**Response to reviewers**

Associate Editor’s comments:

* **page 16, equation 13. what does the plus sign (+) mean?**

The plus sign denotes the Moore-Penrose pseudoinverse. This was mentioned on page 14. The revised manuscript adds the sentence “Superscript ``+’’ denotes the Moore-Penrose pseudoinverse”**.**

* **line 429, 'velar and pharyngeal', do you mean 'pharyngeal'?**

The biomarker values at the bilabial and palatal places did not significantly differ from those at the velar place (p=0.19, p=0.061, see lines 408 and 410) or from those at the pharyngeal place (p=0.96, p=0.5, see lines 412 and 414). The statement on line 429 of the original manuscript was correct. It is not amended in the revised manuscript.

* **Figure 5. I might have missed it, but figure 5 does not seem to be referred to in the text.**

The revised manuscript now uses Figure 5 to explain Equation 26.

* **I noticed that no separate files were submitted for the figures included in the manuscript. If this is the case, please include such figure files in your resubmission.**

The original manuscript included each figure as a separate file. These files were resubmitted with the revised manuscript.

* **Also, please make sure the supplementary materials are truly necessary.**

The revised manuscript removed Supplementary Material #2, which gave the results of the cross-validation for different factor analysis parameterizations, Supplementary Material #3, which estimated precision for different neighborhood sizes in the forward kinematic map, and Supplementary Material #4, which made supplemental graphs for the results of the sensitivity analysis (which remains described in the text). These materials would likely be of limited interest. The study retains Supplementary Materials #1, #5, and #6.

Reviewer #1’s comments:

Links to the theory:

* **The definition of articulator synergy as “a functional grouping of articulators such as the jaw, tongue, and lips whose coordinated movements produce constrictions during speech” is poorly motivated by concurrent definitions.**

We agree that the definition should mention that a reduction of degrees of freedom plays a central role in the definition of a synergy. The definition on page 3, lines 16-17 has been revised. It now reads as follows: “An articulator synergy is a functional grouping of articulators such as the jaw, tongue, and lips whose coordinated movements produce constrictions during speech, and which instantiates a reduction in the number of independent degrees of freedom for controlling a vocal tract movement.”

* **TODO: The proposed model of the speech production system underlying the imaged movements involves a reduction in dimension in two mappings: (a) the mapping from articulator shapes to articulator parameters and (b) the mapping from articulator parameters to constriction degrees. Do these two steps reflect two different levels of redundancy reduction in the motor system? If so, how are these two levels connected to available definitions of articulator synergies?**

The articulator synergy biomarker involves are two forward mappings: the mapping from the spatial coordinates of the articulator contour vertices in the scan plane to the factor scores (i.e., parameters of an articulatory model); and the mapping from the factor scores to constriction task variables at the phonetic places of articulation. Both involve a reduction of dimensionality. The reduction of dimensionality from contour vertices to factor scores provides a compact representation of the data-set. This reduction of dimensionality is not scientifically important per se, but the articulatory model is scientifically meaningful for defining the jaw, tongue, and lip degrees of freedom, and the relationships between them. The reduction of dimensionality from factor scores to constriction task variables (i.e., the reduction in dimensionality performed by the forward kinematic map) is scientifically meaningful. It is a reduction of dimensionality from the articulator degrees of freedom to the space of variables whose values indicate the production of vocal tract constrictions. This allows the forward kinematic map to characterize how a multidimensional set of jaw, tongue, and lip factors influences the evolution of a constriction task variable in time.

The present study showed that the forward kinematic map is a function that can be estimated from MR images. It has not directly addressed the internal forward model, which is the brain’s internal model of the forward kinematic map. As demonstrated by Hogan et al. (2014) in a study using synthetic data, analyses of The Uncontrolled Manifold approach (Scholtz & Schoener, 1999) could be used to test this hypothesis using behavioral data.

As demonstrated by Hogan et al. (2014), the Uncontrolled Manifold approach requires that the coordinate system for the kinematic data be the same as that which the nervous system uses to plan movements. The present study did not provide evidence that the coordinate systems implied by the factors or the constriction task variables is the coordinate system used by the nervous system to plan movements, although the latter is argued for directly in studies of Articulatory Phonology (Browman & Goldstein). The result of a test of the uncontrolled manifold hypothesis is indeterminate without a priori knowledge of the coordinate system used by the nervous system to plan movements. Second, the method requires that the brain controls the constriction task variables. The present study did not test this hypothesis, other studies in Articulatory Phonology directly argue for this point.

The present study used constriction task variables as indicators of motor performance for vocal tract constrictions. It may be

* **The definition should relate quantities obtained from real-time MRI to parameters of the task dynamics model of speech production.**

The revised manuscript now includes a graph of the articulator synergy biomarker as a function of the jaw weight parameter from Task Dynamics (Fig. 6, panel a).

Links to top-down and bottom-up approaches to the study of motor synergies:

* **TODO: How is the proposed biomarker related to the uncontrolled manifold approach?**

The revised manuscript includes a paragraph in the discussion describing the relation to the uncontrolled manifold. (TODO: cite my answer to the question two questions up, when it is in the discussion section)

* **TODO: How is the proposed biomarker related to bottom-up methods of describing synergies?**

The revised manuscript includes a paragraph in the discussion describing the relation to the uncontrolled manifold. (TODO: cite my answer to the question three questions up, when it is in the discussion section)

Validation of the articulator synergy biomarker:

* **The method should be validated by analysis of simulated data. Alternatively, the method could be validated by analysis of real data on the basis of finding patterns that are expected because of available theoretical knowledge of mechanisms.**

Section VI of the revised manuscript estimates the measurement bias of the biomarker using synthetic data.

Potential limits of the study:

* **If the components obtained through guided PCA depend on the analyzed speech material, why is there such a similarity to the components reported in Toutios & Narayanan (2015), based on one speaker from the USC-TIMIT database?**

We have not quantified whether factors differ depending on speech material, and so we cannot confirm that there is similarity between the guided factor analysis results of Toutios & Narayanan (2015) and the guided factor analysis results of the present paper. Provided that the components are extracted from a sufficiently diverse dataset, we believe that the factors should reflect speaker anatomy and efficiencies rather than the speech material. The fact that the limited speech material used in the present study produces factors similar to the much larger set in the 2015 paper supports this view.

* **How does the assumption of linearity in the guided PCA approach affect the results?**

An appropriately optimized non-linear model for compactly representing the vocal tract shapes might do so with a smaller number of components than the linear model of the present study. We used the linear model because of its interpretability in terms of distinct speech articulators.

In order to quantify how the linearity assumption impacts the results, the revised manuscript quantifies the percent variance explained by each factor as a measure of model performance (cf. Fig. 4).

* **What happens when the spatio-temporal pattern of articulator parameters or task variables is not similar for different tokens of the same gesture? Should tokens be sorted into different groups on the basis of their spatio-temporal patterns and be submitted separately to the analysis?**

We respond to this question first by offering an assessment of how substantial the variability between different tokens of the same gesture may be in terms of spatio-temporal patterns.

Variability in the spatio-temporal pattern of articulator parameters is substantial in the sense that vocal tract shaping for a given gesture may be achieved using different factors depending on the participant. Since the factor model is estimated separately for each speaker, the factors of different speakers may have distinct spatial patterns, and the correspondence between factors of different speakers is not known a priori (e.g., the first tongue factor of Speaker A does not necessarily reflect the same deformation of the tongue as the first tongue factor of Speaker B). Intra-speaker variability is also to be expected, and is reflected at different steps of the modeling as distinct factor scores and distinct articulator synergy biomarker values.

Variability in the spatio-temporal pattern of s for different tokens of the same gesture is minor in the sense that the same basic pattern is observed for all tokens of a particular gesture. Specifically, the associated with the consonant closure decreases to a minimum at the point of maximum constriction and then increases again. Subtle variations in this spatio-temporal pattern may be present (e.g., peak velocity, time to peak velocity, etc.), but other techniques with finer temporal and spatial resolution are perhaps better suited to study fine details of the time-course of s (e.g., electromagnetic articulography).

Whether or not it makes sense to group observations into different spatio-temporal patterns and to submit these patterns to the analysis separately depends on which stage of the analysis you are referring to. With regard to the analysis of biomarker precision, one could subset the data according to spatio-temporal pattern instead of by vowel-consonant-vowel sequence type, but different spatio-temporal patterns may contain different numbers of observations from each participant. As this would complicate the analysis, we did not pursue this approach. With regard to the cross-validation of the direct and differential kinematics, the data was divided into subsets by place of articulation. We believe that this is the most intuitive way to subset the data because the error being analyzed was the error in at exactly these places of articulation. Without a sound reason for believing that different spatio-temporal patterns would involve errors of different magnitude, we did not pursue this approach. With regard to testing the main hypothesis (i.e., task-dependence of articulator synergies), our main interest was in determining differences between constrictions of the anterior vocal tract and constrictions of the posterior vocal tract. We subset the data by place of articulation (in addition to scan number and participant) because this straightforwardly reflected the anterior vs. posterior distinction, with bilabial being most anterior and pharyngeal being most posterior. As further subdivisions of the data on the basis of spatio-temporal patterning do not reflect this distinction, we have not pursued it here.

Although we do not believe it makes sense to subset the data on the basis of spatio-temporal patterns in the context of the present study, we believe that future studies might benefit from using the proposed articulator synergy biomarker to distinguish different spatio-temporal patterns.

Presentation of the method:

* **A tutorial style is adopted for some sections, whereas other sections use a terse, descriptive style. More explanation should be given where appropriate (e.g., Eq. 20).**

In the revised manuscript, Sections V.A. and VII have been edited to adopt a more tutorial style.

Textual comments:

* **p.3, lines 16-17: "An articulator synergy is a functional grouping of articulators such as the jaw, tongue, and lips whose coordinated movements produce constrictions during speech (Turvey, 1977)." Motor synergies can be defined in different ways (see main comment I).**

The definition has been revised to the following: “An articulator synergy is a functional grouping of articulators such as the jaw, tongue, and lips whose coordinated movements produce constrictions during speech, and which instantiates a reduction in the effective degrees of freedom from the set of articulators to the space of controlled variables.”

* **p. 3, lines 24 -26: "healthy adult speakers of American English may use the jaw more for anterior constrictions at the bilabial, alveolar, and palatal places of articulation than for posterior constrictions at the velar and pharyngeal places of articulation". Please provide some explanation for this hypothesis.**

This is the result from the cited study that motivates characterizing articulator synergies.

* **p.4 lines 49-51: "The algorithm for computing the articulator synergy biomarker involves a computational model of the direct and differential kinematics of the vocal tract (Lammert et al., 2013a) based on the Task Dynamics model of speech production (Saltzman and Munhall, 1989)" I was a bit puzzled by this passage. As I understood Lammert et al. (2013), the authors of that paper presented a weighted linear regression model (as opposed to a model based on multilayer perceptron) permitting to predict vocal tract constriction degree from articulator positions. If this is the case, wouldn't be more appropriate referring to a "statistical model" rather than to a "computational model"? Moreover, the authors should be more specific about the features shared by the task dynamic model and the weighted linear regression model by Lammert et al (2013). Are the authors referring to the inventories of geometric articulator variables and task variables?**

In the revised manuscript, the quoted passage has been changed to the following: "The algorithm for computing the articulator synergy biomarker involves a statistical model of the direct and differential kinematics of the vocal tract (Lammert et al., 2013a), which uses MRI to estimate a forward kinematic map relating articulatory parameters and s, as in the Task Dynamics model of speech production (Saltzman and Munhall, 1989)”. This change uses the term “statistical model” instead of “computational model” and clarifies that the statistical model of Lammert et al. (2013) is a model of the forward kinematic map and its jacobian matrix from the Task Dynamics model of speech production. Additionally, the revised manuscript has replaced all other instances of the term “computational model” with “statistical model”.

* **p.6 lines 66-75: Is it really necessary to summarize the results in the introduction?** According to the journal guidelines (https://asa.scitation.org/pb-assets/files/publications/jas/jasinfcon-1518211773377.pdf), it is not necessary. Accordingly, we removed the summary of results from the introduction.
* **p.6 line 85. The plan of the paper does not include the discussion and the conclusion sections. Moreover, it makes reference to modelling steps which have not been mentioned before. I strongly suggest to give a rapid overview of the method in the introduction in which the authors presents the pre-processing step (the manually assisted segmentation of the FMRI images) and the two main steps of the algorithm (the guided PCA analysis and the mapping between articulator parameters and the constriction degree).**

The plan of the paper now only refers to modelling steps that have been mentioned previously in the introduction. It reads as follows: “Section II describes the MRI experiment, scanner sequence, participant characteristics, and method for manually annotating the start and end time-points in the real-time MRI. Sections III and IV describe the segmentation of articulator contours in the images and use the segmentation results to estimate of s and parameters of articulator shape and position, which are related by the direct and differential kinematics of the vocal tract. Section V estimates the direct and differential kinematics and evaluates the model through cross-validation. Section VI defines the articulator synergy biomarker and evaluates the precision of the articulator synergy biomarker in a test-retest experiment, which estimated the repeatability of the articulator synergy biomarker in two experiments on the same day under the same experimental conditions. Section VII uses the articulator synergy biomarker to test the task-dependence of articulator synergies. Sections VIII and IX offer discussion and conclusions.”

* **p. 10, line 148, "p is the number of contour vertices": If each vertex is a point on a plane, assuming that X is the vector of the stacked coordinates of the contour vertex, shouldn't p be the number of vertices divided by 2?**

The size of the matrix X is now written as n x 2p, where p is the number of contour vertices.

* **p.13, line 286: Was centering applied on the different dimension of X vectors before PCA?**

Yes. The revised manuscript states that the contour vertex coordinates were centered on zero before factor analysis.

* **p.16: Please add a multi-panel figure showing in each panel the trajectories of the q factor scores during the production of one different syllable. Each panel should also contain the trajectory of the relevant constriction degree. Such a figure will also be also useful when introducing the construction of the mapping between the factor scores and the constriction degree.**

Since the paper does not differentiate the various factors of a given articulator from one another (e.g., distinguishing the contribution of the first and second tongue factors), we believe that it may be confusing to the reader to distinguish them graphically. For this reason, we have proposed a revision to the existing Fig. 5. The revised manuscript adds the palatal to Fig. 5 in order to illustrate the connection between jaw and tongue factor contributions and the s. In particular, this illustrates that the sum of jaw and tongue contributions equals the .

* **p.16, line 239: In order to introduce the construction of the forward map, please consider presenting it as a weighted linear regression problem as in Lammert et al. (2013).**

Section V indicates that the problem is solved using weighted least squares. The estimator is the weighted linear regression estimator (Equation 20), which optimizes the weighted sum of squared errors (Equation 17).

* **p.17, eq. 18: In Lammert et al. (2013) a regularization term was used. What motivates dropping that term?**

In Lammert et al. (2013), the regularization term was used to make the solution robust to the case where there are very few data points in the neighborhood of the query point. In our implementation, the number of data points in the neighborhood of the query point is set by the parameter (see answer to next question for details), and thus there are always a large number of data-points (specifically, data-points) in the neighborhood. Since the problematic case of few data-points does not occur, the proposed estimator does not use a regularization term.

* **p.19, line 245: Are h and f two independent free parameters or is h adjusted to obtain the desired number of neighbours? And by the way, is h constant over different dimensions?**

The revised manuscript clarifies that f is the free parameter, not h by adding the following text: “The parameter sets the percentage of data-points inside the spherical neighborhood within which the forward kinematic map is estimated. The parameter is the radius of the smallest sphere containing exactly data points. The parameter is found using the nearest neighbors algorithm, where .”

* **p.32, lines 454-456, "The study demonstrated low error in the estimator of the direct and differential kinematics and consistent estimation of the articulator synergy biomarker in a test-retest repeatability experiment.": If I am not wrong, the median values of the consistency index for the synergy biomarker were more often below 50% than above this percentage.**

The intra-class correlation coefficient measures the ratio of inter-subject variance to total (inter-subject plus intra-subject) variance. Thus, the reason for low values of the intra-class correlation coefficient may either be low inter-subject variability, high intra-subject variability, or both. Given that the voluntary movement of speech production can be quite variable even within a given subject, low intra-class correlation coefficient may be inevitable. The revised manuscript includes an explanation of this in the discussion, Section VIII.B.

* **p.33, lines 460-462: I am not convinced that the results reported imply that the biomarker has adequate precision to characterize articulator synergies. See main comment III.**

Section VI of the revised manuscript estimates the measurement bias of the biomarker using synthetic data. No measurement bias was detected. The low precision may be an inevitable consequence of high intra-subject variability (see above response).

* **p.33, lines 472-474: This is a cut and paste from p.4 lines 49-51. The same comment applies here.** The revised manuscript changes the text to “The algorithm for computing the articulator synergy biomarker involves a statistical model of the direct and differential kinematics of the vocal tract (Lammert et al., 2013a), which uses MRI to estimate a forward kinematic map relating articulatory parameters and s, as in the Task Dynamics model of speech production (Saltzman and Munhall, 1989).”
* **p.35, lines 521-523: "The implication of the results is that articulator synergies are task-dependent in that they have different patterns of inter-articulator coordination depending on their place of articulation". Here I expected some comparison with other studies based on other approaches and supporting the idea that speech motor synergies are task dependent.**

The revised manuscript now refers to similar results obtained through studies using different experiments and analyses: “This complements findings that articulator synergies are task-dependent in terms of inter-articulator coupling (Lancia & Rosenbaum, 2018) and in terms of response to mechanical perturbation of articulator positions (Kelso et al., 1984).”

* **p.36, lines 534-535: Not all scripts are available as supplemental material. Strictly speaking, in the supplemental material there are wrappers that launch the core scripts residing on a server.** In the original submission, some scripts were obtained by cloning the Git repository “span\_contour\_processing” (https://github.com/usc-sail/span\_contour\_processing). This is no longer the case. All scripts are included in the main repository. In both the original and current submissions, all scripts were/are executed on the machine from which this repository is cloned (i.e., not on another server).

Reviewer #2’s comments:

* **The corpus … is strongly limited to the four isolated vowel-consonant-vowel utterances [apa], [ata], [aka], [aja]. This significantly hinders the validity of the conclusions, and should be clearly mentioned in the concluding section as well as in the abstract.**

The abstract now includes the modified sentence, “The test was conducted on utterances [apa], [ata], [aja], and [aka] with a real-time magnetic resonance imaging biomarker that was computed using a statistical model of the direct and differential kinematics of the vocal tract.” The conclusion now includes the modified sentence, “The analysis showed that articulator synergies vary for the following gestures in the intervocalic context [a]-consonant-[a]: bilabial closure, alveolar closure, palatal approximation, velar closure, and pharyngeal approximation and also vary by speaker.” The conclusion also suggests that future studies investigate a wider range of articulations (namely, different manner classes).

* **Section IV.A does not sufficiently explain the differences between Guided Factor Analysis and Guided PCA (Principal Component Analysis), and why Guided Factor Analysis has been chosen. Indeed, Guided PCA has been abundantly used in the literature for articulatory modelling … It is thus important to quote some of this literature and to emphasize how the present approach is different.**

The terms “Guided Factor Analysis” and “Guided PCA” have both been used to describe Maeda's pioneering work in articulatory modeling. We make no distinction between the two terms. The first paragraph of Section IV.A. now indicates that the origin of the technique lies with Maeda (1990):

* + Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech production and speech modelling* (pp. 131-149). Springer, Dordrecht.
* **In fact, it would seem that these principal axes do capture the variance of tongue and lips that is linearly correlated with the jaw factors, as described in the literature. Could the authors clarify this issue?**

We agree that this was poorly worded. The revised text is simplified to convey only the essential point.

* **Note also that the upper lip might be influenced by jaw in some subjects: this is apparently not taken into account (e.g. Fig. 3, top left), though of minor acoustic importance.**

We acknowledge that the upper lip may be influenced by the jaw. Additionally, as the researchers below (among others) have observed, the velum is lower for open vowels like [a].

* + Fritzell, B. (1969). The velopharyngeal muscles in speech: An electromyographic and cineradiographic study. Acta Oto-laryngologica. Supplement 250.
  + Bell-Berti, F. (1980). Velopharyngeal function: A spatial-temporal model. Speech and language: Advances in basic research and practice, 4, 137-150.

It may be that the velum is influenced by the jaw too. We consider it an open research question how much of the jaw-upper lip and jaw-velum correlations are actually biomechanical and how much are due to motor control. However, it would be somewhat beyond the scope of the present paper to investigate this explicitly, and we would prefer to leave this for future research, which is possible with the proposed method.

* **Finally, it would be very useful to conclude section IV with a comparison of the performances of the Guided Factor Analysis used and more standard Guided PCA.**

The terms “Guided Factor Analysis” and “Guided PCA” have both been used to describe Maeda's pioneering work in articulatory modeling. We make no distinction between the two terms. The first paragraph of Section IV.A. now indicates that the origin of the technique lies with Maeda (1990):

* + Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech production and speech modelling* (pp. 131-149). Springer, Dordrecht.
* **It would be useful to mention more explicitly in section IV.A which organs are really modelled (jaw, tongue, velum, lips?).**

The revised manuscript mentions this explicitly with the following sentence: “The objective of the guided factor analysis was to parameterize the vocal tract contours ... as the linear combination of factors ... such that each factor characterizes spatial variation in the position and shape of an articulator (specifically, the jaw, tongue, lips).”

* **Are the three degrees of freedom expected for a solid object in 2D enough to represent the jaw variance?**

The revised manuscript quantifies the percent variance explained by each jaw factor (cf. Fig. 4, panel a).

* **In lines 52-53, the authors state that “direct kinematics relates the position and shape of articulators to the corresponding degree of constriction”. This definition clearly indicates that these maps are not kinematic, i.e. not related to time, but only represent relations between two partial representations of the geometry of articulators (articulator contours and construction areas). This expression is thus very confusing and should be avoided. Similarly, the authors state that “differential kinematics relates small increments of articulator movement to the resulting changes in the constriction degrees”. Here too, time is not involved either, which is also confusing.**

To the best of our knowledge, the terms “forward kinematics” and “differential kinematics” originated in the field of robotics. The terms were subsequently adopted by the field of computational motor control. A definition of forward kinematics from a recent textbook on motor control reads as follows: “The kinematics maps relate the motions of the hand to the motions of the arm and come in two forms: direct kinematics, from arm’s joint angles to hand position, and inverse kinematics, from hand position to arm’s joint angles.”

* Shadmehr, Reza, and Sandro Mussa-Ivaldi. Biological Learning and Control: How the Brain Builds Representations, Predicts Events, and Makes Decisions, edited by Tomaso A. Poggio, and Terrence J. Sejnowski, MIT Press, 2012. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/socal/detail.action?docID=3339375. Created from socal on 2018-09-23 12:52:04.

The use of the definition is not restricted to this one book, but appears frequently in academic papers on motor control. I sample two such papers below and indicate where in the paper these terms are defined.

* Todorov, E., Li, W., & Pan, X. (2005). From task parameters to motor synergies: A hierarchical framework for approximately optimal control of redundant manipulators. Journal of robotic systems, 22(11), 691-710. DOI: 10.1002/rob.20093 (see mathematical definition in Section 4.2).
* Atkeson, C. G. (1989). Learning arm kinematics and dynamics. Annual review of neuroscience, 12(1), 157-183. DOI: ﻿10.1146/annurev.ne.12.030189.001105 (see Subsection “Kinematic Transformations” in Section “Motor Control involves Transformations”)

The terms are also used in Task Dynamics, the framework for motor control adopted in the present study. In the original Task Dynamics paper, these were simply called “kinematic relationships” rather than “kinematic maps”:

* Saltzman, E., & Kelso, J. A. (1987). Skilled actions: A task-dynamic approach.Psychological Review, 94(1), 84-106. doi:http://dx.doi.org.libproxy2.usc.edu/10.1037/0033-295X.94.1.84 (in particular, see Section “Joint variables and the task-dynamic network”, which defines functions relating “body-space variables” (i.e., controlled variables) as functions of the arm joint angles )

The kinematic maps were then defined for motor control of the vocal tract in speech production in the following Task Dynamics paper:

* Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological psychology*, *1*(4), 333-382. (see in particular Appendix 2, which defines the “direct kinematic relationships” between articulator parameters and controlled “tract variable” parameters )

These terms are still used in more recent papers on Task Dynamics in speech:

* Ramanarayanan, V., Parrell, B., Goldstein, L., Nagarajan, S., Houde, J. (2016) A New Model of Speech Motor Control Based on Task Dynamics and State Feedback. Proc. Interspeech 2016, 3564-3568. DOI: [10.21437/Interspeech.2016-1499](http://dx.doi.org/10.21437/Interspeech.2016-1499) (see Fig. 2 for the graphical description of forward kinematics)

By comparison to the sources cited above in the fields of motor control and speech production, we believe that our usage of “direct kinematics” and “differential kinematics” are consistent with common academic conventions.

* **Finally, the authors write that “factor scores [that] characterize temporal variation in the position and shape of the articulators” (line 151). It is not obvious that these factor scores bear any temporal / kinematic information. The authors are thus invited to clarify this issue.**

The revised manuscript clarified the issue by revising the sentence quoted above to read as follows: “The time-varying coefficients coefficients ... of the linear combination are factor scores that characterize temporal variation in the position and shapes of the articulators. Each image has factor scores, which change from one image to the next as the articulators move and change shape. Thus, changes in the factor scores parameterize articulator motion.”

* **Could the authors explain the relation/interaction between parameters “h” and “f”?**

The revised manuscript clarifies the relation between and by adding the following text: “The parameter sets the percentage of data-points inside the spherical neighborhood within which the forward kinematic map is estimated. The parameter is the radius of the smallest sphere containing exactly data points. The parameter is found using the nearest neighbors algorithm, where .”

* **In lines 46-48, a general definition of “imaging biomarker” is offered. In section VI.A however, no practical definition is proposed before the long technical development that leads to equation 27. The sentence (line 311-2) “The articulator synergy biomarker is the percent contribution of the jaw to narrowing and widening the vocal tract for a constriction.” suggests that a biomarker is a measurable consequence of articulator synergy, more related to latent variables that to visible geometrical parameters that may be deduced from the image. Clarifications about the nature of the biomarkers are thus strongly needed (note that Fig. 5 does not seem to be referred to, though it appears that it could contribute to this clarification).**

The quoted passage has been revised to the following: “The articulator synergy biomarker is the percent contribution of the jaw to narrowing and widening the vocal tract for a constriction. The quantity is the percent contribution of the lips (for the bilabial place) or the tongue (for the alveolar, palatal, velar, and pharyngeal places) to a constriction. Through Equation 26, these quantities are based on the kinematic relations between factor scores and s (i.e., the jacobian of the forward kinematic map), and on how the factor scores and the s evolve in time.” The revised manuscript now also uses Figure 5 to clarify the nature of the biomarker and to explain Equation 26, on which the biomarker is based.

* **Note also that the notion of “elapsed change in constriction degree” needs clarification, in particular “elapsed” (between what instants?).**

As it was used in the original submission, the term “elapsed” was ambiguous. It meant either “elapsed over the time-course of a constriction” or “cumulative sum of” (the latter being an inappropriate use of the term “elapsed”). The revised manuscript clarifies the term “elapsed” in the two ways it appears:

* + “elapsed change in constriction task variables over the time-course of a constriction”
  + “cumulative sum of contributions of the articulator”
* **The use of 10-fold cross-validation is indeed an interesting choice. However, RMSE tends to hide the largest errors. It could thus be very useful to supplement RMSE with another statistical parameter, for instance the 10th - 90th percentile range that gives a more accurate idea of the error extent.**

The revised manuscript uses the 10th-90th percentile range and the median instead of RMSE. As per Reviewer 3’s request, the values in the associated figure have been replaced by an average over all speakers.

* **Statistical results presented at page 30 are a bit confusing. For instance, the authors state that “On average, the percent jaw contribution was 17% less at the velar place compared to the bilabial place (z = 1.9, p = 0.19)”. What is the point to offer a comparison and to mention indirectly that it is not statistically significant? This section would need more specific rewriting.**

The revised manuscript seeks a compromise by reporting the results of all statistical tests in Table 1, while explaining the significant results in the text. This strikes a balance between clarity in the text and established guidelines for reporting statistical results, which we take to indicate that planned statistical tests cannot be omitted from a report because they were not statistically significant.

* **Use of “study personnel”: if this expression refers to the authors of the manuscript and of the work, it sounds peculiar and should be rephrased. If it would refer to personnel who has participated to the work but is not listed as co-authors, it would not appear to comply with standard practice in the academic world.**

The term “study personnel” was replaced with the term “authors”.

* **The notion of “constriction degree” is ambiguous. One might consider that a high degree of constriction refers to a very narrow constriction, but legend in Fig. 5 states “the jaw and tongue produce a narrowing at the palatal place, and constriction degree decreases to a minimum”, which seems to imply the opposite. More specific expressions such as “constriction area” or “constriction size” should be used.**

The term “constriction degree” has been replaced by the term “constriction task variable”. The term “task variable” was the original term for a controlled variable used in Task Dynamics.

* + Saltzman, E., & Kelso, J. A. (1987). Skilled actions: A task-dynamic approach.Psychological Review, 94(1), 84-106. doi:http://dx.doi.org.libproxy2.usc.edu/10.1037/0033-295X.94.1.84

The term “constriction task variable” further qualifies this term to indicate that it is specifically a constriction-related task variable, consistent with other current papers on Task Dynamics of speech, for instance:

* + Ramanarayanan, V., Parrell, B., Goldstein, L., Nagarajan, S., Houde, J. (2016) A New Model of Speech Motor Control Based on Task Dynamics and State Feedback. Proc. Interspeech 2016, 3564-3568. DOI: [10.21437/Interspeech.2016-1499](http://dx.doi.org/10.21437/Interspeech.2016-1499)

The term “constriction task variable” is now defined in Section III with the sentence, “As we use the term in this study, a constriction task variable is defined as the shortest distance between opposing structures at a given place of articulation.”

* **Line 148: Vertices in 2D are specified by two coordinates. This should reflect somehow in the number “p” that could be “2 x p”?**

The contour vertices are now defined as having dimension .

* **Bottom of page 10: “Contour vertices x\_i” have apparently not been defined earlier (cf. remark above).**

The revised manuscript contains a definition of the vocal tract contours in the first sentence of Section IV.A.

* **Line 548: Please, correct to “sensitivity”.**

Done

* **Figure 4. The tick marks on the right Y-axis are hardly visible.**

The revised manuscript has tick marks that are twice as wide. As per Reviewer 3’s request, these have been replaced by an average over all speakers.

* **Figure 6. Lines and symbols are too thin to be neatly visible.**

The revised manuscript has thicker lines and markers in the figure and in the legend.

Reviewer #3’s comments:

1. **First of all, one could say that it is not big news that the jaw contributes more to more anterior than more posterior constrictions. What would be more interesting to know is whether these differences can be explained by a simple view of jaw movement as pure rotation at the condyle - if so it would not be necessary to assume different synergies at different places of articulation, the differences would just fall out of the biomechanical arrangement of the articulators. In other words, do we need to assume, for example, different combinations of jaw rotation and translation at the condyle for different places of articulation?**

We agree that it is not surprising that the jaw contributes more to anterior than to posterior constrictions. Confirming expected results potentially offers a valuable validation of the articulator synergy biomarker beyond the validation provided by estimation of measurement bias (via the analysis of synthetic data) and precision (via the intra-class correlation coefficient).

The specific suggestion that mandibular morphology contributes to articulator synergies is intriguing, and one of our current projects is to uncover morphology-function relationships in speech by analyzing anatomical MRI scans along with 2D real-time MRI.

1. **Also the observation of interspeaker variability in jaw involvement does not, I feel, go as far as it might. That speakers differ is not surprising, but are they consistent within themselves across different places of articulation?**

The intra-class correlation coefficient is the ratio of inter-subject variability to total variability. If the intra-class correlation coefficient is close to 1, variance in the biomarker mostly reflects variability among participants. If the intra-class correlation coefficient is close to 0, variance in the biomarker mostly reflects intra-subject variability (cf. Equation 29 and the text immediately following). Given this interpretation of the intra-class correlation coefficient, we believe that the study quantifies exactly how consistent speakers are within their own productions. The revised submission explains this more clearly in Section VIII.B.

1. **There is also a potential confound in the design of the experiment: the study does not present a pure comparison of place of articulation, but mixes it with manner of articulation (basically approximant for palatal and pharyngeal, but stop elsewhere). In fact, my first reaction to the paper was that it would have been more interesting if it had actually focussed on manner of articulation (this is not mentioned as a possible future direction at the end), since it is well known since e.g. Vatikiotis-Bateson & Ostry (JPhon, 1995) that manner of articulation can have a substantial impact on patterns of jaw movement. Concentrating on manner of articulation at a single place of articulation may also be easier to interpret directly as differences in synergy, given that the complicating factor of distance from the condyle is more or less absent.**

The revised manuscript mentions this potential confound in the conclusion and uses this as motivation for listing manner of articulation as a possible future research direction. The revised manuscript includes the following sentence in the conclusion: “[J]aw usage varies not only by place of articulation, but also by manner of articulation. This introduces a potential confound, and thus is a limitation of the present study. Future research should focus on manner differences in articulator synergy biomarker values.”

1. **I find the basic assumption making the current approach feasible, namely that the relevant relationships are locally linear, perfectly plausible. Nonetheless it would be interesting to have more discussion of the pre-requisites for this assumption, and also whether the present approach can also be used to identify potentially interesting cases where the assumption may start to break down. To take labial constriction as a specific example, it is presumably the case in the present approach that when the constriction size goes to zero then the movements of the lips, based on the external contours of the lips, also cease (so the relationship between constriction, and articulator movement is indeed quite straightforward). However, one could argue that in a more realistic view of the movements of the articulators, the lips continue to move even when the lip aperture has reached zero, and that perhaps other measurement techniques, or other approaches to extracting movement information from the MRI images, might capture this. Would such techniques then actually be less suitable for the present approach (since the relationship between change in articulator position and change in constriction size might then be decidedly non-linear)? Perhaps a similar question from a different point of view: How would the present approach mesh with a somewhat more abstract conceptualization of the constriction target, for example use of negative constriction size (as in TADA) as a target to achieve firm closure in stops?**

Regarding a breakdown in linearity at points of tissue-tissue contact and soft tissue compression, we believe that tagged MRI may be useful for extracting motion information from soft tissue compression as the internal motion of the tongue and lips that occurs when the tongue presses against the palate or when the two lips touch.

The approach to quantifying the articulator synergy biomarkers does not have to deal directly with negative s, as of course they do not occur in real data. In related work, we are using the framework presented in the present study for simulation (as in the validation study using synthetic data) as well as in articulatory speech synthesis experiments. In such use-cases, we can use negative targets, since we have a method to deal with tissue-tissue contact.

Although not directly related to the topics presented in the present study, we view the above two topics as related directions for future work.

1. **It would be worth pointing out in the discussion that the present study only looks at one particular group of synergies, namely those involving the jaw. Many models assume that constrictions at the tongue-tip also involve synergistic movement of tip and dorsum. Would it be feasible to investigate this kind synergy too? And more specifically, can patterns of tip-jaw synergies be interpreted unambigously if at least one further synergy is involved?**

This is listed as a potential future direction for research. Specifically, the conclusion of the revised manuscript now includes the following sentence: “[W]hile the present study focused on jaw-tongue and jaw-lip synergies, the proposed method could be used to study other synergies, such as those between the dorsum and tip of the tongue or between the tongue and velum. Future research studies should characterize biomarker values for these synergies.”

1. **The formulation of the hypotheses is somewhat ad hoc. The initial formulation (p. 5) aims to test for a difference between anterior (bilabial, alveolar, palatal) and posterior (velar, pharyngeal). In fact the motivation for this specific division is not very clear. Might it not be just as plausible to suggest a division based on main active articulator (lips, vs. tongue-tip vs. tongue-dorsum/root)? The formulation on p. 29 is different, since it now refers to the jaw not contributing to posterior constrictions (not just to being different from anterior constrictions). And the formulation of the null hypotheses at the bottom of p.30 is different again.**

The revised manuscript standardizes the way the hypotheses are stated in the text.

* The first mention remains unchanged.
* The second mention reads: “The present study tested the task-dependence of articulator synergies by determining whether the jaw contributes more for anterior constrictions at the bilabial, alveolar, and palatal places of articulation than for posterior constrictions at the velar and pharyngeal places of articulation using a linear mixed effects model fitted with the package lme4 in R.”  
  To this, we add the specific null hypotheses tested: “Specifically, the present study tested the null hypotheses that there is no difference in articulator synergy biomarker values between anterior (i.e., bilabial, alveolar, and palatal) and posterior (i.e., velar and pharyngeal) places of articulation.”
* The third mention is unchanged, since it refers to the specific null hypotheses mentioned above.
* **p. 4 and elsewhere. I think phrases like "12ms temporal resolution" should be used cautiously. Is really the temporal resolution meant, or just something like frame-rate? Given the complex reconstruction algorithms used in real-time MRI the two may not be synonymous.**

The temporal resolution is correct. As mentioned on p. 7, line 118, a single image is reconstructed from the MR signal acquired in two TRs. There is no view-sharing between consecutive images.

* **Fig.1 please spell out the colours for the different contours. For example, green is presumably jaw, but this should be mentioned explicitly since bone and teeth do not image directly in MRI, so the definition of "jaw" is not necessarily obvious. (What then is the yellowish contour around the chin? Is this actually used anywhere?)**

Figure 1 now includes labels. Chin and mandible contours are analyzed together as jaw contours. This is now explicitly described in the “Jaw factors” section.

* **Fig. 2 The legend mentions "velopharyngeal port" in the list of "phonetic places of articulation", but it is not clear to me that the term "velopharyngeal port" has already been introduced in this connection. Also, I don't see the blue line that should apparently be associated to it in Fig. 2.**

Mention of the velopharyngeal port has been removed from the revised manuscript, except in the following sentence where it is used as a landmark used for defining the relevant subsection of the rear pharyngeal wall: “The pharyngeal place was the posterior pharyngeal wall, bounded superiorly by the velopharyngeal port and inferiorly by the larynx.”

* **p. 10 I think the work of Maeda is a sufficiently important part of the background that it should be referred to explicitly, rather than indirectly via the reference to Toutios & Narayanan (2015), since this could help many readers to quickly get to grips with some of the basic concepts of the current approach.**

The revised manuscript now cites Maeda.

* **p. 21, l. 287. The definition of Time 0 and Time T is not completely clear. Presumably Time 0 means onset of movement towards the constriction, not the onset of an aerodynamically relevant consonantal constriction (referred to by some authors as target achievement). (Similarly the earlier description p. 7, l. 107 is rather vague. Perhaps a sketch of a typical closing-opening movement would be useful for defining terms)**

The revised manuscript now refers to Figure 5 in order to explain these terms.

* **p. 19 . 250 typo: "is important parameter"**

The typo “is important parameter” was corrected to “is an important parameter”.

* **p. 21, i. 275 Insert mm after 2.4**

The units “mm” have been added after “2.4”.

* **p. 22 fig. 4 typo: "neighborood" I'm not sure the tiny tick marks in the right-most panels of this figure are useful. Why not just a single (visible) tick corresponding to the average over all speakers? Also, is the figure of 2.4mm really a realistic criterion for discussing error of the differential kinematics? Multiplying by the framerate this corresponds to an error of about 200mm/s, which is really huge (close to peak velocity of many articulator movements), so one would indeed hope that the error is nowhere near this.**

The typo “neighborhood” has been corrected to “neighborhood”. The individual tick marks have been removed and an average over all speakers replaces them. The revised manuscript still uses 2.4mm as the upper limit on all the graphs because it is a convenient value to keep them all on the same scale.

* **p. 37, l. 548 typo "senativity"**

The typo “senativity” has been corrected to “sensitivity”.