Systems/Circuits

Control of Spoken Vowel Acoustics and the Influence of Phonetic Context in Human Speech Sensorimotor Cortex

Kristofer E. Bouchard[1,2,3] **and Edward F. Chang**[1,2,3,4]

[1]Departments of Neurological Surgery and Physiology, University of California, San Francisco, San Francisco, California 94143-0112, [2]Center for Integrative Neuroscience, University of California, San Francisco, San Francisco, California 94158, [3]Center for Neural Engineering and Prosthesis, University of California, San Francisco and University of California, Berkeley, Berkeley, California 94720-3370, and [4]UCSF Epilepsy Center, University of California, San Francisco, San Francisco, California 94143

Speech production requires the precise control of vocal tract movements to generate individual speech sounds (phonemes) which, in turn, are rapidly organized into complex sequences. Multiple productions of the same phoneme can exhibit substantial variability, some of which is inherent to control of the vocal tract and its biomechanics, and some of which reflects the contextual effects of surrounding phonemes ("coarticulation"). The role of the CNS in these aspects of speech motor control is not well understood. To address these issues, we recorded multielectrode cortical activity directly from human ventral sensory-motor cortex (vSMC) during the production of consonant-vowel syllables. We analyzed the relationship between the acoustic parameters of vowels (pitch and formants) and cortical activity on a single-trial level. We found that vSMC activity robustly predicted acoustic parameters across vowel categories (up to 80% of variance), as well as different renditions of the same vowel (up to 25% of variance). Furthermore, we observed significant contextual effects on vSMC representations of produced phonemes that suggest active control of coarticulation: vSMC representations for vowels were biased toward the representations of the preceding consonant, and conversely, representations for consonants were biased toward upcoming vowels. These results reveal that vSMC activity for phonemes are not invariant and provide insight into the cortical mechanisms of coarticulation.

Key words: **ECoG; motor control; sensorimotor cortex; sequences; speech**

Introduction

Communication through spoken language relies on a speaker's ability to articulate sounds that are identifiable to a listener as the meaningful units—consonants and vowels—of a language (Levelt, 1999; MacNeilage, 2011). To maximize clarity in vocal communication, a speaker presumably generates motor commands that differ greatly across distinct phonemes but differ little within repeated renditions of a single phoneme. Indeed, the pattern of acoustic parameters of vowels is more distinct across different vowels than within the same vowel (Maddieson and Disner, 1984). In fluent speakers, the ventral half of the sensory-motor cortex (vSMC) is thought to exert precise control of the vocal tract—control that has likely been optimized through evolution, learning, and extensive practice (Levelt, 1999; Brown et al., 2009; Takai et al., 2010; MacNeilage, 2011; Bouchard et al., 2013). Despite this, multiple utterances of a given phoneme by the same speaker are not identical (Maddieson and Disner, 1984; Perkell and Nelson, 1985; Gracco and Abbs, 1986). Some variability in the production of the same phoneme is likely inherent to repeated production of any behavior, but speech variability also reflects the surrounding phonetic context (Lindblom, 1963; Hillenbrand et al., 1995). To what degree different kinds of speech variability are generated in vSMC is poorly understood.

Addressing this issue is important for understanding speech motor control, but requires analysis of vSMC activity and speech production on a trial-by-trial basis (Churchland et al., 2006a; Sober et al., 2008), which can be difficult to achieve with traditional noninvasive human imaging. However, direct cortical recordings through electrocorticography (ECoG) have sufficient signal-to-noise properties to resolve single-trial activity (Edwards et al., 2010; Leuthardt et al., 2011; Pei et al., 2011). Furthermore, although previous studies have shown that vSMC represents vocal tract articulators (e.g., lips, tongue, jaw, larynx), the internal location of the vocal tract makes it difficult to directly measure its movements (Brown et al., 2009; Bouchard et al., 2013). However, the vocal tract shape is directly reflected by the produced acoustics, especially vowel formants, which are easily studied (Ladefoged and Johnson, 2011). Therefore, we used ECoG to study the relationship between vSMC cortical activity and speech acoustics on a trial-by-trial basis.

Received March 26, 2014; revised July 24, 2014; accepted July 27, 2014.

Author contributions: E.F.C. designed research; E.F.C. performed research; K.E.B. analyzed data; K.E.B. and E.F.C. wrote the paper.

E.F.C. was funded by National Institutes of Health Grants R00-NS065120, DP2-OD00862, and R01-DC012379, and by the Ester A. and Joseph Klingenstein Foundation. We thank Angela Ren for technical help with data collection and preprocessing, Miranda Babiak for audio transcription, and C. Niziolek and K. Chaisanguanthum for helpful comments on this manuscript.

Correspondence should be addressed to Edward F. Chang, Department of Neurological Surgery, University of California, San Francisco, San Francisco, CA 94143-0112. E-mail: changed@neurosurg.ucsf.edu.

K. E. Bouchard's present address: Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720-8150.

DOI:10.1523/JNEUROSCI.1219-14.2014

 0270-6474/14/3412662-16$15.00/0

Our goals were to determine which acoustic features in vowel production are most tightly controlled by vSMC activity, and how surrounding phonemes influenced this control. We examined the degree to which vSMC activity was predictive of acoustics across the production of different vowels ("across-vowel"), as well as the utterance-to-utterance variability in the production of the same vowel ("within-vowel"). Furthermore, because phonemes are rarely produced in isolation, but rather in the context of phoneme sequences, we then focused on an important source of speech variability that arises from the influence of surrounding phonemes, known as "coarticulation" (Hardcastle and Hewlett, 2006). The role of cortex in coarticulation is a central question because it directly addresses the representational nature of phonemes in speech production (Fowler, 1980; Whalen, 1990; Guenther, 1995; Ostry et al., 1996; Guenther et al., 2006; Hardcastle and Hewlett, 2006; Golfinopoulos et al., 2010).

Materials and Methods

The experimental protocol was approved by the Human Research Protection Program at the University of California, San Francisco.

Subjects and experimental task. Three native English-speaking human participants underwent chronic implantation of a high-density, subdural electrocortigraphic (ECoG) array. Our recordings were from the language dominant hemisphere in each patient (as determined with the Wada carotid intra-arterial amybarbital injection). The language dominant hemisphere was left in two subjects and right in one subject, and we did not find clear left/right differences. Participants gave their written informed consent before the day of surgery. Each participant read aloud consonant-vowel syllables (CVs) composed of 18–19 consonants (19 consonants for two participants, 18 consonants for one participant), followed by one of three vowels. Each CV was produced between 15 and 100 times total.

Anatomical location of vSMC. We focused our analysis on the vSMC (the "speech" portion of the sensory-motor cortex). vSMC is anatomically defined as the ventral portions of the precentral and postcentral gyri, as well as the gyral formation at the ventral termination of the central sulcus, known as the subcentral gyrus. Visual examination of coregistered CT and MR scans indicate that the ECoG grid in each patient covered the spatial extent of vSMC of each patient (Figs. 1*c*, 2). The precentral gyrus is thought to be functionally subdivided into a "premotor" and a "primary motor" cortex. However, our task (CV list reading) and the spatial resolution of our ECoG recordings do not allow sufficient sample size or task parameters to meaningfully investigate functional differences between the two. No obvious differences were observed based on hemisphere, perhaps reflecting the language dominance of the grid placements in each subject.

Data acquisition and signal processing. Cortical-surface field potentials were recorded with ECoG arrays and a multichannel amplifier optically connected to a digital signal processor [Tucker-Davis Technologies (TDT)]. The spoken syllables were recorded with a microphone, digitally amplified, and recorded in-line with the ECoG data. ECoG signals were acquired at 3052 Hz. The microphone audio signal was acquired at 22 kHz.

The time series from each channel was visually and quantitatively inspected for artifacts or excessive noise (typically 60 Hz line noise). Artifactual recordings were excluded from analysis, and the raw recorded ECoG signals of the remaining channels were then common average referenced. For each channel, the time-varying analytic amplitude was extracted from eight bandpass filters [Gaussian filters, logarithmically increasing center frequencies (70–150 Hz), and semilogarithmically increasing bandwidths] with the Hilbert transform. The high-gamma (High-γ) power was then calculated by averaging the analytic amplitude across these eight bands, and then this signal was down-sampled to 200 Hz. High-γ power was z-scored relative to the mean and SD of baseline data for each channel. Throughout, when we speak of High-γ power, we refer to this z-scored measure, denoted below as Hγ.

Acoustic feature extraction. The recorded speech signal was transcribed off-line using WaveSurfer (http://www.speech.kth.se/wavesurfer/). We measured the vowel pitch (F_0) and formants, F_1–F_4, as a function of time for each utterance of a vowel using an inverse filter method (Watanabe, 2001; Hamakawa et al., 2007). Briefly, the signal is inverse filtered with an initial estimate of F_2 and then the dominant frequency in the filtered signal is used as an estimate of F_1. The signal is then inverse filtered again, this time with an inverse of the estimate of F_1, and the output is used to refine the estimate of F_2. This procedure is repeated until convergence and is also used to find F_3 and F_4. The inverse filter method converges on very accurate estimates of the vowel formants, without making assumptions inherent in the more widely used linear predictive coding method. For the extraction of F_0 (pitch), we used standard autocorrelation methods.

Correlation coefficient. We used the Pearson product-moment correlation coefficient (R) to quantify the linear relationship between two variables (x and y):

$$R(x,y) = \frac{COV(x,y)}{\sigma_x \sigma_y}, \tag{1}$$

where σ_x and σ_y are the sample SDs of x and y, respectively.

Acoustic analysis. The mean acoustic feature values were extracted from the central 20% of each vowel utterance, and $\log_{10}$ of these values was used for subsequent analysis. We quantified the discriminability of the cardinal vowels ($V \in$ [/a/, /i/, /u/]) based on individual features and feature ratios (F_i, $i \in$ [0 1 2 3 4 1/0 2/1 3/2]) using the d' metric. d' is the difference between the mean of two distributions divided by the square root of the product of their SDs:

$$d'(F_i^v, F_i^{v'}) = \frac{\mu_{F_i^v} - \mu_{F_i^{v'}}}{\sqrt{\sigma_{F_i^v}\ \sigma_{F_i^{v'}}}}. \tag{2}$$

Here, F_i^V, $F_i^{V'}$ denote the values of feature i for vowel V and V'. To summarize the discriminability of a feature, we calculated d' for the three vowel comparisons (e.g., /a/ vs /i/), and averaged these values across comparisons.

Principal components analysis and cortical features. Principal components analysis (PCA) was performed on the set of all vSMC electrodes for dimensionality reduction and orthogonalization. This also ensures that the matrices in the calculation of least mean squared error estimators (from regressions below) were well scaled. PCA was performed independently for each nonoverlapping 10 ms window preceding the acoustic measurement window. First, for each electrode (e_j of which there are n) and syllable utterance (s, of which there are m), we calculated the mean high-gamma activity in 10 ms windows with a nonoverlapping two-sample moving average of the Hγ with time lag τ. The H$\gamma_j(\tau, s)$ values were used as entries in the $n \times m$ data matrix $\mathbf{D}$, with rows corresponding to channels (of which there are n) and columns corresponding to the number of utterances within a recording block (of which there are m). Each electrode's activity was z-scored across utterances to normalize neural variability across electrodes. PCA was performed on the $n \times n$ covariance matrix $\mathbf{Z}$ derived from $\mathbf{D}$. The singular-value decomposition of $\mathbf{Z}$ was used to find the eigenvector matrix $\mathbf{M}$ and associated eigenvalues λ. The principal components (PCs) derived in this way serve as a spatial filter of the electrodes, with each electrode e_j receiving a weighting in PC_i equal to m_{ij}, the i–jth element of $\mathbf{M}$, the matrix of eigenvectors. Because we were interested in examining whether vSMC activity could be used to predict both large acoustic variability (across-vowel variability) and smaller acoustic variability (within-vowel variability), we included the leading 40 eigenvectors in our analysis. For each utterance (s), we projected the vector H$\gamma(\tau,s)$ of high-gamma activity across electrodes into the leading 40 eigenvectors ($\mathbf{M}^{40}$):

$$\Psi(\tau, s) = \mathbf{M}^{40} \cdot \mathrm{H}\boldsymbol{\gamma}(\tau, s). \tag{3}$$

It is important to emphasize that the approach described above identifies principal components (spatial filters) derived only from the spatial structure of the data (structure of Hγ across electrodes); the temporal structure of the Hγ population does not enter into $\mathbf{M}$ in any way. Thus, the

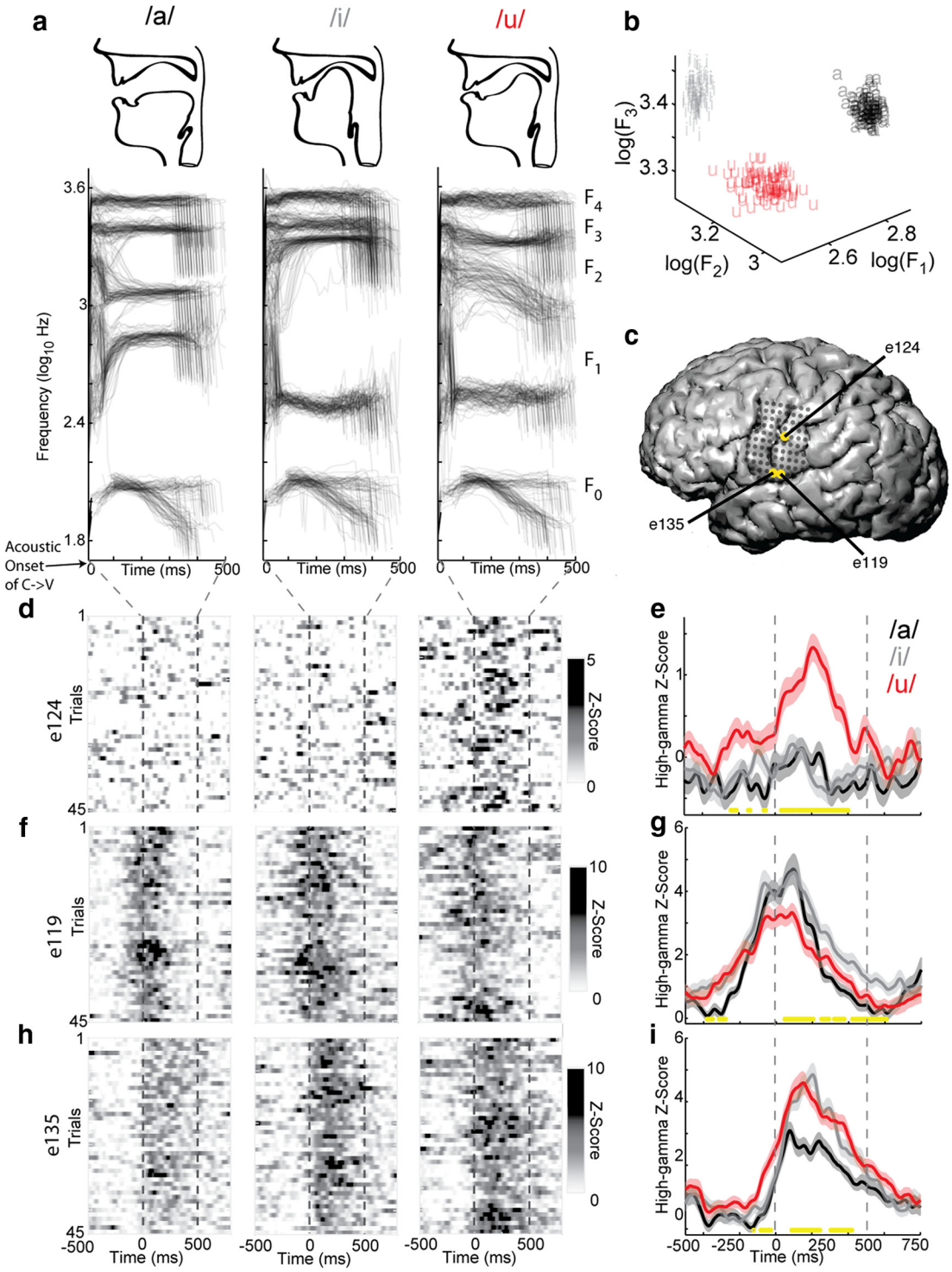

Figure 1. Single-utterance vowel acoustics and vSMC neural activity. ***a***, Schematic vocal tracts (top) and measured formant traces (bottom; F_0:F_4) from >100 utterances each of /a/ (left), /i/ (center), and /u/ (right) from one recording session. Formant values were extracted from the central one-fifth of each vowel utterance. Time point 0 is the acoustic onset of the consonant-to-vowel transition. ***b***, Scatter plot of the vowels in the space formed by F_1, F_2, F_3 ($\log_{10}$ scale) extracted from the traces in ***a***. The vowels /a/ (black), /i/ (gray), and /u/ (red) occupy distinct regions of the formant space. ***c***, Lateral view of the left hemisphere from the same participant with location of electrodes over the ventral half of the sensory-motor cortex highlighted by gray dots. Yellow dots correspond to electrodes in ***d–i***. ***d–i***, Neural activity (high-gamma amplitude, *z*-scored) versus time from the electrodes demarcated in ***c*** during speech production. Heat maps in ***d***, ***f***, and ***h*** display single-trial activity during multiple productions of /ja/ (left), /ji/ (center), and /ju/ (right), while plots in ***e***, ***g***, and ***i*** overlay the across-trial mean ± SE for /ja/ (black), /ji/ (gray), and /ju/ (red). Dashed vertical lines in ***d–i*** demarcate the 500 ms time window displayed in ***a***. Yellow points in ***e***, ***g***, and ***i*** demarcate times at which there is a difference between vowels (rank-sum test, $p < 10^{-3}$).

PCs are completely local in time. Across the three subjects, 40 PCs accounted for ~90% of the variance. We observed that increasing the number of PCs from 20 to 40 resulted in increased decoding performance, particularly for the within-vowel analysis. This can be understood because PCA was performed on the data across all vowels, and the variability across vowels is larger than the variability within a vowel. Since PCA organizes eigenvectors according to decreasing amount z of variance accounted for in the data, intuitively, structure in the neural data

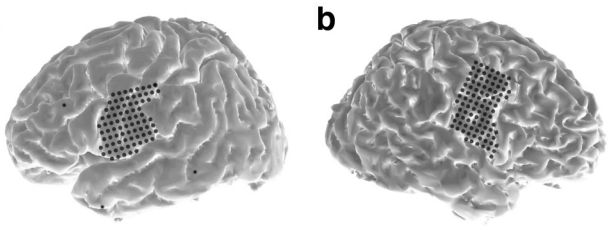

Figure 2. Location of electrodes over vSMC in other participants. ***a***, Lateral view of the left hemisphere from a second participant with location of electrodes over the vSMC highlighted by red dots. ***b***, Lateral view of the left hemisphere from a third participant with location of electrodes over the vSMC highlighted by red dots.

associated with the within-vowel variability will be captured by eigenvectors with smaller eigenvalues.

Acoustic feature decoding model. For each nonoverlapping 10 ms time window (τ) preceding the behavioral measurement, the $\Psi(\tau,s)$ (Eq. 3) served as the basis for training and testing optimal linear predictors of single-trial vowel acoustic features using a fivefold cross-validation procedure (described below). We used a simple linear model to predict the acoustic features ($F_i(s)$) for a syllable (s) from $\Psi(\tau, s)$:

$$\hat{F}_i(s) = \beta \cdot \Psi(\tau, s) + \beta_0, \quad (4)$$

where $\hat{F}_i(s)$ is the best linear estimate of $F_i(s)$ based on the cortical features. The vector of weights β that minimized the mean squared error between $\hat{F}_i(s)$ and $F_i(s)$ was found through multilinear regression and cross-validation with regularization (see below).

Cross-validation and regularization procedure. A cross-validation procedure was used to train and test separate decoding models for within- and across-vowel acoustic features. Separate models were trained/tested for each time point (dt = 10 ms) and recording block. The procedure is as follows. First, to derive null distributions of weights ($\beta^{\star\,\mathrm{rnd}}$) and model performance (R^2_{rnd}), we randomly permuted (200 times) each vowel feature independently relative to Hγ on a trial-by-trial basis, and used an 80/20 cross-validation procedure to derive model weights from training data and model performance on test data. Second, within a 200 iteration bootstrap procedure, random 80% subsets of the data were used to derive linear weights for the models (Eq. 4). From this, we arrived at an estimate of weights ($\beta^{\star\,\mathrm{obs}}$) for each cortical feature predicting acoustic features. We then reduced the dimensionality of the cortical features (Ψ) by comparing the model weights between the observed and randomized datasets to identify cortical features (i.e., PC projections) with weights that were different between the two conditions. Specifically, cortical features (Ψ_j) were retained if the weight magnitudes ($|\beta_j^{\star\,\mathrm{obs}}|$) was greater than the mean + 1 SD of the distribution of weight magnitudes derived from the randomization procedure ($|\beta_j^{\star\,\mathrm{rnd}}|$). Finally, we retrained decoders on the training data based only on this reduced set of cortical features to arrive at optimal weights ($\beta^{\star\,\mathrm{reg}}$) and determined decoding performance (R^2_{reg}) on test data not used in training. This "regularization" resulted in improved decoding performance (up to ~10%) on test data. The choice of threshold (mean + 1 SD of the null distribution) was chosen by visual examination of the weight distributions. An optimization of this threshold may have resulted in better model performance; however, because the chosen threshold resulted in good decoding performance, this optimization was not performed to reduce computational run-time.

The decoding performance for each block and decoding condition was taken as the mean of R^2_{reg} values across bootstrap test samples. This quantifies the expected value of predictive decoding performance across randomly selected training and test samples. The scatter plots shown later in Figures 3*a* and 4*a* were constructed by taking the expected predicted formant values for a given data point from all validation sets that contained that data point (i.e., the average predicted value across different cross-validation randomizations).

We confirmed that the expected value of R^2 under the null hypothesis for our data and procedure was 0 by examining the distributions of R^2_{rnd}. As described above, to derive null distributions of weights ($\beta^{\star\,\mathrm{rnd}}$) and model performance (R^2_{rnd}), we randomly permuted (200 times) each vowel feature independently relative to Hγ on a trial-by-trial basis, and used an 80/20 cross-validation procedure to derive model weights from training data and model performance on test data. Across all blocks, times, and conditions, R^2_{rnd} had a median very close to 0 (median <0.001 for all). We note that, as there are more cortical features in the model used to derive R^2_{rnd} than R^2_{reg}, comparing R^2_{reg} to R^2_{rnd} is a conservative approach for statistical testing. Therefore, we gauged the significance of the across-block distributions of R^2_{reg} for each feature and time window by performing t tests against the null hypothesis of 0. The conclusions of significance were insensitive to different statistical tests.

Analysis of perseverative coarticulation of vowel acoustics. We used a linear model to account for the coarticulation effect of the preceding consonant on the formants of vowels. In agreement with previous studies, we observed that perseverative coarticulation was determined in part by which one of three major oral articulators (lips, coronal tongue, and dorsal tongue) is required for production of the preceding consonant. Therefore, we modeled the effect of perseverative coarticulation for all of the acoustic features of each vowel with a linear model based on a 3×1 binary vector indicating which articulator was engaged by the preceding consonant:

$$\hat{F}_i(s) = \beta \cdot A(s) + \beta_0, \quad (5)$$

This model was fit separately for each vowel, feature, and time during the vowel; performance was quantified by Equation 1: $R^2(\hat{F}_i(s), F_i(s))$. We used a randomization procedure to gauge the expected R^2 values under the null hypothesis. Specifically, we randomly permuted (200 times) each vowel feature independently relative to the articulator vector ($\mathbf{A}(s)$) on a trial-by-trial basis, and used an 80/20 cross-validation procedure to derive null model weights from training data and null model performance on test data.

We removed the linear effects of perseverative coarticulation by calculating the residual formant features:

$$\hat{F}_i^{res}(s) = \hat{F}_i(s) - \hat{F}_i(s). \quad (6)$$

Hence, $F_i^{\mathrm{res}}(s)$ is the best estimate of the formant features unaffected by (the linear effect of) perseverative coarticulation. Linear decoders to predict $F_i^{\mathrm{res}}(s)$ from the cortical features (Ψ) were trained and tested as described above.

Dimensionality reduction. The objective of our dimensionality reduction was to derive a "cortical state-space" to investigate the organization of the vSMC network associated with different consonant-vowel syllables through time. In particular, we wanted to test two related, but not redundant, hypotheses: (1) the identity of adjacent phonemes imparts structure to the state-space representation of activity generating individual phonemes, and (2) that the relative location of single-trial trajectories during one phoneme decays smoothly and overlaps in time with state-space representations of other phonemes. We used specific CV contrasts for this analysis. The state-spaces for specific syllable contrasts designed to test the above hypothesis were derived independently. For the examination of perseverative coarticulation, we contrasted consonants with differing major oral articulators but with the same constriction degree ([/b/ /d/ /g/], [/p/ /t/ /k/], [/w/ /l/ /j/]) transitioning to the different cardinal vowels (e.g., /bu/ vs /du/ vs /gu/). Analogously, for anticipatory coarticulation, we contrasted the different cardinal vowels following each of the consonants (e.g., /ga/ vs /gi/ vs /gu/). Across the three subjects examined here, this resulted in a total of $N = 162$ syllable contrasts for anticipatory coarticulation and $N = 78$ syllable contrasts for perseverative coarticulation.

Specifically, then, the goal of our dimensionality reduction scheme was to find a low-dimensional space derived from single-trial cortical activity that maximized the separability of specific CV contrasts through time. To this end, we devised a two-step dimensionality reduction scheme, in which Gaussian process factor analysis (GPFA; Yu et al., 2009) was followed by linear discriminant analysis (LDA). This approach is similar to that used in previous studies of population neural activity aiming to identify specific axes in state-space (Briggman and Kristan, 2006; Mante et al., 2013). Following dimensionality reduction, we re-

moved DC offsets and differences in scaling across time by *z*-scoring the single-trial distribution of state-space locations across CV contrasts locally at each point in time. This allowed us to examine the temporal profile with which the relative state-space locations at a given time point were correlated with the relative locations at other times. The local *z*-scoring procedure, which is a simple translation and scaling, did not change the qualitative structure of the data, but made the quantification of the conservation of relative state-space locations across time more straightforward.

GPFA is an unsupervised dimensionality reduction algorithm designed to simultaneously perform temporal smoothing (under the assumption of Gaussian process dynamics) and dimensionality reduction (with the factor analysis model). First, because GPFA assumes nonnegative values, we added the across-time minimum to the *z*-scored high-gamma activity for each electrode and trial. Data from all blocks within a participant were combined for this analysis. GPFA was then performed on activity across electrodes and time for specific CV contrasts. Across all CV contrasts, optimal smoothing had a SD of 25 ms and we kept the first 10 latent dimensions for further analysis ($\mathbf{G}^{10}$). We retained the first 10 latent dimensions from the initial round of dimensionality reduction from GPFA. We chose to keep the first 10 dimensions because this occurred at the approximate "elbow" of the percentage variance accounted for curve and cumulatively accounted for ~65% of the variance, and visual examination of the projection time courses confirmed that the lower dimensions contained little structure. The projection of the neural data into these 10 latent dimensions was then subjected to LDA, with three classes.

We observed that the different CV contrasts examined were separable across several latent dimensions from GPFA, and so we applied LDA at consonant and vowel time points, using contrasting phonemes as class identifiers. LDA is a supervised dimensionality reduction algorithm that finds the projection that maximizes the linear discriminability of the (user-defined) clusters. LDA can be thought of as a discrete version of the general linear model decoders we used in the decoding results. Multiclass LDA was performed on the GPFA representation ($\mathbf{G}^{10}$) by computing the matrix $\mathbf{L}^{\star} = \mathbf{L}\boldsymbol{\Sigma}^{-1/2}$, where $\mathbf{L}$ and $\boldsymbol{\Sigma}$ are the class centroids and common within-class covariance matrices, respectively. Classes were determined by the specific phoneme contrasts examined. We then took the singular-value decomposition of the covariance matrix of $\mathbf{L}^{\star}$, and projected $\mathbf{G}^{9}$ into the nine dimensions of the corresponding eigenspace (LDA necessarily results in an *n*-1 dimensional space). As shown later in Figure 9*a*, we choose the first latent dimension from LDA because, by definition, it is the dimension for which the contrasting phonemes are most separable. We used the first two latent dimensions from LDA ($\mathbf{L}^{2}$) for the final state-space. The first two LDA dimensions ($\mathbf{L}^{2}$) were used because we had three categories for each contrast (three vowels for each consonant in the anticipatory coarticulation analysis and three consonants for perseverative coarticulation). Mathematically, an *n*-1 dimensional space is sufficient to linearly separate *n* categories. We therefore performed our quantification in $\mathbf{L}^{2}$.

Quantification of state-space organization. The main thrust of our arguments regarding context dependence in the dynamic state-space organization was that the identity of surrounding phoneme contrasts imparted structure to the state-space during the production of individual phonemes, and that the trajectories for an individual phoneme were biased toward the state-space location for surrounding phonemes. We reasoned that these phenomena would be evidenced as some degree of "clustering" of single-trial state-space trajectories during individual phonemes according to surrounding phoneme contrasts, and an autocorrelation of single-trial state-space trajectories across syllable contrasts that extended through the times for adjacent phoneme segments. Therefore, the goal of our analysis of state-space structure was to derive metrics that quantified these phenomena.

First, we examined the time course of cross-trial phoneme separability (Bouchard et al., 2013) of different consonants and vowels transitioning to/from individual phonemes. This quantifies the difference in the average distance between phonemes and the average distance within a phoneme, so that larger values correspond to tighter distributions within a phoneme and larger distances between phonemes. This measures the degree to which phoneme contrasts imparted categorical structure as a function of time. The empirical null distribution for this metric was found by randomly permuting trial identity 200 times. This distribution was tightly centered on 0, as expected.

Second, we examined the time course with which the exact relative locations of single-trial trajectories during consonants and vowels were correlated while transitioning to a single adjacent phoneme. Here, we averaged the state-space trajectories for single trials over a 25 ms window centered on the times of average peak cluster separability for consonants (Tc) and vowels (Tv). This served as the estimate of state-space region in $\mathbf{L}^{2}$ for particular consonant ($\mathbf{L}^{2}_{\mathrm{Tc}}$) and vowel ($\mathbf{L}^{2}_{\mathrm{Tv}}$) contrasts. The vector of these values across trials associated with the syllable contrasts described above was then correlated through time using Equation 1 (e.g., $R(\mathbf{L}^{2}_{\mathrm{Tc}}(t), \mathbf{L}^{2}_{\mathrm{Tc}}(t-\tau))$; i.e., the state-space autocorrelation function. The empirical null distribution for this metric was found by randomly permuting trial identity across contrasts 200 times. The null distribution was tightly centered on 0, as expected.

Additionally we examined the organization of state-space dynamics by calculating correlation coefficients as described above after randomizing within and across syllables in a given contrast. To examine the time course of single-trial autocorrelations resulting purely from being part of a specific consonant-vowel syllable, we randomized trials (200 times) strictly within a given consonant-vowel syllable, and calculated the state-space autocorrelation function as described above. To examine the time course of correlations between the state-space location for a consonant and the other consonants, and for a given vowel with the other vowels, we randomized trials strictly across the syllables for a given contrast. The correlation for a contrast was taken as the average across randomizations.

Statistical testing. Time points with statistically significant differences in Hγ across vowel comparisons in Figure 1 (e.g., in Fig. 1*i*) were those in which any of three rank-sum tests on pairwise comparisons resulted in a *p*-value of $<10^{-3}$. Statistical significance of the across-block distribution of model performance (R^2, Figs. 3, 4; 8; see Figs. 5, 7) was performed against the null hypothesis of 0 with *t* tests (see Cross-validation and regularization procedure). Comparisons between distributions of different R^2 distributions (Fig. 4, 8) were performed with the Wilcoxon signed rank test. Statistical significance of phoneme separability (see Fig. 9*c*) was determined with *t* tests against the expected value of 0 under the null hypothesis (see Quantification of state-space organization). Unless stated otherwise, the results of statistical tests were deemed significant if the Bonferroni-corrected probability of incorrectly rejecting the null hypothesis was <0.05.

Results

Intracranial cortical recordings from language dominant hemispheres were synchronized with microphone recordings as participants read aloud consonant–vowel syllables (CVs) commonly used in American English (19 consonants followed by either /a/, /u/, or /i/; Bouchard et al., 2013). This task was designed to sample across a broad range of phonetic features for both consonants and vowels (Jakobson et al., 1951; Chomsky and Halle, 1968). In particular, /a/, /i/, and /u/ are considered "cardinal" vowels, because they span the articulatory and acoustic space of all vowels, and are the most conserved in the languages of the world (Jakobson et al., 1951; Maddieson and Disner, 1984; Hillenbrand et al., 1995; Ladefoged and Johnson, 2011). The cardinal vowels were chosen to examine key questions about covariation between cortical activity and vowel production in the context of simple consonant-vowel syllables, not the general encoding/decoding of all vowels.

Single-utterance acoustics of vowels and cortical activity

We focused on determining the role of the ventral sensory-motor cortex (vSMC) in vowel production. vSMC is anatomically defined as the ventral portions of the precentral and postcentral gyri, as well as the gyral formation at the ventral termination of

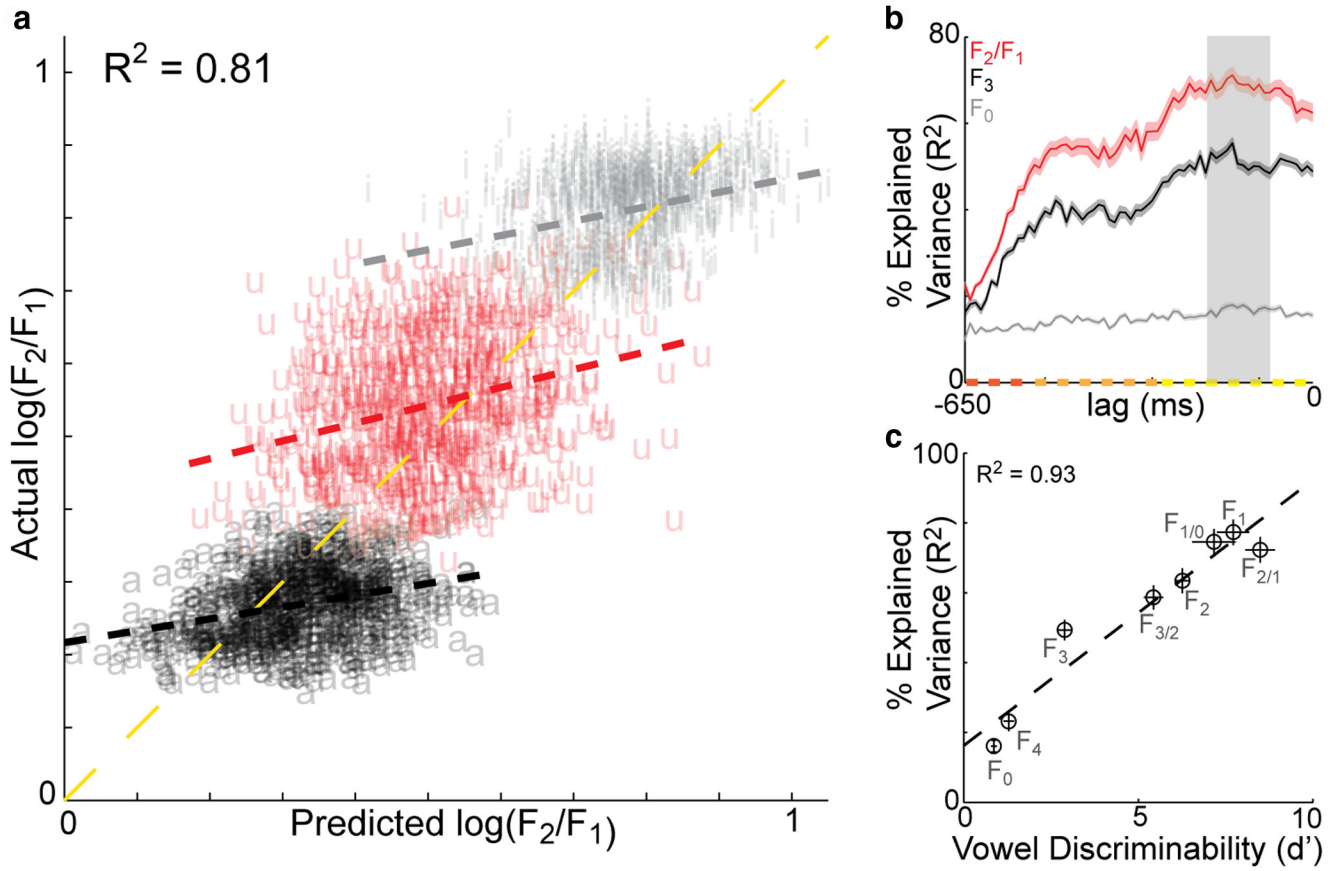

Figure 3. Neural decoding of single-utterance acoustics across vowel. ***a***, Scatter plot of predicted F_2/F_1 versus actual F_2/F_1 for decoders trained to predict the across-vowel variability from the cortical data for /a/ (black), /i/ (gray), and /u/ (red). The long, yellow dashed line is the linear fit across vowels, while the shorter dashed lines correspond to the linear fits within a vowel. ***b***, Time course of decoding performance (% Explained variance, R^2) for F_2/F_1 (red), F_3 (black), and F_0 (gray). Data are presented as mean ± SE from $N = 24$ recording sessions in three participants. Gray shaded area corresponds to time points at which model performance for different features were compared. Colored tick marks on the ordinate demarcate distinct temporal epochs in the structure of the underlying cortical data, corresponding to a vowel epoch (yellow), a consonant epoch (light orange), and early times (dark orange). Time point 0 corresponds to the beginning of the acoustic measurement. ***c***, Vowel discriminability (d') versus decoding performance (% Explained variance, R^2) for the eight formant features. Data are presented as mean ± SE from $N = 24$ sessions for both d' and R^2. The dashed line is the best linear fit between the two.

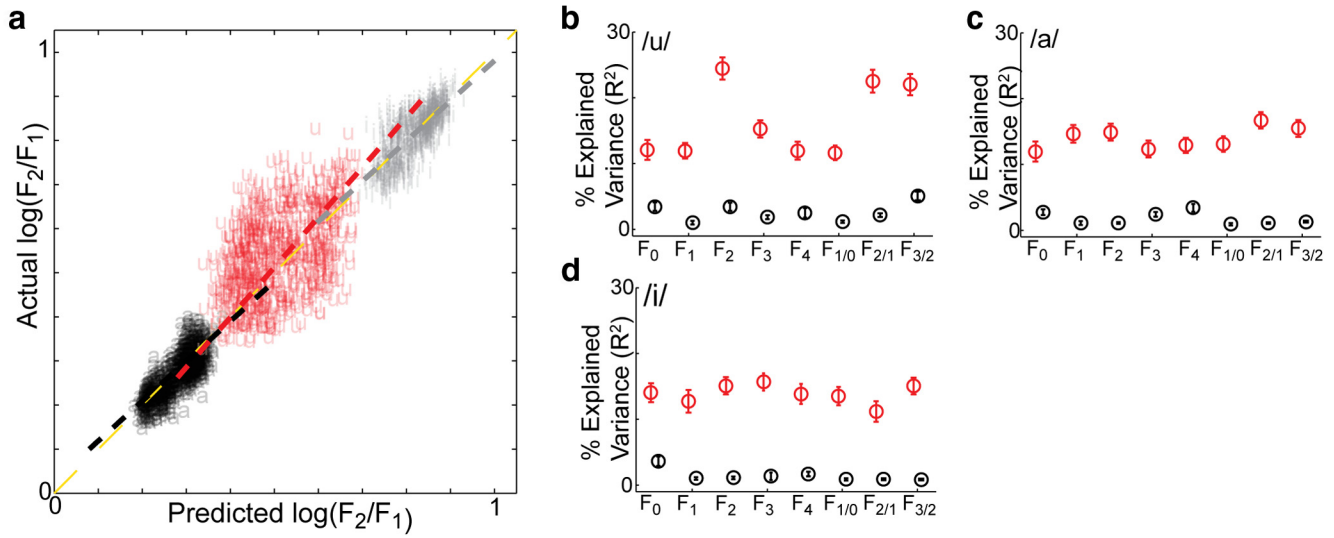

Figure 4. Neural decoding of single-utterance acoustics within vowels. ***a***, Scatter plot of predicted $\log(F_2/F_1)$ versus actual $\log(F_2/F_1)$ for decoders trained to predict the within-vowel variability from the cortical data for /a/ (black), /i/ (gray), and /u/ (red) individually. The yellow dashed line is the linear fit across vowels; the shorter dashed lines correspond to the linear fits within a vowel. ***b–d***, Average decoding performance (% Explained Variance, R^2) across formant features for decoders trained within a vowel (red) and across vowels (black) for /u/ (***b***), /a/ (***c***), and /i/ (***d***). Data are presented as mean ± SE, $N = 24$ recording sessions.

the central sulcus that connects the two, known as the subcentral gyrus (Bouchard et al., 2013). To better understand the relationship between vSMC activity and speech production, we examined the magnitude and dynamics with which the acoustic parameters of cardinal vowels covaried with cortical activity on a single-utterance basis. We extracted the fundamental frequency (F_0) and the first four vowel formants (F_1–F_4) as a function of time (see Materials and Methods), and calculated feature ratios. The bottom row of Figure 1*a* plots single-trial feature traces versus time (individual black lines) from >100 syllables containing -/a/, -/i/, and -/u/ from a male participant during one recording session. Here, acoustic features are aligned to the acoustic onset of the consonant-to-vowel transition ($t = 0$). Vowel acoustics have an extended duration and reach an acoustic steady state, allowing for accurate measurement. The fundamental frequency (F_0) is determined by the vibration frequency of the glottis, whereas the formants F_1 and F_2 and ratios F_1/F_0 and F_2/F_1 are directly related to the physical shape of the vocal tract, which is itself determined by the configuration of the tongue, lips, and jaw (Ladefoged and Johnson, 2011). Specifically, the production of /a/ is accomplished by depressing the tongue toward the bottom of the mouth (Fig. 1*a*, top left); production of /i/ is accomplished by raising the front of the tongue toward the soft palate and is optionally accompanied by narrowing of space between the lips (Fig. 1*a*, top center); production of /u/ is accomplished by raising the back of the tongue toward the soft palate and protruding/rounding the lips (to lengthen the vocal tract; Fig. 1*a*, top right; (Ladefoged and Johnson, 2011). These different vocal tract configurations give rise to formant structures for /a/, /i/, and /u/ that are quite distinct (Fig. 1*a*) and can be visualized by extracting the formant values from the midpoint (average of middle one-fifth) of each vowel utterance (Fig. 1*b*). Across utterances, variability in the acoustics was observed with multiple productions of the different vowels (Fig. 1*a*,*b*), although the means for each feature were generally stable across utterances of a vowel within a recording session.

We recorded neural activity from 80–90 electrodes located directly on the surface of vSMC (Fig. 1*c*; Bouchard et al., 2013). We focused on the high-gamma frequency component of cortical field potentials (70–150 Hz), which correlates well with multi-unit firing rates (Crone et al., 1998; Edwards et al., 2010; Ray and Maunsell, 2011; Bouchard et al., 2013). For each electrode, we normalized the time-varying high-gamma amplitude to baseline statistics by transforming to *z*-scores. These direct cortical recordings yielded robust high-gamma activity that differed between the cardinal vowels on single trials. In Figure 1, *d*–*i*, we present single-trial high-gamma activity recorded from an electrode in the area primarily associated with the lips (e124) and two electrodes in an area primarily associated with the tongue (e119 and e135) during the production of /ja/, /ji/, and /ju/ (/j/ is the American English sound "y"), and the corresponding averages (here, data were temporally aligned to the acoustic onset of the consonant-to-vowel transition). We found that electrode 124 was selectively active during /u/, the production of which involves lip rounding (Fig. 1*d*,*e*; yellow dots demarcate times with $p < 10^{-3}$ from rank-sum tests between vowels). In contrast, although both electrode 119 and electrode 135 are active during all three cardinal vowels, the relative magnitude of activity differentiates one of the vowels from the others. For example, the activity of electrode 119 (Fig. 1*f*,*g*) differed significantly between the vowels at various times, whereas activity of electrode 135 (Fig. 1*h*,*i*) is consistently greater for /i/ and /u/ (both "high-tongue" vowels) than for /a/ (a "low-tongue" vowel), suggesting that e135 is involved in raising the tongue. Together, these examples illustrate that the activity of individual electrodes differentially contributes to vowel production, and emphasize that the vowels are produced by the pattern of activation across multiple cortical sites.

Neural decoding of across-vowel acoustics

To quantify how well vowel formant features could be predicted from the population of recorded vSMC activity, we used cross-validation to train and evaluate regularized linear decoders. The cortical features used for decoding were spatial principal components derived independently at each point in time (see Materials and Methods). We found that such decoders were able to predict single-trial acoustics with high accuracy. In Figure 3*a*, we plot the F_2/F_1 ratio predicted by the decoder versus the actual values for one participant. The predicted values are in good agreement with the actual values; for this participant, 81% of the variability in F_2/F_1 across the cardinal vowels could be accurately predicted from the vSMC population neural activity. Also note that the best-fit lines of predicted versus actual F_2/F_1 within vowels had very shallow slopes (Fig. 3*a*; light gray, red, and black dashed lines correspond to regression within /i/, /u/, and /a/, respectively). This suggests that decoders trained to predict across-vowel acoustic variability do a poor job at predicting within-vowel variability.

We examined the time course of decoding performance by training separate decoders on cortical data at different lag times relative to the measured acoustics (Fig. 3*b*). At each point in time preceding the onset of acoustic measurement (acoustic measurements were taken as the mean values from the central one-fifth of each vowel utterance; onset of the central one-fifth is denoted here as $t = 0$), we quantified decoding performance by calculating the percentage of acoustic variance (R^2) predicted by the optimal decoder for that time. We restricted our decoding analysis to the time before the onset of acoustic measurement to focus on vSMC cortical activity preceding the acoustics. The red trace in Figure 3*b* summarizes decoding accuracy for F_2/F_1 (mean ± SE; $N = 24$ recordings sessions from three participants) as a function of the lag time in cortical data relative to the measured acoustics ($t = 0$). The decoding time course revealed three epochs, with highest decoding during vowel times (i.e., the time when vowel features are represented; $t = 0$:−280, yellow marks along *x*-axis), high decoding performance during consonant times (i.e., when consonant features are represented; $t = -290$: −500, light orange marks), and a sharp drop in decoding performance during prevocal times ($t = -510$: −650, dark orange marks; Bouchard et al., 2013). Note that the very early significant decoding performance may in part reflect the self-paced, list-reading nature of the task. Together, these results demonstrate that the acoustics of cardinal vowels can be predicted with high accuracy from the population of vSMC activity, and high-accuracy decoding extends into consonant times.

Different acoustic features have very different within-vowel versus across-vowel statistics. For example, F_2/F_1 was an acoustic feature which discriminated /a/, /i/, and /u/ from one another and was also among the most variable within individual vowels. Therefore, we examined how decoding performance varied as a function of vowel discriminability. For example, we found that the third formant (F_3; Fig. 3*b*, black trace) had smaller within-vowel variability but reduced cross-vowel discriminability relative to F_2/F_1. F_3 was decoded with less accuracy across all times, although the temporal structure of decoding performance was similar to that of F_2/F_1. In contrast, fundamental frequency (F_0) varies little across vowels and within vowels, and had significantly reduced decoding performance across all times (although still greater than chance). Across features, we found that during peak

vowel decoding times ($t = -80$: -180 ms; Fig. 3*b*, shaded gray), predictive performance varied widely across acoustic features (R^2 range: [0.15 0.80]). In Figure 3*c* we plot the average decoding performance for each acoustic feature as a function of how discriminable the vowels are based on that feature (quantified with the discriminability metric d', $N = 24$ blocks; dashed line is from linear regression). We found that decoding performance of vowel features was well predicted by the discriminability of the feature, with 93% of the decoding performance across features being explained by vowel discriminability. In particular, ~75% of the variance of those features associated with F_1 could be decoded from vSMC activity. From an articulatory standpoint, modulations of F_1 are related to both the height of the tongue in the mouth and the area between the lips (Ladefoged and Johnson, 2011).

Neural decoding of within-vowel acoustics

We found that decoders trained to predict the across-vowel variability did a poor job at predicting within-vowel variability (Fig. 3*a*). For each acoustic feature, the within-vowel variability describes the single-utterance deviations from the mean of each vowel. Therefore, unlike the across-vowel variability, which must be used to acoustically discriminate one vowel from the others, the within-vowel variability is often considered "noise" that interferes with vowel identification (Perkell and Nelson, 1985; Hillenbrand et al., 1995). For speech production, the observed poor performance of cross-vowel decoders to predict within-vowel variability raises the question: is the within-vowel variability related to variability in vSMC during individual vowels, or is it due to peripheral noise (e.g., noisy transmission at the neuromuscular junction) or to variability that arises in the nervous system independently of vSMC (e.g., the cerebellum)? If the vSMC activity generating a vowel is invariant (i.e., categorical), there should be no relationship between the within-vowel acoustic variations and the utterance-by-utterance fluctuations in cortical activity, and so decoding performance during the vowel epoch should be at chance levels (i.e., $R^2 \sim 0$).

We trained different decoders for each vowel and acoustic feature separately and found that even the within-vowel variability can be predicted from vSMC activity, albeit to a lesser extent than the across-vowel acoustics. For example, Figure 4*a* plots the predicted values of F_2/F_1 versus the actual values for the same data in Figure 3*a*, and shows that within-vowel regression lines (dashed light-gray, red, and black lines for /i/, /u/, and /a/, respectively) had slopes near unity (yellow dashed line). Across acoustic features, within-vowel decoders (red, mean ± SE; $N = 24$) for /u/ (Fig. 4*b*), /a/ (Fig. 4*c*), and /i/ (Fig. 4*d*) predicted significantly larger amounts of acoustic variability than expected by chance ($p < 10^{-10}$ for all, t test). Within-vowel decoders also outperformed the across-vowel decoders (black; mean ± SE; $N = 24$; $p < 10^{-10}$ for all, Wilcoxon sign-rank test), which were near chance levels for most features. Across the cardinal vowels, we found that decoding performance was generally largest for /u/ and smallest for /i/. This could reflect the highly configurable nature of articulations involved in /u/, and the reduced articulatory control involved in /i/, which could result in increased controlled variability for /u/ relative to /i/ (Fujimura and Kakita, 1975; Ladefoged and Johnson, 2011). We found that within the cardinal vowels, some features were more accurately decoded than others. In particular, the features associated with F_2 were much more accurately predicted than other features within the vowel /u/, whereas the differentiation of decoding performance across features was less pronounced for /a/ and /i/ (Fig. 4*b–d*). As with F_1, F_2 is also tightly associated with the position of the articulators in the vocal tract. Interestingly, previous linguistic studies have observed that F_2 is also the acoustic feature that is most heavily influenced by surrounding phonemes (Hardcastle and Hewlett, 2006).

Comparison of structure of different decoders

To gain further insight into our decoding models and, thus, the spatial organization of cortical activity, we compared decoders trained to predict different subsets of the data that emphasize different sources of variability in the behavior. We measured the correlation coefficient between the vectors of optimal decoding weights from the time of peak decoding performance. Strong correlations could indicate relationships between the cortical features used to decode different data subsets and, therefore, that similar cortical features were tied to different sources of variability in the acoustics. Conversely, low correlations would indicate that different cortical features are associated with different acoustic features. Note that there are some intrinsic correlations between several of the acoustic features (e.g., F_1 and F_1/F_0, F_2/F_1, etc.).

The plots in Figure 5*a* display average correlation coefficient (averaged across $N = 24$ blocks) between the optimal decoding weights during the times of peak cross-vowel decoding ($t = -180$: -80). Asterisks denote significant correlations at $p < 10^{-4}$. Generally speaking, there is little correlation between decoders trained on different formant features. The pattern of significant correlations between the decoders largely reflects the correlations in the features themselves (e.g., F_2 decoders are correlated with F_2/F_1 decoders). We additionally compared the structure of decoders trained on vowel subsets (i.e., across all vowels or within a particular vowel) for a given acoustic feature (Fig. 5*b*). This analysis revealed that correlations between decoders trained on the same acoustic features across different vowel comparisons were generally very low (no significant correlations at $p < 10^{-3}$). However, a few exceptions are noteworthy: (1) across all vowel set comparisons, correlations were modest for F_0 and F_4; (2) modest correlations were observed for F_2 and F_3/F_2 between the across-vowel decoder and the /u/ decoder; and (3) modest correlations were observed for F_3 between the across-vowel decoder and the /a/ decoder. Across features, the pattern of correlations between the cross-vowel decoders and the single-vowel decoders reflects the degree to which the cross-vowel decoder predicted single-vowel acoustics (Fig. 4*b–d*). Together, these results demonstrate that the weightings of cortical features underlying subtle acoustic variability within a vowel are generally different across the cardinal vowels tested here. This finding is in line with the distinctiveness of articulations involved in their production (Ladefoged and Johnson, 2011) and could reflect precisely tuned control for different cardinal vowels (Todorov and Jordan, 2002). Note that because cortical features correspond to spatial patterns of activity across electrodes, this observation does not imply distinct electrodes for acoustic different features (Bouchard et al., 2013).

Coarticulation in vowel acoustics and vSMC activity

During natural speech, vowels are rarely produced in isolation, but are instead produced in the context of sequences of other phonemes. Previous linguistic studies have shown that vowel formants can differ depending on which articulator is engaged in the production of preceding consonants (Kozhevnikov and Chistovich, 1965; Bell-Berti and Harris, 1975; Hardcastle and Hewlett, 2006). We therefore looked for evidence of perseveratory (i.e.,

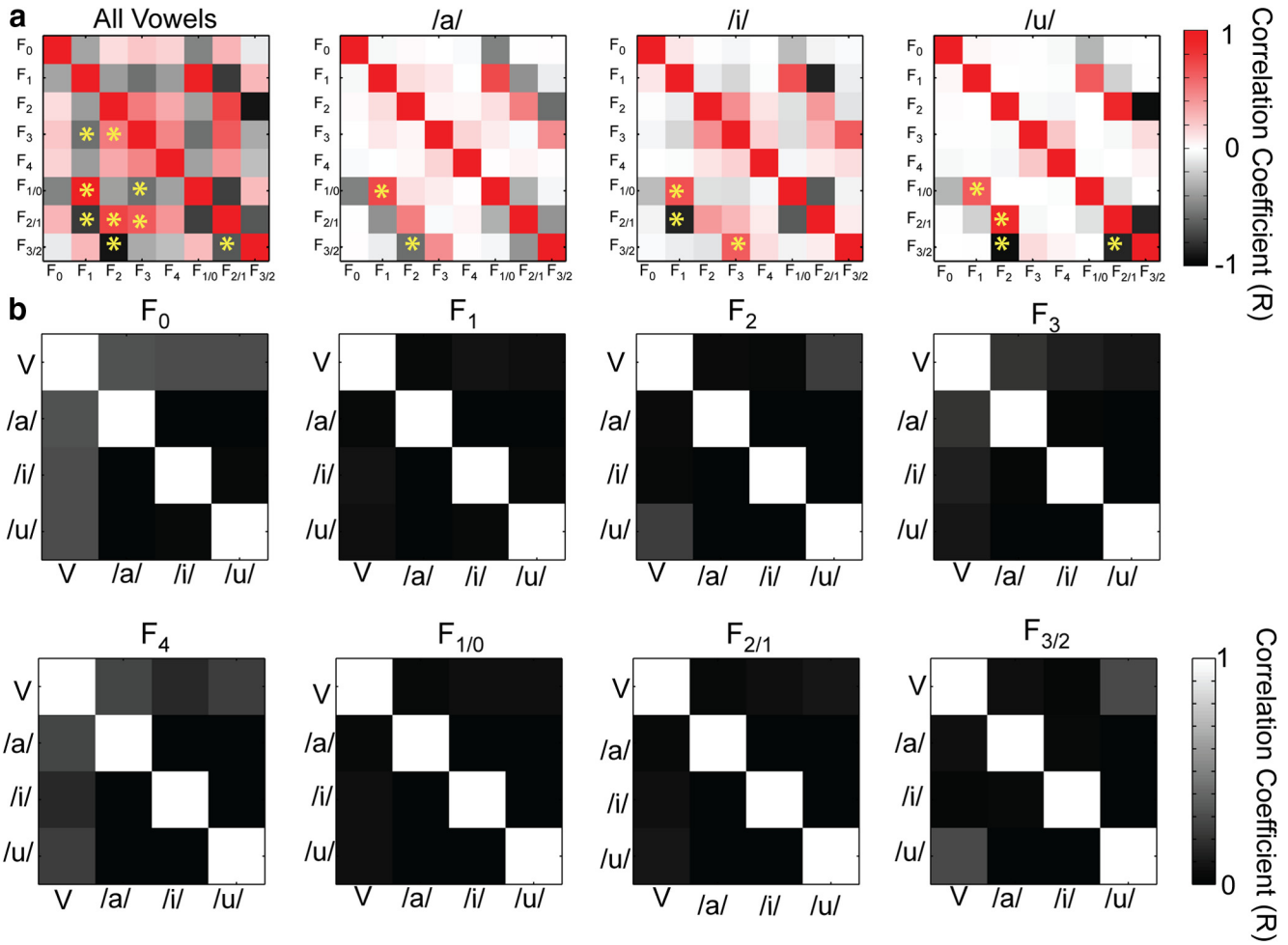

Figure 5. Comparison of structure of different decoders. ***a***, Correlations between decoders trained on different formant features across vowels (far left), and within /a/ (center left), /i/ (center right), and /u/ (far right). Matrices display the average correlation coefficient (averaged across $N = 24$ blocks). ***b***, Correlations of decoders trained for different vowel sets for the same acoustic features. Asterisks in ***a*** denote significant correlations at $p < 10^{-4}$.

carryover) coarticulation in the vowel formants and corresponding vSMC activity. In Figure 6*a*, we plot the single-trial F_2/F_1 traces from one recording session (same as in Fig. 1*a*,*b*) for /u/, /a/, and /i/, with each trace colored according to the major articulator of the preceding consonant [red: labials (e.g., /b/, /p/); green: dorsal tongue (e.g., /g/, /k/); blue: coronal tongue (e.g., /d/, /t/)]. This revealed clear evidence of perseveratory coarticulation in the vowel acoustics attributable to the articulators used by preceding consonants, especially for /u/. For example, F_2/F_1 for /u/ is higher than average when preceded by dorsal tongue consonants, and lower than average when preceded by labial consonants. At each moment in time during the vowel, we estimated the magnitude of perseveratory coarticulation by determining how much of the variability in formant features could be explained by an optimal linear weighting of the major articulator engaged by the preceding consonants (e.g., lips, coronal tongue, dorsal tongue, Fig. 6*b*).

Across /u/, /a/, and /i/, this analysis revealed that the ability to predict acoustic variability in vowels based on the articulator of the preceding consonant peaked 75–125 ms after the consonant-to-vowel transition. During the central one-fifth of each vowel (Fig. 6*b*, gray shaded region), we found that the perseveratory coarticulation effect on F_2/F_1 was large for /u/ (~38%), modest for /a/ (~12%), and minimal for /i/ (~6%). Across vowels and acoustic features, we found a large range of perseveratory coarticulation magnitudes: 3–40% for /u/, 3–17% for /a/, and 3–9% for /i/ (Fig. 6*c*; chance is ~2%). The observed pattern of perseveratory coarticulation in the produced vowel acoustics (Fig. 6*c*) qualitatively resembles the pattern of within-vowel decoding performance across features and vowels (Fig. 4*b*–*d*). This suggests that a portion of the explained acoustic variability may be due to a systematic relationship between cortical activity and perseveratory coarticulation in vowel acoustics.

We observed qualitative evidence of perseveratory coarticulation in the vSMC activity by visually comparing the high-gamma activity during the production of /a/, /i/, and /u/ when preceded by the consonants /b/ (Fig. 7*a*), /d/ (Fig. 7*b*), and /g/ (Fig. 7*c*). The stop-plosive consonants /b/, /d/, and /g/ are produced by the formation and release of complete occlusion of the vocal tract by the lips, coronal tongue, and dorsal tongue, respectively. The plots in Figure 7*a*–*c* show the average high-gamma activity of four electrodes distributed along the dorsal-ventral axis of vSMC (burgundy-to-black with increasing distance from the sylvian fissure). Examination of cortical activity during the production of the vowels revealed qualitative differences across electrodes reflecting a preceding articulator engagement. For example, during /u/ (yellow arrows), electrodes in an area primarily associated with the tongue (bright red and gray traces) were more active when preceded by /d/ relative to /b/ and /g/. Analogously, we can examine whether the cortical activity generating consonants de-

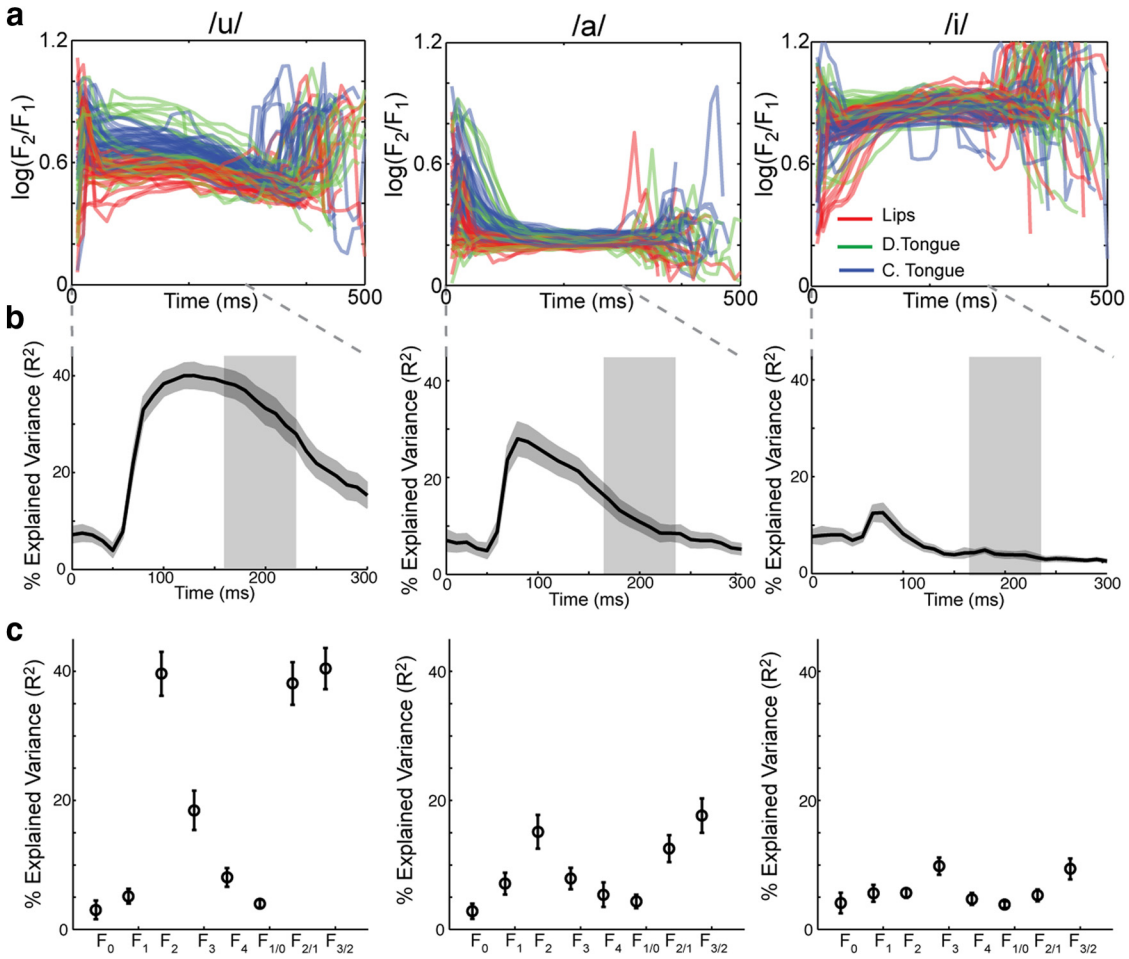

Figure 6. Perseveratory coarticulation of vowel acoustics. ***a***, Traces of F_2/F_1 versus time for /u/, /a/, and /i/ from one recording session. Individual traces are colored according to the articulator engaged during the production of the preceding consonant (red, lips; green, dorsal tongue; blue, coronal tongue). Time point 0 is the acoustic onset of the consonant-to-vowel transition. ***b***, The time course of percentage of variance in F_2/F_1 explained (R^2) by the primary articulator engaged by the preceding consonant for the three vowels (/u/, left; /a/, center; /i/, right). Data are presented as mean ± SE; $N = 24$ recording sessions. Gray-shared area demarcates approximate times of extracted formant values (central one-fifth of each vowel utterance). Time point 0 is the acoustic onset of the consonant-to-vowel transition. ***c***, Percentage variance explained (R^2) by the articulator engaged by the preceding consonant for the eight formant features.

pends on the upcoming vowel. During /b/ (Fig. 7*a*, orange arrows), the electrodes in an area primarily associated with the tongue (bright red and gray traces) are more active when the upcoming vowel is /i/ compared with /a/ and /u/. Furthermore, an electrode in an area primarily associated with the lips is more active during /d/ when the upcoming vowel is /u/ relative to /a/ and /i/ (Fig. 7*b*, orange arrows). This last example may reflect anticipatory lip rounding for the vowel /u/ during the lip-neutral consonant /d/ (Daniloff and Moll, 1968; Noiray et al., 2011). Interestingly, we did not observe activity of these electrodes during the lip-neutral /g/. This lack of activity may reflect coarticulatory strategies specific to this individual that deviate from models that attempt to generalize across all speakers (Hardcastle and Hewlett, 2006). Together, these results suggest that the multielectrode patterns of vSMC activity generating individual phonemes can depend on both the preceding and following phonemic context.

A cortical source for coarticulation

We next quantitatively determined whether the preceding consonant affects the cortical activity generating an individual vowel. We reasoned that if the effects of perseveratory coarticulation observed in vowel formants are mediated purely by the passive dynamical properties (i.e., inertia) of the vocal tract, then these contextual variations in acoustics would not covary with vSMC activity. This premise predicts that removing the effects of perseveratory coarticulation on formant features would increase decoding performance by "de-noising" the acoustics. Conversely, if the vSMC activity for vowels depends on the preceding consonant, then removing perseveratory coarticulatory effects from the acoustics should reduce decoding performance during vowel times by removing a controlled source of variation. Additionally, significant decoding performance during vowel times that persists after accounting for the acoustic consequences of the preceding consonant would imply that the ability to decode within-vowel fluctuations is not purely due to coarticulatory effects. We therefore compared the performance between decoders trained on original acoustic feature values and decoders trained on residual acoustic feature values after removing the effects of perseveratory coarticulation (i.e., the residuals from the linear model described in Fig. 6).

In Figure 8*a*, we plot the time course of within-vowel (/u/, top; /a/, upper middle; /i/, lower middle) R^2 values for decoders trained to predict the original F_2/F_1 values (black), and for decoders trained to predict the residual F_2/F_1 values after subtracting the (linear) effects of the preceding consonant articulator (red). The average decoding performance for all features during the

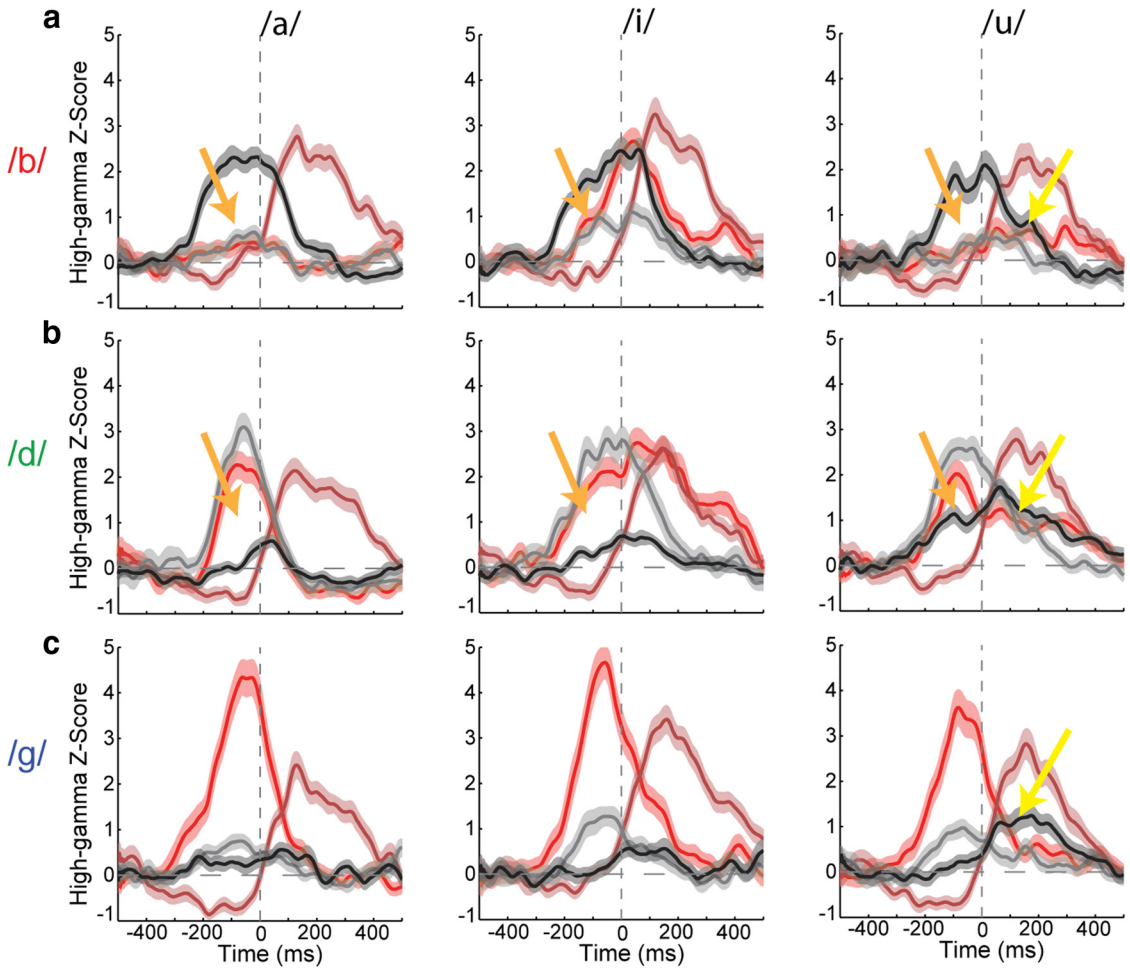

Figure 7. Coarticulation in cortical activity for individual phonemes. ***a–c***, Cortical data (mean ± SE) from four electrodes distributed along the dorsal-ventral extent of vSMC (colored dark red to black with increasing dorsal distance from sylvian fissure) during the production of /ba/, /bi/, /bu/ (***a***); /da/, /di/, /du/ (***b***); and /ga/, /gi/, /gu/ (***c***). Dashed vertical and horizontal lines demarcate 0 for time and high-gamma *z*-scores, respectively. Time point 0 is the acoustic onset of the consonant-to-vowel transition. Yellow arrows demarcate vowel times for /u/ with different activity depending on the preceding consonant; orange arrows demarcate consonant times during /b/ and /d/ with differential activity depending on the upcoming vowel.

vowel phase (Fig. 8*b*; Tv, yellow shaded regions in Fig. 8*a*) and during the consonant phase (Fig. 8*c*; Tc, orange shaded regions in Fig. 8*a*) are also plotted. Focusing on /u/ (Fig. 8*a–c*, top row), for which perseveratory coarticulation was largest, during the consonant phase (Fig. 8*a*, Tc, orange shading), removing the effect of preceding articulators on acoustic features resulted in significant reduction in decoding performance for features associated with F_2 [Fig. 8*c*, $^{*}p < 10^{-3}$, Wilcoxon's sign-rank test (WSRT)], as expected methodologically. Crucially, decoding performance was also significantly reduced during the vowel phase (Fig. 8*a*; Tv, yellow shading) for features associated with F_2 and F_3 (Fig. 8*b*, $^{*}p < 10^{-3}$, WSRT). Qualitatively similar results were observed for /a/ (Fig. 8*a–c*), although with greatly reduced magnitude, whereas no significant effects were found for /i/ (Fig. 8*a–c*). Across vowels and features, decoding performance changes resulting from removal of perseveratory coarticulation were largest for features that were most coarticulated (e.g., F_2 for /u/ and /a/), and smaller for those that were less coarticulated (e.g., F_0). This finding emphasizes that the ability to decode within-vowel feature variability was influenced by the degree to which that feature was coarticulated. These analyses demonstrate that the activity generating a vowel can depend on the preceding consonant.

We found that removing the effects of perseveratory coarticulation equalized decoding performance across features and across vowels, although small differences remained (e.g., F_2 vs F_4 for /u/ during vowel times). Nonetheless, we found that a modest, but significant amount of residual within-vowel acoustic variability (9–15%) could be accurately predicted from vSMC activity (distributions of cross-validated R^2 were all significantly >0). This implies that, after controlling for coarticulation, some of the variability within the production of a vowel has a source in vSMC. However, it is important to keep in mind that our method for quantifying and removing coarticulation is based on a linear model, which likely does not completely remove all such effects.

The high-decoding performance of across-vowel formant features observed during consonant times is strongly suggestive of anticipatory coarticulation in vSMC activity. However, the early decoding performance could reflect perseveratory coarticulation, as observed for within-vowel decoders above (Fig. 8*c*). We reasoned that if vSMC activity during consonant times depends on the identity of upcoming vowels, then removing perseveratory coarticulation from the acoustics should minimally affect cross-vowel decoding performance during consonant times. In the bottom row of Figure 8*a–c*, we compare the across-vowel R^2 values for decoders trained on acoustic features (black) and decoders trained on residual acoustic features after removing perseveratory coarticulation (red). We found that removing the effects of perseveratory coarticulation on

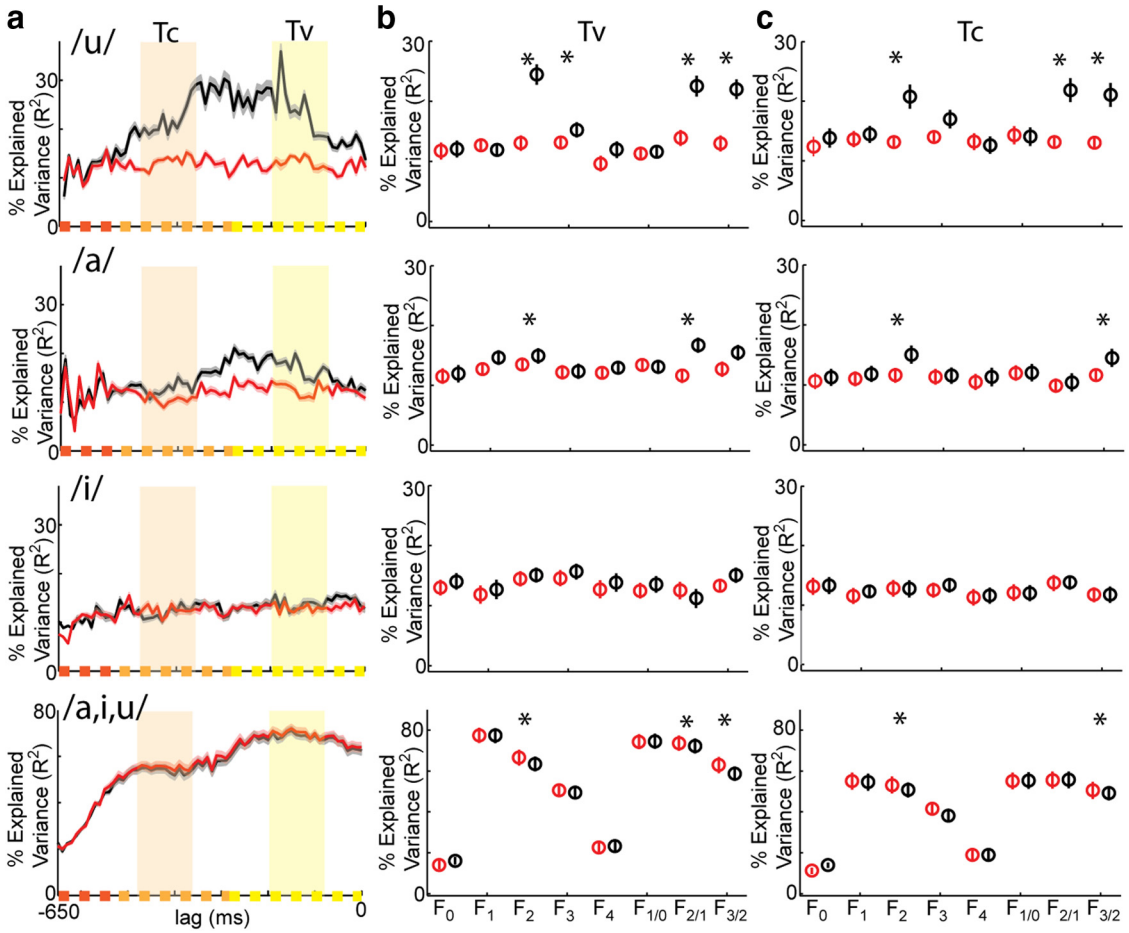

Figure 8. A cortical source for coarticulation. ***a***, Time courses of decoding performance (% Explained Variance, mean ± SE) for raw log(F_2/F_1; black) and residual log(F_2/F_1; red) after removing the effects of perseverative coarticulation that occurs because of the articulator engaged by the preceding consonant. Colored tick marks on the ordinate correspond to distinct temporal epochs in the structure of the underlying cortical data. Colored shaded region corresponds to time points at which decoding performance was calculated for vowel times (Tv, yellow) and consonant times (Tc, light orange). Time point 0 corresponds to the beginning of the acoustic measurement window. ***b***, Decoding performance (mean ± SE) during vowel times (Tv) for raw formant features (black) and the feature residuals after removing the effects of perseveratory coarticulation (red). $^*p < 10^{-3}$ between decoding performance of raw formants and residual formants, Wilcoxon signed-rank test, $N = 24$. ***c***, Decoding performance (mean ± SE) during consonant times (Tc) for raw formant features (black) and the feature residuals after removing the effects of perseveratory coarticulation (red). $^*p < 10^{-3}$ between decoding performance of raw formants and residual formants, Wilcoxon signed-rank test, $N = 24$.

vowel acoustic features minimally affected across-vowel decoding performance during vowel times (Fig. 8*a*, Tv, yellow shading; Fig. 8*b*) and consonant times (Fig. 8*a*, Tv, orange shading; Fig. 8*c*). However, across-vowel decoding of the residuals of F_2-associated features resulted in subtle, but consistent (and therefore statistically significant), increases in decoding performance across vowels (Fig. 8*b*,*c*, bottom, $^*p < 10^{-3}$, WSRT). F_2 is tightly associated with the position of the articulators in the vocal tract (Ladefoged and Johnson, 2011). These analyses demonstrate that the cortical activity generating consonants can be highly influenced by the upcoming vowel.

Anticipatory and perseveratory dynamics across the vSMC network

The decoding results described above, which used data analysis methods that were methodology that was local in time, demonstrates that there is structure in the vSMC network during consonants and vowels that depends on the upcoming/preceding phoneme. However, because of the temporal locality, the network dynamics generating consonant-vowel syllables cannot be uniquely determined from this analysis. Indeed, the observed time courses of decoding performance are consistent with several different dynamic network organizations. To gain insight into network dynamics, we used a combination of unsupervised (GPFA of single-utterance cortical activity) and supervised (LDA projection for specific syllable contrasts) dimensionality reduction techniques (see Materials and Methods) to find a low-dimensional cortical state-space corresponding to spatial activity patterns across the vSMC network (Briggman et al., 2005; Mazor and Laurent, 2005; Afshar et al., 2011; Bouchard et al., 2013; Mante et al., 2013; Shenoy et al., 2013). We examined how surrounding phonemes affect network trajectories for individual phonemes to gain a more mechanistic understanding into coarticulation using specific syllable contrasts (see Materials and Methods).

To understand how the network dynamics for consonants were affected by the identity of the upcoming vowel, we first examined trajectories of individual consonants transitioning to the vowels /a/, /i/, and /u/. Figure 9*a* displays the average (mean ± SE) trajectories derived from single-trial vSMC network activity associated with /ga/ (red), /gi/ (gray), and /gu/ (black). We observed clear separability of the state-space trajectories reflecting the upcoming vowel identity during the latter portion of the /g/-

consonant phase (orange box) as well as earlier times. Furthermore, the relative positions of trajectories during different vowels are preserved through the transition from the preceding consonant. That is, the state-space trajectories through the /g/ "subspace" are biased toward the state-space location of the upcoming vowel. Analogously, to understand how the network dynamics for individual vowels depends on the preceding consonant, we examined trajectories of labial, coronal tongue, and dorsal tongue consonants transitioning to /a/, /i/, or /u/. Figure 9*b* displays average trajectories derived from single-trial cortical activity associated with /bu/ (red), /du/ (gray), and /gu/ (black). During the early vowel time (Tv, yellow box), there is clear separability of state-space trajectories according to the preceding consonant identity, and the relative positions during consonants is preserved through the transition to the upcoming vowel.

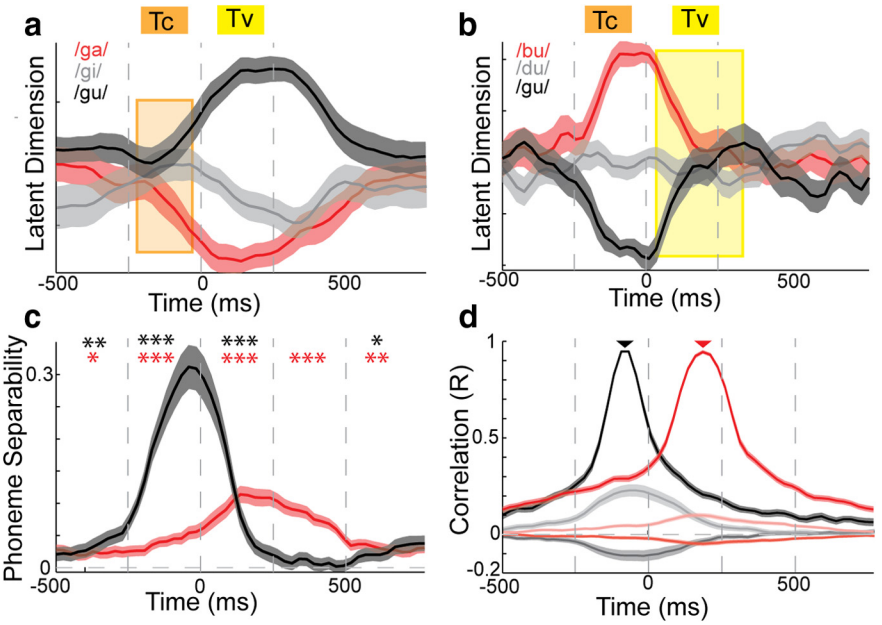

Figure 9. Anticipatory and perseveratory vSMC network dynamics during speech. ***a***, Average (mean ± SE) consonant-vowel trajectories in the single dimension that best separates the vowels derived from dimensionality-reduction of single-trial cortical activity associated with /ga/ (red), /gi/ (gray), and /gu/ (black). Orange box depicts the consonant time. Time point 0 is the acoustic onset of the consonant-to-vowel transition in all plots. ***b***, Average (mean ± SE) consonant-vowel trajectories in the single dimension that best separates the consonants derived from dimensionality-reduction of single-trial cortical activity associated with /bu/ (red), /du/ (gray), and /gu/ (black). Yellow box depicts the vowel time. ***c***, Average (mean ± SE) time course of phoneme separability. Black, Phoneme separability of the labial, coronal tongue, and dorsal tongue consonants transitioning to individual vowels ($n = 78$). Red, Phoneme separability of the vowels /a/, /i/, and /u/ transitioning from individual consonants ($n = 162$). ***d***, Average (mean ± SE) state-space autocorrelations for consonant-vowel syllables. Black, Correlation coefficients for consonant states (black triangle) during the production of individual vowels ($n = 78$). Red, State-space autocorrelation for vowel states (red triangle) during the production of individual consonants ($n = 162$). Also plotted are the correlation for within-syllable randomizations (light gray and light pink), as well as the results of the across-syllable randomizations (dark gray and dark pink).

Across the three subjects examined here, we had a total of $N = 162$ syllable contrasts for anticipatory coarticulation and $N = 78$ syllable contrasts for perseverative coarticulation (including, but not limited to, the examples above; see Materials and Methods). The main thrust of our arguments regarding context dependence in the dynamic state-space organization was that the identity of surrounding phoneme contrasts imparted structure to the state-space during the production of individual phonemes, and that the trajectories for an individual phoneme were biased toward the state-space location for surrounding phonemes. The first two LDA dimensions were used because we had three categories for each contrast (three vowels for each consonant in the anticipatory coarticulation analysis and three consonants for perseverative coarticulation).

We quantified the dynamic organization of the vSMC state-space using two complementary metrics. First, we examined the time course of cross-trial phoneme separability (see Materials and Methods) for individual labial, coronal tongue, and dorsal tongue consonants transitioning to a single vowel (e.g., [/bu/ /du/ /gu/]; Fig. 9*c*, black, mean ± SE, $N = 78$ syllable contrasts), as well the vowels /a/, /i/, and /u/ when transitioning from individual consonants (e.g., [/ga/ /gi/ /gu/]; Fig. 9*c*, red, mean ± SE, $N = 162$ syllable contrasts). This metric quantifies the difference between within-phoneme distances and across-phoneme distances in the cortical state-space for a given contrast, with larger values corresponding to more separable state-space trajectories for an individual phoneme. This analysis showed that the identity of adjacent phonemes imparts significant structure during both the peak consonant times (separability > 0, $p < 10^{-5}$, $N = 162$, t test) and peak vowel times ($p < 10^{-7}$, $N = 78$, t test). Furthermore, significant separability for consonants and vowels extended across multiple production epochs (Fig. 9*c*, $^{***}p < 10^{-5}$, $^{**}p < 0.01$, $^{*}p < 0.05$, t test). This demonstrates that the vSMC network organization for a given phoneme can be structured by the identity of upcoming phonemes (anticipatory coarticulation) and preceding phonemes (perseveratory coarticulation).

We further examined how single-trial state-space locations associated with different consonants/vowels persisted through the transition to/from a single adjacent phoneme. We autocorrelated the vector of single-trial state-space locations for different phoneme contrasts transitioning to/from an adjacent phoneme at different lag times. For example, the vector for consonants could correspond to the location of all single-trial trajectories associated with /bu/, /du/, and /gu/ for a given participant. This analysis tests for systematic biases of single-trial state-space trajectories during a phoneme that reflect the state-space location of adjacent phonemes. The black trace in Figure 9*d* shows the average state-space autocorrelation function derived from single-trial trajectories centered in consonant times (average state across five time points, centered at black triangle) for labial, coronal tongue, and dorsal tongue consonants transitioning to individual vowels (mean ± SE, $N = 78$, Fig. 9*b*). Likewise, the red trace in Figure 9*d* is the average state-space autocorrelation function derived from single-trial trajectories centered in vowel times (average state determined at red triangle) for /a/, /i/, and /u/ transitioning from individual consonants (mean ± SE, $N = 162$, Fig. 9*a*). The average state-space autocorrelations for both consonants and vowels exhibited an initial rapid decay followed by a slower decline. These correlation functions extended through the adjacent phonological segment and beyond.

Finally, we measured the time course of correlations resulting from randomizing single trials across different levels of sequence structure to provide further understanding into the dynamic organization of the vSMC network during CV sequences. We ex-

amined the dynamics of correlations attributable only to being associated with a specific consonant-vowel syllable by randomizing trials strictly within a syllable (see Materials and Methods), which removes single-trial autocorrelations but preserves the mean structure associated with the syllable. The correlations from being in a syllable were much smaller than the observed correlations of single-trial trajectories (Fig. 9*d*, light gray and pink traces). This demonstrates that long-time organization observed in the state-space autocorrelations corresponds to structure above and beyond that attributable to simply being associated with a specific syllable. To examine the correlation between trajectories for a given consonant and the other consonants, and for a given vowel with the other vowels, we randomized trials strictly across the different phonemes for a given contrast (see Materials and Methods). The correlations strictly across phonemes were negative on average and generally small (Fig. 9*d*, dark gray and pink traces). Together, these results demonstrate that the trajectories through phoneme subspaces are biased toward surrounding phoneme subspaces, and that these biases reflect long-time correlations on the level of single trials that are preserved across the consonant-vowel transition.

Discussion

vSMC control of vowels

The cardinal vowels /a/, /i/, and /u/ are important phonemes in speech, as they are found in most human languages and outline the acoustic and articulatory space of all vowels (Jakobson et al., 1951; Maddieson and Disner, 1984; Hillenbrand et al., 1995; Ladefoged and Johnson, 2011). To better understand the cortical processes generating vowel acoustics, we examined the covariation between vSMC neural activity and the acoustics of these vowels on a single-trial basis. Examining the relationship between vSMC and speech on a single-trial basis critically relies on the high spatio-temporal resolution of our ECoG grids, the high signal-to-noise ratio of high-gamma band activity, and the large number of trials in our dataset.

We found that vSMC population activity preceding measured vocalizations could predict large amounts (upwards of 75%) of across-vowel acoustic structure between /a/, /i/, and /u/ (Fig. 3). Performance in predicting different acoustic features was well predicted by the degree to which those features could statistically discriminate the vowels. For example, the produced variability in F_1 and F_2 (which are determined by vocal tract shape through positioning the tongue, lips, and jaw) was well predicted by vSMC activity; F_1 and F_2 are also critical for determining vowel identity (Jakobson et al., 1951; Hillenbrand et al., 1995). In contrast, the variability in pitch (F_0, which is controlled by the larynx) was only modestly predicted by vSMC activity during our particular task in which pitch change or prosodic voice modulation is not an important articulatory goal (Hillenbrand et al., 1995; Ladefoged and Johnson, 2011). This suggests that, during speech production, vSMC is exerting precise control for the generation of task-relevant parameters, while relaxing control for the generation of task irrelevant parameters (Lindblom, 1983; Guenther, 1995; Todorov and Jordan, 2002).

Although we found that vSMC activity was directly linked to the acoustics of produced vowels, it cannot be concluded that the representation of vowels in vSMC is in acoustic coordinates. Our previous results on consonants showed that individual sites in vSMC represent the speech articulators (Bouchard et al., 2013). As a consequence, the organization of vSMC representations of phonemes reflects articulatory relations, not acoustic relations. Indeed, the acoustic features that were most accurately predicted (those associated with F_1 and F_2), are precisely those features that are most directly related to the positioning of the tongue, lips, and jaw (Ladefoged and Johnson, 2011). Our results therefore suggest that, during steady state, a linear transformation of the vocal tract shape to vowel acoustics should be a reasonable approximation. This approximation may prove useful for constructing a speech prosthetic based on vSMC activity. Previous studies of speech decoding have demonstrated that ECoG signals from vSMC can be used to classify different speech sounds from one another at a rate greater than expected by chance (Leuthardt et al., 2011; Pei et al., 2011). In an interesting alternative approach, a study that used an array of penetrating electrodes from the mouth motor cortex of an individual with tetrapalegia demonstrated the ability to decode F_1 and F_2, and to use the subsequent decoded values to control a speech synthesizer (Guenther et al., 2009). Together with these previous studies, the high single-trial decoding performance of speech acoustics demonstrated here suggests that high-density ECoG recordings from vSMC are likely to be a successful strategy for a brain–machine interface (BMI) for a speech prosthetic, perhaps through directly coupling to an artificial vocal tract to generate continuous speech acoustics. This "embodied" approach is more in line with BMIs for limb prosthetics, and potentially allows for the inclusion of somatosensory feedback.

It has traditionally been assumed that variability across repeated movements toward the same target/goal largely reflects noise, perhaps introduced at the neuromuscular junction, or resulting from biomechanical properties of the skeletomuscular system (i.e., the inertia of the articulators; Ostry et al., 1996; Jones et al., 2002). However, recent electrophysiological recordings from motor cortices in non-humans [e.g., macaque premotor and primary motor cortex, and songbird vocal motor cortex (RA)] demonstrate that a portion of movement variability has a source in the CNS, although such studies do not exist for humans (Churchland et al., 2006a; Schoppik et al., 2008; Sober et al., 2008; Afshar et al., 2011). In line with these findings, we found that vSMC population activity could accurately predict a significant fraction of the variability within a given vowel (Fig. 4). As expected, the magnitude of within-vowel decoding was modest compared with across-vowel decoding. Furthermore, a modest but significant fraction of within-vowel variability (9–15%) could be accurately predicted after quantitatively controlling for effects of different preceding consonants on vowel acoustics (perseverative coarticulation, Fig. 8). This observation implies that the ability to decode within-vowel acoustics is not purely a result of coarticulation. Because the ability to decode within-vowel acoustic variability after (linearly) removing the effects of coarticulation was modest, future studies, perhaps combining direct monitoring of the articulators with ECoG in the context of vowel holds, should be performed. Together, these results demonstrate that the cortical activity generating individual vowels is not invariant but, instead, can be linked to utterance-to-utterance fluctuations in their production.

A cortical basis for coarticulation

Speech production is fundamentally a sequential behavior: individual phonemes are organized into syllables, which are sequenced into words, which themselves are flexibly arranged into sentences (Levelt, 1999; MacNeilage, 2011). The precise speech sequence across multiple temporal scales can affect the phonology of individual phonemes (Levelt, 1999; Hardcastle and Hewlett, 2006; Ladefoged and Johnson, 2011; MacNeilage, 2011). At the most local temporal scale, the articulations and acoustics

for individual phonemes can be affected by immediately surrounding phonemes (i.e., coarticulation or coproduction; Kozhevnikov and Chistovich, 1965; Öhman, 1966; Kent and Minifie, 1977; Whalen, 1990; Ostry et al., 1996; Hardcastle and Hewlett, 2006). Coarticulation is the primary reason why connected speech is not simply the concatenation of discrete, segmented units but instead reflects a smoothed trajectory through the speech sequence (Hardcastle and Hewlett, 2006). The cortical basis for such coarticulation/coproduction of individual phonemes has been greatly debated, and despite more than five decades of linguistic research into coarticulation, there is no consensus on its origin or function (Bell-Berti and Harris, 1975; Fowler, 1980; Fowler et al., 1980; Perkell, 1986; Whalen, 1990; Guenther, 1995; Ostry et al., 1996; Hardcastle and Hewlett, 2006; Noiray et al., 2011).

Here, we show that cortical activity generating individual phonemes exhibits both anticipatory and perseverative coarticulation reflecting surrounding phonemes. Specifically, population decoding showed that vSMC activity during the production of consonants was predictive of the upcoming vowel (anticipatory coarticulation), and that the activity generating a vowel depends on the major articulator of the preceding consonant (perseveratory coarticulation; Fig. 8). Furthermore, by examining the dynamic organization of the vSMC network, we found that state-space trajectories through individual "phoneme subspaces" were biased toward the phoneme subspace of the adjacent phoneme (Fig. 9). These two approaches showed a general temporal correspondence between single-trial decoding performance and the internal organization of vSMC network states, and provide complementary insight into cortical functioning during speech production. Together, these data and analyses provide a definitive demonstration that vSMC activity for phonemes exhibits anticipatory and perseveratory biases toward adjacent phonemes during the production of consonant-vowel syllables. Modern theories hypothesize that targets for speech production correspond to continuous regions in acoustic or articulatory ("orosensory") coordinates (i.e., "target windows," "convex regions;" Fowler, 1980; Browman and Goldstein, 1989; Keating, 1990; Guenther, 1995; Guenther et al., 2006; Golfinopoulos et al., 2010). Such regions define the range of representations that will produce identifiable phonemes in a language. In line with a continuous and dynamic representation, our results demonstrate that vSMC representations of individual phonemes are trajectories through network subspaces, and that surrounding phonemes can bias these trajectories.

It is almost certainly the case that higher-order brain areas (e.g., Broca's area, supplementary motor area) contain more invariant representations than those found here for vSMC (Bohland and Guenther, 2006; Bohland et al., 2010). Indeed, the observed perseveratory coarticulation in vSMC activity could result from invariant input command signals. For example, during the production of a consonant-vowel sequence, the articulatory requirements of different consonants could leave the vSMC network in different initial states when the activity for the vowel begins. These differences in "initial conditions" could bias subsequent network states that generate the vowel, and thus result in differences in vowel acoustics that depend on the articulations required for the preceding consonant (a sort of "network inertia" effect; Afshar et al., 2011). Therefore, in addition to the dynamic biomechanical properties of the vocal tract, perseveratory coarticulation may in part reflect the "inertial" properties of the dynamical system instantiated by vSMC during speech production (Guenther, 1995; Ostry et al., 1996; Bohland and Guenther, 2006; Golfinopoulos et al., 2010; Afshar et al., 2011; Shenoy et al., 2013).

That being said, the high decoding performance for vowel acoustics during the consonant phase (Fig. 8), coupled with the observed anticipatory bias in state-space trajectories (Fig. 9), demonstrates that the vSMC network state generating a consonant anticipates the identity of upcoming vowels. Such anticipatory modulations cannot result from network inertia effects that potentially explain results of perseveratory coarticulation. Both the anticipatory and perseverative biases reduce the "distance traveled" between sequentially activated network states. We propose that these biases in network trajectories toward surrounding phoneme states reflect overt, top-down control signals that optimize vSMC for rapid sequence production (speed) and minimize the time-dependent accumulation of behavior-deteriorating neural noise (accuracy) (Todorov and Jordan, 2002; Churchland et al., 2006b). Disentangling the precise contributions of top-down signals versus intrinsic dynamics from observations of the local network alone is very difficult, and is an important direction for future research.

In contrast to the long-time effects of surrounding phonemes on vSMC activity observed here, a previous examination of context-dependent modulations of neural activity in the motor cortex analog RA of songbirds found effects that were much more temporally local (Wohlgemuth et al., 2010). In songbirds, the temporal structure of song is thought to be generated, in part, by a "synaptic chain" mechanism, which, in its simplest form, requires that the sequencing circuit represent one syllable at a time (Long et al., 2010). Thus, the neural activity for a given syllable in downstream motor areas should be similarly temporally local. In both humans and non-human primates, sequence generation is thought to be subserved by a "competitive queuing" mechanism, in which multiple elements in the sequence are coactive and compete with one another for behavioral expression (Averbeck et al., 2002; Rhodes et al., 2004). Our observations are more parsimoniously explained by competitive queuing for sequencing of speech phonemes (Bohland and Guenther, 2006; Bohland et al., 2010).

References

Afshar A, Santhanam G, Yu BM, Ryu SI, Sahani M, Shenoy KV (2011) Single-trial neural correlates of arm movement preparation. Neuron 71:555–564. CrossRef Medline

Averbeck BB, Chafee MV, Crowe DA, Georgopolous AP (2002) Parallel processing of serial movements in prefrontal cortex. Proc Natl Acad Sci U S A 99:13172–13177. CrossRef Medline

Bell-Berti F, Harris KS (1975) Coarticulation in VCV and CVC utterances: some EMG data. J Acoust Soc Am 57:S70. CrossRef

Bohland JW, Guenther FH (2006) An fMRI investigation of syllable sequence production. Neuroimage 32:821–841. CrossRef Medline

Bohland JW, Bullock D, Guenther FH (2010) Neural representations and mechanisms for the performance of simple speech sequences. J Cogn Neurosci 22:1504–1529. CrossRef Medline

Bouchard KE, Mesgarani N, Johnson K, Chang EF (2013) Functional organization of human sensorimotor cortex for speech articulation. Nature 495:327–332. CrossRef Medline

Briggman KL, Kristan WB Jr (2006) Imaging dedicated and multifunctional neural circuits generating distinct behaviors. J Neurosci 26:10925–10933. CrossRef Medline

Briggman KL, Abarbanel HD, Kristan WB Jr (2005) Optical imaging of neuronal populations during decision-making. Science 307:896–901. CrossRef Medline

Browman CP, Goldstein L (1989) Articulatory gestures as phonological units. Haskins Laboratories Status Report on Speech Research 99:69–101.

Brown S, Laird AR, Pfordresher PQ, Thelen SM, Turkeltaub P, Liotti M (2009) The somatotopy of speech: phonation and articulation in the human motor cortex. Brain Cogn 70:31–41. CrossRef Medline

Chomsky N, Halle M (1968) The sound pattern of English. New York: Harper & Row.
Churchland MM, Afshar A, Shenoy KV (2006a) A central source of movement variability. Neuron 52:1085–1096. CrossRef Medline
Churchland MM, Yu BM, Ryu SI, Santhanam G, Shenoy KV (2006b) Neural variability in premotor cortex provides a signature of motor preparation. J Neurosci 26:3697–3712. CrossRef Medline
Churchland MM, Yu BM, Sahani M, Shenoy KV (2007) Techniques for extracting single-trial activity patterns from large-scale neural recordings. Curr Opin Neurobiol 17:609–618. CrossRef Medline
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. CrossRef Medline
Daniloff R, Moll K (1968) Coarticulation of lip rounding. J Speech Hear Res 11:707–721. CrossRef Medline
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. CrossRef Medline
Fowler CA (1980) Coarticulation and theories of extrinsic timing. J Phonetics 8:113–133.
Fowler CA, Rubin P, Remez RE, Turvey ME (1980) Implications for speech production of a general theory of action. In Language production (Butterworth B, ed). New York: Academic.
Fujimura O, Kakita Y (1975) Remarks on quantitative description of the lingual articulation. In: Frontiers of speech communication research (Lindblom B, Öhman SEG, eds). New York: Academic.
Golfinopoulos E, Tourville JA, Guenther FH (2010) The integration of large-scale neural network modeling and functional brain imaging in speech motor control. Neuroimage 52:862–874. CrossRef Medline
Gracco VL, Abbs JH (1986) Variant and invariant characteristics of speech movements. Exp Brain Res 65:156–166. Medline
Guenther FH (1995) Speech sound acquisition, coarticulation, and rate effects in a neural-network model of speech production. Psychol Rev 102: 594–621. CrossRef Medline
Guenther FH, Ghosh SS, Tourville JA (2006) Neural modeling and imaging of the cortical interactions underlying syllable production. Brain Lang 96:280–301. CrossRef Medline
Guenther FH, Brumberg JS, Wright EJ, Nieto-Castanon A, Tourville JA, Panko M, Law R, Siebert SA, Bartels JL, Andreasen DS, Ehirim P, Mao H, Kennedy PR (2009) A wireless brain-machine interface for real-time speech synthesis. PLoS One 4:e8218. CrossRef Medline
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. CrossRef
Hardcastle WJ, Hewlett N (2006) Coarticulation. Cambridge, UK: Cambridge UP.
Hillenbrand J, Getty LA, Clark MJ, Wheeler K (1995) Acoustic characteristics of American English vowels. J Acoust Soc Am 97:3099–3111. CrossRef Medline
Jakobson R, Fant G, Halle M (1951) Preliminaries to speech analysis. The distinctive features and their correlates. Cambridge, MA: MIT.
Jones KE, Hamilton AF, Wolpert DM (2002) Sources of signal-dependent noise during isometric force production. J Neurophysiol 88:1533–1544. CrossRef Medline
Keating P (1990) Phonetic representations in a generative grammar. J Phonetics 18:321–334.
Kent RD, Minifie FD (1977) Coarticulation in recent speech production models. J Phonetics 5:115–133.
Kozhevnikov VA, Chistovich LA (1965) Speech: articulation and perception. Washington, DC: U.S. Joint Publications Research Service.
Ladefoged P, Johnson K (2011) A course in phonetics. Boston: Cengage Learning.
Leuthardt EC, Gaona C, Sharma M, Szrama N, Roland J, Freudenberg Z, Solis J, Breshears J, Schalk G (2011) Using the electrocorticographic speech network to control a brain-computer interface in humans. J Neural Eng 8:036004. CrossRef Medline
Levelt W (1999) Producing spoken language: a blueprint of the speaker. In: The neurocognition of language (Brown CM, Hagoort P, eds), pp 1–40. Oxford, UK: Oxford UP.
Lindblom B (1963) Spectrographic Study of Vowel Reduction. J Acoust Soc Am 35:1773. CrossRef
Lindblom B (1983) Economy of speech gestures. In: The production of speech (MacNeilage PF, ed). New York: Springer.
Long MA, Jin DZ, Fee MS (2010) Support for a synaptic chain model of neuronal sequence generation. Nature 468:394–399. CrossRef Medline
MacNeilage PF (2011) The production of speech. New York: Springer.
Maddieson I, Disner SF (1984) Patterns of sounds. Cambridge studies in speech science and communication. Cambridge, UK: Cambridge UP.
Mante V, Sussillo D, Shenoy KV, Newsome WT (2013) Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature 503:78–84. CrossRef Medline
Mazor O, Laurent G (2005) Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. Neuron 48: 661–673. CrossRef Medline
Noiray A, Cathiard MA, Ménard L, Abry C (2011) Test of the movement expansion model: anticipatory vowel lip protrusion and constriction in French and English speakers. J Acoust Soc Am 129:340–349. CrossRef Medline
Öhman SEG (1966) Coarticulation in VCV utterances: spectrographic measurements. J Acoust Soc Am 39:151–168. CrossRef Medline
Ostry DJ, Gribble PL, Gracco VL (1996) Coarticulation of jaw movements in speech production: is context sensitivity in speech kinematics centrally planned? J Neurosci 16:1570–1579. Medline
Pei X, Leuthardt EC, Gaona CM, Brunner P, Wolpaw JR, Schalk G (2011) Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. Neuroimage 54:2960–2972. CrossRef Medline
Perkell JS (1986) Coarticulation strategies: preliminary implications of a detailed analysis of lower lip protrusion movements. Speech Commun 5:47–68.
Perkell JS, Nelson WL (1985) Variability in production of the vowels /i/ and /a/. J Acoust Soc Am 77:1889–1995. CrossRef Medline
Ray S, Maunsell JH (2011) Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLoS Biol 9:e1000610. CrossRef Medline
Rhodes BJ, Bullock D, Verwey WB, Averbeck BB, Page MP (2004) Learning and production of movement sequences: behavioral, neurophysiological, and modeling perspectives. Hum Mov Sci 23:699–746. CrossRef Medline
Schoppik D, Nagel KI, Lisberger SG (2008) Cortical mechanisms of smooth eye movements revealed by dynamic covariations of neural and behavioral responses. Neuron 58:248–260. CrossRef Medline
Shenoy KV, Sahani M, Churchland MM (2013) Cortical control of arm movements: a dynamical systems perspective. Annu Rev Neurosci 36: 337–359 CrossRef Medline
Sober SJ, Wohlgemuth MJ, Brainard MS (2008) Central contributions to acoustic variation in birdsong. J Neurosci 28:10370–10379. CrossRef Medline
Takai O, Brown S, Liotti M (2010) Representation of the speech effectors in the human motor cortex: somatotopy or overlap? Brain Lang 113:39–44. CrossRef Medline
Todorov E, Jordan MI (2002) Optimal feedback control as a theory of motor coordination. Nat Neurosci 5:1226–1235. CrossRef Medline
Watanabe A (2001) Formant estimation method using inverse-filter control. IEEE Trans Speech Audio Process 2001:317–326.
Whalen DH (1990) Coarticulation is largely planned. J Phonetics 18:3–35.
Wohlgemuth MJ, Sober SJ, Brainard MS (2010) Linked control of syllable sequence and phonology in birdsong. J Neurosci 30:12936–12949. CrossRef Medline
Yu BM, Cunningham JP, Santhanam G, Ryu SI, Shenoy KV, Sahani M (2009) Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. J Neurophysiol 102:614–635. CrossRef Medline