JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Human sensorimotor cortex control of directly-measured vocal tract movements during vowel production

David F. Conant[1,2,3], Kristofer E. Bouchard[4,5], Matthew K. Leonard[1,2] and Edward F. Chang[1,2]

[1]*Department of Neurological Surgery, University of California San Francisco (UCSF), San Francisco, California 94143*

[2]*Center for Integrative Neuroscience, UCSF, San Francisco, California 94143*

[3]*Graduate Program in Neuroscience, UCSF, San Francisco, California 94143*

[4]*Biological Systems and Engineering Division, Lawrence Berkeley National Laboratories (LBNL), Berkeley, CA 94720*

[5]*Helen Wills Neuroscience Institute, University of California, Berkeley (UCB), Berkeley, CA 94720*

DOI: 10.1523/JNEUROSCI.2382-17.2018

Received: 22 August 2017

Revised: 27 January 2018

Accepted: 29 January 2018

Published: 8 February 2018

Author contributions: D.F.C., K.E.B., and E.F.C. designed research; D.F.C. and K.E.B. collected the data; D.F.C. analyzed the data with assistance from K.E.B. and M.K.L.; D.F.C., K.E.B., M.K.L., and E.F.C wrote and edited the manuscript.; E.F.C. conceived and supervised the project.

Conflict of Interest: The authors declare no competing financial interests.

This work was supported by grants from the NIH (U01 NS098971 to E.F.C. and F32-DC013486 to M.K.L.). E.F.C. is a New York Stem Cell Foundation Robertson Investigator. This research was also supported by the New York Stem Cell Foundation, the Howard Hughes Medical Institute, the McKnight Foundation, The Shurl and Kay Curci Foundation, and The William K. Bowes Foundation. KEB was supported by Laboratory Directed Research and Development (LDRD) funding from Berkeley Lab, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Corresponding author: Edward Chang, University of California, San Francisco, Department of Neurological Surgery, 505 Parnassus Ave., M-779, San Francisco, CA 94143-0112, Email: edward.chang@ucsf.edu

Cite as: J. Neurosci ; 10.1523/JNEUROSCI.2382-17.2018

Human sensorimotor cortex control of directly-measured vocal tract movements during vowel production

David F. Conant[1,2,3], Kristofer E. Bouchard[4,5], Matthew K. Leonard[1,2], Edward F. Chang[1,2*]

1. Department of Neurological Surgery, University of California San Francisco (UCSF), San Francisco, California 94143
2. Center for Integrative Neuroscience, UCSF, San Francisco, California 94143
3. Graduate Program in Neuroscience, UCSF, San Francisco, California 94143
4. Biological Systems and Engineering Division, Lawrence Berkeley National Laboratories (LBNL), Berkeley, CA 94720
5. Helen Wills Neuroscience Institute, University of California, Berkeley (UCB), Berkeley, CA 94720

*: Corresponding author
Edward Chang
University of California, San Francisco
Department of Neurological Surgery
505 Parnassus Ave., M-779
San Francisco, CA 94143-0112
Email: edward.chang@ucsf.edu

Author contributions: D.F.C., K.E.B., and E.F.C. designed research; D.F.C. and K.E.B. collected the data; D.F.C. analyzed the data with assistance from K.E.B. and M.K.L.; D.F.C., K.E.B., M.K.L., and E.F.C wrote and edited the manuscript.; E.F.C. conceived and supervised the project.

This work was supported by grants from the NIH (U01 NS098971 to E.F.C. and F32-DC013486 to M.K.L.). E.F.C. is a New York Stem Cell Foundation Robertson Investigator. This research was also supported by the New York Stem Cell Foundation, the Howard Hughes Medical Institute, the McKnight Foundation, The Shurl and Kay Curci Foundation, and The William K. Bowes Foundation. KEB was supported by Laboratory Directed Research and Development (LDRD) funding from Berkeley Lab, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

The authors declare no competing financial interests.

8 Figures

Abstract: 227 Words
Introduction: 606 Words
Discussion: 1315 Words

Abstract

During speech production, we make vocal tract movements with remarkable precision and speed. Our understanding of how the human brain achieves such proficient control is limited, in part due to the challenge of simultaneously acquiring high-resolution neural recordings and detailed vocal tract measurements. To overcome this challenge, we combined ultrasound and video monitoring of the supralaryngeal articulators (lips, jaw and tongue) with electrocorticographic (ECoG) recordings from the cortical surface of 4 subjects (three female, one male) to investigate how neural activity in the ventral sensory motor cortex (vSMC) relates to measured articulator movement kinematics (position, speed, velocity, acceleration) during the production of English vowels. We found that high-gamma activity at many individual vSMC electrodes strongly encoded the kinematics of one or more articulators, but less so for vowel formants and vowel identity. Neural population decoding methods further revealed the structure of kinematic features that distinguish vowels. Encoding and decoding of articulator kinematics was sparsely distributed across time, and primarily occurred during the time of vowel onset and offset. In contrast, encoding was low during the steady state portion of the vowel, despite sustained neural activity at some electrodes. Significant representations were found for all kinematic parameters, but speed was the most robust. These findings enabled by direct vocal tract monitoring demonstrate novel insights to the representation of articulatory kinematic parameters encoded in the vSMC during speech production.

Significance Statement

Speaking requires precise control and coordination of the vocal tract articulators (lips, jaw, and tongue). Despite this impressive proficiency, our understanding of how the brain achieves such control is rudimentary, in part because the movements themselves are difficult to observe. By simultaneously measuring speech movements and the neural activity that gives rise to them, we demonstrate how neural activity in sensorimotor cortex produces complex, coordinated movements of the vocal tract.

Introduction

When we speak, we move the upper vocal tract articulators (lips, jaw, and tongue) to produce vocal tract constrictions of air flow in precise, rapid, and complex ways. These movements result in acoustic events that are highly distinguishable, maximizing communicative utility. In spoken languages, vowels are a major category of speech sounds. Despite their importance it is unknown how cortical activity patterns control the vocal tract articulators to create vowels.

The ventral sensory-motor cortex (vSMC: pre-, post-, and sub-central gyri) is the primary cortical area controlling the speech articulators (Petersen et al., 1988; Lotze et al., 2000; Hesselmann et al., 2004; Brown et al., 2009; Kellis et al., 2010; Pei et al., 2011; Bouchard et al., 2013; Bouchard and Chang, 2014; Mugler et al., 2014; Simonyan et al., 2016; Kumar et al., 2016). Within vSMC, representations of vocal tract articulators are coarsely somatotopically organized, with different neural populations in vSMC being associated with specific articulators (Crone et al., 1998; Brown et al., 2009; Bouchard et al., 2013; Mugler et al., 2014; Herff et al., 2015). However, our understanding of speech motor control in vSMC is incomplete, due to challenges in simultaneously acquiring

neural and behavioral data with sufficient spatial and temporal resolution required to determine the precise correspondence between vSMC activity and the movement of the vocal tract articulators.

The precise movements (kinematics) of the articulators are challenging to measure because many of the vocal tract movements are internal to the mouth and throat, and therefore difficult to monitor externally, especially in the context of neural recordings. As a result, previous studies have used the produced acoustics to infer which articulators are involved, based on extensive linguistic descriptions of speech movements for a given speech sound (Lotze et al., 2000; Crone et al., 2001; Brown et al., 2009; Fukuda et al., 2010; Kellis et al., 2010; Pei et al., 2011; Leuthardt et al., 2011; Grabski et al., 2012; Bouchard et al., 2013; Bouchard and Chang, 2014; Mugler et al., 2014; Herff et al., 2015). Although it is possible to describe the movements of each articulator according to phonetic labels derived from the acoustics, these behavioral descriptions cannot provide exact characterizations of the changing positions of the articulators over time. Moreover, there are many articulator configurations that can result in the same acoustics (Atal et al., 1978; Maeda, 1990; Johnson et al., 1993; Gracco and Lofqvist, 1994; Lofqvist and Gracco, 1999) and considerable across-speaker (Johnson et al., 1993) and across-trial (Perkell and Nelson, 1985) variability in movements that give rise to a particular speech sound. Thus, understanding how the brain produces complex sounds like vowels requires determining how different kinematic parameters of articulatory movements are controlled in vSMC during speech production.

To understand how vSMC neural activity controls precise articulator movements, we have developed a system to simultaneously measure cortical activity using high-

resolution electrocorticography (ECoG) while directly monitoring the lips and jaw with a camera, and the tongue with ultrasound. We previously detailed a technical description of the methods (Bouchard et al., 2016). Here, we examined how vSMC generates articulator kinematics, focusing on the production of American English vowels. We established that articulator kinematics are more strongly represented in vSMC compared to acoustics. We determined that specific kinematic parameters (position, speed, velocity, and acceleration) are all represented, though articulator speed is represented most strongly. Finally, we examined how distinct dynamics of neural activity are related to both movement (from rest to target position) and maintenance of articulators (at target position). By simultaneously measuring speech-related movements and the neural activity generating them, we demonstrate how neural activity in sensorimotor cortex produce complex, coordinated movements of the vocal tract.

Materials and Methods

Experimental Design

Electrocorticography acquisition and signal processing

Four human participants underwent chronic implantation of a high-density subdural electrocorticographic array (ECoG) as part of the clinical treatment of epilepsy (3 female right hemisphere, one male left hemisphere). All subjects were implanted with 256-channel grids over peri-Sylvian cortex (1.17mm diameter electrodes, 4mm pitch, 60x60mm coverage; Integra [Plainsboro NJ, USA]), referenced to scalp electrode. The

total number of vSMC electrodes for individual subjects ranged from 52 to 86 for a total of 270. Cortical-surface electrical potentials were recorded with ECoG arrays and the voltage time series from each electrode was inspected for artifacts or excessive noise. Electrodes with excessive noise and time periods with artifacts were excluded from analysis, and the raw ECoG activity was re-referenced to the common average. For each channel, the time-varying analytic amplitude of the voltage signal in the high-gamma (HG) range (70-150 Hz) was extracted using the Hilbert transform, according to previously published procedures (Edwards et al., 2010). HG correlates well with multi-unit firing (Ray and Maunsell, 2011), and has high spatial and temporal resolution (Muller et al., 2016). The HG signal was down-sampled to 400 Hz for analysis and plotting purposes. HG power was *z*-scored relative to activity recorded during periods of silence during the same recording session. All analyses were limited to the ventral sensory-motor cortex (vSMC), which was anatomically defined as the ventral portions of the pre-central and post-central gyri, as well as the sub-central gyrus.

Task

Participants listened to audio recordings of nine English vowels (*a*/*æ*/*ʌ*/*ɛ*/*ɝ*/*ɪ*/*i*/*ʊ*/*u*) and were instructed to repeat each vowel. On each trial, to ensure they properly identified the vowel, they first heard it in an /h-V-d/ context (e.g. 'hood'), and then they heard the vowel in isolation. After a 1-1.5 sec delay, participants were presented with a visual cue to produce the isolated vowel. They were not explicitly instructed to hold the vowel for a specific amount of time. The median duration of production was 1.66 seconds (STD = 0.35 s). For each participant, between 15 and 30 repetitions of each vowel were collected over the course of 3-6 recording sessions.

Articulator Tracking

We developed a system to record the movements of the main supra-laryngeal articulators while participants performed the vowel production task (Figure 1A), the details of which have been described previously (Bouchard et al., 2016). Briefly, to capture the movement of the lips and jaw, a camera was placed in front of the participant's mouth. The participant's lips were painted blue, and red dots were painted on the tip of the nose and the chin to simplify the process of extracting the shape and position of these articulators. The camera captured video at 30 frames per second. To image the tongue, an ultrasound transducer was held firmly under the participant's chin with the plane-of-view capturing the midline of the tongue. The ultrasound recorded images at 30 frames per second, and the data were aligned to the lips/jaw video according to the peak of the cross-correlation of the audio signals from each video. Using hue thresholding, we extracted the lips and jaw automatically from these videos as binary masks (Figure 1B). From these binary masks, we extracted the locations of the four corners of the mouth (upper/lower lip, left/right corners) and the jaw. For the tongue, we used EdgeTrak to extract 100 points of the mid-sagittal contour, which were then down sampled to 3 points by taking the median x and y value for the front, middle, and back thirds of the contour (Li et al., 2004). Since video and ultrasound were collected in orthogonal spatial planes, x and y positions in the lips/jaw images reflect left/right and top/bottom, whereas x and y positions in the tongue images reflect front/back and top/bottom. To correct for differences in the relative position of the camera and ultrasound transducer with respect to the participant, we referenced each articulatory

point to the neutral starting position at the beginning of each trial. From the measured position of each articulatory feature (X) we also derived movement parameters including velocity (X'), speed (|X'|), and acceleration (X'') of that articulator. We refer to these parameters collectively as the articulator kinematics. While the lips and jaw were both included in all analyses, we found that lip opening and jaw height were correlated for this vowel production task (cross-subject average correlation: $r = 0.73 \pm 0.12$). Therefore, to simplify visualizations we only show results for the lips.

Statistical Analysis

Acoustic feature extraction

Speech sounds were recorded using a Sennheiser microphone placed in front of the participant's mouth and recorded at 22 kHz (Figure 1A). The speech signal was transcribed offline using Praat (Boersma, 2001). For each vowel, we extracted the formants (F_1-F_4) using an inverse filter method (Watanabe, 2001; Ueda et al., 2007; Bouchard et al., 2016).

Trial duration standardization

To standardize the durations of the vowels across trials and participants, we linearly resampled each trial to be the median duration across vowels and subjects (1.66 seconds). Behavior and neural signals changed with rapid and stereotyped dynamics around onset and offset; resampling the entire trial would systematically change those dynamics based on vowel duration. Therefore, to preserve onset and offset dynamics, we

only resampled data in the time window from 250ms after the onset of the acoustics to 250ms before the offset: corresponding to the steady-state hold. Trials with durations less than half or greater than twice the median were excluded from analysis (26 in total across all subjects). Final analyses utilized an average of 15.3+/- 5.7 trials per vowel per subject.

Permutation tests

To evaluate statistical significance in each analysis, we used permutation tests. A permutation distribution for a given model was constructed by randomly permuting the trial labels of the observed data, and then training and testing the model using this shuffled data. This process was repeated 500 times, and the performance of these shuffled models comprised the permutation distribution. A model was considered significant if its performance on test data was greater than the 99^{th} percentile of its corresponding permutation distribution. For the correlations in Figure 2D, we test if $|r| > 99^{th}$ percentile of $|r_{null}|$.

Correlations with articulatory position

To evaluate the relationship between vSMC HG activity and individual articulators, we correlated HG activity at individual electrodes with the measured trial-to-trial position for each articulatory feature. HG activity averaged over a 200ms window centered at acoustic onset was correlated with the mean position of each articulator taken from a 200ms window centered at the midpoint of the vowel. Electrodes were labeled according to whether they had significant correlations with zero, one, or multiple

articulator positions. Electrodes in 2D are examples of electrodes with significant correlations with only one articulator.

Encoding of kinematics, formants, and vowel categories

We compared the representation of articulator kinematics, vowel formants, and vowel category at each electrode using L_1-regularized linear regression (Lasso). These models predict HG activity at each time point 500ms before acoustic onset to 500ms after acoustic offset from a sparse linear combination the behavior:

$$HG_e = \sum_{i=1}^{n} \beta_i \quad (1)$$

Where HG_e is the HG power at a given electrode, β are the linear weights that describe the mapping, and i is a vowel category (n=9), vowel formant feature (n=10), articulator kinematic feature (n=40), or all feature sets jointly (n=59). The formant features were F1-F4, as well as all pairwise ratios of F1-F4. The articulator kinematic parameters were position, speed, velocity, and acceleration for lip opening, lip width, jaw height, and the front, middle, and back tongue. Vowel identity was parameterized as 9 binary vectors corresponding to the vowel being produced during vocalization. Formant, articulatory, and vowel identity features were lagged +100ms relative to HG, corresponding to the causal direction of neural activity generating behavior. This lag was determined empirically by optimizing model performance over a range of lag values (-500ms to +500ms).

To train and test linear models, we used L_1-regularized linear regression in a leave-one-trial-out cross-validation procedure. We calculated the correlation between the observed and predicted HG values, averaged across cross-validations. Electrodes were

included in visualizations and summary statistics only if their performance passed the permutation test described above for at least one of these models (i.e., formants, kinematics, vowel identity, or combined). To compare models with different numbers of parameters, we calculated the adjusted R^2:

$$R^2_{adj} = R^2 - (1 - R^2)\frac{p}{n-p-1} \quad (2)$$

Where R^2 is the unadjusted coefficient of determination of the model, n is the number of observations the model was trained on, and p is the number of parameters.

Organization of vowels in behavioral and neural spaces

To examine the similarity of vowels in behavioral and neural representation spaces, we used multi-dimensional scaling (MDS). MDS provides a low-dimensional projection of the data that preserves the relative distances (similarities) between points in a higher-dimensional space. For each feature set (formants, articulator position, and neural) we extracted the median value for each vowel from a 200ms window centered at the midpoint of the vowel (formants and articulator position) or the onset (neural), and then z-scored that value across vowels. We applied MDS to the distance matrix computed on these measurements for each feature set separately. To measure the differences in the organization of vowels between the formant, articulator, and neural spaces, we calculated the pairwise distances between the vowels in each space. We quantified the similarity between the neural and kinematic or formant spaces by calculating a bootstrapped correlation between the pairwise distances for each feature set. We performed agglomerative hierarchical clustering on the pairwise distances to visually organize the results.

Encoding of kinematic parameters across time

To assess the relative encoding of different kinematic parameters, we used the measured position of each articulator on each trial (X) to derive the velocity (X'), speed (|X'|), and acceleration (X'') of that articulator on that trial. To examine the encoding of these parameters independent of one another, we removed the shared variance between these parameters using semi-partial correlation. For each time point we first used linear regression to predict the values of one kinematic parameter, y, from a linear combination of the remaining 3 parameters, X:

(3) $$\hat{y} = \beta X$$

Where β are the weights that describe the linear relationship, and $\hat{y}$ is the model's prediction of that kinematic parameter. We then calculated the linearly independent component of the kinematic parameter, y_{idp}, by subtracting predicted parameter values from the observed:

(4) $$y_{idp} = y - \hat{y}$$

We then used L_1-regularized linear encoding models to predict HG activity from the kinematic parameters (position, speed, velocity, and acceleration) of the lips, jaw, and tongue. However, instead of including the entire trial time-course in each model, we trained and tested models within 100ms non-overlapping windows that tiled the trial. Articulator kinematics were lagged +100ms relative to HG to evaluate the causal nature of neural activity on behavior. Models were trained and tested independently for each time window. Performance was measured by the correlation between the observed and predicted HG values, averaged across cross-validations. Electrodes were included in

visualizations and summary statistics only if their performance passed the permutation test described above for at least 3 contiguous time windows at any point in the trial.

Decoding of kinematic parameters from vSMC HG

To determine the degree to which variations in individual articulatory parameters could be predicted from vSMC population activity, we used linear decoding models. Similar to the encoding models above, we built L_1-regularized linear models to relate vSMC HG to articulator kinematics within 100ms time windows that tile the trial. However, instead of predicting the HG activity at a single electrode from a combination of all articulator parameters, we predicted the trial-to-trial variance of each articulator parameter from a combination of all vSMC electrodes. As with the encoding models, articulator kinematic features were lagged 100ms relative to vSMC HG, and models were trained and tested independently for each time window using a leave-one-out cross-validation procedure. The resulting models thus express how well the vSMC population can predict each resulting articulator kinematic feature as a function of time within the vocalization. To simplify visualization, we averaged performance across subjects and similar articulators.

Description of vSMC HG dynamics

To characterize the major physiological response types in HG dynamics, we used non-negative matrix factorization (NMF). NMF is a dimensionality reduction technique that extracts a predetermined number (i.e., rank, k) of bases ($B \in \mathbb{R}^{m\mathrm{x}k}$) and weights ($W \in \mathbb{R}^{n\mathrm{x}k}$) that linearly combine to reconstruct the non-negative data ($A \in \mathbb{R}^{m\mathrm{x}n}$), such

that k < min(n,m) under the constraint that both the bases and weights are strictly non-negative:

$$A \approx BW^T;\ B,W \geq 0 \quad (5)$$

The solutions B and W are found by solving the (bi-convex) constrained optimization problem:

$$\widehat{B},\widehat{W} = \min_{B,W} \frac{1}{2}\|A - BW^T\|_F^2;\ s.t. B,W \geq 0 \quad (6)$$

NMF is particularly useful for decomposing data into 'parts' that have interpretable meanings (e.g., transient vs. sustained response types) (Lee and Seung, 1999; Donoho and Stodden, 2004; Bouchard et al., 2016; Leonard et al., 2016). The HG activity for each vSMC electrode across all participants was averaged across trials, offset by the minimum value (such that all values were positive), and NMF was applied to the matrix of time courses x electrodes. To determine a parsimonious number of bases, we calculated the reconstruction error when projecting the data onto the NMF bases:

$$err = \frac{1}{2}\|A - BW^T\|_F^2 \quad (7)$$

We then found the number of bases (i.e., rank k) beyond which reconstruction error only marginally reduced (i.e., the elbow of the curve): five bases were used. The first two bases resembled the transient and sustained activity observed in Figure 7A. Electrodes with sustained activity were defined as those that had weighting for basis 1 greater than for basis 2. The width (HG_w) of the HG activity for sustained electrodes was derived as follows:

$$HG_w = argmin \int (HG_{e,t} - \overline{HG_{e,t}}) - argmax \int (HG_{e,t} - \overline{HG_{e,t}}) \quad (8)$$

Where $HG_{e,t}$ is the HG activity at given sustained electrode for a given trial. This measure was calculated for each sustained vSMC electrode, for each trial. We assessed

spatial organization by measuring the Euclidean distance between electrodes organized according to their maximum NMF weight (i.e., transient or sustained). We compared distributions of intra-parameter distances and cross-parameter distances to randomized distributions derived by shuffling the labeling of the electrodes. If the HG dynamic variability across vSMC is spatially organized, the distribution of intra-parameter and cross-parameter distances should differ from the distributions of the random distributions.

Results

Participants produced nine English vowels in isolation (*a/æ/ʌ/ɛ/ɝ/ɪ/i/ʊ/u*) (e.g., the vowels pronounced as in the following set of words: “calm”, “cat”, ”send”, ”fun”, ”heard”, ”sit”, ”need”, ”should”, ”boot”) while neural activity was recorded from ventral sensorimotor cortex (vSMC) and the movements of the supra-laryngeal articulators were monitored. These vowels densely cover both the acoustic and kinematic space of all American English vowels, and are a basic and essential component of all languages. We specifically studied vowels in isolation for several reasons. First, the associated movements of the speech articulators consist of a single displacement from rest, to the target position, and back to rest. This simplicity provides the opportunity to study isolated movements of the speech articulators free from the context of surrounding phonemes. The task was also designed to minimize variability in the lower vocal tract (e.g. larynx), which we did not explicitly monitor. Using the recorded acoustics, we verified that subjects produced the vowels with little trial-to-trial variation in either pitch or intensity. Additionally, the movements occur at distinct epochs, allowing us to resolve the neural representation of the movement to the target, from the maintenance of that target, from the return to the resting configuration.

Articulator tracking during vowel production

We simultaneously tracked the movements (Figure 1A) of the major supralaryngeal articulators (i.e., lips, jaw, and tongue; Figure 1B) while recording neural activity directly from the cortical surface (Figure 1C; see Methods, Bouchard et al., 2016). We first verified that by extracting the positions of the articulators, we observed characteristic vocal tract configurations that reflect distinct vowels. For example, the vowel /a/ is characterized by lowering the front tongue, raising the back tongue, and opening the lips, while the vowels /i/ and /u/ have different configurations (Figure 2A). The measured articulatory movements captured these characteristics, and clearly discriminated vowel categories (Figure 1B, 2B). We also used the produced acoustics as a behavioral measure of vowel discriminability. By extracting the formants from the acoustic signal, we observed distinct relative patterns of acoustic power for different vowels. For example, /a/ is characterized by high F1 and low F2, whereas /i/ and /u/ have different formant profiles (Figure 2C).

Representation of articulator kinematics in ventral sensory-motor cortex

These descriptions demonstrate that both articulator kinematics and acoustic formants provide rich descriptions of the same behavior. However, although kinematics and acoustics are causally related, their relationship is not one-to-one (Atal et al., 1978; Maeda, 1990; Johnson et al., 1993; Gracco and Lofqvist, 1994; Lofqvist and Gracco, 1999), nor are they perfectly correlated (in the present data set, $rho_{kin,acoust}=0.53 \pm 0.17$). For example, producing the vowel /uw/ (“h<u>oo</u>t”) involves raising the back of the tongue

393 towards the soft palate while rounding the lips. However, those movements can be
394 compensatory. The vowel /u/ can be produced with less pronounced lip movements
395 accompanied by greater tongue movements, or vice-versa (Perkell et al., 1993).
396 Therefore, we asked whether articulator kinematics or acoustic formants are the
397 behavioral characterization of vowels represented in vSMC.

398 First, we quantified how well the positions of speech articulators or vowel
399 formants explain the variance of HG at individual vSMC electrodes (i.e., encoding
400 strength). We recorded cortical electrical potentials from a total of 270 electrodes from
401 the surface of vSMC across 4 subjects (Figure 1C). The high-gamma (HG) activity at
402 many vSMC electrodes was elevated above baseline during the speech movements, and
403 was significantly correlated with the trial-to-trial position of the speech articulators
404 (Figure 2D). We observed a clear relationship between articulator position and HG
405 activity. For illustration, we identified representative electrodes where activity was most
406 correlated with a single articulator. For example, the HG activity of electrode 1 at the
407 time of vowel onset was significantly correlated with only the back tongue. Likewise,
408 electrode 2 showed greater activity for higher front tongue positions. Electrode 3 was
409 correlated with the opening of the lips. To examine whether HG activity at these
410 electrodes was similarly correlated with the produced acoustics, we binned the activity by
411 formant values (Figure 2E). We observed weaker correlations with formants compared to
412 articulator position, demonstrating more robust encoding of articulatory representations.

413 We were specifically interested in whether vSMC activity is best explained by
414 articulator kinematics, vowel formants, or vowel identity. We used linear encoding
415 models to predict neural activity from kinematic or acoustic features, or the vowel

identity (see Methods). Across electrodes, we found that articulator kinematics provided significantly better model fits compared to vowel formants (U = 5.3, p = 1.0e-8; Wilcoxon rank sum), or vowel identity (U = 8.7, p = 4.1e-18; Wilcoxon rank sum) (Figure 3A). We used nested models to examine how much additional neural variance is explained by predicting HG from both articulator kinematics and vowel formants. We found that the joint model explained no more variance than the articulator kinematics alone (U = 0.4, p = 0.7; Wilcoxon rank sum) suggesting that the performance of the formant models was likely driven by variance shared with the kinematics. Therefore, we find no evidence for encoding of vowel formants separate from their articulatory origin. Furthermore, these results demonstrate that the production of distinct vowels is grounded in direct control of articulator kinematics.

Across all vSMC electrodes, we found that 27% (74 out of 270) were significantly correlated with movements of one or more articulators (correlations > 99^{th} percentile of permutation distributions). We observed a clear spatial organization to the articulator correlations, with lips/jaw more dorsal than the tongue (Figure 3B), consistent with previously-described somatotopy (Penfield and Roberts, 1959; Brown et al., 2009; Bouchard et al., 2013; Huang et al., 2013; Conant et al., 2014). Within the ventral region, we observed electrodes that more strongly reflected either the front or back of the tongue. Both front and back tongue electrodes were distributed throughout the broader tongue region. Finally, we observed 45 electrodes that had significant correlations with multiple articulators, which were distributed throughout vSMC. Together, these results extend our understanding of speech-motor cortex somatotopy by demonstrating that the dominant

encoding scheme in these neural populations reflects the specific movements of the preferred articulators.

Organization of vowels in vSMC population activity

To understand how vSMC encoding of articulator kinematics contributes to our ability to produce distinct vowels, we examined the organization of behavioral and neural activity in relation to all nine vowels. In addition to articulator kinematic representations at individual electrodes, population activity in vSMC may reflect the coordinated movements of the vocal tract that produce vowels. Furthermore, because vowel formants arise from the relative positions of multiple articulators, it may be the case that while articulators are most strongly represented at single electrodes, population activity may reflect a different, emergent representation. We examined the organization of speech representations at the population level by comparing the relative distances of vowels in acoustic, articulatory, and neural space. We performed multi-dimensional scaling (MDS) on the vowel centroids (see Methods) measured by vowel formants, articulator position, or vSMC neural activity across all participants. In this analysis, vowel tokens that are near each other in MDS space have similar formant, kinematic, or HG values. Consistent with previous behavioral and linguistic descriptions of vowel production, the formant and kinematic MDS projections replicate the classic vowel space 'trapezoid' (Figure 4A-C) (Hillenbrand et al., 1995; Ladefoged and Johnson, 2011). The HG neural MDS projection also closely resembled the acoustic/kinematic organization of the vowels. For example, the vowel /a/ (as in h<u>a</u>ll) is near the vowel /Λ/ (as in h<u>u</u>t), but far from /i/ (as in h<u>ea</u>t).

To characterize the difference in organization of vowels across these feature spaces, we calculated the pairwise distances between the vowels in MDS space, visualized as confusion matrices (Figure 4D-F, right). We additionally performed hierarchical clustering of the pairwise distances and organized the confusion matrices by the derived hierarchical organization. The pairwise distances and hierarchical clustering reaffirm the classic vowel organization, but also highlight the specific differences between the feature spaces. For example, /i/ is distant from the other vowels in the formant space, but closer in the articulator and neural spaces. We found that the organization of vowels in vSMC HG activity is significantly more correlated with the organization of vowels in the articulator space compared to the acoustic space ($U = 9,5$, $p=1.3e-21$, Wilcoxon rank sum), although both representations were significantly correlated with vowel organization in the HG neural space (acoustic: $r=0.56$, $p=2.8e=-4$, kinematic: $r=0.73$, $p=5.9e=-5$).

Encoding of articulator kinematic parameters

In the above analyses, we considered the joint encoding of multiple kinematic parameters for individual articulators. However, it is unknown whether kinematic encoding reflects particular aspects of the articulator movements. The movements of the articulators can be described according to a variety of different kinematic parameters (e.g. position, speed or velocity, acceleration, etc.,). For each kinematic parameter, we used L1-regularized encoding models to explain vSMC HG from the moment-to-moment measurements of position, speed, velocity, and acceleration. Since all four kinematic parameters are correlated with one another, we removed shared variance between the

parameters using semi-partial correlations in order to better interpret their relative encoding performances.

We found electrodes that significantly encoded the trial-to-trial variability in position (Figure 5A), speed (Figure 5B), velocity (Figure 5C), and acceleration (Figure 5D). Speed was the most robustly encoded parameter at most vSMC electrodes, with significant encoding at more electrodes and a higher average correlation compared to the other parameters ($U = 1720$ to 2735, $p = 3.3e\text{-}9$ to $1.1e\text{-}14$; Wilcoxon rank sum; Figure 5E).

To understand the timing of kinematic parameter encoding throughout the production of vowels, we also examined models that predicted HG neural activity from the joint combination of all four parameters simultaneously. These models were evaluated over a sliding 100ms window to characterize the kinematic parameter encoding during different phases of the trial (i.e., movement initiation, target position, steady state maintenance, and movement back to the starting position). We observed a peak in encoding for most electrodes around the onset of the movement (91% of electrodes), with some electrodes also showing a peak around the offset (9%; Figure. 5F). There was no spatial organization associated with electrodes that specifically encoded particular parameters (intra-parameter, $p = 0.31$; cross-parameter, $p = 0.08$; see Methods), nor was there a significant relationship between electrodes that encoded specific kinematic parameters and specific articulators (χ^2 (9, N=155) = 9.26, $p = 0.4$; Chi-square). Strikingly, encoding during the steady state was near zero for all kinematic parameters.

To understand how these individual electrode kinematic representations relate to the population representations of articulator kinematics and dynamics, we additionally

used L1-regularized decoding models to predict the articulator kinematics from the population of vSMC HG electrodes. As with the encoding analyses, these models were constructed from a small (100ms) sliding window of time, resulting in a description of how much of the trial-to-trial variability of the articulator position (Figure 6A), speed (Figure 6B), velocity (Figure 6C), and acceleration (Figure 6D) can be explained by vSMC HG activity. The time course of decoding strength was similar to the encoding models, with peaks around the onset and offset and near-zero values while the vowel was being held. Across kinematic parameters, articulator speed was the best-predicted parameter ($U = 2.6$ to 3.3, $p = 8.1e\text{-}3$ to $9.0e\text{-}4$; Wilcoxon rank sum).

Together, these results demonstrate a strikingly sparse representation of kinematic parameters across time, despite the fact that there continues to be trial-to-trial variability in both kinematic and neural features throughout vowel (Figure 5). Only 56% of time points had significant encoding performance at any electrode, and individual significant electrodes had an average of 15% (±1) significant time points. In particular, we did not observe any electrodes that exhibited significant kinematic parameter encoding during the steady state of the vowel.

Onset vs. steady state HG activity and kinematic encoding

The temporal sparsity of neural representations described above is particularly notable given that many electrodes showed sustained HG activity during the steady state portion of the vowel, independent of the particular articulatory movements that occurred (Figure 7A). These electrodes contrast with other HG activity that is only transiently increased around the onset and/or offset of the vowel (Figure 7A). To characterize these

529 response types, we used non-negative matrix factorization (NMF) to derive basis
530 functions that best describe vSMC HG temporal profiles across all electrodes (Hamilton
531 et al., 2017). Our motivation for using NMF was not to provide a complete description of
532 HG dynamics, but rather to provide an unsupervised method of quantifying the transient
533 and onset/offset responses across electrodes. We found that the first two bases (i.e., the
534 most important bases), captured the sustained and onset/offset response types we
535 observed qualitatively (Figure 7B). Organizing all vSMC electrodes by the degree to
536 which their activity is reconstructed by the first or second NMF bases (i.e., the NMF
537 weights), we observed a continuum of HG dynamics: some electrodes showed sustained
538 activity throughout the utterance, while others showed transient increases in activity only
539 at onset and offset of the utterance (Figure 7C). Some electrodes showed a combination
540 of sustained and transient components. There was no apparent spatial organization (intra-
541 response type, $p = 0.13$; cross-response type $p = 0.09$; see Methods) or relationship
542 between response type and the articulators represented at each electrode (χ^2 (3, N=155) =
543 2.3, $p = 0.5$; Chi-square).

544 We separately considered electrodes that showed stronger weights for the
545 sustained NMF basis (basis 1 in Figure 7B). The average HG activity at these electrodes
546 indeed showed sustained activity throughout the vowel, however there was not a
547 concomitant sustained encoding of kinematic parameters (Figure 7D). This dissociation
548 between activity and encoding was apparent even at the single trial level (Figure 8A).
549 Thus, although some electrodes exhibit sustained HG activity throughout the production
550 of the vowel, there is not a systematic relationship between the trial-to-trial variability of
551 that activity and the kinematics of the articulators. Instead, encoding of kinematics at

552 electrodes with sustained activity was prevalent only around the onset and offset of
553 movement. We hypothesized that although activity during the steady state does not relate
554 to kinematic variability, it still reflects an important aspect of the task, namely the
555 duration of each utterance. Across sustained electrodes we found that the duration of the
556 HG timecourse was significantly correlated with the duration of the vocalization
557 (Spearman's rho = 0.61, p = 2e-153; Figure 8B). Thus, at a minimum the sustained
558 activity was associated with vowel production.
559
560 Discussion
561 We report a detailed description of how activity in speech-motor cortex controls
562 the precise movements of the vocal tract articulators to produce vowels. By
563 simultaneously measuring the movements of the articulators, recording the acoustic
564 consequences of those movements, and recording the neural activity in vSMC, we are
565 able to establish that the dominant representation in vSMC is articulator kinematics. The
566 precise control of these movements allows speakers to create specific configurations of
567 the mouth, which lead to distinct categories of sounds.
568 Without simultaneous measurements of the articulators, previous studies of the
569 neural basis of speech production have relied on approximate, categorical phonemic-
570 based descriptions of speech behavior (Crone et al., 2001; Fukuda et al., 2010; Kellis et
571 al., 2010; Leuthardt et al., 2011; Pei et al., 2011; Bouchard et al., 2013; Bouchard and
572 Chang, 2014; Mugler et al., 2014; Herff et al., 2015). Although the produced acoustics
573 and categorical vowel descriptions reflect the ultimate (perceptual) outcome of vocal tract
574 movements, the many-to-one relationship between kinematics and vowels (Atal et al.,

1978; Maeda, 1990; Johnson et al., 1993; Gracco and Lofqvist, 1994; Lofqvist and Gracco, 1999) means that it was not possible to understand the precise nature of the neural representation in vSMC. Previous studies have implicated vSMC in articulator kinematic control in several ways. First, stimulation to sites in vSMC elicits involuntary activations of the orofacial muscles (Penfield and Boldrey, 1937; Huang et al., 2013). Second, neurons in these and other sensorimotor regions are often tuned to movement kinematics (Georgopoulos et al., 1986; Paninski et al., 2004; Arce et al., 2013). Finally, the spatio-temporal patterns of HG activity in vSMC are consistent with the engagement of the articulators (Bouchard et al., 2013). The present results confirm these interpretations by showing directly that kinematic descriptions of speech behavior are more closely related to neural activity compared to acoustic or categorical vowel descriptions. Further, we find no evidence that vSMC activity encodes either produced acoustics or vowel category distinct from their correlations with the articulator kinematics. Crucially, we observed that this encoding scheme exists both at single electrodes, and across the spatially distributed set of electrodes. For spatially distributed activity patterns, we demonstrate the neural existence of the classic vowel 'trapezoid', which has dominated linguistic descriptions of speech (Harshman et al., 1977; Alfonso and Baer, 1982; Hillenbrand et al., 1995).

Furthermore, by characterizing the movements of the articulators according to a variety of kinematic parameters (position, speed, velocity, and acceleration), we demonstrated that neural activity encodes each of the examined parameters independent of one another. Previous studies examining arm movements using analogous parameters have also found significant encoding of these parameters (Georgopoulos et al., 1982,

598 1984; Ashe et al., 1994; Moran and Schwartz, 1999; Paninski et al., 2004). While we find
599 electrodes that significantly encode each parameter examined, speed is by far the most
600 robustly encoded parameter. Furthermore, the dominant kinematic parameter at
601 individual electrodes was not significantly related to the articulator representation of
602 those electrodes. The predominance of speed over other parameters is somewhat
603 surprising; previous studies of the single-unit representation of kinematic parameters
604 during arm reaching typically find that velocity and direction are the most commonly
605 encoded parameter (Moran and Schwartz, 1999). Similar results were also observed in a
606 recent ECoG study, which found that movement speed was predominately represented
607 during arm reaching in humans (Hammer et al., 2016). The predominance of speed
608 encoding was interpreted in the context of a model in which the summed activity of many
609 velocity-tuned neurons with random directional tuning resembles speed tuning. Thus it
610 may be the case that individual vSMC neurons are actually representing mostly velocity,
611 but the summed activity observed with ECoG electrodes reflects the magnitude of
612 movement without direction (i.e. speed).

613 By studying vowels, we were able to examine the dynamics of kinematic
614 encoding that are associated with movements to specific vocal tract articulators. We
615 found that articulator kinematics were encoded around the time of movement onset
616 and/or offset, but not while the vocal tract configuration was being held to maintain the
617 vowel. Encoding of articulator kinematics only during movement onset and offset
618 suggests that control of speech articulators is accomplished primarily through control of
619 changes to the plant, rather than moment-to-moment maintenance of the vocal tract
620 configuration. This is consistent with models of speech production that utilize changes to

621 the plant as the primary mechanism by which sensorimotor cortex receives input from,
622 and sends commands to, the vocal tract (Houde and Nagarajan, 2011; Tourville and
623 Guenther, 2011). Furthermore, these dynamics have been observed in studies with
624 analogous behavior from different body parts, including arm reaching. These studies have
625 found individual neurons in motor cortex that exhibit transient firing patterns, where
626 firing rates are high around movement onset and offset (Crammond and Kalaska, 1996;
627 Shalit et al., 2012; Arce et al., 2013; Shadmehr, 2017).

628 We also found a subset of electrodes that exhibited sustained neural activity
629 during the steady-state portion of the vowel which was not correlated with any measured
630 kinematic features. Instead, we found that the duration of the sustained activity correlated
631 well with trial-by-trial vowel length. At a minimum, this suggests that this sustained
632 activity co-varies with whether the subject is vocalizing. One possibility is that sustained
633 activity represents an articulatory parameter that has little variability in our task, such as
634 respiration. However, a more intriguing possibility is that sustained activity may
635 represent a non-specific signal for holding the vocal tract configuration, which does not
636 directly encode the articulatory kinematics like position. Such a signal combined with the
637 onset/offset encoding of kinematics may provide sufficient information for encoding the
638 observed behavior. Further studies utilizing tasks with more variability in manner of
639 articulation are necessary to resolve these possibilities.

640 It is important to emphasize that these analyses focus on the neural representation
641 of the supra-laryngeal articulators. While the movements of these articulators are critical
642 to the production of vowels, the lower vocal tract (e.g. larynx, pharynx, and diaphragm)
643 is also necessary to produce voiced sounds. It is likely that sub-regions of vSMC are

involved in the control of the lower vocal tract (Brown et al., 2008; Bouchard et al., 2013; Conant et al., 2014), but the limitations of our vocal tract kinematic monitoring system and the lack of across-trial variability in pitch and intensity preclude a detailed examination of the representations of these articulators in the present experiment.

Further, we are not able to make evaluate whether the activity we observe is due to feed-forward signals originating in vSMC, or sensory feedback signals. Our models performed optimally at a neural-leading lag of approximately 100ms, implying that the representations we observed were driven more by feed-forward activity. However, the relatively simple movements examined here exhibit temporal auto-correlation, which makes it difficult to dissociate feed-forward activity from feedback. Examining speech tasks with faster, less stereotyped movements (e.g. naturally produced words or sentences) would make it possible to disentangle feed-forward and feedback signals, and is an interesting and important future direction (Chang et al., 2013; Greenlee et al., 2013; Kingyon et al., 2015; Behroozmand et al., 2016; Li et al., 2016; Cao et al., 2017).

Finally, while we observed qualitatively similar results across patients regardless of the hemisphere in which the electrodes were implanted, three of the four participants had grid placements on the right (non-dominant) hemisphere. It is presently unknown whether there are differences in the representation of articulator movements between left and right hemisphere, and the results presented here may not fully address the extent to which such differences exist.

We found that the representation of spoken vowels in vSMC is directly explained by the movements of speech articulators. The encoding of multiple kinematic parameters is present for the articulators, most prominently speed. Articulator kinematic encoding

667 was primarily observed at the onset and offset of vowel production and not while the
668 vowel is being held. Together, these findings provide insight into how neural activity in
669 primary sensorimotor cortex results in the precise production of vowels. Future work will
670 address how these encoding properties operate in the context of natural continuous
671 speech.

672

673 **References**
674
675 Alfonso PJ, Baer T (1982) Dynamics of Vowel Articulation. Lang Speech 25:151–173.
676 Arce FI, Lee J-C, Ross CF, Sessle BJ, Hatsopoulos NG (2013) Directional information
677 from neuronal ensembles in the primate orofacial sensorimotor cortex. J
678 Neurophysiol 110:1357–1369.
679 Ashe J, Georgopoulos AP, Medical VA (1994) Movement Parameters and Neural
680 Activity in Motor Cortex and Area 5. Cereb Cortex 6:590–600.
681 Atal BS, Chang JJ, Mathews M V, Tukey JW (1978) Inversion of articulatory-to-acoustic
682 transformation in the vocal tract by a computer-sorting technique. J Acoust Soc Am
683 63:1535–1553.
684 Behroozmand R, Oya XH, Nourski K V, Kawasaki H, Larson CR, Brugge JF, Howard
685 MA, Greenlee JDW (2016) Neural Correlates of Vocal Production and Motor
686 Control in Human Heschl's Gyrus. J Neurosci 36:2302–2315.
687 Boersma P (2001) Praat, a system for doing phonetics by computer. Glot Int 5:341–345.
688 Bouchard KE, Chang EF (2014) Control of Spoken Vowel Acoustics and the Influence of
689 Phonetic Context in Human Speech Sensorimotor Cortex. J Neurosci 34:12662–
690 12677.
691 Bouchard KE, Conant DF, Anumanchipalli GK, Dichter B, Chaisanguanthum KS,
692 Johnson K, Chang EF (2016) High-resolution, non-invasive imaging of upper vocal
693 tract articulators compatible with human brain recordings. PLoS One 11:1–30.
694 Bouchard KE, Mesgarani N, Johnson K, Chang EF (2013) Functional organization of
695 human sensorimotor cortex for speech articulation. Nature 495:327–332.
696 Brown S, Laird AR, Pfordresher PQ, Thelen SM (2009) The somatotopy of speech:
697 Phonation and articulation in the human motor cortex. Brain Cogn 70:31–41.
698 Cao L, Thut G, Gross J (2017) The role of brain oscillations in predicting self-generated
699 sounds. Neuroimage 147:895–903.
700 Chang EF, Niziolek C a, Knight RT, Nagarajan SS, Houde JF (2013) Human cortical
701 sensorimotor network underlying feedback control of vocal pitch. Proc Natl Acad
702 Sci U S A 110:2653–2658.
703 Conant D, Bouchard KE, Chang EF (2014) Speech map in the human ventral sensory-
704 motor cortex. Curr Opin Neurobiol 24:63–67.
705 Crammond D, Kalaska JF (1996) Differential relation of discharge in primary motor
706 cortex and premotor cortex to movements versus actively maintained postures

during a reaching task. Exp brain Res 108:45–61.
Crone NE, Hao L, Hart J, Boatman D, Lesser RP, Irizarry R, Gordon B (2001) Electrocorticographic gamma activity during word production in spoken and sign language. Neurology 57:2045–2053.
Crone NE, Miglioretti DL, Gordon B, Lesser RP (1998) Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. Brain 121:2301–2315.
Donoho D, Stodden V (2004) When does non-negative matrix factorization give a correct decomposition into parts? Proc Adv Neural Inf Process Syst 16:1141–1148.
Edwards E, Nagarajan SS, Dalal SS, Canolty RT, Kirsch HE, Barbaro NM, Knight RT (2010) Spatiotemporal imaging of cortical activation during verb generation and picture naming. Neuroimage 50:291–301.
Fukuda M, Rothermel R, Juhász C, Nishida M, Sood S, Asano E (2010) Cortical gamma-oscillations modulated by listening and overt repetition of phonemes. Neuroimage 49:2735–2745.
Georgopoulos a P, Kalaska JF, Caminiti R, Massey JT (1982) On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. J Neurosci 2:1527–1537.
Georgopoulos AP, Caminiti R, Kalaska JF (1984) Static Spatila Effects in Motor Cortex and Area 5: Quantitative Relations in a Two-Dimensional Space. Exp brain Res:446–454.
Georgopoulos AP, Schwarz AB, Ketiner RE (1986) Neuronal population coding of movement direction. Science (80-) 233:1416–1419.
Grabski K, Lamalle L, Vilain C, Schwartz J-L, Vallée N, Tropres I, Baciu M, Le Bas J-F, Sato M (2012) Functional MRI assessment of orofacial articulators: neural correlates of lip, jaw, larynx, and tongue movements. Hum Brain Mapp 33:2306–2321.
Gracco VL, Lofqvist A (1994) Speech motor coordination and control: evidence from lip, jaw, and laryngeal movements. J Neurosci 14:6585–6597.
Greenlee JDW, Behroozmand R, Larson CR, Jackson AW, Chen F, Hansen DR, Oya H, Kawasaki H, Howard M a (2013) Sensory-motor interactions for vocal pitch monitoring in non-primary human auditory cortex. PLoS One 8:e60783.
Hamilton LS, Edwards E, Chang EF (2017) Parallel streams define the temporal dynamics of speech processing across human auditory cortex. bioRxiv.
Hammer J, Pistohl T, Fischer J, Kr P, Tomá M, Marusic P, Schulze-bonhage A, Aertsen A, Ball T (2016) Predominance of Movement Speed Over Direction in Neuronal Population Signals of Motor Cortex : Intracranial EEG Data and A Simple Explanatory Model. Cereb Cortex 26:2863–2881.
Harshman R, Ladefoged P, Goldstein L (1977) Factor analysis of tongue shapes. J Acoust Soc Am 62:693–713.
Herff C, Heger D, De pesters A, Telaar D, Brunner P, Schalk G, Schultz T (2015) Brain-to-text: Decoding spoken phrases from phone representations in the brain. Front Neuroeng 8.
Hesselmann V, Sorger B, Lasek K, Guntinas-Lichius O, Krug B, Sturm V, Goebel R, Lackner K (2004) Discriminating the cortical representation sites of tongue and up movement by functional MRI. Brain Topogr 16:159–167.
Hillenbrand J, Getty L a., Clark MJ, Wheeler K (1995) Acoustic characteristics of

753 American English vowels. J Acoust Soc Am 97:3099–3111.
754 Houde JF, Nagarajan SS (2011) Speech production as state feedback control. Front Hum
755 Neurosci 5:1–14.
756 Huang CS, Sirisko MA, Hiraba H, Murray GM, Sessle J (2013) Organization of the
757 primate face motor cortex as revealed by intracortical microstimulation and
758 electrophysiological identification of afferent inputs and corticobulbar projections
759 Organization of the Primate Face Motor Cortex as Revealed by Intracortical. J
760 Neurophysiol 59:796–818.
761 Johnson K, Ladefoged P, Lindau M (1993) Individual differences in vowel production. J
762 Acoust Soc Am 94:701–714.
763 Kellis S, Miller K, Thomson K, Brown R, House P, Greger B (2010) Decoding spoken
764 words using local field potentials recorded from the cortical surface. J Neural Eng
765 7:56007–56016.
766 Kingyon J, Behroozmand R, Kelley R, Oya H, Kawasaki H, Narayanan NS, Greenlee
767 JDW (2015) High-gamma band fronto-temporal coherence as a measure of
768 functional connectivity in speech motor control. Neuroscience 305:15–25.
769 Kumar V, Croxson PL, Simonyan K (2016) Structural Organization of the Laryngeal
770 Motor Cortical Network and Its Implication for Evolution of Speech Production. J
771 Neurosci 36:4170–4181.
772 Ladefoged P, Johnson K (2011) A Course in Phonetics. Boston: Cengage Learning.
773 Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix
774 factorization. Nature 401:788–791.
775 Leonard MK, Cai R, Babiak MC, Ren A, Chang EF (2016) Brain & Language The peri-
776 Sylvian cortical network underlying single word repetition revealed by
777 electrocortical stimulation and direct neural recordings. Brain Lang:1–15.
778 Leuthardt EC, Gaona C, Sharma M, Szrama N, Roland J, Freudenberg Z, Solis J,
779 Breshears J, Schalk G (2011) Using the electrocorticographic speech network to
780 control a brain-computer interface in humans. J Neural Eng 8.
781 Li M, Kambhamettu C, Stone M (2004) Automatic contour tracking in ultrasound
782 images. Clin Linguist Phon 19:545–554.
783 Li W, Chen Z, Yan N, Jones JA, Guo Z, Huang X (2016) Temporal Lobe Epilepsy Alters
784 Auditory-motor Integration For Voice Control. Sci Rep 6:1–13.
785 Lofqvist A, Gracco V (1999) Interarticulator programming in VCV sequences: Lip and
786 tongue movements. J Acoust Soc Am 105:1864–1876.
787 Lotze M, Seggewies G, Erb M, Grodd W, Birbaumer N (2000) The representation of
788 articulation in the primary sensorimotor cortex. Neuroreport 11:2985–2989.
789 Maeda S (1990) Compensatory articulation during speech: Evidence from the analysis
790 and synthesis of vocal-tract shapes using an articulatory model. In: Speech
791 production and speech modeling, pp 131–149. Netherlands: Springer.
792 Moran DW, Schwartz AB (1999) Motor Cortical Representation of Speed and Direction
793 During Reaching. J Neurophysiol 82:2676–2692.
794 Mugler EM, Patton JL, Flint RD, Wright Z a, Schuele SU, Rosenow J, Shih JJ,
795 Krusienski DJ, Slutzky MW (2014) Direct classification of all American English
796 phonemes using signals from functional speech motor cortex. J Neural Eng
797 11:35015.
798 Muller L, Hamilton LS, Edwards E, Bouchard KE, Chang EF (2016) Spatial resolution

dependence on spectral frequency in human speech cortex electrocorticography. J Neural Eng 13.
Paninski L, Fellows MR, Hatsopoulos NG, Donoghue JP (2004) Spatiotemporal tuning of motor cortical neurons for hand position and velocity. J Neurophysiol 91:515–532.
Pei X, Barbour DL, Leuthardt EC, Schalk G (2011) Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. J Neural Eng 8.
Penfield W, Boldrey E (1937) Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. Brain 60:389–443.
Penfield W, Roberts L (1959) Speech and brain mechanisms. Princeton.
Perkell JS, Matthies ML, Svirsky M a, Jordan MI (1993) Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: a pilot "motor equivalence" study. J Acoust Soc Am 93:2948–2961.
Perkell JS, Nelson WL (1985) Variability in production of the vowels /i/ and /a/. Acoust Soc Am 77:1889–1895.
Petersen S, Fox PT, Posner M, M M, Raichle M (1988) Positron emissiion tomographic studies of the cortical anatomy of single-word Processing.
Ray S, Maunsell JHR (2011) Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLoS Biol 9.
Shadmehr R (2017) Distinct neural circuits for control of movement vs . holding still. J Neurophysiol 117:1431–1460.
Shalit U, Zinger N, Joshua M, Prut Y (2012) Descending Systems Translate Transient Cortical Commands into a Sustained Muscle Activation Signal. Cereb Cortex 22:1904–1914.
Simonyan XK, Ackermann H, Chang EF, Greenlee JD (2016) New Developments in Understanding the Complexity of Human Speech Production. J Neurosci 36:11440–11448.
Tourville J a, Guenther FH (2011) The DIVA model: A neural theory of speech acquisition and production. Lang Cogn Process 26:952–981.
Ueda Y, Hamakawa T, Sakata T, Hario S, Watanabe A (2007) A real-time formant tracker based on the inverse filter control method. Acoust Sci Technol 28:271–274.
Watanabe a. (2001) Formant estimation method using inverse-filter control. IEEE Trans Speech Audio Process 9:317–326.

Figure Legends

Figure 1: Experimental setup and articulator monitoring. **A**, Schematic of the articulatory tracking system. A video camera placed in front of the subject recorded the movements of the lips while an ultrasound transducer under the jaw captured the tongue contour. **B**, Example images of the from the video (top) and ultrasound (bottom) imaging during production of the corner vowels /a/, /i/, and /u/. The lips and tongue contour were extracted from these images, and the resulting binary masks are shown in color on top of the raw images. **C**, Magnetic resonance imaging (MRI) reconstruction of the brains of

the four subjects included in the study. Co-registered ECoG electrodes are plotted on the cortical surface, with dark points denoting electrodes over vSMC.

Figure 2: Articulatory and acoustic behavior correlates with single electrode vSMC neural activity. **A**, Prototypical articulator positions for the corner vowels /a/, /i/, and /u/. **B**, Average (± s.e.m.) timecourses of measured articulator displacements during production of the corner vowels. For illustration, the kinematic parameter of vertical position is shown, however both vertical and horizontal measurements are used for subsequent analyses. **C**, Average formant values (F1-F4) for the corner vowels during the same productions as in (B). **D**, HG activity in three example vSMC electrodes. Each electrode was selected to be representative for the articulator shown in the top subplot (median configurations shown for back tongue, front tongue, and lips). Trials for each electrode are binned by the displacement of the articulator which best correlates with the HG values at acoustic onset (± 100 ms). Yellow bars mark timepoints of significant difference between the bins (e1: $F_{(2,75)} < 5.1$, $p<0.01$; e2: $F_{(2,102)} < 4.9$, $p<0.01$; e3: $F_{(2,105)} < 4.8$, $p<0.01$; ANOVA). **E**) HG activity in the same electrodes binned according to formant values (F1/F2 ratio).

Figure 3: vSMC activity primarily encodes speech articulators. **A**, Performance of encoding models predicting vSMC HG using vowel identify, acoustic formants, articulator kinematics, or all three. Articulator kinematics explain vSMC activity better than vowel identity and acoustic formants (*** $p<0.01$; Wilcoxon rank sum). Furthermore, the joint model does not explain more variance than the kinematic model alone, indicating that the vowel identity and acoustic formant models are likely driven by variance shared with the kinematics. **B**, Electrodes over vSMC from three right hemisphere subjects were warped onto a common brain and color-coded according to articulator selectivity. Empty circles mark electrodes with no significant selectivity for any articulator; black electrodes are selective for more than one articulator, and blue, red, and green electrodes are selective for front tongue, back tongue, or lips, respectively.

Figure 4: vSMC activity reflects articulator kinematic organization of vowels. **A-C**, Multidimensional Scaling (MDS) representations of (**A**) acoustic formants, (**B**) articulator position, and (**C**) vSMC HG activity. Each letter marks the position of the median production of the indicated vowel in MDS space across all subjects. The relative organization of the vowels is similar across spaces. For example, the low-back vowel /a/ is always near the mid-back vowel /ʌ/, but far from the high-front vowel /i/. **D-F**, Hierarchical clustering (left) and confusion matrices (right) derived from the pairwise distances between vowels in the MDS spaces of (D) acoustic formants, (E) articulator position, and (F) vSMC HG.

Figure 5: Representations of position, speed, velocity, and acceleration kinematic features over time. **A-D**, Top: Example kinematic parameters of position (**A**), speed (**B**), velocity (**C**), and acceleration (**D**) for all utterances of /a/ for one subject. Thin lines mark individual trials, while the thick line is the across-trial average. Bottom: Performance of encoding models predicting vSMC HG from articulator position (**A**), speed (**B**), velocity (**C**), or acceleration (**D**). vSMC electrodes with significant

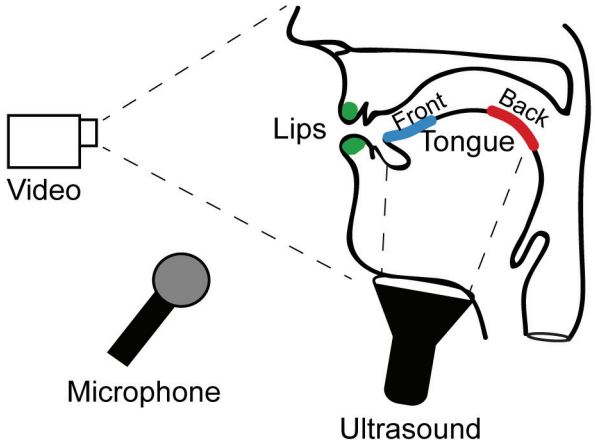
performance are marked by red dashes at the time peak encoding performance. Electrodes in A-D are plotted in order of their peak encoding times in the joint model (F). Vertical black lines mark the onset (solid) and offset (dashed) of vowel acoustics. **E**, Comparison of the number of significant electrodes (black) and average peak performance (grey) of position, speed, velocity, and acceleration encoding models. Speed is significantly encoded at more electrodes, with a higher average model performance (*** $p<0.01$; Wilcoxon rank sum). **F**, Performance of encoding models predicting vSMC HG from all articulator kinematics jointly. **G**, Average performance across significant electrodes for the joint and independent parameter models.

Figure 6: Time course of kinematic parameter decoding. **A-D**, Performance of decoding models predicting articulator position (**A**), speed (**B**), velocity (**C**), or acceleration (**D**) from HG at all vSMC electrodes. Features are averaged across subjects and within articulators. Black lines denote the onset and offset of vowel acoustics.

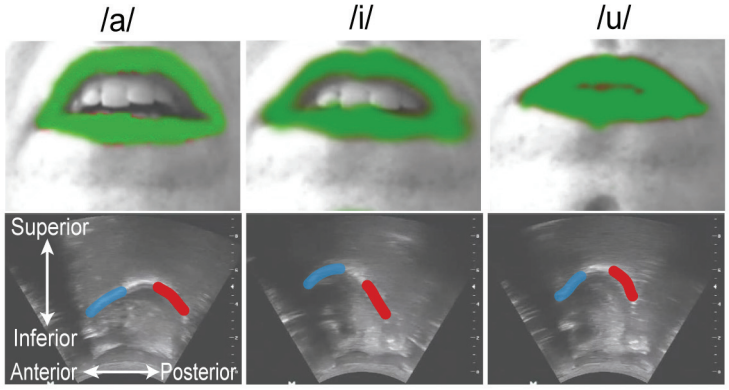
Figure 7: Relationship of vSMC HG dynamics and kinematic encoding. **A**, HG activity at several example electrodes illustrating the diverse dynamics during the same behavior, especially during the time period when the vowel is being held. **B**, The first two NMF bases extracted from HG dynamics across all electrodes. These bases recapitulate the key differences in dynamics seen in the example electrodes, and serve as an unbiased quantification of the HG dynamics seen across vSMC. **C**, HG activity at all vSMC electrodes ordered by the ratio of NMF bases used to reconstruct their activity (*(NMF1-NMF2)/(NMF1+NMF2)*). Arrow denotes the example electrode used in Figure 8. **D**, Average HG activity across all sustained (NMF1>NMF2) electrodes (solid) plotted alongside the average encoding performance (dashed) across time. During the steady state of the vowel, there is elevated activity, but almost no significant encoding of articulator kinematics.

Figure 8: Sustained HG activity is related to vowel duration. **A**, An example electrode that is characterized by sustained HG activity. Trials are ordered by vowel production duration. Red lines mark the onset and offset of vowel acoustics. **B**, The duration of the sustained HG activity at each trial is plotted against the trial duration for all sustained electrodes. Larger grey markers denote observations from the example electrode in A.

A Articulator Tracking Setup



B Extracted Articulators



C ECoG Electrodes