INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Characterizing vocal tract dynamics across speakers using real-time MRI

Tanner Sorensen[12], *Asterios Toutios*[1], *Louis Goldstein*[2], *Shrikanth Narayanan*[12]

[1]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA
[2]Department of Linguistics, University of Southern California, Los Angeles, CA, USA
tsorense@usc.edu, toutios@usc.edu, louisgol@usc.edu, shri@sipi.usc.edu

Abstract

Real-time magnetic resonance imaging (rtMRI) provides information about the dynamic shaping of the vocal tract during speech production and valuable data for creating and testing models of speech production. In this paper, we use rtMRI videos to develop a dynamical system in the framework of Task Dynamics which controls vocal tract constrictions and induces deformation of the air-tissue boundary. This is the first task dynamical system explicitly derived from speech kinematic data. Simulation identifies differences in articulatory strategy across speakers ($n = 18$), specifically in the relative contribution of articulators to vocal tract constrictions.

Index Terms: task dynamics, real time MRI, factor analysis

1. Introduction

Real-time magnetic resonance imaging (rtMRI) provides information about the dynamic shaping of the vocal tract during speech production [1]. Task Dynamics [2] views speech production as involving controlled dynamics of this shaping to make constrictions of the vocal tract. That is, Task Dynamics characterizes the relation between vocal tract shape and vocal tract constrictions. Using rtMRI we can directly observe and measure salient articulatory details such as vocal tract shape and constriction degree which reflect the state of the vocal tract as a dynamical system. This makes it possible to estimate parameters of the task dynamical system from rtMRI videos.

This paper presents a task dynamical method for quantifying differences by speaker in how much each individual articulator (i.e., each independently controllable degree of freedom) contributes to a constriction. Here we give special consideration to the tongue, lips, and jaw. In the Task Dynamics framework, how much each articulator contributes to constrictions is determined by manually assigning weights to the articulators [2] or based on theoretical considerations [3]. We present a way to estimate these weights from rtMRI videos and to quantify variability in the relative contributions of the tongue, lips, and jaw. This variability may depend on speech task [4], sociolinguistic factors, sex [5], and anatomy [6, 7].

Section 2 describes data acquisition, postprocessing, dynamical system parameter estimation, and simulations. Section 3 presents the simulation results, namely, a systematic quantification of how much each subject uses their tongue, lips, and jaw to make a constriction.

2. Method

2.1. Experiment and imaging

Eighteen (9 m, 9 f) speakers of American English participated. The upper airways were imaged while the subject lay supine in the MRI scanner with their head firmly but comfortably padded at the temples to minimize motion. The subject read two repetitions of four prose passages (i.e., the rainbow passage [8], the grandfather passage [9], the North Wind passage [10], and a passage based on the dialect (shibboleth) sentences of the DARPA-TIMIT corpus [11]) off a back-projection screen from inside the scanner bore without moving their head. Speaker M2 produced only two repetitions of the rainbow passage. The nature of the experiment and the protocol was explained to the subject before they entered the scanner. The subject was paid for their time upon completion of the session. The USC Institutional Review Board has previously approved the data collection procedures.

Data were acquired at Los Angeles County Hospital on a Signa Excite HD 1.5T scanner (GE Healthcare, Waukesha WI) with gradients capable of 40 mT/m amplitude and 150 mT/m/ms slew rate. A body coil was used for radio frequency (RF) signal transmission. A custom upper airway receiver coil array was used for RF signal reception. This 4-channel array included two anterior coil elements and two coil elements posterior to the head and neck. However, only the two anterior coils were used for data acquisition (the posterior coils of this hardware are not used because they have been previously shown to result in aliasing artifacts).

The rtMRI acquisition protocol is based on a spiral fast gradient echo sequence. This is a scheme for sampling the spatial frequency domain (k-space) in which data are acquired in spiraling patterns. Thirteen interleaved spirals together form a single image. Each spiral is acquired over 6.164 ms (repetition time, TR, which includes slice excitation, readout, and gradient spoiler) and thus every image comprises information spanning $13 \times 6.164 = 80.132$ ms. A sliding window technique is used to allow for view sharing and thus to increase frame rate [1]. The TR-increment for view sharing is seven acquisitions, which results in the generation of an MRI movie with a frame rate of $1/(7 \times \text{TR}) = 1/(7 \times 6.164 \text{ ms}) = 23.18$ frames/sec [1, 12, 13]. The imaging field of view is 200×200 mm, the flip angle is 15°, and the receiver bandwidth ± 125 kHz. Slice thickness is 5 mm, located midsagittally; image resolution in the sagittal plane is 68×68 pixels (2.9 mm^2/pixel). Scan plane localization of the midsagittal slice is performed using RTHawk (HeartVista, Inc., Los Altos, CA), a custom real-time imaging platform [14].

MR image reconstruction was performed using MATLAB. Images from each of the two anterior coils of the 4-channel coil array were formed using gridding reconstruction [15, 12] and the two images were combined by taking their root sum-of-squares in order to improve image signal-to-noise ratio and spatial coverage of the vocal tract.

http://dx.doi.org/10.21437/Interspeech.2016-583

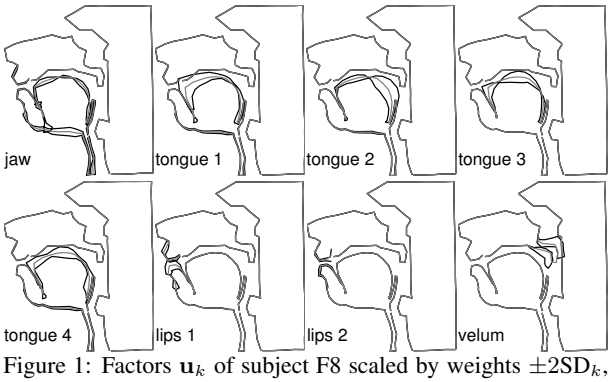

Figure 1: Factors $\mathbf{u}_k$ of subject F8 scaled by weights $\pm 2\text{SD}_k$, where SD_k is the standard deviation of weight w_k

2.2. Air-tissue segmentation

We used the unsupervised algorithm of [16] to segment airway from tissue. We manually constructed a midsagittal morphological template of the upper airway for each speaker. A hierarchical gradient descent procedure registered this template to each rtMRI video frame to approximate the sagittal air-tissue boundaries as polylines. We stacked the xy-coordinates of the polyline vertices on top of one another to represent the air-tissue boundary of frame n as the vector $\mathbf{a}_n \in \mathbf{R}^{2d}$, where d is the number of polyline vertices. The data matrix $\mathbf{A}$ had rows $\mathbf{a}_1^\intercal, \mathbf{a}_2^\intercal, \ldots, \mathbf{a}_q^\intercal$ ($q = 6792 \pm 921$, mean $\pm$ SD, not counting M2, for whom $q = 1628$).

2.3. Factor analysis of vocal tract shapes

In order to distinguish movements of the tongue, lips, and jaw, we parameterized the air-tissue boundary of each speaker as a combination of independent tongue, lip, and jaw factors using a guided factor analysis [17].

First we found the jaw factor. We computed the first principal component $\mathbf{t}_1$ of $\mathbf{A}$ after setting to zero the columns of $\mathbf{A}$ corresponding to non-jaw polyline vertices. While $\mathbf{t}_1$ captured jaw movement, it did not capture jaw-related tongue or lip movement. We proceeded to compute $\mathbf{R} = \mathbf{A}^\intercal\mathbf{A}/q$ after setting to zero the columns of $\mathbf{A}$ corresponding to non-tongue, non-lip, and non-jaw polyline vertices. We scaled $\mathbf{t}_1$ to have unit variance as $\mathbf{h}_1 = \mathbf{t}_1/\sqrt{v}$ for $v = \mathbf{t}_1^\intercal\mathbf{R}\mathbf{t}_i$ [18]. The jaw factor is $\mathbf{u}_1 = (\mathbf{h}_1^\intercal\mathbf{R})^\intercal$, and it captures jaw movement as well as tongue and lip movement which is due to jaw [19].

We then subtracted the contribution of the jaw factor to obtain $\mathbf{A}' = \mathbf{A} - \mathbf{A}\mathbf{u}_1\mathbf{u}_1^\dagger$, where $\dagger$ denotes pseudoinverse, and identified four tongue factors as follows. We computed four principal component $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3, \mathbf{t}_4$ of $\mathbf{A}'$ after setting to zero the columns of $\mathbf{A}'$ corresponding to non-tongue polyline vertices. Tongue factors were iteratively derived from these principal components. The first tongue factor $\mathbf{u}_2 = (\mathbf{h}_2^\intercal\mathbf{R}_2)^\intercal$, where $\mathbf{R}_2 = \mathbf{A}'^\intercal\mathbf{A}'/q$, $\mathbf{h}_2 = \mathbf{t}_2/\sqrt{v_2}$, and $v_2 = \mathbf{t}_2^\intercal\mathbf{R}_2\mathbf{t}_2$. The variance due to $\mathbf{u}_2$ is factored out of $\mathbf{R}_3 = \mathbf{R}_2 - \mathbf{u}_2\mathbf{u}_2^\intercal$ before finding the next factor $\mathbf{u}_3$. We continued iteratively to find four tongue factors $\mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4, \mathbf{u}_5$ and then repeated the process to obtain two lip factors $\mathbf{u}_6, \mathbf{u}_7$. In this way we factored out the contribution of the jaw to tongue and lip movements and decomposed tongue postures into linearly independent tongue and jaw factors and lip postures into linearly independent lip and jaw factors. We also computed a velum factor $\mathbf{u}_8$. Figure 1 shows the factors of speaker F8 over the mean vocal tract shape of F8.

For each speaker, we approximated the air-tissue boundary in each frame n as the linear combination $\hat{\mathbf{a}}_n = (\mathbf{w}_n^\intercal\mathbf{U}^\dagger)^\intercal$,

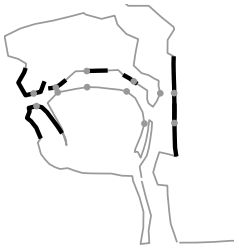
Figure 2: Places of constriction for subject F8

where the columns of $\mathbf{U}$ are $\mathbf{u}_1, \ldots, \mathbf{u}_{10}$, the weights $\mathbf{w}_n = \mathbf{U}^\intercal\mathbf{a}_n$ vary by frame n. We calculated the root mean square error (RMSE) of jaw, lips, tongue, epiglottis, pharynx, and velum air-tissue boundaries $\hat{\mathbf{a}}_n$ reconstructed from factor weights compared to the air-tissue boundaries $\mathbf{a}_n$ segmented directly from rtMRI video [16]. Average RMSE over the eighteen speakers was 1.7 ± 0.2 mm.

2.4. Constriction degree measurement

We used the algorithm of [20] to measure the minimal distance (constriction degree) between each of the six regions shown in Figure 2 and the opposing vocal tract surface. This involved manually annotating for each speaker the places of articulation as polylines of the air-tissue boundaries (lips, alveolar ridge, most superior part of the hard palate, soft palate, posterior velopharyngeal wall, and posterior pharyngeal wall). For each frame n, we computed constriction degrees $\mathbf{z}_n \in \mathbf{R}^6$ at each place as the minimum distance between two polylines.

2.5. Forward map

Having linearly decomposed the midsagittal air-tissue boundary into tongue, lip, and jaw factors and measured constriction degrees, we relate these by constructing a forward map from factors of vocal tract shape to constriction degrees. This relation quantifies how much a speaker uses their tongue, lips, and jaw to make constrictions.

Each rtMRI video has both a path $\mathbf{z}_n \in \mathbf{R}^6$ of constriction degree vectors and a path $\mathbf{w}_n \in \mathbf{R}^8$ of factor weight vectors sampled at each frame n. We developed an algorithm to estimate the forward map $\mathbf{g} : \mathbf{R}^8 \to \mathbf{R}^6$ which maps weight vectors to constriction degree vectors [21, 22]. For each speaker, the algorithm computes a tree whose root node is the set of all observed weight vectors in $\mathbf{R}^8$. A k-means subroutine starts at the root and iteratively breaks nodes in two (i.e., $k = 2$). Children in this tree are disjoint subsets of the parent and the union of siblings is the parent. Nodes stop breaking either when a child would contain fewer than seven weight vectors (to prevent rank-deficiency in least squares estimation of $\mathbf{g}$) or when $\mathbf{g}$ maps the weight vectors of that node to constriction degree vectors in $\mathbf{R}^6$ approximately linearly (i.e., when $\mathbf{g}(\mathbf{w})$ estimates $\mathbf{z}$ with RMSE less than 0.24 mm). Averaging over speakers, there were 142 ± 69 terminal nodes, 84% of which are volumes of $\mathbf{R}^8$ in which $\mathbf{g}$ was approximately linear, and 16% of which were too small to continue breaking. Each terminal node ℓ has a center $\mathbf{c}_\ell \in \mathcal{C}$, where $\mathcal{C} \subset \mathbf{R}^8$ is the set of centers in the factor weight space. Within terminal node ℓ, the algorithm uses least squares to estimate the matrix representation $\mathbf{G}_\ell$ of $\mathbf{g}$, the jacobian $\mathbf{J}_\ell$ of $\mathbf{g}$, and the time derivative $\dot{\mathbf{J}}_\ell$ of the jacobian. These are linear approximations to $\mathbf{g}$, $\mathbf{J}(\mathbf{w})$, and $\dot{\mathbf{J}}(\mathbf{w}, \dot{\mathbf{w}})$ in the neighborhood of $\mathbf{c}_\ell$.

For each observed constriction degree vector $\mathbf{z}_n$ we approximated the corresponding weight vector $\mathbf{w}_n$ as $\hat{\mathbf{w}}_n = \mathbf{G}_\ell^\dagger\mathbf{z}_n$

using the pseudoinverse $\mathbf{G}_\ell^\dagger$ of $\mathbf{g}$ whose center $\mathbf{c}_\ell$ is nearest to $\mathbf{w}_n$ in the weight space by the Euclidean distance metric and approximated the jaw, lips, tongue, epiglottis, pharynx, and velum air-tissue boundaries $\mathbf{a}_n$ as the linear combination $\hat{\mathbf{a}}_n = \mathbf{U}\mathbf{G}^\dagger\mathbf{z}_n = \mathbf{U}\hat{\mathbf{w}}_n$, where the columns of $\mathbf{U}$ are $\mathbf{u}_1, \ldots, \mathbf{u}_8$. For each speaker, we calculate RMSE compared to the air-tissue boundaries segmented directly from rtMRI video [16]. The average RMSE over speakers was 2.3 ± 0.5 mm.

2.6. Dynamical system simulation

The forward map quantifies how vocal tract shape relates to constrictions. We used the forward maps to parameterize speaker-specific task dynamical systems and ran simulations to quantify how much each speaker used the tongue, lips, and jaw to make constrictions of the vocal tract.

Following Task Dynamics [2], we describe change in the vector $\mathbf{z}$ of constriction degrees over time as

$$\ddot{\mathbf{z}} = -\mathbf{K}(\mathbf{z} - \mathbf{z}_0) - \mathbf{B}\dot{\mathbf{z}} \tag{1}$$

where $\mathbf{z}_0$ is a vector of six constriction degree targets and $\mathbf{K}$ and $\mathbf{B}$ are 6×6 diagonal matrices of stiffness and damping coefficients, respectively [2]. The matrices $\mathbf{K}$ and $\mathbf{B}$ are constant for a given constriction. When the constriction is for z_i, then stiffness $k_{ii} = \omega^2$ and damping $b_{ii} = 2\omega$ for natural frequency ω. The other entries of $\mathbf{K}$ and $\mathbf{B}$ are zero. We set $\omega = 25$ Hz arbitrarily. The relative contributions of the tongue, lips, and jaw are independent of time-scale. The constriction degree target z_{0i} depends on the constriction: $z_{01} = 0$ mm for bilabial closure, $z_{02} = 0$ mm for alveolar closure, $z_{03} = 4$ mm for palatal approximation, $z_{04} = 0$ mm for velar closure, $z_{05} = 7$ mm for pharyngeal approximation, and $z_{06} = 0$ for velopharyngeal closure.

Given model parameters $\mathbf{K}, \mathbf{B}, \mathbf{z}_0$ and initial conditions $\mathbf{z}(0), \dot{\mathbf{z}}(0)$, the solution $\mathbf{z}(t)$ to Equation 1 is unique for $0 \leq t \leq T$. This solution does not uniquely determine paths $\mathbf{w}(t)$ of factor weights [23]. We must further specify the model parameters $\mathbf{g}$, $\mathbf{J}(\mathbf{w})$, $\dot{\mathbf{J}}(\mathbf{w}, \dot{\mathbf{w}})$, which we estimated from rtMRI videos in Section 2.5, and initial conditions $\mathbf{w}(0), \dot{\mathbf{w}}(0)$ such that $\mathbf{g}(\mathbf{w}(0)) = \mathbf{z}(0)$ and $\mathbf{J}(\mathbf{w})\dot{\mathbf{w}}(0) = \dot{\mathbf{z}}$. Then the solution to the following equation is unique for $0 \leq t \leq T$.

$$\ddot{\mathbf{w}} = \mathbf{J}^\dagger(\mathbf{w})(-\mathbf{B}\mathbf{J}(\mathbf{w})\dot{\mathbf{w}} - \mathbf{K}(\mathbf{g}(\mathbf{w}) - \mathbf{z}_0)) - \mathbf{J}^\dagger(\mathbf{w})\dot{\mathbf{J}}(\mathbf{w}, \dot{\mathbf{w}})\mathbf{w} \tag{2}$$

This follows from the change of variables $\mathbf{z} = \mathbf{g}(\mathbf{w})$, $\dot{\mathbf{z}} = \mathbf{J}(\mathbf{w})\dot{\mathbf{w}}$, $\ddot{\mathbf{z}} = \mathbf{J}(\mathbf{w})\ddot{\mathbf{w}} + \dot{\mathbf{J}}(\mathbf{w}, \dot{\mathbf{w}})\dot{\mathbf{w}}$ and from the pseudoinverse $\mathbf{J}^\dagger(\mathbf{w})$ of $\mathbf{J}(\mathbf{w})$ [2].

We establish a grid of 18 speakers $\times$ 6 constrictions $\times$ 20 initial conditions. Varying initial conditions gives the system response over a range of initial vocal tract postures. For each of the 2160 vertices on the grid, we numerically approximate the corresponding solution to Equation 2. We now describe the construction of this grid.

For each speaker and for each of the six constrictions (i.e., for each entry z_i of $\mathbf{z}$), we establish a grid of 20 initial conditions as follows. We take four percentile values of z_i, starting at the 30-percentile and ending at the 90-percentile of z_i observed for that speaker. As these initial conditions for constriction degree z_i do not uniquely determine initial conditions for factor weights, we choose five initial conditions which best satisfy the equation $\mathbf{w}(0) = \operatorname{argmin}_{\mathbf{c}_\ell \in \mathcal{C} \subset \mathbf{R}^8} \sqrt{(g_i(\mathbf{c}_\ell) - z_{0i})^2}$,where $g_i(\mathbf{c}_\ell)$ is entry i of $\mathbf{g}(\mathbf{c}_\ell)$. The result is a grid of 18 speakers $\times$ 6 constrictions $\times$ 13 constriction degree initial conditions $\times$ 5 factor weight initial conditions for 2160 total grid vertices.

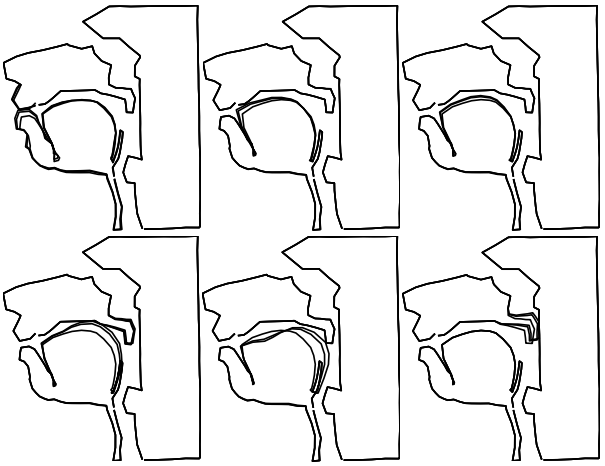

Figure 3: Time-lapses of the air-tissue boundaries during bilabial closure, alveolar closure, palatal approximation, velar closure, pharyngeal approximation, and velopharyngeal closure

At each grid vertex we solve Equation 2 by linearizing the equation over a grid of time points from 0 to $T = 0.2$ (step-size $h_1 = 0.02$ s). For $(k-1)h \leq t \leq kh$ we have a forward map $\mathbf{G}_\ell$, jacobian $\mathbf{J}_\ell$, and time-derivative $\dot{\mathbf{J}}_\ell$ of the jacobian from the terminal node ℓ whose center $\mathbf{c}_\ell$ is nearest to $\mathbf{w}((k-1)h)$ in the weight space by the Euclidean distance metric. We then have the linear equation

$$\ddot{\mathbf{w}} = \mathbf{J}_\ell^\dagger(-\mathbf{B}\mathbf{J}_\ell\dot{\mathbf{w}} - \mathbf{K}(\mathbf{G}_\ell\mathbf{w} - \mathbf{z}_0)) - \mathbf{J}_\ell^\dagger\dot{\mathbf{J}}_\ell\dot{\mathbf{w}} \tag{3}$$

whose solution $\mathbf{w}(t)$, $(k-1)h \leq t \leq kh$ we approximate numerically at a grid of time points (step-size $h_2 = 0.002$ s).

The panels of Figure 3 show time-lapses of the air-tissue boundaries at 0 s, 0.1 s, and 0.2 s for six solutions to Equation 2 where the system makes constrictions for each of the constriction degrees $z_1, z_2, \ldots, z_6$.

We now turn to summarize the relative contributions of active articulators to a vocal tract constriction relative to the total contribution of all articulators. That is, we seek to summarize the contribution of an articulator to the elapsed change in z_i over the course of a vocal tract constriction as a proportion of the total elapsed change in z_i.

By change of variables, we have the system $\dot{\mathbf{z}} = \mathbf{J}(\mathbf{w})\dot{\mathbf{w}}$ of 6 equations in the 8 factor weights. We integrate equation i for $0 \leq t \leq T$ to obtain the equation for elapsed change in constriction degree z_i.

$$\int_0^T \dot{z}_i \, \mathrm{d}t = \int_0^T \mathbf{J}_i(\mathbf{w})\dot{\mathbf{w}} \, \mathrm{d}t \tag{4}$$

$\mathbf{J}_i(\mathbf{w})$ is row i of $\mathbf{J}(\mathbf{w})$. The integral can be broken down into the contribution of each weight w_k as the sum

$$\int_0^T \dot{z}_i \, \mathrm{d}t = \sum_{k=1}^{8} \int_0^T \mathbf{J}_i(\mathbf{w}_n)\mathbf{P}_k\dot{\mathbf{w}}_n \, \mathrm{d}t \tag{5}$$

where $\mathbf{P}_k$ is the 8×8 diagonal projection matrix whose entry $p_{kk} = 1$ and all other entries equal zero. Term k of the outer summation is the theoretical contribution of factor k to elapsed change in z_i. We approximate the integral numerically as

$$\sum_{k=1}^{8} \int_0^T \mathbf{J}_i(\mathbf{w}_n)\mathbf{P}_k\dot{\mathbf{w}}_n \, \mathrm{d}t \sim h_2 \sum_{k=1}^{8} \sum_{n=0}^{N} \mathbf{J}_i(\mathbf{w}_n)\mathbf{P}_k\dot{\mathbf{w}}_n \tag{6}$$

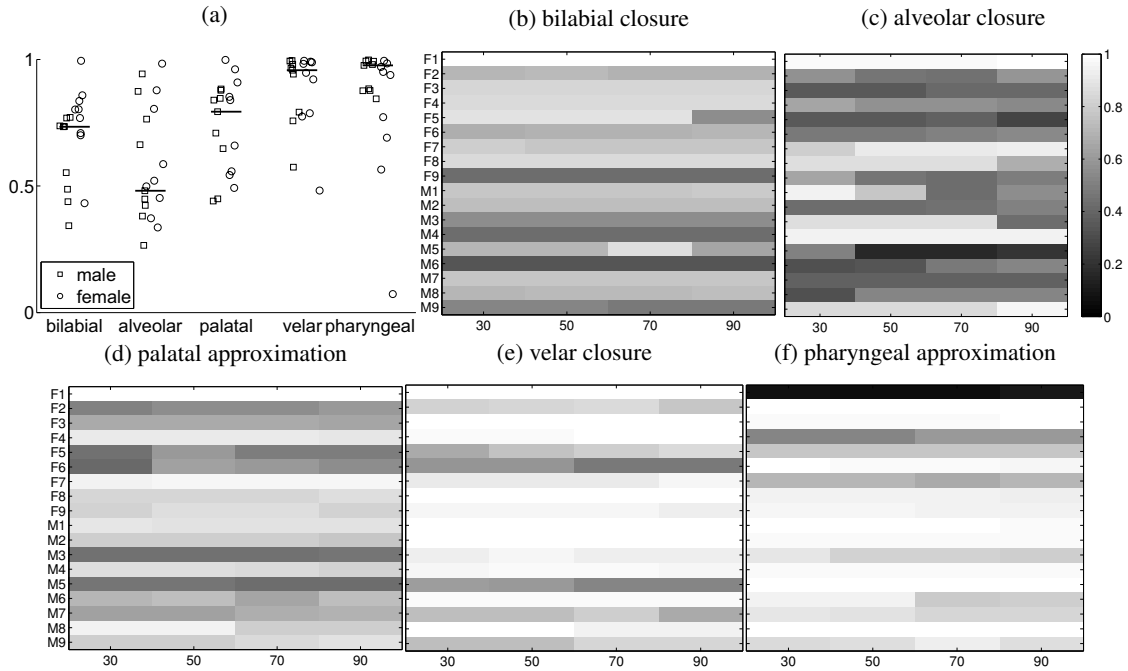

Figure 4: **a.** Ratio of lip and tongue contributions to total change in constriction degree (y-axis) by constriction (x-axis) and by speaker (points), averaging over initial conditions; line is median. **b-f.** Ratio of lip and tongue contributions to total change in constriction degree (cells each average five initial postures) by initial constriction degree (30 to 90 percentile along x-axis) and by speaker (y-axis).

where $N = T/h_2 = 100$ is the total number of time steps over the interval $[0, T]$. The proportion of elapsed change in constriction degree z_i due to an articulator with factors $k \in \mathcal{U}$ relative to total elapsed change in z_i is

$$\frac{\sum_{k\in\mathcal{U}} \sum_{n=0}^{N} \mathbf{J}_i(\mathbf{w}_n)\mathbf{P}_k\dot{\mathbf{w}}_n}{\sum_{k=1}^{8} \sum_{n=0}^{N} \mathbf{J}_i(\mathbf{w}_n)\mathbf{P}_k\dot{\mathbf{w}}_n} \tag{7}$$

For example, setting $i = 2$ and $\mathcal{U} = \{2, 3, 4, 5\}$ gives the proportion of elapsed change in alveolar constriction degree due to the tongue (i.e., due to factors $\mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4, \mathbf{u}_5$).

3. Results

Figure 4 shows how much each speaker used their lips or tongue compared to their jaw in making vocal tract constrictions, starting from a range of initial vocal tract postures. Panel 4a shows that the jaw contributed more on average to constrictions of the anterior vocal tract than to constrictions of the posterior vocal tract. Speakers nevertheless varied substantially in how much they used their tongue, lips, and jaw. For example, speakers F1, F3, F4, F5, F8, and M1 made much greater use of their lips than their jaw during bilabial closure, and so their rows are light-colored in Panel 4b. In contrast, speakers F9, M4, M6, and M9 made greater use of their jaw than their lips and so their rows are dark-colored in Panel 4b.

For the bilabial closure task, the lips contributed more than the jaw in 81% of the simulations. On average, the lips contributed more than the jaw for 14 of the 18 speakers.

For the alveolar closure task, the tongue contributed more than the jaw in 57% of the simulations. On average, the tongue contributed more than the jaw for 11 of the 18 speakers.

For the palatal approximation task, the tongue contributed more than the jaw in 83% of the simulations. On average, the tongue contributed more than the jaw for 16 of the 18 speakers.

For the velar closure task, the tongue contributed more than the jaw in 97% of the simulations. On average, the tongue contributed more than the jaw for all 18 speakers.

For the pharyngeal closure task, the tongue contributed more than the jaw in 94% of the simulations. On average, the tongue contributed more than the jaw for 17 of the 18 speakers.

4. Discussion

We have presented a technique for extracting from rtMRI videos information about the articulatory strategies adopted by individuals, and in particular about how much each subject uses their tongue, lips, and jaw to make vocal tract constrictions. How much each speaker used their tongue or lips compared to their jaw differed by constriction type. Specifically, the jaw tended to contribute more to constrictions of the anterior vocal tract than to constrictions of the posterior vocal tract, which indicates functional specificity of articulatory organization. How much each speaker used their tongue or lips compared to their jaw varied by initial vocal tract posture, especially for alveolar closure.

This is the first task dynamical system explicitly derived from speech kinematic data. The system can be used to quantitatively validate model structure and parameter values through error analysis. With ongoing large-scale acquisition of speech MRI data at our site, an extension of the proposed modeling methodology is to develop a scalable pipeline for summarizing individual articulatory function in large speech MRI datasets with the goal of relating articulatory function to anatomical structure.

5. Acknowledgements

Work supported by NIH grant R01DC007124 and NSF grant 1514544.

6. References

[1] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *Journal of the acoustical society of America*, vol. 115, no. 4, pp. 1771–1776, 2004.

[2] 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.

[3] J. Simko and F. Cummins, "Embodied task dynamics." *Psychological review*, vol. 117, no. 4, p. 1229, 2010.

[4] J. S. Kelso, B. Tuller, E. Vatikiotis-Bateson, and C. A. Fowler, "Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 10, no. 6, p. 812, 1984.

[5] M. Weirich, A. Simpson, S. Fuchs, R. Winkler, and P. Perrier, "Mumbling is morphology?" in *10th International Seminar on Speech Production (ISSP 2014)*. Köln Universität, 2014, pp. 457–460.

[6] A. Lammert, M. Proctor, and S. Narayanan, "Interspeaker variability in hard palate morphology and vowel production," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 6, pp. S1924–S1933, 2013.

[7] ——, "Morphological variation in the adult hard palate and posterior pharyngeal wall," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 2, pp. 521–530, 2013.

[8] G. Fairbanks, *Voice and Articulation Drillbook*. New York: Harper & Row, 1960.

[9] A. E. Aronson and J. R. Brown, *Motor speech disorders*. Philadelphia: WB Saunders Company, 1975.

[10] International Phonetic Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[11] J. Garofolo, L. F. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "The darpa timit acoustic-phonetic continuous speech corpus, nist order number pb91-100354," 1993.

[12] E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 123–132, 2008.

[13] Y.-C. Kim, S. S. Narayanan, and K. S. Nayak, "Flexible retrospective selection of temporal resolution in real-time speech mri using a golden-ratio spiral view order," *Magnetic resonance in medicine*, vol. 65, no. 5, pp. 1365–1371, 2011.

[14] J. M. Santos, G. A. Wrigh, and J. M. Pauly, "Flexible real-time magnetic resonance imaging framework," in *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, vol. 1. IEEE, 2004, pp. 1048–1051.

[15] J. I. Jackson, C. H. Meyer, D. G. Nishimura, and A. Macovski, "Selection of a convolution function for fourier inversion using gridding [computerised tomography application]," *Medical Imaging, IEEE Transactions on*, vol. 10, no. 3, pp. 473–478, 1991.

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

[17] A. Toutios and S. S. Narayanan, "Factor analysis of vocal-tract outlines derived from real-time magnetic resonance imaging data," in *18th International Congress of Phonetic Sciences (ICPhS)*, 2015.

[18] J. E. Overall, "Orthogonal factors and uncorrelated factor scores," *Psychological Reports*, vol. 10, no. 3, pp. 651–662, 1962.

[19] J. Cai, Y. Laprie, J. Busset, and F. Hirsch, "Articulatory modeling based on semi-polar coordinates and guided pca technique," in *10th Annual Conference of the International Speech Communication Association-INTERSPEECH 2009*, 2009.

[20] V. Ramanarayanan, L. Goldstein, D. Byrd, and S. S. Narayanan, "An investigation of articulatory setting using real-time magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 510–519, 2013.

[21] A. Lammert, L. Goldstein, S. Narayanan, and K. Iskarous, "Statistical methods for estimation of direct and differential kinematics of the vocal tract," *Speech communication*, vol. 55, no. 1, pp. 147–161, 2013.

[22] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 444–460, 2005.

[23] N. A. Bernstein, *The co-ordination and regulation of movements*. Pergamon Press Ltd., 1967.