Representation of directly measured speech movements in human sensorimotor cortex

by

David Conant

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Neuroscience

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Acknowledgements

I would like to thank my advisor, Dr. Edward Chang, for the advice, mentorship and opportunities through the past 6 years, and my committee, Drs. John Houde, Keith Johnson, and Michael Brainard, for the ongoing guidance and encouragement.

I would also like to thank the entire Chang lab, and especially Matt Leonard and Kristofer Bouchard for their contributions to this work.

Some material presented in this work is published in peer-reviewed journals:

Conant, David, Kristofer E Bouchard, and Edward F Chang (2014). "Speech map in the human ventral sensory-motor cortex". In: Curr. Opin. Neurobiol. 24, pp. 63–67.

Bouchard, Kristofer E. et al. (2016). "High-resolution, non-invasive imaging of upper vocal tract artic- ulators compatible with human brain recordings". In: PLoS One 11.3, pp. 1–30.

Abstract

Representation of directly measured speech movements in human sensorimotor cortex

David CONANT

During speech production, we make vocal tract movements with remarkable precision and speed. Starting with the earliest cortical stimulation studies, we have learned much about what brain regions are involved with speech motor control. However, our understanding of how activity in these regions gives rise to the movements made is limited, in part due to the challenge of simultaneously acquiring high-resolution neural recordings and detailed vocal tract measurements. A complete neurobiological understanding of speech motor control requires determination of the relationship between simultaneously recorded neural activity and the kinematics of the speech articulators (i.e, lips, jaw, and tongue). Recent advances in human electrophysiological recordings allow us to observe neural activity in these regions with unprecedented resolution, but without concurrently measuring the speech articulators it is difficult to interpret this activity. To overcome this challenge, we combined ultrasound and video monitoring of the supralaryngeal articulators (lips, jaw and tongue) with electrocorticographic (ECoG) recordings from the cortical surface to investigate how neural activity relates to measured articulator movement kinematics (position, speed, velocity, acceleration) during the production of English vowels. In this document, we first provide a review of the functional organization of primary speech motor cortex, also called ventral sensory motor cortex (vSMC). Next, we describe and validate methods for a noninvasive, multi-modal imaging system to monitor vocal tract kinematics that is compatible with bedside human neurophysiological recordings. Last, we use these methods to examine the relationship between activity in vSMC and the kinematics of speech articulator movements. These findings demonstrate novel insights into how articulatory kinematic parameters are encoded in vSMC during speech production.

Contents

List of Figures

Chapter 1

Spatial organization of speech articulators in human ventral sensory-motor cortex

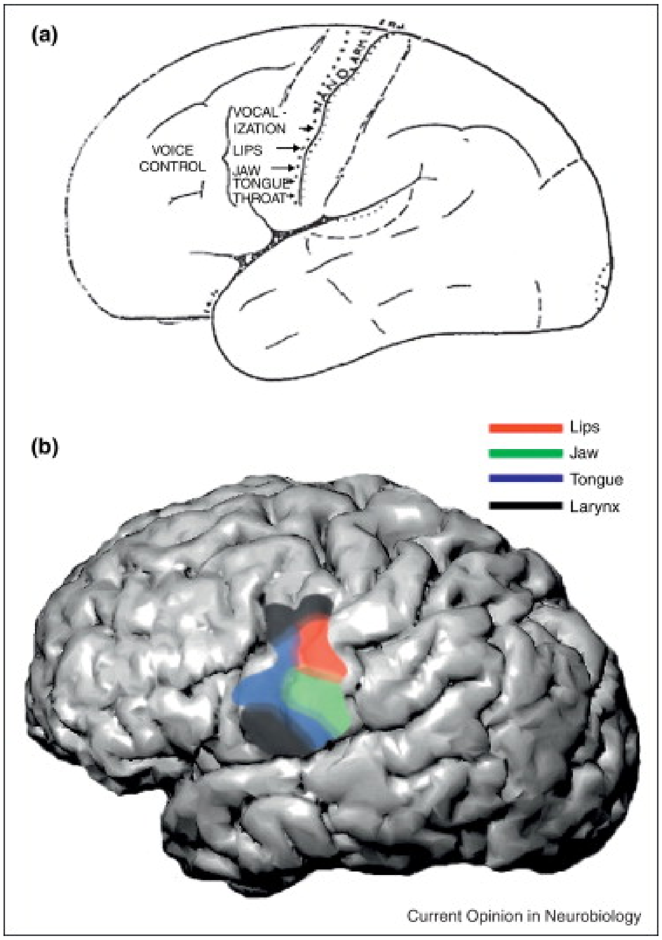
1.1 Introduction

Speaking is a unique and defining human behavior. It is carried out by precise, controlled movements of different parts of the vocal tract, known as articulators, which are closely coordinated with the larynx and respiration. Speech articulation is often described as the most complex motor behavior because over 100 muscles are involved, and the movements occur on an extremely rapid time scale. Despite its complexity, nearly all of us learn to master this skill to speak fluently and effortlessly (Kent and Moll, 1972).A key brain area in the neural control of articulation is the ventral portion of the sensory-motor cortex (vSMC). Injuries to this area produce motor deficits in articulation, called dysarthria (Penfield and Roberts, 1959). In comparison to the dorsal sensorimotor cortical regions involved in arm reaching and hand function, the neurobiology of vSMC is relatively understudied. The vSMC features some important anatomic and functional differences from dorsal sensory-motor cortex, while sharing others. For example, in contrast to the dorsal areas, vSMC projects via the corticobulbar tract to the oro-facial motor nuclei, and ultimately to the articulatory muscles. vSMC has connections with higher-order cortical areas such as the anterior cingulate and supplementary motor area, basal ganglia, and cerebellum.

In classic studies, the vSMC has been described by its somatotopic organization of face and oropharynx representations.

These areas are involved in controlling such non-speech movements as facial expressions, tongue movements, and swallowing. However, over the past decade we have begun to learn more about how this same cortical area mediates a totally different functional purpose in the production of vocal speech.

The goal of this chapter is to address the functional organization of the vSMC in the context of speaking, broadly focused on three central topics: firstly somatotopy of speech articulator representations, secondly potential neuroanatomical specializations for speech in humans, and thirdly organization of distributed spatial patterns of cortical activity during speech.



Current Opinion in Neurobiology

FIGURE 1.1: **Somatotopic organization of vSMC** ***(a)*** Spatial organization of the lips, jaw, tongue in the 'homunculus' as described by classic early stimulation studies. Adapted from Penfield (1959). ***(b)*** Functional organization of the vSMC derived using electrocorticographic recordings of 3 subjects during speech. The overall ordering of representations of the vocal tract is the same as previously described by Penfield, except that two laryngeal areas were identified in the dorsal-most and ventral-most aspects of the vSMC. The layout of speech articulators was more fractured and overlapping than previous depictions.

1.2 The somatotopy of speech articulators in vSMC

Electrical stimulation studies provided the earliest description of the human vSMC somatotopy from Foerster and Penfield (Penfield and Boldrey, 1937). The popular conception of vSMC organization features a highly stereotyped, discretely ordered progression of representations for the lips, jaw, tongue, and pharynx/larynx, respectively, along the dorsal-to-ventral orientation of the central sulcus (Figure 1.1a) (Penfield and Boldrey, 1937).

However, the full details of their qualitative descriptions actually portray a more complex picture of organization. Cortical regions representing separate, but neighboring, body parts occupied overlapping regions of cortex such that a given point on vSMC may fall within the region for several, neighboring body parts. Generally, there was a strong bias for motor responses on the precentral gyrus and somatosensory responses on the postcentral, but this boundary is not absolute: motor and sensory responses have been described on both gyri (Welker et al., 1957; Penfield and Boldrey, 1937).

Some examples of motor responses evoked by cortical stimulation are contralateral pulling of the mouth, twitching of the lips, simple opening or closing of the mouth, or swallowing. Sensory responses were usually reported as tingling in a given body part, sometimes with extreme precision, for example, the tip of the contralateral upper tooth. Responses rarely if ever corresponded to proprioceptive sensation or the perception of movement (Penfield and Boldrey, 1937).

Utilizing the increased spatial resolution of intracortical microstimulation (ICMS) in non-human primates (NHPs), which applies a small amount of current at varying depths in the cortex, the localization of individual muscles rather than body parts was possible. This technique confirmed the somatotopic organization, but revealed that individual muscles did not appear to have a somatotopic organization, and there were multiple loci that evoked movement from the same muscle (Huang et al., 2013). More recent ICMS studies in NHPs have shown that stimulating motor regions for a relatively long time scale (500 ms) results in complex movements of muscle groups (e.g., rhythmic jaw movements), as opposed to the simple twitches resulting from shorter stimulation. Nonetheless, it is important to note that linguistically meaningful sounds such as simple syllables or words have never been evoked during stimulation (Penfield and Boldrey, 1937).

With the advent of functional imaging such as PET and fMRI it became possible to noninvasively study humans during vocalization, with enough spatial resolution to investigate somatotopic maps. These studies have generally recapitulated the stimulation findings about the cortical representation of the lips, jaw and tongue (Petersen et al., 1988; Lotze et al., 2000; Hesselmann et al., 2004; Brown et al., 2009; Grabski et al., 2012).

While there is agreement about the general somatopic layout of the lips, jaw and tongue in vSMC, there have been inconsistencies in the localization of the larynx representation. Some studies have placed it at the most ventral position of vSMC, which is similar to the conclusions of many studies in both humans and primates (Grabski et al., 2012; Brown et al., 2009; Huang et al., 2013; Guenther, Ghosh, and Tourville, 2006). Others have noted a laryngeal motor area just dorsal of the lip representation (Brown et al., 2009; Grabski et al., 2012; Simonyan et al., 2009). This more dorsal location has not been described in NHPs, but it is located near sites that vocalization has been elicited using stimulation in humans (Penfield and Boldrey, 1937). While the existence of somatotopy in vSMC is fairly clear, its consequences for control of speech production are not clear.

Recently, electrocorticography (ECoG) was used to investigate the functional organization of ventral sensorimotor cortex during a task in which patients produced a large number of consonant-vowel syllables (Bouchard et al., 2013). ECoG in humans can be carried out in specific clinical conditions and involves the surgical implantation of an array of electrodes directly on the cortical surface, thereby providing high spatial and temporal resolution. Unlike the unnatural and simple movements of single articulators evoked by electrical stimulation, the production of meaningful speech sounds requires the precisely coordinated control of multiple articulators. The authors leveraged the variability in articulatory patterns associated with this large corpus of speech sounds to quantitatively assign a dominant articulator (lips, jaw, tongue, or larynx) representation to the cortical activity recorded at each electrode. Because of the superior temporal resolution of ECoG, cortical activity could be parsed out at the level of phonemes.

In this fashion, the cortical organization of all articulators could be derived without the need to isolate the movement of articulators using non-speech tasks or a limited set of carefully chosen speech sounds with constrained production. Although articulator representations were partially overlapping in both space and time, a dorsal-to-ventral organization of articulator representations was found (Figure 1.1b). This organization featured two separate representations of the larynx, with one site located ventral to the tongue, and the other dorsal to the lips. The dorsal representation is approximately the same as those seen in fMRI (Brown et al., 2009; Simonyan et al., 2009), but the ventral representation is similar to sites for throat seen in human stimulation studies (Penfield and Boldrey, 1937). The presence of this more dorsal site which was found over the precentral gyrus has not been described in NHP, and raises an interesting question about differences between humans and NHPs that may have a role in the production of speech. Evidence using transcranial magnetic stimulation in humans suggests a potential differentiation between localized representations of laryngeal muscles, with the cricothryroid muscle dorsally and the vocalis ventrally. Evoked movements of the vocalis in the ventral region have been confirmed using direct cortical stimulation as well.

1.3 Specializations within human vSMC for speech

NHPs have largely homologous orofacial anatomical structures and do vocalize, but do not have the capacity to produce the same repertoire of speech sounds as humans. The functional and anatomical differences between humans and NHPs with respect to speech may inform what features of oro-facial sensorimotor cortex are integral to speech production. One such difference has been evoked vocalization observed in human cortical stimulation studies. This was typically described as a prolonged phonation, sounding like 'ahhh. . . ', which continues throughout the duration of the stimulation (Penfield and Boldrey, 1937). Within vSMC, these sites are clustered along the central sulcus just dorsal to the representation of the lips on the precentral gyrus (Penfield and Boldrey, 1937). In NHP studies, a region in the ventral-most premotor cortex has been identified that, when stimulated, produces vocal fold movement. However, vocalization has never been produced from cortical stimulation of this or any other sensorimotor area in NHP (Jürgens, 2009).

Anatomically, two descending pathways exist in primates: a direct, bi-lateral projection between motor cortex and the oro-facial motor nuclei in the pontine and medullar level of the brain stem, and another indirect projection to the oro-facial motor nuclei via interneurons within the reticular formation. This indirect pathway interfaces with other descending cortical areas involved in vocal production at the reticular formation, such as the anterior cingulate cortex (Huang et al., 2013; Jürgens, 2009). Although the direct path is found in all primates, humans have an additional direct connection from larynx motor cortex to the nucleus ambiguus, which innervates the laryngeal muscles (Jürgens, 2002).

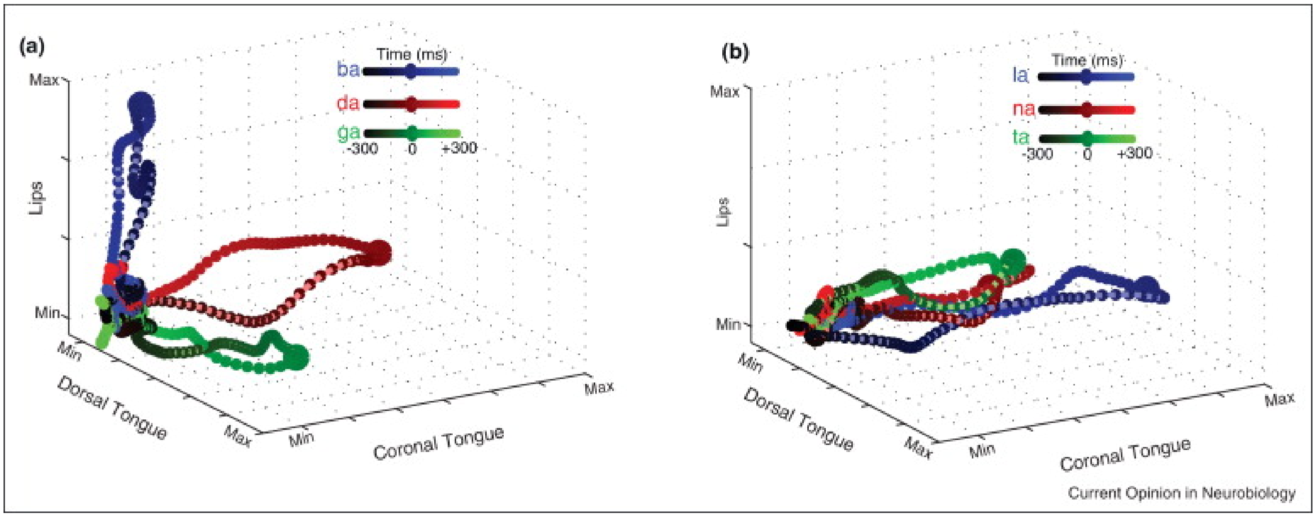


FIGURE 1.2: **vSMC electrode dynamics** Each axis corresponds to high gamma activity from a given electrode representing selected speech articulators (e.g., lips, dorsal tongue, coronal tongue). These plots help visualize the trajectory of the 'cortical state' across time during the production of a speech sound. Speech sounds that each have a different primary articulator (e.g., labial, coronal tongue, and dorsal tongue, in /ba/, /da/, and /ga/, respectively) ***(a)*** show divergent trajectories across the timecourse of the production, while speech sounds that have the same primary articulator (e.g., the coronal tongue in /na/, /la/, /ta/) ***(b)*** have very similar trajectories.

Given the short synaptic distance between vSMC and the muscles of the speech articulators, activity in vSMC is likely closely tied to the generation of movement of the speech articulators. Furthermore, the additional descending pathway from ventral laryngeal motor cortex found only in humans may be a neuroanatomical specialization for speech.

In humans, patients with lesions of unilateral oro-facial sensory-motor cortex suffer temporary dysarthria, or 'thickness of speech', but this improves with time until there is no noticeable deficit [2]. This is often also associated with deficits in nonspeech function as well, such as contralateral weakness of facial or tongue protrusion. However, bilateral loss of vSMC results in complete loss of voluntary control of the speech articulators (Jürgens, 2002). Together, lesion studies in humans suggest that vSMC is necessary for speech, and that there is some degree of redundant bilateral control. Lesion studies from NHPs suggest a specific role for ventral motor cortex in producing learned vocalizations. In NHPs, oro-facial motor cortex can be removed without affecting unlearned species-specific vocalizations(Sutton, Larson, and Lindeman, 1974). However, if the animal is trained to do a task that involves precise volitional oro-facial control (e.g., produce constant force with the tongue), deactivation of sensorimotor cortex results in pronounced deficits. This implies that oro-facial motor cortex is specifically recruited in the control of learned, volitional oro-facial tasks, and not more innate vocal behaviors (Jürgens, 2002). This discrimination between innate and learned vocal behaviors is thought to arise at the level of the direct versus indirect pathway; the direct is necessary for volitional articulator control while the indirect is necessary for innate orofacial behaviors (Simonyan et al., 2009; Jürgens, 2002).

A great deal of the evidence above points towards a special role of the larynx representation within vSMC for speech. More so than any other part of vSMC, the larynx seems to carry inconsistencies between humans and NHP that are relevant to speech. It has been proposed that the functional and anatomical differences in laryngeal motor representation may underlie some differences in capacity for speaking (Simonyan et al., 2009; Bouchard et al., 2013; Jürgens, 2002). It appears to represent an important exception to the general principles of somatotopic organization of the sensorimotor cortex and warrants further investigation.

1.4 What is the functional organization of speech sounds?

The somatotopic maps up to this point describe the representation of individual articulators on the cortical surface. However, the generation of speech sounds is not accomplished through the simple movement of a single articulator, but rather the precise coordination of multiple articulators. Therefore, in order to understand the functional organization of speech in vSMC it is necessary to move away from static descriptions of somatotopy and instead analyze the population-derived spatial patterns of cortical activity during unconstrained production of a variety of speech sounds. Bouchard et al. used principal component analysis to transform the population neural activity into a ‘cortical state-space’ that best describes the cortical patterns associated with the produced syllables. Capitalizing upon the high temporal resolution of ECoG, it was possible to temporally disambiguate the cortical activity associated with consonants and vowels (Figure 1.2).

An examination of the organization of both consonants and vowels in this cortical state-space revealed that different phonemes were clustered according to the major oral articulators engaged during production (i.e., the dorsal tongue, coronal tongue, and lips). Furthermore, a detailed analysis of phoneme representations revealed a rich, hierarchical organization of 'phonetic features', which also emphasized the major oral articulators, but additionally demonstrated that the place of constriction within a given articulator was the secondary organizing principle, followed by the degree of constriction. Therefore, the spatial patterns of cortical activity across multiple speech articulators were used to understand the organization of phoneme representations across the vSMC network. This organization likely reflects the coordinative patterns across articulatory motions during speech. The somatotopic and phonemic feature maps during speech production are both important principles of vSMC mesoscale spatial organization. Deriving the mathematical mapping from somatotopic organization to phonemic feature organization in this way is critical to understanding the role of somatotopy in speech production.

1.5 Conclusions

Previous research has described the basic organization of maps within human sensorimotor cortex, but we are only beginning to understand the functional significance of vSMC somatotopy in speech. Many of the same questions that were investigated decades ago are still relevant to the study of speech production today. What is the relevance of somatotopy to models of speech motor control? Where does the precise coordination of articulators originate? How does vSMC functionally relate to other speech areas? To what degree is the vSMC activity for a phoneme categorical and to what degree does it depend on surrounding phonemes? New tools that afford increased spatial and temporal pre- cision to record brain activity, combined with more detailed monitoring of speech articulators, will allow us to more fully address these questions in the near future.

Chapter 2

High-Resolution, Non-Invasive Imaging of Upper Vocal Tract Articulators Compatible with Human Brain Recordings

2.1 Introduction

Speech sounds are produced by the coordinated movements of the speech articulators, namely the lips, jaw, tongue, and larynx. Each articulator itself has many degrees of freedom resulting in a large number of vocal tract configurations. The precise shape of the vocal tract dictates the produced acoustics- however, at a coarse level, the same phoneme can be produced by many vocal tract configurations (Gick, Wilson, and Derrick, 2012; Johnson, Ladefoged, and Lindau, 1993; Perkell et al., 1993). For example, normal production of the vowel /u/ involves raising the back of the tongue towards the soft palate while protruding/ rounding the lips. Furthermore, the shape and size of individuals' vocal tracts can vary significantly (Narayanan et al., 2004), and as a result there is not a general (i.e. cross-subject) mapping from vocal tract configuration and resulting acoustics that is valid across speakers (Narayanan et al., 2004; Johnson, Ladefoged, and Lindau, 1993). Therefore, the precise shape of the vocal tract cannot be determined from observation of the acoustics alone. Furthermore, not all vocal tract movements have simultaneous acoustic consequences. For example, speakers will often begin moving their vocal tract into position before the acoustic onset of an utterance (Lofqvist and Gracco, 1999; Gracco and Lofqvist, 1994). Thus, the timing of movements cannot be derived from the acoustics alone. This ambiguity in both position and timing of articulator movements makes studying the precise cortical control of speech production from acoustics measurements alone very difficult. To study the neural basis of such a complex task requires monitoring cortical activity at high spatial and temporal resolution (on the order of tens of milliseconds) over large areas of sensorimotor cortex. To achieve the simultaneous high-resolution and broad coverage requirements in humans, intracranial recording technologies such as electrocorticography (ECoG) have become ideal methods for recording spatio-temporal neural signals (Bouchard and Chang, 2014; Bouchard et al., 2013; Herff et al., 2015; Kellis et al., 2010; Mugler et al., 2014; Pei et al., 2011). Recently, our understanding of the cortical control of speech articulation has been greatly enriched by the utilization of electrocorticography (ECoG) in neurosurgical patients However, previous studies have only been able to examine speech motor control as it relates to the produced speech tokens, canonical descriptions of articulators, or measured acoustics, rather than the actual articulatory movements (Bouchard and Chang, 2014; Bouchard et al., 2013; Herff et al., 2015; Kellis et al., 2010; Mugler et al., 2014; Pei et al., 2011). To date there have been no studies that relate neural activity in ventral sensorimotor cortex (vSMC) to simultaneously collected vocal tract movement data, primarily because of the difficulty of combining high-resolution vocal tract monitor with ECoG recordings at the bedside. The inability to directly relate to articulator kinematics is a serious impediment to the advancement of our understanding of the cortical control of speech.

In this study, our primary goal was to develop and validate a minimally invasive vocal tract imaging system. Additionally, we use novel, data-driven analytic approaches to better capture the shape of the articulators; synthesize perceptible speech from kinematic measurements; and combine our articulator tracking system with ECoG recordings to demonstrate continuous decoding of articulator movements. We collected data from six normal speakers during the production of isolated vowels (e.g. /ɑ/, /i/, /u/, /ɝ/) while simultaneously monitoring the lips, jaw, tongue, and larynx utilizing a video camera, ultrasound, and electroglottogram (EGG), respectively. We categorically related the measured kinematics to vowel identity and continuously mapped these measurements to the resulting acoustics, which revealed both shared as well as speaker specific patterns of vowel production. Application of unsupervised, non-negative matrix factorization (NMF) extracted bases that were often found to be associated with a particular vowel, and moreover allowed for a more accurate classification of vowels than traditional point-based parameterization of the articulators. Additionally, A central long-term goal of our work is to produce a speech prosthetic that transforms recorded brain signals into perceptually meaningful acoustics of speech. As speech production is mediated in the brain through control of the articulators, a first goal is to reconstruct intelligible speech from articulator measurements. Therefore, we synthesized auditory speech from the measured kinematic features and shows that these synthesize sounds are perceptually identifiable by humans. Finally, we demonstrated the feasibility of combining our noninvasive lip/jaw tracking system with ECoG recordings in a neurosurgical patient and demonstrate continuous decoding of lip-aperture using neural activity from ventral sensorimotor cortex. Together, our results suggest the methods described here could be used to synthesize perceptually identifiable speech from ECoG recordings.

2.2 Methods

This study was approved by the UCSF Committee on Human Research. All participants gave their written informed consent before participating. The individual in this manuscript has given written informed consent to publish these case details.

2.2.1 Task

Six speakers (5 males, 1 female) participated in this experiment. In order to mimic the conditions of a hospital bed, speakers sat at an incline with a laptop positioned at eye level 0.25– 0.5m away. To validate our articulatory monitoring system we had speakers produce vowels because they are well studied in the phonetics literature and much is known about their acoustics and articulatory bases (Harshman, Ladefoged, and Goldstein, 1977; Alfonso and Baer, 1982; Baer et al., 1991; Hillenbrand et al., 1995; Lindblom and Sundberg, 1971). Furthermore, the relationship between the shape of the vocal tract and the produced acoustics is the most direct for held vowels. Speakers were randomly presented with an audio recording (speaker KJ) of one of nine vowels (/ɑ/ae/ʌ/ɛ/ɝ/ɪ/i/ʊ/u) in an hVd context (e.g. 'hood') and then in isolation. Speakers repeated these tokens as they were presented following a brief (1 s) delay. For each speaker, between 30 and 50 repetitions of each vowel were collected. Only the isolated condition is examined here.

2.2.2 Data Acquisition and Analysis

During the production of each sound, we simultaneously tracked the produced acoustics, as well as the movement of the vocal tract (lips, jaw, tongue, and larynx) employing three imaging methods. First, in order to capture the movement of the lips and jaw, the lips of the speaker were painted blue and red dots were painted on the nose and chin (Fig 2.1Ai) and a camera (FPS = 30) was placed in front of the speaker's face such that all painted regions were contained within the frame and the lips are approximately centered (Noiray et al., 2011). In each frame of the video, lips and jaw position were determined using a hue threshold to extract the blue and red face regions, resulting in binary masks (Fig 2.1Aii). From the binary masks, we extracted the location of the jaw and the four corners of the mouth (upper/lower lip, left/right corners). The x and y position of these points were extracted as a time varying signal (Fig 2.1Aiii).

To image the tongue, an ultrasound transducer (Mindray M7 with C5-2s transducer) was held firmly under the speaker's chin such that the tongue was centered in the frame (Fig 2.1Bi). Video for both the camera and the ultrasound was captured at 30 frames per second (fps). The tongue contour for each frame was extracted using EdgeTrak, which uses a deformable contour model and imposes constraints of smoothness and continuity in order to extract the tongue from noisy ultrasound images (Li, Kambhamettu, and Stone, 2004). The output is an x and y position of 100 evenly placed points along the tongue surface (Fig 2.1Bii). Except where stated otherwise, our analyses parameterize tongue position as the vertical position of three equidistant points representing the front, middle, and back tongue regions (Fig 2.1Biii).

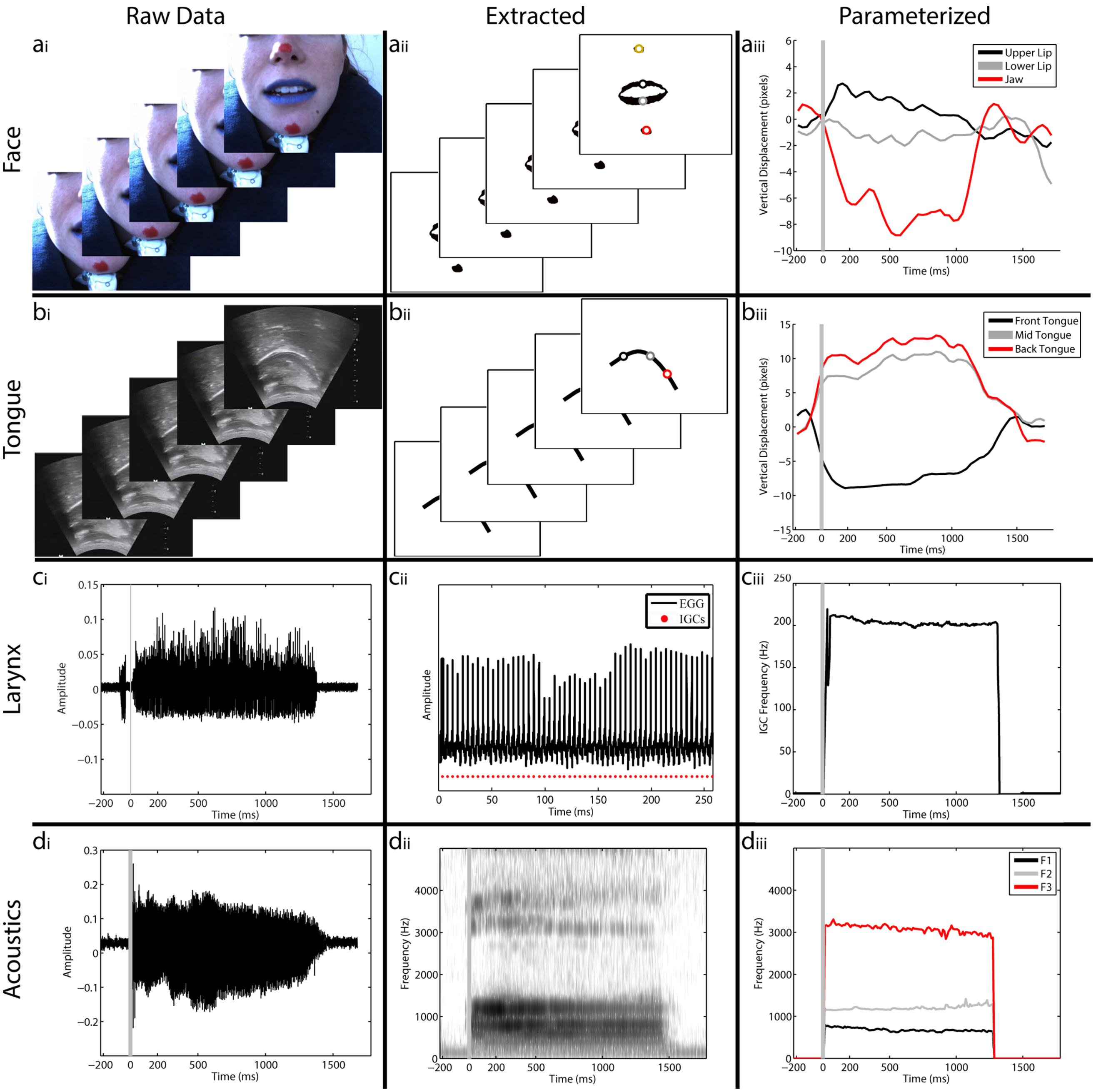

Raw Data
Extracted
Parameterized
Face
Tongue
Larynx
Acoustics
ai
aii
aiii
Upper Lip
Lower Lip
Jaw
Vertical Displacement (pixels)
Time (ms)
bi
bii
biii
Front Tongue
Mid Tongue
Back Tongue
Vertical Displacement (pixels)
Time (ms)
Ci
Cii
Ciii
Amplitude
Time (ms)
EGG
IGCs
Amplitude
Time (ms)
IGC Frequency (Hz)
Time (ms)
di
dii
diii
Amplitude
Time (ms)
Frequency (Hz)
Time (ms)
F1
F2
F3
Frequency (Hz)
Time (ms)
−200
0
200
500
1000
1500

FIGURE 2.1: **Data processing steps for facial, lingual, laryngeal, and acoustic data** ***(a)*** The lips of the speaker were painted in blue and dots were painted in red on the nose and chin. A camera was then placed in front of the speaker's face such that all painted regions were contained within the frame and the lips are approximately centered. Video was captured at 30 frames per second (fps) during speaking (i). Each frame of the video was thresholded based on hue value, resulting in a binary mask. Points were defined based upon the upper, lower, left and right extents of the lip mask and the centroids of the nose and jaw masks (ii). The X and Y position of these points was extracted as a time varying signal (iii). Grey lines mark the acoustic onset. ***(b)*** The tongue was monitored using an ultrasound transducer held firmly under the speaker's chin such that the tongue is centered in the frame of the ultrasound image. Video output of the ultrasound was captured at 30 fps (i). The tongue contour for each frame was extracted using EdgeTrak, resulting in an X and Y position of 100 evenly placed points along the tongue surface (ii). From these 100 points, three equidistant points were extracted, representing the front, middle, and back tongue regions which comprises our time varying signal (iii). ***(c)*** Instances of glottal closure were measured using an electroglottograph placed with contacts on either side of the speaker's larynx. The instances of glottal closure were tracked by changes in the impedance between the electrodes using the SIGMA algorithm (Thomas and Naylor, 2009). ***(d)*** Speech acoustics were recorded using a microphone placed in front of the subject's mouth (though not blocking the video camera) and recorded at 22 kHz (Fig 2.1di). We measured the vowel formants, F1-F4, as a function of time for each utterance of a vowel using an inverse filter method. For the extraction of F0 (pitch), we used standard auto-correlation methods.

The larynx was monitored using an electroglottogram (EGG) (EG2-PCX, Glottal Enterprises). The subject wore a band around the neck, and the EGG measured the electrical impedance across the larynx with electrodes in the neckband on either side of the thyroid. The opening and closing of the glottis during voiced speech creates changes in the impedance (Fig 2.1Ci). The instants of glottal closure (IGCs) in the EGG signal were found using the SIGMA algorithm (Fig 2.1Cii) (Thomas and Naylor, 2009). EGG recordings were collected on 3 of the 6 speakers.

Speech sounds were recorded using a Sennheiser microphone placed in front of the subject's mouth (though not blocking the video camera) and recorded at 22 kHz (Fig 2.1Di). The recorded speech signal was transcribed off-line using Praat (http://www.fon.hum.uva.nl/praat/). We measured the vowel formants, F1-F4, as a function of time for each utterance of a vowel using an inverse filter method (Fig 2.1Diii) (Ueda et al., 2007; Watanabe, 2001). Briefly, the signal was inverse filtered with an initial estimate of F2 and then the dominant frequency in the filtered signal was used as an estimate of F1. The signal was then inverse filtered again, this time with an inverse of the estimate of F1, and the output was used to refine the estimate of F2. This procedure was repeated until convergence and was also used to find F3 and F4. The inverse filter method converges on very accurate estimates of the vowel formants, without making assumptions inherent in the more widely used linear predictive coding (LPC) method. For the extraction of F0 (pitch), we used standard auto-correlation methods. Instants of glottal closure (IGCs) in the acoustic signal were estimated from the acoustics using the DYPSA algorithm (Naylor et al., 2007). To adjust for differences in utterance duration, we used linear interpolation to temporally warp each trial for all extracted features (not the raw signals) such that it was equal to the median trial duration.

2.2.3 Correlation Coefficient

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

$$R(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y} \tag{2.1}$$

Where σx and σy are the sample standard deviations of x and y, respectively.

2.2.4 Registration of Kinematic Data

To maximize the amount of data collected, subjects were recorded on multiple ‘blocks’, each lasting 10-15 minutes (as in the clinical setting). Different blocks could be collected on different days, and could be set-up by different experimenters. Furthermore, during a given block, there could be substantial movement of the subject, as well as the experimenter holding the ultrasound transducer. All of these are potential sources of artifactual variability (noise) in the data. Therefore, a necessary pre-processing step is to ‘register’ all the data so as to minimize this noise. For each trial (across all recording sessions), the optimal affine transformation (shifts, rotations, and scaling) was found to maximize the overlap (minimize the difference) of each pre-vocalization image to the median of all pre-vocalization images. This transform was then applied to all subsequent time points. We found this optimal transform in two ways (each with their potential shortcomings): 1) grid-search over parameter ranges of entire binary images, 2) analytic calculation of transform for extracted features; both methods gave similar results. The details of these different approaches are described below.

The first approach was to calculate the optimal translation (τ) for a grid of rotations (θ = [-20°: 1 :20°]) and uniform scaling's (β = [0.8: 0.01 :1.2]). For each image (Xi), first we convolved the binary image with radial Gaussian (N(0,3)) for smoothing. For the face data, we then centered each image by assuming that the center of the mouth is in the same location. After this initial data processing, we calculated the reference image (X*) as the grand median across all the pre-vocalization images. We refined X* over three iterations of the following procedure. For each Xi, we looped over both θ and β ranges, applied the transformations to the image, and performed 2D cross-correlations against X* to find the translation τ. For each image Xi, the [θ, β, τ triplet that gave the maximum cross-correlation values was taken as the optimal transformation:

$$\max_{\beta,R,\tau} R(X^*, X_i) \tag{2.2}$$

This was applied to the data and then we re-calculated the reference image as the grand median of this transformed data, and re-ran the above procedure (using the non-transformed images). This was done to ensure the use of an optimal reference image, which is important for our reference based registration method. From this optimal reference image, the optimal [θ, β, τ] triplet was found as described above, and applied to the data. While slow, this method considers all the data in the image and so is not sensitive to feature extraction variability. Indeed, this method does not require any features to be extracted from the data, and so could be run with out defining ROI's required for feature extraction. Furthermore, this method has lots of natural parallelization, both over parameters and trials, so compute time should decrease very well on cluster computing platforms.

The second approach for registering the data was to perform a so-called Procrustes analysis of the extracted landmarks in the images. For this analysis, we used the x,y coordinates of the extracted tongue contours (100 points) and the 6 landmark points on the face (nose, jaw, upper/lower lip, left/right corners of lip). The Procrustes problem is to find the uniform scaling (β), rotation matrix (θ), and translation (τ) that minimizes the difference between a set of points (Xi) and a reference set (X*):

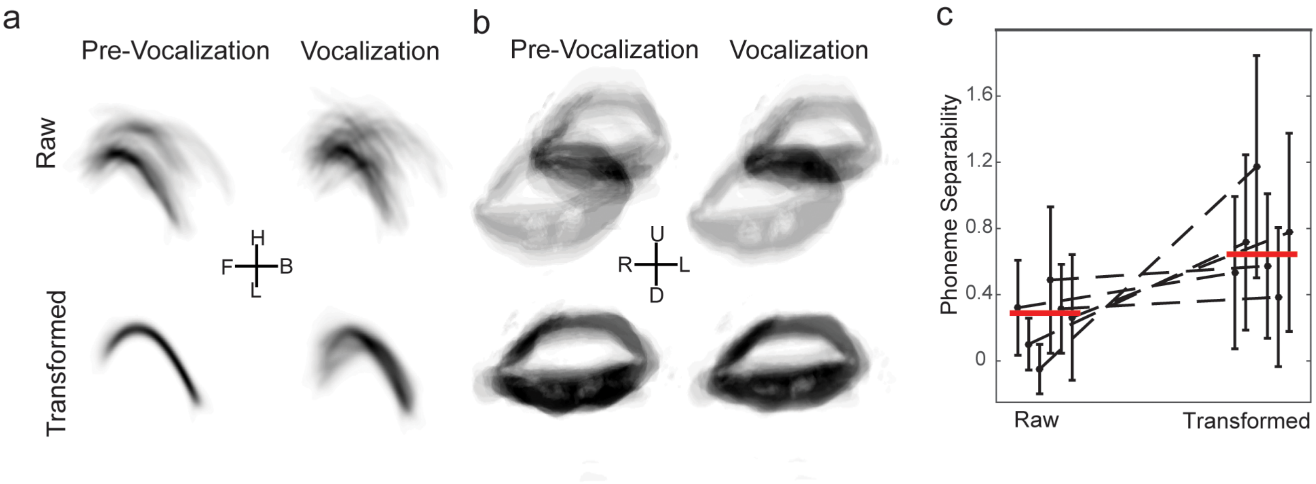


FIGURE 2.2: **Utterance-to-utterance registration of vocal tract data *(a)*** All tongue data from one subject. Left, top: overlay of raw tongue data during the pre-vocalization period for all trials; left, bottom: overlay of the same data after applying the optimal transformation. Right, top: overlay of raw tongue data during vocalization for all trials; right, bottom: overlay of the same data after applying the transformation from the pre-vocalization time. ***(b)*** All lip data from one subject. Left, top: overlay of raw lip data during the pre-vocalization period for all trials; left, bottom: overlay of the same data after applying the optimal transformation. Right, top: overlay of raw lip data during vocalization for all trials; right, bottom: overlay of the same data after applying the transformation from the pre-vocalization time.***(c)*** Quantification of efficacy of applying transform from pre-vocalization data to vocalization times: enhanced separability. We calculated the separability between vowels based on articulatory features during vocalization before (raw) and after (transformed) applying the transformation optimized for the pre-vocalization time. The transformed data consistently had increased separability. Black points: mean ± s.d. across vowel comparisons for a subject, red line: median across subjects.

$$\min_{\beta,R,\tau} \frac{1}{2}||X^* - \beta X_i\theta + 1\tau^T||\frac{2}{F} \tag{2.3}$$

Where $||X||\frac{2}{F} = trace(X^T X)$ denotes the squared Frobenius matrix norm.

Analogously to the grid-search procedure described above, we used the average shape across all pre-vocalization times as the reference set (X*) and we determined the Procrustes average of the pre-data data using an iterative procedure. First, we initialized X* to be the grand mean of the data points. Second, we solved the Procrustes problem (Eq.3) for each trial (Xi) using the reference X*. Third, we applied the transform to each trial, and updated X* as the grand mean of the transformed data points. This procedure was repeated until numerical convergence of X* (here, taken to be ΔX < 10-10). With this optimized average shape calculated, we then solved and store the transformations from the Procrustes problem (Eq.3) for each trial, and applied to subsequent time points for that trial. This method has the advantage of being computationally fast, but for the face, depends on good extraction of features.

We examined the cross-utterance phoneme separability (Bouchard et al., 2013) of different vowels based on extracted parameter values before and after this registration. This quantifies the difference in the average distance between vowels and the average distance within a vowel, so larger values correspond to tighter distributions within a vowel and larger distances between vowels. We also used this metric to visually compare the time-courses of vowel identity structure in the acoustic and kinematic data. (Fig 2.3). For a given feature set, phoneme separability (PS) is defined as the average difference of the median cross-cluster distances (AC) and median within-cluster distances (WC).

$$PS = (\overline{AC - WC}) \tag{2.4}$$

This measure expresses the average distance between vowels categories and the tightness of the category clusters.

2.2.5 Classification

As described below, we used Naïve Bayes classifiers based on a latent dimensional representation (found by LDA) of acoustic, kinematic, and 'shape' features. These classifiers were trained and tested using a bootstrapped cross-validation procedure.

We applied Linear Discriminant Analysis (LDA) to the mean data from the central 1/5th of timepoints, using vowels as class identifiers. LDA is a supervised dimensionality reduction algorithm that finds the projection that maximizes the linear discriminability of the (user defined) clusters. LDA can be thought of as a discrete version of the general linear models used for continuous mapping (see below). Multiclass LDA was performed on the (central 1/5th of) vowel acoustics (Fig 2.4), articulator features (Fig 2.4), and NMF representations (Fig 2.6) by computing the matrix $L^* = LS^{-1/2}$, where L and S are the class centroids and common with-in class covariance matrices, respectively. Classes were the different vowels. We then took the singular-value decomposition of the covariance matrix of L^*, and projected the data into the corresponding eigenspace (note that LDA necessarily results in a n-1 dimensional space). To ease comparisons across different speech representations, the first 3 latent dimensions (L^3) were used for all feature sets (e.g. kinematic, acoustic, NMF).

We trained and tested Naïve Bayes classifiers to predict vowel identity based on different features of the produced vowels. The Naïve Bayes approach makes the simplifying assumption that each of the input features (in our case, projections of data into latent dimensions found by LDA) are conditionally independent, given the class identity (in our case, vowel identity). Under this assumption, the posterior probability of the class (V_k) given the features (F_i) is:

$$P(V_k|F_i) = \frac{P(V_k)}{P(F_i)}\Pi_i P(F_i|V_k) \tag{2.5}$$

We used the maximum a posterior (MAP) as an estimator of class identity:

$$\hat{V} = \max_{k \in K} \frac{P(V_k)}{P(F_i)}\Pi_i P(F_i|V_k) \tag{2.6}$$

Here, we used a Naïve Bayes classifier as opposed to the posterior probabilities from the LDA because it gave slightly better performance.

2.2.6 Bootstrapped Cross-validation Procedure

To train and test the Naïve Bayes classifiers, we used cross-validation on randomly selected (with replacement) subsets of the data. Specifically, within a 50 iteration bootstrap procedure, random 80% subsets of the data were used to train the classifiers, and model performance was tested on the 20% of data not used in (any part) of the training procedure. Performances are reported as the statistics (e.g. mean) across these bootstrap samples.

2.2.7 Non-negative Matrix Factorization

A common method for unsupervised learning of reduced basis sets is principal components analysis (PCA), which finds an orthogonal basis set that optimally captures the directions of highest variance in the data. However, a critique of PCA is that the bases often bear little resemblance to the data from which they were derived (Lee and Seung, 1999). Although this may be of little consequence if quantitative performance is the primary interest (as is often the case in machine learning), when understanding the bases is important (as is often the case in science), this lack of resemblance to data can hinder interpretability (Lee and Seung, 1999). Non-negative matrix factorization (NMF) has been used to extract 'meaningful' bases from data that consist of only positive values, such as images and movies (as in our data set) (Lee and Seung, 1999; Donoho and Stodden, 2003; Kim and Park, 2008). Additionally, NMF can be formally related to K-means clustering, and can result in clusters that are more robust than K-means (Kim and Park, 2008). NMF is a dimensionality reduction technique that extracts a predetermined number of bases ($B \in \mathbb{R}^{mxk}$) and weights ($W \in \mathbb{R}^{nxk}$) that linearly combine to reconstruct the non-negative data ($A \in \mathbb{R}^{mxn}$), such that $k < \min(n, m)$ under the constraint that both the bases and weights are strictly non-negative:

$$A \approx BW^T; B, W \geq 0 \tag{2.7}$$

The solutions B and W are found by solving the constrained optimization problem:

$$\hat{B}, \hat{W} = \min_{B,W} \frac{1}{2} ||A - BW^T|| \frac{2}{F}; s.t.B, W \geq 0 \tag{2.8}$$

This matrix factorization has no closed form solution, and so is often found through numerical approximation (here, we used the Matlab 'nnmf' function). Additionally, NMF does not have a unique solution, and so good initialization is important. Therefore, we performed an initial search using the computationally less expensive (but less robust) multiplicative method with 20 random initializations and terminated the optimization procedure at 10 iterations or a numerical tolerance in the solution of

10^{-10}. The result of this search that had minimal final reconstruction error was used as the initial conditions in a more exhaustive search using the non-negative alternating least-squares approach. Here, the procedure terminated after 100 iterations or a numerical tolerance of 10^{-16}.

We applied this procedure separately to the (registered) face and tongue images from the central 1/5th of each utterance. We applied NMF to the data for faces and tongues separately because the NMF objective function (Eq. 8) finds solutions that directly minimize the pixel level reconstruction of the raw data. Therefore, because the number of pixels associated with the tongues and lip images are very different, a combined analysis would result in differential weighting of these articulators in the extraction of the basis. Indeed, the number of non-zero pixels in the lip data is much greater than the tongue data, indicating that lips would be given a much larger importance in the NMF reconstruction, which is undesirable.

2.2.8 Linear Mapping

To understand the continuous relationship between articulator position and resulting acoustics, we utilized general linear models. For each trial, the average value over the middle fifth of the vowel was calculated for each articulatory and acoustic feature. These averages were then z-scored across trials to remove differences in scaling between recording modalities. We then used the Boostrapped Adaptive Threshold Selection (BoATS) algorithm to estimate regularized linear models using an 80-10-10 cross-validation procedure to derive model weights from training data (80% of data), determine an optimal regularization parameter (10% of data) and calculate model performance on test data (10% of data) (Bouchard and Chang, 2014; Bouchard, 2015). Briefly, first, to derive null distributions of weights (β^{*rnd}) and model performance (R^2_{reg}), we randomly permuted (200 times) each input feature independently relative to the output feature on a trial-by-trial basis. Second, within a 200 iteration bootstrap procedure, random 80% subsets of the data were used to derive linear weights for the models using the equation:

$$y = \beta X \tag{2.9}$$

Where y is the predicted feature, X is the set of predictor features, and β are the weights that describe the linear relationship. From this, we arrived at an estimate of weights (β^{*obs}) for each input feature predicting and output feature. We then reduced the dimensionality of the input features (X) by comparing the model weights between the observed and randomized data sets to identify input features with weights that were different between the two conditions. Specifically, features (X_j) were retained if the weight magnitude ($|\beta^{*obs}|$) was greater than the mean plus a threshold multiple (regularization parameter) of the standard deviation of the distribution of weight magnitudes derived from the randomization procedure ($|\beta^{*rnd}|$. Finally, we re-trained models on the training data based only on this reduced set of cortical features to arrive at optimal weights (β^{*reg}) and determined decoding performance (R^2_{reg}) on test data (10%) not used in training. The choice of threshold was chosen to optimize the predictive performance on the regularization data (10%). The model performance was taken as the mean of R^2_{reg} values across bootstrap test samples. This quantifies the expected value of predictive performance across randomly selected training and test samples.

2.2.9 Speech Synthesis

Statistical parametric approaches are the dominant approach for speech synthesis in recent years for their flexibility in mapping arbitrary feature descriptions of speech and language to intelligible speech (Zen, Tokuda, and Black, 2009). In traditional speech synthesis, text or sequences of phonemes are input, which are then analyzed to get relevant linguistic and contextual information into building a supervised model that optimizes the prediction of speech given its context. Since this approach assumes noiseless inputs of linguistic categories, it is not usable, as it is, for the current task. The contextual information in this study is continuous and articulatory, and noisy. This requires building a speech synthesizer that can work on such inputs to optimally predict speech. Clustering and Regression Trees (CART) (Breiman, 2001) is a widely used model in statistical speech synthesis for mapping contextual feature representations into a synthesizable feature representation of speech. Speech parameters were extracted for each trial of vocalization from each subject. This comprises joint vectors of Fundamental

Frequency (F0), Mel-Cepstral Coefficients, excitation strengths and voicing. This description is sufficient to resynthesize perceptually lossless speech (Black, 2006). In the current setting of estimating these representations from articulatory features, the context comprises continuous-valued questions about the spatial co-ordinates of various tracked features in the vocal tract (e.g lip-width > 3.8 units?). Based on the configuration of vocal tract considered, the articulatory streams include points on the tongue or lips or a combination of both to model the produced acoustics. These articulatory feature streams were resampled at the same frequency of the speech, so as to create aligned vectors for training and so that the synthesized acoustics were at the same sampling rate as the produced acoustics for perceptual comparisons.

The CART model itself is a decision tree that hierarchically clusters data in subsets that are optimally described by categorical or continuous valued questions about given aspects of the training data. Questions about the appropriate articulatory features were greedily chosen in CART training to best reduce the variance in the data due to the split. The variance reduction $I_v(N)$ for a node N, that split the acoustic data S into subsets S_t and S_f is given by:

$$I_v(N) = \frac{1}{|S|}\Sigma_{i\in S}\Sigma_{j\in S}\frac{1}{2}(x_i - x_j)^2 - (\frac{1}{|S_t|}\Sigma_{i\in S_t}\Sigma_{j\in S_t}\frac{1}{2}(x_i - x_j)^2 + \frac{1}{|S_f|}\Sigma_{i\in S_f}\Sigma_{j\in S_f}\frac{1}{2}(x_i - x_j)^2) \quad (2.10)$$

The decision tree was recursively grown until a criterion is met, like the minimum number of data points within a cluster. A stop value of 50 was used, as the minimum number of data points within a subcluster. The mean and variance statistics of the subset of data points within the final clusters were stored at the leaf nodes of the trees. At runtime for synthesizing speech from given articulatory trajectories, the CART trees were traversed and the mean speech vectors at the leaf nodes were sequentially picked at the frame rate of the articulatory data and synthesized.

We report the performance of the model both objectively and subjectively. For objective quantification, we used the Mel-Cepstral Distortion (MCD) (Toda, Black, and Tokuda, 2008), which is a normalized sum Euclidian distance between a sequence of synthesized cepstral features and those of the corresponding reference acoustic stimuli. MCD for a reference 24-dimensional mel-cepstral vector and an estimate is given by:

$$MCD = \frac{10}{ln(10)}\sqrt{\Sigma_{0<d<25}(mc_d^y - mc_d^{\hat{y}})^2} \tag{2.11}$$

For subjective evaluation, we used human subjective evaluation via the Amazon Mechanical Turk.

2.2.10 Mechanical Turk for Subjective Assessment of Synthesis

Perceptual listening tests were conducted on the Amazon Mechanical Turk. The Mechanical Turk is a crowdsourcing portal where paid online volunteers perform tasks like annotations, perceptual judgments etc., called HITs (Human Intelligence Tasks). It is possible to constrain the task to be assigned to volunteers from a geographical region or those with a desired skill set or HIT success rate. To evaluate the speech synthesis outputs of different articulatory representations, a held out set of articulatory trajectories is synthesized and HITs are created such that qualified Turkers judge each synthesized audio stimulus. The task itself is vowel identification based on the audio of each stimulus. In this forced choice identification task, for each audio stimulus, Turkers were asked to choose one among nine vowels that best identifies the vowel as they perceived it. Illustrative examples of each vowel (e.g., /ae/ as in /CAT/) were also provided to help those without formal phonetic knowledge. While quality control is hard, some metrics like the HIT response time can be thresholded to weed out spammers among the volunteers. Unless reported otherwise, all listening tests were conducted with no restrictions on the location of the Turker. A HIT success rate of 80% was used to select only the genuine Turkers. HITs were randomly created and assigned such that each stimulus was identified by at least 10 Turkers. A HIT response time threshold of 30 seconds was used to filter out spurious Turkers.

2.2.11 ECoG Subjects and Experimental Task

One native English speaking human participant underwent chronic implantation of a high-density, subdural electrocortigraphic (ECoG) array. Our recordings were from the language dominant hemisphere (as determined with the Wada carotid intraarterial amybarbital injection), which was the right hemisphere in this patient. Participants gave their written informed consent before the day of surgery. The participant read aloud a set of words and pseudo-words (e.g. 'Leakst Skoot') (Noiray et al., 2011).

2.2.12 Anatomical location of vSMC

We focused our analysis on the ventral ("speech") portion of the sensory-motor cortex (vSMC). vSMC is anatomically defined as the ventral portions of the pre-central and post-central gyri, as well as the gyral formation at the ventral termination of the central sulcus, known as the sub-central gyrus. Visual examination of co-registered CT and MR scans indicate that the ECoG grid in the patient covered the spatial extent of vSMC (Bouchard and Chang, 2014; Bouchard et al., 2013).

2.2.13 ECoG Data Acquisition and Signal Processing

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

The time series from each channel was visually and quantitatively inspected for artifacts or excessive noise (typically 60 Hz line noise). Artifactual recordings were excluded from analysis, and the raw recorded ECoG signal of the remaining channels were then common average referenced. For each channel, the time-varying analytic amplitude was extracted from eight bandpass filters (Gaussian filters, logarithmically increasing center frequencies (70-150 Hz) and semi-logarithmically increasing band-widths) with the Hilbert transform. The high-gamma (HG) power was then calculated by averaging the analytic amplitude across these eight bands, and then this signal was down-sampled to 200

Hz. HG power was z-scored relative to the mean and standard deviation of baseline (i.e. subject at rest in silence) data for each channel. Throughout, when we speak of HG power, we refer to this z-scored measure, denoted below as HG.

Principal components analysis (PCA) was performed on the set of all vSMC electrodes for dimensionality reduction and orthogonalization. This also ensures that the matrices in the calculation of least mean squared error estimators (from regressions below) were well scaled. For each electrode (e_j of which there are n, n = 59) and time point (t, of which there are m), we calculated the high-gamma power. The HG(e,t) were used as entries in the nxm data matrix D, with rows corresponding to channels (of which there are n) and columns corresponding to the number of time points (of which there are m). Each electrode's activity was z-scored across time to normalize neural variability across electrodes. PCA was performed on the nxn covariance matrix Z derived from D. The singular-value decomposition of Z was used to find the eigenvector matrix M and associated eigenvalues λ. The PCs derived in this way serve as a spatial filter of the electrodes, with each electrode ej receiving a weighting in PCi equal to mij, the i-jth element of M, the matrix of eigenvectors. For each point in time, we projected the vector $HG(e,t)$ of high-gamma activity across electrodes into the leading 40 eigenvectors (M^{40}):

$$\Psi(t) = M^{40} \bullet HG(e,t) \tag{2.12}$$

We used 40 PCs, which accounted for 90% of the variance.

2.2.14 Kinematic Feature Decoding Model

The $\Psi(t)$ (equation 12) served as the basis for training and testing optimal linear predictors of lip aperture ($A(t)$) over time using BoATS algorithm described above. We used a simple linear model to predict the lip aperture from $\Psi(t-\tau)$:

$$\hat{A}(t) = \beta \bullet \Psi(t-\tau) + \beta_0 \tag{2.13}$$

Where $\hat{A}(t)$ is the best linear estimate of $A(t)$ based on the cortical features. The vector of weights β that minimized the mean squared error between $\hat{A}(t)$ and $A(t)$ was found through multi-linear regression and cross-validation with regularization (see above). Based on our previous work [14], we used τ = 100ms.

2.2.15 Statistical Testing

Results of statistical tests were deemed significant if the probability of incorrectly rejecting the null-hypothesis was less than or equal to 0.05. We used paired Wilcoxon sign-rank tests (WSRT) for all statistical testing.

2.3 Results

For the initial characterization and validation of these methods we focused on data collected from speakers during the production of American English vowels, as these are a well studied and understood subset of speech sounds that engage the articulators monitored here. Specifically, we performed a variety of analyses to validate our methodology by comparing to previous results across a variety of domains, and propose new methods for measuring, parameterizing, and characterizing vocal tract movements. First, we describe techniques for reducing artifacts from recorded articulator videos, allowing us to combine data across different recording sessions. Next, we show the measured acoustics and articulator position time courses, and quantify the extent to which acoustic and kinematic features can discriminate vowel category, both of which are in good agreement with classical studies of vowel production. In line with the categorical conceptualization of speech, we describe a data-driven approach to extract vocal tract shape using non-negative matrix factorization (NMF). This method discovers 'shapes' that allow for more accurate classification of vowels than a priori defined parametric descriptions of the articulator positions. We then transition from categorical to continuous mappings between articulators and acoustics. Using the measured articulator positions, we assessed how articulatory features and acoustics linearly map to one another. Next, we synthesized speech from articulator positions and demonstrate that the processed articulatory trajectories retain sufficient signal to synthesize audio that can be perceived as the intended vowel. Finally, to illustrate the potential of combining articulatory tracking with brain recordings, we demonstrate robust decoding of a speech articulatory movement using multi-linear methods.

Our goal was to develop an articulatory tracking system compatible with electrocorticgraphy (ECoG) recordings at the bedside. This imposes several strong constraints on our experimental protocol. In particular, because our ECoG recordings are taken from neurosurgical patients, it is not possible to secure any apparatus to the patients' head. Additionally, only a limited amount of data can be collected in a given recording session, and so data are often taken on multiple recording sessions. Finally, the recording equipment must be as electrically quiet as possible so as to not interfere with the electrical recordings from the brain. Thus, our recordings in non-clinical speakers were subject to the same experimental constraints and multi-session recordings. Our approach combined the simultaneous use of ultrasonography to track the tongue, videography to monitor the mouth and jaw, and electroglotiography to measure the larynx. The raw data from this system and initial extraction of vocal tract articulators and parametric tracking is displayed in Fig 2.1. We collected data from six American English speakers (5 male, 1 female) during the production of hVd (e.g. "hood") words and sustained production of the corresponding vowels. The results presented here are focused on the vowel hold segment of the task, which included 1813 vocalizations (N = 292, 290, 197, 270, 391, 373 for the six speakers).

2.3.1 Utterance-to-utterance Registration of Vocal Tract Data

For both the ultrasound and videography recordings, a major source of artifactual variability introduced by our constraints was inconsistency in the position of the sensors (ultrasound transducer and camera), resulting in translations, rotations, and scaling differences in the plane of recordings. For example, the images shown in the top row of Fig 2.2a and b display the mean tongue and lip shapes extracted from the raw data at pre-vocalization times, as well as during the center (central 1/5th) of the vocalized vowels (N = 292 for all plots). In all plots, there are clear translations, rotations, and scaling differences between frames. (e.g. the translation and scaling of the mouth). These experimental artifacts are clearly a serious impediment to analyzing the data.

To correct for these experimental aberrations, we registered the images based on pre-vocalization frames with three simplifying assumptions: (1) the vocal tract is assumed to be the same across all vocalizations during pre-vocalization, (2) the position of the sensors is stable on the time-scale of a single vocalization, and (3) transformations are assumed to be affine (rotation, translation, and scaling). Ultrasound and videography from each trial were registered by first finding the transformations (translation, rotation, and scaling) that maximized the overlap of the pre-vocalization data (on an utterance-to-utterance bases), and then applying these transformations to subsequent time points. The details of this procedure are described in the Methods. Briefly, for each trial (across all recording sessions), the optimal affine transformation (translation, rotation, and scaling) was found to maximize the overlap of the pre-vocalization images to the median image (after an initial centering operation). This transform was then applied to all subsequent time points. We found this optimal transform in two ways (each with their potential shortcomings): 1) brute force search of binary images, 2) analytic calculation of affine transform for extracted features (i.e. 'Procrustes Analysis'); both methods gave similar results.

We found that image registration removed much of the obviously artifactual variability in the images. The images shown in the bottom row of Fig 2.2a and b display the mean tongue and lip shapes after registration for pre-vocalization times, as well as during the center of the vocalized vowels. For the pre-vocalization data, the mean of the transformed images is clearly less variable then the mean of the unregistered extracted images for both the tongue (Fig. 2.2a, left column) and the lips (Fig. 2.2b, right column). For example, the large translation and scaling of the mouth have largely been removed. This validates that our procedure is working as expected. Importantly, applying the transformation optimized on pre-vocalization data to data during vocalization times greatly cleaned up these images as well (Fig. 2.2a,b, right columns).

We checked that the transformations derived from the pre-vocalization data were removing artifactual variability while preserving signal useful for discriminating the different vowels. For this, we extracted articulatory features (e.g. lip aperture, front tongue height) from the central 1/5th of each utterance. Based on these features, we calculated the separability between the different vowels, which measures the distance between the data for different vowels relative to the tightness of the data for the same vowel. In Fig 2.2c, we plot the separability for each subject before (raw) and after (transformed) applying the optimal transformation from the pre-vocalization data (black: mean ± s.d. for individual subjects; red lines, median across subjects). For each subject, the average separability of the vowels was enhanced by the application of the transformation. Importantly, the subjects with the greatest enhancement were those that had the worst separability before the transformation, indicating that the degree of enhancement scales with the amount of artifact present. Together, these analyses demonstrate that our method of data registration removes artifactual variability due to data acquisition and enhances the signal useful for differentiating the vowels. This allows us to combine data acquired across different recording sessions.

2.3.2 Articulatory and Acoustic Feature Time-courses and Classification

To examine how the articulatory and acoustic measures change over the course of vowel production, we produced time-courses for each feature. This visualization allows for initial validation that our articulator monitoring system is producing meaningful measurements. For each trial, we extracted acoustic features (F0-F4) and articulatory features (front, mid, and back tongue, lip aperture and lip width) over the time-course of the vowel utterance. To eliminate differences in scale between features, we first z-scored each feature across all trials. For each speech feature, the average and variance was calculated across trials of the same vowel.

The acoustic and articulatory features for speaker 1 are plotted in Fig 2.3a-b. Each color shows the average trace for a different vowel, and error bars show standard error. The grey region marks the central 1/5th of the vocalizations. For the acoustic features, F1 through F4 all exhibit considerable separation between vowels during steady state production, and the relative magnitudes between vowels are consistent with previous literature on vowel acoustics [2,24,26]. As an example, the vowel /i/ (black) has the highest F2, but very low F1. F0 (pitch) shows little separability between vowels, but does show a consistent pitch lowering for which is consistent with previous literature on intrinsic pitch. For the articulatory features, tongue height and lip aperture also demonstrate clear vowel category structure, while lip width did not vary consistently between vowels. The lack of category structure in lip width measurements may partially reflect the fact that movements in lip width were small relative to lip opening and tongue movements. Interestingly, while tongue height measures all reach a steady state during production of the vowel, lip aperture continuously changes during the trial, reaching maximal opening around onset, and beginning to close during the production of the vowel. The positions of the speech articulators during vowel production are in line with previous descriptions (Gick, Wilson, and Derrick, 2012; Baer et al., 1991; Lindblom and Sundberg, 1971; Alfonso and Baer, 1982; Kent and Moll, 1972). For example, production of the vowel /ɑ/ resulted in lowering of the front and mid tongue points, and raising the back tongue, all consistent with the description of /ɑ/ being a 'low-front' vowel. The timing of movement onset is also similar to previous descriptions: most articulatory movement started shortly before acoustic onset, reached steady state shortly thereafter, and remained in position well past acoustic offset (i.e. end of phonation) [43]. Additionally, lip movements tend to precede tongue movements.

To quantify the dynamics of vowel category structure for both acoustic and kinematic features, we first performed linear discriminants analysis (LDA) on all the acoustic and kinematic features (see Methods). LDA is a dimensionality reduction method that finds the lower-dimensional manifold that allows best linear discriminability of the (pre-defined) categories (here, the vowels). We projected the acoustic and kinematic features into the first three dimensions of the LDA space, and calculated the phoneme separability at each time point. This allows comparison between acoustic and kinematic features in the same number of dimensions in an orthogonalized space (so Euclidean distance is valid). The phoneme separability was calculated for each time point for each subject, then averaged across subjects. In Fig 2.3c,d, we present the average separability for the acoustic (Fig 2.3c) and kinematic features (Fig 2.3d)(mean ± s.e.m. across subjects). For acoustic features (Fig 2.3c), phoneme separability between vowels exhibits steep onsets and offsets and remains steady during vocalization, which is expected given that there are no measureable acoustics outside of the vocalized region (by definition).

For kinematic features (Fig 2.3d), phoneme separability is lower in magnitude and has a more gradual time course, rising before acoustic onset, peaking shortly after, and falling slowly after offset. Together, these time courses suggest that our method produces articulatory and acoustic measurements with reasonable timing and magnitude for each of the vowels measured. However, the difference between acoustic and kinematic time courses is an important issue to consider for understanding the cortical control of speech production. Specifically, there are clear movements of the articulators with no simultaneous acoustic consequences, emphasizing the importance of explicitly measuring articulator kinematics.

As the identity of a vowel is defined not by a single feature, but by the relationships amongst multiple features, we next visualized how the vowels clustered in multi-dimensional acoustic and kinematic spaces. We took the average feature value during the steady state portion of each vocalization (central 1/5th) for each articulatory and acoustic feature and labeled each trial according to the vowel spoken. In the acoustic space (Fig 2.3e), the different vowels shows very little overlap. In the kinematic space (Fig 2.3f) there are distinct regions for each vowel, but there is large overlap between vowels. The difference in overlap between kinematics and acoustics may partially be due to a larger degree of noise in the kinematic recordings. To quantitatively identify the features that best discriminate between vowels, we determined the contribution of each acoustic and articulatory feature to each of latent dimensions in the LDA space. On average, the acoustic LDs primary contributions were from F2 for LD1, F1 for LD2, and F3 for LD3. The first two articulatory LDs are dominated by tongue height, while the third is predominantly lip opening.

Finally, to quantify the extent to which acoustic and kinematic features can discriminate vowel category, we used a naïve Bayes classifier (see Methods) to predict vowel identity from the first 3 LDs derived from vowel acoustics, lip features alone, tongue features alone, and all kinematics (Fig 2.3g: black: mean and standard error for individual subjects; red lines: median across speakers). Acoustics are the best predictor of vowel category, with on average 88% correct classification, and classification based on the lips alone (24%), tongue alone (43%), and all kinematic features combined (52%), all performed significantly better than chance (11%) (*: $P < 0.05$, WSRT, $N = 6$). Importantly, performance of all kinematic features is significantly higher than either lip or tongue features alone (*: $P < 0.05$, WSRT, $N = 6$) demonstrating that there is non-redundant information between the lips and tongue. All of these findings are consistent with classic descriptions of the articulatory and acoustic bases of vowel, and provide further validation of our recording system and registration methods (Hillenbrand et al., 1995; Alfonso and Baer, 1982; Maddieson and Disner, 1984).

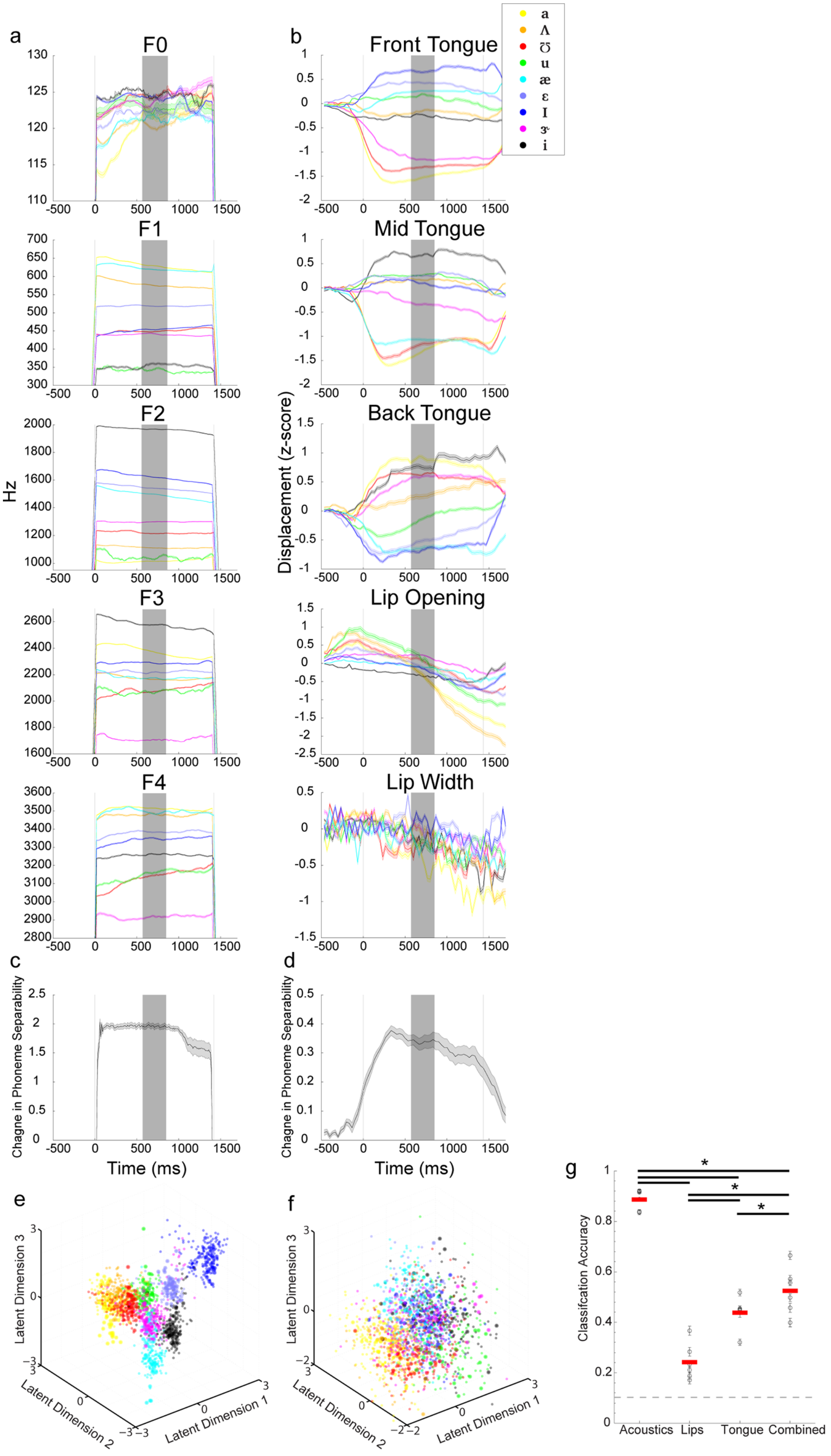

a
F0
130
125
120
115
110
F1
700
650
600
550
500
450
400
350
300
F2
2000
1800
1600
1400
1200
1000
Hz
F3
2600
2400
2200
2000
1800
1600
F4
3600
3500
3400
3300
3200
3100
3000
2900
2800
-500
0
500
1000
1500
b
Front Tongue
Mid Tongue
Back Tongue
Lip Opening
Lip Width
Displacement (z-score)
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
● a
● ʌ
● ʊ̆
● u
● æ
● ɛ
● I
● ɝ
● i
c
Chagne in Phoneme Separability
2.5
2
1.5
1
0.5
0
Time (ms)
d
Chagne in Phoneme Separability
0.5
0.4
0.3
0.2
0.1
0
Time (ms)
e
Latent Dimension 3
Latent Dimension 2
Latent Dimension 1
3
0
−3
f
Latent Dimension 3
Latent Dimension 2
Latent Dimension 1
3
0
−2
g
Classifcation Accuracy
1
0.8
0.6
0.4
0.2
0
*
*
*
Acoustics
Lips
Tongue
Combined

FIGURE 2.3: **Articulatory and acoustic feature time-courses and classification** ***(a-b)*** Average time course of formant values(a) and articulator position (b) for each of the 9 vowels examined. Traces shown are for a single subject (speaker 1). Each trial was warped using linear interpolation so that all trials were of equal length. Grey lines mark the acoustic onset and offset. Error bars denote standard error. Shaded region marks the time window used for LDA and classification analyses in (e-g). ***(c-d)*** Change in cluster separability across the trial for acoustic (c) and articulatory (d) features. Error bars denote standard error.***(e-f)*** LDA projections of formant values (e) and articulator position (f) drawn from the middle fifth of each trial across all speakers. All values are z-scored across all trials. Each dot marks the values for a single trial. Color denotes vowel spoken during the trial. Larger dots mark trials from a single speaker (same as in Fig 2.4i). ***(g)*** Classification performance resulting from running a 50x cross-validated naïve Bayes classifier on the mid-vowel acoustic and kinematic measurements. Each dot denotes an individual speaker with error bars denoting standard error across the cross-validations. Red line marks the median performance across speakers. Horizontal lines denote statistical significance ($P < 0.05$, WSRT, $N = 6$).

2.3.3 Unsupervised Extraction of Vocal Tract Shape with Non-negative Matrix Factorization Improves Vowel Classification

The modest classification performance of vowel identity based on articulator kinematics could be due to a number of causes. For example, it is likely that, even with image registration, the measurement noise in our articulatory imaging system and extraction procedures is larger than that of the collected acoustics and formant extraction procedures. Alternatively, the transformation from articulator configurations to acoustics could be highly non-linear, or very small differences in the vocal tract shape could lead to large differences in the acoustic output. However, the poor performance could also reflect the parameterization we chose to describe the articulators. Although well motivated by the literature, this parameterization was not entirely data driven (i.e. it was determined by the experimenters, not derived from the data de novo), and does not capture the richness of the full articulator shapes.

One reasonable choice of basis images that describes the vocal tract would be the mean image associated with each vowel. For example, in Fig 2.4a,c, we plot the mean tongue and lip images (center 1/5th of each vocalization) associated with each of the nine vowels in our data set from one speaker. These 'bases' clearly reveal that /ɑ/ is produced by a low-back tongue shape with an open lip configuration, /i/ is produced by a high-front tongue shape and a more narrow lip configuration, and /u/ is produced by a high-back tongue shape and a narrow lip configuration. This 'basis set' has the advantage (by definition) that each individual basis can be readily associated with a given vowel, making them easily interpretable. However, from a mathematical perspective, using the mean tongue and lip images as bases have several undesirable properties: (1) they are supervised, requiring the vowel labels to be known beforehand, (2) they reflect the full variability of the data set, and thus can be sensitive to individual trials (e.g. light gray traces in /i/), (3) many of the average images are quite similar, and therefore it is unlikely that these images correspond to a parsimonious description of the data. This last point can be formalized by simply measuring the coefficient of determination (R^2) between each image. The similarity matrices at the bottom of the Fig 2.4a plot the R^2 values of all pair-wise comparisons of images. Several of the tongue images (e.g. Fig 2.4a bottom: /ɑ/ vs. /ʌ/ and /i/ and /ɪ/), and most of the lip images (Fig 2.4c bottom) have high-degrees of similarity.

We therefore used unsupervised learning methods to extract structure from the images that capture the entire shape of the tongue and lips with a reduced number of bases. A common method for unsupervised learning of reduced basis sets is principal components analysis (PCA), which finds an orthogonal basis set that optimally captures the directions of highest variance in the data. However, a critique of PCA is that the bases often bear little resemblance to the data from which they were derived (Lee and Seung, 1999). Although this may be of little consequence if quantitative performance is the primary interest (as is often the case in machine learning), when understanding the bases is important (as is often the case in science), this lack of resemblance to data can hinder interpretability. Non-negative matrix factorization (NMF) has been used to extract 'meaningful' bases from data that consist of only positive values, such as images and movies (as in our data set) (Lee and Seung, 1999). NMF is a dimensionality reduction technique that extracts a predetermined number of bases (B) and weights (W) that linearly combine to reconstruct the data, under the constraint that both the bases and weights are strictly non-negative (Lee and Seung, 1999).

As our study primarily focuses on examining the steady-state configurations of the vocal tract during the production of vowels, we applied NMF to the lip and tongue images extracted from the center of the vowel (see Methods). The plots in the top of Fig 2.4b display the leading nine NMF bases derived for the tongue data, while Fig 2.4d displays the leading four NMF bases derived from the lip data.

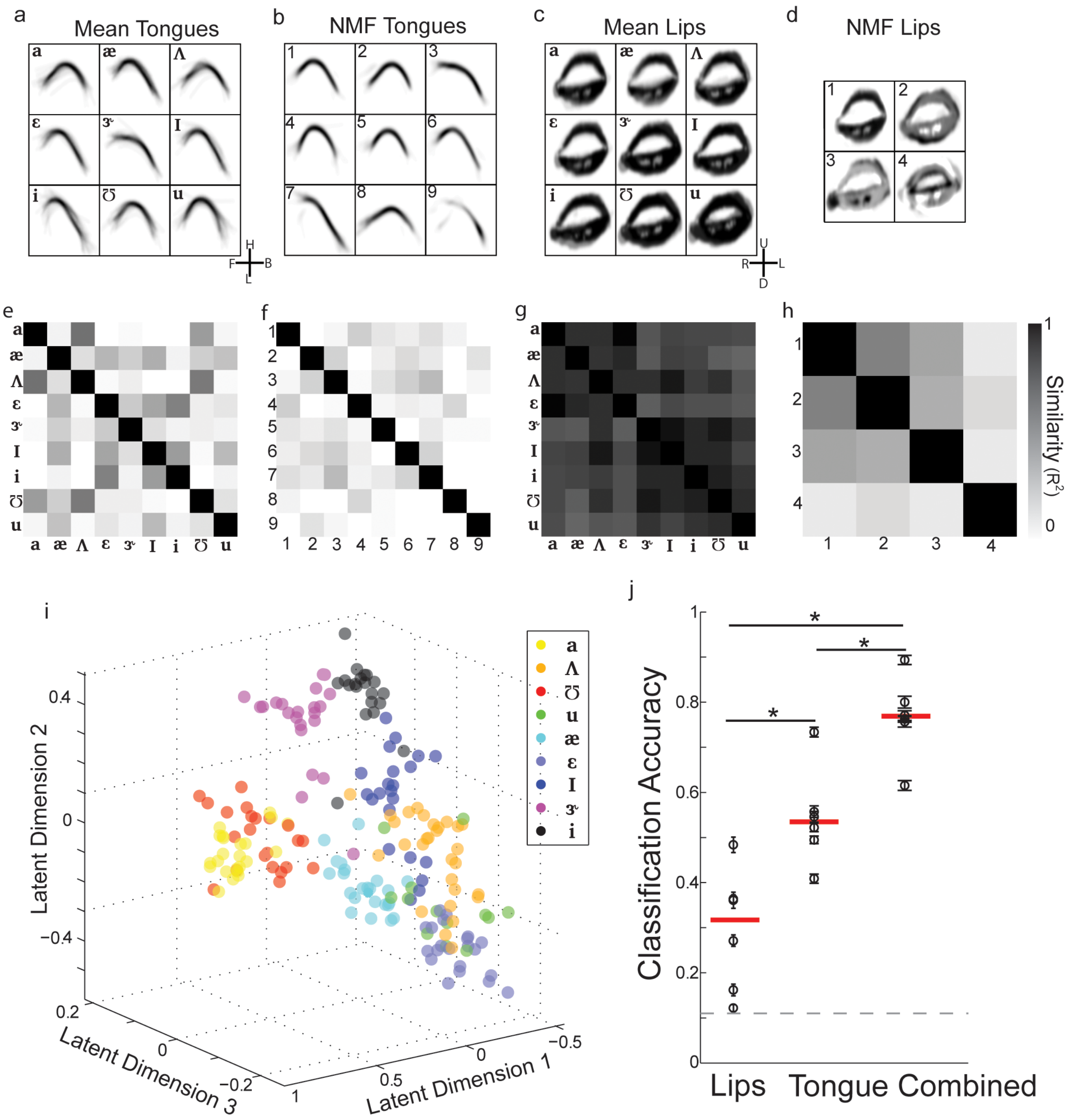


FIGURE 2.4: **Unsupervised extraction of vocal tract shape with non-negative matrix factorization improves vowel classification** ***(a)*** Mean tongue shape for each vowel from one subject. ***(b)*** Non-negative matrix bases blindly extracted from the tongue data for all vowels from one subject. ***(c)*** Mean lip shape for each vowel from one subject. ***(d)*** Non-negative matrix bases blindly extracted from the lip data for all vowels from one subject. ***(e-f)*** Similarity (R2) between mean tongue shapes (e), tongue non-negative matrix components (f), mean lip shapes (g), and lip non-negative matrix components (h). ***(i)*** Scatter plot of all vowels in the first 3 linear discriminant dimensions for one subject (same as Fig 2.3a-d). ***(j)*** Cross-validated classification accuracy of vowels from vocal tract shapes across all subjects. Naïve Bayes classifiers were trained to predict vowel identity based on NMF reconstruction weights for the lips and tongue individually, as well as from both lips and tongue. The combined model out-performs the individual models. Furthermore, the classification accuracy is enhanced relative to using the pre-defined articulatory features.

Focusing first on the tongue, we found that nine NMF bases could accurately and parsimoniously reconstruct the single utterance images (reconstruction error plateaued at nine NMFs). Furthermore, NMF extracted many bases shapes that could readily be associated with a particular vowel. For example, basis 1 in Fig 2.4b is very similar to the mean image for /ae/, while basis 3 resembles the mean image for /ɝ/. We note that, although this is an intuitive solution for the algorithm, it is not guaranteed mathematically. Additionally, not only were many of these bases interpretable, but they also appeared to contain less 'contamination' from single trials than mean images. Finally, the different NMF bases were, generally speaking, less similar to each other than the mean images, which is quantified for this subject by the similarity matrix at the bottom of Fig 2.4b.

In contrast to the tongue, we found that only four bases were needed to parsimoniously reconstruct the single trial images of the lips (reconstruction error plateaued at four NMFs). This likely reflects the fact that the tongue is the primary articulator responsible for shaping the vocal tract during vowel production, while the lip is a secondary articulator for vowels (e.g. Fig 2.3f,g). Furthermore, the contribution of these bases to the reconstruction of the different vowels is apparent. For example, basis 1 would contribute to vowels with large lip openings (e.g. /ɑ/), while the basis 4 likely contributed to vowels with more narrow lip openings (e.g. /u/). However, in general, the lip NMF bases did not have the same qualitative degree of one-to-one correspondence to the mean lip images. Instead, weighted combinations of several lip NMFs would likely have to contribute to the reconstruction of single images. Nonetheless, the different NMF bases for the lips were much less similar to each other than were the mean images for the vowels (Fig 2.4g,h). Across all 6 speakers examined here, similar results for number of bases and similarity of bases were found.

We examined if the NMF bases are useful for classifying vowels. To this end, we reconstructed each single utterance tongue and lip images as an optimal weighted combination of the NMF bases. This 13-dimensinsional reconstruction weight vector describes the contribution of a given bases to a specific utterance, and can be thought of as the 'representation' of that utterance in the NMF bases space. We then took the weight vectors for all utterances within a subject, and used linear discriminants analysis to find the three-dimensional latent space in which the vowels were most linearly separated (as in Fig

2.3), and projected the data for all vowels into this space. The plot in Fig 2.4i displays the organization of the vowels in this latent space for the same subject as emphasized in Fig 2.3. Visual comparison to the plot in Fig 2.3 suggests that the vowels could be more accurately assigned to distinct classes when using the NMF representation.

Similar results were observed across all subjects. Analogously to the analysis of acoustics and the parametric description of articulator position (Fig 2.3g), we trained a Naïve Bayes classifier to predict vowel identity based on the projection of the NMF reconstruction weights into the top three latent dimensions from LDA. This was done for the lips and tongue individually, as well as from combined lip and tongue data. The plot in Fig 2.4j shows the cross-validated classification accuracy of vowels from NMF bases across all subjects. As with the parametric description of the articulators (Fig 2.3g), the combined model out-performs the individual models (Fig. 2.4j, *: $P < 0.05$, WSRT, $N = 6$ for each comparison; median accuracies are 25%, 52%, and 77%). Furthermore, the average classification accuracy utilizing NMF bases was significantly greater than when using pre-defined points ($P < 0.05$, WSRT, $N = 6$). Therefore, NMF discovers bases that allow for more accurate classification of vowels than using a priori defined parametric descriptions of the articulator positions.

2.3.4 Continuous linear relationship between vowel acoustics and articulator position

In addition to the relationship between vowel category and articulatory or acoustic features, we wanted to assess how articulatory features and acoustics continuously map directly to one another. Understanding this relationship is critical because the degeneracies in the transformation between articulatory movements and resulting acoustics are what motivate our methods for explicitly monitoring the articulators. Utilizing regularized linear modeling (see Methods), we evaluated how well a given articulatory feature can be estimated by a linear combination of all acoustic features, and vice versa. An example model prediction is illustrated in Fig 2.5a–c. In the most direct relationship, pitch and glottal closure (Fig. 2.5a) are so closely correlated (R^2 = 0.99) that the link between articulation and acoustics is directly apparent. For the upper vocal tract, back tongue height (Fig. 2.5b) exhibits a more complex relationship with the acoustics, as a linear combination of all acoustic features is capable of modest prediction (R^2 = 0.36). Similarly F1 is fairly well predicted from all articulatory features (Fig. 2.5c, R^2 = 0.62).

We systematically performed this analysis across all articulatory and acoustic features for all subjects. Performance, as quantified by R^2, is plotted for each kinematic and acoustic feature, for each speaker, in Fig 2.5d-e (black: individual subjects; red line: median across speakers). The performance of these models illustrates that, on average, front, mid, and back tongue height are best predicted by the acoustics with R^2 all around 0.40. Lip opening and lip width show more modest values (opening R^2 = 0.20, width R^2 = 0.16). For speakers that had glottal recordings collected (N = 3), glottal closure is extremely well predicted by the acoustics (R^2 = 0.98). F1 and F2 are best predicted by articulator position (R^2 = 0.62 and 0.59) and F3 and F4 show moderate correlations (R^2 = 0.24 and 0.31). Additionally, F0 is predicted very well, but only for those subjects with glottal closure measurement. In general, the features that are well predicted by the linear models are the same features that are the primary contributors to vowel identity (as measured by LDA weights). Importantly, there is significant variability between speakers, suggesting that there is some variability in how each speaker is achieving roughly the same acoustic result. This cross-subject variability has been described before (Ladefoged and Johnson, 2011; Johnson, Ladefoged, and Lindau, 1993; Perkell et al., 1993; Borden and Gay, 1979), and reiterates the need to explicitly measure articulator movements rather than relying on canonical descriptions of the vocal tract during speech.

2.3.5 Statistical Synthesis of Speech from Articulator Positions

A central long-term goal of our work is to produce a speech prosthetic that transforms recorded brain signals into perceptually meaningful acoustics of speech. As speech production is mediated in the brain through control of the articulators, a first goal is to reconstruct intelligible speech from articulator measurements. This also provides further validation to the usability of the articulator measurements and preprocessing routines described. We use speech synthesis as a tool to evaluate a variety of increasingly rich descriptions of the vocal tract to find the optimal parameter space to generate intelligible and discriminable speech for the vowels considered. Using statistical parametric speech synthesis

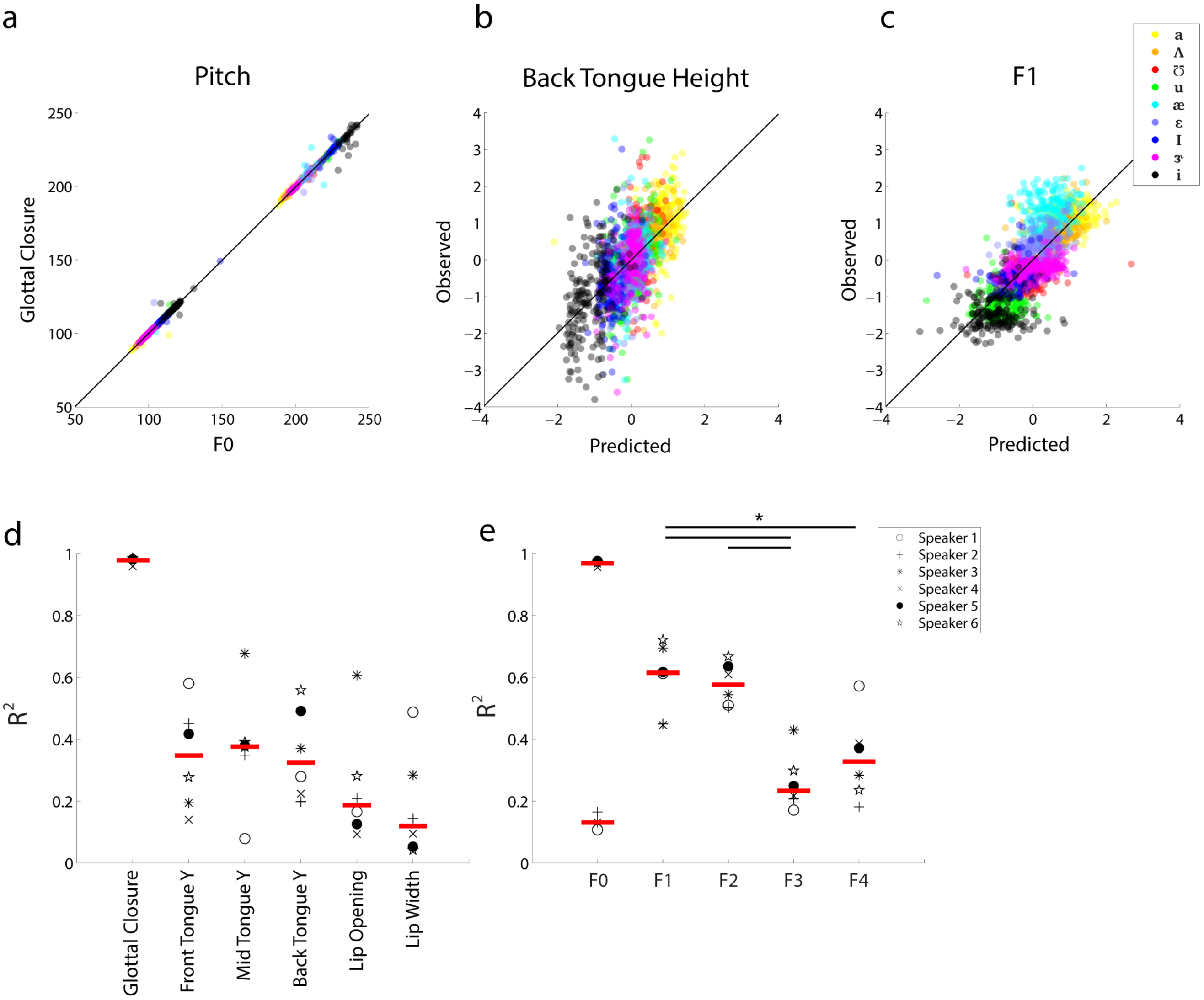


FIGURE 2.5: **Continuous linear relationship between vowel acoustics and articulator position** ***(a)*** Scatter plot of pitch values (F0) vs. frequency of glottal closures from 3 subjects (2 males, 1 female). Color corresponds to vowel identity. ***(b)*** Linear prediction of (z-scored) back tongue height from all acoustic features vs. the observed values for all nine vowels and six speakers. ***(c)*** Linear prediction of (z-scored) F1 from all kinematic features vs. the observed values for all nine vowels and six speakers. ***(d)*** Acoustic-to-articulator mappings. Amount of explained variance (R^2) for six kinematic features from all acoustic features for each subject. Subjects are identified by symbol. ***(e)*** Articulator-to-acoustic mappings. Amount of explained variance (R^2) for five acoustic features from all kinematic features for each subject.

(see Methods) several articulatory models are evaluated for articulator-to-acoustic conversion. The vocal tract parameterizations considered are i) tongue represented as sequence of equally spaced points (Tongue-based synthesis), ii) features of the lips (Lips-based synthesis), and iii) combined optimal parameterizations of tongue and lip features from (i) and (ii) (Combined model). As a visual illustration of synthesized speech, Fig 2.6a shows the spectrograms of speech synthesized from each articulatory model considered, shown against a prototypical spectrogram for the reference phoneme /ɑ/. It is interesting to note that both the lips and tongue-based models have visible errors in the spectrogram marked by physiologically impossible jumps in spectral energy. However, the comibined model does not make these errors, indicating that there is complimentary information between the lip and tongue measurements.

To objectively characterize the contributions of individual articulators, cross-validated Mel-Cepstral distortion (MCD) of the predicted speech features on an unseen test set of trials is computed across different models. Fig 2.6b shows the performance of models solely based on increasing number of points on the tongue. These trends suggest a dense representation using 10 points on the tongue to explain acoustic variability across subjects. Fig 2.6c compares the individual performances of the optimal tongue and lip based (two mid-sagittal and two coronal extremities of the lips, and derived lip width and lip height as predictors) models. Across subjects, lips and tongue articulators contribute complementary information as shown by superior performance of the combined model. All these comparisons are statistically significant ($P < 0.005$, WSRT, $N = 6$).

The ultimate test for speech synthesis is perceptual intelligibility by human listeners. For the case of vowel synthesis here, the right test is a perceptual judgment task to classify each synthesized stimulus into one of the nine possible vowel categories considered. We utilized crowdsourcing to conduct this subjective task (see Methods). 30 samples of unseen trials were synthesized and judged by human listeners on the Amazon Mechanical Turk. Participants were instructed to listen to each sample and identify which of nine vowels they heard. Fig 2.6d(i) summarizes the results of the perceptual tests as the confusion matrices of the perceived vs. true identities of vowel sounds as reported by listeners in the United States. The same result for listeners not restricted to just the United States is shown in Fig

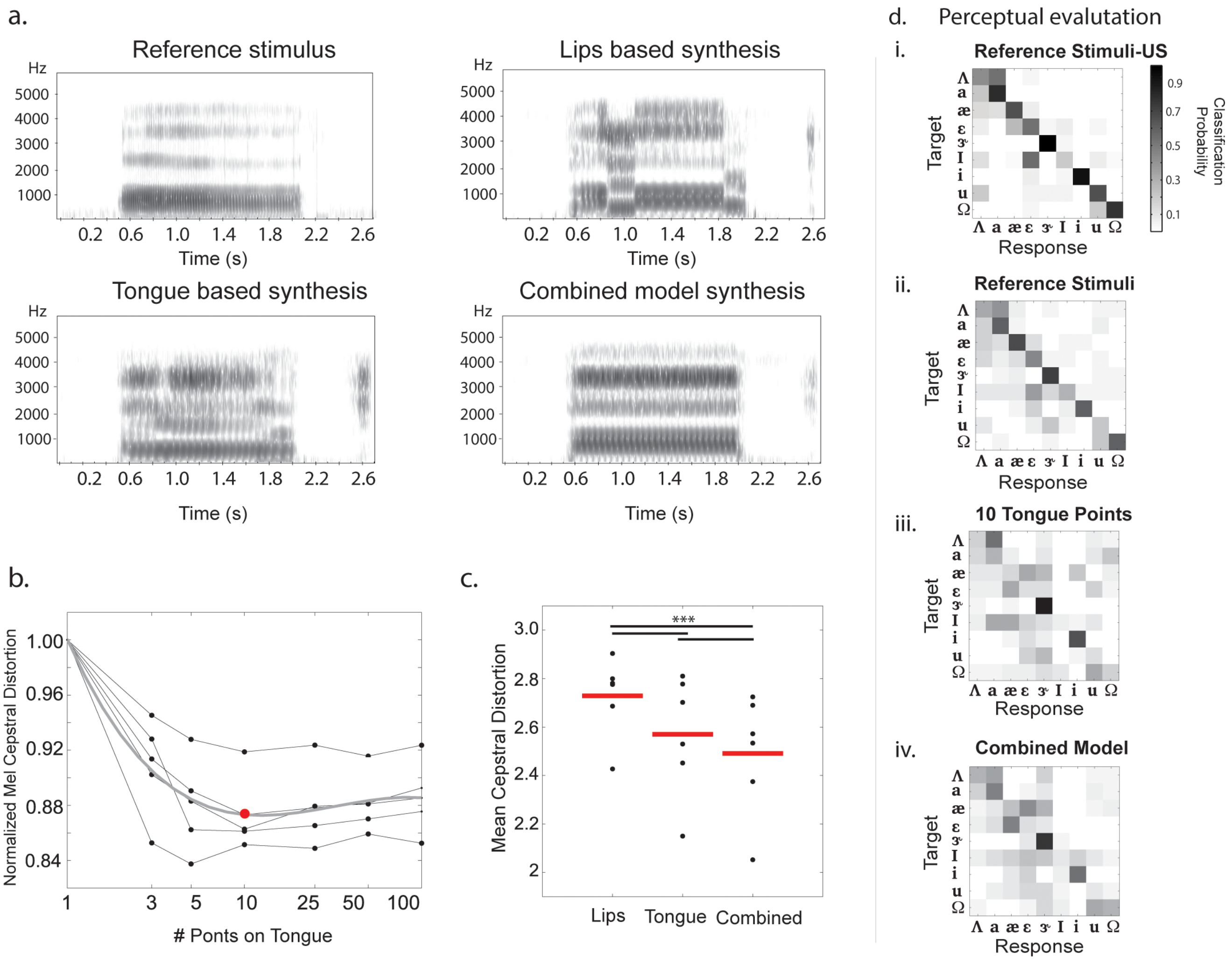


FIGURE 2.6: **Synthesis of speech from articulator kinematics *(a)*** Examples of reference and synthesized stimuli using various articulatory feature sets as identified for articulation of a phoneme /ɑ/ by one speaker. ***(b)*** Error on synthesized acoustics, measured as a mean cepstral distortion measure with increasing number of tongue points. Each subject is one black line, overlaid with a grey trend for average across subjects. Red dot at 10 points is selected as the optimal set of tongue points. ***(c)*** Prediction error with different sets of articulators used for synthesis, each subject is one dot and the mean across subjects is marked with the red line segment. ***(d)*** Reference natural stimuli as perceived by listeners based in the United States (i) and Turkers globally (ii) Stimuli synthesized using 10 tongue points (iii) and using both tongue and lip kinematics (iv).

2.6d(ii). While it is apparent that prior exposure to the target phonemes, as in the case of American listeners, improves the perceived accuracy, the assessment is still comparable to Fig 2.6d(i). Even the confusions made are along articulatory lines (e.g. confusions among the low front tongue vowels /ɛ/, and /ae/; confusions among short and long vowels /ɪ/ and /i/, /ʊ/ and /u/ respectively). Hence, subsequent tests are done with no restriction of selecting only American Turkers, since global listeners are still seem to perceive the acoustics (as shown in Fig. 2.6d(ii)) but form a less systematically biased and stricter listener population to assess the identity of these vowels.

We conducted two perceptual experiments on synthetic speech using 10 tongue points (Fig. 2.6d(iii)), and the combined model including 10 tongue points and the lips features (Fig. 2.6d(iv)). The classification accuracies of the synthesized speech are 31% and 36% respectively. It is interesting to note that perceptual classification of natural stimuli is 56% (Fig. 2.6d(ii)) accurate by Turkers around the globe, while the same number restricted to American Listeners is at 64%. It is evident that the overall performance increasingly matches that of natural stimuli with richer descriptions of articulators as anticipated. These also conform to the conclusions of the objective analysis reported in Figs 2.3g, 2.4j and 2.6c; best perceptual identification is obtained with the configuration using the combined tongue and lip features. These perceptual judgment results meet the aforementioned goal of intelligible and discriminable vowels synthesis across speakers, using only their articulatory trajectory information. Thus, from predictions of all articulators, we should be able to synthesize speech.

2.3.6 Decoding of lip aperture from ECoG recordings during production of words

We have described/validated a system for simultaneous monitoring of all speech articulators, and demonstrated that continuous linear models based on articulator measurements can be used to both predict acoustic features of vowels and to synthesize perceptually identifiable vowel sounds. One of our goals is to use this system to study the neural control of speech articulation by combining the articulatory tracking with simultaneously recorded neural signals from electrocorticography. This is a critical step towards developing a continuously controlled speech prosthetic.

To demonstrate the potential of combining articulatory tracking with ECoG recordings, we conducted a preliminary experiment in a neurosurgical patient with our face tracking system. We recorded the cortical field potential from ECoG electrodes placed directly over the ventral sensorimotor cortex (vSMC), an area of the human brain intimately involved in the control of speech articulation and orofacial movements (Bouchard and Chang, 2014; Bouchard et al., 2013; Mugler et al., 2014; Arce et al., 2013). Fig 2.7a plots a reconstruction of the electrode l ocations over vSMC in this subject (black dots are electrode locations). At each electrode, we extracted the time-varying high-gamma amplitude (70-150Hz), which likely reflects multi-unit firing rates (Ray and Maunsell, 2011). We extracted lip aperture from the face tracking system while the patient produced short words. For example, the red trace in Fig 2.7b plots lip aperture over time during a 35 second segment of the recordings. The lip contours from two vocalizations with different lip apertures are also plotted (Fig 2.7b, lip aperture is demarcated by the red vertical line in each image). We found that the moment-to-moment aperture of the lips could be well predicted from an optimal linear decoder of the vSMC high-gamma activity (R^2 = 0.55, Fig 2.7c). Although it is clear that much more can be done with these recordings, these preliminary results demonstrate the ability to successfully combine our articulator measurement system with ECoG recordings which will allow for studying the neural basis of speech production in unprecedented detail.

2.4 Discussion and Conclusions

We have developed a multi-modal system for simultaneously monitoring the lips, jaw, tongue, and larynx that is compatible with bedside human electrophysiology. To provide initial characterization and validation of our system, we collected and analyzed data from six speakers during the prolonged production vowels. We introduced methods to remove movement artifacts that are a consequence of the recording setting and validated these methods by classifying vowels using canonical descriptions of vowel production. We then applied unsupervised non-negative matrix factorization to derive novel parameterizations of articulator shape and show improved classification accuracy. We complement these categorical analyses by examining the continuous (linear) mappings between acoustics and articulations, and synthesized perceptually identifiable speech acoustics from articulator measurements.

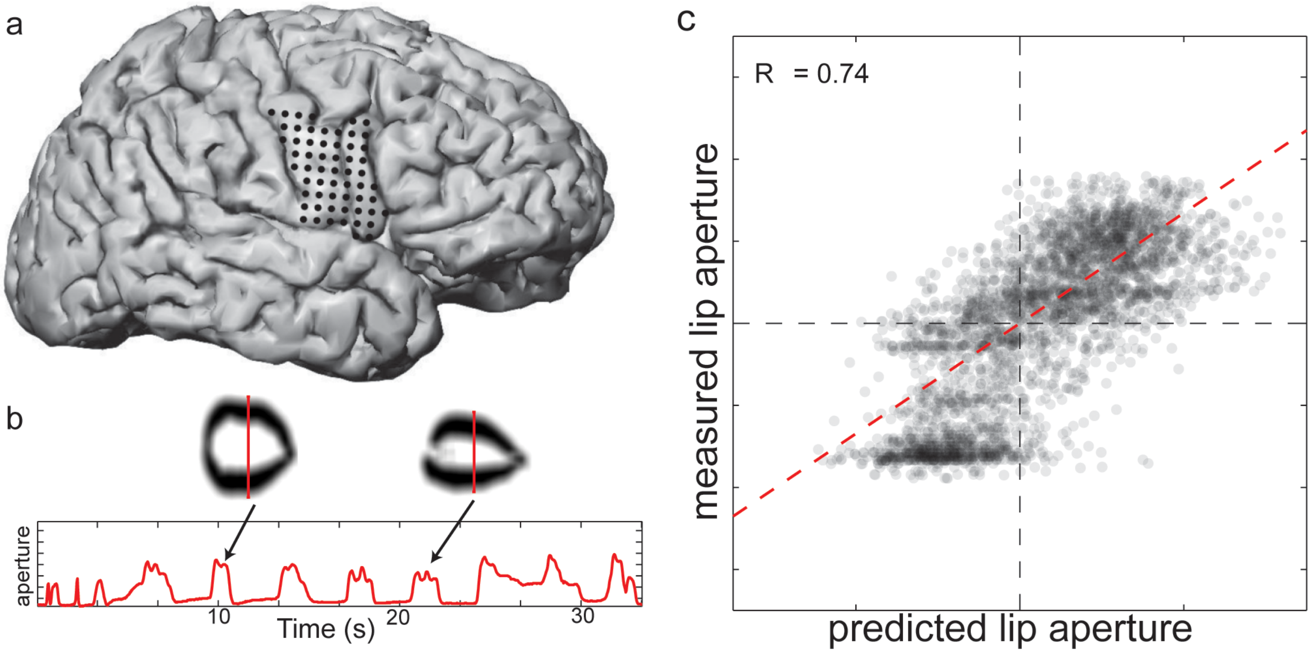


FIGURE 2.7: **Decoding of lip aperture from ECoG recordings during production of words** ***(a)*** Lateral view of the right hemisphere of a neurosurgical patient. The location of ECoG electrodes over the ventral sensorimotor cortex are demarcated with grey disks. ***(b)*** Example lip shape and vertical aperture during production of words in this subject. ***(c)*** Predicted lip aperture based on linear decoding of ECoG data vs. the actual aperture. Each dot is a time-point; red-dashed line is best linear fit.

Finally, we demonstrate the ability of this system to be used in conjunction with ECoG recordings by robustly decoding measured articulator kinematics from neural activity. Not only will this system allow for unprecedented insight into the neural basis of speech motor control, but also the methods outlined here are relevant for articulator monitoring in any setting where bulky and constrictive equipment is impractical.

2.4.1 System and methods for data collection and registration

Exploring the relationship between neural activity and measured speech articulator movements is critical to understanding speech motor control. The inability to unambiguously infer movements from acoustics means that we need to explicitly monitor the speech articulators simultaneously with neural activity and acoustics if we are to achieve a complete neurobiological understanding of speech motor control.

A variety of systems have been developed for the purpose of monitoring the speech articulators, but most are impractical for use in conjunction with bedside human electrophysiological recordings; they are either too invasive (e.g. photoglottography [50]), not portable (e.g. real-time MRI Narayanan et al., 2004), likely to create electrical artifact on the neural recordings (e.g electromagnetic midsagittal articulography, or some combination thereof. To address this, we developed a system to simultaneously monitor the lips and jaw with a video camera, the tongue with ultrasound, and the larynx with electroglottography. These methods provide high spatial and temporal resolution, and are fully compatible with beside ECoG recordings in a hospital setting. Another important consideration when recording in the clinic is that data can only be collected in short sessions, and it is not possible to ensure that the recording apparatus is positioned in the same way relative to the speaker in each session, or even across vocalizations within a session. We note that this is a common issue for all those who work with multi-channel, multi-site recordings of behavior. To address this general issue, we have developed two algorithms to register data across sessions and conditions. A critical assumption of our registration methods is that the position of the articulators several hundred milliseconds before acoustic onset can be used as a neutral reference for subsequent time points. Another assumption is that the registration transformations are restricted to scalings, translations and rotations (i.e. an affine transform).

We thoroughly validated this system and methods of registration by employing them on six subjects speaking nine vowels. We show that our system measures articulator movements that are consistent with previous studies of vowel production. First, the vowels organize according to kinematic and acoustic features as previously described (e.g. /i/ is characterized as a 'high front vowel'). Second, when relating articulator kinematics to produced acoustics, the classic features of tongue height and frontness prove to be the best predictors of vowel acoustics. Lastly, we consistently see variability between individual speakers. Together, these results validate our system and reiterate the need to explicitly measure articulation rather than relying on categorical descriptions of speech production.

It should be noted that our system does have some limitations. First, we are not able to image the entirety of the vocal tract, and therefore unable to describe the vocal tract as completely as some techniques allow. Furthermore we are imaging a two-dimensional sagittal plane of the tongue, and while this plane has been shown to be a good descriptor of tongue kinematics, it is known that the non-sagittal shape of the tongue has an important impact on the resulting acoustics. Lastly, the placement of the ultrasound transducer under the speaker's jaw slightly restricts natural jaw motion, which likely results in some degree of compensatory movement from the other articulators. These limitations are unavoidable given the clinical constraints of the recording setting, and are shared with most other dynamic articulator monitoring techniques. In this study we have chosen to primarily focus on the vowels for validation purposes. The articulations involved in vowel production correspond to a reduced sub-space of the total articulatory space of speech. Our methods should be extendable to the analysis to the entire English inventory, which is an important future direction. Furthermore, it will be important to derive time varying relationships between articulators and acoustics. This could be done using a combination of autoregressive models and canonical correlation analysis.

2.4.2 Non-negative matrix factorization extraction of vocal tract bases

The generation of even the simplest speech sounds requires the precise coordination of multiple articulators to achieve a vocal tract shape that produces the target sound. Indeed, it has been argued that speech motor control should be viewed in terms of dynamic spatial configurations of articulators [(Browman and Goldstein, 1990; Bouchard and Chang, 2014). Non-negative matrix factorization (NMF) has been used to extract 'meaningful' bases from data that consist of only positive values, such as images and movies, or recordings of electromyographic recordings (Lee and Seung, 1999). Recently, a variant of NMF (sparse convolutional NMF) was used to model real-time MRI data of human speech production for extraction of time-varying spatial configurations of the vocal tract (Ramanarayanan, Katsamanis, and Narayanan, 2011). Here, we applied NMF to the lip and tongue data during the center of the vowel, and found that the extracted bases had several desirable qualities: 1) several bases resembled the mean shapes associated with specific vowels (Fig 2.4), 2) the individual bases were less similar to each other than the means shapes of individual vowels (Fig 2.4), and 3) NMF bases could be used to improve classification performance over a priori defined point-based descriptions (Fig 2.4). Indeed, using the NMF bases, the accuracy for classification of vowel identity approached the accuracy based on the acoustics. Together, these results demonstrate the utility of NMF for data-driven extraction of vocal tract bases, and suggest it would be a useful approach for understanding other types of behavioral data.

Our utilization of NMF to extract purely spatial bases was motivated both to parallel the analysis of kinematic/acoustic features during the center of the vowel, but also to provide clear demonstration of the utility of this method to extract interpretable bases that reflect important vocal tract shapes. In this work, we extracted bases for the lips and tongue separately, and found that this resulted in readily identifiable bases for each articulator. We derived separate bases for lips and tongue because NMF attempts to explicitly reconstruct every point in the data, and the number of data points associated with lips was much larger than the number of data points associated with the tongue. Therefore, in a combined analysis, NMF would have 'weighted' the lips more heavily than the tongue, making it difficult to interpret the bases. Generally speaking, consideration of how the objective function of an algorithm interacts with the statistics of the data is critical for interpreting its outputs. Nonetheless, in combination with previous studies, our results strongly suggest that NMF will be a fruitful analytic approach for understanding speech production (Ramanarayanan, Katsamanis, and Narayanan, 2011). A critically important direction of future research is to use data driven descriptions of the vocal tract to understand the cortical control of speech through direct encoding/decoding analysis of simultaneously collected neural activity from multiple subjects.

2.4.3 Statistical speech synthesis from articulators to acoustics

The speech synthesis experiments reported in this work show that the processed articulatory trajectories retain sufficient information to synthesize audio that can be perceived as the intended vowel. Also noteworthy is the result that not all points on the tongue are necessary to accomplish this goal, suggesting that a high spatial resolution of tongue need not be tracked or estimated for intelligible synthesis. These findings are also shown to be consistent across subjects setting a valid precedent for synthesis of all possible phonemes using only articulatory data. At the implementation level, the advantage of the statistical model used for speech synthesis is that it is not constrained to a predefined geometrical or physiological model of the vocal tract (which may not be tractable), but instead models the salient relationships between articulatory and acoustic feature streams, as inferred from the data. Another advantage is that statistical models can be bootstrapped and adapted across speakers, potentially reducing the amount of data required to train the synthesizers (Yamagishi et al., 2009). It remains to be shown that this success can also be replicated on synthesizing consonants, where place of constriction (e.g. velar, palatal etc.) and the manner (affricate, plosive etc.) play additional roles along with the overall shape of the tongue and lips used here. Nonetheless, our results imply that decoding trajectories of the critical articulators from neural activity can be sufficient to produce speech understandable by most listeners (although whether this can be expanded from vowels to all of speech remains to be seen).

2.4.4 Decoding of speech kinematics

Here, we provide the first direct demonstration that the moment-to-moment kinematics of a vocal tract feature can be well predicted from a statistical mapping of the vSMC high-gamma activity (Fig 2.7). Brain-machine interface approaches to speech prosthetics hold promise to dramatically improve the communication abilities of the profoundly disabled. Together with previous studies, our results strongly suggest that combining continuous statistical mappings of vSMC activity to articulator kinematics with mappings of kinematics to acoustics is likely to be a successful strategy for a brain-machine interface for a speech prosthetic. A continuous transformation approach could be combined with a categorical decoder for a hybrid prosthetic system. The advantages of such a hybrid system would be the capacity to simultaneously model the continuous transformation of vSMC activity into speech (which is the ‘natural’ transformation), while capitalizing upon existing massive data sets to incorporate a language model for classification (as has been powerfully applied to automated speech recognition systems). Future studies combining ECoG recordings with simultaneous measurement of multiple vocal tract articulators, as permitted by our system, would allow unraveling the cortical coordination underlying multi-articulator control for speech production. Understanding whether neural representations in different brain areas underlying speech production are acoustic, articulatory, or phonetic is a central challenge in developing a cortical theory of speech production. The ability to measure speech behavior on the acoustic, articulatory and phonetic basis with the simultaneously collected high spatio-temporal resolution cortical activity on an individual subject and single-utterance level may provide insight into this issue.

Chapter 3

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

3.1 Introduction

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

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

The precise movements (kinematics) of the articulators are challenging to measure because many of the vocal tract movements are internal to the mouth and throat, and therefore difficult to monitor externally, especially in the context of neural recordings. As a result, previous studies have used the produced acoustics to infer which articulators are involved, based on extensive linguistic descriptions of speech movements for a given speech sound (Lotze et al., 2000; Crone et al., 1998a; Brown et al., 2009; Fukuda et al., 2010; Kellis et al., 2010; Pei et al., 2011; Leuthardt et al., 2011; Grabski et al., 2012; Bouchard et al., 2013; Bouchard and Chang, 2014; Mugler et al., 2014; Herff et al., 2015). Although it is possible to describe the movements of each articulator according to phonetic labels derived from the acoustics, these behavioral descriptions cannot provide exact characterizations of the changing positions of the articulators over time. Moreover, there are many articulator configurations that can result in the same acoustics (Atal et al., 1978; Maeda, 1990; Gracco and Lofqvist, 1994) and considerable across-speaker (Johnson, Ladefoged, and Lindau, 1993) and across-trial (Perkell and Nelson, 1985) variability in movements that give rise to a particular speech sound. Thus, understanding how the brain produces complex sounds like vowels requires determining how different kinematic parameters of articulatory movements are controlled in vSMC during speech production.

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

3.2 Methods

3.2.1 Electrocorticography Acquisition and Signal Processing

Four human participants underwent chronic implantation of a high-density subdural electrocorticographic array (ECoG) as part of the clinical treatment of epilepsy (3 female right hemisphere, one male left hemisphere). All subjects were implanted with 256-channel grids over peri-Sylvian cortex (1.17mm diameter electrodes, 4mm pitch, 60x60mm coverage; Integra [Plainsboro NJ, USA]), referenced to scalp electrode. The total number of vSMC electrodes for individual subjects ranged from 52 to 86 for a total of 270. Cortical-surface electrical potentials were recorded with ECoG arrays and the voltage time series from each electrode was inspected for artifacts or excessive noise. Electrodes with excessive noise and time periods with artifacts were excluded from analysis, and the raw ECoG activity was re-referenced to the common average. For each channel, the time-varying analytic amplitude of the voltage signal in the high-gamma (HG) range (70-150 Hz) was extracted using the Hilbert transform, according to previously published procedures (Edwards et al., 2010). HG correlates well with multi-unit firing(Ray and Maunsell, 2011), and has high spatial and temporal resolution Muller et al., 2016. The HG signal was down-sampled to 400 Hz for analysis and plotting purposes. HG power was z-scored relative to activity recorded during periods of silence during the same recording session. All analyses were limited to the ventral sensory-motor cortex (vSMC), which was anatomically defined as the ventral portions of the pre-central and post-central gyri, as well as the sub-central gyrus.

3.2.2 Task

Participants listened to audio recordings of nine English vowels (/ɑ/ae/ʌ/ɛ/ɝ/ɪ/i/ʊ/u) and were instructed to repeat each vowel. On each trial, to ensure they properly identified the vowel, they first heard it in an /h-V-d/ context (e.g. ‘hood’), and then they heard the vowel in isolation. After a 1-1.5 sec delay, participants were presented with a visual cue to produce the isolated vowel.

They were not explicitly instructed to hold the vowel for a specific amount of time. The median duration of production was 1.66 seconds (STD = 0.35 s). For each participant, between 15 and 30 repetitions of each vowel were collected over the course of 3-6 recording sessions.

3.2.3 Articulator Tracking

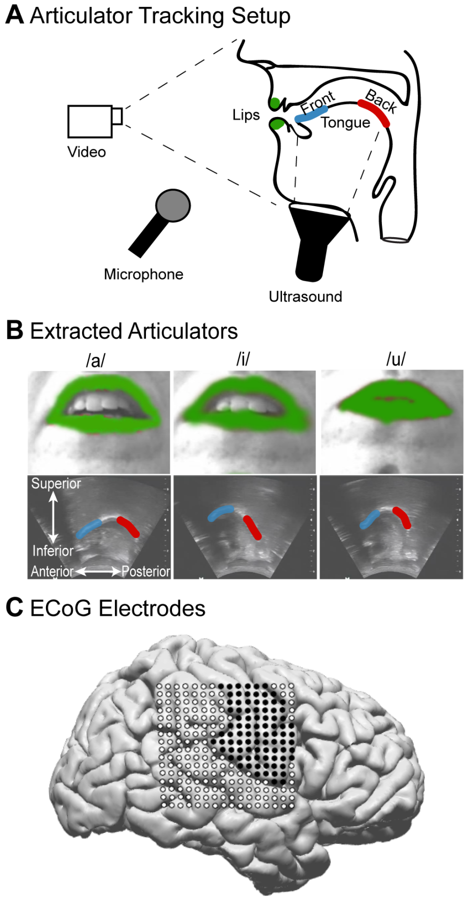


FIGURE 3.1: **Experimental setup and articulator monitoring *(a)*** Schematic of the articulatory tracking system. A video camera placed in front of the subject recorded the movements of the lips while an ultrasound transducer under the jaw captured the tongue contour. ***(b)*** Example images of the from the video (top) and ultrasound (bottom) imaging during production of the corner vowels /a/, /i/, and /u/. The lips and tongue contour were extracted from these images, and the resulting binary masks are shown in color on top of the raw images. ***(c)*** Magnetic resonance imaging (MRI) reconstruction of the brains of the four subjects included in the study. Co-registered ECoG electrodes are plotted on the cortical surface, with dark points denoting electrodes over vSMC.

We developed a system to record the movements of the main supra-laryngeal articulators while participants performed the vowel production task (Figure 3.1A), the details of which have been described previously (Bouchard et al., 2016). Briefly, to capture the movement of the lips and jaw, a camera was placed in front of the participant's mouth. The participant's lips were painted blue, and red dots were painted on the tip of the nose and the chin to simplify the process of extracting the shape and position of these articulators. The camera captured video at 30 frames per second. To image the tongue, an ultrasound transducer was held firmly under the participant's chin with the plane-of-view capturing the midline of the tongue. The ultrasound recorded images at 30 frames per second, and the data were aligned to the lips/jaw video according to the peak of the cross-correlation of the audio signals from each video. Using hue thresholding, we extracted the lips and jaw automatically from these videos as binary masks (Figure 3.1B). From these binary masks, we extracted the locations of the four corners of the mouth (upper/lower lip, left/right corners) and the jaw. For the tongue, we used EdgeTrak to extract 100 points of the mid-sagittal contour, which were then down sampled to 3 points by taking the median x and y value for the front, middle, and back thirds of the contour (Li, Kambhamettu, and Stone, 2004). Since video and ultrasound were collected in orthogonal spatial planes, x and y positions in the lips/jaw images reflect left/right and top/bottom, whereas x and y positions in the tongue images reflect front/back and top/bottom. To correct for differences in the relative position of the camera and ultrasound transducer with respect to the participant, we referenced each articulatory point to the neutral starting position at the beginning of each trial. From the measured position of each articulatory feature (X) we also derived movement parameters including velocity (X'), speed ($|X'|$), and acceleration (X'') of that articulator. We refer to these parameters collectively as the articulator kinematics. While the lips and jaw were both included in all analyses, we found that lip opening and jaw height were correlated for this vowel production task (cross-subject average correlation: r = 0.73 ± 0.12). Therefore, to simplify visualizations we only show results for the lips.

3.2.4 Acoustic Feature Extraction

Speech sounds were recorded using a Sennheiser microphone placed in front of the participant's mouth and recorded at 22 kHz (Figure 3.1A). The speech signal was transcribed offline using Praat (Boersma, 2001). For each vowel, we extracted the formants (F1-F4) using an inverse filter method (Ueda et al., 2007; Watanabe, 2001; Bouchard et al., 2016).

3.2.5 Trial Duration Standardization

To standardize the durations of the vowels across trials and participants, we linearly resampled each trial to be the median duration across vowels and subjects (1.66 seconds). Behavior and neural signals changed with rapid and stereotyped dynamics around onset and offset; resampling the entire trial would systematically change those dynamics based on vowel duration. Therefore, to preserve onset and offset dynamics, we only resampled data in the time window from 250ms after the onset of the acoustics to 250ms before the offset: corresponding to the steady-state hold. Trials with durations less than half or greater than twice the median were excluded from analysis (26 in total across all subjects). Final analyses utilized an average of 15.3± 5.7 trials per vowel per subject.

3.2.6 Permutation Tests

To evaluate statistical significance in each analysis, we used permutation tests. A permutation distribution for a given model of neural representation was constructed by randomly permuting the trial labels of the observed data, and then training and testing the model using this shuffled data. This process was repeated 500 times, and the performance of these shuffled models comprised the permutation distribution. A model was considered significant if its performance on test data was greater than the 99th percentile of its corresponding permutation distribution. For the correlations in Figure 3.2D, we test if $|r| >$ 99th percentile of $|r_{null}|$.

3.2.7 Correlations with Articulatory Position

To evaluate the relationship between vSMC HG activity and individual articulators, we correlated HG activity at individual electrodes with the measured trial-to-trial position for each articulatory feature. HG activity averaged over a 200ms window centered at acoustic onset was correlated with the mean position of each articulator taken from a 200ms window centered at the midpoint of the vowel. Electrodes were labeled according to whether they had significant correlations with zero, one, or multiple articulator positions. Electrodes in 2D are examples of electrodes with significant correlations with only one articulator.

3.2.8 Encoding of Kinematics, Formants, and Vowel Categories

We compared the representation of articulator kinematics, vowel formants, and vowel category at each electrode using L1-regularized linear regression (Lasso). These models predict HG activity at each time point 500ms before acoustic onset to 500ms after acoustic offset from a sparse linear combination the behavior:

$$HG_e = \Sigma_{i=1}^{n} \beta_i X_i \tag{3.1}$$

Where HG_e is the HG power at a given electrode, X are the feature sets (vowel, formant, kinematics, or joint), β are the linear weights that describe the mapping, and i is a vowel category (n=9), vowel formant feature (n=10), articulator kinematic feature (n=40), or all feature sets jointly (n=59). The formant features were F1-F4, as well as all pairwise ratios of F1-F4. The articulator kinematic parameters were position, speed, velocity, and acceleration for lip opening, lip width, jaw height, and the front, middle, and back tongue. Vowel identity was parameterized as 9 binary vectors corresponding to the vowel being produced during vocalization. Formant, articulatory, and vowel identity features were lagged +100ms relative to HG, corresponding to the causal direction of neural activity generating behavior. This lag was determined empirically by optimizing model performance over a range of lag values (-500ms to +500ms). To train and test linear models, we used L1-regularized linear regression in a leave-one-trial-out cross-validation procedure. We calculated the correlation between the observed and predicted HG values, averaged across cross-validations. Electrodes were included in visualizations and summary statistics only if their performance passed the permutation test described above for at least one of these models (i.e., formants, kinematics, vowel identity, or combined). To compare models with different numbers of parameters, we calculated the adjusted R^2:

$$R^2_{adj} = R^2 - (1 - R^2)\frac{p}{n - p - 1} \tag{3.2}$$

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

3.2.9 Organization of vowels in behavioral and neural spaces

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

3.2.10 Encoding of kinematic parameters across time

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

$$\hat{y} = \beta X \tag{3.3}$$

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

$$y_{idp} = y - \hat{y} \tag{3.4}$$

We then used L1-regularized linear encoding models to predict HG activity from the kinematic parameters (position, speed, velocity, and acceleration) of the lips, jaw, and tongue. However, instead of including the entire trial time-course in each model, we trained and tested models within 100ms non-overlapping windows that tiled the trial. Articulator kinematics were lagged +100ms relative to HG to evaluate the causal nature of neural activity on behavior. Models were trained and tested independently for each time window. Performance was measured by the correlation between the observed and predicted HG values, averaged across cross-validations. Electrodes were included in visualizations and summary statistics only if their performance passed the permutation test described above for at least 3 contiguous time windows at any point in the trial.

3.2.11 Description of vSMC HG dynamics

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

$$A \approx BW^T; B, W \geq 0 \tag{3.5}$$

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

$$\hat{B}, \hat{W} = \min_{B,W} \frac{1}{2} ||A - BW^T|| \frac{2}{F}; s.t. B, W \geq 0 \tag{3.6}$$

NMF is particularly useful for decomposing data into 'parts' that have interpretable meanings (e.g., transient vs. sustained response types) (Lee and Seung, 1999; Donoho and Stodden, 2003; Bouchard et al., 2016). The HG activity for each vSMC electrode across all participants was averaged across trials, offset by the minimum value (such that all values were positive), and NMF was applied to the matrix of time courses x electrodes. To determine a parsimonious number of bases, we calculated the reconstruction error when projecting the data onto the NMF bases:

$$err = \frac{1}{2}||A - BW^T||\frac{2}{F} \tag{3.7}$$

We then found the number of bases (i.e., rank k) beyond which reconstruction error only marginally reduced (i.e., the elbow of the curve): five bases were used. The first two bases resembled the transient and sustained activity observed in Figure 3.7A. Electrodes with sustained activity were defined as those that had weighting for basis 1 greater than for basis 2. The width (HG_w) of the HG activity for sustained electrodes was derived as follows:

$$HG_w = argmin \int (HG_{e,t} - \overline{HG_{e,t}}) - argmax \int (HG_{e,t} - \overline{HG_{e,t}}) \tag{3.8}$$

Where $HG_{e,t}$ is the HG activity at given sustained electrode for a given trial. This measure was calculated for each sustained vSMC electrode, for each trial. We assessed spatial organization by measuring the Euclidean distance between electrodes organized according to their maximum NMF weight (i.e., transient or sustained). We compared distributions of intra-parameter distances and cross-parameter distances to randomized distributions derived by shuffling the labeling of the electrodes. If the HG dynamic variability across vSMC is spatially organized, the distribution of intra-parameter and cross-parameter distances should differ from the distributions of the random distributions.

3.3 Results

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

3.3.1 Articulator tracking during vowel production

We simultaneously tracked the movements (Figure 3.1A) of the major supralaryngeal articulators (i.e., lips, jaw, and tongue; Figure 3.1B) while recording neural activity directly from the cortical surface (Figure 3.1C; see Methods, Chapter 2, Bouchard et al., 2016). We first verified that by extracting the positions of the articulators, we observed characteristic vocal tract configurations that reflect distinct vowels. For example, the vowel /ɑ/ is characterized by lowering the front tongue, raising the back tongue, and opening the lips, while the vowels /i/ and /u/ have different configurations (Figure 3.2A). The measured articulatory movements captured these characteristics, and clearly discriminated vowel categories (Figure 3.1B, 3.2B). We also used the produced acoustics as a behavioral measure of vowel discriminability. By extracting the formants from the acoustic signal, we observed distinct relative patterns of acoustic power for different vowels. For example, /ɑ/ is characterized by high F1 and low F2, whereas /i/ and /u/ have different formant profiles (Figure 3.2C).

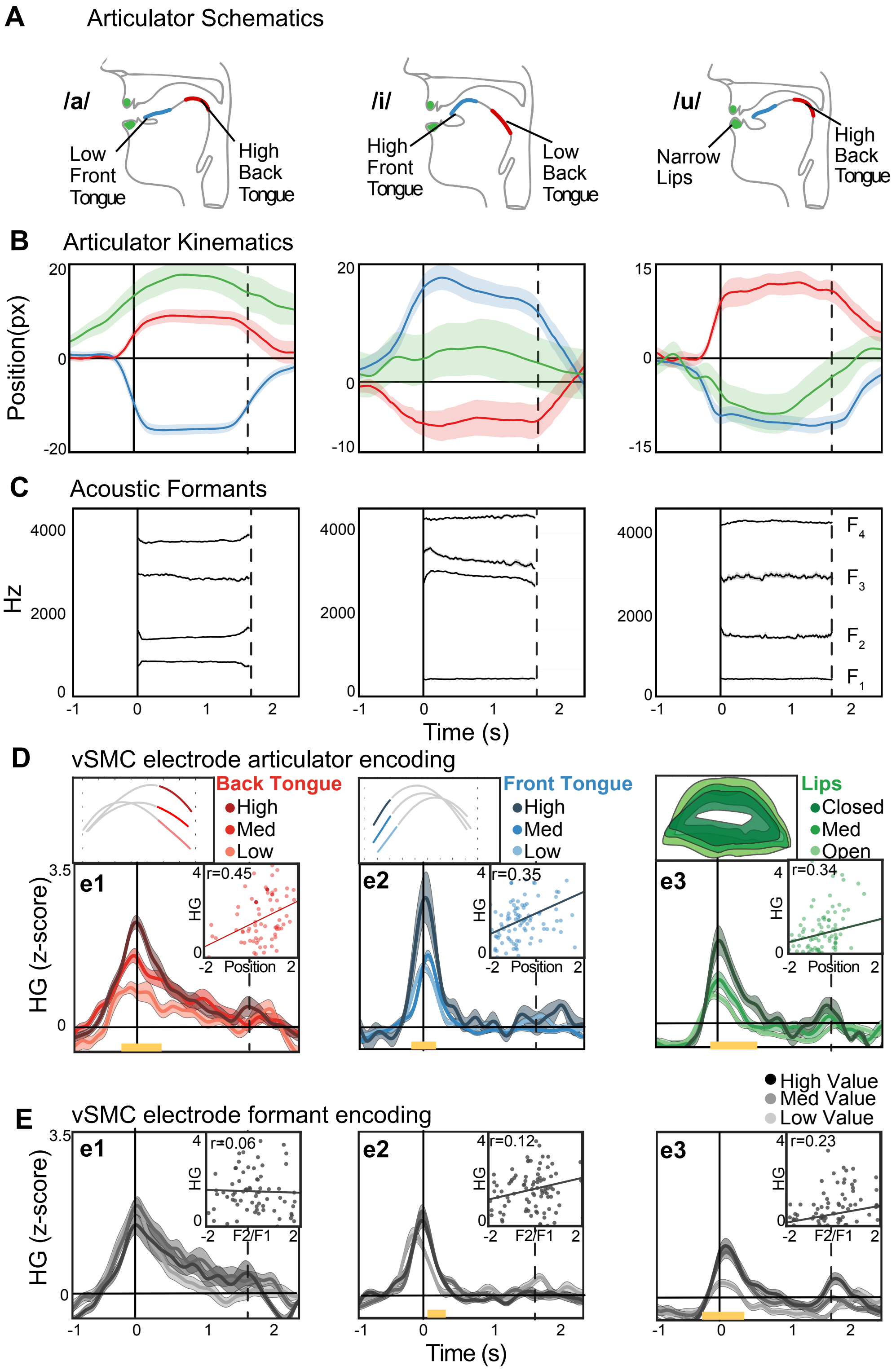

A
Articulator Schematics
/a/
Low
Front
Tongue
High
Back
Tongue
/i/
High
Front
Tongue
Low
Back
Tongue
/u/
Narrow
Lips
High
Back
Tongue
B
Articulator Kinematics
Position(px)
20
0
-20
20
0
-10
15
0
-15
C
Acoustic Formants
Hz
4000
2000
0
-1
0
1
2
Time (s)
F_4
F_3
F_2
F_1
D
vSMC electrode articulator encoding
Back Tongue
High
Med
Low
Front Tongue
High
Med
Low
Lips
Closed
Med
Open
HG (z-score)
3.5
0
e1
r=0.45
HG
4
0
-2
Position
2
e2
r=0.35
e3
r=0.34
High Value
Med Value
Low Value
E
vSMC electrode formant encoding
e1
r=0.06
F2/F1
e2
r=0.12
e3
r=0.23
Time (s)

FIGURE 3.2: **Articulatory and acoustic behavior correlates with single electrode vSMC neural activity** ***(a)*** Prototypical articulator positions for the corner vowels /a/, /i/, and /u/. ***(b)*** Average (± s.e.m.) timecourses of measured articulator displacements during production of the corner vowels. For illustration, the kinematic parameter of vertical position is shown, however both vertical and horizontal measurements are used for subsequent analyses..***(c)*** Average formant values (F1-F4) for the corner vowels during the same productions as in (b). ***(d)*** HG activity in three example vSMC electrodes. Each electrode was selected to be representative for the articulator shown in the top subplot (median configurations shown for back tongue, front tongue, and lips). Trials for each electrode are binned by the displacement of the articulator which best correlates with the HG values at acoustic onset (± 100 ms). Yellow bars mark timepoints of significant difference between the bins (e1: $F(2,75) < 5.1$, $p<0.01$; e2: $F(2,102) < 4.9$, $p<0.01$; e3: $F(2,105) < 4.8$, $p<0.01$; ANOVA). ***(e)*** HG activity in the same electrodes binned according to formant values (F1/F2 ratio).

3.3.2 Representation of articulator kinematics in ventral sensory-motor cortex

These descriptions demonstrate that both articulator kinematics and acoustic formants provide rich descriptions of the same behavior. However, although kinematics and acoustics are causally related, their relationship is not one-to-one (Atal et al., 1978; Maeda, 1990; Gracco and Lofqvist, 1994), nor are they perfectly correlated (in the present data set, $r_{kin-acoust}$=0.53 ± 0.17). For example, producing the vowel /u/ ("hoot") involves raising the back of the tongue towards the soft palate while rounding the lips. However, those movements can be compensatory. The vowel /u/ can be produced with less pronounced lip movements accompanied by greater tongue movements, or vice-versa (Perkell et al., 1993). Therefore, we asked whether articulator kinematics or acoustic formants are the behavioral characterization of vowels represented in vSMC. First, we quantified how well the positions of speech articulators or vowel formants explain the variance of HG at individual vSMC electrodes (i.e., encoding strength). We recorded cortical electrical potentials from a total of 270 electrodes from the surface of vSMC across 4 subjects (Figure 3.1C). The high-gamma (HG) activity at many vSMC electrodes was elevated above baseline during the speech movements, and was significantly correlated with the trial-to-trial position of the speech articulators (Figure 3.2D). We observed a clear relationship between articulator position and HG activity. For illustration, we identified representative electrodes where activity was most correlated with a single articulator. For example, the HG activity of electrode 1 at the time of vowel onset was significantly correlated with only the back tongue. Likewise, electrode 2 showed greater activity for higher front tongue positions. Electrode 3 was correlated with the opening of the lips. To examine whether HG activity at these electrodes was similarly correlated with the produced acoustics, we binned the activity by formant values (Figure 3.2E). We observed weaker correlations with formants compared to articulator position, demonstrating more robust encoding of articulatory representations.

We were specifically interested in whether vSMC activity is best explained by articulator kinematics, vowel formants, or vowel identity. We used linear encoding models to predict neural activity from kinematic or acoustic features, or the vowel identity (see Methods). Across electrodes, we found that articulator kinematics provided significantly better model fits compared to vowel formants (U = 5.3, p = 1.0e-8; Wilcoxon rank sum), or vowel identity (U = 8.7, p = 4.1e-18; Wilcoxon rank sum) (Figure 3.3A). We used nested models to examine how much additional neural variance is explained by predicting HG from both articulator kinematics and vowel formants. We found that the joint model explained no more variance than the articulator kinematics alone (U = 0.4, p = 0.7; Wilcoxon rank sum) suggesting that the performance of the formant models was likely driven by variance shared with the kinematics. Therefore, we find no evidence for encoding of vowel formants separate from their articulatory origin. Furthermore, these results demonstrate that the production of distinct vowels is grounded in direct control of articulator kinematics.

Across all vSMC electrodes, we found that 49% (133 out of 270) were significantly correlated with movements of one or more articulators (correlations > 99th percentile of permutation distributions). We observed a clear spatial organization to the articulator correlations, with lips/jaw more dorsal than the tongue (Figure 3.3B), consistent with previously-described somatotopy (Penfield and Roberts, 1959; Brown et al., 2005; Bouchard et al., 2013; Huang et al., 2013; Conant, Bouchard, and Chang, 2014). Within the ventral region, we observed electrodes that more strongly reflected either the front or back of the tongue. Both front and back tongue electrodes were distributed throughout the broader tongue region. Finally, we observed 72 electrodes that had significant correlations with multiple articulators, which were distributed throughout vSMC. Together, these results extend our understanding of speech-motor cortex somatotopy by demonstrating that the dominant encoding scheme in these neural populations reflects the specific movements of the preferred articulators.

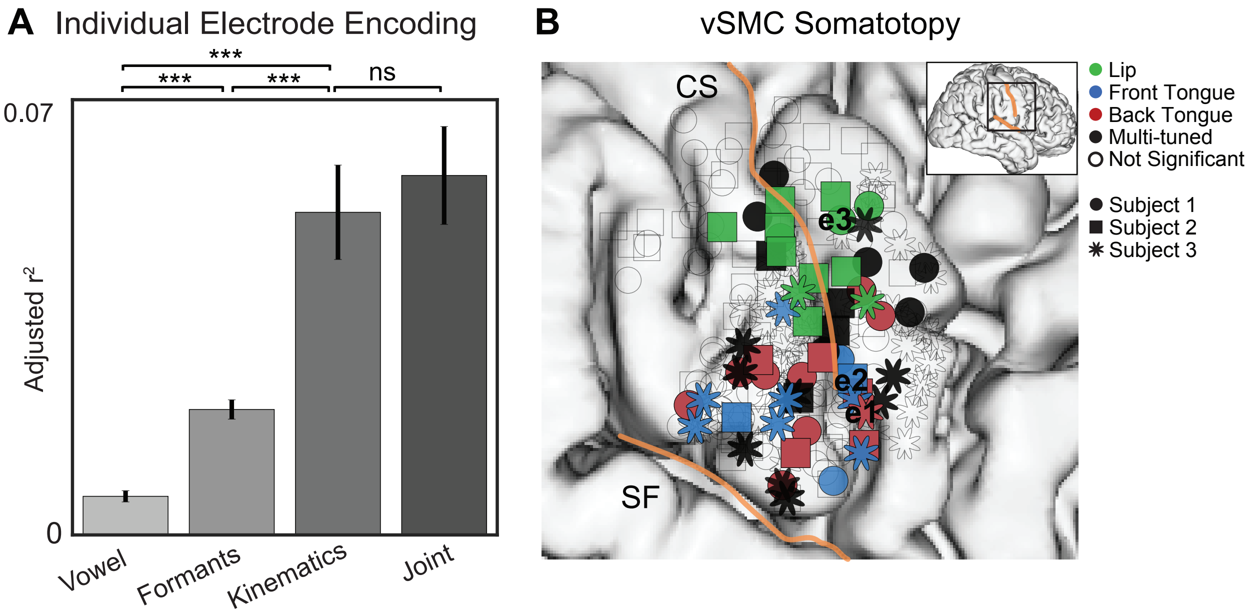


FIGURE 3.3: **vSMC activity primarily encodes speech articulators** ***(a)*** Performance of encoding models predicting vSMC HG using vowel identify, acoustic formants, articulator kinematics, or all three. Articulator kinematics explain vSMC activity better than vowel identity and acoustic formants (*** p<0.01; Wilcoxon rank sum). Furthermore, the joint model does not explain more variance than the kinematic model alone, indicating that the vowel identity and acoustic formant models are likely driven by variance shared with the kinematics.***(b)*** Electrodes over vSMC from three right hemisphere subjects were warped onto a common brain and color-coded according to articulator selectivity. Empty circles mark electrodes with no significant selectivity for any articulator; black electrodes are selective for more than one articulator, and blue, red, and green electrodes are selective for front tongue, back tongue, or lips, respectively.

3.3.3 Organization of vowels in vSMC population activity

To understand how vSMC encoding of articulator kinematics contributes to our ability to produce distinct vowels, we examined the organization of behavioral and neural activity in relation to all nine vowels. In addition to articulator kinematic representations at individual electrodes, population activity in vSMC may reflect the coordinated movements of the vocal tract that produce vowels. Furthermore, because vowel formants arise from the relative positions of multiple articulators, it may be the case that while articulators are most strongly represented at single electrodes, population activity may reflect a different, emergent representation. We examined the organization of speech representations at the population level by comparing the relative distances of vowels in acoustic, articulatory, and neural space. We performed multi-dimensional scaling (MDS) on the vowel centroids (see Methods) measured by vowel formants, articulator position, or vSMC neural activity across all participants. In this analysis, vowel tokens that are near each other in MDS space have similar formant, kinematic, or HG values. Consistent with previous behavioral and linguistic descriptions of vowel production, the formant and kinematic MDS projections replicate the classic vowel space 'trapezoid' (Figure 3.4A-C) (Hillenbrand et al., 1995; Ladefoged and Johnson, 2011). The HG neural MDS projection also closely resembled the acoustic/kinematic organization of the vowels. For example, the vowel /ɑ/ (as in hall) is near the vowel /ʌ/ (as in hut), but far from /i/ (as in heat).

To characterize the difference in organization of vowels across these feature spaces, we calculated the pairwise distances between the vowels in MDS space, visualized as confusion matrices (Figure 3.4D-F, right). We additionally performed hierarchical clustering of the pairwise distances and organized the confusion matrices by the derived hierarchical organization. The pairwise distances and hierarchical clustering reaffirm the classic vowel organization, but also highlight the specific differences between the feature spaces. For example, /i/ is distant from the other vowels in the formant space, but closer in the articulator and neural spaces. We found that the organization of vowels in vSMC HG activity is significantly more correlated with the organization of vowels in the articulator space compared to the acoustic space (U = 9,5, p=1.3e-21, Wilcoxon rank sum), although both representations were significantly correlated with vowel organization in the HG neural space (acoustic: r=0.56, p=$2.8e^{-4}$, kinematic: r=0.73, p=$5.9e^{-5}$).

3.3.4 Encoding of articulator kinematic parameters

In the above analyses, we considered the joint encoding of multiple kinematic parameters for individual articulators. However, it is unknown whether kinematic encoding reflects particular aspects of the articulator movements. The movements of the articulators can be described according to a variety of different kinematic parameters (e.g. position, speed or velocity, acceleration, etc.,). For each

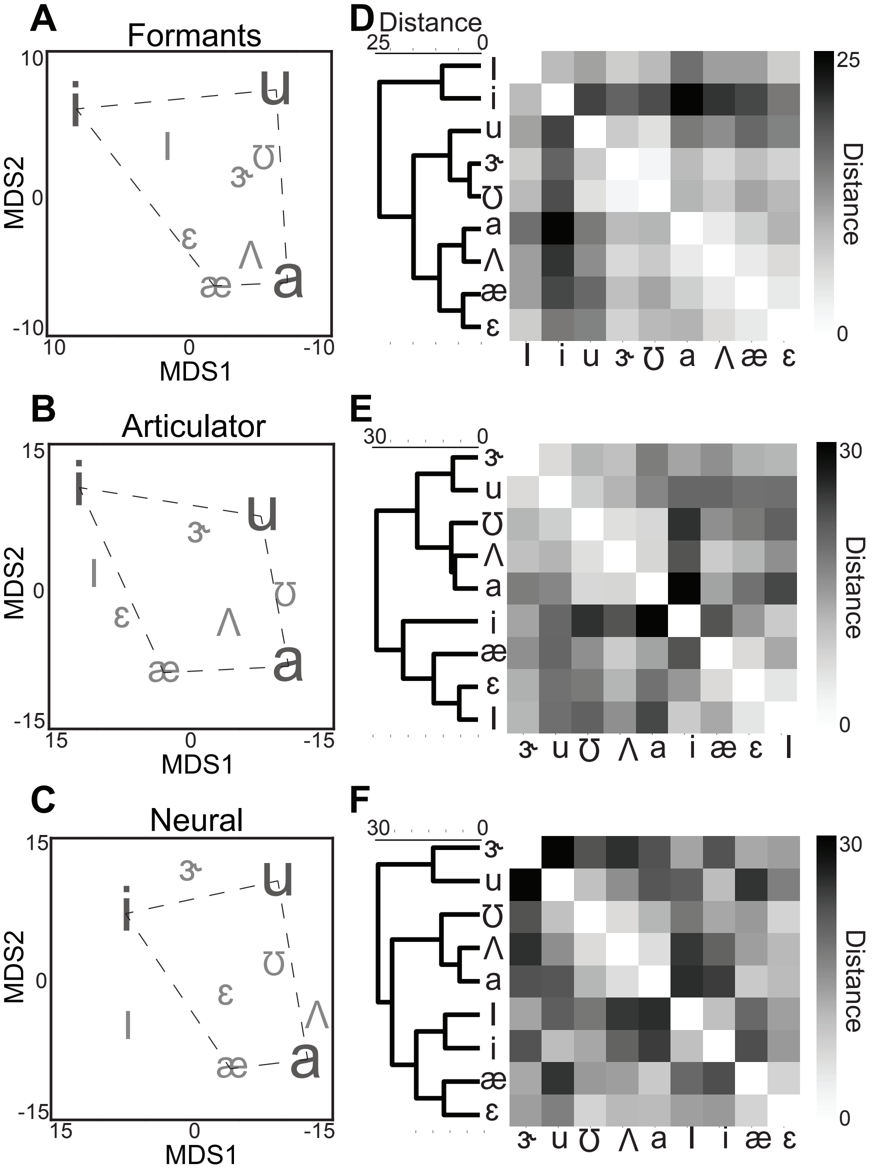


FIGURE 3.4: **vSMC activity reflects articulator kinematic organization of vowels *(a-c)*** Multidimensional Scaling (MDS) representations of (A) acoustic formants, (B) articulator position, and (C) vSMC HG activity. Each letter marks the position of the median production of the indicated vowel in MDS space across all subjects. The relative organization of the vowels is similar across spaces. For example, the low-back vowel /a/ is always near the mid-back vowel /ʌ/, but far from the high-front vowel /i/. ***(d-f)*** Hierarchical clustering (left) and confusion matrices (right) derived from the pairwise distances between vowels in the MDS spaces of (D) acoustic formants, (E) articulator position, and (F) vSMC HG.

kinematic parameter, we used L1-regularized encoding models to explain vSMC HG from the moment-to-moment measurements of position, speed, velocity, and acceleration. Since all four kinematic parameters are correlated with one another, we removed shared variance between the parameters using semi-partial correlations in order to better interpret their relative encoding performances.

We found electrodes that significantly encoded the trial-to-trial variability in position (Figure 3.5A), speed (Figure 3.5B), velocity (Figure 3.5C), and acceleration (Figure 3.5D). Speed was the most robustly encoded parameter at most vSMC electrodes, with significant encoding at more electrodes and a higher average correlation compared to the other parameters (U = 1720 to 2735, p = 3.3e-9 to 1.1e-14; Wilcoxon rank sum; Figure 3.5E).

To understand the timing of kinematic parameter encoding throughout the production of vowels, we also examined models that predicted HG neural activity from the joint combination of all four parameters simultaneously. These models were evaluated over a sliding 100ms window to characterize the kinematic parameter encoding during different phases of the trial (i.e., movement initiation, target position, steady state maintenance, and movement back to the starting position). We observed a peak in encoding for most electrodes around the onset of the movement (91% of electrodes), with some electrodes also showing a peak around the offset (9%; Figure 3.5F,G). There was no spatial organization associated with electrodes that specifically encoded particular parameters (intra-parameter, p = 0.31; cross-parameter, p = 0.08; see Methods), nor was there a significant relationship between electrodes that encoded specific kinematic parameters and specific articulators (χ^2 (9, N=155) = 9.26, p = 0.4; Chi-square). Strikingly, encoding during the steady state was near zero for all kinematic parameters.

To visualize how individual articulators contribute to the performance of these encoding models, we looked at the weights derived by these models. For each significant electrode we took the weighting of each kinematic feature at the peak encoding performance and performed a permutation test. Kinematic features whose weights exceeded the 99th percentile of the permuted distribution were considered significant. We found electrodes that were primarily driven by individual articulators, as well as those that had significant weighting for more than one articulator (black electrodes in Figure 3.6). We did not observe a consistent relationship between the primary articulator contributing to an electrodes

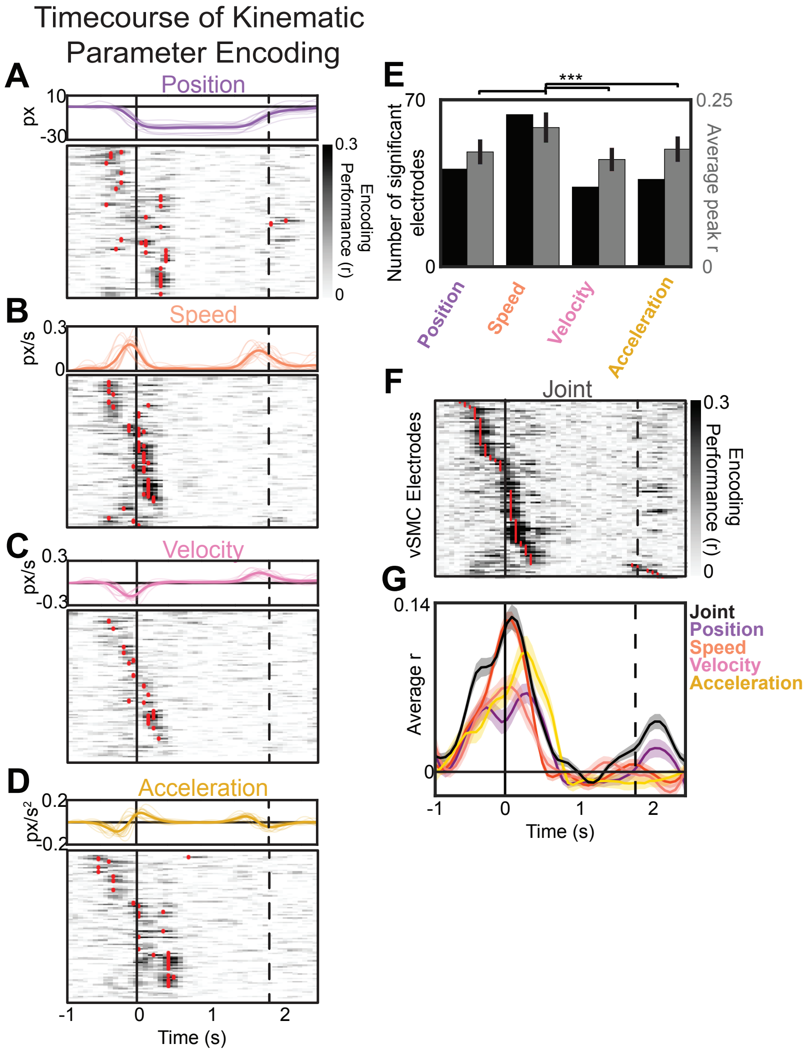


FIGURE 3.5: **Representations of position, speed, velocity, and acceleration kinematic features over time *(a-d)*** Top: Example kinematic parameters of position (A), speed (B), velocity (C), and acceleration (D) for all utterances of /a/ for one subject. Thin lines mark individual trials, while the thick line is the across-trial average. Bottom: Performance of encoding models predicting vSMC HG from articulator position (A), speed (B), velocity (C), or acceleration (D). vSMC electrodes with significant performance are marked by red dashes at the time peak encoding performance. Electrodes in A-D are plotted in order of their peak encoding times in the joint model (F). Vertical black lines mark the onset (solid) and offset (dashed) of vowel acoustics. ***(e)*** Comparison of the number of significant electrodes (black) and average peak performance (grey) of position, speed, velocity, and acceleration encoding models. Speed is significantly encoded at more electrodes, with a higher average model performance (*** $p<0.01$; Wilcoxon rank sum). ***(f)*** Performance of encoding models predicting vSMC HG from all articulator kinematics jointly. ***(g)***Average performance across significant electrodes for the joint and independent parameter models.

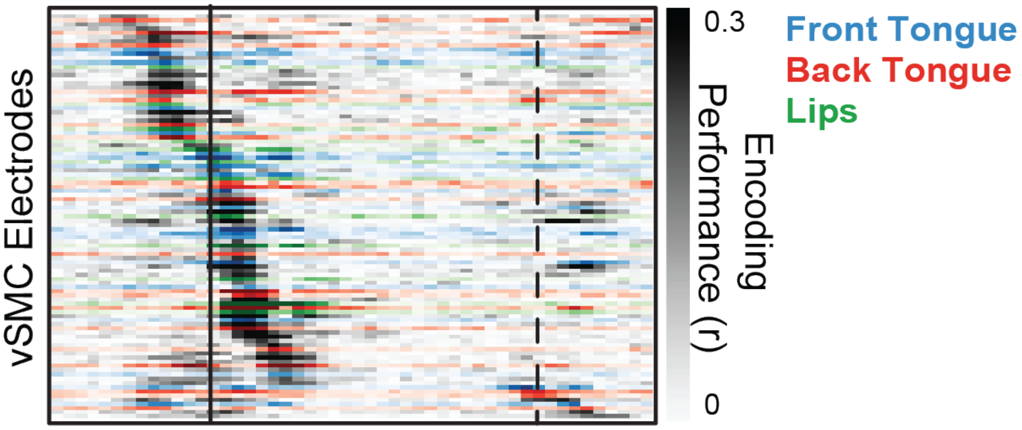


FIGURE 3.6: **Role of articulator representation in kinematic encoding *(a)*** Joint encoding models (same as in Figure 5f) colored by the articulator(s) that significantly contributed to the model of that electrode at its peak encoding performance. Black marks electrodes which had more than one articulator significantly contribute.

performance, and the timing of that electrodes significance.

To understand how these individual electrode kinematic representations relate to the population representations of articulator kinematics and dynamics, we additionally used L1-regularized decoding models to predict the articulator kinematics from the population of vSMC HG electrodes. As with the encoding analyses, these models were constructed from a small (100ms) sliding window of time, resulting in a description of how much of the trial-to-trial variability of the articulator position (Figure 3.7A), speed (Figure 3.7B), velocity (Figure 3.7C), and acceleration (Figure 3.7D) can be explained by vSMC HG activity. The time course of decoding strength was similar to the encoding models, with peaks around the onset and offset and near-zero values while the vowel was being held. Across kinematic parameters, articulator speed was the best-predicted parameter (U = 2.6 to 3.3, p = 8.1e-3 to 9.0e-4; Wilcoxon rank sum).

Together, these results demonstrate a strikingly sparse representation of kinematic parameters across time, despite the fact that there continues to be trial-to-trial variability in both kinematic and neural features throughout vowel (Figure 3.5). Only 56% of time points had significant encoding performance at any electrode, and individual significant electrodes had an average of 15% (±1) significant time

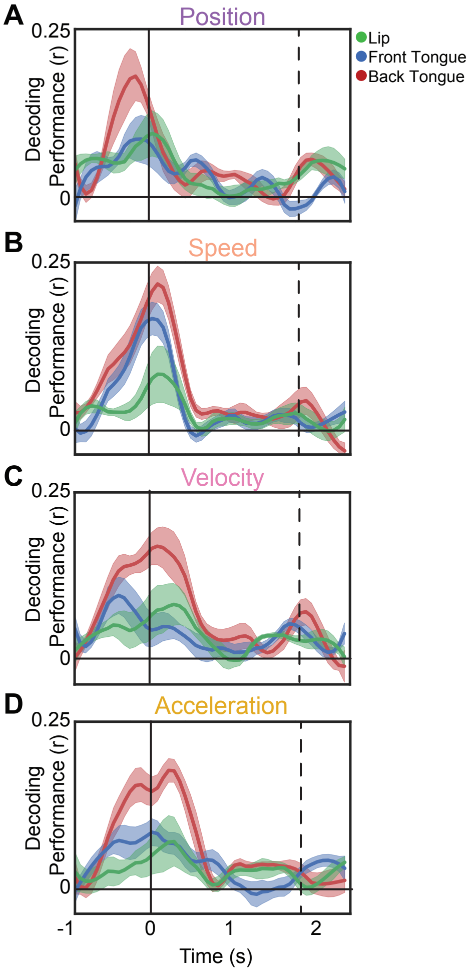


FIGURE 3.7: **Time course of kinematic parameter decoding *(a-d)*** Performance of decoding models predicting articulator position (A), speed (B), velocity (C), or acceleration (D) from HG at all vSMC electrodes. Features are averaged across subjects and within articulators. Black lines denote the onset and offset of vowel acoustics.

points. In particular, we did not observe any electrodes that exhibited significant kinematic parameter encoding during the steady state of the vowel.

3.3.5 Onset vs. steady state HG activity and kinematic encoding

The temporal sparsity of neural representations described above is particularly notable given that many electrodes showed sustained HG activity during the steady state portion of the vowel, independent of the particular articulatory movements that occurred (Figure 3.8A). These electrodes contrast with other HG activity that is only transiently increased around the onset and/or offset of the vowel (Figure 3.8A). To characterize these response types, we used non-negative matrix factorization (NMF) to derive basis functions that best describe vSMC HG temporal profiles across all electrodes (Hamilton, Edwards, and Chang, 2017). Our motivation for using NMF was not to provide a complete description of HG dynamics, but rather to provide an unsupervised method of quantifying the transient and onset/offset responses across electrodes. We found that the first two bases (i.e., the most important bases), captured the sustained and onset/offset response types we observed qualitatively (Figure 3.8B). Organizing all vSMC electrodes by the degree to which their activity is reconstructed by the first or second NMF bases (i.e., the NMF weights), we observed a continuum of HG dynamics: some electrodes showed sustained activity throughout the utterance, while others showed transient increases in activity only at onset and offset of the utterance (Figure 3.8C). Some electrodes showed a combination of sustained and transient components. There was no apparent spatial organization (intra-response type, p = 0.13; cross-response type p = 0.09; see Methods) or relationship between response type and the articulators represented at each electrode (χ^2 (3, N=155) = 2.3, p = 0.5; Chi-square).

We separately considered electrodes that showed stronger weights for the sustained NMF basis (basis 1 in Figure 3.8B). The average HG activity at these electrodes indeed showed sustained activity throughout the vowel, however there was not a concomitant sustained encoding of kinematic parameters (Figure 3.8D). This dissociation between activity and encoding was apparent even at the single trial level (Figure 3.9A). In other words, although some electrodes exhibit sustained HG activity throughout the production of the vowel, there is not a systematic relationship between the trial-to-trial variability

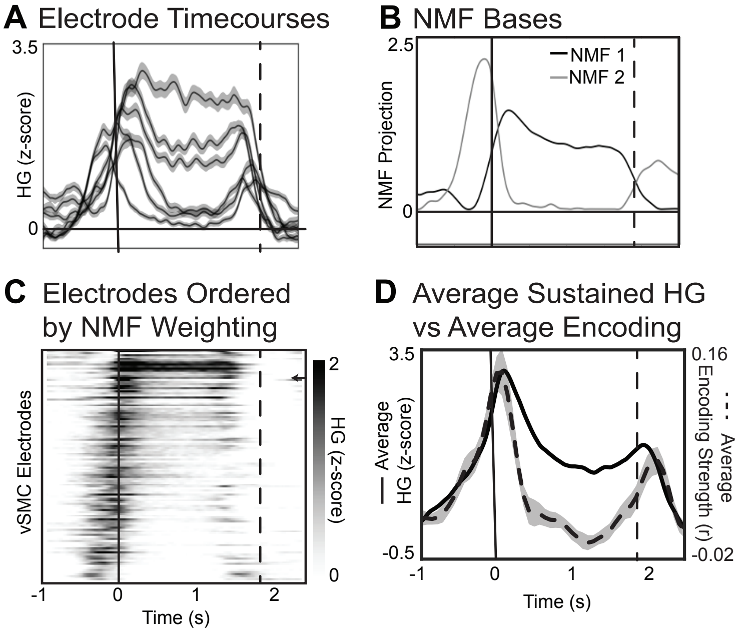


FIGURE 3.8: **Relationship of vSMC HG dynamics and kinematic encoding** ***(a)*** HG activity at several example electrodes illustrating the diverse dynamics during the same behavior, especially during the time period when the vowel is being held. ***(b)*** The first two NMF bases extracted from HG dynamics across all electrodes. These bases recapitulate the key differences in dynamics seen in the example electrodes, and serve as an unbiased quantification of the HG dynamics seen across vSMC. ***(c)*** HG activity at all vSMC electrodes ordered by the ratio of NMF bases used to reconstruct their activity ((NMF1-NMF2)/(NMF1+NMF2)). Arrow denotes the example electrode used in Figure 8. ***(d)***Average HG activity across all sustained (NMF1>NMF2) electrodes (solid) plotted alongside the average encoding performance (dashed) across time. During the steady state of the vowel, there is elevated activity, but almost no significant encoding of articulator kinematics.

of that activity and the kinematics of the articulators. Instead encoding of kinematics at electrodes with sustained activity was prevalent only around the onset and offset of movement. We hypothesized that although activity during the steady state does not relate to kinematic variability, it still reflects an important aspect of the task, namely the duration of each utterance. Across sustained electrodes we found that the duration of the HG timecourse was significantly correlated with the duration of the vocalization (Spearman's ρ = 0.61, p = 2e-153; Figure 9B). Thus, the sustained activity was associated with at least one aspect of vowel production.

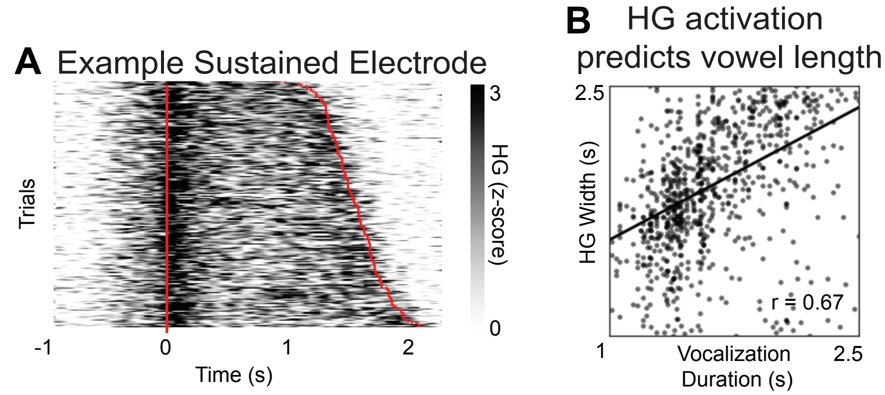


FIGURE 3.9: **Sustained HG activity is related to vowel duration** ***(a)*** An example electrode that is characterized by sustained HG activity. Trials are ordered by vowel production duration. Red lines mark the onset and offset of vowel acoustics. ***(b)*** The duration of the sustained HG activity at each trial is plotted against the trial duration for all sustained electrodes. Larger grey markers denote observations from the example electrode in A.

3.4 Discussion

We report a detailed description of how activity in speech-motor cortex controls the precise movements of the vocal tract articulators to produce vowels. By simultaneously measuring the movements of the articulators, recording the acoustic consequences of those movements, and recording the neural activity in vSMC, we are able to establish that the dominant representation in vSMC is articulator kinematics. The precise control of these movements allows speakers to create specific configurations of the mouth, which lead to distinct categories of sounds.

Without simultaneous measurements of the articulators, previous studies of the neural basis of speech production have relied on approximate, categorical phonemic-based descriptions of speech behavior (Crone et al., 1998a; Fukuda et al., 2010; Kellis et al., 2010; Pei et al., 2011; Leuthardt et al., 2011; Bouchard et al., 2013; Bouchard and Chang, 2014; Mugler et al., 2014; Herff et al., 2015). Although the produced acoustics and categorical vowel descriptions reflect the ultimate (perceptual) outcome of vocal tract movements, the many-to-one relationship between kinematics and vowels (Atal et al., 1978; Maeda, 1990; Gracco and Lofqvist, 1994; Johnson, Ladefoged, and Lindau, 1993; Perkell and Nelson, 1985) means that it was not possible to understand the precise nature of the neural representation in vSMC. Previous studies have implicated vSMC in articulator kinematic control in several ways. First, stimulation to sites in vSMC elicits involuntary activations of the orofacial muscles (Penfield and Boldrey, 1937; Huang et al., 2013). Second, neurons in these and other sensorimotor regions are often tuned to movement kinematics (Georgopoulos, Schwarz, and Ketiner, 1986; Paninski et al., 2004; Arce et al., 2013). Finally, the spatio-temporal patterns of HG activity in vSMC are consistent with the engagement of the articulators (Bouchard et al., 2013). The present results confirm these interpretations by showing directly that kinematic descriptions of speech behavior are more closely related to neural activity compared to acoustic or categorical vowel descriptions. Further, we find no evidence that vSMC activity encodes either produced acoustics or vowel category distinct from their correlations with the articulator kinematics. Crucially, we observed that this encoding scheme exists both at single electrodes, and across the spatially distributed set of electrodes. For spatially distributed activity patterns, we demonstrate the neural existence of the classic vowel 'trapezoid', which has dominated linguistic descriptions of speech (Harshman, Ladefoged, and Goldstein, 1977; Alfonso and Baer, 1982; Hillenbrand et al., 1995).

Furthermore, by characterizing the movements of the articulators according to a variety of kinematic parameters (position, speed, velocity, and acceleration), we demonstrated that neural activity encodes each of the examined parameters independent of one another. Previous studies examining arm movements using analogous parameters have also found significant encoding of these parameters (Georgopoulos et al., 1982; Georgopoulos, Schwarz, and Ketiner, 1986; Paninski et al., 2004; Arce et al., 2013; Moran and Schwartz, 1999). While we find electrodes that significantly encode each parameter examined, speed is by far the most robustly encoded parameter. Furthermore, the dominant kinematic parameter at individual electrodes was not significantly related to the articulator representation of those electrodes. The predominance of speed over other parameters is somewhat surprising; previous studies of the single-unit representation of kinematic parameters during arm reaching typically find that velocity and direction are the most commonly encoded parameter (Moran and Schwartz, 1999). Similar results were also observed in a recent ECoG study, which found that movement speed was predominately represented during arm reaching in humans (Hammer et al., 2016). The predominance of speed encoding was interpreted in the context of a model in which the summed activity of many velocity-tuned neurons with random directional tuning resembles speed tuning. Thus it may be the case that individual vSMC neurons are actually representing mostly velocity, but the summed activity observed with ECoG electrodes reflects the magnitude of movement without direction (i.e. speed). By studying vowels, we were able to examine the dynamics of kinematic encoding that are associated with movements to specific vocal tract articulators. We found that articulator kinematics were encoded around the time of movement onset and/or offset, but not while the vocal tract configuration was being held to maintain the vowel. Encoding of articulator kinematics only during movement onset and offset suggests that control of speech articulators is accomplished primarily through control of changes to the plant, rather than moment-to-moment maintenance of the vocal tract configuration. This is consistent with models of speech production that utilize changes to the plant as the primary mechanism by which sensorimotor cortex receives input from, and sends commands to, the vocal tract (Houde and Nagarajan, 2011; Tourville and Guenther, 2011). Furthermore, these dynamics have been observed in studies with analogous behavior from different body parts, including arm reaching. These studies have found individual neurons in motor cortex that exhibit transient firing patterns, where firing rates are high around movement onset and offset (Crammond and Kalaska, 1996; Arce et al., 2013; Shadmehr, 2017).

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

It is important to emphasize that these analyses focus on the neural representation of the supralaryngeal articulators. While the movements of these articulators are critical to the production of vowels, the lower vocal tract (e.g. larynx, pharynx, and diaphragm) is also necessary to produce vocalized sounds. It is likely that sub-regions of vSMC are involved in the control of the lower vocal tract, but this relationship is not presently examined (Brown, Martinez, and Parsons, 2006; Bouchard et al., 2013; Conant, Bouchard, and Chang, 2014).

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

We found that the representation of spoken vowels in vSMC is directly explained by the movements of speech articulators. The encoding of multiple kinematic parameters is present for the articulators, most prominently speed. Articulator kinematic encoding was primarily observed at the onset and offset of vowel production and not while the vowel is being held. Together, these findings provide insight into how neural activity in primary sensorimotor cortex results in the precise production of vowels. Future work will address how these encoding properties operate in the context of natural continuous speech.

Bibliography

Alfonso, Peter J and Thomas Baer (1982). “Dynamics of Vowel Articulation”. In: *Lang. Speech* 25.2, pp. 151–173.

Arce, F I et al. (2013). “Directional information from neuronal ensembles in the primate orofacial sensorimotor cortex.” In: *J. Neurophysiol.* 110.6, pp. 1357–69. ISSN: 1522-1598. DOI: `10.1152/jn.00144.2013`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/23785133`.

Atal, B S et al. (1978). “Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique.” In: *J. Acoust. Soc. Am.* 63.May 1978, pp. 1535–1553. ISSN: 00014966. DOI: `10.1121/1.381848`.

Baer, T et al. (1991). “Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels.” In: *J. Acoust. Soc. Am.* 90.2 Pt 1, pp. 799–828. ISSN: 0001-4966. URL: `http://www.ncbi.nlm.nih.gov/pubmed/1939886`.

Black, A. W. (2006). “{CLUSTERGEN}: A Statistical Parametric Synthesizer Using Trajectory Modeling”. In: *Interspeech*, pp. 1762–1765.

Boersma, Paul (2001). “Praat, a system for doing phonetics by computer”. In: *Glot Int.* 5.9, pp. 341–345.

Borden, Gloria and Thomas Gay (1979). “Temporal Aspects of Articulatory Movements for /s/-Stop Clusters”. In: *Phonetica* 36, pp. 21–31.

Bouchard, K. E. and E. F. Chang (2014). “Control of Spoken Vowel Acoustics and the Influence of Phonetic Context in Human Speech Sensorimotor Cortex”. In: *J. Neurosci.* 34.38, pp. 12662–12677. ISSN: 0270-6474. URL: `http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.1219-14.2014`.

Bouchard, Kristofer E (2015). *Bootstrapped Adaptive Threshold Selection for Statistical Model Selection and Estimation*. URL: `arxiv:1505.03511`.

Bouchard, Kristofer E et al. (2013). “Functional organization of human sensorimotor cortex for speech articulation.” In: *Nature* 495.74441, pp. 327–332. ISSN: 1476-4687. URL: `http://www.ncbi.nlm.nih.gov/pubmed/23426266`.

Bouchard, Kristofer E. et al. (2016). “High-resolution, non-invasive imaging of upper vocal tract articulators compatible with human brain recordings”. In: *PLoS One* 11.3, pp. 1–30. ISSN: 19326203. DOI: `10.1371/journal.pone.0151327`.

Breiman, L (2001). “Random Forests”. In: *Mach. Learn.* 45.1, pp. 5–32.

Browman, C and L Goldstein (1990). “Gestural Specification Using Dynamically-defined Articulatory Structures”. In: *J. Phoneticsetics* 18, pp. 299–320.

Brown, Steven, Michael J Martinez, and Lawrence M Parsons (2006). “Music and language side by side in the brain: a PET study of the generation of melodies and sentences.” In: *Eur. J. Neurosci.* 23.10, pp. 2791–803. ISSN: 0953-816X. DOI: `10.1111/j.1460-9568.2006.04785.x`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/16817882`.

Brown, Steven et al. (2005). "Stuttered and fluent speech production: an ALE meta-analysis of functional neuroimaging studies." In: *Hum. Brain Mapp.* 25.1, pp. 105–17. ISSN: 1065-9471. DOI: 10.1002/hbm.20140. URL: http://www.ncbi.nlm.nih.gov/pubmed/15846815.

Brown, Steven et al. (2009). "The somatotopy of speech: Phonation and articulation in the human motor cortex". In: *Brain Cogn.* 70.1, pp. 31–41. DOI: 10.1016/j.bandc.2008.12.006.The.

Conant, David, Kristofer E Bouchard, and Edward F Chang (2014). "Speech map in the human ventral sensory-motor cortex". In: *Curr. Opin. Neurobiol.* 24, pp. 63–67. ISSN: 09594388. URL: http://linkinghub.elsevier.com/retrieve/pii/S0959438813001748.

Crammond, Donald and John F Kalaska (1996). "Differential relation of discharge in primary motor cortex and premotor cortex to movements versus actively maintained postures during a reaching task". In: *Exp. brain Res.* 108, pp. 45–61.

Crone, N E et al. (1998a). "Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. I. Alpha and beta event-related desynchronization." In: *Brain* 121 (Pt 1, pp. 2271–99. ISSN: 0006-8950. URL: http://www.ncbi.nlm.nih.gov/pubmed/9874480.

Crone, N E et al. (1998b). "Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band." In: *Brain* 121, pp. 2301–15. ISSN: 0006-8950. URL: http://www.ncbi.nlm.nih.gov/pubmed/9874481.

Donoho, David and Victoria Stodden (2003). "When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts?"

Fukuda, Miho et al. (2010). "Cortical gamma-oscillations modulated by listening and overt repetition of phonemes". In: *Neuroimage* 49.3, pp. 2735–2745. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2009.10.047. arXiv: NIHMS150003. URL: http://dx.doi.org/10.1016/j.neuroimage.2009.10.047.

Georgopoulos, a P et al. (1982). "On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex." In: *J. Neurosci.* 2.11, pp. 1527–37. ISSN: 0270-6474. URL: http://www.ncbi.nlm.nih.gov/pubmed/7143039.

Georgopoulos, Apostolos P, Andrew B Schwarz, and Ronald E Ketiner (1986). "Neuronal population coding of movement direction". In: *Science (80-.).* 233, pp. 1416–1419.

Gick, Bryan, Ian Wilson, and Donald Derrick (2012). *Articulatory Phonetics*. Malden: Wiley & Sons, p. 272.

Grabski, Krystyna et al. (2012). "Functional MRI assessment of orofacial articulators: neural correlates of lip, jaw, larynx, and tongue movements." In: *Hum. Brain Mapp.* 33.10, pp. 2306–21. ISSN: 1097-0193. DOI: 10.1002/hbm.21363. URL: http://www.ncbi.nlm.nih.gov/pubmed/21826760.

Gracco, V L and A Lofqvist (1994). "Speech motor coordination and control: evidence from lip, jaw, and laryngeal movements". In: *J. Neurosci.* 14.11, pp. 6585–6597. ISSN: 0270-6474.

Guenther, Frank H, Satrajit S Ghosh, and Jason a Tourville (2006). "Neural modeling and imaging of the cortical interactions underlying syllable production." In: *Brain Lang.* 96.3, pp. 280–301. ISSN: 0093-934X. DOI: 10.1016/j.bandl.2005.06.001. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1473986{\&}tool=pmcentrez{\&}rendertype=abstract.

Hamilton, Liberty S, Erik Edwards, and Edward F Chang (2017). "Parallel streams define the temporal dynamics of speech processing across human auditory cortex". In: *bioRxiv*.

Harshman, R, P Ladefoged, and L Goldstein (1977). "Factor analysis of tongue shapes." In: *J. Acoust. Soc. Am.* 62.3, pp. 693–713. ISSN: 0001-4966. URL: `http://www.ncbi.nlm.nih.gov/pubmed/903511`.

Herff, C. et al. (2015). "Brain-to-text: Decoding spoken phrases from phone representations in the brain". In: *Front. Neuroeng.* 8.6.

Hesselmann, Volker et al. (2004). "Discriminating the cortical representation sites of tongue and up movement by functional MRI." In: *Brain Topogr.* 16.3, pp. 159–67. ISSN: 0896-0267. URL: `http://www.ncbi.nlm.nih.gov/pubmed/15162913`.

Hillenbrand, James et al. (1995). "Acoustic characteristics of American English vowels". In: *J. Acoust. Soc. Am.* 97.5, pp. 3099–3111. ISSN: 00014966.

Houde, John F and Srikantan S Nagarajan (2011). "Speech production as state feedback control". In: *Front. Hum. Neurosci.* 5.October, pp. 1–14. DOI: `10.3389/fnhum.2011.00082`.

Huang, C S et al. (2013). "Organization of the primate face motor cortex as revealed by intracortical microstimulation and electrophysiological identification of afferent inputs and corticobulbar projections Organization of the Primate Face Motor Cortex as Revealed by Intracortical". In: *J. Neurophysiol.* 59.3, pp. 796–818.

Johnson, K, P Ladefoged, and M Lindau (1993). "Individual differences in vowel production." In: *J. Acoust. Soc. Am.* 94.2, pp. 701–14. ISSN: 0001-4966. URL: `http://www.ncbi.nlm.nih.gov/pubmed/8370875`.

Jürgens, U (2009). "The neural control of vocalization in mammals: a review." In: *J. Voice* 23.1, pp. 1–10. ISSN: 1557-8658. DOI: `10.1016/j.jvoice.2007.07.005`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/18207362`.

Jürgens, Uwe (2002). "Neural pathways underlying vocal control." In: *Neurosci. Biobehav. Rev.* 26.2, pp. 235–58. ISSN: 0149-7634. URL: `http://www.ncbi.nlm.nih.gov/pubmed/11856561`.

Kellis, Spencer et al. (2010). "Decoding spoken words using local field potentials recorded from the cortical surface." In: *J. Neural Eng.* 7.5, pp. 56007–16. ISSN: 1741-2560. DOI: `10.1088/1741-2560/7/5/056007`.

Kent, RD and KL Moll (1972). "Tongue Body Articulation during Vowel and Diphthong Gestures". In: *Folia Phoniatr. Logop.* 24.4, pp. 278–300.

Kim, Jingu and Haesun Park (2008). *Sparse Nonnegative Matrix Factorization for Clustering*. Tech. rep. URL: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.4302{\&}rep=rep1{\&}type=pdf`.

Ladefoged, Peter and Keith Johnson (2011). *A Course in Phonetics*. Boston: Cengage Learning.

Lee, D D and H S Seung (1999). "Learning the parts of objects by non-negative matrix factorization." In: *Nature* 401.6755, pp. 788–791. ISSN: 0028-0836. DOI: `10.1038/44565`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/10548103`.

Leuthardt, Eric C et al. (2011). "Using the electrocorticographic speech network to control a brain-computer interface in humans." In: *J. Neural Eng.* 8.3. ISSN: 1741-2552. DOI: `10.1088/1741-2560/8/3/036004`. arXiv: `NIHMS150003`. URL: `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3701859{\&}tool=pmcentrez{\&}rendertype=abstract`.

Li, Min, Chandra Kambhamettu, and Maureen Stone (2004). "Automatic contour tracking in ultrasound images." In: *Clin. Linguist. Phon.* 19.6-7, pp. 545–554. ISSN: 0269-9206. DOI: `10.1080/02699200500113616`.

Lindblom, E F and E F Sundberg (1971). "Acoustical Consequences of Lip, Tongue, Tongue, Jaw, and Larynx Movement". In: *J. Acoust. Soc. Am.* 50.4, pp. 1166–1179.

Lofqvist, Anders and Vincent Gracco (1999). "Interarticulator programming in VCV sequences: Lip and tongue movements". In: *J. Acoust. Soc. Am.* 105.3, pp. 1864–1876. ISSN: 00014966.

Lotze, M et al. (2000). "The representation of articulation in the primary sensorimotor cortex." In: *Neuroreport* 11.13, pp. 2985–9. ISSN: 0959-4965. URL: `http://www.ncbi.nlm.nih.gov/pubmed/11006980`.

Maddieson, I and SF Disner (1984). *Patterns of Sounds: Cambridge Studies in Speech Science and Communication*. Cambridge: Cambridge University Press.

Maeda, Shinji (1990). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model". In: *Speech Prod. speech Model.* Netherlands: Springer, pp. 131–149.

Moran, Daniel W and Andrew B Schwartz (1999). "Motor Cortical Representation of Speed and Direction During Reaching". In: *J. Neurophysiol.* 82.5, pp. 2676–2692.

Mugler, Emily M et al. (2014). "Direct classification of all American English phonemes using signals from functional speech motor cortex." In: *J. Neural Eng.* 11.3, p. 035015. ISSN: 1741-2552. DOI: `10.1088/1741-2560/11/3/035015`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/24836588`.

Muller, Leah et al. (2016). "Spatial resolution dependence on spectral frequency in human speech cortex electrocorticography". In: *J. Neural Eng.* 13.

Narayanan, Shrikanth et al. (2004). "An approach to real-time magnetic resonance imaging for speech production." In: *J. Acoust. Soc. Am.* 115.4, pp. 1771–1776. ISSN: 00014966. DOI: `10.1121/1.1652588`.

Naylor, Patrick a. et al. (2007). "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm". In: *IEEE Trans. Audio, Speech Lang. Process.* 15.1, pp. 34–43. ISSN: 15587916. DOI: `10.1109/TASL.2006.876878`.

Noiray, Aude et al. (2011). "Test of the movement expansion model: anticipatory vowel lip protrusion and constriction in French and English speakers." In: *J. Acoust. Soc. Am.* 129.1, pp. 340–349. ISSN: 1520-8524. DOI: `10.1121/1.3518452`. URL: `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3055290{\&}tool=pmcentrez{\&}rendertype=abstract`.

Paninski, Liam et al. (2004). "Spatiotemporal tuning of motor cortical neurons for hand position and velocity." In: *J. Neurophysiol.* 91.1, pp. 515–32. ISSN: 0022-3077. URL: `http://www.ncbi.nlm.nih.gov/pubmed/13679402`.

Pei, Xiaomei et al. (2011). "Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans." In: *J. Neural Eng.* 8.4. ISSN: 1741-2560. DOI: `10.1088/1741-2560/8/4/046028`.

Penfield, W and L Roberts (1959). *Speech and brain mechanisms*. Princeton.

Penfield, Wilder and Edwin Boldrey (1937). "Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation". In: *Brain* 60.4, pp. 389–443. ISSN: 00068950.

Perkell, J S et al. (1993). “Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: a pilot "motor equivalence" study”. In: *J. Acoust. Soc. Am.* 93.5, pp. 2948–61. ISSN: 0001-4966. URL: http://www.ncbi.nlm.nih.gov/pubmed/8315158.

Perkell, Joseph S and Winston L Nelson (1985). “Variability in production of the vowels /i/ and /a/”. In: *Acoust. Soc. Am.* 77.5, pp. 1889–1895.

Petersen, SE et al. (1988). “Positron emissiion tomographic studies of the cortical anatomy of single-word Processing”. In:

Ramanarayanan, Vikram, Athanasios Katsamanis, and Shrikanth Narayanan (2011). “Automatic data-driven learning of articulatory primitives from real-time MRI data using convolutive nmf with sparseness constraints”. In: *INTERSPEECH*, pp. 61–64. ISSN: 19909772.

Ray, Supratim and John H R Maunsell (2011). “Different origins of gamma rhythm and high-gamma activity in macaque visual cortex”. In: *PLoS Biol.* 9.4. ISSN: 15449173. DOI: 10.1371/journal.pbio.1000610.

Shadmehr, Reza (2017). “Distinct neural circuits for control of movement vs . holding still”. In: *J. Neurophysiol.* 117, pp. 1431–1460. DOI: 10.1152/jn.00840.2016.

Simonyan, Kristina et al. (2009). “Functional but not structural networks of the human laryngeal motor cortex show left hemispheric lateralization during syllable but not breathing production.” In: *J. Neurosci.* 29.47, pp. 14912–23. ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.4897-09.2009. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2805075{\&}tool=pmcentrez{\&}rendertype=abstract.

Sutton, D, C Larson, and R C Lindeman (1974). “Neocortical and limbic lesion effects on primate phonation.” In: *Brain Res.* 71.1, pp. 61–75. ISSN: 0006-8993. URL: http://www.ncbi.nlm.nih.gov/pubmed/4206919.

Thomas, Mark R P and Patrick a. Naylor (2009). “The sigma algorithm: A glottal activity detector for electroglottographic signals”. In: *IEEE Trans. Audio, Speech Lang. Process.* 17.8, pp. 1557–1566. ISSN: 15587916. DOI: 10.1109/TASL.2009.2022430.

Toda, Tomoki, Alan W Black, and Keiichi Tokuda (2008). “Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model”. In: *Speech Commun.* 50.3, pp. 215–227.

Tourville, Jason a and Frank H Guenther (2011). “The DIVA model: A neural theory of speech acquisition and production.” In: *Lang. Cogn. Process.* 26.7, pp. 952–981. ISSN: 0169-0965. DOI: 10.1080/01690960903498424. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3650855{\&}tool=pmcentrez{\&}rendertype=abstract.

Ueda, Yuichi et al. (2007). “A real-time formant tracker based on the inverse filter control method”. In: *Acoust. Sci. Technol.* 28.4, pp. 271–274. ISSN: 1346-3969.

Watanabe, a. (2001). “Formant estimation method using inverse-filter control”. In: *IEEE Trans. Speech Audio Process.* 9.4, pp. 317–326. ISSN: 10636676.

Welker, W I et al. (1957). “Motor effects of stimulation of cerebral cortex of squirrel monkey (Saimiri sciureus).” In: *J. Neurophysiol.* 20.4, pp. 347–64. ISSN: 0022-3077. URL: http://www.ncbi.nlm.nih.gov/pubmed/13439407.

Yamagishi, Junichi et al. (2009). “Robust speaker-adaptive HMM-based text-to-speech synthesis”. In: *IEEE Trans. Audio, Speech Lang. Process.* 17.6, pp. 1208–1230.

Zen, Heiga, Keiichi Tokuda, and Alan W. Black (2009). “Statistical parametric speech synthesis”. In: *Speech Commun.* 51, pp. 1039–1064. ISSN: 01676393. DOI: `10.1016/j.specom.2009.04.004`.

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

12/3/2017

Date