Title
Cortical encoding and decoding models of speech production

Author
Chartier, Josh

Publication Date
2019

Peer reviewed|Thesis/dissertation

Cortical encoding and decoding models of speech production

by

Josh Chartier

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
AND
UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:

DocuSigned by:
Edward Chang
4B5B40824E04415...

Edward Chang

Chair

DocuSigned by:
John Houde

John Houde

DocuSigned by:
Richard Ivry
51E0C0B952AA468...

Richard Ivry

Committee Members

Acknowledgments

I would like to thank my advisor, Dr. Edward Chang, for the opportunity to work under his mentorship, as well as my committee, Dr. John Houde, and Dr. Rich Ivry for their advice and encouragement.

I would also like to thank Dr. Gopala Anumanchipalli for working along side me, every step of the way. Lastly, I would like to thank the members of the Chang lab, my family, and friends for their encouragement through this exciting journey.

Material presented in this work is published in peer-reviewed journals:

Chartier, J.*, Anumanchipalli, G. K.*, Johnson, K., & Chang, E. F. (2018). Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. *Neuron*, 98(5), 1042-1054.

Anumanchipalli, G. K.*, Chartier, J.*, & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753), 493.

* Denotes equal contributions.

Abstract

Cortical encoding and decoding models of speech production

Josh Chartier

To speak is to dynamically orchestrate the movements of the articulators (jaw, tongue, lips, and larynx), which in turn generate speech sounds. It is an amazing mental and motor feat that is controlled by the brain and is fundamental for communication. Technology that could translate brain signals into speech would be transformative for people who are unable to communicate as a result of neurological impairments. This work first investigates how articulator movements that underlie natural speech production are represented in the brain. Building upon this, this work also presents a neural decoder that can synthesize audible speech from brain signals. Data to support these results were from direct cortical recordings of the human sensorimotor cortex while participants spoke natural sentences. Neural activity at individual electrodes encoded a diversity of articulatory kinematic trajectories (AKTs), each revealing coordinated articulator movements towards specific vocal tract shapes. The neural decoder was designed to leverage the kinematic trajectories encoded in the sensorimotor cortex which enhanced performance even with limited data. In closed vocabulary tests, listeners could readily identify and transcribe speech synthesized from cortical activity. These findings advance the clinical viability of using speech neuroprosthetic technology to restore spoken communication.

Table of Contents

List of Figures

Chapter 1. Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex

Chapter 2. Speech synthesis from neural decoding of spoken sentences

List of Tables

Chapter 2. Speech synthesis from neural decoding of spoken sentences

Introduction

Speaking to one another is a distinctly human process. It forms the basis for communication, which is perhaps the single most important action we do. Communication allows us to share experiences and to connect with one another. Without communication, we would also feel helpless and isolated from one another. Given the core significance of communication, it is remarkable how little we know about the extraordinary mechanisms that give rise to our ability to communicate, that is, our ability to speak.

At a fundamental level, speaking is the act of using our vocal tract to create vibrations in air to transmit information in the form of sounds to another individual. We accomplish this feat by coordinating the activations of nearly 100 muscles to dynamically position the speech articulators: lips, tongue, jaw, larynx, etc. First, we exhale air that blows across tensioned vocal folds; causing them to vibrate with a specific frequency we associate with pitch. Then with our supralaryngeal articulators, we further modulate the exhalation by superimposing additional perturbations. For instance, when creating a vowel sound, the jaw and tongue move to change the volume of the oral cavity to introduce resonant frequencies into the output sound. Consonants are formed by constricting the oral cavity at various points in time to disrupt the more continuous nature of these harmonic sounds. For example a /b/ sound is formed by blocking the oral cavity with the lips and then releasing the lips after some pressure is built up.

To speak, we must rapidly generate sequences of perceptually distinct sounds, called phonemes, thereby requiring immense coordination and rapid successive movements of the vocal tract. The sequences of phonemes must also adhere to a set of mutually understood rules that dictate how sounds can be combined to form words. Despite the vast complexity of

speaking, it often feels automatic. We rarely need to stop and consider what words we want to say and how to say each word. In fact, many of us often do not heed the proverb, “think before you speak.” This begs the question, “What processes occur in the human brain that enable the amazing mental and motor feat of speaking?” Over the years, this singular, unifying question has brought forth a collection of partial answers that build upon one another from psycholinguists, speech scientists, neuroscientists, and others alike. This dissertation offers another partial answer, which addresses the role of the sensorimotor cortex in relation to vocal tract movements during speaking through the lens of electrocortiographic (ECoG) recordings.

Like all methods of investigating brain activity, ECoG offers a unique and distinctly limited perspective of the brain. ECoG grants the capability of measuring multi-unit firing patterns of neuronal populations with high temporal precision and targeted spatial coverage, both of which are necessary for the rapid process of speaking. By using high-density ECoG electrode arrays, it is possible to measure the changes in voltages associated with neural firing at a distance of 4mm between electrodes, often in a grid of 16 x 16 electrodes. ECoG is an invasive neural recording technique because the electrodes are placed sub-durally, directly on the surface of the cortex offering direct insight to activity patterns of cortical neurons. ECoG grids are clinically implanted for use in epilepsy treatment to isolate seizure epicenters for resection. This dissertation is built upon data collected from patients undergoing this treatment who graciously volunteered their time to read several hundred sentences aloud while being monitored with ECoG.

ECoG grids are sometimes placed over brain areas important to speech production and can be used to help limit resection of areas crucial to speaking. One brain area involved in speech production is the sensorimotor cortex, which is often regarded as the command center for movement. It has long since been mapped out where each cortical section is associated with a

part of the human body. A natural progression from understanding where each body part is represented in the sensorimotor cortex has been to understand how the movement of each body part can be generated from the brain. The studies that followed often isolated their focus to understanding the relationship between single unit recordings in non-human primates and arm movements. Many fascinating theories have emerged, but little has been done to study the neural basis for movements of the vocal tract. Unlike arm reaching, movements of the inner vocal tract are notoriously difficult to observe, and neural recordings must be done in humans as opposed to non-human primates.

In Chapter 1, we made use of electromagnetic articulography (EMA) to study vocal tract movements. Sensors are placed on the upper and lower lips, on the tip, blade, and dorsum of the tongue, and on upper and lower incisors. A magnetic field is projected across the vocal tract and as the sensors move, an electrical current is induced that can be associated with their position in space. Unfortunately, due to clinical restrictions this recording apparatus could not be used to track the vocal tract movements of the subjects undergoing ECoG recordings. Instead, we used simultaneous recordings of EMA and audio recordings from other subjects to construct and inversion algorithm that could predict EMA traces from audio recordings. We then used the audio recordings from subjects with ECoG and audio recordings to infer what the EMA traces would be for these subjects. Finally, we fit a linear encoding model that predicted electrode activity from vocal tract movements. Individual electrodes encoded a diversity of articulatory kinematic trajectories (AKTs), each revealing coordinated articulator movements toward specific vocal tract shapes. AKTs captured a wide range of movement types, yet they could be differentiated by the place of vocal tract constriction. Additionally, AKTs manifested out-and-back trajectories with harmonic oscillator dynamics. While AKTs were functionally stereotyped across different sentences, context-dependent encoding of preceding and following movements

during production of the same phoneme demonstrated the cortical representation of coarticulation.

The articulatory movements encoded in sensorimotor cortex during speech production offered us clues into what a brain-computer interface for speaking might look like. Often when we think about interacting with a computer we think of typing on a keyboard or mobile phone. Naturally, we could expect that a brain-computer interface would allow one to type using brain signals. However, with the findings from Chapter 1 in mind, the speech sensorimotor cortex does not represent letters, but rather the vocal tract movements that make up a word. A brain-computer interface that directly decodes speech in terms of vocal tract movement would allow a user to communicate at the rate of natural speech production, a rate 3-4 times faster than typing on a keyboard.

In Chapter 2, we detail a proof-of-concept speech brain-computer interface that would restore spoken communication to people with severe speech and movement impairments. We again used ECoG signals from participants who underwent intracranial monitoring for epilepsy treatment as they spoke several hundreds of sentences aloud. We designed a neural decoder that explicitly leverages kinematic and sound representations encoded in human cortical activity to synthesize audible speech. Recurrent neural networks first decoded directly recorded cortical activity into representations of articulatory movement, and then transformed these representations into speech acoustics. In closed vocabulary tests, listeners could readily identify and transcribe speech synthesized from cortical activity. Intermediate articulatory dynamics enhanced performance even with limited data. Decoded articulatory representations were highly conserved across speakers, enabling a component of the decoder to be transferrable across participants. Furthermore, the decoder could synthesize speech when a participant silently

mimed sentences. These findings advance the clinical viability of using speech neuroprosthetic technology to restore spoken communication.

Chapter 1. Encoding of articulatory kinematics trajectories in human speech sensorimotor cortex

Josh Chartier,[1,2,3,5] Gopala K. Anumanchipalli,[1,2,5] Keith Johnson,[4] and Edward F. Chang[1,2,6]
[1]Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA 94158, USA [2]Department of Neurological Surgery, University of California, San Francisco, San Francisco, CA 94143, USA [3]Joint Program in Bioengineering, University of California, Berkeley and University of California, San Francisco, Berkeley, CA 94720, USA [4]Department of Linguistics, University of California, Berkeley, Berkeley, CA 94720, USA
[5]These authors contributed equally

Abstract When speaking, we dynamically coordinate movements of our jaw, tongue, lips, and larynx. To investigate the neural mechanisms underlying articulation, we used direct cortical recordings from human sensorimotor cortex while participants spoke natural sentences that included sounds spanning the entire English phonetic inventory. We used deep neural networks to infer speakers' articulator movements from produced speech acoustics. Individual electrodes encoded a diversity of articulatory kinematic trajectories (AKTs), each revealing coordinated articulator movements toward specific vocal tract shapes. AKTs captured a wide range of movement types, yet they could be differentiated by the place of vocal tract constriction. Additionally, AKTs manifested out-and-back trajectories with harmonic oscillator dynamics. While AKTs were functionally stereotyped across different sentences, context dependent encoding of preceding and following movements during production of the same phoneme demonstrated the cortical representation of coarticulation. Articulatory movements encoded sensorimotor cortex give rise to the complex kinematics underlying continuous speech production.

Introduction

To speak fluently, we perform an extraordinary movement control task by engaging nearly 100 muscles to rapidly shape and reshape our vocal tract to produce successive speech segments to form words and phrases. The movements of the articulators—lips, jaw, tongue, and larynx—are precisely coordinated to produce particular vocal tract patterns (Fowler et al., 1980; Bernstein, 1967). Previous research that has coded these movements by linguistic features (e.g. phonemes—well studied units of sound) has found evidence that the neural encoding in the ventral sensorimotor cortex (vSMC) is related to the presumed kinematics underlying speech sounds (Bouchard, et al., 2013, Lotte et al., 2015, Carey et al., 2017). However, there

are two key challenges that have precluded a complete understanding of how vSMC neural populations represent the actual articulatory movements underlying speech production.

The first challenge is to move beyond the experimentally convenient approach, taken in most studies, of studying the vSMC during isolated speech segments (Grabski et al., 2012, Bouchard, et al., 2013, Carey et al., 2017), towards studying the richer, complex movement dynamics in natural, continuous speech production. The second challenge is to go beyond categorical linguistic features (e.g. phonemes or syllables), towards describing the precise representations of movement, that is, the actual speech kinematics. Overcoming these challenges is critical to understanding the fluid nature of speech production. While speech is often described as the combination of discrete components with local invariances (i.e. vocal tract gestures (Browman & Goldstein, 1989) or phonemes), at any given time, the articulatory movements underlying the production of a speech segment may be influenced by previous and upcoming speech segments (known as coarticulation) (Hardcastle & Hewitt, 1999). For example, in “cool,” lip rounding necessary for /u/ is also present in /k/ while in “keep” /k/ is palatalized in anticipation of /i/. A central question remains as to whether cortical control invokes combinations of these primitive movement patterns to perform more complicated tasks (Bernstein, 1967, Bizzi et al., 1991, Bizzi & Cheung, 2013).

To address these challenges, we recorded high-density intracranial electrocorticography (ECoG) signals while participants spoke aloud full sentences. Our focus on continuous speech production allowed us to study the dynamics and coordination of articulatory movements not well captured during isolated syllable production. Furthermore, since a wide range of articulatory movements is possible in natural speech, we used sentences to cover nearly all phonetic and articulatory contexts in American English. Our approach allowed us to characterize sensorimotor cortical activity during speech production in terms of vocal tract movements.

A major obstacle to studying natural speech mechanisms is that the inner vocal tract movements can only be monitored for extended durations with specialized tools for tracking tongue movements with high spatial and temporal resolution, most of which are not practically compatible with intracranial recordings nor suitable for capturing naturalistic speech patterns. We overcame this obstacle by developing a statistical approach to derive the vocal tract movements from the produced acoustics. Then, we used the inferred articulatory kinematics to determine the neural encoding of articulatory movements, in a manner that was model independent and agnostic to pre-defined articulatory and acoustic patterns used in speech production (e.g. phonemes, gestures, etc.). By learning how combinations of articulator movements mapped to electrode activity, we estimated articulatory kinematic trajectories (AKTs) for single electrodes, and characterized the heterogeneity of movements that were represented through the speech vSMC.

Results

Inferring articulatory kinematics

To estimate the articulatory kinematics during natural speech production, we built upon recent advances in acoustic-to-articulatory inversion (AAI) to obtain reliable estimates of vocal tract movements from only the produced speech acoustics (Richmond, 2001; Afshan et al., 2015, Mitra et. al., 2017). While existing methods for AAI work well in situations where simultaneously recorded acoustic and articulatory data are available to train for the target speaker, there are few successful attempts for AAI in which no articulatory data is available from the target speaker. Specifically for this purpose, we developed an approach for Speaker-Independent Acoustic-to-Articulatory Inversion (AAI). We trained the AAI model using publicly available multi-speaker articulatory data recorded via Electromagnetic Midsagittal Articulography (EMA), a

reliable vocal tract imaging technique well suited to study articulation during continuous speech production (Berry, 2011). The training dataset comprised simultaneous recordings of speech acoustics and EMA data from 8 participants reading aloud sentences from the MOCHA-TIMIT dataset (Wrench, 1999; Richmond, et. al., 2011). EMA data for a speech utterance consisted of six sensors that tracked the displacement of articulators, critical to speech articulation (Figure 1.1A) in the caudo-rostral (x) and dorso-ventral (y) directions. We approximated laryngeal function by using the fundamental frequency (f0) of produced acoustics and whether or not the vocal folds were vibrating (voicing) during the production of any given segment of speech. In all, a 13 dimensional feature vector described articulatory kinematics at each time point (Figure 1.1B).

We modified the deep learning approach by Liu et. al., 2015 by incorporating phonological context to capture context dependent variance. Additionally, we spectrally warped training speakers to sound like the target (or test) speaker to improve cross-speaker generalizability (Toda et al., 2007). With these modifications, our AAI method performed markedly better than the current state-of-the-art methods within the speaker independent condition, and proved to be a reliable method to estimate articulatory kinematics. Using leave-one-participant-out cross validation, the mean correlation of inferred trajectories with ground truth EMA for a held out test participant was 0.68 ± 0.11 across all articulators and participants (0.53 correlation reported by Afshan et al., 2015). Figure 1.1B shows the inferred and ground truth EMA traces for each articulator during an example utterance for an unseen test speaker. There was a high degree of correlation across all articulators between the reference and inferred movements. Figure 1.9A shows a detailed breakdown of performance across each of the 12 articulators.

To investigate the ability of our AAI method to infer acoustically relevant articulatory movements, we trained identical deep recurrent networks to perform articulatory synthesis, i.e., predicting the acoustic spectrum (coded as 24 dimensional mel-cepstral coefficients and

energy) from articulatory kinematics, for both the real and inferred EMA. We found on average that there was no significant difference (p=.4, Figures S1B and C) in the resulting acoustic spectrum of unseen utterances when using either the target speaker's real EMA or those inferred via from the AAI method. This suggests that the difference between inferred and real EMA may largely be attributed to kinematic excursions that do not have significant acoustic effects. Other factors may also include differences in sensor placement, acquisition noise, and other speaker/recording specific artifacts that may not have acoustic relevance.

To further validate the AAI method, we examined how well the inferred kinematics preserved phonetic structure. To do so, we analyzed the phonetic clustering resulting from both real and inferred kinematic descriptions of phonemes. For one participant's real and inferred EMA, a 200 millisecond window of analysis was constructed around the kinematics for each phoneme onset. We then used linear discriminant analysis (LDA) to model the kinematic differences between phonemes from the real EMA data. We projected the both real and inferred EMA data for phonemes into this two dimensional LDA space to observe the relative differences in phonetic structure between real and inferred EMA. We found that the phonetic clustering and relative distances between phonemes centroids were largely preserved (Figure 1.1C) between inferred and real kinematic data (correlation r = 0.97 for consonants and 0.9 for vowels, p<.001). Together, these results demonstrate that using kinematic, acoustic, and linguistic metrics, it is possible to obtain high-resolution descriptions of vocal tract movements from easy-to-record acoustic data.

Encoding of articulatory kinematic trajectories at single vSMC electrodes

Using AAI, we inferred vocal tract movements as traces from EMA sensor locations (Figure 1.1A) while participants read aloud full sentences during simultaneous recording of acoustic and high-density intracranial electrocorticography signals. To describe the relationship between

vocal tract dynamics and sensorimotor cortical activity, we used a trajectory encoding model (Saleh et al., 2012) to predict each electrode's high gamma (70 – 150 Hz) activity (z-scored analytic amplitude) (Crone et al., 2001) as a weighted sum of articulator kinematics over time. Similar to models describing spectro-temporal receptive fields (Theunissen et al., 2001), a widely used tool to describe acoustic selectivity, we used ridge regression to model high gamma activity for a given electrode from time-varying estimated EMA sensor positions. In Figure 1.2, we show for an example electrode (Figure 1.2A), the weights learned (Figure 1.2C) from the linear model act as a spatio-temporal filter that we then convolved with articulator kinematics (Figure 1.2B) to predict electrode activity (Figure 1.2D).

The resulting filters described specific patterns of articulatory kinematic trajectories (AKTs) (Figure 1.2C), which are the vocal tract dynamics that best explain each electrode's activity. By validating on held-out data, we found that the AKT model significantly explained neural activity for electrodes active during speech in the vSMC (108 electrodes across 5 participants, mean $r = 0.25 \pm 0.08$ up to 0.5, $p<.001$) compared to AKT models constructed for electrodes in other anatomical regions ($p<.001$, Wilcoxon signed rank tests, Figure S7).

To provide a more intuitive understanding of these filters, we projected the X and Y coordinates of each trajectory onto a midsagittal schematic view of the vocal tract (Figure 1.2E). Each trace represents a kinematic trajectory of an articulator with a line that thickens with time to illustrate the time course of the filter. For the special case of the larynx, we did not estimate actual movements because they are not measured with EMA, and therefore used voicing-related pitch modulations that were represented along the y-axis with the x-axis providing a time course for visualization.

We observed a consistent pattern across articulators where each exhibited a trajectory that moved away from the starting point in a directed fashion before returning to the starting point. The points of maximal movement describe a specific functional vocal tract shape involving the coordination of multiple articulators. For example, the AKT (Figure 1.2E) for the electrode in Figure 1.2A exhibits a clear coordinated movement of the lower incisor and the tongue tip in making a constriction at the alveolar ridge. Additionally, the tongue blade and dorsum move frontward to facilitate the movement of the tongue tip. The upper and lower lips remain open and the larynx is unvoiced. The vocal tract configuration corresponds to the classical description of an alveolar constriction (e.g., production of /t/, /d/, /s/, /z/, etc.). The tuning of this electrode to this particular phonetic category is apparent in Figure 1.2D, where both the measured and predicted high gamma activity increased during the productions /st/, /dɪs/, and /nz/, all of which require an alveolar constriction of the vocal tract.

While vocal tract constrictions have typically been described as the action of one primary articulator, the coordination among multiple articulators is critical for achieving the intended vocal tract shape (Kelso, et al., 1984). For example, in producing a /p/, if the lower lip moves less than it usually does (randomly, or because of an obstruction) the upper lip compensates and the lip closure is accomplished (Abbs & Gracco, 1984). This coordination may arise from the complex and highly overlapping topographical organization of articulator representation in the vSMC (Meier et al., 2008, Grabski et al., 2012). We asked whether, like the coordinated limb movements encoded motor cortex (Aflalo & Graziano, 2006; Saleh et al., 2012), the encoded AKTs were the result of coordinated articulator movements. Alternatively, high gamma activity could be related to a single articulator trajectory with the rest of articulators representing irrelevant correlated movements. To evaluate these hypotheses, we used a cross-validated, nested regression model to compare the neural encoding of a single articulator trajectory with the AKT model. Here, we refer to one articulator as one EMA sensor. The models were trained

on 80% of the data and tested on the remaining 20% data. For each electrode, we fit single articulatory trajectory models using both X and Y directions for each estimated EMA sensor and chose the single articulator model that performed best for our comparison with the AKT model. Since each single articulator model is nested in the full AKT model, we used a general linear F-test to determine whether the additional variance explained by adding the rest of the articulators at the cost of increasing the number of parameters was significant. After testing each electrode on the data held-out from the training set, we found that the multi-articulatory patterns described by the AKT model explained significantly more variance compared to the single articulator trajectory model ($F(280, 1820)>1.31$, $p <.001$ for 96 of 108 electrodes, mean F-statistic=6.68, $p<.001$, Wilcoxon signed rank tests, Figure 1.11, mean change in R2: 99.55% ± 8.63%, Figure 1.12). This means that activity of single electrodes is more related to vocal tract movement patterns involving multiple articulators than those of a single articulator.

One potential explanation for this result is that single electrode neural activity in fact encodes the trajectory of a single articulator, but could appear to be multi-articulatory because of the correlated movements of other articulators due to the biomechanical properties of the vocal tract. While we would expect some coordination among articulator movements due to the intrinsic dynamics of the vocal tract, it is possible that further coordination could be cortically encoded. To evaluate these hypotheses, we examined the structure of correlations among articulators during periods of high and low neural activity for each speech-active electrode. If the articulator correlation structures were same regardless of electrode activity, the additional articulator movements were solely the result of governing biomechanical properties of the vocal tract. However, we found that articulator correlation structures differed according to whether high gamma activity was high or low (threshold at 1.5 standard deviations) ($p<.001$ for 108 electrodes, Bonferroni corrected) indicating that, in addition to coordination due to biomechanical properties of the vocal tract, coordination among articulators was reflected in

changes of neural activity. Contrary to popular assumptions of a one-to-one relationship between a given cortical site and articulator in the homunculus, these results demonstrate that, similar to cortical encoding of coordinated movements in limb control (Saleh, et al., 2012), neural activity at a single electrode encodes the specific, coordinated trajectory of multiple articulators.

Kinematic organization of vSMC

In our previous work, we used hierarchical clustering of electrode selectivity patterns to reveal the phonetic organization of the vSMC (Bouchard et al., 2013). We next wanted to examine whether clustering based upon all encoded movement trajectories, i.e. grouping of kinematically similar AKTs, yielded similar organization. Because the AKTs were mostly out-and-back in nature, we extracted the point of maximal displacement for each articulator along their principal axis of movement (see methods) to concisely summarize the kinematics of each AKT. We used hierarchical clustering to organize electrodes by their condensed kinematic descriptions (Figure 1.3A). To interpret the clusters in terms of phonetics, we fit a phoneme encoding model for each electrode. Similar to the AKT model, electrode activity was explained as a weighted sum of phonemes in which the value each phoneme was either 1 or 0 depending on whether it was being uttered at a given time. For each electrode, we extracted the maximum encoding weight for each phoneme. The encoded phonemes for each electrode were shown in the same order as the kinematically clustered electrodes (Figure 1.3B).

There was a clear organizational structure that revealed shared articulatory patterns among AKTs. The first level organized AKTs by their direction of jaw movement (lower incisor goes up or down). Sub-levels manifested four main clusters of AKTs with distinct coordinative articulatory patterns. The AKTs in each cluster were averaged together yielding a representative AKT for each cluster (Figure 1.3C). Three of the clusters described constrictions of the vocal

tract: coronal, labial, and dorsal, which broadly cover all consonants in English. The other cluster described a vocalic (vowel) AKT involving laryngeal activation and a jaw opening motion.

Instead of distributed patterns of electrode activity representing individual phonemes, we found that electrodes exhibited a high degree of specificity towards a particular group of phonemes. Electrodes within each AKT cluster also primarily encoded phonemes that had the same canonically defined place of articulation. For example, an electrode within the coronal AKT cluster was selective for /t/, /d/, /n/, /ʃ/, /s/, and /z/, all of which have a similar place of articulation. However, there were differences within clusters. For instance, within the coronal AKT cluster (Figures 3A and B, green), electrodes that exhibited a comparatively weaker tongue tip movement (less purple) had phonetic outcomes less constrained to phonemes with alveolar places of constriction (less black for phonemes in green cluster).

Hierarchical clustering was also performed on the phoneme encoding weights to identify phoneme organization to both compare with and help interpret the clustering of AKTs. These results confirm our previous description of phonetic organization of the vSMC (Bouchard, et al., 2013), as phonetic features defined by place-of-articulation were dominant. We found a strong similarity in clustering when electrodes were described by their AKTs and phonemes (Figures 3A and B), which is not surprising given that AKTs reflected specific locations of vocal tract constrictions (Figure 1.3C).

We observed broad groupings of electrodes that were sensitive to place-of-articulation, but within those groupings, we found differences in encoding for manner and voicing in consonant production. Within the coronal cluster, electrode encoding weights were highest for fricatives, then affricates, and followed by stops ($F(3) = 36.01$, $p < .001$, ANOVA). Conversely, bilabial stops were more strongly encoded than labiodental fricatives ($p < .001$, Wilcoxon signed rank

tests). Additionally, we found that consonants (excluding liquids) were clustered entirely separately from vowels. This is an important distinction from our previous work (Bouchard et al., 2013), where clustering was performed independently for the consonants and vowels in a CV syllable. Again, the vocalic AKTs were defined by both laryngeal action (voicing) and jaw opening configuration. Vowels were organized by three primary clusters which correspond to low vowels, mid/high vowels, and high front vowels.

To understand how kinematically and phonetically distinct each AKT cluster was from one another, we quantified the relationship between within-cluster and between-cluster similarities for each AKT cluster using the silhouette index as a measure of clustering strength (Figure 1.13). The degrees of clustering strength of AKT clusters for kinematic and phonetic descriptions were significantly higher compared to shuffled distributions indicating that clusters had both similar kinematic and phonetic outcomes ($p < .01$, Wilcoxon signed rank tests).

We also examined the anatomical clustering of AKTs across vSMC for each participant. While the anatomical clusterings for coronal and labial AKTs were significant ($p < .01$, Wilcoxon signed rank tests), clusterings for dorsal and vocalic AKTs were not. We found that only one participant had more than two dorsal AKT electrodes so we could not justly quantify the clustering strength of this cluster. Furthermore, vocalic AKTs were not well clustered because two spatial locations (dorsal and ventral LMC) were found, as previously seen in Bouchard et al., 2013. To further investigate the anatomical locations of AKT clusters, we projected electrode locations from all participants onto a common brain (Figure 1.4). Previous research has suggested that somatomotor maps of place of articulation are organized along the dorsal-ventral axis of the vSMC with labial constrictions were more dorsal and velar constrictions more ventral (Bouchard et al., 2013, Carey et al., 2017). We found that this coarse somatotopic organization was present for AKTs, which were spatially localized according to kinematic function and place

of articulation. Since AKTs encoded coordinated articulatory movements, we did not find single articulator localization. For example, with detailed descriptions of articulator movements, we found lower incisor movements were not localized to a single region, but rather opening and closing movements were represented separately as seen in vocalic and coronal AKTs, respectively.

Damped oscillatory dynamics of trajectories

Similar to motor cortical neurons involved in limb control, we found that the encoded kinematic properties were time-varying trajectories (Hatsopoulos et al., 2007). However, in contrast to the variety of trajectory patterns found during limb control from single neurons, we observed that each AKT exhibited an out-and-back trajectory from single ECoG electrode recordings. To further investigate the trajectory dynamics of every AKT, we analyzed phase portraits (velocity and displacement relationships) for each articulator. In Figure 1.5A, we show the encoded position and velocity of trajectories of each articulator, along its principal axis of displacement, for AKTs of 4 example electrodes, each representative of a main AKT cluster. The trajectory of each articulator was determined by the encoding weights from each AKT. All trajectories moved outwards and then returned to the same position as the starting point with corresponding increases and decreases in velocity forming a loop. This was true even for articulators that only made relatively small movements. In Figure 1.5B, we show the trajectories for each articulator from all 108 AKTs, which again illustrate the out-and-back trajectory patterns. Trajectories for a given articulator did not exhibit the same degree of displacement, indicating a level of specificity for AKTs within a particular cluster. Qualitatively, we observed that trajectories with more displacement also tended to correspond with high velocities.

While each AKT specifies time-varying articulator movements, the governing dynamics dictating how each articulator moves, may be time-invariant. In articulator movement studies, the time-

invariant properties of vocal tract gestures have been described by damped oscillatory dynamics (Saltzman & Munhall, 1989). Just like a pendulum, descriptors of movement (i.e. velocity and position) are related to one another independent of time. We found that there was a linear relationship between peak velocity and displacement for every articulator described by the AKTs (Figure 1.5C, r: 0.85, 0.77. 0.83, 0.69, 0.79, 0.83 in respective order, $p < .001$), demonstrating that AKTs also exhibited damped oscillatory dynamics. Furthermore, the slope associated with each articulator revealed the relative speed of that articulator. The lower incisor and upper lip moved the slowest (0.65 and 0.65 slopes) and the tongue varied in speed along the body with the tip moving fastest (0.66, 0.78, 0.99 slopes, respectively). These dynamics indicate that an AKT makes a stereotyped trajectory to form a single vocal tract configuration, a sub-syllabic speech component, acting as a building block for the multiple vocal tract configurations required to produce single syllables. While we were unable to dissociate whether the dynamical properties of single articulators were centrally planned or resulted from biomechanical properties of the vocal tract (Fuchs & Perrier, 2005), the velocity-position relationship strongly indicates that the AKT model encoded movements for each articulator corresponding to the intrinsic dynamics of continuous speech production.

Coarticulated kinematic trajectories

Some of the patterns observed in the detailed kinematics of speech result from interactions between successive vocal tract constrictions, a phenomenon known as coarticulation (Farnetani, 1997). Depending on the kinematic constraints of upcoming or previous vocal tract constrictions, some vocal tract constrictions may require anticipatory or carryover modifications to be optimally produced. Despite these modifications, each vocal tract constriction is often thought of as an invariant articulatory unit of speech production in which context-dependent kinematic variability results from the co-activation (i.e. temporal overlap) of vocal tract constrictions (Fowler, 1980; Browman & Goldstein, 1989; Saltzman & Munhall, 1989). We

investigated whether the vSMC shared similar invariant properties by studying how vSMC representations of vocal tract AKTs interacted with one another during varying degrees of anticipatory and carryover coarticulation.

During anticipatory coarticulation, kinematic effects of upcoming phonemes may be observed during the production of the present phoneme. For example, consider the differences in jaw opening (lower incisor goes down) during the productions of /æz/ (as in 'has') and /æp/ (as in 'tap') (Figure 1.6A). The production of /æ/ requires a jaw opening but the degree of opening is modulated by the upcoming phoneme. Since /z/ requires a jaw closure to be produced, the jaw opens less during /æz/ to compensate for the requirements of /z/. On the other hand, /p/ does not require a jaw closure and the jaw opens more during /æp/. In each context, the jaw opens during /æ/, but to differing degrees based the compatibility of the upcoming movement.
To investigate whether anticipatory coarticulation is neurally represented, we investigated the change in neural activity during the production /æz/ and /æp/, two contexts with differing degrees of coarticulation. While vSMC activity at the electrode population level is biased towards surrounding contextual phonemes (Bouchard & Chang, 2014), we investigated the representation of coarticulation at single electrodes. We studied high gamma of an electrode that encoded a vocalic AKT, crucial for the production of /æ/ (high phonetic selectivity index for /æ/, see methods). In Figure 1.6B, the AKT for electrode 120, describes a jaw opening and laryngeal vocal tract configuration. Time locked to the acoustic onset of /æ/, high gamma for electrode 120 was higher during /æp/ than /æz/ (Figure 1.6C). To quantify this difference, we compared the median high gamma activity during 50 ms centered at point of peak discriminability for all phonemes ($p<.05$, Wilcoxon signed rank tests). We also found that the predicted high gamma from the AKT was similarly higher during /æp/ than /æz/ ($p<.001$, Wilcoxon signed rank tests) (Figure 1.6D). For this electrode, we found that high gamma activity

reflected changes in kinematics, as predicted by the AKT, due to anticipatory coarticulation effects.

We then examined whether coarticulatory effects were present in all vSMC electrodes during all the anticipatory contexts of every phoneme. To quantify this effect, we fit a mixed-effects model to study how high gamma for a given electrode changed during the production of a phoneme with different following phonemes. In particular, we expected that for an electrode with an AKT heavily involved in producing a given phoneme, the kinematic compatibility of the following phoneme would be reflected in its peak high gamma. The model used cross-random effects to control for differences across electrodes and phonemes and a fixed effect of predicted high gamma from the AKT to describe the kinematic variability to which each electrode is sensitive. In Figure 1.6E, each line shows the relationship between high gamma and coarticulated kinematic variability for a given phoneme and electrode in all following phonetic contexts with at least 25 instances. For example, one line indicates how high gamma varied with the kinematic differences during /tæ/, /tɑ/, …, /ts/, etc. Kinematic variability due to following phonemes was a significant effect of the model indicating that neural activity associated with particular articulatory movements is modulated by the kinematic constraints of the following articulatory context (β = 0.30, SE = 0.04, $\chi 2(1) = 38.96$, p = 4e–10).

In a similar fashion, we also investigated the neural representation of carryover articulation, in which kinematic effects of previously produced phonemes are observed. In Figure 1.6F, we again show two coarticulated contexts with varying degrees of compatibility: /æz/ (as in ‘has’) and /iz/ (as in ‘ease’). /æ/ involves a large jaw opening while /i/ does not. However, in both contexts the jaw is equally closed for /z/ and the major difference between /æz/ and /iz/ is how much the jaw must move to make the closure. While the target jaw position for /z/ was achieved in both contexts, we found that for an electrode with a coronal AKT involved in producing /z/

(Figure 1.6G), the difference in high gamma reflected the kinematic differences between the two preceding phonemes (Figures 6H and I). Again, we used a mixed-effects model to examine the effects of carryover coarticulation in all vSMC electrodes to find that neural activity reflected carried-over kinematic differences in electrodes with AKTs for making the present phoneme (β = 0.32, SE = 0.04, $\chi2(1) = 42.58$, p = 6e–11) (Figure 1.6J). These results indicate that electrodes involved in producing a particular vocal tract configuration reflect kinematic variability due to anticipatory and carryover coarticulation.

Comparison with other encoding models

To evaluate how well AKTs are encoded in the vSMC, we compared i) the AKT model's encoding performance with respect to other cortical regions, and ii) vSMC encoding models for alternative representations of speech.

To determine how specific AKTs are to the vSMC, we compared AKT model performance (Pearson's r on held-out data) of every cortical region recorded from across participants (Figure 1.7A). Besides electrodes from middle frontal gyrus (MFG) and pars orbitalis (n = 4), the AKT model significantly explained some of the variance for all recorded cortical regions above chance level (p<.001, Wilcoxon rank-sum test). However, for the considered electrodes in this study (EIS)—i.e., the speech active electrodes in the vSMC—the AKT model explained neural activity markedly better than in other cortical areas (p<1e–15, Wilcoxon rank-sum test). The other cortical areas we examined were all previously shown to be involved in different aspects of speech processing: acoustic and phonological processing (STG & MTG) (Mesgarani et al., 2013), and articulatory planning (IFG) (Flinker et al., 2015). Therefore, it was expected that cortical activity in these regions would have some correlation to the produced kinematics. The higher performance of the AKT model for EIS indicates that studying the neural correlates of kinematics may best focused in the vSMC.

While AKTs were best encoded in vSMC, there may be alternative representations of speech that may better explain vSMC activity. We evaluated vSMC encoding of both acoustics (described here by using the first three formants: F1, F2, and F3) and phonemes with respect to the AKT model. Each model was fit in the same manner as the AKT model and performance compared on held-out data from training. If each vSMC electrode represented acoustics or phonemes, we would expect a higher model fit for that representation than the AKT model. Due to the similarity of these representations, we expected the encoding models to be highly correlated. It is worth noting that the inferred articulator movements are unable to provide an account of movements without correlations to acoustically significant events, a key property that would be invaluable for differentiating between models. Furthermore, while acoustics and phonemes are both complete representations of speech, the midsagittal movements of a few vocal tract locations captured by EMA are a partial description of speech relevant movements of the vocal tract in that we are missing palate, lateral and oropharyngeal movements. Even so, we found that articulator movements were encoded markedly better than both the acoustic and phoneme encoding models despite the limitations of the AKT model (Figure 1.7B & C, $p<1e–20$, Wilcoxon rank-sum test).

These comparisons were consistent with previous findings that vSMC encoding is tuned to articulatory features (Bouchard et al., 2013; Cheung et al., 2015). During single vowel production, vSMC showed encoding of directly measured kinematics over phonemes and acoustics (Conant et al., 2018). Furthermore, vSMC is also responsible for non-speech voluntary movements of the lips, tongue, and jaw, in behaviors such as swallowing, kissing, oral gestures. While vSMC is critical for speech production, it is not the only vSMC function. Indeed, when vSMC is injured, patients have facial and tongue weakness, in addition to dysarthria.

When vSMC is electrically stimulated, we observe movements -- not speech sounds, phonemes, or auditory sensations (Penfield & Boldrey, 1937; Breshears et al., 2015).

Decoding articulator movements

Given that we could determine encoding of AKTs at single electrodes, we next wanted to understand how well we could decode vocal tract movements from the population of electrodes. We decoded articulatory movements during sentence production with a long short-term memory recurrent neural network (LSTM), an algorithm well suited for time series regression (Hochreiter & Schmidhuber, 1997). The performance of the decoder was high, especially in light of the articulatory variance lost due to process of inferring kinematics and the neural variance unrecorded by the ECoG grid (i.e. within the central sulcus or at a resolution finer than the capability of the electrodes). For an example sentence (Figure 1.8A), the predicted articulator movements from the decoder closely matched with the inferred articulator movements from the acoustics. All of the articulator movements were well predicted across 100 held-out sentences significantly above chance (mean r: 0.43, $p<.001$) (Figure 1.8B). Prior work has demonstrated the possibility of decoding phonemes from ECoG recordings (Mugler et al., 2014) with automatic speech recognition techniques to decode full sentences (Herff et al., 2015) in addition to phrase classification with non-invasive recordings (Wang et al. 2017). Here, we show that decoding articulator movements directly from neural signals may be an additional approach for decoding speech.

Discussion

Our goal was to demonstrate how neural activity in human sensorimotor cortex represents the movements of vocal tract articulators during continuous speech production. We used a novel acoustic-to-articulatory inversion (AAI) method to infer vocal tract movements, which we then

related directly to high-resolution neural recordings. By describing vSMC activity with respect to detailed articulatory movements, we demonstrate that discrete neural populations encode articulatory kinematic trajectories (AKTs), a level of complexity that has not been observed using simpler syllable-level speech tasks in our previous work.

There are two important features of the AKTs that are encoded in the vSMC. First, encoded articulator movements are coordinated to make a specific vocal tract configuration. While the structure of coordination across articulators has been shown shown to be task-specific (e.g. different coordinative patterns during /p/ versus /z/) (Kelso et al., 1984), cortical control of this coordination has not been previously studied. However, studies in limb control have discovered single motor cortical neurons that encode complex coordinated movements involving both the arm and hand with specific functions (Aflalo & Graziano, 2006; Saleh et al., 2012). While previous studies have investigated vSMC activity on the basis of whether or not a given articulator is involved (Bouchard et al., 2013), we studied vSMC activity using detailed articulatory trajectories that suggest, similar to limb control, coordinated movements across articulators for specialized vocal tract configurations are encoded at the single electrode level. For example, the coordinated movement to close the lips is encoded rather than individual lip movements. This finding is consistent with studies where stimulation of localized neural populations in non-human primates has revealed functional action maps of complex arm and hand movements (Graziano et al., 2002). For speech, we found four major clusters of AKTs that were differentiated by place of articulation and covered the main vocal tract configurations that comprise American English. At the sampling level of ECoG, cortical populations encode sub-syllabic coordinative movements of the vocal tract.

The second important feature of AKTs is the trajectory profile itself. Encoded articulators moved in out-and-back trajectories with damped oscillatory dynamics. During limb control, single motor

cortical neurons have been also found to encode time-dependent kinematic trajectories, but the patterns were very heterogeneous and did not show clear spatial organization (Hatsopoulos et al., 2007). It is possible that individual neurons encode highly specific movement fragments that combine together to form larger movements represented by ensemble activity at the ECoG scale of resolution. For speech, these larger movements correspond to canonical vocal tract configurations. While motor cortical neurons encoded a variety of trajectory patterns, we found that AKTs only exhibited out-and-back profiles which may be a fundamental movement motif in continuous speech production.

With both coordinative and dynamical properties, each AKT appeared to encode the movement necessary to make a specific vocal tract configuration and return to a neutral position. Although we have described neural activity associated with articulatory movements without regard to any particular theory of speech production, the AKTs discovered here bear a striking resemblance to the vocal tract gestures theorized to be the articulatory units of speech production (Fowler et al., 1980; Browman & Goldstein, 1989). Each vocal tract gesture is described as a coordinated articulatory pattern to make a vocal tract constriction. Like the AKTs, each vocal tract gesture has been characterized as a time-invariant system with damped oscillatory dynamics (Saltzman and Munhall, 1989).

Articulatory theories suggest that each vocal tract gesture is an invariant unit and that the variability in the kinematics of continuous speech directly results from the temporal overlapping of successive gestures (Saltzman and Munhall, 1989). A particularly interesting phenomenon is that some vocal tract gestures are incompatible with one another in that the two vocal tract patterns require opposing movements of the articulators. This incompatibility results in a coarticulated compromise of target vocal tract patterns while compatible gestures are able to combine without inhibiting any necessary articulator movements (Farnetani, 1991; Farnetani &

Faber, 1992). Despite the theorized invariance of vocal tract gestures, we found that AKTs encoded in vSMC neural activity reflected kinematic differences due to constraints of the phonetic or articulatory context. While the invariant properties of vocal tract gestures may be represented elsewhere in higher order speech processes, the AKTs encoded in the vSMC represent coarticulation of successive AKTs.

The neural encoding of coarticulation also suggests that the vSMC does not locally encode phonemes. Phonemes by definition are segmental, perceptually defined, discrete units of sound. We would expect that an electrode encoding a particular set of phonemes as features would exhibit the same patterns of activation during the production of the same phoneme regardless of preceding or following phonemes and the accompanying kinematic constraints. However, we found that not only was there a difference in neural activity between productions of the same phoneme in different contexts, but also that the differences in kinematics partially explained the changes in neural activity. Furthermore, a direct comparison showed that AKTs were better encoded than both phoneme and acoustic models at single electrodes. We find the neural encoding of coarticulation to offer compelling support for AKTs as dominant features encoded in the speech sensorimotor cortex.

In summary, we described the cortical encoding of the movements underlying the rich dynamics of continuous speech production. These findings paint a new picture about the cortical basis of speech, and perhaps other sequential motor tasks. Coordinated articulator trajectories are locally encoded and fluidly combine while taking into account the surrounding movement context to produce the wide range of vocal tract movements we require to communicate. The insights gained by understanding the vSMC in terms of articulatory movements will help frame new questions of higher order planning and its realization as speech, or more broadly, movement.

Methods

Experimental model and subject details

Subjects. Five human subjects (Female, ages: 30, 31, 43, 46, 47) underwent chronic implantation of high-density, subdural electrode array over the lateral surface of the brain as part of their clinical treatment of epilepsy (2 left hemisphere grids, 3 right hemisphere grids). Subjects gave their written informed consent before the day of the surgery. No subjects had a history of any cognitive deficits that were relevant to the aims of the present study. All subjects were fluent in English. All procedures were approved by the University of California, San Francisco Institutional Review Board.

Method details

Experimental Task. Subjects read aloud 460 sentences from the MOCHA-TIMIT database (Wrench, 1999). Sentences were recorded in 9 blocks (8 of 50, and 1 of 60 sentences) spread across several days of patients' stay. Within each block, sentences are presented on a screen, one at a time, for the subject to read out. The order was random and subjects were given a few seconds of rest in between. MOCHA-TIMIT is a sentence-level database, a subset of the TIMIT corpus designed to cover all phonetic contexts in American English. Each subject read each sentence 1-10 times. Microphone recordings were obtained synchronously with the ECoG recordings.

Data acquisition and signal processing. Electrocorticography was recorded with a multi-channel amplifier optically connected to a digital signal processor (Tucker-Davis Technologies). Speech was amplified digitally and recorded with a microphone simultaneously with the cortical recordings. ECoG electrodes were arranged in a 16 x 16 grid with 4 mm pitch. The grid

placements were decided upon purely by clinical considerations. ECoG signals were recorded at a sampling rate of 3,052 Hz. Each channel was visually and quantitatively inspected for artifacts or excessive noise (typically 60 Hz line noise). The analytic amplitude of the high-gamma frequency component of the local field potentials (70 - 150 Hz) was extracted with the Hilbert transform and down-sampled to 200 Hz. Finally, the signal was z-scored relative to a 30 second window of running mean and standard deviation, so as to normalize the data across different recording sessions. We studied high-gamma amplitude because it has been shown to correlate well with multi-unit firing rates and has the temporal resolution to resolve fine articulatory movements (Crone et al., 2006).

Phonetic and phonological transcription. For the collected speech acoustic recordings, transcriptions were corrected manually at the word level so that the transcript reflected the vocalization that the subject actually produced. Given sentence level transcriptions and acoustic utterances chunked at the sentence level, hidden Markov model based acoustic models were built for each subject so as to perform sub-phonetic alignment (Prahallad et. al., 2006). Phonological context features were also generated from the phonetic labels, given their phonetic, syllabic and word contexts.

Speaker-Independent Acoustic-to-Articulatory Inversion (AAI). To perform articulatory inversion for a target subject for whom only acoustic data is available, we developed a method, we refer to as "Speaker-Independent AAI", where parallel EMA and speech data were simulated for the target speaker. In contrast to earlier approaches for speaker-independent AAI, where normalization is performed to remove speaker identity from acoustics, we accomplished the opposite goal of transforming the 8 EMA subjects' spectral properties to match those of the target speaker for whom we want to estimate vocal tract kinematics. To transform the acoustics of all data to the target speaker, we applied voice conversion (as proposed in Toda et al., 2007)

to transform the spectral properties of each EMA speaker to match those of the target subject. This method assumes acoustic data corresponding to the same sentences for the two subjects. When parallel acoustic data was not available across subjects in our case (the mngu0 corpus uses a different set of sentences than the MOCHA-TIMIT corpus), concatenative speech synthesis were used to synthesize comparable data across subjects (Hunt and Black '94).

Since there was no information about the target speaker's kinematics, we back off to using a subject and articulator normalized average of the 8 speakers' articulatory space. For cross-subject utilization of kinematic data, for each of the training speakers, we use an articulator specific z-scoring across each subject's EMA data. This ensured that the target speaker's kinematics were an unbiased average across all available EMA subjects. The kinematics were described by 13 dimensional feature vectors (12 dimensions to represent X and Y coordinates of 6 vocal tract points and fundamental frequency, F0, representing the Laryngeal function).

We used 24 dimensional mel-cepstral coefficients as the spectral features. Both kinematics and acoustics were sampled at a frequency 200 Hz (each feature vector represented a 5 ms segment of speech). Additionally, phonetic and phonological information corresponding to each frame of speech was coded as one-hot vectors and padded onto the acoustic features. These features included phoneme identity, syllable position, word part of speech, positional features of the current and of the neighboring phoneme and syllable states. We found that contextual data provided complementary information to acoustics and improved inversion accuracies.

Using these methods for each EMA subject-to-target subject pair, we were able to create a simulated dataset of parallel speech and EMA data, that were both customized for the target subject. For training the inversion model itself, we used a deep recurrent neural network based articulatory inversion technique (replicating Liu. et al., 2015) to learn a mapping from spectral

and phonological context to a speaker generic articulatory space. Following (Liu., et. al., 2015) an optimal network architecture with a 4 layer deep recurrent network with two feedforward layers (200 hidden nodes) and two bidirectional LSTM layers (with 100 LSTM cells) was chosen. The trained inversion model was then applied to all speech produced by the target subject to infer articulatory kinematics in the form of Cartesian X and Y coordinates of articulator movements. The network was implemented using Keras (Chollet et. al., 2015), a deep learning library running on top of a Tensorflow backend.

Electrode selection. We selected electrodes located on either the precentral and postcentral gyri that had distinguishable high gamma activity during speech production. We measured the separability of phonemes using the ratio of between-class to within-class variability (F statistic) for a given electrode across time. We chose electrodes with a maximum F statistic of 8 or greater. This resulted in a total of 108 electrodes across the 5 subjects with robust activity during speech production.

Encoding models. To uncover the kinematic trajectories represented in electrodes, we used linear encoding models to describe the high gamma activity recorded at each electrode as a weighted sum of articulator kinematics over time. This model is similar to the spectrotemporal receptive field, a model widely used to describe selectivity for natural acoustic stimuli (Theunissen et al., 2001). However, in our model, articulator X and Y coordinates are used instead spectral components. The model estimates the time series $x_i(t)$ for each electrode i as the convolution of the articulator kinematics A, comprised of kinematic parameters k, and a filter H, which we refer to as the articulatory kinematic trajectory (AKT) encoding of an electrode.

$$\hat{x}_i(t) = \sum_{k}^{K} \sum_{\tau}^{T} H_i(k,\tau) A(k, t-\tau)$$

Since our task was not designed to differentiate between motor commands and somatosensory feedback, we designed our filter to use a 500 ms window of articulator movements centered about the high gamma sample to be predicted. Movements occurring before the sample of high gamma are indicated by a negative lag while movements occurring after the high gamma sample are indicated by a positive lag. The 500 ms window was chosen to both maximize the performance of the AKT model (Figure 1.14) and allow full visualization of the AKTs. While Figure 1.14, indicates the filters need only be 200 ms long for optimal performance, we found that extending filters to 500 ms with appropriate regularization ensured that we could visualize every AKT in its entirety. Some AKTs encoded movements occurring well before or after the corresponding neural activity resulting AKTs cutoff using a 200 ms window. L2 regularization ensured that weights from time points not encoding an articulatory trajectory (e.g. at 250 ms before the neural sample) had no weighting and did not affect interpretability of the AKTs.

Additionally, we fit acoustic and phoneme encoding models to electrode activity. Instead of articulator X and Y coordinates, we used formants (F1, F2, and F3) as a description of acoustics and a binary description of the phonemes produced during a sentence. Each feature indicated whether a particular phoneme was being produced or not with a 1 or 0, respectively.

The encoding models were fit using ridge regression and trained using cross-validation with 70% of the data used for training, 10% of the data held-out for estimating the ridge parameter, and 20% held out as a final test set. The final test set consisted of sentences produced during entirely separate recording sessions from the training sentences. Performance was measured as the correlation between the predicted response of the model and the actual high gamma measured in the final test set.

Hierarchical clustering. We used Ward's method for agglomerative hierarchical clustering. Clustering of the electrodes was carried out solely on the kinematic descriptions for encoded kinematic trajectory of each electrode. To develop concise kinematic descriptions for each kinematic trajectory, we extracted the point of maximal displacement for each articulation. We used principal components analysis on each articulator to extract the direction of each articulator that explained the most variance. We then projected the filter weights onto each articulator's first principal component and chose the point with the highest magnitude. This resulted in length 7 vector with each articulator described by the maximum value of the first principal component. Phonemes were clustered based on the phoneme encoding weights for each electrode. For a given electrode, we extracted the maximum encoding weight for each phoneme during a 100 ms window centered at the point of maximum phoneme discriminability (peak F statistic) for the given electrode.

Cortical surface extraction and electrode visualization. To visualize electrodes on the cortical surface of a subject's brain, we used a normalized mutual information routine in SPM12 to co-register the preoperative T1 MRI with a postoperative CT scan containing electrode locations. We used Freesurfer to make pial surface reconstructions. To visualize electrodes across subjects on a common MNI brain, we performed nonlinear surface registration using a spherical sulcal-based alignment in Freesurfer, aligned to the cvs avg35 inMNI152 template (Fischl et al., 1999). While the geometry of the grid is not maintained, the nonlinear alignment ensures that electrodes on a gyrus in the subject's native space will remain on the same gyrus in the atlas space.

Decoding model. To decode articulatory movements, we trained a long short-term memory (LSTM) recurrent neural network to learn the mapping from high gamma activity to articulatory movements. LSTM are particularly well suited for learning mappings with time-dependent

information (Hochreiter & Urgen Schmidhuber, 1997). Each sample of articulator position was predicted by the LSTM using a window of 500 ms of high gamma activity, centered about the decoded sample, from all vSMC electrodes. The decoder architecture was a 4 layer deep recurrent network with two feedforward layers (100 hidden nodes each) and two bidirectional LSTM layers (100 cells). Using Adam optimization and dropout (40% of nodes), we trained the network to reduce mean squared error of the decoded and actual output. The network was implemented using Keras (Chollet et. al., 2015), a deep learning library running on top of a Tensorflow backend.

Quantification and statistical analyses

Nested encoding model comparison. We used a nested regression model to compare the neural encoding of a single articulator trajectory with the AKT model (Allen, 1997). For each electrode, we fit single articulatory trajectories models using both X and Y directions for each EMA sensor and chose the single articulator model that with the lowest residual sum of squares (RSS) on held-out data. From RSS values for the full (2) and nested (1) models, we compared the significance of the explained variance by calculating an F statistic for each electrode.

$$F = \frac{\left(\frac{RSS_1 - RSS_2}{p_2 - p_1}\right)}{\frac{RSS_2}{n - p_2}}$$

p and n are the number of model parameters and samples used in RSS computation, respectively. An F statistic greater than the critical value defined by the number of parameters in both models and confidence interval indicates that the full model (AKT) explains statistically significantly explains more variance than the nested model (single articulator) after accounting for difference in parameter numbers.

Correlation structure comparison. To test whether the correlational structure of articulators (EMA points) was different between periods of low and high gamma activity for a speech responsive electrode, we split the inferred articulator movements into two data sets based on whether the z-scored high gamma activity of given electrode for that sample was above the threshold (1.5). We then randomly sampled 1000 points of articulator movement from each data set to construct two cross-correlational structures between articulators. To quantify the difference between the correlational structures, we computed the Euclidean distance between the two structures. We then sampled an additional 1000 points from the below threshold data set to quantify the difference between correlational structures within the sub-threshold data. We repeated this process 1000 times for each electrode and compared the two distributions of Euclidean distances with a Wilcoxon rank sum test (Bonferroni corrected for multiple comparisons) to determine whether correlational structures of articulators differed in relation to high or low high gamma activity of an electrode.

Silhouette analysis. To assess cluster separability, we computed the silhouette index for each electrode to compare how well each electrode matched its own cluster based on the given feature representation. The silhouette index for an electrode is calculated by taking the difference between the average dissimilarity with all electrodes within the same cluster and the average dissimilarity with electrodes from the nearest cluster. This value is then normalized by taking the maximum value of the previous two dissimilarity measures. A silhouette index close to 1 indicates that the electrode is highly matched to its own cluster. 0 indicates that that the clusters may be overlapping, while -1 indicates that the electrode may be assigned to the wrong cluster.

Phoneme Selectivity Index (PSI). To determine the phoneme selectivity of each electrode, we use the statistical framework as described in Mesgarani et al., 2014 to test whether the high

gamma activity of an electrode is significantly different during the productions of two different phonemes. For a phoneme pair and a given electrode, we created two distributions of high gamma activity from data acoustically aligned to each phoneme. We used a 50 ms window of activity centered on the time point with the peak F statistic for that electrode. We used a non-parametric statistical hypothesis test (Wilcox rank-sum test) to assess whether these distributions have different medians ($p<0.001$). The PSI is the number of phonemes that have statistically distinguishable high gamma activity for a given electrode. A PSI of 0 indicates that no other phonemes have a distinguishable high gamma activity. Whereas, a PSI of 40 indicates that all other phonemes have distinguishable high gamma activity.

Mixed effects model. To examine the relationship between high gamma and coarticulated kinematics, we used a mixed-effects model with several crossed random effects. In particular, for a given electrode, we computed the "peak activity" by taking the median high gamma activity during a 50 ms window centered about the peak F statistic for that electrode (see PSI method) during the production of a target phoneme. We then took the mean peak activity for each unique phoneme pair (target phoneme preceded by context phoneme). For each electrode, we only considered phoneme pairs with at least 25 instances and a target PSI > 25. This helped stabilize the means and targeted electrodes that presumably encoded the AKT necessary to produce the target phoneme. In Figure 1.6C,D,H,I, we extended /z/ to include /z/ and /s/, and /p/ to include /p/ and /b/ since, from an EMA standpoint, the articulation is nearly identical and it increased the number of coarticulated instances we could analyze, thus decreasing biases from other contextual effects and variability from noise. In a similar fashion to high gamma, we computed high gamma activity predicted by the AKT model to provide insight into the kinematics during the production of a particular phoneme pair. Our mixed-effects model described high gamma from a fixed effect of kinematically predicted high gamma with crossed random effects (random slopes and intercepts) controlling for difference in electrodes, and target and context

phonemes (Barr et al., 2013). To determine model goodness, we used ANOVA to compare the model with a nested model that retained the crossed random effects but removed the fixed effect. The mixed-effects model was fit using the lme4 package in R (Baayen et al., 2008).

Acknowledgements

We thank Matthew Leonard, Neal Fox, Ben Dichter, Claire Tang, Jon Kleen, and Kristofer Bouchard for their helpful comments on the manuscript. This work was supported by grants from the NIH (DP2 OD008627 and U01 NS098971-01). E.F.C. is a New York Stem Cell Foundation-Robertson Investigator. This research was also supported by the New York Stem Cell Foundation, the Howard Hughes Medical Institute, the McKnight Foundation, the Shurl and Kay Curci Foundation, and the William K. Bowes Foundation. We grate- fully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

Author Contributions

Conception, J.C., G.K.A., and E.F.C.; AAI Programming, G.K.A.; Encoding Analyses, J.C.; Data Collection, G.K.A., E.F.C., and J.C.; Writing, J.C., G.K.A., K.J., and E.F.C.; Project Supervision, E.F.C.

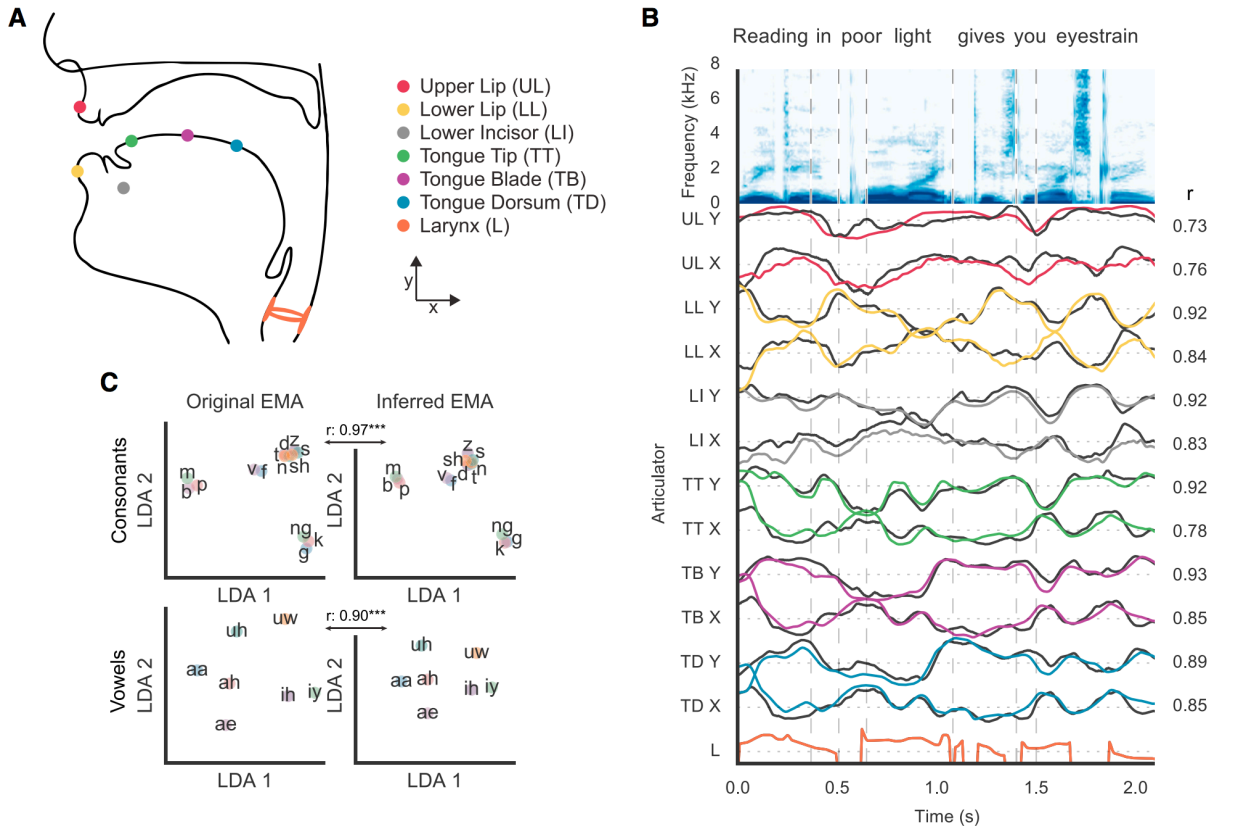


Figure 1.1. Inferred Articulator Kinematics

(**A**) Approximate sensor locations for each articulator during EMA recordings. Midsagittal movements represented as Cartesian x and y coordinates. (**B**) Midsagittal articulator movements inferred from both acoustic and phonetic features (in color). The trace of each reference sensor coordinate is also shown (in black). The larynx was approximated by fundamental frequency (f0) modulated by whether the segment of speech was voiced. (**C**) Recorded articulator movements (EMA) representing consonants and vowels projected into a low-dimensional (LDA) space. Inferred articulator movements projected into the same space were highly correlated with the original EMA. Correlations were pairwise distances between phonemes (consonants, $r = 0.97$, $p < 0.001$; vowels, $r = 0.90$, $p < 0.001$).

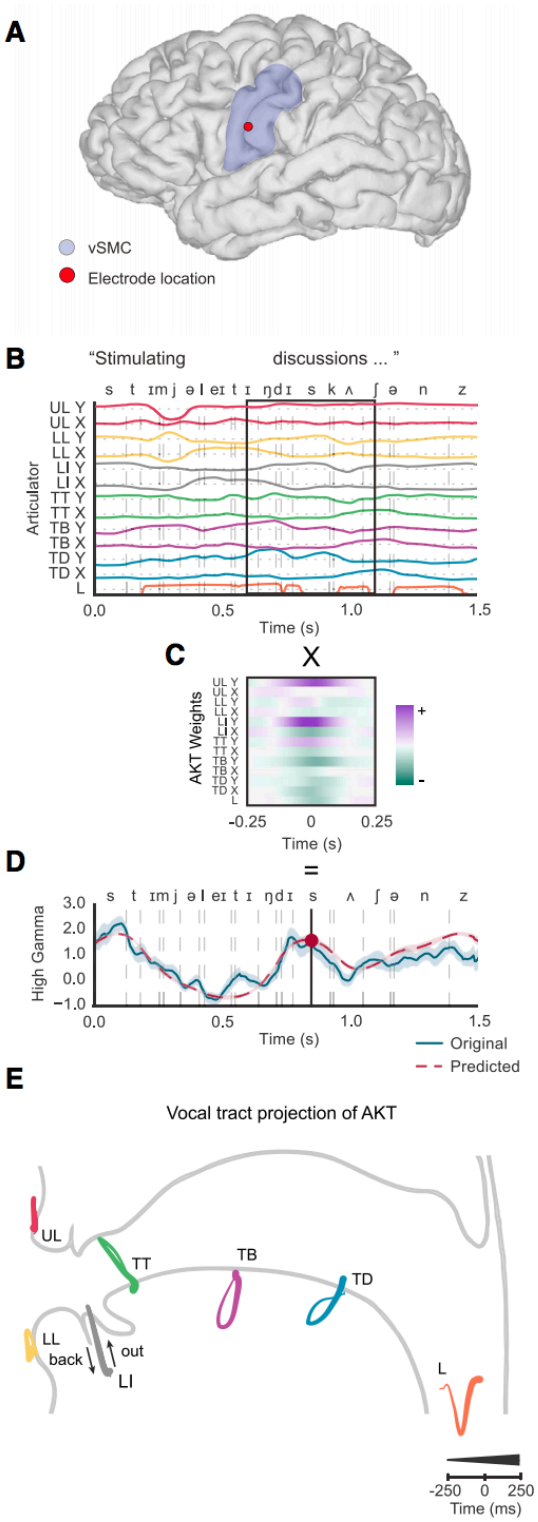


Figure 1. Neural Encoding of Articulatory Kinematic Trajectories

(**A**) Magnetic resonance imaging (MRI) reconstruction of single participant's brain where an example electrode is shown in the ventral sensorimotor cortex (vSMC). (**B**) Inferred articulator movements during the production of the phrase ''stimulating discussions.'' Movement directions are differentiated by color (positive x and y directions, purple; negative x and y directions, green), as shown in Figure 1.1A. (**C**) Spatiotemporal filter resulting from fitting articulator movements to explain high gamma activity for an example electrode. Time 0 represents the alignment to the predicted sample of neural activity. (**D**) Convolving the spatiotemporal filter with articulator kinematics explains high gamma activity as shown by an example electrode. High gamma from 10 trials of speaking ''stimulation discussions'' was dynamically time warped based on the recorded acoustics and averaged together to emphasize peak high gamma activity throughout the course of a spoken phrase. (**E**) Example electrode-encoded filter weights projected onto a midsagittal view of the vocal tract exhibits speech-relevant articulatory kinematic trajectories (AKTs). Time course of trajectories is represented by thin-to-thick lines. Larynx (pitch modulated by voicing) is one dimensional along the y axis, with the x axis showing time course.

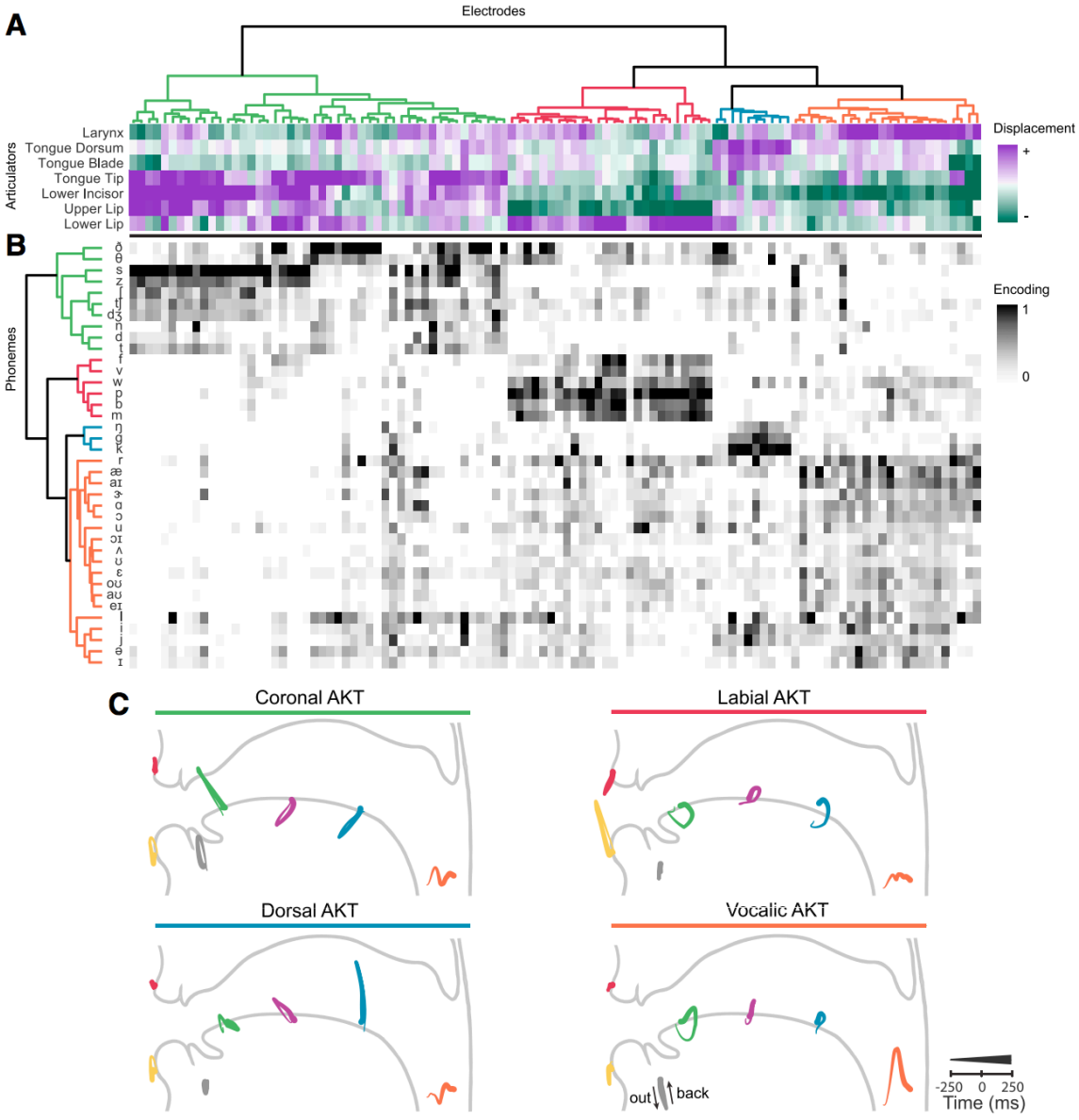


Figure 1.3. Clustered Articulatory Kinematic Trajectories and Phonetic Outcomes.

(**A**) Hierarchical clustering of encoded articulatory kinematic trajectories (AKTs) for all 108 electrodes across 5 participants. Each column represents one electrode. The kinematics of AKTs were described as a seven-dimensional vector by the points of maximal displacement along the principal movement axis of each articulator. Electrodes were hierarchically clustered by their kinematic descriptions resulting in four primary clusters. (**B**) A phoneme-encoding model was fit for each electrode. Kinematically clustered electrodes also encoded four clusters of encoded phonemes differentiated by place of articulation (alveolar, bilabial, velar, and vowels). (**C**) Average AKTs across all electrodes in a cluster. Four distinct vocal tract configurations encom- passed coronal, labial, and dorsal constrictions in addition to vocalic control.

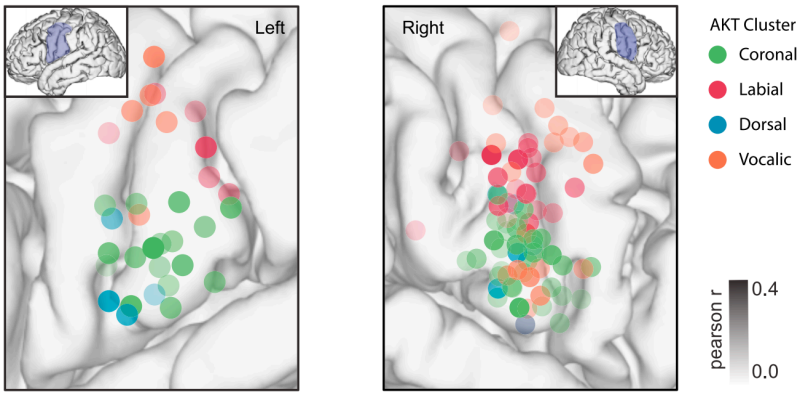


Figure 1.4. Spatial Organization of Vocal Tract Gestures.

Electrodes from five participants (two left and three right hemisphere) colored by kinematic cluster warped to the vSMC location on common MRI-reconstructed brain. Opacity of electrode varies with Pearson's correlation coefficient from the kinematic trajectory encoding model.

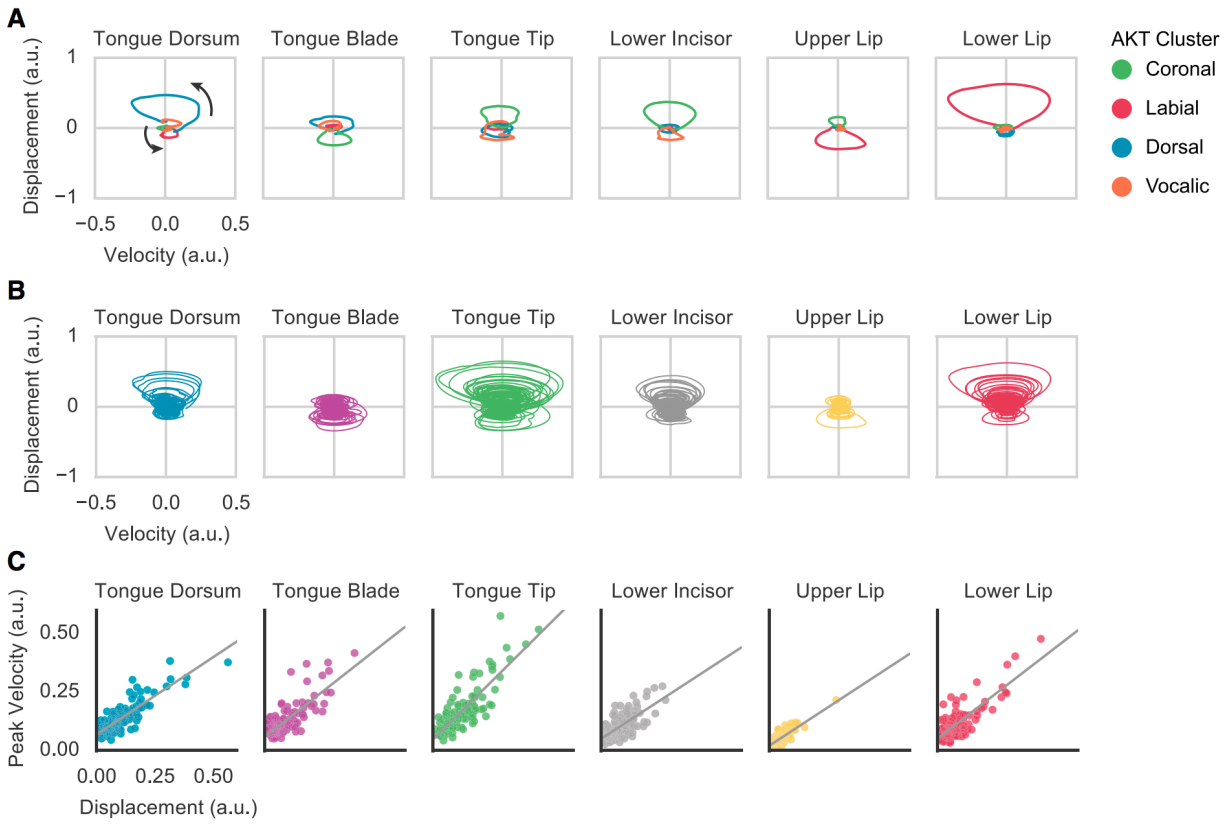


Figure 1.5. Damped Oscillatory Dynamics of Kinematic Trajectories.

(**A**) Articulator trajectories from encoded AKTs along the principal movement axes for example electrodes from each kinematic cluster. Positive values indicate a combination of upward and frontward movements. (**B**) Articulator trajectories for all 108 encoded kinematic trajectories across 5 participants. (**C**) Linear relationship between peak velocity and articulator displacement (r = 0.85, 0.77. 0.83, 0.69, 0.79, and 0.83 in respective order; $p < 0.001$). Each point represents the peak velocity and associated displacement of an articulator from the AKT for an electrode.

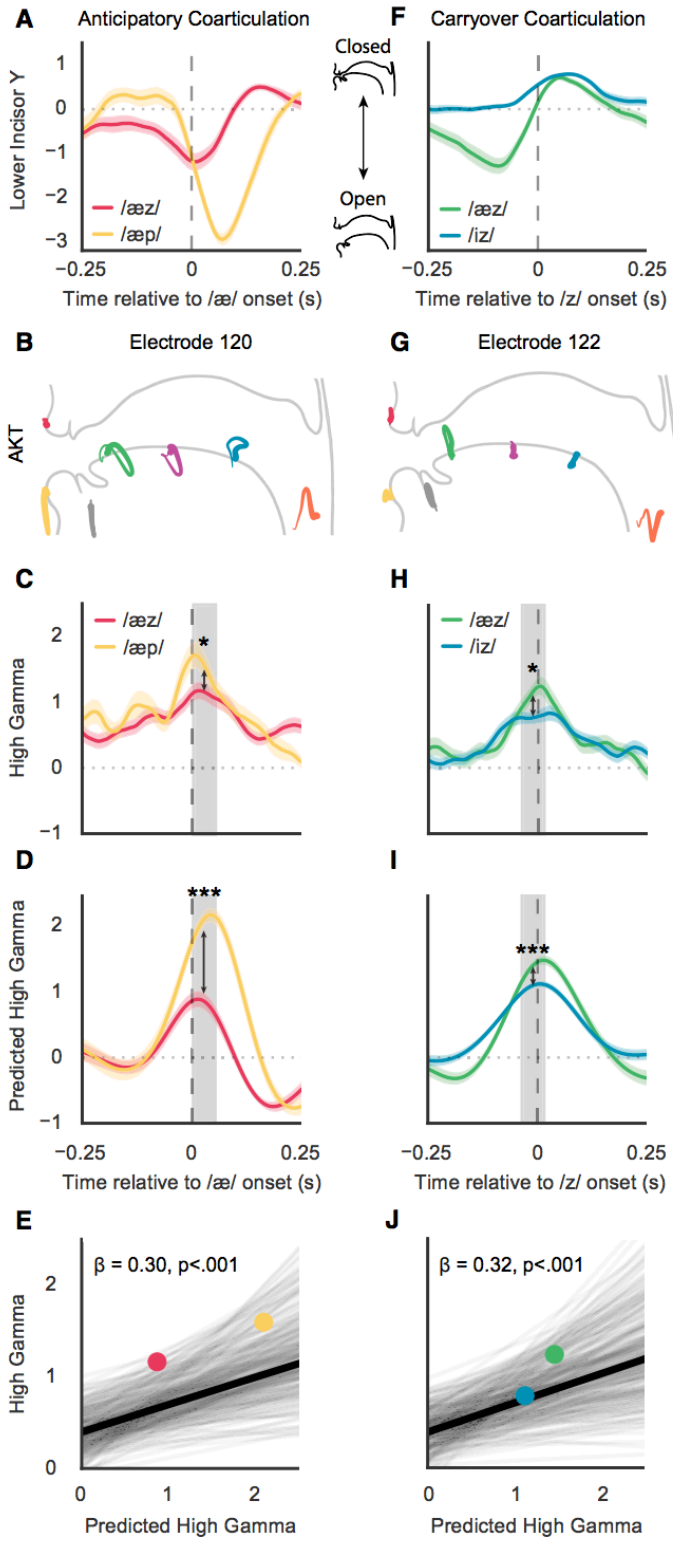


Figure 1.6. Neural Representation of Coarticulated Kinematics.

(**A**) Example of different degrees of anticipatory coarticulation for the lower incisor. Average traces for the lower incisor (y direction) are shown for /æz/ and /æp/ aligned to the acoustic onset of /æ/. (**B**) Electrode 120 is crucially involved in the production of /æ/ with a vocalic AKT (jaw opening and laryngeal control) and has a high phonetic selectivity index for /æ/. (**C**) Average high gamma activity for electrode 120 during the productions of /æz/ and /æp/. Median high gamma during 50 ms centered at the electrode's point of peak phoneme discriminability (gray box) is significantly higher for /æp/ than /æz/ ($p < 0.05$, Wilcoxon signed-rank tests). (**D**) Average predicted high gamma activity predicted by AKT in (B). Median predicted high gamma is significantly higher for /æp/ than /æz/ ($p < 0.001$, Wilcoxon signed-rank tests). (**E**) Mixed-effect model shows relationship of high gamma with kinematic variability due to anticipatory coarticulatory effects of following phonemes for all electrodes and phonemes (b = 0.30, SE = 0.04, $c^2(1) = 38.96$, p = 4e 10). Each line shows the relationship between high gamma and coarticulated kinematic variability for a given phoneme and electrode in all following phonetic contexts with at least 25 instances. Relationships from (C) and (D) for /æz/ (red) and /æp/ (yellow) are shown as points. Electrodes in all participants were used to construct the model. (**F**) Example of different degrees of carryover coarticulation for the lower incisor. Average traces for the lower incisor (y direction) are shown for /æz/ and /iz/ aligned to the acoustic onset of /z/. (**G**) Electrode 122 is crucially involved in the production of /z/ with a coronal AKT and has a high phonetic selectivity index for /z/. (**H**) Average high gamma activity for electrode 122 during the productions of /æz/ and /iz/. Median high gamma is significantly higher for /æz/ than /iz/ ($p < 0.05$, Wilcoxon signed-rank tests). (I) Average predicted high gamma activity predicted by AKT in (G). Median predicted high gamma is significantly higher for /æz/ than /iz/ ($p < 0.001$, Wilcoxon signed-rank tests). (**J**) Mixed-effect model shows relationship of hi gh gamma with kinematic variability due to carryover coarticulatory effects of preceding phonemes for all electrodes (in all participants) and phonemes (b = 0.32, SE = 0.04, $c^2(1) = 42.58$, p = 6e 11). Relationships from (H) and (I) for /æz/ (green) and /iz/ (blue) are shown as points.

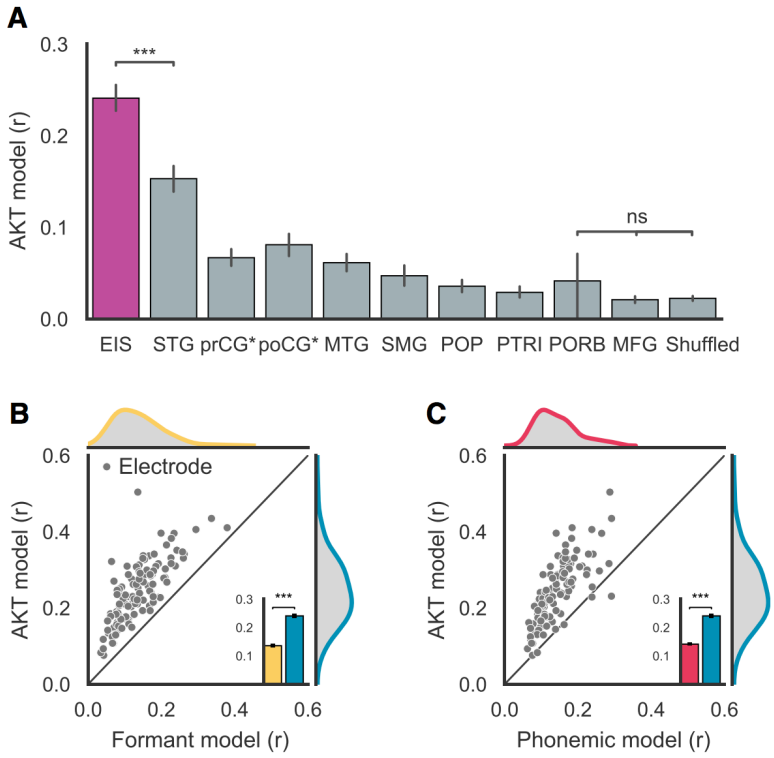


Figure 1.7. Neural-Encoding Model Evaluation.

(**A**) Comparison of AKT encoding performance across electrodes in different anatomical regions. Anatomical regions compared: electrodes in study (EIS), superior temporal gyrus (STG), precentral gyrus* (preCG*), postcentral gyrus* (postCG*), middle temporal gyrus (MTG), supramarginal gyrus (SMG), pars opercularis (POP), pars triangularis (PTRI), pars orbitalis (PORB), and middle frontal gyrus (MFG). Electrodes in study were speech selective electrodes from pre- and post-central gyri while preCG* and postCG* only included electrodes that were not speech selective. EIS encoding performance was significantly higher than all other regions ($p < 1e\ 15$, Wilcoxon signed-rank test). (**B**) Comparison of AKT- and formant-encoding models for electrodes in the study. Using F1, F2, and F3, the formant-encoding model was fit in the same manner as the AKT model. Each point represents the performance of both models for one electrode. (**C**) Comparison of AKT- and phonemic-encoding models. The phonemic model was fit in the same manner as the AKT model, except that phonemes were described as one hot vector. The best single phoneme predicting elec- trode activity was said to be the encoded phoneme of that particular electrode, and that r value was reported along with the r value of the AKT model. Pear- son's r was computed on held-out data from training for all models. In both comparisons, the AKT performed significantly better ($p < 1e\ 20$, Wilcoxon signed-rank test). Error bars represent SEM.

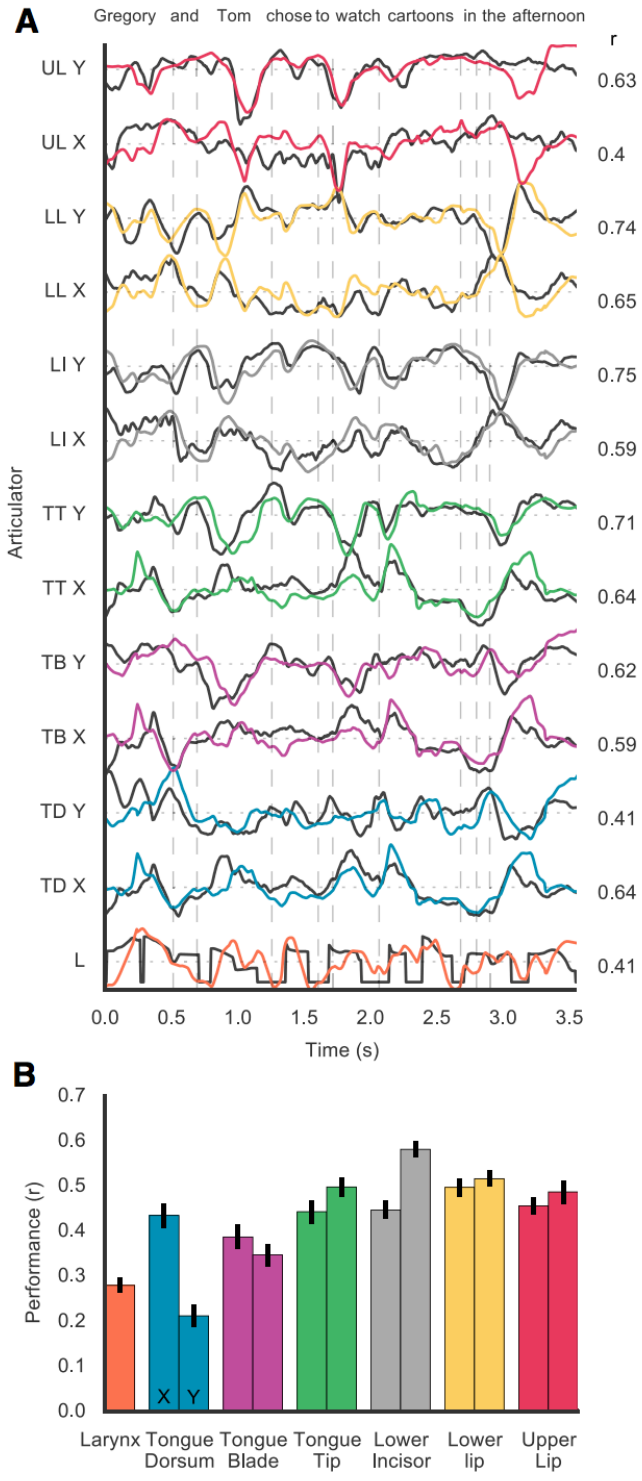
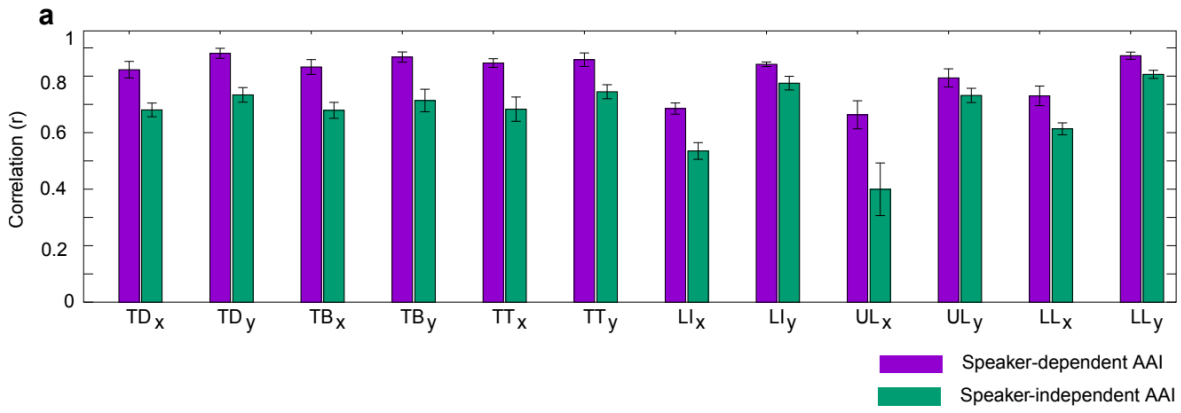
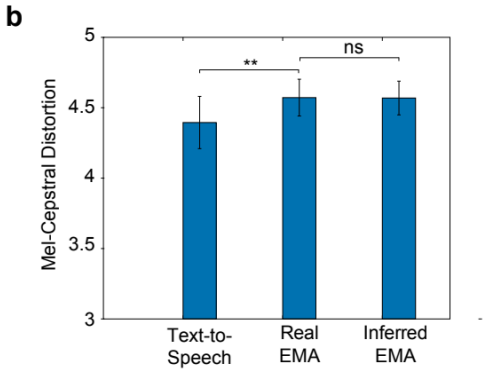


Figure 1.8. Decoded Articulator Movements from vSMC Activity.

(**A**) Original (black) and predicted (colored) x and y coordinates of articulator movements during the production of an example held-out sentence. Pearson's correlation coefficient (r) for each articulator trace. (**B**) Average performance (correlation) for each articulator for 100 sentences held out from training set. Error bars represent SEM.





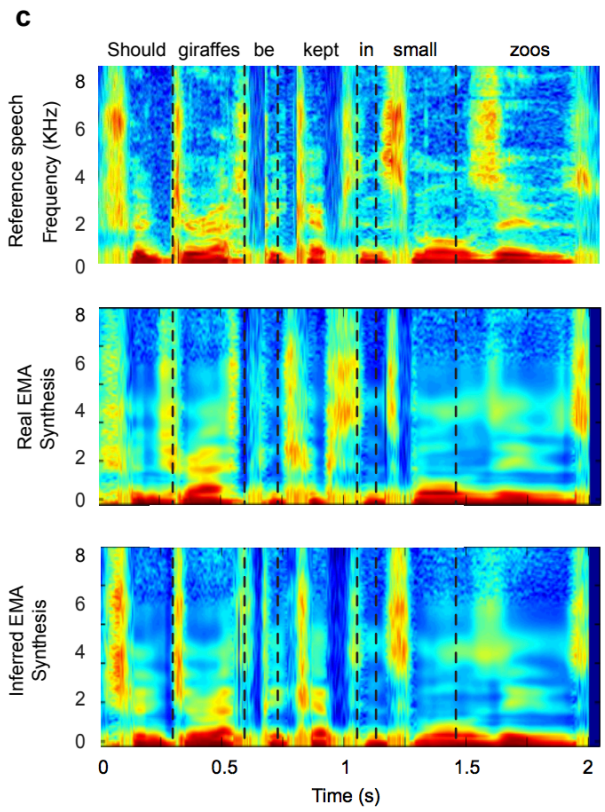


Figure S1. Acoustic-to-Articulatory Inversion, related to Figure 1. a, Articulator-wise breakdown of performance (pearson's correlation coefficient) of inferred articulatory trajectories. Speaker-dependent AAI refers to AAI using articulatory data from the speaker to train the model while speaker-independent AAI refers to AAI that does not use any articulatory data from the speaker to train the model (used in present study). **b**, Performance of articulatory synthesis in terms of mel-cepstral distortion comparing EMA based synthesis for both real and inferred EMA. General text-to-speech synthesis (phonemes, context, duration) shown for reference. **c**, Comparison of an example synthesized utterance based on real EMA and inferred EMA. Top most spectrogram is of the speaker's actual production for the utterance.

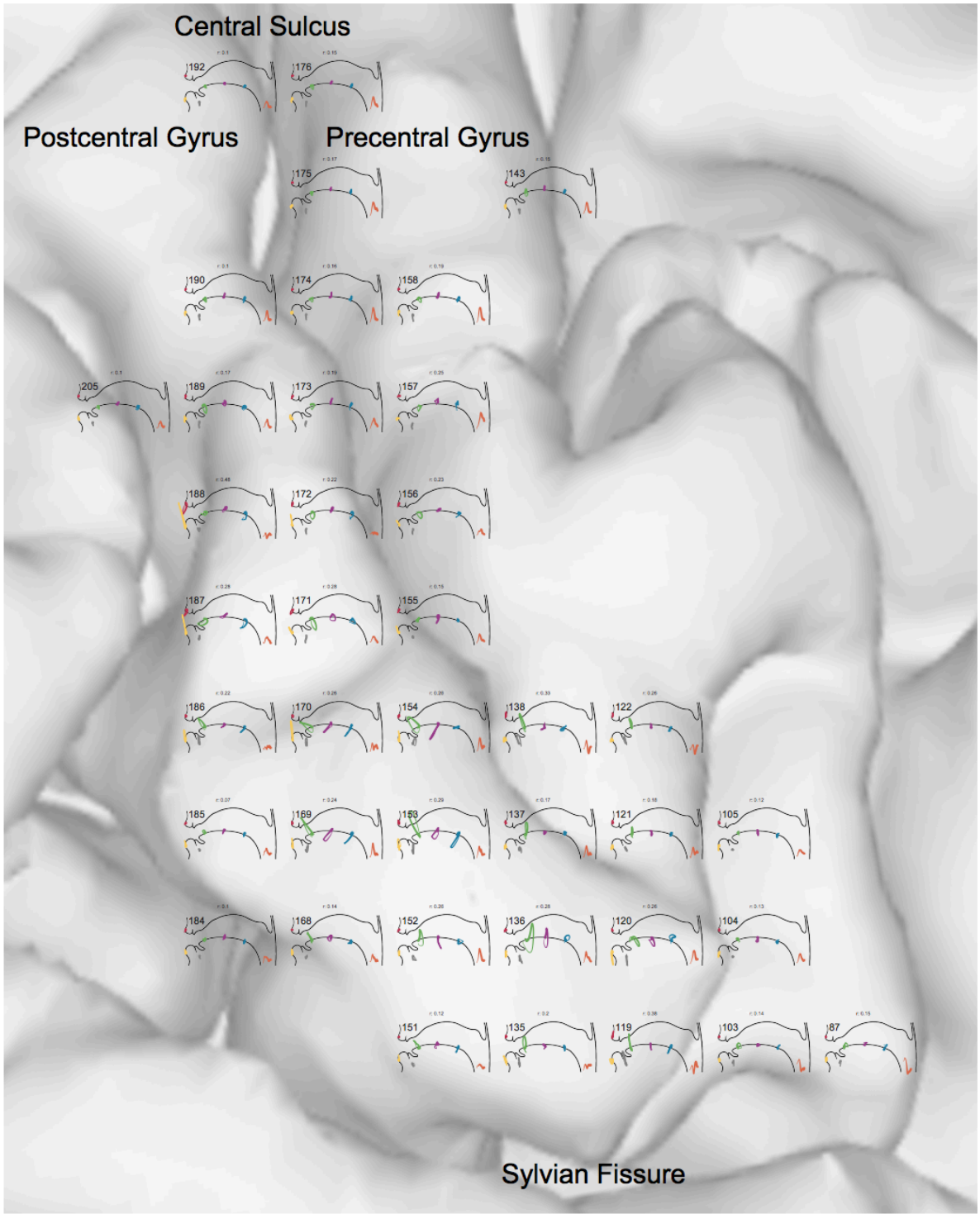


Figure 1.10. Articulatory kinematic trajectories for an individual subject, related to Figure 1.2.

For one subject with a right hemisphere ECoG grid, we plotted the encoded articulatory kinematic trajectories (AKTs) for each electrode in correspondence to its location on the cortical surface. Electrodes not active during speech were not shown. Each vocal tract plot shows the encoding model filter weights. Thin to thick lines indicate the time course of each articulator trajectory. We found AKTs were encoded in both the precentral and postcentral gyri. Furthermore,

AKTs with similar trajectory shapes appeared to spatially close to one another. r values for the correlation of each encoding model with high gamma are shown above the vocal tract plots.

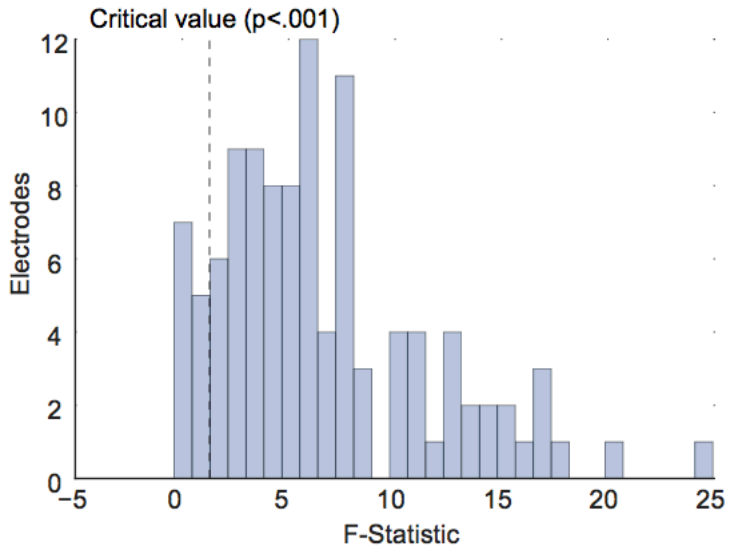


Figure 1.11. Electrodes encode coordinated articulatory movements involving multiple articulators, related to Quantification and Statistical Analyses. We used a nested regression model to compare whether the additional variance explained by trajectories of multiple articulatory was significant when compared to the variance explained by the trajectory of a single articulator. Since differences in parameter numbers can influence the explained variance solely by changing model complexity, we computed an F statistic on held-out data for each electrode to statistically test for model significance. Here, we plotted the distribution of F statistics and found that 96 out of 108 electrodes had F statistic greater than the critical value ($F(280, 1820)>1.31$, $p<.001$). The mean F statistic was 6.68 indicating that single electrodes encoded coordinated trajectories of
multiple articulators.

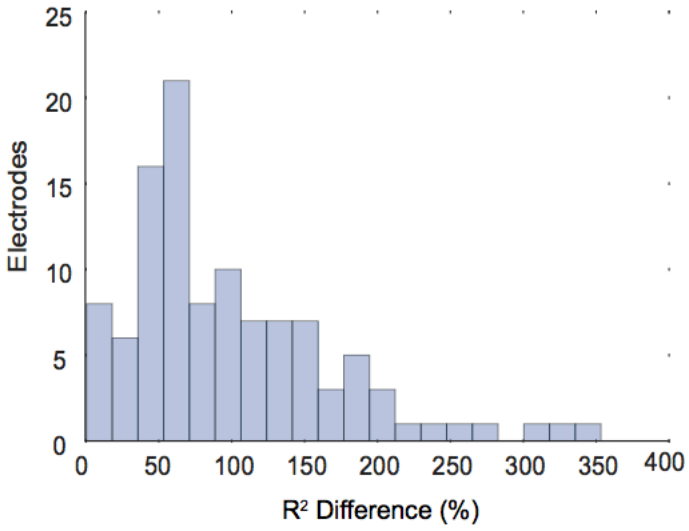


Figure 1.12. Comparison of variance explained by AKT model over single articulator model, related to Quantification and Statistical Analyses. The change in explained variance on held-out data by using the AKT model (all articulators) instead of the single articulatory trajectory model (X and Y for one articulator) is shown for each electrode as a percentage. The mean increase in explained variance from the AKT model was 99.55% +/- 8.63%.

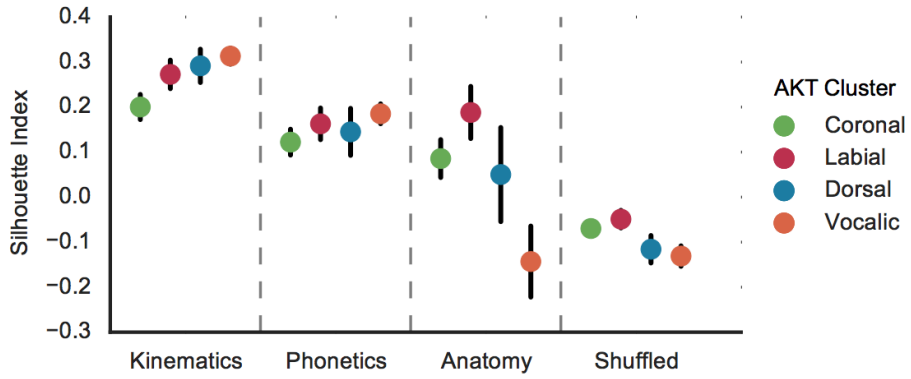


Figure 1.13. Silhouette analysis of clusters, related to Figures 3 and 4. We quantified the relationship between within-cluster and between-cluster similarities for each AKT cluster using the silhouetteindex as a measure of clustering strength. A silhouette index close to 1 indicates that the electrode is highly matched to its own cluster. 0 indicates that that the clusters may be overlapping, while -1 indicates that the electrode may be assigned to the wrong cluster. The degrees of clustering strength of AKT clusters for kinematic and phonetic descriptions were statistically significant above clusters of shuffled AKTs indicating that clusters had both similar kinematic and phonetic outcomes ($p < .01$, Wilcoxon signed rank tests). However, only the spatial clusterings for coronal and labial AKTs were statistically significant ($p < .01$, Wilcoxon signed rank tests). Only one subject had more than two dorsal AKT electrodes and we could not justly quantify the clustering strength of this cluster. The low silhouette index for anatomical clustering of the vocalic AKTs was expected because two spatial clusters were later found.

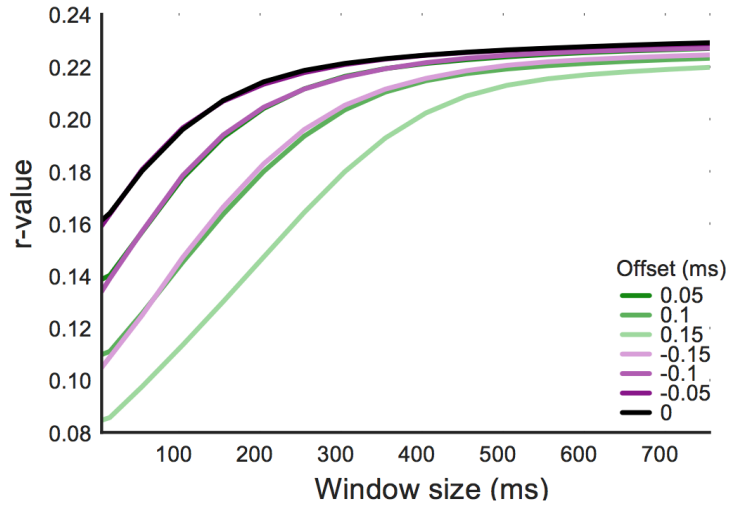


Figure 1.14. Effect of window size and offset from neural activity on AKT model performance, related to Method Details. Using varying window sizes of articulatory movements and offsets of those movements to electrode activity, the AKT model was fit and then tested on data held-out from training to compute mean correlations (r-value) between high gamma and predicted high gamma activity across all electrodes from all subjects in the study.

References

Abbs, J.H. & Gracco, V.L. (1984) Control of complex motor gestures: Orofacial muscles responses to load perturbation of the lip during speech. Journal of Neurophysiology, 51, 705-723.

Aflalo, T. N., & Graziano, M. S. (2006). Partial tuning of motor cortex neurons to final posture in a free-moving paradigm. Proceedings of the National Academy of Sciences of the United States of America, 103(8), 2909–14.

Allen, M.P. (1997) Testing hypotheses in nested regression models. In: Understanding Regression Analysis. Springer, Boston, MA

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255–278.

Bernstein, N. (1967). The co-ordination and regulation of movements. New York: Pergamon Press.

Berry, J.J. (2011) Accuracy of the NDI Wave Speech Research System. Journal of Speech, Language, and Hearing Research 54: 1295-1301.

Bizzi, E.,Mussa-Ivaldi,F. A.,and Giszter, S. (1991). Computations underly- ing the execution of movement: a biological perspective. Science 253, 287–291.

Bizzi, E., & Cheung, V. C. K. (2013). The neural origin of muscle synergies. Frontiers in Computational Neuroscience, 7(April), 51.

Bouchard, K. E., Mesgarani, N., Johnson, K., & Chang, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. Nature, 495(7441), 327–32.

Bouchard, K. E., & Chang, E. F. (2014). Control of spoken vowel acoustics and the influence of phonetic context in human speech sensorimotor cortex. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 34(38), 12662–77.

Breshears, J.D., Molinaro, A.M., & Chang, E.F. (2015). A probabilistic map of the human ventral sensorimotor cortex using electrical stimulation. Journal of Neurosurgery, 123: 340-349.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. Phonology, 6, 201-251.

Carey, D., Krishnan, S., Callaghan, M. F., Sereno, M. I., & Dick, F. (2017). Functional and Quantitative MRI Mapping of Somatomotor Representations of Human Supralaryngeal Vocal Tract. Cereb Cortex 2017; 27 (1): 265-278.

Cheung, C., Hamilton, L. S., Johnson, K., & Chang, E. F. (2016). The auditory representation of speech sounds in human motor cortex. Elife, 5.

Chollet, F., et al (2015), Keras, Github repository: https://github.com/fchollet/keras

Conant, D. F., Bouchard, K. E., Leonard, M. K., & Chang, E. F. (2018). Human sensorimotor cortex control of directly-measured vocal tract movements during vowel production. Journal of Neuroscience, 2382-17.

Crone, N. E., Sinai, A., & Korzeniewska, A. (2006). High-frequency gamma oscillations and human brain mapping with electrocorticography. Progress in brain research, 159, 275-295.

Crone, N. E., Hao, L., Hart, J., Boatman, D., Lesser, R. P., Irizarry, R., & Gordon, B. (2001). Electrocorticographic gamma activity during word production in spoken and sign language. Neurology, 57(11), 2045–2053.

Farnetani, E. (1991), Coarticulation and reduction in coronal consonants: Comparing isolated words and continuous speech. PERILUS XIV, Stockholm University, pp. 11-15.

Farnetani, E., & Faber, A. (1992). Tongue-jaw coordination in vowel production: Isolated words versus connected speech. Speech Communication, 11(4–5), 401–410.

Farnetani, E. (1997). Coarticulation and connected speech processes, The handbook of phonetic sciences, pp. 371–404.

Fischl, B., Sereno, M.I., Tootell, R.B.H., and Dale, A.M. (1999). High-Resolution Intersubject Averaging and a Coordinate System for the Cortical Surface. Hum. Brain Mapp. 8, 272-284.

Flinker, A., Korzeniewska, A., Shestyuk, A. Y., Franaszczuk, P. J., Dronkers, N. F., Knight, R. T., & Crone, N. E. (2015). Redefining the role of Broca's area in speech. Proceedings of the National Academy of Sciences, 112(9), 2871-2875.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. Journal of Phonetics.

Fowler, C. A., Rubin, P. E., Remez, R. E., & Turvey, M. T. (1980). Implications for speech production of a general theory of action. In B. Butterworth (Ed.), Language Production, Vol. I: Speech and Talk (pp. 373–420). New York: Academic Press.

Fuchs, S., & Perrier, P. (2005). On the complex nature of speech kinematics. ZAS Papers in Linguistics, 42, 137–165.

Grabski, K., Lamalle, L., Vilain, C., Schwartz, J. L., Vallée, N., Tropres, I., … Sato, M. (2012). Functional MRI assessment of orofacial articulators: Neural correlates of lip, jaw, larynx, and tongue movements. Human Brain Mapping, 33(10), 2306–2321.

Graziano, M. S. A., Taylor, C. S. R., & Moore, T. (2002). Complex movements evoked by microstimulation of precentral cortex. Neuron, 34(5), 841–851.

Hardcastle, W.J. & Hewlett, N. (1999). Coarticulation: Theory, Data, and Techniques. Cambridge University Press.

Hatsopoulos, N. G., Xu, Q., & Amit, Y. (2007). Encoding of movement fragments in the motor cortex. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 27(19), 5105–5114.

Herff, C., Heger, D., de Pesters, A., Telaar, D., Brunner, P., Schalk, G., & Schultz, T. (2015). Brain-to-text: decoding spoken phrases from phone representations in the brain. Frontiers in Neuroscience, 9(June), 1–11.

Hochreiter, S., & Urgen Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.

Liu, P., Yu, Q., Wu, Z., Kang, S., Meng, H., & Cai, L. (2015, April). A deep recurrent approach for acoustic-to-articulatory inversion. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on (pp. 4450-4454). IEEE.

Lotte, F., Brumberg, J. S., Brunner, P., Gunduz, A., Ritaccio, A. L., Guan, C., & Schalk, G. (2015). Electrocorticographic representations of segmental features in continuous speech. Frontiers in Human Neuroscience, 9(97), 1–13.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E.F. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. Science 1-6.

Meier, J. D., Aflalo, T. N., Kastner, S., & Graziano, M. S. (2008). Complex organization of human primary motor cortex: a high-resolution fMRI study. Journal of neurophysiology, 100(4), 1800-1812.

Mitra, V., Sivaraman, G., Bartels, C., Nam, H., Wang, W., Wilson, Á. C. E., … Park, M. (2017). Joint Modeling of Articulatory and Acoustic Spaces for Continuous Speech Recognition, ICASSP 5205–5209.

Mugler E. M., Patton J. L., Flint R. D., Wright Z. A, Schuele S. U., Rosenow J., Shih Jerry J., Krusienski D. J., Slutzky M. W. Direct classification of all American English phonemes using signals from functional speech motor cortex. Journal of Neural Engineering. 2014;11(3):035015.

Ostry, D.J.; Gribble, P.L. and Gracco, V.L (1996) Coarticulation of Jaw Movements in Speech Production: Is Context Sensitivity in Speech Kinematics Centrally Planned? J. Neuroscience 16 (4), 1570-9.

Penfield, W., & Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. Brain, 60(4), 389-443.

Prahallad, K., Black, A. W, and Mosur, R., (2006) Sub-Phonetic Modeling For Capturing Pronunciation Variations For Conversational Speech Synthesis, IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, 2006

Richmond, K. (2001) Estimating articulatory parameters from the acoustic speech signal. PhD Thesis, University of Edinburgh.

Richmond, K. (2011) Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In Proc. Interspeech, pages 1505-1508, Florence, Italy, August 2011

Saleh, M., Takahashi, K., & Hatsopoulos, N. G. (2012). Encoding of coordinated reach and grasp trajectories in primary motor cortex. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 32(4), 1220–32.

Saltzman, E. L. & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. Ecological Psychology, 1, 333-382.

Theunissen, F.E., David, S. V, Singh, N.C., Hsu, A., Vinje, W.E., and Gallant, J.L. (2001). Estimating spatiotemporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network 12, 289-316

Toda, T, Black, A. W, and Tokuda, K., "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222-2235.

Wang, J., Kim, M., Hernandez-Mulero, A. H., Heitzman, D., & Ferrari, P. (2017). Towards decoding speech production from single-trial Magnetoencephalography (MEG) signals, Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, 3036-3040.

Wrench, A. (1999) MOCHA: MultiChannel Articulatory database: English.

http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

Chapter 2. Speech synthesis of neural decoding from spoken sentences

Gopala K. Anumanchipalli,[1,2,4] Josh Chartier,[1,2,3,4] and Edward F. Chang[1,2]
[1]Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA 94158, USA [2]Department of Neurological Surgery, University of California, San Francisco, San Francisco, CA 94143, USA [3]Joint Program in Bioengineering, University of California, Berkeley and University of California, San Francisco, Berkeley, CA 94720, USA
[4]These authors contributed equally

Abstract Technology that translates neural activity into speech would be transformative for people who are unable to communicate as a result of neurological impairments. Decoding speech from neural activity is challenging because speaking requires very precise and rapid multi-dimensional control of vocal tract articulators. Here we designed a neural decoder that explicitly leverages kinematic and sound representations encoded in human cortical activity to synthesize audible speech. Recurrent neural networks first decoded directly recorded cortical activity into representations of articulatory movement, and then transformed these representations into speech acoustics. In closed vocabulary tests, listeners could readily identify and transcribe speech synthesized from cortical activity. Intermediate articulatory dynamics enhanced performance even with limited data. Decoded articulatory representations were highly conserved across speakers, enabling a component of the decoder to be transferrable across participants. Furthermore, the decoder could synthesize speech when a participant silently mimed sentences. These findings advance the clinical viability of using speech neuroprosthetic technology to restore spoken communication.

Introduction

Neurological conditions that result in the loss of communication are devastating. Many patients rely on alternative communication devices that measure residual nonverbal movements of the head or eyes (Fager et al., 2019), or on brain–computer interfaces (BCIs) (Brumberg et al., 2018; Pandarinath et al., 2017) that control a cursor to select letters one-by-one to spell out words. Although these systems can enhance a patient's quality of life, most users struggle to transmit more than 10 words per min, a rate far slower than the average of 150 words per min of natural speech. A major hurdle is how to overcome the constraints of current spelling-based approaches to enable far higher or even natural communication rates.

A promising alternative is to directly synthesize speech from brain activity (Guenther et al., 2009; Bocquelet et al., 2016). Spelling is a sequential concatenation of discrete letters, whereas speech is a highly efficient form of communication produced from a fluid stream of overlapping, multi-articulator vocal tract movements (Browman & Goldstein, 1992). For this reason, a biomimetic approach that focuses on vocal tract movements and the sounds that they produce may be the only means to achieve the high communication rates of natural speech, and is also likely to be the most intuitive for users to learn (Sadtler et al., 2014; Golub et al., 2018). In patients with paralysis—caused by for example, amyotrophic lateral sclerosis or brainstem stroke—high-fidelity speech-control signals may only be accessed by directly recording from intact cortical networks.

Our goal was to demonstrate the feasibility of a neural speech prosthetic by translating brain signals into intelligible synthesized speech at the rate of a fluent speaker. To accomplish this, we recorded high-density electrocorticography (ECoG) signals from five participants who underwent intracranial monitoring for epilepsy treatment as they spoke several hundreds of sentences aloud. We designed a recurrent neural network that decoded cortical signals with an explicit intermediate representation of the articulatory dynamics to synthesize audible speech.

Results

Speech decoder design

The two-stage decoder approach is shown in Fig. 2.2.1a–d. Stage 1, a bidirectional long short-term memory (bLSTM) recurrent neural network (Graves et al., 2005), decodes articulatory kinematic features from continuous neural activity (high-gamma amplitude envelope (Crone et al., 2001) and low frequency component (Nourski et al., 2015; Pesaran et al., 2018), see Methods) recorded from ventral sensorimotor cortex (vSMC) (Bouchard et al., 2013), superior

temporal gyrus (STG) (Mesgarani et al., 2014) and inferior frontal gyrus (IFG) (Flinker et al., 2015) (Fig. 2.1a, b). Stage 2, a separate bLSTM, decodes acoustic features (pitch (F0), mel-frequency cepstral coefficients (MFCCs), voicing and glottal excitation strengths) from the decoded articulatory features from stage 1 (Fig. 2.1c). The audio signal is then synthesized from the decoded acoustic features (Fig. 2.1d). To integrate the two stages of the decoder, stage 2 (articulation-to-acoustics) was trained directly on output of stage 1 (brain-to-articulation) so that it not only learns the transformation from kinematics to sound, but also corrects articulatory estimation errors made in stage 1.

A key component of our decoder is the intermediate articulatory representation between neural activity and acoustics (Fig. 2.1b). This step is crucial because the vSMC exhibits robust neural activations during speech production that predominantly encode articulatory kinematics (Chartier et al., 2018; Mugler et al., 2018). Because articulatory tracking of continuous speech was not feasible in our clinical setting, we used a statistical approach to estimate vocal tract kinematic trajectories (movements of the lips, tongue and jaw) and other physiological features (for example, manner of articulation) from audio recordings. These features initialized the bottleneck layer within a speech encoder–decoder that was trained to reconstruct a participant's produced speech acoustics (see Methods). The encoder was then used to infer the intermediate articulatory representation used to train the neural decoder. With this decoding strategy, it was possible to accurately reconstruct the speech spectrogram.

Synthesis performance

Overall, we observed detailed reconstructions of speech synthesized from neural activity alone (see Supplementary Video 1). Figure 2.1e, f shows the audio spectrograms from two original spoken sentences plotted above those decoded from brain activity. The decoded spectrogram retained salient energy patterns that were present in the original spectrogram and correctly

reconstructed the silence in between the sentences when the participant was not speaking. Figure 2.5a, b illustrates the quality of reconstruction at the phonetic level. Median spectrograms of original and synthesized phonemes—units of sound that distinguish one word from another—showed that the typical spectrotemporal patterns were preserved in the decoded examples (for example, resonant frequency bands in the spectrograms called formants F1–F3 in vowels /i:/ and /æ/; and key spectral patterns of mid-band energy and broadband burst for consonants /z/ and /p/, respectively).

To understand to what degree the synthesized speech was perceptually intelligible to naive listeners, we conducted two listening tasks that involved single-word identification and sentence-level transcription, respectively. The tasks were run on Amazon Mechanical Turk (see Methods), using all 101 sentences from the test set of participant 1.

For the single-word identification task, we evaluated 325 words that were spliced from the synthesized sentences. We quantified the effect of word length (number of syllables) and the number of choices (10, 25 and 50 words) on speech intelligibility, since these factors inform optimal design of speech interfaces (Huggins et al., 2011). Overall, we found that listeners were more successful at word identification as syllable length increased, and the number of word choices decreased (Fig. 2.2a), consistent with natural speech perception (Luce et al., 1998).

For sentence-level intelligibility, we designed a closed vocabulary, free transcription task. Listeners heard the entire synthesized sentence and transcribed what they heard by selecting words from a defined pool (of either 25 or 50 words) that included the target words and random words from the test set. The closed vocabulary setting was necessary because the test set was a subset of sentences from MOCHA-TIMIT (Wrench, 1999), which was primarily designed to

optimize articulatory coverage of English but contains highly unpredictable sentence constructions and low-frequency words.

Listeners were able to transcribe synthesized speech well. Of the 101 synthesized trials, at least one listener was able to provide a perfect transcription for 82 sentences with a 25-word pool and 60 sentences with a 50-word pool. Of all submitted responses, listeners transcribed 43% and 21% of the trials perfectly, respectively (Fig. 2.6). Figure 2b shows the distributions of mean word error rates (WER) of each sentence. Transcribed sentences had a median 31% WER with a 25-word pool size and 53% WER with a 50-word pool size. Table 2.1 shows listener transcriptions for a range of WERs. Median level transcriptions still provided a fairly accurate, and in some cases legitimate, transcription (for example, ‘mum’ transcribed as ‘mom’). The errors suggest that the acoustic phonetic properties of the phonemes are still present in the synthesized speech, albeit to the lesser degree (for example, ‘rabbits’ transcribed as ‘rodents’). This level of intelligibility for neurally synthesized speech would already be immediately meaningful and practical for real world application.

We then quantified the decoding performance at a feature level for all participants. In speech synthesis, the spectral distortion of synthesized speech from ground-truth is commonly reported using the mean mel-cepstral distortion (MCD) (Kominek et al., 2011). Mel-frequency bands emphasize the distortion of perceptually relevant frequency bands of the audio spectrogram (Davis & Mermelstein, 1990). We compared the MCD of neurally synthesized speech to a reference synthesis from articulatory kinematics and chance-level decoding (a lower MCD is better; Fig. 2.2c). The reference synthesis simulates perfect neural decoding of the kinematics. For our five participants (participants 1–5), the median MCD scores of decoding speech ranged from 5.14 dB to 6.58 dB ($P < 1 \times 10^{-18}$, Wilcoxon signed-rank test, for each participant).

We also computed the correlations between original and decoded acoustic features. For each sentence and feature, the Pearson's correlation coefficient was computed using every sample (at 200 Hz) for that feature. The sentence correlations between the mean decoded acoustic features (consisting of intensity, MFCCs, excitation strengths and voicing) and inferred kinematics across participants are plotted in Fig. 2.2d. Prosodic features such as pitch (F0), speech envelope and voicing were decoded well above the level expected by chance ($r > 0.6$, except F0 for participant 2: $r = 0.49$ and all features for participant 5; $P < 1 \times 10^{-10}$, Wilcoxon signed-rank test, for all participants and features in Fig. 2.2d). Correlation decoding performance for all other features is shown in Fig. 2.8a, b.

Decoder characteristics

The following analyses were performed on data from participant 1. When designing a neural decoder for clinical applications, there are several key considerations that determine model performance. First, in patients with severe paralysis or limited speech ability, training data may be very difficult to obtain. Therefore, we assessed the amount of data that would be necessary to achieve a high level of performance. We found a clear advantage in explicitly modelling articulatory kinematics as an intermediate step over decoding acoustics directly from the ECoG signals. The 'direct' decoder was a bLSTM recurrent neural network that was optimized for decoding acoustics (MFCCs) directly from same ECoG signals as used in an articulatory decoder. We found robust performance could be achieved with as little as 25 min of speech, but performance continued to improve with the addition of data (Fig. 2.2e). Without the articulatory intermediate step, the direct ECoG to acoustic decoding MCD was offset by 0.54 dB (0.2 dB is perceptually noticeable[21]) using the full dataset (Fig. 2.3a; $n = 101$, $P = 1 \times 10^{-17}$, Wilcoxon signed-rank test).

This performance gap between the two approaches persisted with increasing data sizes. One interpretation is that aspects of kinematics are more preferentially represented by cortical activity than acoustics (Chartier et al., 2018), and are thus learned more quickly with limited data. Another aspect that may underlie this difference is that articulatory kinematics lie on a low-dimensional manifold that constrains the potential high-dimensionality of acoustic signals (Sadtler et al., 2014; Golub et al., 2018; Gallego et al., 2017) (Fig. 2.9). Therefore, separating out the high-dimensional translation of articulation to speech, as done in stage 2 of our decoder may be critical for performance. It is possible that with sufficiently large datasets both decoding approaches would converge on one another.

Second, we wanted to understand the phonetic properties that were preserved in synthesized speech. We used Kullback–Leibler divergence to compare the distribution of spectral features of each decoded phoneme to those of each ground-truth phoneme to determine how similar they were (Fig. 2.10). We expected that, in addition to the same decoded and ground-truth phoneme being similar to one another, phonemes with shared acoustic properties would also be characterized as similar to one another.

Hierarchical clustering based on the Kullback–Leibler divergence of each phoneme pair demonstrated that phonemes were clustered into four main groups. Group 1 contained consonants with an alveolar place of constriction (for example, /s/ and /t/). Group 2 contained almost all other consonants (for example, /f/ and /g/). Group 3 contained mostly high vowels (for example, /i/ and /u/). Group 4 contained mostly mid and low vowels (for example, /ɑ/ and /æ/). The difference between groups tended to correspond to variations along acoustically significant dimensions (frequency range of spectral energy for consonants and formants for vowels). Indeed, these groupings explain some of the confusions reflected in listener transcriptions of these stimuli. This hierarchical clustering was also consistent with the acoustic similarity matrix

of only ground-truth phoneme pairs (Fig. 2.11; cophenetic correlation (Sokal & Rohlf, 1962) = 0.71, $P = 1 \times 10^{-10}$).

Third, because the success of the decoder depends on the initial electrode placement, we quantified the contribution of several anatomical regions (vSMC, STG and IFG) that are involved in continuous speech production (Brumberg et al., 2016. Decoders were trained in a leave-one-region-out fashion, for which all electrodes from a particular region were held out (Fig. 2.2f). Removing any region led to some decrease in decoder performance (Fig. 2.2g; $n = 101$, $P = 3 \times 10^{-4}$, Wilcoxon signed-rank test). However, excluding the vSMC resulted in the largest decrease in performance (MCD increase of 1.13 dB).

Fourth, we investigated whether the decoder generalized to novel sentences that were never seen in the training data. Because participant 1 produced some sentences multiple times, we compared two decoders: one that was trained on all sentences (not the particular instances in the test set), and one that was trained excluding every instance of the sentences in the testing set. We found no significant difference in decoding performance of the sentences for both MCD and correlations of spectral features ($P = 0.36$ and $P = 0.75$, respectively, $n = 51$, Wilcoxon signed-rank test; Fig. 2.12). Notably, this suggests that the decoder can generalize to arbitrary words and sentences that the decoder was never trained on.

Synthesizing mimed speech

To rule out the possibility that the decoder is relying on the auditory feedback of participants' vocalization, and to simulate a setting in which subjects do not overtly vocalize, we tested our decoder on silently mimed speech. We tested a held-out set of 58 sentences in which the participant 1 audibly produced each sentence and then mimed the same sentence, making the same articulatory movements but without making sound. Even though the decoder was not

trained on mimed sentences, the spectrograms of synthesized silent speech demonstrated similar spectral patterns to synthesized audible speech of the same sentence (Fig. 2.3a–c). With no original audio to compare, we quantified performance of the synthesized mimed sentences with the audio from the trials with spoken sentences. We calculated the spectral distortion and correlation of the spectral features by first dynamically time-warping the spectrogram of the synthesized mimed speech to match the temporal profile of the audible sentence (Fig. 2.3d, e) and then comparing performance. Although synthesis performance for mimed speech was inferior to the performance for audible speech—which is probably due to absence of phonation signals during miming—this demonstrates that it is possible to decode important spectral features of speech that were never audibly uttered ($P < 1 \times 10^{-11}$ compared to chance, n = 58; Wilcoxon signed-rank test) and that the decoder did not rely on auditory feedback.

State–space of decoded speech articulation

Our findings suggest that modelling the underlying kinematics enhances the decoding performance, so we next wanted to better understand the nature of the decoded kinematics from population neural activity. We examined low-dimensional kinematic state–space trajectories, by computing the state–space projection using principal components analysis onto the articulatory kinematic features. The first ten principal components (of 33 components in total) captured 85% of the variance and the first two principal components captured 35% (Fig. 2.9).

We projected the kinematic trajectory of an example sentence onto the first two principal components (Fig. 2.4a, b). These trajectories were well decoded, as shown in the example (Pearson's correlation: r = 0.91 and r = 0.91, principal components 1 and 2, respectively; Fig. 2.4a, b), and summarized across all test sentences and participants (median r > 0.72 for all participants except participant 5, where r represents the mean r of first two principal

components, Fig. 2.4e). Furthermore, state–space trajectories of mimed speech were well decoded (median r = 0.6, P = 1×10^{-5}, n = 38, Wilcoxon signed-rank test; Fig. 2.4e).

The state–space trajectories appeared to manifest the dynamics of syllabic patterns in continuous speech. The time courses of consonants and vowels were plotted on the state–space trajectories and tended to correspond with the troughs and peaks of the trajectories, respectively (Fig. 2.4a, b). Next, we sampled from every vowel-to-consonant transition (n = 22,453) and consonant-to-vowel transition (n = 22,453), and plotted 500-ms traces of the average trajectories for principal components 1 and 2 centred at the time of transition (Fig. 2.4c, d). Both types of trajectories were biphasic in nature, transitioning from the 'high' state during the vowel to the 'low' state during the consonant and vice versa. When examining transitions of specific phonemes, we found that principal components 1 and 2 retained their biphasic trajectories of vowel or consonant states, but showed specificity towards particular phonemes, indicating that principal components 1 and 2 do not necessarily describe only jaw opening and closing, but rather describe global opening and closing configurations of the vocal tract (Fig. 2.13). These findings are consistent with theoretical accounts of human speaking behaviour, which postulate that high-dimensional speech acoustics lie on a low-dimensional articulatory state–space (Browman & Goldstein, 1992).

To evaluate the similarity of the decoded state–space trajectories, we correlated productions of the same sentence across participants that were projected onto their respective kinematic state–spaces (only participants 1, 2 and 4 had comparable sentences). The state–space trajectories were highly similar (r > 0.8; Fig. 2.4f), suggesting that the decoder is probably relying on a shared representation across speakers, a critical basis for generalization.

A shared kinematic representation across speakers could be very advantageous for someone who cannot speak as it may be more intuitive and faster to learn to use the kinematics decoder (stage 1), while using an existing kinematics-to-acoustics decoder (stage 2) trained on speech data collected independently. We show synthesis performance when transferring stage 2 from a source participant (participant 1) to a target participant (participant 2) (Fig. 2.4g). The acoustic transfer performed well, although less than when both stage 1 and stage 2 were trained on the target (participant 2), probably because the MCD metric is sensitive to speaker identity.

Discussion

Here we demonstrate speech synthesis using high-density, direct cortical recordings from the human speech cortex. Previous strategies for neural decoding of speech production focused on direct classification of speech segments such as phonemes or words (Martin et al., 2014; Mugler et al., 2014); however, these approaches are generally limited in their ability to scale to larger vocabulary sizes and communication rates. Meanwhile, sensory decoding of auditory cortex has been promising for speech sounds (Herff et al., 2015; Moses et al., 2016; Pasley et al., 2012) or for auditory imagery (Akbari et al., 2019) in part because of the direct relationship between the auditory encoding of spectrotemporal information and the reconstructed spectrogram. An outstanding question has been whether motor decoding of vocal tract movements during speech production could be used for generating high-fidelity acoustic speech output.

Previous work focused on understanding movement that was encoded at single electrodes (Chartier et al., 2018); however, a fundamentally different challenge for speech synthesis is decoding the population activity that addresses the complex mapping between vocal tract movements and sounds. Natural speech production involves over 100 muscles and the

mapping from movement to sounds is not one-to-one. Our decoder explicitly incorporated this knowledge to simplify the translation of neural activity to sound by first decoding the primary physiological correlate of neural activity and then transforming to speech acoustics. This statistical mapping permits generalization with limited amounts of training.

Direct speech synthesis has several major advantages over spelling-based approaches. In addition to the capability to communicate unconstrained vocabularies at a natural speaking rate, it captures prosodic elements of speech that are not available with text output, such as pitch intonation (Dichter et al., 2018). Furthermore, a practical limitation for current alternative communication devices is the cognitive effort required to learn and use them. For patients in whom the cortical processing of articulation is still intact, a speech-based BCI decoder may be far more intuitive and easier to learn to use (Sadtler et al., 2014; Golub et al., 2018).

BCIs are rapidly becoming a clinically viable means to restore lost function. Neural prosthetic control was first demonstrated in participants without disabilities (Wessberg et al., 2000; Serruya et al., 2002; Taylor et al., 2002) before translating the technology to participants with tetraplegia (Hochberg et al., 2006; Collinger et al., 2013; Aflalo et al., 2015; Ajiboye et al., 2017). Our findings represent one step forward for addressing a major challenge posed by patients who are paralysed and cannot speak. The generalization results presented here demonstrate that speakers share a similar kinematic state–space representation that is speaker-independent and that it is possible to transfer model knowledge about the mapping of kinematics to sound across subjects. Tapping into this emergent, low-dimensional representation of neural activity from a coordinated population in the intact cortex may be a critical for bootstrapping a decoder (Gallego et al., 2017), as well facilitating BCI learning (Sadtler et al., 2014). Our results may be an important next step in realizing speech restoration for patients with paralysis.

Methods

Participants and experimental task. Five human participants (30 F, 31 F, 34 M, 49 F, 29 F) underwent chronic implantation of high-density, subdural electrode array over the lateral surface of the brain as part of their clinical treatment of epilepsy (right, left, left, left, left) hemisphere grids, respectively, Figure 2.11). Participants gave their written informed consent before the day of the surgery. All participants were fluent in English. All protocols were approved by the Committee on Human Research at UCSF and experiments/data in this study complied with all relevant ethical regulations. Each participant read and/or freely spoke a variety of sentences. P1 read aloud two complete sets of 460 sentences from the MOCHA-TIMIT (Wrench, 1999) database. Additionally, P1 also read aloud passages from the following stories: Sleeping Beauty, Frog Prince, Hare and the Tortoise, The Princess and the Pea, and Alice in Wonderland. P2 read aloud one full set of 460 sentences from the MOCHA-TIMIT database and further read a subset of 50 sentences an additional 9 times each. P3 read 596 sentences describing three picture scenes and then freely described the scene resulting in another 254 sentences. P3 also spoke 743 sentences during free response interviews. P4 read two complete sets of MOCHA-TIMIT sentences, 465 sentences drawn of scene descriptions and 399 sentences during free response interviews. P5 read one set of MOCHA-TIMIT sentences and 360 sentences of scene descriptions. In addition to audible speech, P1 also read 10 sentences 12 times each alternating between audible and silently mimed (i.e. making the necessary mouth movements) speech. Microphone recordings were obtained synchronously with the ECoG recordings.

Data acquisition and signal processing. Electrocorticography was recorded with a multi-channel amplifier optically connected to a digital signal processor (Tucker-Davis Technologies). Speech was amplified digitally and recorded with a microphone simultaneously with the cortical

recordings. The grid placements were decided upon purely by clinical considerations. ECoG signals were recorded at a sampling rate of 3,052 Hz. Each channel was visually and quantitatively inspected for artifacts or excessive noise (typically 60 Hz line noise). The analytic amplitude of the high-gamma frequency component of the local field potentials (70 - 200 Hz) was extracted with the Hilbert transform and down-sampled to 200 Hz. The low frequency component (1-30 Hz) was also extracted with a 5th order Butterworth bandpass filter, down-sampled to 200 Hz and parallelly aligned with the high-gamma amplitude. Finally, the signals were z-scored relative to a 30 second window of running mean and standard deviation, so as to normalize the data across different recording sessions. We studied high-gamma amplitude because it has been shown to correlate well with multi-unit firing rates and has the temporal resolution to resolve fine articulatory movements (Crone et al., 2001). We also included a low frequency signal component due to the decoding performance improvements note for reconstructing perceived speech from auditory cortex (Nourski et al., 2015; Pesaran et al., 2018). Decoding models were constructed using all electrodes from vSMC, STG, and IFG except for electrodes with bad signal quality as determined by visual inspection. We removed 8 electrodes for P1, 7 electrodes for P2, and 16 electrodes for P3. No electrodes were removed for P4 or P5. The decoder uses both high-gamma amplitude and raw low-frequency signals together as input to the model. For instance, n electrodes will result as n * 2 input features.

Phonetic and phonological transcription. For the collected speech acoustic recordings, transcriptions were corrected manually at the word level so that the transcript reflected the vocalization that the participant actually produced. Given sentence level transcriptions and acoustic utterances chunked at the sentence level, hidden Markov model based acoustic models were built for each participant so as to perform sub-phonetic alignment (Prahallad et al., 2006) within the Festvox (Anumanchipalli et al., 2006) framework. Phonological context features were also generated from the phonetic labels, given their phonetic, syllabic and word contexts.

Cortical surface extraction and electrode visualization. We localized electrodes on each individual's brain by co-registering the preoperative T1 MRI with a postoperative CT scan containing the electrode locations, using a normalized mutual information routine in SPM12. Pial surface reconstructions were created using Freesurfer. Final anatomical labeling and plotting was performed using the img_pipe python package (Hamilton et al., 2017).

Inference of articulatory kinematics. Among the most accurate methods to record vocal tract kinematics is called Electromagnetic Midsagittal Articulography (EMA). The process involves gluing small sensors to the articulators, generally 3 sensors on the tongue, 1 on each lip, 1 on each incisor. A magnetic field is projected at the participant's head and as the participant speaks, each sensor can be precisely tracked as it moves through the magnetic field. Each sensor has a wire leading out of the participant's mouth and connected to a receiver to record measurements.

Because of the above requirements, we did not pursue using EMA in the setting of our ECoG recordings because potential disruption of medical instruments by the magnetic field, long setup time conflicted with limited recording session time with patients, the setup procedure was too uncomfortable. Instead, we developed a model to infer articulatory kinematics from audio recordings. The articulatory data used to build the articulatory inference models was from MOCHA-TIMIT (Wrench, 1999) and MNGU0 corpora (Richmond et al., 2011).

The articulatory kinematics inference model comprises a stacked deep encoder-decoder, where the encoder combines phonological (linguistic and contextual features, resulting from the phonetic segmentation process) and acoustic representations (25 dimensional MFCC vectors sampled at 200 Hz) into a latent articulatory representation (also sampled at 200 Hz) that is then

decoded to reconstruct the original acoustic signal. The latent representation is initialized with inferred articulatory movement and appropriate manner features.

We performed statistical subject-independent acoustic-to-articulatory inversion (Chartier et al., 2918) to estimate 12 dimensional articulatory kinematic trajectories (x and y displacements of tongue dorsum, tongue blade, tongue tip, jaw, upper lip and lower lip, as would be measured by EMA) using only the produced acoustics and phonetic transcriptions. Since EMA features do not describe all acoustically consequential movements of the vocal tract, we append complementary speech features that improve reconstruction of original speech. First, to approximate laryngeal function, we add pitch, voicing (binary value indicating if a frame is voiced or not), and speech envelope, i.e., the frame level intensity computed as the sum total power within all the Mel scale frequencies within a 25 millisecond analysis window, computed at a shift of 5 milliseconds. Next, we added place-manner tuples (represented as continuous [0-1] valued features) to bootstrap the EMA with what we determined were missing physiological aspects in EMA. There were 18 additional values to capture the following place-manner feature tuples (palatal approximant, labial stop etc., see Supplemental Information (a) for the complete list). We used an existing annotated speech database (Wall Street Journal Corpus; Paul & Baker, 1992) and trained speaker independent deep recurrent network regression models to predict continuous valued place-manner vectors only from the acoustics features, the phonetic labels were used to determine the ground truth values for these labels (e.g., the dimension “labial stop” would be 1 for all frames of speech that belong to the phonemes /p/, /b/ and so forth). However, with a regression output layer, predicted values were not constrained to the binary nature of the input features. The network architecture was 3 feedforward layers followed by one bLSTM layer to predict each time point of these manner descriptors from a 100 millisecond window of acoustic features. Combined with the EMA trajectories, these 33 feature vectors form the initial articulatory feature estimates.

To ensure that the articulatory representation has the potential to reliably reconstruct speech for the target subject, we designed a stacked encoder-decoder network to optimize these initial estimates for these values. Specifically, a recurrent neural network encoder is trained to convert phonological and acoustic features to the articulatory representation and then a decoder that converts the articulatory representation back to the acoustic features (original MFCC). The encoder is implemented as 2 feedforward layers followed by 2 bLSTM layers. The decoder is implemented as 3 feedforward layers. Software implementation was done using Keras Functional API within Tensorflow (Martin et al., 2015). The stacked network is re-trained optimizing the joint mean squared error loss on acoustic and EMA parameters using the ADAM optimizer, with an initial learning rate set at 0.001. For regularization 40% dropout was allowed in all feedforward layers. After convergence, the trained encoder is used to estimate the final articulatory kinematic features that act as the articulatory intermediate to decode acoustic features from ECoG.

Neural decoder. The decoder maps ECoG recordings to MFCCs via a two stage process by learning intermediate mappings between ECoG recordings and articulatory kinematic features, and between articulatory kinematic features and acoustic features. All data (ECoG, kinematics, and acoustics) are sampled and processed by the model at 200 Hz. We implemented this model using TensorFlow in python. In the first stage, a stacked 3-layer bLSTM (Graves &Schmidhuber, 2005) learns the mapping between 300 ms (60 time points) sequences of high-gamma and LFP signals and a corresponding single time point (sampled at 200 Hz) of the 33 articulatory features. In the second stage, an additional stacked 3-layer bLSTM learns the mapping between the output of the first stage (decoded articulatory features) and 32 acoustic parameters (200 Hz) for full sentences sequences. These parameters are 25 dimensional MFCCs, 5 sub-band voicing strengths for glottal excitation modelling, log(F0), voicing.

During testing, a full sentence sequence of neural activity (high-gamma and low-frequency components) is processed by the decoder. The first stage processes 300 ms of data at a time, sliding over the sequence sample by sample, until it has returned a sequence of kinematics that is equal length to the neural data. The neural data is padded with an additional 150 ms of data before and after the sequence to ensure the result is the correct length. The second stage processes the entire sequence at once, returning an equal length sequence of acoustic features. These features are then synthesized into an audio signal.

At each stage, the model is trained using the Adam optimizer to minimize mean-squared error. The optimizer was initialized with learning rate=0.001,beta1=0.9, beta2=0.999, epsilon=1e-8. Models were stopped from training after the validation loss no longer decreased. Dropout rate is set to 50% in stage 1 and 25% in stage 2 to suppress overfitting tendencies of the models. There are 100 hidden units for each LSTM cell. Each model employed 3 stacked bLSTMs with an additional linear layer for regression. We use a bLSTM because of their ability to retain temporally distant dependencies when decoding a sequence (Hochreither & Schmidhuber, 1997).

In the first stage, the batch size for training is 256, and in the second stage the batch size is 25. Training and testing data were randomly split based off of recording sessions, meaning that the test set was collected during separate recording sessions from the training set. The training and testing splits in terms of total speaking time (minutes:seconds) are as follows: P1 – training: 92:15, testing: 4:46 (n=101); P2 – training: 36:57, testing: 3:50 (n=100); P3 – training: 107:42, testing: 4:44 (n=98); P4 – training: 27:39, testing 3:12 (n=82).; P5 – training 44:31, testing 2:51 (n=44). n=number of sentences in test set.

For shuffling the data to test for significance, we shuffled the order of the electrodes that were fed into the decoder. This method of shuffling preserved the temporal structure of the neural activity.

The "direct" ECoG to acoustics decoder described in Figure 2e a similar architecture as the stage 1 articulatory bLSTM except with an MFCC output. Originally we trained the direct acoustic decoder as a 6-layer bLSTM that mimics the architecture of the 2 stage decoder with MFCCs as the "intermediate layer" and as the output. However, we found performance was better with a 4-layer bLSTM (no intermediate layer) with 100 hidden units for each layer, 50% dropout and 0.005 learning rate using Adam optimizer for minimizing mean-squared error. Models were coded using Python's version 1.9 of Tensorflow.

Speech synthesis from acoustic features. We used an implementation of the Mel-log spectral approximation algorithm with mixed excitation (Maia et al., 2007) within Festvox to generate the speech waveforms from estimates of the acoustic features from the neural decoder.

Mel-Cepstral Distortion (MCD). To examine the quality of synthesized speech, we calculated the Mel-Cepstral Distortion (MCD) of the synthesized speech when compared the original ground-truth audio. MCD is an objective measure of error determined from MFCCs and is correlated to subjective perceptual judgments of acoustic quality (Kominek et al, 2008). For reference acoustic features $mc_d^{(y)}$ and decoded features $\hat{mc_d^{(y)}}$,

$$MCD = \frac{10}{ln(10)} \sqrt{\sum_{0<d<25} (mc_d^{(y)} - \hat{mc_d^{(y)}})^2}$$

Intelligibility Assessment. Listening tests using crowdsourcing are a standard way of evaluating the perceptual quality of synthetic speech (Wolters et al., 2010). To comprehensively assess the intelligibility of the neurally synthesized speech, we conducted a series of identification and transcription tasks on the Amazon Mechanical Turk. The unseen test set from P1 (101 trials of 101 unique sentences, shown in Supplemental Information (b)) was used as the stimuli for listener judgments. For the word level identification tasks, we created several cohorts of words grouped by the number of syllables within. Using the time boundaries from the ground truth phonetic labelling, we extracted audio from the neurally synthesized speech into four classes of 1-syllable, 2-syllable, 3-syllable and 4-syllable words. We conducted tests on each of these groups of words that involve identification of the synthesized audio from a group of i) 10 choices, ii) 25 choices, and iii) 50 choices of what they think the word is. The presented options included the true word and the remaining choices randomly drawn from the other words within the class (see Supplemental Information (c) for class sizes across these conditions). All words within the word groups were judged for intelligibility without any further sub-selection.

Since the content words in the MOCHA-TIMIT data are largely low frequency words to assess sentence-level intelligibility, along with the neurally synthesized audio file, we presented the listeners a pool of words that may be in the sentence. This makes it task a limited vocabulary free response transcription. We conducted two experiments where the transcriber is presented with pool of i) 25 word choices, and ii) 50 word choices that may be used the sentence (a sample interface is shown in Supplemental Information (d)). The true words that make up the sentence are included along with randomly drawn words from the entire test set and displayed in alphabetical order. Given that the median sentence is only 7 words long (std=21., min=4, max=13), this task design allows for reliable assessment of intelligibility. Each trial was judged by 10-20 different listeners. Each intelligibility task was performed by 47-187 unique listeners (a total of 1755 listeners across 16 intelligibility tasks, see supplemental information (e) for

breakdown per task) making all reported analyses statistically reliable. All sentences from the test set were sent for intelligibility assessment without any further selection. The listeners were required to be English speakers located in the United States, with good ratings (>98% rating from prior tasks on the platform). For the sentence transcription tasks, an automatic spell checker was employed to correct misspellings. No further spam detection, or response rejection was done in all analyses reported. Word Error Rate (WER) metric computed on listener transcriptions is used to judge the intelligibility of the neurally synthesized speech. Where I is the number of word insertions, D is the number of word deletions and S is the number of word substitutions for a reference sentence with N words, WER is computed as

$$WER = \frac{I + D + S}{N}$$

Data limitation analysis. To assess the amount of training data affects decoder performance, we partitioned the data by recording blocks and trained a separate model for an allotted number of blocks. In total, 8 models were trained, each with one of the following block allotments: [1, 2, 5, 10, 15, 20, 25, 28]. Each block comprised an average of 50 sentences recorded in one continuous session.

Quantification of silent speech synthesis. By definition, there was no acoustic signal to compare the decoded silent speech. In order to assess decoding performance, we evaluated decoded silent speech in regards to the audible speech of the same sentence uttered immediately prior to the silent trial. We did so by dynamically time-warping (Brendt & Clifford, 1994) the decoded silent speech MFCCs to the MFCCs of the audible condition and computing Pearson's correlation coefficient and Mel-cepstral distortion.

Phoneme acoustic similarity analysis. We compared the acoustic properties of decoded phonemes to ground-truth to better understand the performance of our decoder. To do this, we sliced all time points for which a given phoneme was being uttered and used the corresponding time slices to estimate its distribution of spectral properties. With principal components analysis (PCA), the 32 spectral features were projected onto the first 4 principal components before fitting the gaussian kernel density estimate (KDE) model. This process was repeated so that each phoneme had two KDEs representing either its decoded and or ground-truth spectral properties. Using Kullback-Leibler divergence (KL divergence), we compared each decoded phoneme KDE to every ground-truth phoneme KDE, creating an analog to a confusion matrix used in discrete classification decoders. KL divergence provides a metric of how similar two distributions are to one another by calculating how much information is lost when we approximate one distribution with another. Lastly, we used Ward's method for agglomerative hierarchical clustering to organize the phoneme similarity matrix.

To understand whether the clustering of the decoded phonemes was similar to the clustering of ground-truth phoneme pairs (Figure 2.11), we used the cophenetic correlation (CC) to assess how well the hierarchical clustering determined from decoded phonemes preserved the pairwise distance between original phonemes, and vice versa (Sokal & Rohlf, 1962). For the decoded phoneme dendrogram, the CC for preserving original phoneme distances was 0.71 as compared to 0.80 for preserving decoded phoneme distances. For the original phoneme dendrogram, the CC for preserving decoded phoneme distances was 0.64 as compared to 0.71 for preserving original phoneme distances. $p<1e-10$ for all correlations.

State-space kinematic trajectories. For state-space analysis of kinematic trajectories, principal components analysis (PCA) was performed on the 33 kinematic features using the training data set from P1. Figure 4a,b shows kinematic trajectories (original, decoded (audible

and mimed) projected onto the first two principal components (PCs). The example decoded mimed trajectory occurred faster in time by a factor of 1.15 than the audible trajectory so we uniformly temporally stretched the trajectory for visualization. The peaks and troughs of the decoded mimed trajectories were similar to the audible speech trajectory ($r=0.65$, $r=0.55$) although the temporal locations are shifted relative to one another, likely because the temporal evolution of a production, whether audible or mimed, is inconsistent across repeated productions. To quantify the decoding performance of mimed trajectories, we used the dynamic time-warping approach described above, although in this case, temporally warping with respect to the inferred kinematics (not the state-space) (Figure 4e).

For analysis of state-space trajectories across participants (Figure 4f), we measured the correlations of productions of the same sentence, but across participants. Since the sentences were produced at different speeds, we dynamically time-warped them to match and compared against correlations of dynamically time-warped mismatched sentences.

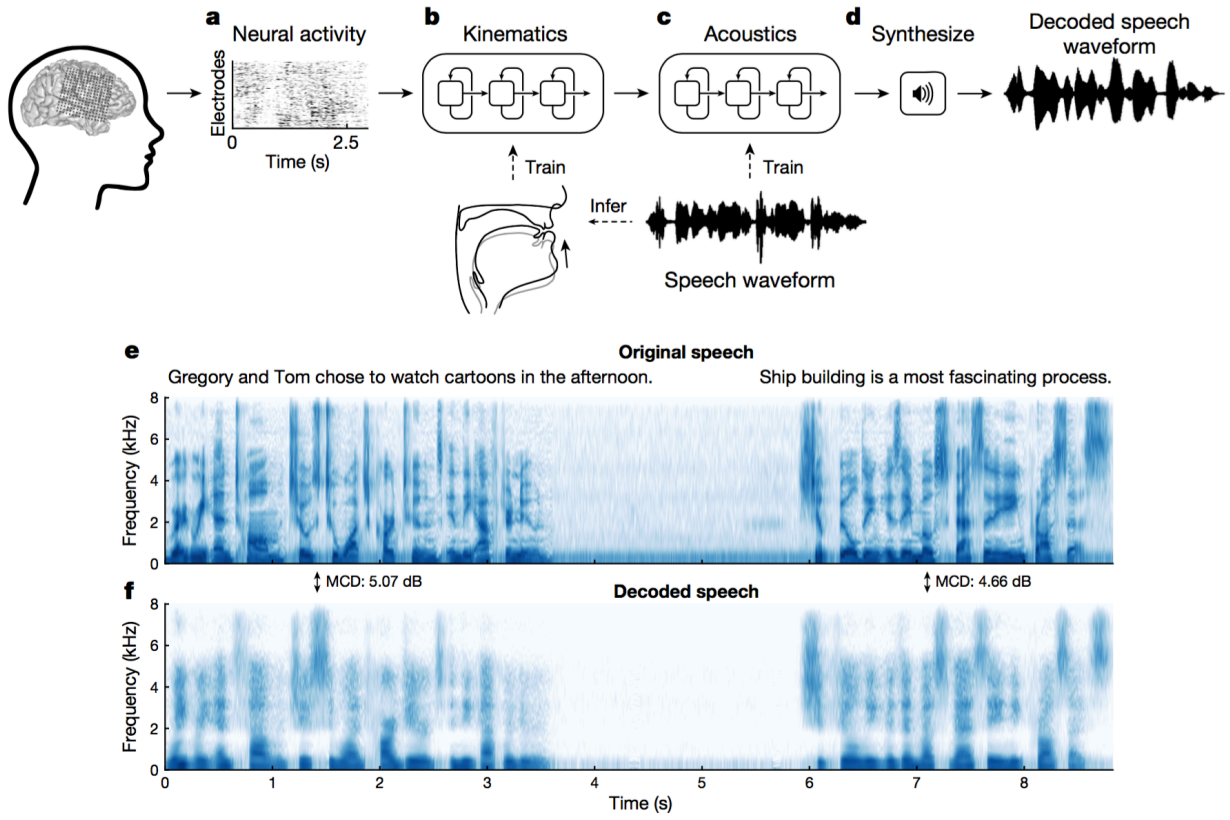


Fig. 2.1 | **Speech synthesis from neurally decoded spoken sentences. a**, The neural decoding process begins by extracting relevant signal features from high-density cortical activity. **b**, A bLSTM neural network decodes kinematic representations of articulation from ECoG signals. **c**, An additional bLSTM decodes acoustics from the previously decoded kinematics. Acoustics are spectral features (for example, MFCCs) extracted from the speech waveform. **d**, Decoded signals are synthesized into an acoustic waveform. **e**, Spectrogram shows the frequency content of two sentences spoken by a participant. **f**, Spectrogram of synthesized speech from brain signals recorded simultaneously with the speech in **e** (repeated five times with similar results). MCD was computed for each sentence between the original and decoded audio. Fivefold cross-validation was used to find consistent decoding.

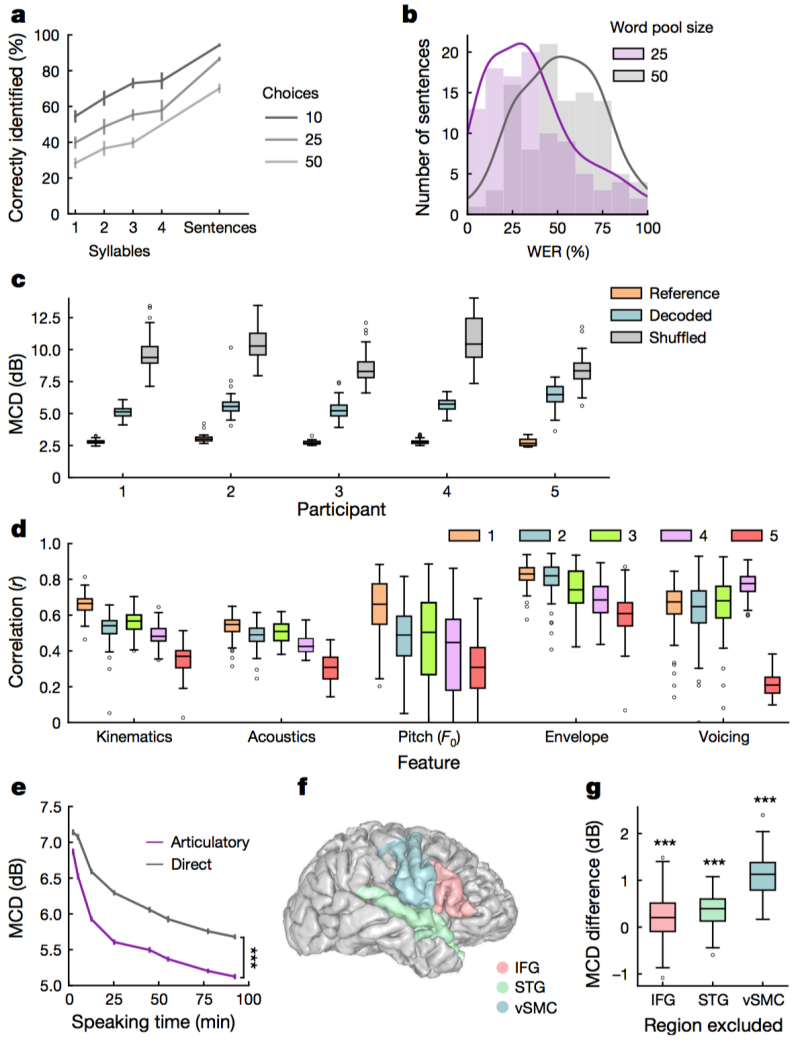


Fig. 2.2 | Synthesized speech intelligibility and feature-specific performance. a, Listening tests for identification of excerpted single words (n = 325) and full sentences (n = 101) for synthesized speech from participant 1. Points represent mean word identification rate. Words were grouped by syllable length (n = 75, 158, 68 and 24, respectively, for one, two, three and four syllables). Listeners identified speech by selecting from a set of choices (10, 25 or 50 words). Data are mean ± s.e.m. **b**, Listening tests for closed vocabulary transcription of synthesized sentences (n = 101). Responses were constrained in word choice (25 or 50), but not in sequence length. Outlines are kernel density estimates of the distributions. **c**, Spectral distortion, measured by MCD (lower values are better), between original spoken sentences and neurally decoded sentences (n = 101, 100, 93, 81 and 44, respectively, for participants 1–5). Reference MCD refers to the synthesis of original (inferred) kinematics without neural decoding. **d**, Correlation of original and decoded kinematic and acoustic features (n = 101, 100, 93, 81 and 44 sentences, respectively, for participants 1–5). Kinematic and acoustic values represent mean correlation of 33 and 32 features, respectively. **e**, Mean MCD of sentences (n = 101) decoded from models trained on varying amounts of training data. The neural decoder with an articulatory intermediate stage (purple) performed better than the direct ECoG to acoustics decoder (grey). All data sizes: n = 101 sentences; $P < 1 \times 10^{-5}$, Wilcoxon signed-rank test. **f**, Anatomical reconstruction of the brain of participant 1 with the following regions used for neural decoding: ventral sensorimotor cortex (vSMC), superior temporal gyrus (STG) and inferior frontal gyrus (IFG). **g**, Difference in median MCD of sentences (n = 101) between decoder trained on all regions and decoders trained on all-but-one region. Exclusion of any region resulted in decreased performance. n = 101 sentences; $P < 3 \times 10^{-4}$, Wilcoxon signed-rank test. All box plots depict median (horizontal line inside box), 25th and 75th percentiles (box), 25th or 75th percentiles ±1.5× interquartile range (whiskers) and outliers (circles). Distributions were compared with each as other as indicated or with chance-level distributions using two-tailed Wilcoxon signed-rank tests. ***P < 0.001.

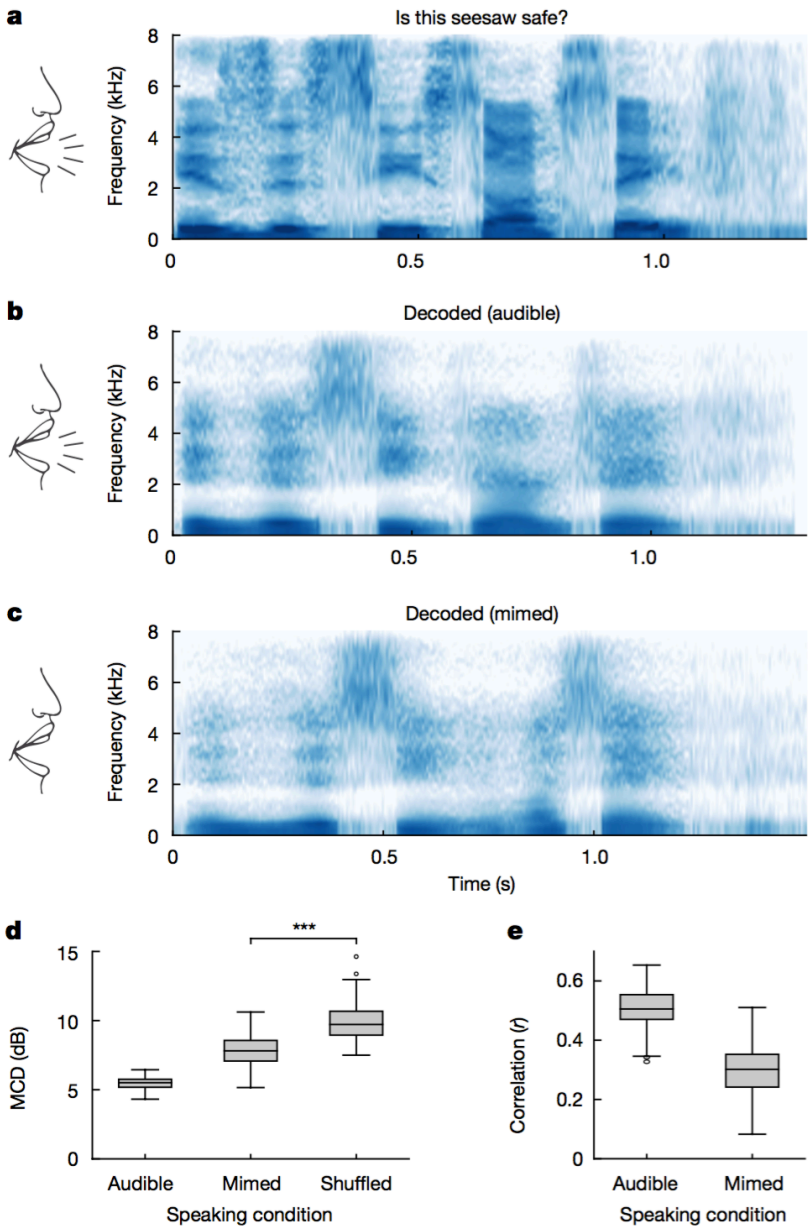


Fig. 2.3 | **Speech synthesis from neural decoding of silently mimed speech. a**–**c**, Spectrograms of original spoken sentence (**a**), neural decoding from audible production (**b**) and neural decoding from silently mimed production (**c**) (repeated five times with similar results). **d**, **e**, MCD (**d**) and correlation of original and decoded spectral features (**e**) for audibly and silently produced speech (n = 58 sentences). Decoded sentences were significantly better than chance-level decoding for both speaking conditions. n = 58; audible, $P = 3 \times 10^{-11}$; mimed, $P = 5 \times 10^{-11}$, Wilcoxon signed-rank test. Box plots as described in Fig. 2.2. ***P < 0.001.

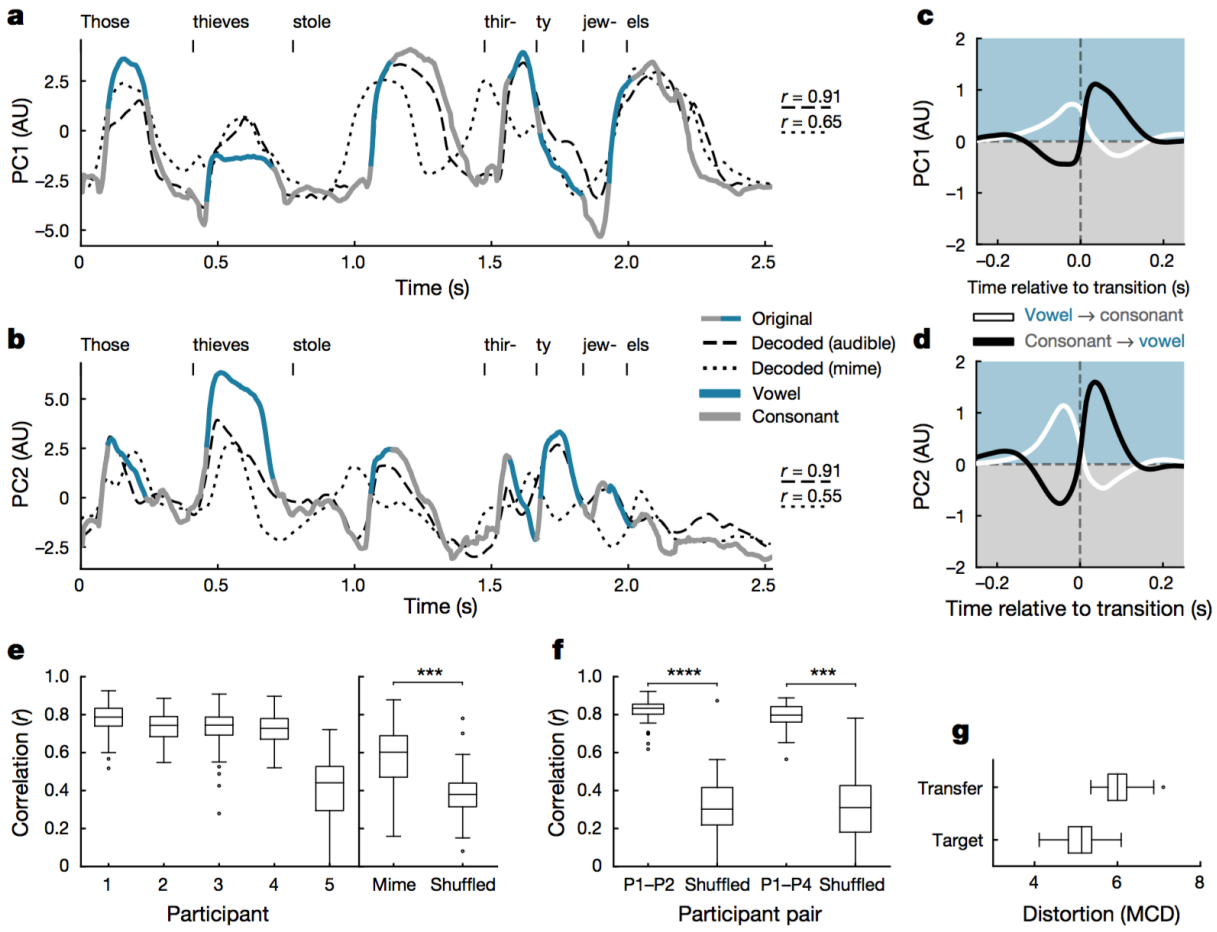


Fig. 2.4 | **Kinematic state–space representation of speech production. a**, **b**, A kinematic trajectory (grey–blue) from a single trial (participant 1) projected onto the first two principal components—principal components (PC)1 (**a**) and 2 (**b**)—of the kinematic state–space. Decoded audible (dashed) and mimed (dotted) kinematic trajectories are also plotted. Pearson's *r*, *n* = 510 time samples. The trajectory for mimed speech was uniformly stretched to align with the audible speech trajectory for visualization as it occurred at a faster time scale. **c**, **d**, Average trajectories for principal components 1 (**c**) and 2 (**d**) from **a** and **b**, respectively, for transitions from a vowel to a consonant (black, *n* = 22,453) and from a consonant to a vowel (white, *n* = 22,453). Time courses are 500 ms. **e**, Distributions of correlations between original and decoded kinematic state–space trajectories (averaged across principal components 1 and 2) (*n* = 101, 100, 93, 81, 44 sentences, respectively, for participants 1–5). Pearson's correlations for mimed trajectories were calculated by dynamically time-warping to the audible production of the same sentence and then compared to correlations of the dynamically time-warping of a randomly selected sentence trajectory. *n* = 58 sentences; ***$P = 1 \times 10^{-5}$, Wilcoxon signed-rank test. **f**, Distributions of correlations for state–space trajectories of the same sentence across participants. Alignment between participants was done by dynamically time-warping and compared to correlations of dynamically time-warping of unmatched sentence pairs. *n*=92;****$P=1\times10^{-16}$ and *n*=44;***$P=1\times10^{-8}$, respectively, Wilcoxon signed-rank test. **g**, Comparison between acoustic decoders (stage 2) (*n* = 101 sentences). 'Target' refers to an acoustic decoder trained on data from the same participant as the kinematic decoder (stage 1) is trained on (participant 1). 'Transfer' refers to an acoustic decoder that was trained on kinematics and acoustics from a different participant (participant 2). Box plots as described in Fig. 2.2.

Table 1 | Listener transcriptions of neurally synthesized speech

Word error rate	Original sentences and transcriptions of synthesized speech
0%	o: Is this seesaw safe t: Is this seesaw safe
~10%	o: Bob bandaged both wounds with the skill of a doctor t: Bob bandaged full wounds with the skill of a doctor
~20%	o: Those thieves stole thirty jewels t: Thirty thieves stole thirty jewels
	o: Help celebrate brother's success t: Help celebrate his brother's success
~30%	o: Get a calico cat to keep the rodents away t: The calico cat to keep the rabbits away
	o: Carl lives in a lively home t: Carl has a lively home
~50%	o: Mum strongly dislikes appetizers t: Mom often dislikes appetizers
	o: Etiquette mandates compliance with existing regulations t: Etiquette can be made with existing regulations
>70%	o: At twilight on the twelfth day we'll have Chablis t: I was walking through Chablis

Examples are shown for several word error rate levels. The original text is indicated by 'o' and the listener transcriptions are indicated by 't'.

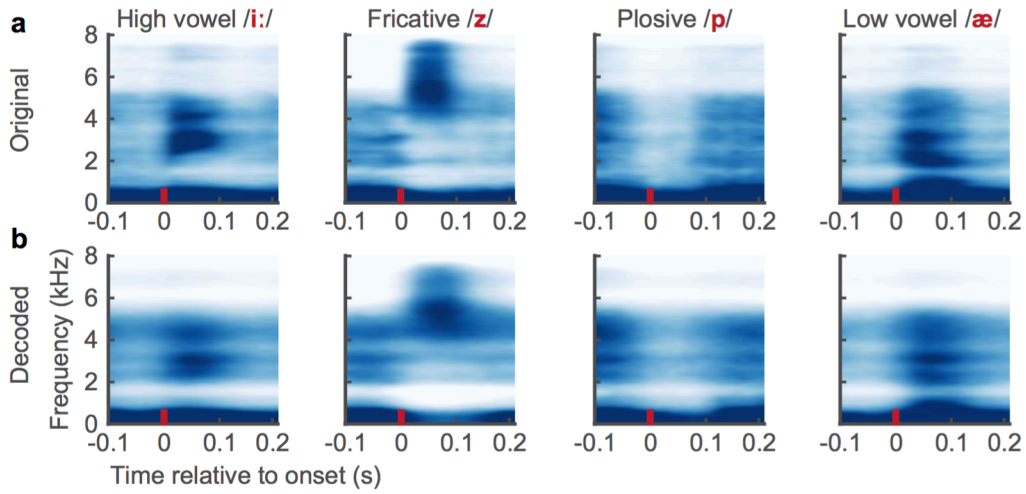


Fig. 2.5 | **Median original and decoded spectrograms. a**, **b**, Median spectrograms, time-locked to the acoustic onset of phonemes from original (**a**) and decoded (**b**) audio (/i/, *n* = 112; /z/, *n* = 115; /p/, *n* = 69, /ae/, *n* = 86). These phonemes represent the diversity of spectral features. Original and decoded median phoneme spectrograms were well-correlated (Pearson's $r > 0.9$ for all phonemes, $P = 1 \times 10^{-18}$).

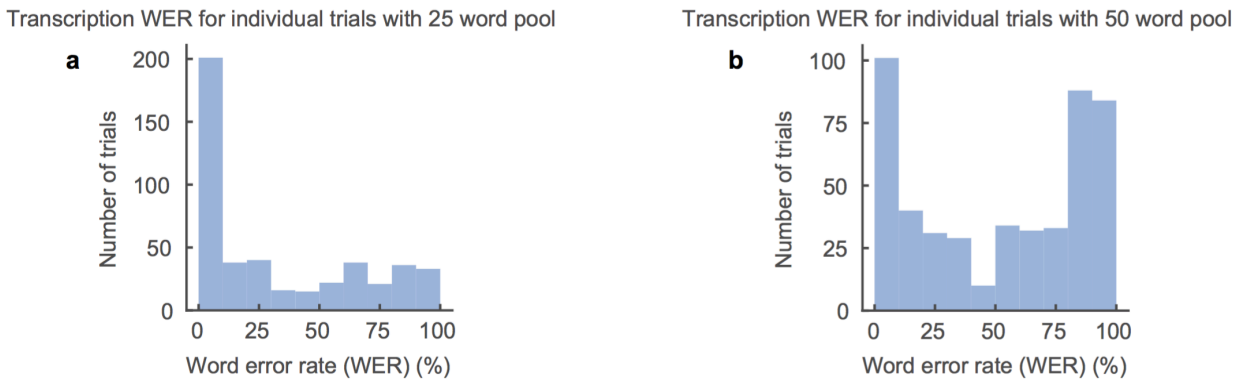


Fig. 2.6 | **Transcription WER for individual trials. a**, **b**, WERs for individually transcribed trials for pools with a size of 25 (**a**) or 50 (**b**) words. Listeners transcribed synthesized sentences by selecting words from a defined pool of words. Word pools included correct words found in the synthesized sentence and random words from the test set. One trial is one transcription of one listener of one synthesized sentence.

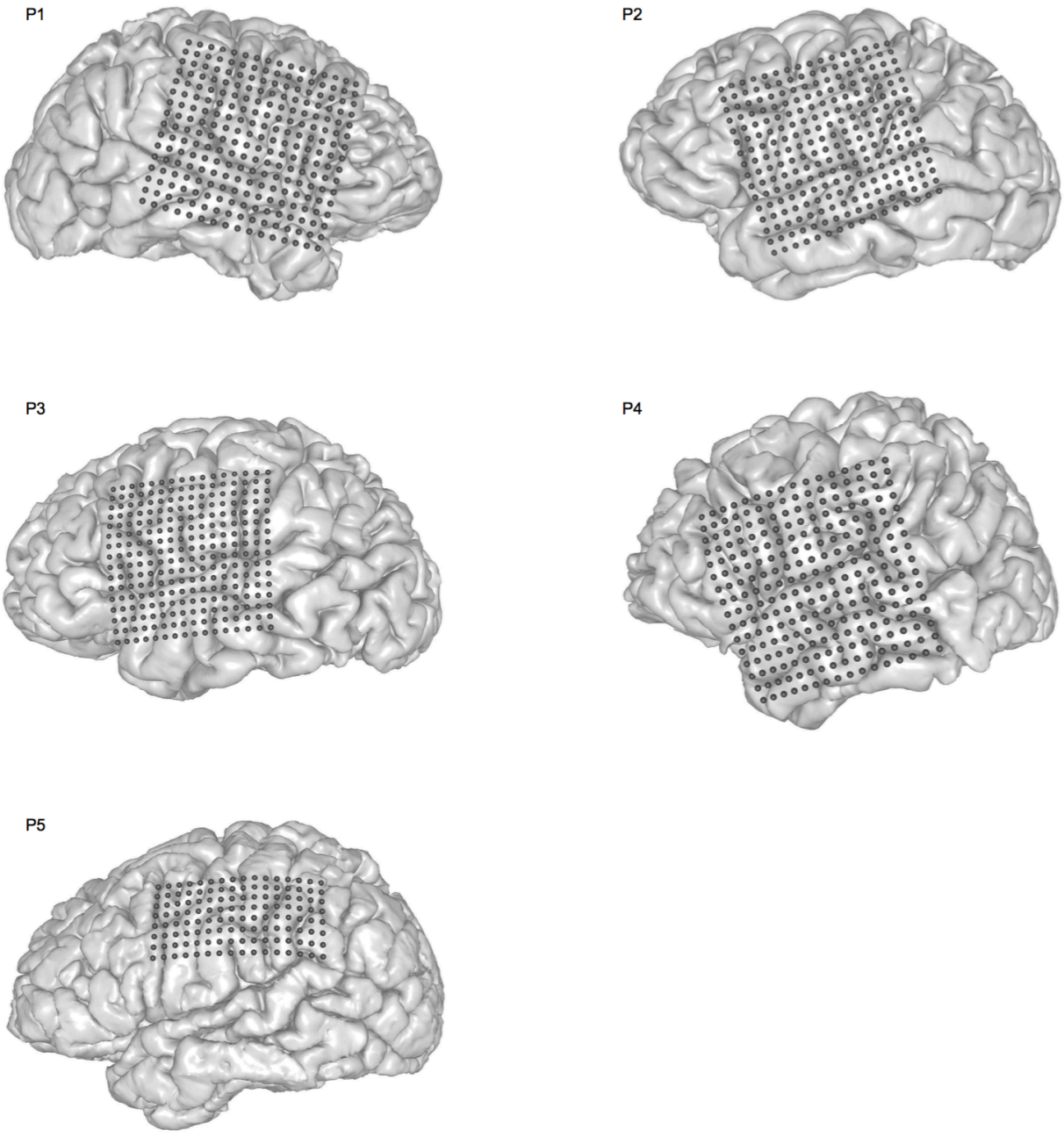


Fig. 2.7 | **Electrode array locations for participants.** MRI reconstructions of participants' brains with overlay of electrocorticographic electrode (ECoG) array locations. P1–5, participants 1–5.

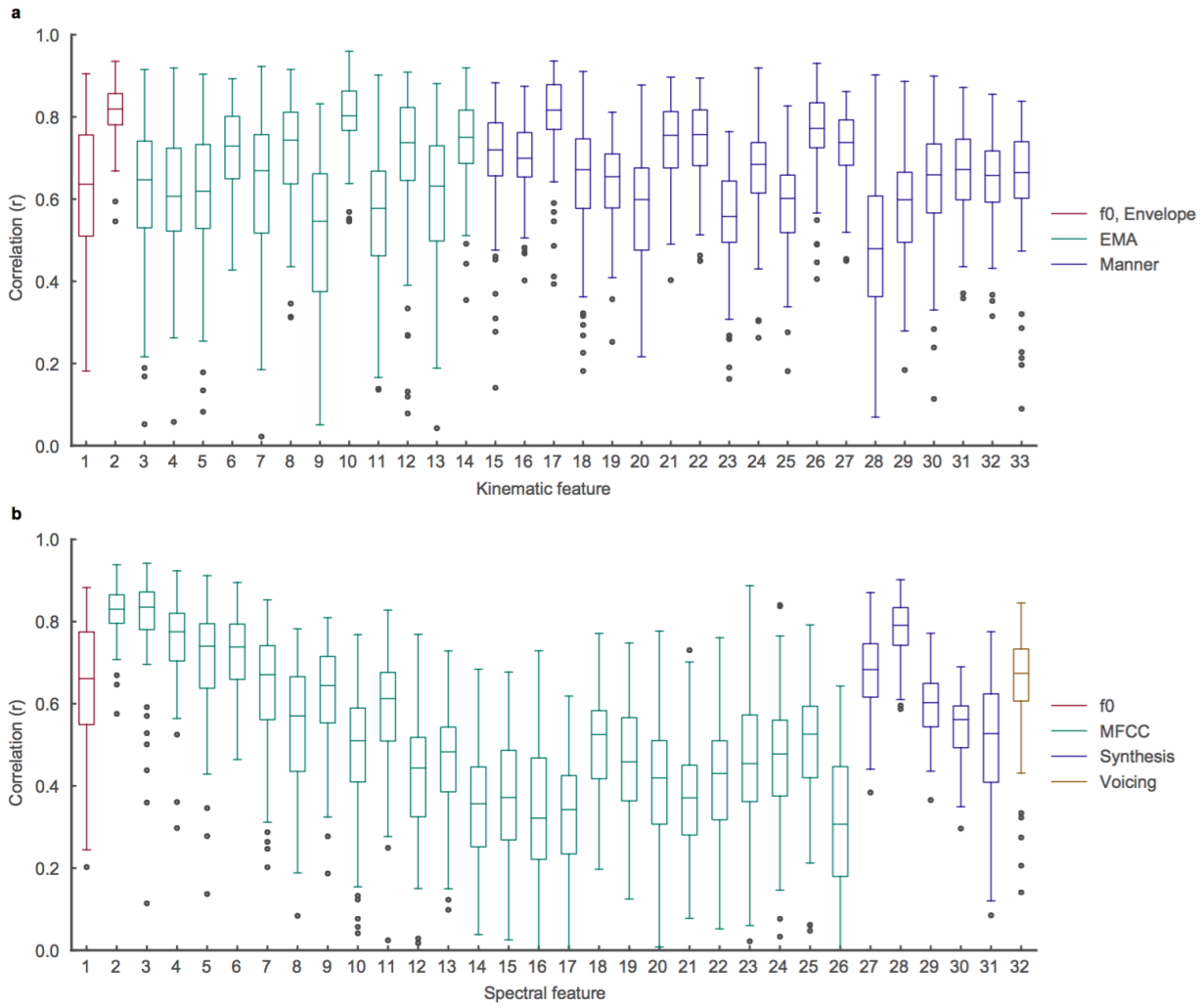


Fig. 2.8 | **Decoding performance of kinematic and spectral features.** Data from participant 1. **a**, Correlations of all 33 decoded articulatory kinematic features with ground-truth (*n* = 101 sentences). EMA features represent *x* and *y* coordinate traces of articulators (lips, jaw and three points of the tongue) along the midsagittal plane of the vocal tract. Manner features represent complementary kinematic features to EMA that further describe acoustically consequential movements. **b**, Correlations of all 32 decoded spectral features with ground-truth (*n* = 101 sentences). MFCC features are 25 mel-frequency cepstral coefficients that describe power in perceptually relevant frequency bands. Synthesis features describe glottal excitation weights necessary for speech synthesis. Box plots as described in Fig. 2.2.

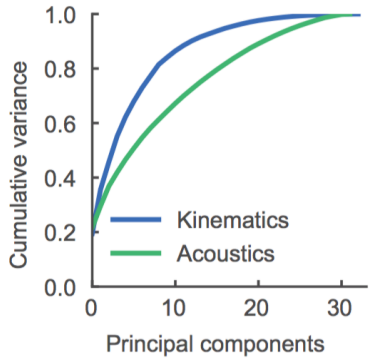


Fig. 2.9 | **Comparison of cumulative variance explained in kinematic and acoustic state–spaces.** For each representation of speech— kinematics and acoustics—a principal components analysis was computed and the explained variance for each additional principal component was cumulatively summed. Kinematic and acoustic representations had 33 and 32 features, respectively.

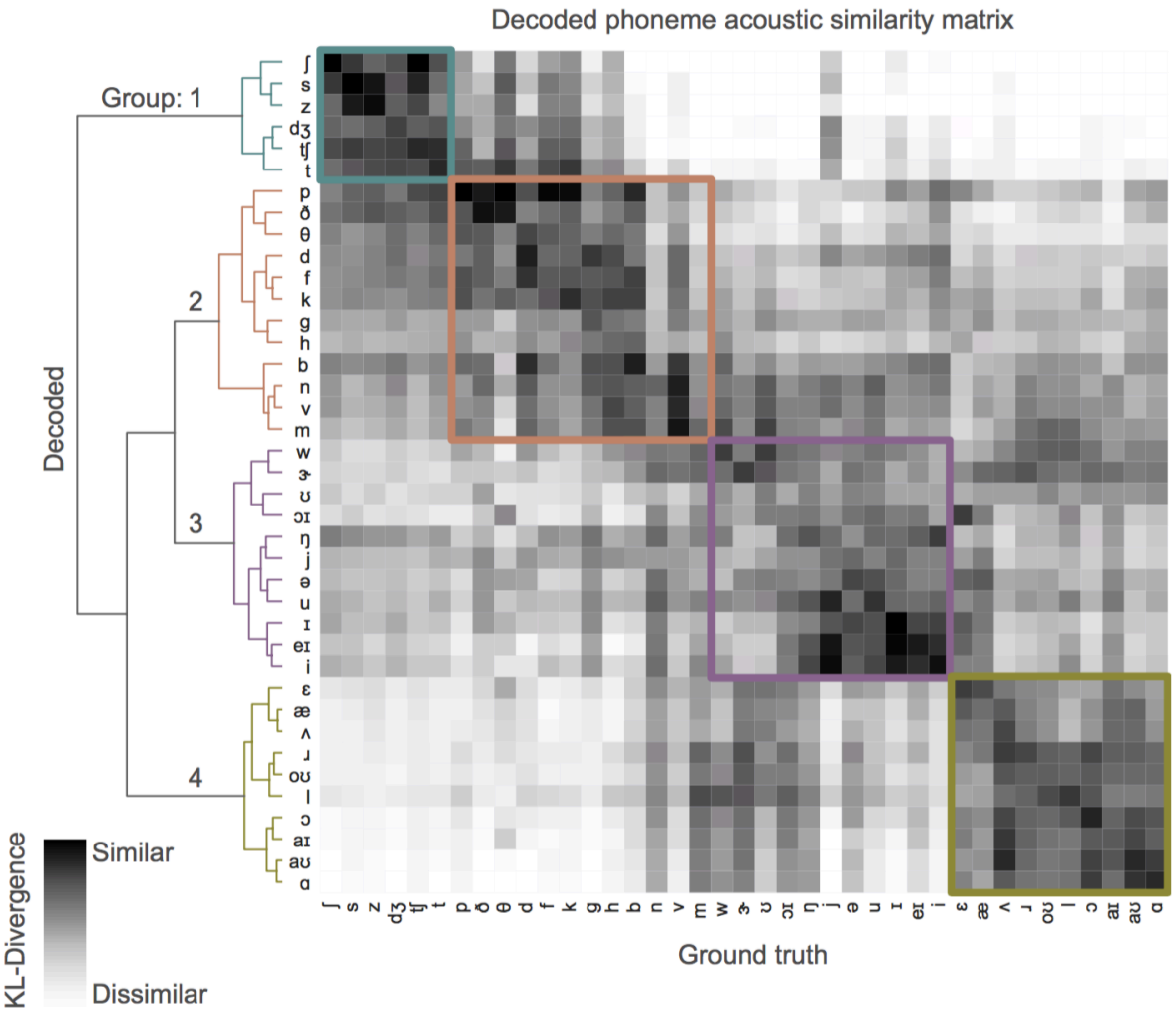


Fig. 2.10 | **Decoded phoneme acoustic similarity matrix.** Acoustic similarity matrix compares acoustic properties of decoded phonemes and originally spoken phonemes. Similarity is computed by first estimating a Gaussian kernel density for each phoneme (both decoded and original) and then computing the Kullback–Leibler (KL) divergence between a pair of decoded and original phoneme distributions. Each row compares the acoustic properties of a decoded phoneme with originally spoken phonemes (columns). Hierarchical clustering was performed on the resulting similarity matrix. Data from participant 1.

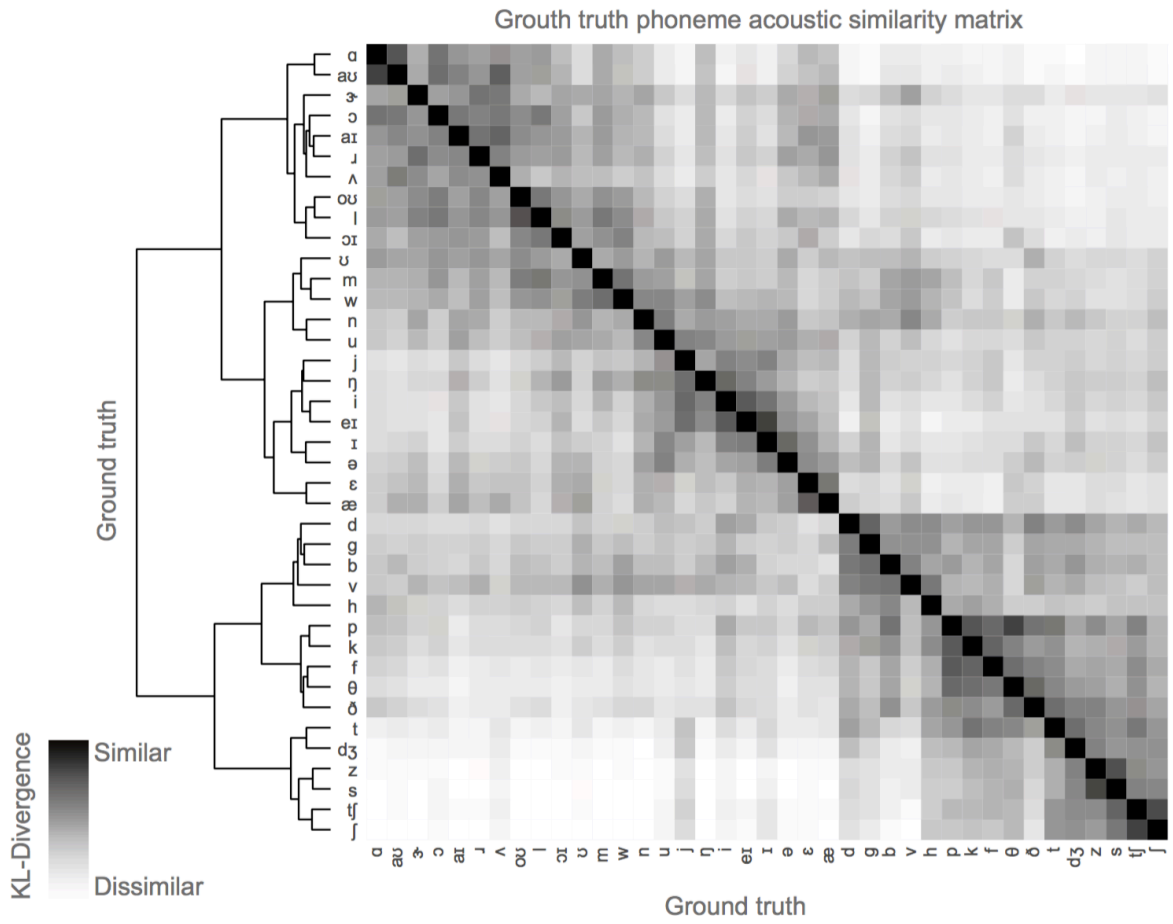


Fig. 2.11 | **Ground-truth acoustic similarity matrix.** The acoustic properties of ground-truth spoken phonemes are compared with one another. Similarity is computed by first estimating a Gaussian kernel density for each phoneme and then computing the Kullback–Leibler divergence between a pair of a phoneme distributions. Each row compares the acoustic properties of two ground-truth spoken phonemes. Hierarchical clustering was performed on the resulting similarity matrix. Data from participant 1.

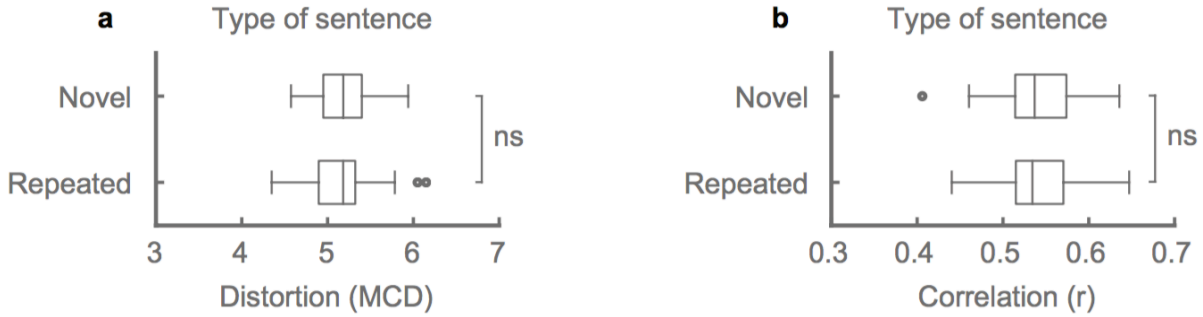


Fig. 2.12 | **Comparison between decoding novel and repeated sentences. a**, **b**, Comparison metrics included spectral distortion (**a**) and the correlation between decoded and original spectral features (**b**). Decoder performance for these two types of sentences was compared and no significant difference was found ($P = 0.36$ (**a**) and $P = 0.75$ (**b**), $n = 51$ sentences, Wilcoxon signed-rank test). A novel sentence consists of words and/or a word sequence not present in the training data. A repeated sentence is a sentence that has at least one matching word sequence in the training data, although with a unique production. Comparison was performed on participant 1 and the evaluated sentences were the same across both cases with two decoders trained on differing datasets to either exclude or include unique repeats of sentences in the test set. ns, not significant; $P > 0.05$. Box plots as described in Fig. 2.

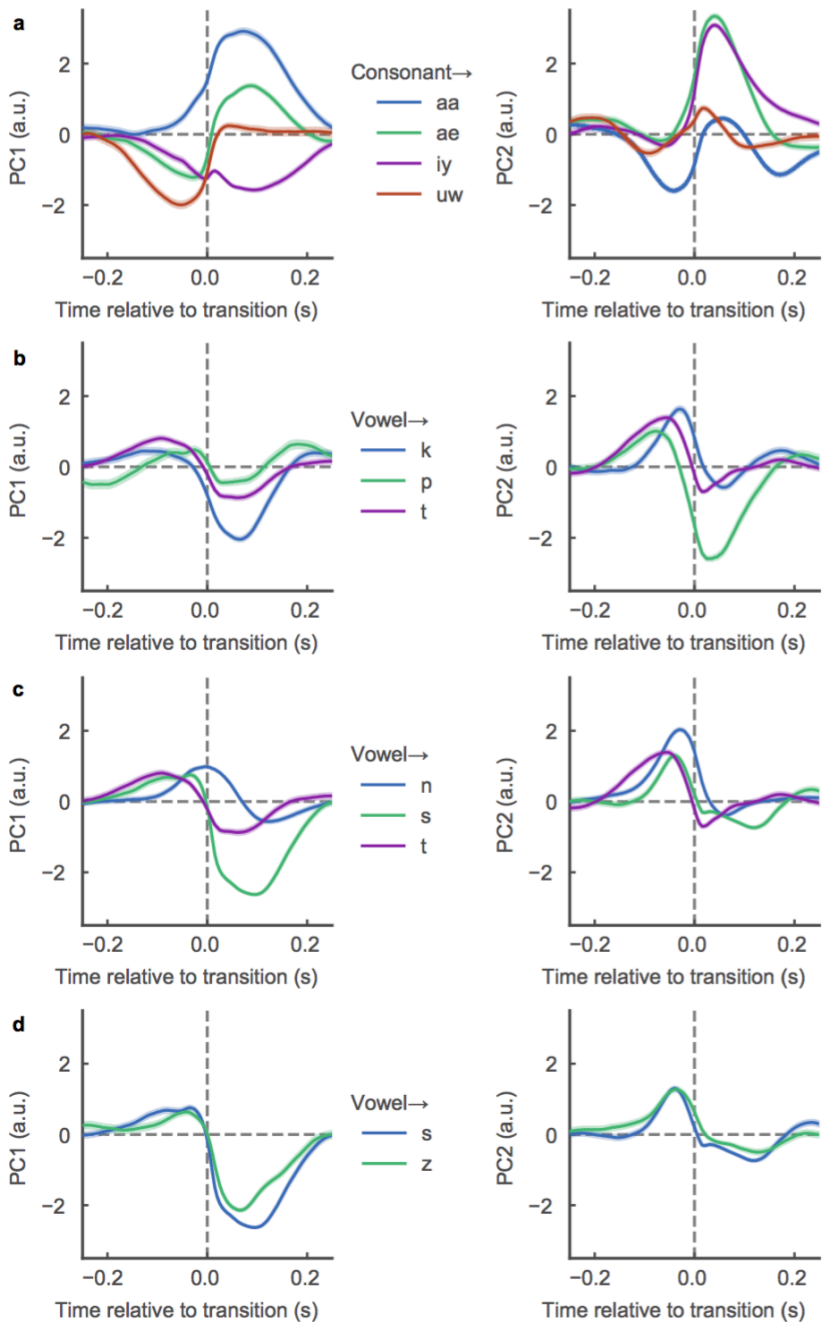


Fig. 2.13 | **Kinematic state–space trajectories for phoneme- specific vowel–consonant transitions.** Average trajectories of principal components 1 (PC1) and 2 (PC2) for transitions from either a consonant or a vowel to specific phonemes. Trajectories are 500 ms and centred at transition between phonemes. **a**, Consonant to corner vowels (*n* = 1,387, 1,964, 2,259, 894, respectively, for aa, ae, iy and uw). PC1 shows separation of all corner vowels and PC2 delineates between front vowels (iy, ae) and back vowels (uw, aa). **b**, Vowel to unvoiced plosives (*n* = 2,071, 4,107 and 1,441, respectively, for k, p and t). PC1 was more selective for velar constriction (k) and PC2 for bilabial constriction (p). **c**, Vowel to alveolars (*n* = 3,919, 3,010 and 4,107, respectively, for n, s and t). PC1 shows separation by manner of articulation (nasal, plosive or fricative) whereas PC2 is less discriminative. **d**, PC1 and PC2 show little, if any, delineation between voiced and unvoiced alveolar fricatives (*n* = 3,010 and 1,855, respectively, for s and z).

Acknowledgements

We thank M. Leonard, N. Fox and D. Moses for comments on the manuscript and B. Speidel for his help reconstructing MRI images. This work was supported by grants from the NIH (DP2 OD008627 and U01 NS098971-01). E.F.C. is a New York Stem Cell Foundation-Robertson Investigator. This research was also supported by The William K. Bowes Foundation, the Howard Hughes Medical Institute, The New York Stem Cell Foundation and The Shurl and Kay Curci Foundation.

Author contributions

G.K.A., J.C. and E.F.C. conceived the study; G.K.A. inferred articulatory kinematics; G.K.A. and J.C. designed the decoder; J.C. performed decoder analyses; G.K.A., E.F.C. and J.C. collected data and prepared the manuscript; E.F.C. supervised the project.

References

Aflalo, T., Kellis, S., Klaes, C., Lee, B., Shi, Y., Pejsa, K., ... & Andersen R. A. (2015). Decoding motor imagery from the posterior parietal cortex of a tetraplegic human. *Science*, *348*(6237), 906-910.

Ajiboye, A. B., Willett, F. R., Young, D. R., Memberg, W. D., Murphy, B. A., Miller, J. P., ... & Peckham, P. H. (2017). Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration. The Lancet, 389(10081), 1821-1830.

Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., & Mesgarani, N. (2019). Towards reconstructing intelligible speech from the human auditory cortex. *Scientific reports*, *9*(1), 874.

Anumanchipalli, G. K., Prahallad, K., & Black., A. W. (2011). Festvox: Tools for creation and analyses of large speech corpora, Workshop on Very Large Scale Phonetics Research, UPenn, Philadelphia. http://www.festvox.org

Berndt, D. J., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In KDD workshop (Vol. 10, No. 16, pp. 359-370).

Bocquelet, F., Hueber, T., Girin, L., Savariaux, C., & Yvert, B. (2016). Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. *PLoS computational biology*, *12*(11), e1005119.

Bouchard, K.E., Mesgarani, N., Johnson, K., and Chang, E.F. (2013). Functional organization of human sensorimotor cortex for speech articulation. Nature 495, 327–332.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. Phonetica, 49(3-4), 155-180.

Brumberg, J. S., Krusienski, D. J., Chakrabarti, S., Gunduz, A., Brunner, P., Ritaccio, A. L., & Schalk, G. (2016). Spatio-Temporal Progression of Cortical Activity Related to Continuous Overt and Covert Speech Production in a Reading Task. *PloS one*, *11*(11), e0166872. doi:10.1371/journal.pone.0166872

Brumberg, J.S., Pitt, K.M., Mantie-Kozlowski, A., & Burnison, J.D. (2018). Brain–computer interfaces for augmentative and alternative communication: A tutorial. *American Journal of Speech-Language Pathology*, 27, 1–12. doi:10.1044/2017_AJSLP-16-0244

Chartier, J., Anumanchipalli, G. K., Johnson, K., & Chang, E. F. (2018). Encoding of Articulatory Kinematic Trajectories in Human Speech Sensorimotor Cortex. Neuron, 98(5), 1042–1054.e4. https://doi.org/10.1016/j.neuron.2018.04.031

Collinger, J. L., Wodlinger, B., Downey, J. E., Wang, W., Tyler-Kabara, E. C., Weber, D. J., ... & Schwartz, A. B. (2013). High-performance neuroprosthetic control by an individual with tetraplegia. The Lancet, 381(9866), 557-564.

Crone, N.E., Hao, L., Hart, J., Jr., Boatman, D., Lesser, R.P., Irizarry, R., and Gordon, B. (2001). Electrocorticographic gamma activity during word production in spoken and sign language. Neurology 57, 2045–2053.

Davis, S. B., & Mermelstein, P. (1990). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In Readings in speech recognition (pp. 65-74).

Dichter B. K., Breshears J. D., Leonard M. K., and Chang E. F. (2018) The Control of Vocal Pitch in Human Laryngeal Motor Cortex. Cell, 174, 21–31.

Fager, S. K., Fried-Oken, M., Jakobs, T., & Beukelman, D. R. (2019). New and emerging access technologies for adults with complex communication needs and severe motor impairments: State of the science, *Augmentative and Alternative Communication,* DOI: 10.1080/07434618.2018.1556730

Flinker, A., Korzeniewska, A., Shestyuk, A. Y., Franaszczuk, P. J., Dronkers, N. F., Knight, R. T., & Crone, N. E. (2015). Redefining the role of Broca's area in speech. Proceedings of the National Academy of Sciences, 112(9), 2871-2875.

Gallego, J. A, Perich, M., G, Miller, L., E, Solla, S., A, (2017) Neural manifolds for the control of movement., *Neuron,* 94(5), 978-984.

Golub, M. D., Sadtler, P. T., Oby, E. R., Quick, K. M., Ryu, S. I., Tyler-Kabara, E. C., ... & Yu, B. M. (2018). Learning by neural reassociation. Nat. Neurosci., 21.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5-6), 602-610.

Guenther, F. H., Brumberg, J. S., Joseph Wright, E., Nieto-Castanon, A., Tourville, J. A., Panko, M., … Kennedy, P. R. (2009). A wireless brain-machine interface for real-time speech synthesis. *PLoS ONE*, 4(12). https://doi.org/10.1371/journal.pone.0008218

Hamilton, L. S., Chang, D. L., Lee, M. B., & Chang, E. F. (2017). Semi-automated Anatomical Labeling and Inter-subject Warping of High-Density Intracranial Recording Electrodes in Electrocorticography. Frontiers in Neuroinformatics, 11, 62. http://doi.org/10.3389/fninf.2017.00062

Herff, C., Heger, D., de Pesters, A., Telaar, D., Brunner, P., Schalk, G., and Schultz, T. (2015). Brain-to-text: decoding spoken phrases from phone representations in the brain.

Hochberg, L. R., Serruya, M. D., Friehs, G. M., Mukand, J. A., Saleh, M., Caplan, A. H., ... & Donoghue, J. P. (2006). Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature, 442(7099), 164

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Comput. 9, 1735-1780.

Huggins, J. E., Wren, P. A., Gruis, K. L. (2011) What would brain-computer interface users want? Opinions and priorities of potential users with amyotrophic lateral sclerosis. *Amyotroph Lateral Scler.* 2011 Sep;12(5):318-24. doi: 10.3109/17482968.2011.572978.

Kominek, J., Schultz, T., and Black, A. (2008). "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion", In *SLTU-2008*, 63-68.

Luce, P. A. & Pisoni, D. B. Recognizing spoken words: the neighborhood activation model. *Ear Hear.* **19**, 1–36 (1998).

Maia, R., Toda, T., Zen, H., Nankaku, Y., Tokuda, K., 2007. An excitation model for HMM-based speech synthesis based on residual modeling. In: Proc. ISCA SSW6, pp. 131–136.

Martin, S., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N. E., Rieger, J., Schalk, G., Knight, R.T., Pasley, B.N. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. Front. Neuroeng. 7:14.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. http://www.tensorflow.org

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. Science, 343(6174), 1006-1010.

Moses, D. A., Mesgarani, N., Leonard, M. K., & Chang, E. F. (2016). Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity. Journal of neural engineering, 13(5), 056004.

Mugler, E.M., Patton, J.L., Flint, R.D., Wright, Z.A., Schuele, S.U., Rosenow, J., Shih, J.J., Krusienski, D.J., and Slutzky, M.W. (2014). Direct classification of all American English phonemes using signals from functional speech motor cortex. J. Neural Eng. 11, 035015.

Mugler, E. M., Tate, M. C., Livescu, K., Templer, J. W., Goldrick, M. A., & Slutzky M. W. (2018) Differential Representation of Articulatory Gestures and Phonemes in Precentral and Inferior Frontal Gyri. *J Neurosci.* 38(46):9803-9813. doi: 10.1523/JNEUROSCI.

Nourski, K. V., Steinschneider, M., Rhone, A. E., Oya, H., Kawasaki, H., Howard III, M. A., & McMurray, B. (2015). Sound identification in human auditory cortex: Differential contribution of local field potentials and high gamma power as revealed by direct intracranial recordings. *Brain and language*, *148*, 37-50.

Pandarinath, C., Nuyujukian, P., Blabe, C. H., Sorice, B. L., Saab, J., Willett, F. R., … Henderson, J. M. (2017). High performance communication by people with paralysis using an intracortical brain-computer interface. ELife, 6, 1–27. doi:10.7554/eLife.18554

Pasley, B. N., David, S. V, Mesgarani, N., Flinker, A., & Shamma, S. A. (2012). Reconstructing Speech from Human Auditory Cortex. *PLoS Biol*, *10*(1), 1001251. https://doi.org/10.1371/journal.pbio.1001251

Paul, B., D, & Baker, M., J. (1992). The design for the wall street journal-based CSR corpus. In Proceedings of the workshop on Speech and Natural Language (HLT '91). Association for Computational Linguistics, Stroudsburg, PA, USA, 357-362. DOI: https://doi.org/10.3115/1075527.1075614

Pesaran, B., Vinck, M., Einevoll, G. T., Sirota, A., Fries, P., Siegel, M., ... & Srinivasan, R. (2018). Investigating large-scale brain dynamics using field potential recordings: analysis and interpretation. *Nature neuroscience*.

Prahallad, K., Black, A.W., & Mosur, R. (2006). Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. In Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. I–I.

Richmond, K, Hoole, P, & King, S. (2011). Announcing the electromagnetic articulography (Day 1) subset of the mngu0 articulatory corpus Proceedings of *Interspeech 2011*, Florence, Italy

Sadtler, P. T., Quick, K. M., Golub, M. D., Chase, S. M., Ryu, S. I., Tyler-Kabara, E. C., ... & Batista, A. P. (2014). Neural constraints on learning. *Nature*, *512*(7515), 423.

Serruya MD, Hatsopoulos NG, Paninski L, Fellows MR, Donoghue JP (2002) Instant neural control of a movement signal. Nature 416: 141–142.

Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 33-40.

Taylor DM, Tillery SI, Schwartz AB (2002) Direct cortical control of 3D neuroprosthetic devices. Science 296: 1829–1832.

Wessberg J, Stambaugh CR, Kralik JD, Beck PD, Laubach M, et al. (2000) Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. Nature 408: 361–365.

Wolters, M. K., Isaac, Renals, S., Evaluating Speech Synthesis intelligibility using Amazon Mechanical Turk. (2010) In proceedings of ISCA speech synthesis workshop (SSW7), 2010.

Wrench, A. (1999). MOCHA: multichannel articulatory database. http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

9/8/2019
Date