 ms) were achieved for both speakers.

Since the rtMRI recordings of /spæn/ is not available in our dataset, we did not have the baseline to which the synthesized /spæn/ can be compared. But the comparison was possible for /baɪt/ and /bjut/. As shown in Fig. 4, the trajectories of constriction degrees for the synthesized /baɪt/ are a good approximation of the real trajectories. This suggests that our manual fitting of gestural patterns to recorded vocal tract dynamics was successful.

Fig. 5 shows the recorded and synthesized constriction degree trajectories for /bjut/ before and after optimization. Before optimization, the synthesized trajectories (orange dotted line) were quite deviated from the real trajectories (blue dashed line), indicating that the manual fitting of gestural scores was not very successful. With optimization implemented, the resulting synthesized trajectories (red solid lines) approximate the real trajectories better, especially in the cases of alveolar, palatal, and pharyngeal constrictions.

4. Conclusion

In the present paper we discussed the progress toward a framework for articulatory synthesis directly informed by real-time MRI data, which includes the conversion of gestural spatiotemporal specifications to sequences of mid-sagittal vocal tract slices. We proposed an updated implementation of the framework in Python highlighting the use of discretization and enhanced gestural specifications. We also added an optimization loop which can adjust the hypothesized gestural scores according to the recorded vocal tract dynamics using gradient descent. The synthesis results of some complex utterances suggest the improvement of the synthesizer compared to its earlier versions. Further work will include designing gestural scores for more utterances, and eventually developing an inventory of gestural score templates for English utterances.

5. Acknowledgement

Work supported by NSF grant 1908865.

6. References

[1] R. Alexander, T. Sorensen, A. Toutios, and S. Narayanan, "A modular architecture for articulatory synthesis from gestural specification," *The Journal of the Acoustical Society of America*, vol. 146, no. 6, pp. 4458–4471, 2019.

[2] A. Toutios, D. Byrd, L. Goldstein, and S. Narayanan, "Advances in vocal tract imaging and analysis." in *The Routledge Handbook of Phonetics*, W. Katz and P. Assmann, Eds. London and New York: Routledge, 2019.

[3] E. Bresch and S. S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 28, no. 3, pp. 323–338, Mar. 2009.

[4] A. Toutios and S. S. Narayanan, "Factor analysis of vocal-tract outlines derived from real-time magnetic resonance imaging data," in *International Congress of Phonetic Sciences (ICPhS)*, Glasgow, UK, Aug. 2015.

[5] C. P. Browman and L. Goldstein, "Articulatory Phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[6] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, 1989.

[7] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, Jan. 2018.

[8] Y. Lim, A. Toutios, Y. Bliesener, Y. Tian, S. G. Lingala, C. Vaz, T. Sorensen, M. Oh, S. Harper, W. Chen, Y. Lee, J. Töger, M. L. Montesserin, C. Smith, B. Godinez, L. Goldstein, D. Byrd, K. S. Nayak, and S. S. Narayanan, "A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images," *arXiv:2102.07896*, 2021.

[9] T. Sorensen, A. Toutios, L. Goldstein, and S. Narayanan, "Characterizing vocal tract dynamics across speakers using real-time MRI," in *Interspeech*, San Francisco, CA, 2016.

[10] ——, "Task-dependence of articulator synergies," *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. 1504–1520, Mar. 2019.