

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

Neural signals and control of the larynx

**Permalink**

<https://escholarship.org/uc/item/0134s162>

**Author**

Dichter, Benjamin K

**Publication Date**

2017

Peer reviewed|Thesis/dissertation

Neural signals and control of the larynx

by

Benjamin K. Dichter

DISSERTATION

Submitted in partial satisfaction of the requirements of the degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO



## Acknowledgments

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

I would also like to thank my mother, father and twin sister for the advice and support they have given me throughout this journey, and my uncle Dr. Marc Dichter, for fostering my interest in computational neuroscience since high school.

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

Joseph G. Makin, Benjamin K Dichter, Philip N Sabes. “Learning to Estimate Dynamical State with Probabilistic Population Codes.” Published in PLoS Computational Biology, 2015.

Benjamin K. Dichter, Kristofer E. Bouchard, Edward F. Chang. “Dynamic Structure of Neural Variability in the Cortical Representation of Speech Sounds.” Published in Journal of Neuroscience, 2016.

## Neural signals and control of the larynx

Benjamin Dichter

The ability of the human brain to represent senses reliably and command a motor response is central to our ability to respond to the world around us. Here, I aim to understand these processes through simulation and experimental analysis. First I developed a recurrent neural network as a model of the brain receiving approximate sensory input. By learning to capture the distribution of the representation of a sense, the network learns the dynamics of the underlying stimulus and learns to integrate information near optimally over time, using recent estimates of position and velocity to inform the current estimate of the state of the object.

Next, I analyzed the cortical representation of auditory speech. Syllables were played to human subjects while recording voltage fluctuations directly from their brains using electrocorticography. I found that the variability of the neural activity in the superior temporal gyrus, an auditory cortical region, was “quenched” upon stimulus presentation. Furthermore, this decrease in variability is coincident with stimulus representation, and enables the brain to represent a stimulus more accurately.

Then, I examined the cortical control of laryngeal functions in humans using electrocorticography during produced speech. I found that the dorsal laryngeal motor cortex controls modulations of vocal pitch. Activity in that region is correlated with pitch in speech and in song. The representation of pitch in this region is separable from voicing, showing multiple dimensions of control represented in the cortex. Through cortical stimulation, I show that activity in this area caused proportional laryngeal muscle activation. I discuss how these findings may add important information furthering our understanding the evolution of speech in humans. Finally, I discuss how these signals can be used to decode prosodic patterns directly from neural activity for use in a speech prosthetic.

## Table of Contents

Introduction.....	1
Chapter 1: Learning to Estimate Dynamical State with Probabilistic Population Codes .....	5
Introduction .....	6
Results.....	9
Discussion .....	22
Related Work .....	23
Methods.....	29
Acknowledgements and Author Contributions.....	44
References .....	53
Chapter 2: Dynamic Structure of Neural Variability in the Cortical Representation of Speech Sounds .....	58
Introduction .....	59
Materials and Methods .....	61
Results.....	68
Discussion .....	77
References .....	88
Chapter 3: The Control of Vocal Pitch in Human Laryngeal Motor Cortex.....	94
Introduction .....	95
Results.....	96
Discussion .....	106
Methods.....	107
Supplemental Figures.....	114

References .....	118
Chapter 4: Decoding Prosody .....	123
Pitch Decoding .....	124
Decoding Word-of-Emphasis.....	125
References .....	128
Conclusion.....	129

## List of Figures

### Chapter 1:

Figure 1. Dynamical system and the neural network that learns it .....	45
Figure 2. Mean squared errors (MSEs) for various dynamical systems .....	46
Figure 3. Dynamical system with efference copy .....	47
Figure 4. Box-and-whisker plot .....	48
Figure 5. Position and velocity receptive fields of hidden units .....	49
Figure 6. Network sensitivity to instantaneous reliability .....	50
Figure 7. Emergence of position and velocity receptive during training .....	50
Figure 8. The weight matrices .....	51
Figure 9. Training and testing: in the model .....	52

### Chapter 2:

Figure 1: Example response and correlation .....	82
Figure 2: Reduction in mean dependence of variance .....	83
Figure 3: Temporal and spatial correlation .....	84
Figure 4: Factor analysis .....	85
Figure 5: Noise correlations and decoding .....	86
Figure 6: Multiband change in variability and encoding .....	87

### Chapter 3:

Figure 1: Human cortical encoding of vocal pitch .....	98
Figure 2: Cortical representation of pitch contour components in speech .....	100
Figure 3: Pitch encoding during singing. ....	102
Figure 4: Electrical stimulation of dLMC .....	105



Figure S1: Electrode coverage .....	114
Figure S2: Pitch partial correlation analysis .....	114
Figure S3: f0 tuning in MOCHA sentence production.....	115
Figure S4: Singing performance.....	116
Figure S5: Stimulation-evoked vocalizations.....	117

Chapter 4:

Figure 1.....	125
Figure 2.....	126

## Introduction

Brains are machine that sense the body and the outside world, process the observation, form a plan of action, and send a command to muscles that carry out the plan. In order to achieve understanding of sensory inputs robustly in an ever-changing environment, the brain must respond in a systematic way to sensory stimuli. This relationship between the senses and the corresponding neural activity is often referred to as a “representation,” and much of systems neuroscience is devoted to understanding the nature of these representations, and the neural transformations thereof that facilitate appropriate responses.

All of our senses must be captured by the electrochemical signals of neurons and transformations must be mediated by the connectivity between neurons. Although sensory observations generally exist in continuous, multidimensional spaces, the brain must represent them within the physical constraints of neuro-anatomy. A representational scheme commonly found in the brain is a grid of “receptive fields.” Each neuron has a maximal rate of action potentials for a specific sensory value (e.g. pitch, arm position, etc.) and the firing rate falls gradually as the sense deviates more from that “preferred” value. A population of neurons can faithfully capture a sensory space if the neurons have receptive fields that densely cover the range of that sense.

The senses, though rich, often give organisms incomplete information about the state of their surroundings. In fact, sensory information is often better thought of as *evidence* about the state of the environment. Brains must receive, weigh, and combine evidence about the state of the world, and formulate a plan based on the probability distribution. A corollary of this fact is that the neural representations of sensory evidence must represent not point estimates (i.e. a best guess) of the state of the world, but probability distributions over possible states. It may seem unintuitive- most of us are not aware that the brain is considering a spectrum of world states and performing the complicated mathematics necessary to transform these distributions to a plan.

However, numerous psychophysical experiments have elicited behavior that would require the use of probability distributions and are incompatible with point estimate representations. A common example is multisensory integration: when given two conflicting cues from different senses about the location of target, subjects will estimate the position to be closer to the more reliable sense. This requires the brain to weigh senses according to the likelihood distributions determined by the reliability of each sense.

A popular theory for the representation of probability distributions in the brain is the use of “probabilistic population codes.” In this scheme, the receptive field grid of neurons model the likelihood of states directly from the firing rates of neurons. This theory imposes assumptions about the shape of the receptive fields, the density of the coverage in sensory space, and the variability of response. Using these assumptions, a likelihood can be derived using Bayes rule, which describes a probability distribution of a state in the body or world given the firing rates of the neurons which represent it.

There is yet another challenge faced by the brain: the world is constantly changing, causing rapid fluctuations in the senses that must be represented concurrently by the brain. Thus, the representational scheme employed to represent a phenomenon must be appropriate for the dynamics of that sense. A scheme that requires 10 seconds to represent the probability distribution of position of a limb would be useless for guiding a gate. It might seem, then, that in the face of these changing states, the brain must infer the state of the world only on the current instantaneous sensory information. In Chapter 1, we explore an alternative. We show that there is a tool that can be used – and is in fact used – by the brain to integrate information over time.

Just as the brain can probabilistically combine information from different senses, it can also combine information from the past state estimates and the current sensory information. To do this, the brain needs to build a model of the dynamics of that state (e.g. the probability

distribution of position and velocity of the arm now given the position and velocity 1 second ago). In this way, information can be combined through time similarly to multi-sensory integration. In order to employ this strategy, the brain must build a representational map (state  $\rightarrow$  neural activity) as well as a dynamical model (earlier state  $\rightarrow$  current state). Estimating these models must be done indirectly, because, as established above, the brain never knows exact states, only probability distributions over states.

This type of indirect inference is a well-studied problem in statistical learning theory, and is called a *latent variable problem*. Effective techniques have been developed for inference of these models (most notably expectation maximization). In this context, the goal of inference is to estimate these relationships based solely on the observed patterns of neural activation.

However, established techniques are unlikely to be used by the brain. In the first chapter, I explore how the brain might perform dynamical state estimation by modeling the distribution of data. The approach extends successful computational models of multisensory integration. I show how the brain could use established neural learning rules to build the functional connections necessary to integrate probabilistic codes of state estimates over time, and I prove that this learning technique would allow the brain to learn a dynamical model with higher dynamics than those represented in the instantaneous sensory evidence.

The representation of speech sounds in the human brain is a particularly interesting case of neural representation. The human brain is so specialized and successful at representing speech sounds that we can perform the computationally difficult task of speech perception in a variety of noisy environments with ease. In the second chapter, I examine the neural representation of these sounds using electrocorticography (ECoG). Surgically implanted electrodes record from the cortex and monitor neural activation as subjects listen to speech sounds. I show that the variability of neural activity is reduced when the brain is representing a sound, and that this

reduction in variability helps the brain faithfully capture that sound so that it can be distinguished from other speech sounds.

In chapter three, I close the behavior loop by using ECoG to study the neural representation of vocal pitch as we speak and sing. I establish neural representation of vocal pitch in humans, and show that it is distinct from voicing, another behavior of the larynx. I also use electrical cortical stimulation to establish a causal relationship between activation of this region and excitation of the laryngeal muscles necessary to produce and modulate vocal pitch.

Finally, I take a more engineering approach, showing how one might use ECoG signals to determine the intended emphasis pattern. In my first approach, I use a Kalman Filter to integrate neural information over time. This approach is not unlike the neural network explored in Chapter 1. Here, dynamical state estimation is used in an engineering application inferring the intention of a subject rather than as a model of how the brain works. In my second approach, I explore decoding of word-emphasis discretely. Here, I appreciate the phonological and semantic aspects of pitch by decoding specifically the essential information of extracting the meaning of the prosodic contour.

# Chapter 1. Learning to Estimate Dynamical State with Probabilistic Population Codes

Joseph G. Makin<sup>\*1,2</sup>, Benjamin K. Dichter<sup>\*1,3</sup>, Philip N. Sabes<sup>1,2,3</sup>

<sup>1</sup>Center for Integrative Neuroscience, University of California, San Francisco, San Francisco, California, United States of America,

<sup>2</sup>Department of Physiology, University of California, San Francisco, San Francisco, California, United States of America, <sup>3</sup>UC Berkeley-UCSF Graduate Program in Bioengineering, University of California, San Francisco, San Francisco, California, United States of America. \*These authors contributed equally to this work.

**Abstract** Tracking moving objects, including one's own body, is a fundamental ability of higher organisms, playing a central role in many perceptual and motor tasks. While it is unknown how the brain learns to follow and predict the dynamics of objects, it is known that this process of state estimation can be learned purely from the statistics of noisy observations. When the dynamics are simply linear with additive Gaussian noise, the optimal solution is the well-known Kalman filter (KF), the parameters of which can be learned via latent-variable density estimation (the EM algorithm). The brain does not, however, directly manipulate matrices and vectors, but instead appears to represent probability distributions with the firing rates of population of neurons, “probabilistic population codes.” We show that a recurrent neural network—a modified form of an exponential family harmonium (EFH)—that takes a linear probabilistic population code as input can learn, without supervision, to estimate the state of a linear dynamical system. After observing a series of population responses (spike counts) to the position of a

moving object, the network learns to represent the velocity of the object and forms nearly optimal predictions about the position at the next time-step. This result builds on our previous work showing that a similar network can learn to perform multisensory integration and coordinate transformations for static stimuli. The receptive fields of the trained network also make qualitative predictions about the developing and learning brain: tuning gradually emerges for higher-order dynamical states not explicitly present in the inputs, appearing as delayed tuning for the lower-order states.

**Author Summary** A basic task for animals is to track objects—predators, prey, even their own limbs—as they move through the world. Because the position estimates provided by the senses are not error-free, higher levels of performance can be, and are, achieved when the velocity and acceleration, as well as the position, of the object are taken into account. Likewise, tracking of limbs under voluntary control can be improved by considering the motor command that is (partially) responsible for its trajectory. Engineers have built tools to solve precisely these problems, and even to learn dynamical features of the object to be tracked. How does the brain do it? We show how artificial networks of neurons can learn to solve this task, simply by trying to become good predictive models of their incoming data—as long as some of those data are the activities of the neurons themselves at a fixed time delay, while the remainder (imperfectly) report the current position. The tracking scheme the network learns to use—keeping track of past positions; the corresponding receptive fields; and the manner in which they are learned, provide predictions for brain areas involved in tracking, like the posterior parietal cortex.

## Introduction

Over the last decade, neuroscience has come increasingly to believe that sensory systems represent not merely stimuli, but probability distributions over them. This conclusion follows from two observations. The first is that the apparent stochasticity of the response,  $R$ , of a population

of neurons inherently represents the likelihood of the stimulus  $R : p(r|s)$  [1]. The second is that certain common computations essential to the function of many animals require keeping track of probability distributions over stimuli, rather than mere point estimates. For example, primates integrate information from multiple senses by weighting each sense by its reliability (inverse variance) [5, 6]. This framework has been used to hand-wire neural networks that integrate spatial information across sensory modalities and across time [2, 7, 8]. The more challenging problem faced by the brain, however, is to learn to perform these tasks.

We have recently shown [4, 9] that the problem of learning to integrate information about a common stimulus from multiple, unisensory populations of neurons can be solved by a neural network that implements a form of unsupervised learning called density estimation. Such a network learns to represent the joint probability density of the unisensory responses—to build a good model for these data—in terms of the activities of its downstream, multisensory units. For example [4], an exponential family harmonium (EFH) [3] trained on the activities of two populations of Gaussian-tuned, Poisson neurons (linear probabilistic population codes [2]) that tile their respective sensory spaces (visual and proprioceptive, e.g.) will learn to extract the “common cause” of these populations, encoding the stimulus in its hidden layer. In this case, the unisensory information available on a “trial” can be characterized by two means (best estimates) and two variances (inverse reliabilities); and the estimate extracted by the hidden units of the trained network is precisely the inverse-variance-weighted convex combination that primates appear in psychophysical studies to use.

Ecologically, however, the critical challenge is not typically to estimate the location of a static object, but to track the state of a dynamically changing environment. This task likewise requires reliability-weighted combination of information, in this case of the current sensory evidence and the current best estimate of the state given past information. But it is considerably more difficult,



since its solution requires learning a predictive model of the dynamics, which is not explicitly encoded in the sensory reports. In the case of Gaussian noise and linear dynamics (LDS), this recursive process is described by the Kalman filter, the parameters of which can be acquired with well-known iterative learning schemes. How the brain learns to solve this problem, however, is unknown.

Here we propose a neural model that accomplishes this task. We show that by adding recurrent connections to an EFH similar to that used in [4], the network can learn to estimate the state of a dynamical system. For concreteness, we consider the problem of tracking the dynamical state of the upper limb, a necessary computation for accurate and precise movement planning and control. In this case, the neural circuit corresponds to the posterior parietal cortex (PPC), which appears to subserve state estimation [10, 11]; and its inputs are taken to be a population of proprioceptive neurons. The network's performance can be quantified precisely by restricting our view to linear-Gaussian dynamics, where the filtering and learning problems have known optimal solutions (respectively, the Kalman filter and expectation-maximization, a maximum-likelihood algorithm). And indeed, performance approaches that optimum.

We then extend the network to controlled dynamical systems. Under the assumption that the controls are provided by motor cortex, these too are observed only noisily by PPC, in the form of efference copy, which the network must then learn to interpret as motor commands. State estimation is again close to optimal. In addition, the network is neurally plausible in both its representation of stimulus probabilities [2] and in the unsupervised learning procedure, which relies only on pairwise correlations between firing rates of connected neurons [12, 13]. Finally, the network makes two predictions about neural circuits that learn to perform state estimation: (1) During learning, position receptive fields will emerge before velocity receptive fields; or more generally, receptive fields will develop from lower- to higher-order states, especially when explicit information about the higher-order states is not in the inputs. (2) Filtering is implemented

by tuning to past positions (or more generally, lower-order states), rather than tuning directly to velocity (or more generally, higher-order states).

## Results

### *Network performance*

*The filtering problem.* We present results for an uncontrolled and a controlled dynamical system. For both, the basic dynamical system is second-order (position, velocity), discrete time, stochastic, and linear. The noise in state transitions is additive, white, and Gaussian. The “observation” at time  $t$  is the response  $(R_t^\theta)$  of a population of Poisson neurons, with Gaussian (bell-shaped) tuning curves that smoothly tile position. Since these neurons are taken to be reporting the proprioceptive sense, the “position” variable is the angle of a (single) joint,  $\Theta$ . For the controlled system, there is a second population of Poisson neurons—carrying the “efference copy,”  $R_t^u$ —that smoothly tiles a space of input torques,  $\mathbf{U}$ . Details appear in the methods section “Input-data generation.”

The task of tracking an object (estimation) is to provide, at time  $t$ , the best estimate of the location that can be computed from all the noisy observations from time 0 up to  $t$ . When current position depends on only a finite number  $n$  of past positions, this problem can be solved recursively: rather than retaining a history of all past observations, it is necessary to maintain only the current best estimate of the state (a vector whose dimension is set by  $n$ ), and the reliabilities of these estimates (a covariance matrix). By artful design of the system (see Methods), we have arranged for the optimal estimate at time  $t$  to be computable in closed form (see below). We emphasize that this computation does not approximate the firing statistics of the Poisson observations as Gaussian; see the section The optimal filtering distribution for details.

The task of learning is to acquire the parameters that make it possible to carry out the estimation task. These parameters correspond to a dynamical system (e.g., the state transition matrix and the covariance of the state transition noise), and a model of how the states of this system give rise to observations. In our network, however, the parameters that are learned directly are the synaptic connections (weights) and the bias of each unit's response function; the correspondence with the parameters of the dynamical system is not transparent (we explore it below).

*The network.* The central idea behind training our network, the “recurrent, exponential-family harmonium” (rEFH), is the choice of input data. In particular, each input vector consists of both the proprioceptive response to the current joint angle, and the activities of the hidden units at the previous time step. (For the case of a controlled dynamical system, the input vector also contains a noisy copy of the efferent motor command.) Biologically, this could be implemented via an additional population that simply reports, at a single time-step delay, the activities of the hidden units (Fig 1B, heavy black arrows; see also the section Cortical implementation below).

Conceptually, this choice of input data reflects the fact that filtering can be expressed as a “multisensory integration,” not between (e.g.) proprioceptive and visual inputs (cf. [4]), but between proprioceptive inputs and a running best estimate of the state. We hypothesize that, because the hidden units learn to extract all the information available in their inputs [9], the hidden vector at time  $t - 1$  will accumulate the information of the filtering distribution,

$p(\theta_{t-1} | r_0^\theta, \dots, r_{t-1}^\theta)$ . Then at the next time step, the network will “integrate” this information with

the current proprioceptive information about joint angle. (See S2 Text for a longer discussion of this point.)

*Experiments.* We therefore test our network by decoding, for all  $t$ , hand position at time  $t$ , from its hidden units. Rather than directly compare this estimate to the optimal estimate, which would provide no sense of scale, we compute error statistics for both. That is, we take the difference between the network's estimate and the true hand location; compute the mean and variance of this error across time steps (0 to  $T = 1000$ ) and trajectories ( $N_{traj} = 40$ ); and then compare these statistics for the rEFH and the optimum (OPT). We also compare error statistics from a "naïve" decoder (PROP) that simply decodes the current proprioceptive population, ignoring dynamics. It is the optimal decoder for data with no temporal dependencies.

The rEFH had to learn to solve the estimation problem, and in practice, learned solutions will always be somewhat suboptimal, because of finite, noisy data and a nonconvex problem space. Therefore, a perhaps more useful point of comparison is the set of the error statistics from another model that has been trained on the same data. In particular, it is possible in certain cases to derive optimal parameter-update equations for a learning procedure ("expectation-maximization," EM) that is guaranteed to reach at least local optima. Again by design of the dynamical system, and although the rEFH is not in theory limited to such data, such update equations are available (they are derived in S1 Text). We therefore generate error statistics for this model (EM), as well. We emphasize, however, that EM was given additional information not provided to the rEFH: the order of the dynamical system, as well as the parameters of the observation model. The latter includes the best estimate of the stimulus given the population response, and the reliability of that "observation"—whereas the rEFH had to learn how to infer these values from the population response itself. Since EM is sensitive to the initial (random) values of the parameters it is to estimate, we present results for the best model from 20 random restarts (see Methods); the same was done for the rEFH. To determine what order of dynamics the rEFH has learned, we also compare against lower-order models trained with EM. The order of these models is denoted with a superscript (e.g., EM<sup>2</sup>).

*Uncontrolled dynamical system.* The generative model for the data appears in Fig 1A.

Conceptually, the problem is to track the shoulder joint (Fig 1C). To encourage second-order behavior, the parameters of the system were chosen make it underdamped (as e.g. when the arm hangs downward and acts as a pendulum; see Methods), yielding trajectories like those shown in black in Fig 1D, and the proprioceptive responses shown in orange. The rEFH was trained as a density estimator on these responses and the recurrent activity of its own hidden units at the previous time step (Fig 1B).

Error statistics for the various decoders are shown in Fig 1E. Performance of the rEFH exceeds that of the naïve, purely sensory decoder (PROP), and approaches that of the optimum and the (second-order) EM-trained model ( $EM^2$ ). The rEFH also outperforms the first-order, EM-trained model,  $EM^1$ , showing that it has learned to keep track not just of past positions, but of past velocities as well. We explore below how it encodes position and velocity (Learned receptive fields and connectivity).

We chose the dynamics of the underlying stimulus because they let us clearly see that the rEFH can learn a second-order model; that is, it learns to track the lawful changes in velocity, as well as position—even though only position information was available at each time step. But the results are robust across various dynamical models. Fig 2 shows results for 36 different dynamical systems which were created by varying the oscillator's Fig 2A stiffness, Fig 2B damping, or Fig 2C moment of inertia (colors as in Fig 1E). For all of them, the rEFH outperforms the first-order model ( $EM^1$ ), and performs close to the second-order model ( $EM^2$ ).

*Controlled dynamical system.* We now consider a system with inputs. Whereas the uncontrolled dynamical system, above, corresponds to the case where the arm is moved only passively by external forces, the controlled system corresponds to the more general case of self-motion. Controls are issued to the muscles of the arm by motor cortex, but a copy of these efferent

signals is also fed back to posterior parietal cortex, Fig 3A [21]. This reference copy, being transmitted by a population of neurons, is assumed to be a noisy representation of the true control. Nor, presumably, is its role as a control signal explicitly given; rather, the network must learn, without supervision, to interpret it as such. This is precisely the learning problem faced by our network model (Fig 1B).

Realistic controls are correlated through time, so the control signal was given its own dynamics: a random walk with a very mild decay towards zero (see Methods). This resulted in trajectories like that in Fig 3E. The effect on angle can be seen in the corresponding trajectory of Fig 3C: here, the control is driving the trajectory increasingly negative (cf. the first and second trough) in spite of the damping and the restoring force. In general, since the control is a random walk rather than merely white noise, the changes it effects on the position trajectory tend to accumulate.

Mean squared errors (MSEs) for the various filters of the controlled system are shown in Fig 3D. The naïve model that ignores dynamics (PROP) is again the worst, as expected. Here the optimal EM-trained model is third-order ( $EM^3$ ), since the second-order dynamical system is driven by a control with first-order dynamics. Again the rEFH performs close to this model, and outperforms the best lower-order model ( $EM^2$ ). No trained model quite matches the true model's performance (OPT), which result appears to be robust (see error bars), and presumably owes to the shape of the objective function (e.g., the optimal solution may be separated from suboptimal local maxima by deep valleys of low likelihood solutions).

So the rEFH has learned a third-order system. However, this does not per se show that it has learned to use the efference-copy population; it might, for example, simply attribute all input to the system as white noise. To demonstrate that it does learn to use the controls, we compare it to the best state estimates that can be made without efference copy. To produce such

estimates, we fit a sixth filter, OBS, via regression, with full access to the state as well as the proprioceptive responses, but forced to assume zero control input. The resulting performance (Fig 3D, yellow bar) is clearly inferior to the other dynamical models.

Since the control signal is, like the state, only noisily observed by the network (via the efference copy), it is sensible to ask how well it can be decoded from the various models as well. And since the signal has its own (first-order) dynamics, it is possible for these models to make better estimates of the control at time  $t$  than can be made from the efference copy alone at  $t$ . Fig 3F shows that this is indeed the case. All the filters perform about equally well, in comparison with the non-dynamical decoding of the efference-copy population (“EfCp”), although the harmonium is slightly inferior.

*Distributions of performance* across initializations and hidden-layer size. One known limitation on the learning capacity of the harmonium is the number of hidden units: the network requires sufficient representational power in this vector to encode the cumulants of the input distribution—in this case, the filter distribution. To determine what network size is necessary for maximal performance, we test a series of networks with systematically increasing numbers of hidden units. Because, however, the number of recurrent units must equal the number of hidden units, the ratio of hidden units to input units (recurrent, proprioceptive, and efference-copy) is necessarily upper bounded at unity. This asymptote presumably diminishes the returns of additional hidden units, even beyond those limitations imposed by the learning algorithm or the difficulty of the filtering task.

For each of twelve sizes, we train 20 networks de novo, test them on a single fixed data set, and compute mean squared errors. The results are shown in Fig 4, where network sizes (abscissae) are given by the number of hidden units. For the uncontrolled network, Fig 4A, MSE of the best network (out of 20) clearly diminishes, asymptotically, with increasing numbers of hidden units.

Beyond about 180 hidden units, no improvement in MSE is produced. For the controlled network, which has more to learn, the effect is even more severe (Fig 4B): increasing the number of hidden units beyond about 180 results in decreasing performance for even the best-performing networks at each network size. This suggests a limit to the complexity of the dynamical systems learned by this architecture, or with this learning procedure (for details of which, see Methods).

On the other hand, rEFH learning appears to be more robust than EM learning, as can be seen in the box plots for the EM-based models (Fig 4). As with the rEFHs, 20 of each of the four EM-based models were trained from scratch, resulting in the distributions shown in light red and blue in Fig 4A and 4B, (the narrowness of some of these distributions results in some very thin boxes). Although the EM algorithm guarantees convergence, it is only to a local (rather than global) optimum; which optimum is determined by the (random) initial parameters. The lower order EM benchmarks (light red boxes; EM<sup>1</sup> in Fig 4A and EM<sup>2</sup> in Fig 4B) do indeed learn robustly, achieving nearly the same performance for all initializations. But the models with the true dynamical order (blue; EM<sup>2</sup> in Fig 4A and EM<sup>3</sup> in Fig 4B) exhibit a large performance distribution. These models are capable of outperforming the rEFH, but the runs that do are outliers. Thus, the large majority of the true-order EM-based models, for both the controlled and uncontrolled dynamical system, perform only about as well as their lower-order counterparts, which is inferior to all 20 of the 180-unit rEFH models. These rEFHs show comparatively little variation in performance—although that variance increases with the number of hidden units beyond 180.

How does the rEFH track the state? Optimal (or nearly optimal) position estimation for these dynamical systems requires tracking velocity and position, so we plot receptive fields (RFs) in position-velocity space. Now, for oscillatory dynamics, high speeds rarely co-occur with positions far from zero (equilibrium), which leaves the “corners” of such RFs empty. This



obscures the pattern of RFs and the corresponding state-estimation scheme learned by the rEFH. Therefore, for simplicity, we present results from a network trained on a third dynamical model (“nospring”): uncontrolled, and with no spring force (see Methods). (Similar results, albeit less clean, are observed in the corresponding analyses for oscillatory dynamics; see S3 Text in the supporting material.) In Fig 5A, the position-velocity receptive fields are plotted for all 225 hidden units of this rEFH, arranged in a  $15 \times 15$  grid. The ordinate of each subsquare corresponds to position (increasing from top to bottom), and the abscissa to velocity (increasing from left to right). The large majority of receptive fields are negatively sloped “stripes” in this space. Interestingly, they resemble in this the receptive fields of neurons in MSTd of a rhesus macaque trained to track moving stimuli [22]—although in that work there are positively-sloped stripes as well. Interpretation of these receptive fields is facilitated by an observation. If the velocity ( $\omega$ ) is roughly constant over  $n$  time steps of length  $\Delta$ , then:

$$\omega_{t-1} \approx \frac{\theta_t - \theta_{t-n\Delta}}{n\Delta}$$

$$\Rightarrow \theta_t \approx n\Delta\omega_{t-1} + \theta_{t-n\Delta}$$

the equation of a line in position-velocity space. Hence, such fields could be produced by neurons tuned simply for position at a delay, where the size of the delay ( $n\Delta$ ) determines the slope (negative because we have plotted position as increasing from top to bottom, to match the corresponding figure in [22]), and the “preferred” position determines the y-intercept. The equation is exact for  $n = 1$ ; but to the extent that velocity is not constant, the receptive fields will be diminished—as seen in the more irregular and faded character of receptive fields with greater slopes.

We therefore re-plot the receptive fields as a function of position only, but each at the time delay that maximizes mutual information between position and that hidden unit’s response (Fig 5C).

Units are ordered by preferred position (whereas units in Fig 5A were ordered by time delay). The resulting position tuning curves appear to tile space uniformly, with roughly constant receptive-field widths, suggesting that this is a concise description of the tuning curves that captures the computation being performed. As a final method of verification, we use these lagged-position receptive fields to generate idealized position-velocity receptive fields (Fig 5B; see Tuning analysis for details.) The match with Fig 5A is apparent.

This result is, perhaps, unexpected. It is possible to represent (an estimate of) the current state of a second-order dynamical system compactly by encoding current position and the current velocity. But it is also possible to represent it—seemingly less efficiently—in terms of the past positions alone, with the weight on each position decaying exponentially as a function of the number of time steps into the past. The rEFH appears to have learned a representation of this second type.

To determine if the weighting function applied to past positions by the rEFH does indeed correspond to such a scheme, we examine the distribution of “preferred” lags across hidden units (Fig 5D, center panel). Unsurprisingly, most units are tuned to the recent past, with an apparently monotonic decline into the past. Superimposed is the autocorrelation of the dynamical system on which the network was trained (see Methods), normalized to have the same integral as the histogram. Evidently, the distribution of lags is well tuned to the dynamics. To confirm the robustness of this finding, we trained four new rEFHs on four different dynamical systems, identical to the one under discussion up to the damping coefficient (frictional force). From left to right in Fig 5D, the dynamical systems were increasingly damped, resulting in longer autocorrelations (thick black lines). The temporal tuning of the rEFHs trained on these dynamical systems appears well matched in each case.

*Responsiveness to instantaneous reliability.* Thus, the rEFH tracks stimuli by encoding its position at various lags, with the number of units assigned to each lag decreasing exponentially with distance into the past, according to the autocorrelation of the signal. What is nevertheless not clear from Fig 5D is whether the rEFH's weighting of past positions takes into account the instantaneous reliability of the proprioceptive encoding of joint angle. In our models, as (presumably) in the brain, instantaneous reliability varies because the proprioceptive report of joint angle is corrupted by Poisson noise. For Poisson spiking, the reliability—i.e., the inverse variance of the posterior distribution over joint location, conditioned on the spiking of all proprioceptive neurons—is proportional to the total number of spikes produced by the population at that time step. In the results discussed above, we also increased the fluctuations in this number by additionally varying the “gain” of the proprioceptive population, i.e., the single parameter that sets the height of all the tuning curves (see Input-data generation). This is meant to model other random changes in the reliability of the proprioceptive report. Thus the optimal weighting of past position information, although on average an exponentially decaying function of time delay, will at any particular moment vary as a function of the recent, unpredictable, history of reliabilities: higher weights should be assigned to those time steps when the proprioceptive units had a collectively higher average firing rate, and vice versa. A network that ignores such fluctuations will perform well, but suboptimally, perhaps explaining the (small) discrepancy between the MSEs for the rEFH and  $EM^2$  seen in Fig 1E. Here we show that the rEFH is indeed sensitive to instantaneous reliability. We retest the network of the section Uncontrolled dynamical system on noiseless sensory data, i.e., using the mean spike counts of the proprioceptive neurons rather than Poisson samples drawn from those means. This allows us to set the total spike count essentially directly. In this noiseless test, a higher total spike count does not correspond to a more reliable signal: the signal is perfectly reliable for all spike counts. Hence for any total spike count, minimal (zero) error could be achieved by a “filter” that relied entirely on the current sensory input. Nevertheless, the optimal filter for the data on which

the rEFH was trained, viz., OPT, will not rely entirely upon this current sensory information, but rather will weight it in proportion to the total spike count. In consequence, OPT will achieve lower MSEs on data with greater total spike counts, since for such data it will lean more heavily on the perfect sensory information. This is the pattern observed in Fig 6A. If the rEFH is also treating total spike count correctly—i.e., if it is properly sensitive to instantaneous reliability—its MSEs will exhibit a similar pattern. And indeed, they do (Fig 6B).

Here we emphasize again that, for the optimal filter (as well as OBS and EM<sup>n</sup>), we provided the transformation from total spike count to sensory reliability, whereas the the rEFH had to learn this transformation. Likewise, in engineered solutions to tracking problems, the Kalman filter is usually simplified to learn a single, average reliability for all time. We have demonstrated that the rEFH is not similarly limited, since it can learn to use instantaneous indicators of reliability if they are present in the observations.

*Emergence of receptive fields.* Since velocity is not reported directly by the sensory (“proprioceptive”) population, the network will not immediately develop tuning for it. Fig 7 illustrates its emergence across training trials. (We return here to the “no-spring” model for comparison with Fig 5A.) Since position information is a useful “feature” for explaining the proprioceptive inputs, the hidden units learn to extract it after just 100 batches of training (Fig 7A). But information that is in the hidden units will also appear in the input, at a one-step time delay, via the recurrent units (see Fig 1B). So at this point in training, the input contains information about both the current position (in the proprioceptive population) and about the past position (in the recurrent population). This makes extraction of velocity information possible. It is useful because the stimuli obey second-order dynamics: knowing the relationship between past and present position allows each to provide information about the other, yielding overall superior estimates.

And indeed, by 200 batches of learning, some velocity tuning appears, evidenced by the sloping of receptive fields in position-velocity space, Fig 7B.

But these units look back no more than a few steps in time. By 1000 batches (Fig 7C), strong velocity tuning is evident, although the full distribution of lags (slopes) has yet to emerge (cf. Fig 5A, which is after some 100,000 batches of training).

*Organization of the learned weight matrix.* The network learns to model the dynamics by making changes to the synaptic connection strengths, summarized by the weight matrices  $W_{prop}$  (sensory to hidden) and  $W_{fb}$  (recurrent to hidden). To understand better the mechanism of the network we examine these matrices (Fig 8), again for the “no-spring” model. (The corresponding figure for the non-zero stiffness model, slightly more difficult to interpret, is shown in S3 Text.) In the arbitrarily ordered form in which they are learned, they are difficult to decipher, but interesting patterns emerge when they are reordered by the parameters of the receptive fields. The reordering is applied to both the hidden and the recurrent units. (The original, topographical ordering of the sensory units is retained.) First, reordering by the hidden units’ preferred stimulus angles (“PA”), Fig 8A, reveals that the difference in PA between two hidden units dictates the sign of their connection. Hidden units with similar PAs have positive connections, and units with “out of phase” (recall that the stimulus is a circular variable) PAs have negative connections. This results from the continuity of the stimulus trajectories: the network “expects” the stimulus to move from any given position (encoded in the recurrent units) to a nearby one (encoded in the hidden units). Second, we reorder the units by “preferred lag,”  $\tau$  — i.e., the time delay at which mutual information between sensory input and hidden-unit activity is maximal (Fig 8B). Again, units with similar  $\tau$  are preferentially wired together.

### *Cortical implementation*

More than one cortical area is thought to subserve object tracking. Since we have in this study focused on the task of tracking one's own limbs, we consider posterior parietal cortex (PPC), which is thought to be responsible for this task [10, 11]. The computation may well be distributed across the PPC, but we focus on just one that has been particularly implicated [11], Brodmann Area 5. Our aim is to show that our neural network and its learning scheme are consistent with what is known about the connectivity of Area 5, both inter laminar and interareal. In particular, we consider its connections with the primary motor area (M1) and primary somatosensory cortex (S1). Our proposed implementation is speculative and not the only one possible; e.g., we identify the “recurrent” units with another layer of Area 5, but they might alternatively correspond to another area of PPC.

Fig 9A summarizes the training procedure from an algorithmic perspective (see Methods for details). In Fig 9B, as in Fig 9A, input comes from two sources. Feedforward, proprioceptive input ( $R_t^\theta$ ) from primary somatosensory cortex, S1 (especially BA3a), projects to layer IV [23]. A copy of the efferent command ( $R_t^u$ ) feeds back from M1 to layer I of Area 5 [23]. Layer II/III of Area 5 in turn projects forward to M1 [24]. Layer I is not believed to contain cell bodies [25], so we take these to be the terminal branches of the apical dendrites of layer II/III cells (which are also lightly labeled by anterograde tracers injected in M1 [23]). Within Area 5, we propose that the temporally delayed recurrency ( $Z_{t-1}$ ) of the rEFH is provided by the loop from layer II/III down to VI, then up to V, before modulating the activity of layer II/III neurons, consistent with the anatomy of Area 5 [25]. Layer IV and III, as well as V and III, also have reciprocal connections [25], as required for the rEFH training procedure. The latter loop has in fact been hypothesized to give rise to rhythmic activity in rat parietal cortex [26].

According to the learning and filtering schemes of our model, the temporal flow of information is as follows. Sensory input  $(r_t^\theta)$  and efference copy  $(r_t^u)$  arrive at, respectively, layer IV of BA5 and the feedback layer (presumably VI) of M1. At the same time, a “copy” (which could be any information-preserving transformation) of activity from layer II/III of BA5 ( $z_{t-1}$ ) passes down to layer V. Next, the spiking in these layers (M1 layer VI, BA5 layer IV, BA5 layer V) drives spiking ( $z_t$ ) in BA5 layer II/III. These responses encode, according to the model, the optimal estimate of the limb, and this information will ultimately become the temporally delayed recurrent activities identified above. For learning, however, it is also necessary that this activity drive spiking in M1 ( $\hat{r}_t^u$ ), BA5 layer IV ( $\hat{r}_t^y$ ), and BA5 layer V ( $\hat{z}_t$ ), through the reciprocal connectivity lately noted. A “copy” of the layer II/III activity ( $z_t$ ) is simultaneously propagated down to layer VI. Lastly, the activities in M1, BA5 layer IV, and BA5 layer V again drive activity ( $\hat{z}_t$ ) in BA5 layer II/III. At the same time, the “copy” of layer II/III activity ( $z_t$ ) is communicated up to layer V.

## Discussion

### *Summary of results*

We have shown that a neural network (the “rEFH”) with a biologically plausible architecture and synaptic-plasticity rule can learn to track moving stimuli, in the sense that its downstream (“hidden”) units learn to encode (nearly) the most accurate estimate of stimulus location possible, given the entire history of incoming sensory information (Figs 1 and 2). This requires learning a model of the stimulus dynamics. This is (as far as we know) the first biologically plausible model that has been shown to learn to solve this task. Moreover, the network learns the reliability of the sensory signal: the trained network leans more heavily on the internal model when the sensory signal is less reliable, and more heavily on the sensory signal when it is more reliable (Fig 6).

We are particularly interested in tracking the state of one’s own limbs. Here, additional information about stimulus location is thought to be available in the form of a “copy,” relayed to the posterior parietal cortex, of the efferent motor command [21]. And indeed, when such signals are available to our network, it learns to make use of them appropriately to track the arm more precisely—in spite of the fact that none of the incoming signals is “labeled” according to its role (Fig 3). Although an expectation-maximization (EM) algorithm can sometimes learn a Kalman filter that noticeably outperforms the best rEFH on these data, it usually does not (Fig 4). That is, learning in the rEFH is more robust than EM in the sense that the variance in performance across models trained de novo is smaller, albeit at the price of a bias towards worse models. Finally, and surprisingly, the downstream neurons of the trained network track a moving stimulus by encoding its position at various time lags (Fig 5).

### **Related work**

The earliest implementation of dynamical state estimation (“filtering”) in neural architecture comes from Rao and Ballard [27]. Their model, like ours, assigns a central role to recurrent connections, but as predictive coders rather than simply delayed copies of previous neural states. Likewise, the network connectivity is acquired with an unsupervised and local learning rule, a variant on EM. However, the authors do not train their network on moving objects or moving images, presumably because convergence of the neural state under their learning scheme is slow compared with any plausible stimulus dynamics. Instead, the connectivity is acquired on static images. Performance on state-estimation tasks is not tested.

Several groups have hand wired neural networks to act as state estimators [7, 8, 28]. Although these papers do not address our central concern, the learning problem, it is nevertheless useful to compare the resulting architectures with our rEFH. For example, Beck and colleagues constructed a neural network to implement the Kalman filter on linear probabilistic population



codes, as in this work, and showed its performance (measured in information loss) to be nearly optimal. From analytical considerations, the authors showed that the required operations on neural firing rates are weighted summation (as in our network) and a quadratic operation (that acts like a divisive normalization in the steady state). In our rEFH, on the other hand, the only nonlinearities are element-wise: interaction between inputs is always in the form of a weighted sum. That the rEFH can nevertheless filter (nearly) optimally is possible because we do not require, as they do, that the output population encode information in the same way as the inputs (sc., that the posterior distribution over the stimulus have linear sufficient statistics; see S4 Text). This critical difference provides the basis for an experimental discrimination between the respective models. Likewise, filters have been hand wired into attractor networks [28] and spike-based (rather than rate-based) networks [8]. The latter in particular argues that the precise arrival time of spikes contains information about the stimulus, rather than the average rate across time, as in in our model.

An approach that does include learning comes from Huys, Zemel, Natarajan, and Dayan [29, 30]. The authors formulate the problem in terms very similar to ours, but they allow more general dynamical systems generated by Gaussian processes, and the basic unit of information is spikes rather than spike counts (although approximations that ignore precise arrival times lose little information [29]). The most significant difference with our work is that the authors learn the parameters of their network with a supervised, non-local rule, which they do not consider to be a biological mechanism. But again the comparison is instructive. We are able to formulate an unsupervised rule because we approach the filtering problem indirectly: Natarajan and colleagues require the posterior distribution, conditioned on hidden-unit activities, to be factorizable over hidden-unit spikes (so that a third layer can consider those spikes separately), and then force it to match the true filtering distribution by directly descending the KL divergence between them [30]. We, on the other hand, force the network to be a good model of its incoming

data—which, when some of those data are past hidden-unit activities, achieves the same end. In the machine-learning literature, Hinton and colleagues have proposed three variants on a theme quite similar to ours [31–33], although different in important ways. Most importantly, in all three, the past hidden-unit activities are treated by the learning rule as (fixed) biases rather than as input data; i.e., they cannot be modified during the “down pass” of contrastive-divergence training. That these activities ought to be treated as data, we argue more rigorously in a forthcoming work.

The earliest variant [31], the “spiking Boltzmann machine,” is, like ours, a temporal extension of the restricted Boltzmann machine that is trained with the contrastive-divergence rule. Hidden units are directly influenced by past hidden-unit activities, as with the rEFH, but possibly from temporal distances  $\tau$  that are greater than one time step (contra the rEFH). However, the weights from a particular “past” hidden unit at various delays (e.g., from  $z_{t-n\tau}^i, n \in \{1, 2, 3, \dots\}$  to  $z_t^j$  are constrained to be identical up to a fixed (not learned) exponential decay.

The motivation was to model the influence of past spikes in a biologically plausible way: Whereas in our rEFH, the (one-time-step delayed) past hidden activities are maintained in a separate population of neurons (Fig 9), in the “spiking Boltzmann machine” their effect on current hidden units is interpreted simply as the decaying influence of their original arrival. This makes it plausible, unlike in the rEFH, to include influences at delays greater than one time step. On the other hand, it necessitates treating those effects as biases rather than data. It is difficult to judge the limitations this imposes on the model, since the authors do not quantify its performance. However, they do investigate more thoroughly performance of a similar, but more powerful network. The “temporal restricted Boltzmann machine” (TRBM) [32] is a spiking Boltzmann machine without the constraint that the weights decay exponentially backwards in time; instead, they are learned freely and independently for all time. The order of the dynamical

system that can be learned by this network turns out, unlike ours, to be tied to  $\tau$ : TRBMs with  $\tau = 1$  (like the rEFH) can learn only random-walk behavior (first-order dynamics) [33]. This can (presumably) be overcome by including connections back as many time steps as the order of the system to be learned, but it is not obvious what biological mechanism could maintain copies of past activities at distant lags, or determine a priori how many such lags to maintain. The same authors show that this problem can be alleviated with a variant architecture, the “recurrent temporal RBM” (RTRBM) [33], but it requires a non-causal learning rule (backpropagation through time), again making it a poor model for neural function. For neither model do the authors precisely quantify its filtering performance; we do in a forthcoming study.

### *Implications of the model*

Our simulations demonstrate three things: First, the rEFH is capable of learning to “track” moving stimuli, i.e. to estimate their dynamical state, and nearly as well as an optimal algorithm, as has been seen behaviorally in humans [34]. In fact, the network learns to encode the full posterior distribution over the stimulus, rather than just its peak: although we did not show it directly, it must, since the variance of this (Gaussian) distribution is required to combine properly the previous best estimate with the current sensory information. And rather than relying on a fixed estimate of sensory reliability, the network learns to take into account instantaneous changes in it (Fig 6).

Second, the network does not require a special architecture or ad hoc modifications. It is, rather, identical, up to the choice of input populations, to the network and learning rule in our previous work [4]. Thus, if the input populations are proprioceptive and recurrent units, it will learn to estimate dynamical state; if they also include efference copy, it will learn the influence of motor commands on stimulus dynamics. If they are proprioceptive and visual reports of a common stimulus, it will learn to perform multisensory integration; if a gaze-angle-reporting population is also present, to transform the visual signal by that angle before integrating (“coordinate

transformations”); if the stimulus distribution is non-uniform, to encode that distribution [4]. (We have shown elsewhere, in terms of information theory, why this is the case [9]. For further discussion of the relationship between the static and dynamical computations, see S2 Text.) Thus, the network provides a very general model for posterior parietal cortex, where some combination of all of these signals is often present. Third, the model makes some predictions about the encoding scheme, receptive fields, and connectivity of cortical areas that track objects. As with all models, we take certain elements of ours to be essential and others to be adventitious. That learning in posterior-parietal circuits can be well described as a form of latent-variable density estimation, for example, is central to our theory; but the precise form of the learning rule (“one-step contrastive divergence”), although plausible, is not. Our theory requires that sensory neurons encode distributions over stimulus position, but the representation scheme need not be probabilistic population codes of the Pougetian variety [2]. Here we list three predictions that do follow from essential aspects of the network.

1. The network learns to track by encoding past positions. This is a non-obvious scheme (it is not, e.g., the one used by the Kalman filter) and apparently results from the fact that only position information is reported by the sensory afferents. It is possible that such receptive fields (Fig 5A) are in fact found in MSTd of monkeys that have been trained to track moving objects [22]. Now, in many circuits, velocity is detected at early stages. But even when velocity is directly reported by the inputs to an rEFH, tuning to past positions still appears, albeit with lower prevalence (see S3 Text). More generally, we predict that higher derivatives (e.g., acceleration), especially those not directly available in sensory input, will be encoded via delayed, lower derivatives (e.g., velocities)— as long as those higher-order states have lawful dynamics.
2. During learning, receptive fields for position emerge before those for velocity. This is a necessary consequence of density estimation on recurrent units. A similar proviso

attaches: where velocity information is directly reported, it is acceleration-coding that will emerge over time.

3. The use of delayed, feedback connections in neural circuits is a mechanism for learning dynamical properties of stimuli. Under this prediction, primary sensory areas that process information with very little temporal structure—e.g., smell—will lack the dense feedback found in, e.g., visual areas. Alternatively, the recurrency might be identified with interlaminar, rather than interareal, structure, as we have hypothesized (Fig 9B)—which would explain why piriform cortex only needs three layers.

#### *Neural computation in posterior parietal cortex*

More generally, our investigation was motivated by two main ideas. The first is that populations of neurons, in virtue of their natural variability, encode probability distributions over stimuli (rather than point estimates) [1, 2]. Encoding certainty or “reliability” is a necessity for optimal integration of dynamic sensory information, since it determines the relative weight given to (a) current sensory information and (b) the prediction of the internal model. But rather than explicitly encoding the reliability of the stimulus location—e.g., via neurons that are “tuned” to reliability, as other neurons are tuned to location itself—this reliability is identified with the inverse variance of the posterior distribution over the stimulus,  $p(s|r)$ , conditioned on the population activity [2]. This distribution arises as a natural consequence of the (putative) fact that neural responses are noisy, and can therefore be characterized by a likelihood,  $p(r|s)$ [1]. If reliability were not encoded this way, our learning scheme would not work: it would have no way of knowing what to do with those reliabilities, which would be to it indistinguishable from (e.g.) the location of another stimulus.

The second idea is that higher sensory areas, like posterior parietal cortex and MSTd, can encode more precise distributions over the location (e.g.) of a stimulus than that provided by their sensory afferents at any given moment in time. This is due, essentially, to the continuity of the physical world: at successive moments in time, objects tend to remain near their previous positions. More precise localizations can consequently be achieved by a form of averaging that, because objects do move, accounts for the predictable changes in position from moment to moment. This requires learning a model of those predictable changes. The rich statistical structure of the sensory afferents—including efference copy of motor commands that may be influencing the evolution of the stimulus to be tracked, as when tracking one's own limbs—makes it possible to learn the model from those inputs alone. This unsupervised learning is a much more efficient approach than trying to use the few bits of information that may be available in the form of reward: very few rewards can be reaped before an animal can control its own limbs. In the special case of linear dynamics and Gaussian noise, these two problems—learning a dynamical model, and filtering in that model—have known algorithmic solutions: an expectation-maximization algorithm and the Kalman filter, respectively. Rather than try to map operations on vectors and matrices directly onto neural activity and learning rules, we have taken a more general approach, showing how a rather general neural-network architecture that tries to build good models for its inputs can learn to solve the problem, if those inputs are suitably chosen: temporally delayed recurrent activity from downstream units must be among the inputs. Our network learns by a local, Hebbian rule operating on spike-count correlations, although it remains to relate these to more specific biological learning rules, like STDP.

## **Methods**

Notation is standard: capital letters for random variables, lowercase for their realizations; boldfaced font for vectors, italic for scalars. Capitalized italics are also used for matrices (context distinguishes them from random scalars).

### *Input-data generation*

We describe the most general dynamical system and observation model to be learned: a controlled, second-order, discrete-time, stochastic, linear dynamical system, whose “observations” or outputs come in the form of linear probabilistic population codes [2]; cf. Fig 3A. The uncontrolled model of the section Uncontrolled dynamical system (Fig 1A) is a special case (see below). We interpret the plant to be a rotational joint, so distance is in units of radians; and the control to be a torque, hence in Joules/radian.

The primary rationale for our choice of dynamics and observation model was to show what kinds of computational issues the recurrent, exponential-family harmonium (rEFH) can overcome— issues which it must overcome if it is to be a good model for the way cortex learns to solve the problem. In particular, it might appear that the rEFH can learn relationships only between its current inputs and the previous ones, since its recurrent inputs are from the previous time step only (see Fig 9A). Therefore, we let the inputs report position only, but make the (hidden) dynamics second-order: velocity, as well as position, depends on previous position and velocity. If the rEFH can learn to associate only current and previous inputs, it can learn only first-order dynamics from these data. Furthermore, to clearly distinguish models that have learned second-order dynamics from those that have learned only a first-order approximation, we let the true dynamics be a (damped) oscillator (first-order systems cannot oscillate). Although the demonstration is in terms of positions and velocities, the point is more general: if the rEFH can learn second-order dynamics from position reports, it can learn higher temporal dynamics from lower-order data more generally.

The controlled, single-joint limb obeys:

$$p(\theta_{t+1} | \theta_t, u_t) = N(A\theta_t + bu_t + \mu_o, \Sigma_o) \quad (1)$$

where the vector random variable  $\theta_t$  consists of angle and angular velocity. The control signal (torque) has itself first-order dynamics:

$$p(u_{t+1} | u_t) = N(\alpha u_t + \mu_u, \sigma_u^2) \quad (2)$$

making the combined system third-order. The initial state and control are also normally distributed:

$$p(\theta_o, u_o) = N(\nu_o, \Upsilon_o) \quad (3)$$

The current (time t) joint position and control are noisily encoded in the spike counts of populations of neurons, whose Gaussian-shaped tuning curves (fi) smoothly tile their respective spaces, proprioceptive (angle) and control (torque). Spike counts are drawn from (conditionally) independent Poisson distributions:

$$p(r_t^o | \theta_t, g_t^\theta) = \prod_i Pois[r_{i,t}^o | g_t^o f_i(C\theta_t)], \quad p(r_t^u | u_t, g_t^u) = \prod_i Pois[r_{i,t}^u | g_t^u f_i(hr_t)] \quad (4)$$

with  $C = [1 \ 0]$  and  $h = 1$ . Here the  $g_t$  are “gains,” scaling factors for the mean spike count [2, 4]. Because the signal-to-noise ratio increases with mean for Poisson random variables, these gains essentially scale (linearly) the reliability of each population. Therefore, in order to model instant-to-instant changes in sensory reliability, the gains of each population were chosen independently and uniformly:

$$p(g_t^\theta) = u(6.4, 9.6), \quad p(g_t^u) = u(6.4, 9.6). \quad (5)$$

Since the discrete time interval for a single draw from Eq. 4 is 0.05 s (see below), these gains correspond to maximal firing rates between 130 and 192 spikes/second, reasonable rates for



neurons in cortex. The joint distribution of the states, controls, their observations, and the gains is the product of Eqs 1–5, multiplied across all time.

In accordance with the broad tuning of higher sensory areas, the “standard deviation,”  $\sigma_{tc}$ , of the tuning curves,

$$f_i(x) = \exp\left\{-\frac{(x - \xi_i)^2}{2\sigma_{tc}^2}\right\},$$

was chosen so that the full-width at half maximum is one-sixth of the space of feasible joint angles/torques, for all preferred stimuli  $\xi_i$ . However, joints and torques can in fact leave these “feasible spaces”: Although the system was designed to be stable (eigenvalues of the state-transition matrix are within the unit circle), trajectories are nevertheless unbounded, since the input noise is unbounded (normally distributed). We chose not to impose hard joint and torque limits, because this would make the dynamics nonlinear, vitiating the optimality calculations. Instead, stimuli beyond the feasible space simply “wrap” onto the opposite side of encoding space; that is, each population tiles its corresponding stimulus modulo the length of its feasible space.

But for the dynamical systems on which model performance was tested, parameters were chosen to make wrapping unlikely (but cf. the “no-spring” model described below). In particular, we used the discrete-time approximation to a damped harmonic oscillator, i.e.,

$$m\ddot{\theta} + c\dot{\theta} + k\theta = u :$$

$$A = \begin{bmatrix} 1 & \Delta \\ -\frac{k}{m}\Delta & 1 - \frac{c}{m}\Delta \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ \frac{\Delta}{m} \end{bmatrix},$$

with moment of inertia  $m = 5 \text{ J}\cdot\text{s}^2 / \text{rad}^2$ , viscous damping  $c = 0.25 \text{ J}\cdot\text{s} / \text{rad}^2$ , ideal-spring stiffness  $k = 3 \text{ J} / \text{rad}^2$ , and sampling interval  $\Delta = 0.05 \text{ s}$ . This makes the system stable and under-damped (oscillatory). The control decay,  $\alpha$ , in Eq 2 was set to 0.9994, making the dynamics close to a random walk, but mildly decaying towards zero.

These parameters and the noise variances were chosen so that the system could not be well approximated by a lower-order one—i.e., so that the uncontrolled and controlled systems were “truly” second- and third-order (respectively). This was accomplished by ensuring that the Hankel singular values [14] for the system, with output matrix  $C = [1 \ 0]$  and input matrix set by the noise variances, were within one order of magnitude of each other; that is, ensuring that the transfer function from noise to joint angle had roughly equal power in all modes. For the uncontrolled system, this was achieved with  $\Sigma_o = \text{diag}([5_E - 7, 5_E - 5])$ ; for the controlled system,  $\Sigma_o = \text{diag}([5_E - 5, 1_E - 6])$  and  $\sigma_u^2 = 7.5_E - 4$ . While this last choice of noise is large enough to ensure that the control’s contribution to the dynamics is significant, it is also small enough to keep wrapping rare. This facilitates the comparison between the benchmark models (see below), which are acquired from non-wrapped trajectories, and the rEFH, which learns from sensory inputs with periodic tuning curves. That is, for fast enough trajectories on a circle, the dynamics would no longer be locally linear, and the learning and filtering tasks no longer comparable.

The only other difference between the uncontrolled and controlled dynamical systems was that the former had, of course, no control signal (or simply  $b = 0$ ) and no control observations (efference copy). For all models, the bias terms were set to zero:  $\mu_\theta = 0$  and  $\mu_u = 0$ . The initial positions for all trajectories were drawn from a uniform distribution across joint space (shoulder  $\theta \in [-\pi/3, \pi/3]$  radians; Fig 1C), up to a margin of 0.05 radians from the joint limits (to

discourage state transitions out of the feasible space); for EM learning (see below), this was treated as an infinite-covariance Gaussian centered in the middle of joint space. The initial velocity and initial control were normally distributed very tightly about zero, with a standard deviation of  $\sqrt{5_E} - 5$  (rad/s and J/rad, resp.). Hence  $v_o = 0, Y_o = \text{diag}([\infty, 5_E - 10])$ . The range (modulus) of feasible controls is  $u \in [-1.25, 1.25] J / \text{rad}$ .

For the receptive-field (RF) analyses, we used a third dynamical system. In the harmonic oscillator, whether driven or undriven, the non-zero stiffness ( $k$  above) couples velocity to position, making high speeds and far-from-zero positions unlikely to co-occur. This makes the RF analysis unreliable in the “corners” of position-velocity space, and the overall velocity-encoding harder to interpret. For the analyses presented in Figs 5, 7 and 8, therefore, we trained a (third) rEFH on a simplified version (“no-spring”) of the uncontrolled dynamics, setting the spring constant to zero (eliminating oscillations). To encourage full exploration of the space, the variance of the state-transition noise was also increased by a factor of 50. The more and less auto-correlated variants of Fig 5D were created by simply scaling up or down the damping coefficient: from left to right,  $c = 0.25/4, 0.25/2, 0.25, 0.25 * 2, 0.25 * 4$ . For completeness, we nevertheless include, in the Supplement, the harder-to-interpret RF analyses for the rEFH trained on the (undriven) harmonic oscillator (S3 Text).

### *The recurrent, exponential-family harmonium (rEFH)*

The network is very similar to that in [4], but we repeat the description here briefly. The harmonium is a generalization of the restricted Boltzmann machine (RBM) beyond Bernoulli units to other random variables in the exponential family [3]. That is, it is a two-layer network with full interlayer connections and no intralayer connections, which can be thought of as a Markov random field (undirected graphical model) or as a neural network. In our implementation (see Figs 1B and 3B), hidden units (turquoise,  $Z_t$ ) and recurrent units (dark turquoise,  $Z_{t-1}$ ) are

binary (spike/no spike), and the “proprioceptive” (orange,  $R_t^o$ ) and “efference-copy” (purple,  $R_t^u$ ) populations are non-negative integers (spike counts). For all networks, the number of recurrent units is the same as the number of downstream or “hidden” units, because recurrent units at time  $t$  carry the activities of the hidden units at time  $t - 1$ —making the harmonium recurrent through time (rEFH). We chose  $N_{\text{hid}} = N_{\text{recurrent}} = 240$  for the network trained on the uncontrolled system, and  $N_{\text{hid}} = N_{\text{recurrent}} = 180$  for the controlled system. We used fifteen proprioceptive units ( $N_{\text{prop}}$ ) and, for the network trained on the controlled system, fifteen efference-copy units ( $N_{\text{efcp}}$ ), so the total number of “observed” (or “input”) variables was  $255 = N_{\text{recurrent}} + N_{\text{prop}}$  for the uncontrolled model and  $210 = N_{\text{recurrent}} + N_{\text{prop}} + N_{\text{efcp}}$  for the controlled model.

During training and testing, the layers of the rEFH reciprocally drive each other, yielding samples from the following distributions:

$$Z_t \sim q(z_t | z_{t-1}, r_t^\theta, r_t^u) = \prod_i^{N_{\text{hid}}} \text{Bern} \left[ \{z_t\}_i \mid \sigma \left( \{W_{fb} z_{t-1} + W_{prop} r_t^\theta + W_{ctrl} r_t^r + b_{hid}\}_i \right) \right] \quad (6a)$$

$$Z_{t-1} \sim q(z_{t-1} | z_t) = \prod_i^{N_{\text{hid}}} \text{Bern} \left[ \{z_{t-1}\}_i \mid \sigma \left( \{W_{fb}^T z_t + b_{fb}\}_i \right) \right] \quad (6b)$$

$$R_t^\theta \sim q(r_t^\theta | z_t) = \prod_i^{N_{\text{prop}}} \text{Pois} \left[ \{r_t^\theta\}_i \mid \exp \left( \{W_{prop}^T z_t + b_{prop}\}_i \right) \right] \quad (6c)$$

$$R_t^u \sim q(r_t^u | z_t) = \prod_i^{N_{\text{efcp}}} \text{Pois} \left[ \{r_t^u\}_i \mid \exp \left( \{W_{efcp}^T z_t + b_{efcp}\}_i \right) \right] \quad (6d)$$

which corresponds to Gibbs sampling from the joint distribution represented by the harmonium,

$q(z_t, z_{t-1}, r_t^\theta, r_t^u; W, b)$ . The letter  $q$  is used for the probability density function assigned by the rEFH to distinguish it from the true distribution over the observed variables,  $p(r_t^\theta, r_t^u)$ .

Here the notation  $\{x\}_i$  means the  $i^{\text{th}}$  element of the vector  $x$ ; the matrices  $W$  and vectors  $b$  are the synaptic connection strengths (“weights”) and biases, respectively; and the neural nonlinearities,

the logistic ( $\sigma(x) = 1 / (1 + e^x)$ ) and exponential functions, were chosen to produce means for each distribution that are in the appropriate interval ( $[0, 1]$  and  $^{\circ} +$ , resp). The entire procedure is depicted graphically in Fig 9A.

*Training.* Although the ultimate goal of training is to make the network able to solve the filtering problem, this is achieved indirectly by making the harmonium a good model for the data on which it was trained. That is, the harmonium should assign probability ( $q$ ) to the observed data ( $Y$ ) that matches the probability with which they actually appear ( $p$ ); in short, the goal is to achieve:

$$q(y; W, b) = p(y) \quad (7)$$

equality between the “model distribution” and “data distribution,” by adjustment of the weights and biases of the network. In our case,  $Y = [Y_0, \dots, Y_T]$ , a collection of observations across time, where intuitively the observations at time  $t$  are the responses of the proprioceptive and efference-copy populations,  $Y_t = [R_t^\theta, R_t^u]$ . However, these random variables are not independent across time; that is  $p(r_0^\theta, r_0^u, \dots, r_T^\theta, r_T^u) \neq \prod_t p(r_t^\theta, r_t^u)$ . In order, then, to make possible incremental training—weight changes without first collecting population responses for all time,  $[0, \dots, T]$ —we train on the augmented observation vector:

$$Y_t = [Z_{t-1}, R_t^\theta, R_t^u], \quad (8)$$

where  $Z_{t-1}$  are the hidden-unit activities at the previous time step. Intuitively, the addition of these recurrent activities renders the data independent because they recursively accumulate all the information contained in their inputs [9].

Weight changes are made proportional to the approximate gradient of a function (“onestep contrastive divergence,” CD<sub>1</sub>) that has Eq 7 at its minimum [12, 13]. In exponential family harmoniums, following this gradient is particularly simple: Stimuli in the world drive the input populations ( $y$ , Eq 4), which drive the hidden units ( $z$ , Eq 6a), which reciprocally drive the input populations ( $\hat{y}$ , Eqs 6b, 6c and 6d), which drive the hidden units once more ( $\hat{z}$ , Eq 6a); after which parameters are changed according to:

$$\Delta W \propto yz^T - \hat{y}\hat{z}^T, \quad \Delta b_y \propto y - \hat{y}, \quad \Delta b_z \propto z - \hat{z} \quad (9)$$

Note that the learning rule is local and Hebbian (correlational). The entire procedure amounts to taking a full step of Gibbs sampling in a Markov chain that has been initialized at a vector sampled from the “data distribution”  $p$ , and then changing weights so as to penalize the network for drifting away from the data distribution. In practice, we depart from Eq 9 by using “momentum” and “weight decay” [15], as is standard in neural-network training. Our choice of momentum and decay make this equivalent to low-pass filtering the learning signal (the right-hand sides of the equations) with an overdamped second-order system before making weight changes. Biologically, it corresponds to changes in synaptic strength having their own intrinsic dynamics.

Training took place in “epochs.” Data in each epoch consisted of 40,000 vectors: 40 trajectories of 1000 time steps apiece, each vector consisting of the current sensory response (proprioceptive and efference-copy) and the previous hidden-unit activities (“recurrent”; see Fig 3B). On the initial time step, the recurrent units were set to all zeros and drove no weight changes. In order to accelerate convergence, and although biological implausible, weight changes were made on “minibatches” of 40 input vectors, each of which corresponded to the same time point, but from the 40 different trajectories. Fresh data (40 new trajectories) were generated every five epochs. Learning rates ( $\epsilon$ ) also decayed across epochs. For the rEFH

trained on an uncontrolled dynamical system, the total number of epochs was 120, and the decay was exponential: for the  $k$ th epoch,  $\hat{U}_k^{-1} = 1.1^k \hat{U}_0^{-1}$ . For the case of controlled dynamics, the

network was trained for 1200 epochs, with the reciprocal learning rate growing according to a

sigmoidal function:  $U_k^{-1} = \left( \frac{1000}{1 + \exp\{-k/8 + 7.5\}} + \frac{1}{2} \right) U_0^{-1}$  (The numbers were chosen so that the

sigmoid approximately matches the exponential growth for the first 120 epochs, although their exact values are not critical.)

*Testing.* Filtering was tested on a new set of 40 trajectories (40,000 vectors). At each time step, the current “sensory” (proprioception and efference copy) and recurrent responses were fed forward to the hidden layer of the network, as in training. Unlike training, however, no samples were taken from this vector of means; instead, the real-valued vector was itself returned as the recurrent response. This is equivalent to taking several ( $\sim 15$ ) samples and averaging [4]; the means themselves were used to simplify presentation of the results, since they correspond to the maximum achievable performance of the network.

Formally, the solution to the filtering problem is the optimal posterior distribution over the current stimulus location, given all the observations up to this point in time:  $p(\theta_t | r_0^\theta, \dots, r_t^\theta, r_0^u, \dots, r_t^u)$ . For

the controlled dynamical system, we also ask about the posterior distribution over the controls,

$p(u_t | r_0^\theta, \dots, r_t^\theta, r_0^u, \dots, r_t^u)$ , since they are observed only noisily at each time step. We discuss

optimality below, but note here that in our case these distributions are Gaussian, so their only non-zero cumulants are mean and covariance. Generically, proving optimality of the harmonium would require showing that both these cumulants can be recovered from its hidden units at every point in time; but in the present case it is only necessary to decode the posterior mean, since it is impossible for the network to keep track of the mean without also keeping track of the

covariance: incorrect estimates of the latter would result in mis-weighting of the relative reliability of current sensory information and current filter estimate, resulting in suboptimal inference of the mean at the next time step. Decoding the rEFH’s hidden units exploits a trick [4]. The representational space of the hidden units is obscure; therefore, the hidden unit activities (a real-valued vector) are passed back down through the network, i.e. into the space of the inputs. Here, the optimal decoding scheme is known: it is the center of mass of each noisy hill of activity [4]. This decoder was applied to hidden units at each time step, for each of the 40 testing trajectories, from which errors from the actual joint angle and control input were computed.

#### *The optimal filtering distribution*

For the graphical models in Figs 1A and 3A, the solution to the filtering problem can be assimilated to a variant on the Kalman filter, and therefore computed in closed form. This is because, although the emission  $p(r_i^\theta | \theta_i)$  is not a Gaussian distribution over  $r_i^\theta$ , it is a Gaussian function of  $\theta_i$  [4, 7] (i.e., the likelihood is an unnormalized Gaussian over  $\theta_i$ )—or more precisely, of  $C \theta_i$ , with  $C$  the observation matrix (see Eq. 4)—and this is the critical requirement for the derivation of Kalman’s recursive solution. The resulting modification is small: Where the emission variance and the (Gaussian-distributed) emission appear in the standard KF equations, we substitute, respectively, the scaled tuning-curve width,  $\sigma_{ic}^2 / \sum_i r_i^\theta$ , and the center of mass of the population response,  $\sum_i \xi_i r_i^\theta / \sum_i r_i^\theta$  [16].

The same applies, mutatis mutandis, to the controls. In fact, the “controlled” case provides no additional complexity, since it corresponds to an uncontrolled third-order system (since the control has its own dynamics) whose state  $X_t$  is the concatenation of  $\theta_t$  and  $u_t$ :



$$\mathcal{N} p(x_{t+1} | x_t) = \mathcal{N} (\Gamma x_t + \mu_x, \Sigma_x), \quad (10)$$

With

$$\Gamma := \begin{bmatrix} 1 & \Delta & 0 \\ -\frac{k}{m} \Delta & 1 - \frac{c}{m} \Delta & \frac{\Delta}{m} \\ 0 & 0 & \alpha \end{bmatrix}, \quad \mu_x := \begin{bmatrix} \mu_\theta \\ \mu_u \end{bmatrix}, \quad \Sigma_x := \begin{bmatrix} \Sigma_\theta & 0 \\ 0 & \sigma_u^2 \end{bmatrix}$$

In both cases, then, the posterior (filtering) distribution over the state is always Gaussian, so at every time step, one computes the posterior mean and covariance, which can be expressed in terms of the filtering distribution at the previous time step, and of the current sensory information. A full derivation appears in S1 Text.

Eq 10 ignores some independence statements asserted by the graph of Fig 3A. In fact, an EM algorithm that accounts for them can be derived; but in our experiments, this algorithm does not achieve superior results to the “agnostic” version that tries to learn unconstrained versions of  $\Gamma$ ,  $\mu_x$ , and  $\Sigma_x$ . Therefore, results for EM<sup>3</sup> throughout use the unconstrained version of the algorithm.

*Benchmark models.* Error statistics for the rEFH are compared to those from four types of model. It is simplest to think of all four types using the same filtering algorithm—the KF, modified as described to account for the Poisson emissions—but running that filter on different generative models for the observed data  $(r_t^\theta, r_t^u)$ .

- **PROP and EfCp:** In the simplest benchmark model, joint-angle (PROP) and control (EfCp) estimates are made simply via the center of mass on the current “sensory” population. This is the optimal decoder for populations of smoothly tiled, Gaussian-tuned, Poisson neurons [17], under the assumption of independence through time (no dynamics). It can also be

thought of as a Kalman filter applied to a generative model with infinite state-transition covariance,  $\Sigma_x$ . (This severs the horizontal connections in Figs 1A and 3A.)

- **OPT:** The “optimal” model runs the KF on the true generative model for the data, i.e., using the true parameters,  $\{A, b, C, \alpha, h, \Sigma_\theta, \sigma_u^2, \nu_0, \Upsilon_0\}$ .
- **EM<sup>n</sup>:** The rEFH was not, of course, given its parameters, but had to learn them, and only from the noisy population responses,  $R_0^\theta, \dots, R_T^\theta, R_0^u, \dots, R_T^u$ . A useful point of comparison, then, is the performance of a (Kalman) filter that has learned those parameters from the same noisy data, but following a learning procedure that is known to be optimal. For our linear, time-invariant systems, such learning rules can be derived: they are an implementation of expectation-maximization (EM), an algorithm that guarantees convergence to at least local optima. Applying EM to linear dynamical systems requires both a forward pass through the data, filtering, and a backward pass for smoothing, i.e., computing the probability of the hidden state given the observations for all time. These equations are derived in S1 Text. Likewise, the algorithm must be told the order (number of states) of the latent dynamical system. Which order it was told is indicated by a superscript (e.g., EM<sup>2</sup>). We emphasize that access to the backward pass of observations, and knowledge of the order of the latent dynamics, are advantages the EM-trained models (EM<sup>n</sup>) enjoyed over the harmonium.
- **OBS:** For the controlled system (Controlled dynamical system), one would also like to know how useful knowledge of the controls is to inference of the joint angle. The optimal model that ignores controls, again under the assumption of linear dynamics, can be constructed by training on fully observed data, where the model parameters  $\{A, C, \Sigma_\theta, \nu_0, \Upsilon_0\}$  are learned from  $\theta_0, \dots, \theta_T, R_0^\theta, \dots, R_T^\theta$ , essentially via linear regression (see S1 Text). This corresponds to

using the generative model of Fig 1A, even though the true data were generated by the model of Fig 3A. OBS is therefore suboptimal, but tells us how much suboptimality accrues by ignoring the controls. (We fit the parameters of OBS, rather than providing them, because the fit model can actually outperform one based on the true dynamical-system parameters. This is because OBS can compensate to some extent for the missing controls by overestimating the state-transition noise. If the model were forced to use the true state-transition noise, but still assume zero control input, it would be worse at explaining state transitions.)

The benchmark models also enjoyed another advantage over the rEFH. The sufficient statistics of the emission for the state are the population center of mass and the scaled tuning-curve width (or simply the scaling factor), as alluded to above. These were given to the benchmark models, rather than learned. (In the standard model, linear-Gaussian emissions with fixed variance, learning that variance with EM is straightforward [18]. For our more complicated emission model, it is not, which is why we decided simply to provide it to the benchmark models.) The harmonium, on the other hand, had to learn what to do with the vectors of raw spike counts,  $r^{\theta}, r^u$ .

Different training runs, like different testing sets, will yield slightly different models. Thus, for each type of model to be trained, including the rEFH, we selected the best of 20 different networks, each trained from scratch with random initialization. Then we repeated this procedure itself twelve times, and from these twelve tokens of each model type, generated the error bars for the MSEs in Figs 1E, 3D and 3F. Each of the twelve tokens of a particular model type was tested on a different testing set, but the same testing set was used for matching tokens of different types (so, e.g., the fourth rEFH token was tested on the same data as the fourth EM2 token).

### *Tuning analysis*

In the section Learned receptive fields and connectivity, in order to determine how the network has learned to solve the filtering problem, we sort hidden units by their “preferred” lags and “preferred” angles. These were computed as follows. First, we generated a new set of 40 trajectories of 1000 time steps apiece. Then we computed hidden-unit mean activities, i.e., their probability of firing (these are the same quantity because the hidden units are conditionally Bernoulli random variables). Angular positions for all 40,000 time points were then discretized into 30 bins of uniform width spanning the feasible joint space.

For each hidden unit, the following calculation was then performed. First, the empirical mutual information (MI) was computed, according to the standard formula [19], between the two discrete random variables: the discretized position (30 bins) and the binary (spike/no spike) hidden-unit response. Next, to reject spurious MI (which will anyway be rare, given the number of data), for each of 20 reshuffles, the unit’s response was shuffled in time and the MIs recalculated. If the unit’s unshuffled MI fell below the 95th percentile of its shuffled MIs, the unshuffled MI was set to zero. The entire procedure was then repeated with the response time shifted forward by one step, for each of 40 steps. Finally, the “preferred” lag was selected to be the time shift for which MI was maximized. These were used to sort the receptive fields in Figs 5A and 5B, 7 and 8B.

For each unit, a “lagged” tuning curve can be constructed by considering its mean responses to past (discretized) stimuli; in particular, to stimuli at that unit’s preferred lag. These are the curves plotted as a heat map in Fig 5C, where they have been sorted by the locations of the tuning curves’ peaks. The same locations were used to sort the weight matrix in Fig 8A. Inverting the process, one can ask how well these tuning curves explain the receptive fields in the space of non-delayed position and velocity (Fig 5A): apply each tuning curve to each of the 40000 stimuli,

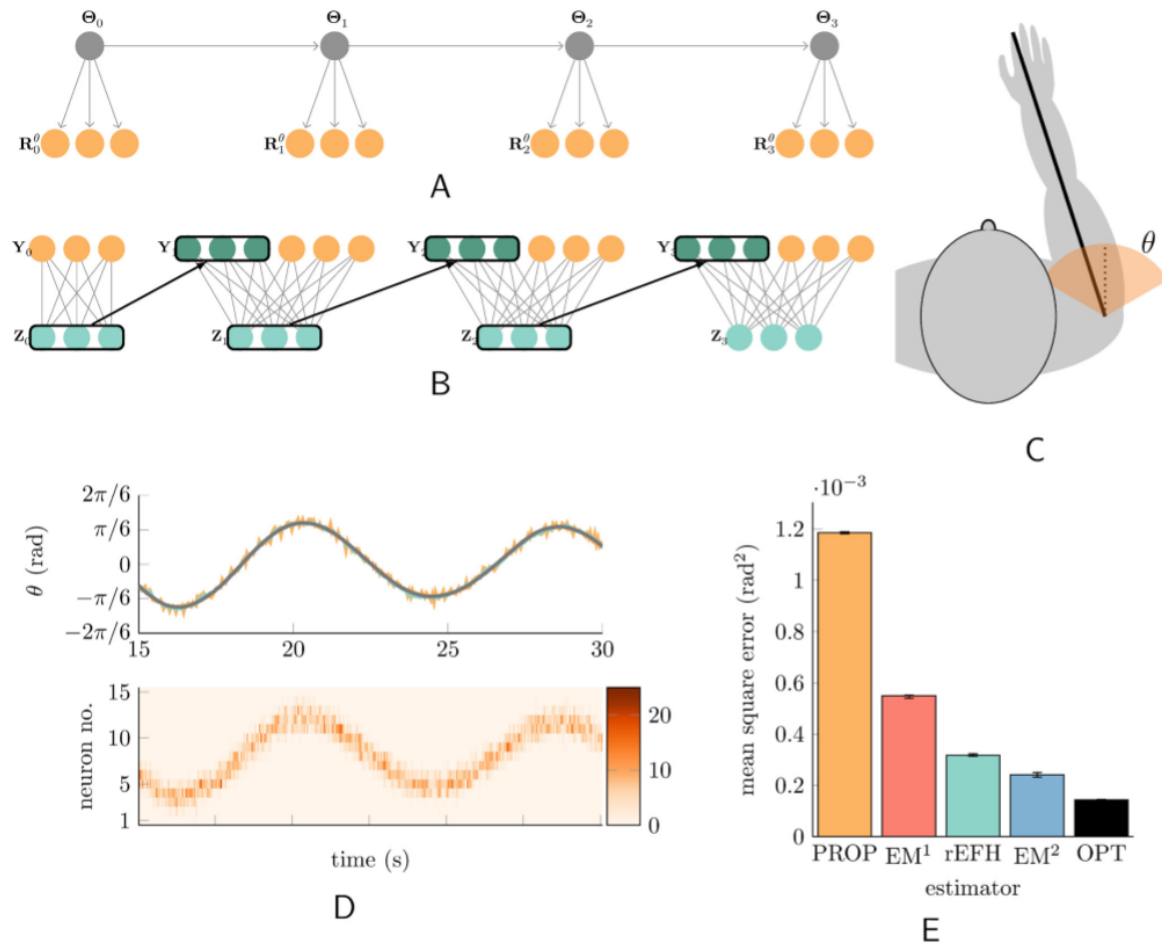
delay the responses by the units' preferred lags, and then compute receptive fields with these responses. This is how Fig 5B was constructed.

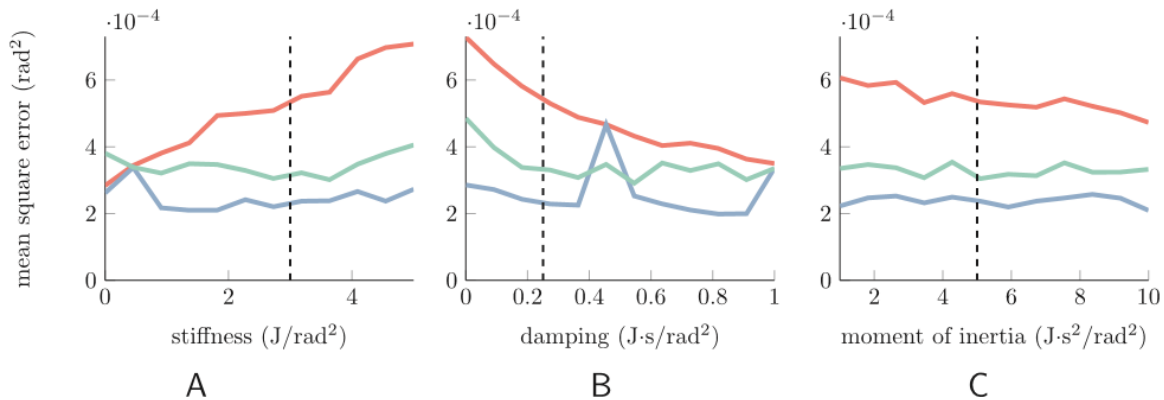
Finally, comparing the distribution of preferred lags (Fig 5D) to the autocorrelation of the stimulus required computing the autocorrelation of a circular variable (angle). We used the angular-angular correlation measure given by Zar [20].

### **Acknowledgments and Author Contributions**

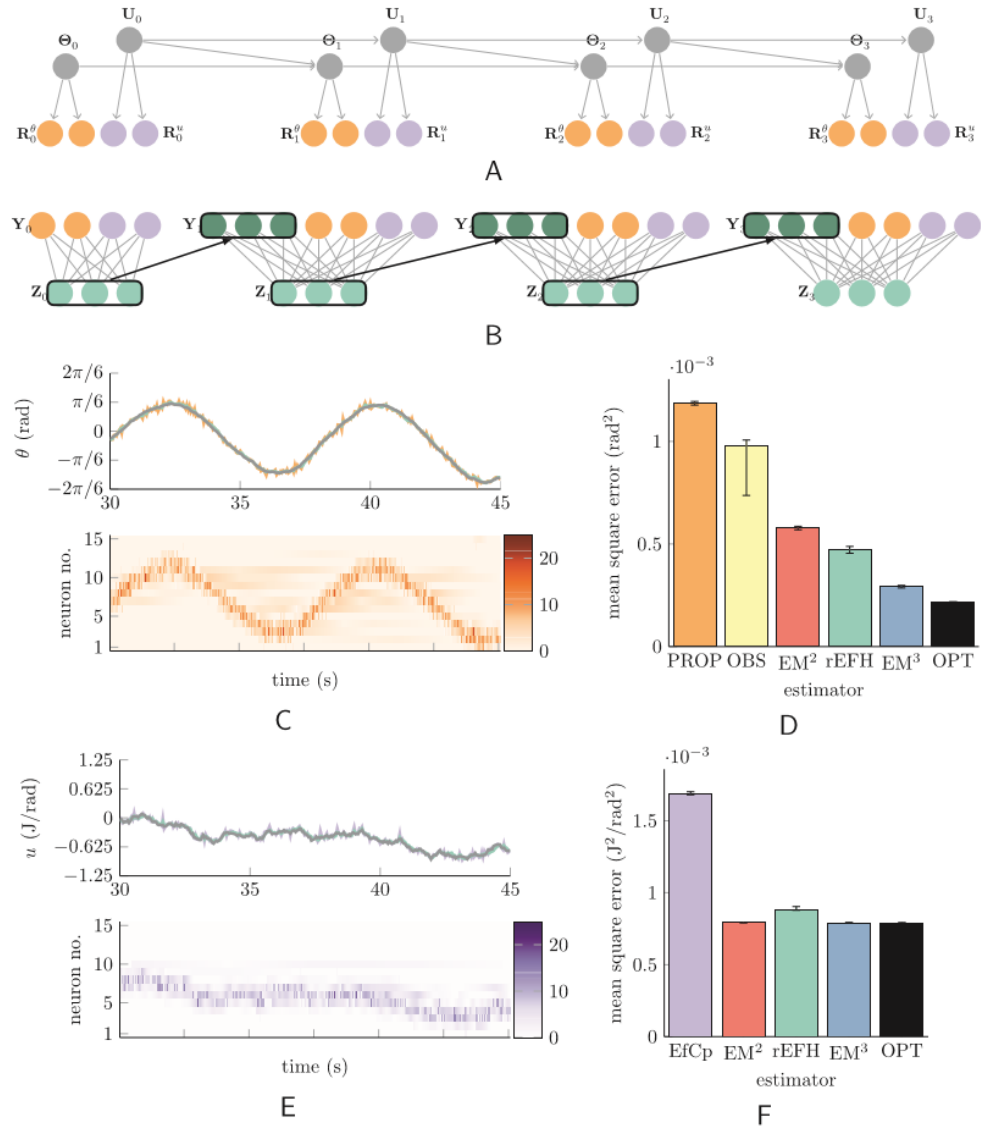
Some of the EFHs were trained using Tesla K40 GPUs, the generous donation of the Nvidia Corporation.

Conceived and designed the experiments: JGM BKD PNS. Performed the experiments: JGM BKD. Analyzed the data: JGM BKD PNS. Wrote the paper: JGM BKD PNS. Proposed the technique that extends the harmonium to dynamical data: BKD.



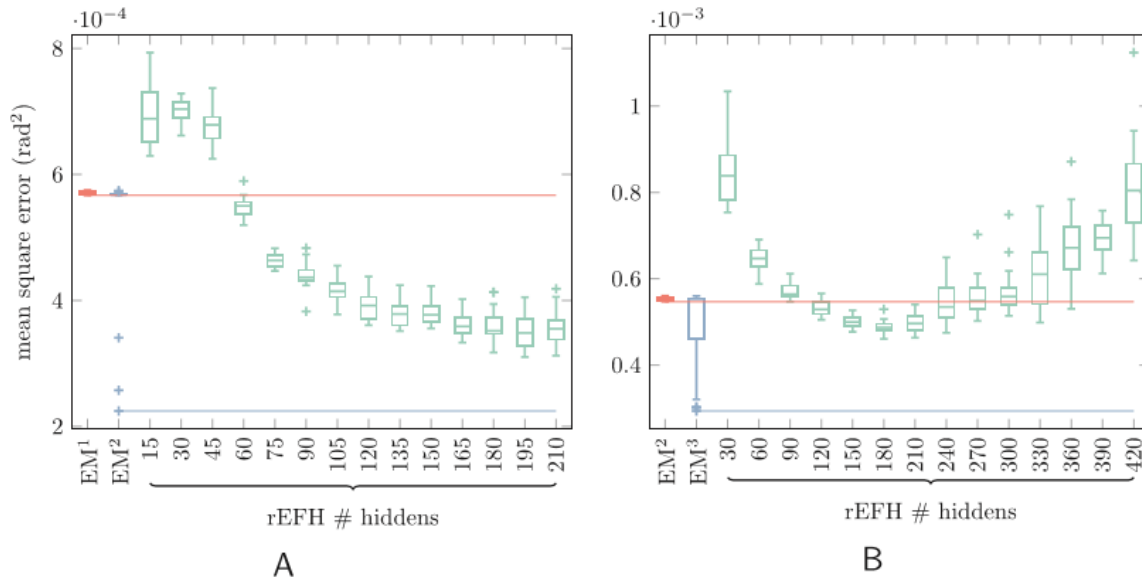


**Figure 2. Mean squared errors (MSEs) for various dynamical systems.** To show the flexibility of the rEFH, we train one for each of several different second order dynamical systems. Each sub plot shows the results for twelve different systems in which a single parameter has been varied across a range of values; otherwise the systems are identical to the undriven model of Fig1. For each of those twelve dynamical systems, 20 rEFHs were trained (each with 150 hidden units), MSEs calculated, and the best selected (turquoise). The same was then done for  $EM^1$  (light red) and  $EM^2$  (blue), i.e., Kalman filters trained with EM, assuming either first- or second-order dynamics. The vertical black line in each plot indicates the dynamical system used in Fig1 in the main text. **(A)** Varying the spring constant. The left most datum ( $k=0$ ) corresponds to the “no-spring” model from which the RFs are analyzed below (although with lower transition noise); at this point, the dynamics can be well approximated by a first-order model. **(B)** Varying the damping coefficient. The spike in  $EM^2$  at  $c=0.4545$  indicates that EM failed to find the second-order solution in any of its 20 attempts. As  $c$  increases and the systems approach critical damping, however, first-order approximations are increasingly adequate. **(C)** Varying the moment of inertia.

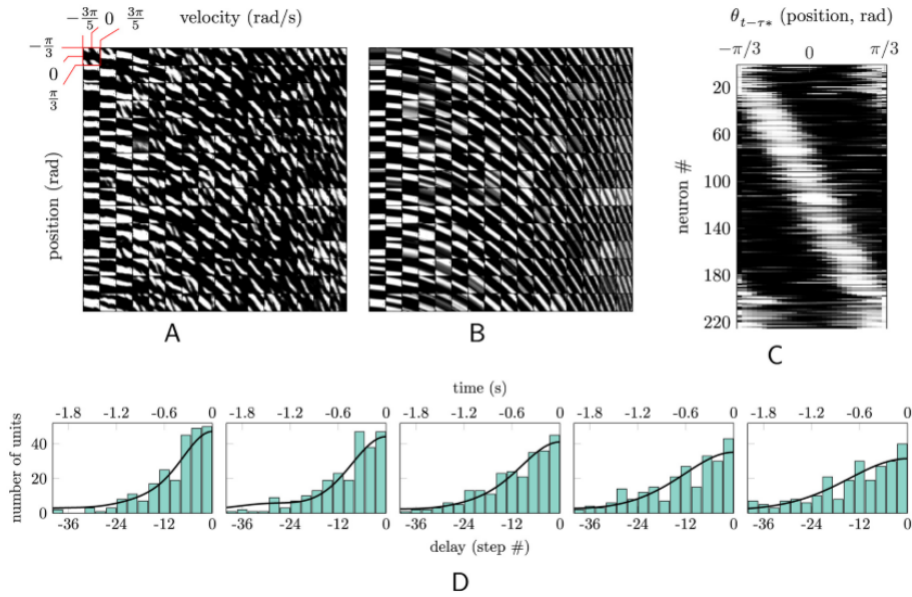


**Figure 3. Dynamical system with efference copy.** A reprise of Fig 1 for the controlled system. The evolution is now third-order, since the control signal is (close to) a random walk. **(A)** Angle (but not angular velocity) and the control signal are each noisily reported by populations of neurons (orange and purple; two of the fifteen neurons are depicted). **(B)** The training data for this harmonium at each time step are these two populations, and recurrent activity from the hidden units. **(C)** (Top) Fifteen seconds of a typical trajectory (black) and the trajectory decoded from the sensory population (orange), and from the hidden units (turquoise). The influence of the control can be seen in the (mildly) increasing amplitude of the oscillation. (Bottom) The activity of the fifteen sensory neurons is reported as a heat map. **(D)** Error statistics for joint-angle decoding under different models (see text). **(E)** The same as in (C), but the neurons carry “efference copy” rather than sensory information. **(F)** Error statistics for control-signal decoding under different models (see text). Error bars mark the first and third quartile across twelve de novo trained and tested networks of each type.

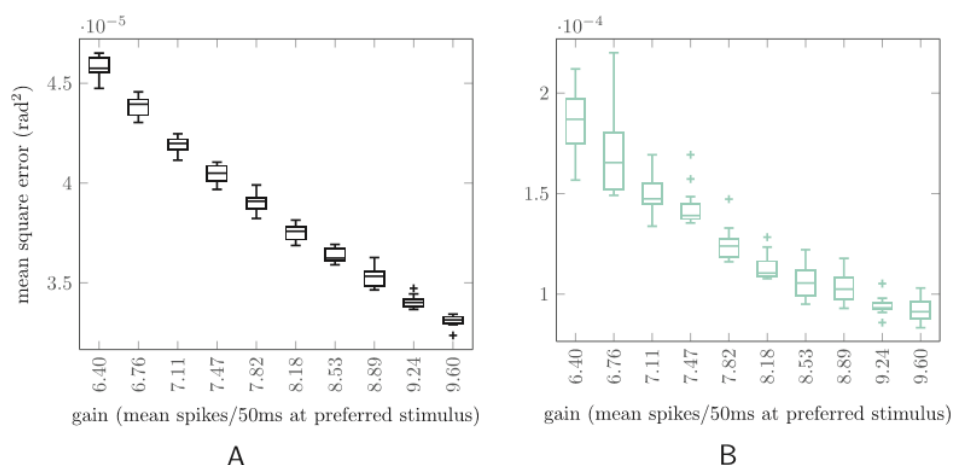




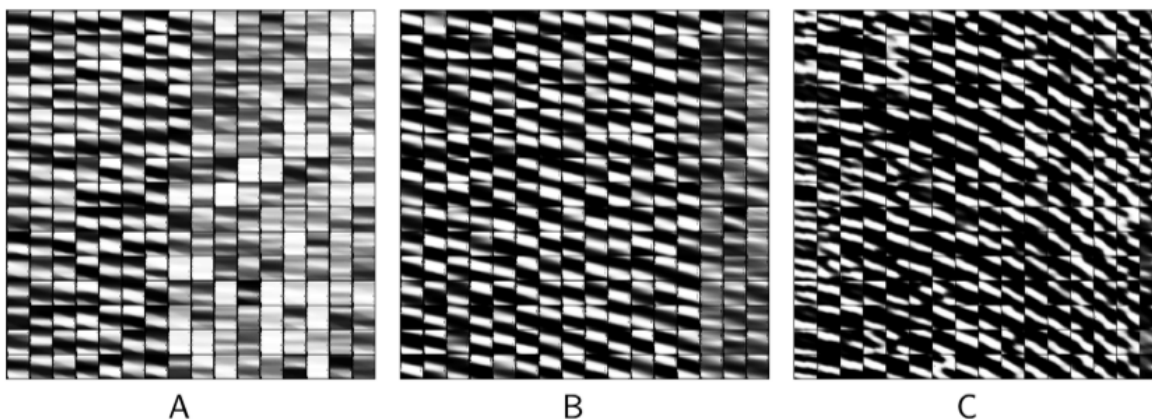
**Figure 4. Box-and-whisker plot of MSEs for EM-based models and for rEFHs of various sizes.** Each box corresponds to 20 networks trained de novo and tested on a common data set. Median MSE for each model is marked with a horizontal line; the box contains the interquartile range; whiskers extend to  $1.5\times$  the interquartile range, beyond which outliers are marked with plus signs. Performance among the EM-learned linear dynamical systems (the first two boxes, light red and blue) varies comparatively little, although large outliers are sometimes produced. In fact, the higher-order (best performing) models are all outliers. To facilitate comparison with the rEFHs, a line extends from the best EM-based models across the entire plot. The remaining twelve boxes (turquoise) are rEFHs with different numbers of hidden units, listed on the abscissae. All have the same number of recurrent units as hidden units, and have a fixed number of “sensory units”: (A) 15 proprioceptive, or (B) 15 proprioceptive and 15 efference-copy. **(A)** The uncontrolled dynamical system. Overall, MSEs decline with increasing number of hidden/recurrent units, but appear to asymptote by 180 hidden units (ratio =  $180/15 = 12$ ). **(B)** The controlled dynamical system. The optimal number of hidden units is about 180 (ratio =  $180/30 = 6$ ), after which mean and variance (across networks) of MSE increases.



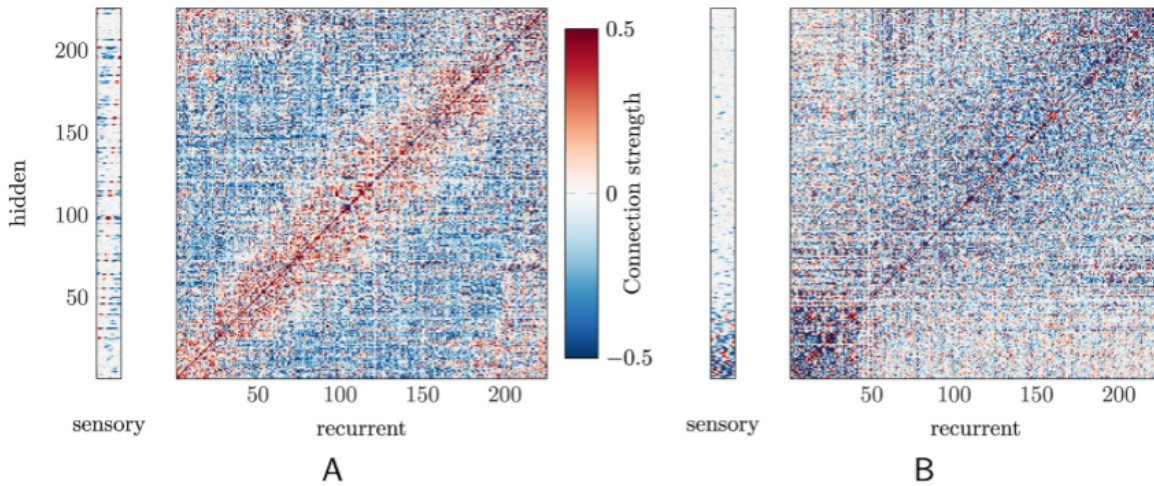
**Figure 5. Position and velocity receptive fields of hidden units.** In (A)-(C), pure white corresponds to a firing probability of one; pure black to zero. **(A)** Receptive fields for all 225 hidden units of the spring-free model (see text) in the space of (angular) position (ordinates) and velocity (abscissae). The angle limits and angular-velocity limits, indicated on the first (upper-left) receptive field, are the same for all units. **(B)** The predicted position-velocity receptive fields of units that have only the lagged-position tuning given by (C). The match with (A) is excellent for all but the anomalous 25 units at the right. **(C)** The same 225 units, each now plotted as a function of the time-lagged position with which that unit has maximal mutual information. Units have been arranged in order of increasing preferred position, whereas the units in (A) and (B) are arranged in order of maximally informative lags: from top to bottom and left to right, units are tuned for more temporally distant positions. This tuning gives rise to “stripes” in position-velocity space. For (A)-(C), the 25 units that do not appear to be well modeled by tuning to past positions have been placed at the end. **(D)** Histograms of the “preferred” lags, in terms both of time and (equivalently) discrete time steps, for five different networks. The normalized autocorrelation of the underlying dynamical system is superimposed. The central panel corresponds to the network analyzed in (A)-(C). The other four panels correspond to networks trained on observations from dynamical systems with different autocorrelations. From left to right panel, the dynamical systems get slower.



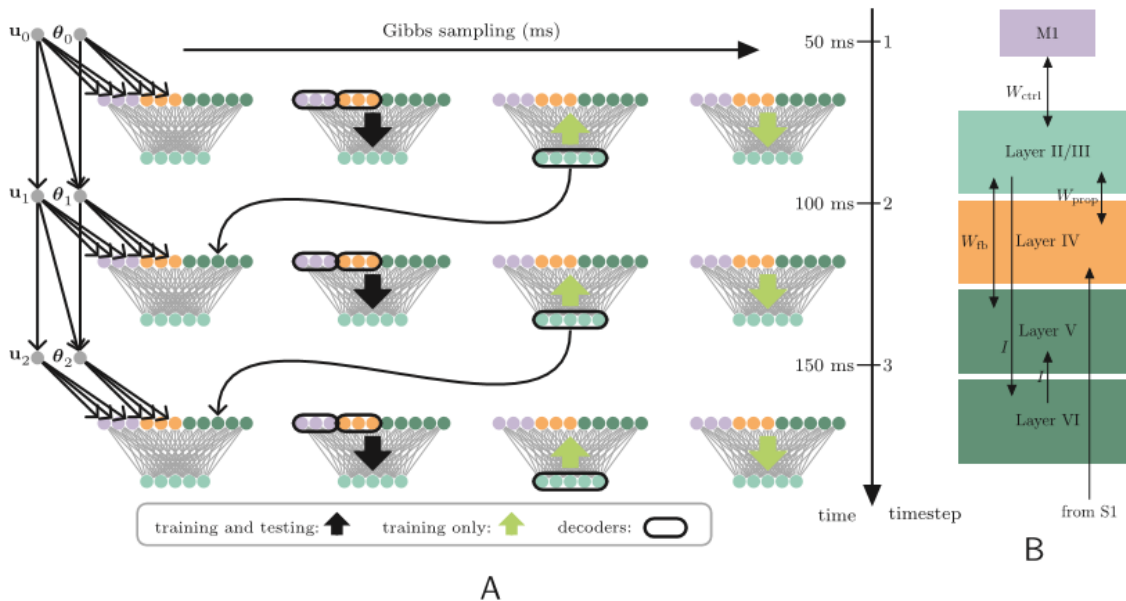
**Figure 6. Network sensitivity to instantaneous reliability.** The instantaneous (one-time-step) reliability of sensory information is determined by the total number of spikes across the sensory population within one time step. An optimal filter will up-weight sensory information that is more reliable (and vice versa). If such a filter is run on noiseless sensory data, then its errors will be smaller for sensory input with more total spikes (higher gain), since it will up-weight the perfect sensory information. **(A)** Box-and-whisker plot (interpretation as in Fig 4) of mean squared errors for the optimal model (OPT), when tested on noiseless sensory data and a range of gains. For each gain on the abscissa, the filter was tested on 12 sets of 320 trajectories apiece, for which the sensory gain was fixed throughout. Higher-gain trajectories yield lower mean errors, as expected. **(B)** The same plot for the network (rEFH). The magnitude of the MSEs is larger than for the optimal filter, as in Fig 1E, but the pattern is the same, showing that the rEFH has indeed learned to treat higher-gain (more- spikes) sensory information as more reliable.



**Figure 7. Emergence of position and velocity receptive during training.** Velocity tuning takes longer to emerge than position tuning, because velocity information is only available in the inputs after position has already been learned by the hidden units—and subsequently fed back. **(A)** After 100 batches of training, stripes are mostly horizontal or very shallowly sloped in position-velocity space—no velocity tuning. **(B)** By 200 batches, velocity tuning is evident across most units. **(C)** Later, at 1000 batches, the slopes of the “stripes” have increased, indicating position tuning for more temporally distant (past) stimuli.



**Figure 8. The weight matrices.** The organization of connections between inputs (sensory or recurrent units) and the hidden layer can be visualized by sorting units by **(A)** preferred stimulus angle or **(B)** by preferred lag (increasing from lower left to upper right). Here we analyze the “no-spring” network (see text). In both subfigures, both the recurrent and hidden units have been re-sorted; the sensory units remain organized by preferred angle. Note that self-connections (along the diagonal) are in fact more than 0.5, but the plot saturates at this value to make the other connections more visible.



**Figure 9. Training and testing: in the model, and in a cortical implementation. (A)** The training and testing procedure in the model. Three discrete time steps are arrayed vertically. At each one, the current arm position ( $\theta_t$ ) is reported by the proprioceptive population (orange) with Poisson-distributed spike counts, as shown in the first column. The current motor command is likewise reported by an “efference-copy” population (purple). Second column: this spiking, along with recurrent activities (dark turquoise), stochastically drives single spikes in the hidden layer (turquoise). Third column: these spikes in turn drive the three “input” populations (this is not required during testing); a “copy” of the hidden vector is also saved to serve as recurrent activity at the next time step (curved black arrow). Fourth column: finally, the input populations drive the hidden layer once more, after which the weights are changed according to Eq. 9 (also not required during testing). At every time step, the current joint angle and current control are decoded naïvely from the current activities of their respective input populations; with Kalman filters that are recursively updated based on these activities (not depicted in this figure); and from the hidden units of the rEFH. **(B)** A possible cortical implementation of the rEFH. The cortical layers of Brodmann Area 5 and its inputs are identified with elements of the rEFH by color. The synaptic connections are denoted by the arrows and their corresponding weight matrices. Primary somatosensory cortex (S1) provides feedforward proprioceptive input to layer IV, while primary motor cortex (M1) provides feedback input—a copy of the efferent command—to layer II/III. The recurrent signal of the rEFH (heavy curved arrow in (A)) is identified with a reverberatory loop from II/III to VI, to V, and back to II/III.

## References

- Földiák P. The “Ideal Homunculus”: Statistical Inference from Neural Population Responses. In: Eeckman FH, Bower JM, editors. *Computation and Neural Systems*. Norwell, MA: Norwell, MA: Kluwer Academic Publishers; 1993. p. 55–60.
- Ma WJ, Beck JM, Latham PE, Pouget A. Bayesian Inference with Probabilistic Population Codes. *Nature Neuroscience*. 2006; 9:1423–1438.
- Welling M, Rosen-Zvi M, Hinton GE. Exponential Family Harmoniums with an Application to Information Retrieval. In: *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*; 2005. p. 1481–1488.
- Makin JG, Fellows MR, Sabes PN. Learning Multisensory Integration and Coordinate Transformation via Density Estimation. *PLoS Computational Biology*. 2013; 9(4):1–17. doi: 10.1371/journal.pcbi. 1003035
- Van Beers RJ, Sittig A, van Der Gon JJD. Integration of Proprioceptive and Visual Position-Information: An Experimentally Supported Model. *Journal of Neurophysiology*. 1999; 81:1355–1364. PMID: 10085361
- Ernst MO, Banks MS. Humans Integrate Visual and Haptic Information in a Statistically Optimal Fashion. *Nature*. 2002; 415(January):429–433. doi: 10.1038/415429a PMID: 11807554
- Beck JM, Latham PE, Pouget A. Marginalization in Neural Circuits with Divisive Normalization. *Journal of Neuroscience*. 2011 oct; 31(43):15310–9. doi: 10.1523/JNEUROSCI.1706-11.2011 PMID: 22031877

- Boerlin M, Denève S. Spike-Based Population Coding and Working Memory. *PLoS Computational Biology*. 2011 feb; 7(2):e1001080. doi: 10.1371/journal.pcbi.1001080  
PMID: 21379319
- Makin JG, Sabes PN. Sensory Integration and Density Estimation. *Advances in Neural Information Processing Systems 27: Proceedings of the 2014 Conference*. 2015;p. 1–9.
- Wolpert DM, Goodbody SJ, Husain M. Maintaining Internal Representations: the Role of the Human Superior Parietal Lobe. *Nature Neuroscience*. 1998; 1:529–533. doi: 10.1038/2245 PMID: 10196553
- Mulliken GH, Musallam S, Andersen RA. Decoding trajectories from posterior parietal cortex ensembles. *Journal of Neuroscience*. 2008 nov; 28(48):12913–26. doi: 10.1523/JNEUROSCI.1463-08.2008 PMID: 19036985
- Hinton GE, Osindero S, Teh YW. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*. 2006; 18:1527–1554. doi: 10.1162/neco.2006.18.7.1527 PMID: 16764513
- Hinton GE. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*. 2002; 14:1771–1800. doi: 10.1162/089976602760128018 PMID: 12180402
- Scherpen JMA. Chapter 4: Balanced Realizations, Model Order Reduction, and the Hankel Operator. In: Levine WS, editor. *The Control Handbook*. 2nd ed. CRC Press; 2015. p. 4–1–4–24.
- Hinton GE. *A Practical Guide to Training Restricted Boltzmann Machines*. Toronto: University of Toronto; 2010.

- Makin JG, Fellows MR, Sabes PN. Learning Multisensory Integration and Coordinate Transformation via Density Estimation—Supporting Material. *PLoS Computational Biology*. 2013; 9(4):1–9. doi: 10.1371/journal.pcbi.1003035
- Dayan P, Abbott LF. *Theoretical Neuroscience*. MIT Press; 2001.
- Ghahramani Z, Hinton GE. *Parameter Estimation for Linear Dynamical Systems*. University of Toronto; 1996.
- Cover TM, Thomas JA. *Elements of Information Theory*. Wiley; 2006.
- Zar JH. *Biostatistical Analysis*. Prentice Hall; 1999.
- Andersen RA, Snyder LH, Bradley DC, Xing J. Multimodal Representation of Space in the Posterior Parietal Cortex and its Use in Planning Movements. *Annual Review of Neuroscience*. 1997; 20:303– 330. doi: 10.1146/annurev.neuro.20.1.303 PMID: 9056716
- Egger SW, Britten KH. Linking sensory neurons to visually guided behavior: relating MST activity to steering in a virtual environment. *Visual neuroscience*. 2013 nov; 30(5–6):315–30. doi: 10.1017/S0952523813000412 PMID: 24171813
- Kuenzle H. Cortico-Cortical Efferents of Primary Motor and Somatosensory Regions of the Cerebral Cortex in *Macaca Fascicularis*. *Neuroscience*. 1978; 3:25–39. doi: 10.1016/0306-4522(78)90151-3
- Ghosh S, Brinkman C, Porter R. A Quantitative Study of the Distribution of Neurons Projecting to the Precentral Motor Cortex in the Monkey (*M. Fascicularis*). *The Journal of Comparative Neurology*. 1987; 259:424–444. doi: 10.1002/cne.902590309 PMID: 3584565



Burchinskaya LF. Neuronal Composition and Interneuronal Connection of Area 5 in the Cat Parietal Association Cortex. *Neirofiziologiya*. 1979; 11(1):35–42.

Carracedo LM, Kjeldsen H, Cunningham L, Jenkins a, Schofield I, Cunningham MO, et al. A Neocortical Delta Rhythm Facilitates Reciprocal Interlaminar Interactions via Nested Theta Rhythms. *Journal of Neuroscience*. 2013; 33(26):10750–10761. doi: 10.1523/JNEUROSCI.0735-13.2013 PMID: 23804097

Rao RPN, Ballard DH. Dynamic Model of Visual Recognition Predicts Neural Response Properties in the Visual Cortex. *Neural Computation*. 1997 may; 9(4):721–63. doi: 10.1162/neco.1997.9.4.721 PMID: 9161021

Denève S, Duhamel JR, Pouget A. Optimal Sensorimotor Integration in Recurrent Cortical Networks: A Neural Implementation of Kalman Filters. *Journal of Neuroscience*. 2007; 27(21):5744–5756. doi: 10.1523/JNEUROSCI.3985-06.2007 PMID: 17522318

Huys QJM, Zemel RS, Natarajan R, Dayan P. Fast Population Coding. *Neural Computation*. 2007; 19:404–441. doi: 10.1162/neco.2007.19.2.404 PMID: 17206870

Natarajan R, Huys QJM, Dayan P, Zemel RS. Encoding and decoding spikes for dynamic stimuli. *Neural Computation*. 2008; 20:2325–2360. doi: 10.1162/neco.2008.01-07-436 PMID: 18386986

Hinton GE, Brown A. Spiking Boltzmann Machines. *Advances in Neural Information Processing Systems 12: Proceedings of the 1999 Conference*. 2000;12.

Sutskever I, Hinton GE. Learning Multilevel Distributed Representations for High-Dimensional Sequences. In: *AISTATS*; 2007. p. 1–8.

Sutskever I, Hinton GE, Taylor G. The Recurrent Temporal Restricted Boltzmann Machine. In:  
Advances in Neural Information Processing Systems 21: Proceedings of the 2008  
Conference; 2009. p. 1–8.

Wolpert DM, Ghahramani Z, Jordan MI. An Internal Model for Sensorimotor Integration.  
Science. 1995; 269(5232):1880. doi: 10.1126/science.7569931 PMID: 7569931

## Chapter 2. Dynamic Structure of Neural Variability in the Cortical Representation of Speech Sounds

Benjamin K. Dichter,<sup>1,2,3</sup> Kristofer E. Bouchard,<sup>1,2,4,5</sup> and Edward F. Chang<sup>1,2,3,4,6</sup>

<sup>1</sup>Departments of Neurological Surgery and Physiology, University of California–San Francisco, San Francisco, California 94143-0112, <sup>2</sup>Center for Integrative Neuroscience, University of California–San Francisco, San Francisco, California 94158, <sup>3</sup>University of California–Berkeley and University of California–San Francisco Joint Program in Bioengineering, Berkeley, California 94720-3370, <sup>4</sup>Center for Neural Engineering and Prosthesis, University of California–San Francisco and University of California–Berkeley, Berkeley, California 94720-3370, <sup>5</sup>Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, and <sup>6</sup>University of California–San Francisco Epilepsy Center, San Francisco, California 94143

**Abstract:** Accurate sensory discrimination is commonly believed to require precise representations in the nervous system; however, neural stimulus responses can be highly variable, even to identical stimuli. Recent studies suggest that cortical response variability decreases during stimulus processing, but the implications of such effects on stimulus discrimination are unclear. To address this, we examined electrocorticographic cortical field potential recordings from the human nonprimary auditory cortex (superior temporal gyrus) while subjects listened to speech syllables. Compared with a prestimulus baseline, activation variability decreased upon stimulus onset, similar to findings from

microelectrode recordings in animal studies. We found that this decrease was simultaneous with encoding and spatially specific for those electrodes that most strongly discriminated speech sounds. We also found that variability was predominantly reduced in a correlated subspace across electrodes. We then compared signal and variability (noise) correlations and found that noise correlations reduce more for electrodes with strong signal correlations. Furthermore, we found that this decrease in variability is strongest in the high gamma band, which correlates with firing rate response. Together, these findings indicate that the structure of single-trial response variability is shaped to enhance discriminability despite non-stimulus-related noise.

**Key words:** ECoG; encoding; noise correlations; speech; superior temporal gyrus; variability

**Significance Statement:** Cortical responses can be highly variable to auditory speech sounds. Despite this, sensory perception can be remarkably stable. Here, we recorded from the human superior temporal gyrus, a high-order auditory cortex, and studied the changes in the cortical representation of speech stimuli across multiple repetitions. We found that neural variability is reduced upon stimulus onset across electrodes that encode speech sounds.

## Introduction

The human superior temporal gyrus (STG) represents speech sounds with spatiotemporal patterns of neural activity across populations that are tuned to specific acoustic features of the sounds (Formisano et al., 2008; Chang et al., 2010; Obleser et al., 2010; Steinschneider, 2011; Mesgarani et al., 2014; Nourski et al., 2014). However, neural responses to sensory stimuli are variable, and the brain responds differently to the same stimulus each time it is encountered (Faisal et al., 2008). Speech perception fundamentally involves classifying instances of sounds as members of specific linguistic categories (e.g., phonemes, words, etc.) (Liberman et al., 1957, 1967; Perkell and Klatt, 1986; Diehl et al., 2004), although the acoustics of these sounds

can vary in pitch, location, intensity, etc. The classification problem is compounded by the presence of different neural responses to physically identical sounds (Kisley and Gerstein, 1999). However, despite this variability, human listeners perceive speech effortlessly. This reliability in sensory perception despite variability in the neural response holds across sensory domains: visual (Schiller et al., 1976; Heggelund and Albus, 1978; Rose, 1979; Churchland et al., 2010), somatosensory (Whitsel et al., 1977), and as such, understanding neural variability is essential to understanding neural representations in general (Averbeck et al., 2006; Churchland et al., 2011).

Recent studies have identified a variety of factors that modulate trial-to-trial neural response variability in single-neuron firing rate. For example, the variability of neural responses to sensory stimuli is modulated by attentional state (Mitchell et al., 2009; Downer et al., 2015), and the difference in variability accounts for a large change in discriminability of stimuli (Cohen and Maunsell, 2009). Furthermore, neural response variability changes dynamically during the time course of stimulus presentation, with a reduction time-locked to stimulus onset (Cohen and Maunsell, 2009; Churchland et al., 2010). The potential for variability in both single neuron and population activity to hinder perception suggests that its modulation plays an important role in neural signal processing (Shadlen and Newsome, 1998; Abbott and Dayan, 1999; Churchland et al., 2011; Hu et al., 2014; Moreno-Bote et al., 2014).

It is unclear how reduction in neural variability affects stimulus representation and discriminability at the mesoscale of aggregate neural populations (i.e., field potentials). With a few exceptions (He and Zempel, 2013), most of the literature on the dynamics of neural response variability focuses on individual or multiple single-unit recordings, in which at most a few hundred neurons are simultaneously observed, a small subset of those neurons active in the sensory task. Previous studies have found strong encoding of acoustic-phonetic features in the STG using high-density electrocorticography (ECoG) (Mesgarani et al., 2014), where each

electrode records from populations several orders of magnitude greater than those observed in multi-neuron recordings (Chang, 2015). It is unclear how variability dynamics found in single and multiunit recordings extend to these larger neural populations, and how this affects sensory processing.

To address these questions, we recorded cortical field potentials using ECoG from human STG while subjects listened to simple speech sounds. We first determined whether cortical field potentials shared the trends found in firing rate responses: that the variance is correlated with the mean of the response (Tolhurst et al., 1983; Vogels et al., 1989) and that variability decreases after stimulus onset. Furthermore, we explored the relationship between variability and stimulus encoding, testing the hypothesis that the shape of the changes in variability of population neural responses depends on cortical sound representations.

## **Materials and Methods**

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

### *Subjects and experimental task.*

Three native English-speaking human participants (one female) underwent implantation of a high-density, subdural ECoG array as part of their clinically indicated neurosurgical treatment for epilepsy. Participants gave their written informed consent before the day of surgery. Implanted ECoG grids were each 256-channel grids with 2.3-mm-diameter electrodes at 4 mm center-to-center spacing and were placed in the language dominant hemisphere in each patient (as determined with the Wada carotid intra-arterial amobarbital injection), which was left in two subjects and right in one subject. Each participant listened to a recording of consonant-vowel (CV) syllables. Sixteen consonants (/b/, /d/, /g/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /sh/, /t/, /v/, /w/, /y/, /z/) combined with 3 vowels (/a/, /i/, /u/) were spoken by 6 speakers (3 female), resulting in 288

unique auditory stimuli total. The stimuli had a mean duration of 0.43 s and SD of 0.093 s. The inter stimulus interval was jittered across trials, with a mean of 1 s and a SD of 0.15 s. Stimuli were recorded in-house and played with speakers. Each stimulus was presented between 17 and 21 times to each subject.

#### *Anatomical location of STG.*

We focused our analysis on the STG, a non-primary auditory area that responds to speech sounds. Visual examination of co-registered CT and MR scans indicate that the ECoG grid covered the spatial extent of the STG of each patient. STG electrodes were identified through inspection of this co-registration and only those electrodes were used for analysis (see Fig. 1; number of electrodes in STG: S1:51; S2:48; S3:72).

#### *Data acquisition and signal processing.*

Cortical-surface field potentials were recorded referenced to scalp with a multichannel PZ2 amplifier optically connected to a RZ2 digital signal processor (Tucker-Davis Technologies). ECoG signals were acquired at 3052 Hz. The speaker signal was split and recorded in-line with the ECoG data to ensure synchronization. The time series voltage trace of each channel was visually and quantitatively inspected for artifacts or excessive noise (typically 60 Hz line noise or movement artifacts). Recordings with many artifacts were excluded from analysis, and the signal of the remaining channels were then common average referenced for each 16-channel ECoG strip to remove electrical noise shared across electrodes. The signals of the remaining channels were bandpass filtered using Gaussian bandpass filters with logarithmically increasing center frequencies and semilogarithmically increasing band widths from 4 to 200 Hz. The Hilbert transform was then calculated for each band, and the analytic amplitude at 400 Hz tracked the activation in each of the filter bands. The high-gamma power was calculated by averaging the analytic amplitude across the eight bands between 70 and 150 Hz. To mitigate intersession

changes in signal strength, high-gamma power was z-scored relative to the mean and SD of the recording session for each channel. Throughout, when we speak of high-gamma power, we refer to this z-scored measure, denoted as  $H\gamma$

### *Measuring variability.*

We wish to measure the time course of variability in neural activity. If we were to simply measure variance, the dynamics of the variability would be swamped by the effect of increased activity, which tends to increase following stimulus presentation and is strongly correlated with variance (Tolhurst et al., 1983). To study the effect of a stimulus on response variability, we observe how it changes the relationship between mean and variance of response. Previous studies have accounted for this relationship by using the Fano factor (FF) (Churchland et al., 2010) and varCE (Churchland et al., 2011), both of which are variability measures designed for firing rates that assume a linear relationship between mean activity and variance, and both were used to show a variability reduction following stimulus onset. We establish a similar metric, but cannot use FF or varCE because (z-scored)  $H\gamma$  has nonzero variance at zero activation. To account for this difference, we modify previous methods to include a y-intercept and quantify the variability as the slope of affine regression. We quantify the relationship between variance of responses ( $v$ ) and mean of responses ( $m$ ) with an affine function as follows:

$$v = Dm + C \tag{1}$$

where  $D$ , the slope, is our measure of variability, and  $C$  is a y-intercept, which is discarded.

Using this method, we are able to account for the effect of mean on variance.  $D$  is our analog for FF. It is not the same metric, but it serves the same purpose of quantifying the relationship between the mean and variance of neural activation.



### *Mean matching.*

To ensure that the variability reduction was not due to differences in the mean distribution, we use a “mean-matching” procedure adapted from Churchland et al. (2010). For each subject, electrode stimuli sets were removed so that each time point has the same mean distribution of  $H\gamma$  responses. The mean  $H\gamma$  responses for each stimulus on each electrode are binned into 30 equally spaced bins for each of the time points in the 400 Hz  $H\gamma$  signal spanning 300 ms before to 700 ms after the stimulus onset. A maximum common histogram of mean responses was then constructed, which was where all of the histograms through time overlapped. At each time point, stimulus-electrode responses were removed randomly from bins with mean  $H\gamma$  in excess of the maximum common histogram until each histogram was equal to the maximum common histogram. The mean matching procedure resulted in data that had the same mean distribution through time, so differences in the regression coefficient cannot be attributed to differences in mean activity.

### *Temporal encoding.*

The relationship between acoustics and  $H\gamma$  was modeled by a token stimulus-encoding model, where the  $H\gamma$  response was modeled for each of the 288 different CV sound stimuli independently. The adjusted  $R^2(\bar{R}^2)$  (Theil, 1961) was used to determine the degree of encoding through time as follows:

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-p-1} \quad (2)$$

Where  $n$  is the sample size and  $p$  is the number of unique stimuli.  $\bar{R}^2$  is similar to  $R^2$  but is altered to adjust for bias in the estimation of variance due to a small sample size.  $\bar{R}^2$  was calculated independently for each  $H\gamma$  measurement in time, then averaged across electrodes

to obtain a trend across the STG. Qualitatively similar time courses for all three subjects were obtained with a linear discriminant classifier decoding stimulus identity from  $H\gamma$  across all electrodes in a subject.

*Spatial variability and encoding.*

Electrode-wise encoding and variability changes were analyzed by a comparison of the mean of  $D$  during a representation baseline (300 to 0 ms) and a stimulus-encoding period (100–400 ms). Here, encoding strength was calculated as  $\bar{R}^2$  but not averaged. The variance-mean regression was performed individually for each electrode and not mean-matched.

*Factor analysis.*

Next, we studied how the decrease in variability affected correlated noise across electrodes. Factor analysis (FA) was used to separate the variability of responses into shared variance and private noise. FA is an unsupervised machine learning algorithm that models the data as being generated by a Gaussian distribution on a lower dimensional space ( $x$ ), corrupted with private (uncorrelated) noise  $Q$  into the full dimensionality of the data as follows:

$$\begin{aligned} x &\sim \mathcal{N}(0, I) \\ y &\sim \mathcal{N}(Cx, Q) \end{aligned} \tag{3}$$

where  $x$  is a vector of latent variables and  $y$  is a vector of the observed  $H\gamma$ ,  $Q$  is the noise covariance matrix and is constrained to be diagonal, and  $C$  is the loadings matrix, which maps from the latent space to the observed space, and  $\mathcal{N}(\mu, \Sigma)$  denotes a multivariate normal distribution with mean  $\mu$  and covariance  $\Sigma$ . The shared component of network variability captured by FA was calculated by  $C^T C$ , and a private uncorrelated component was the diagonal matrix  $Q$ . To have enough trials to robustly fit the FA model, stimuli were combined across the six CV speakers of the stimulus set. To ensure unique and informative stimulus responses, only electrodes with  $\bar{R}^2 > 0.5$  during the encoding period were used for each subject, approximately

half of the STG electrodes for each subject.  $H\gamma$  was z-scored across trials for each set of stimulus responses on each electrode, stimulus, and time point. We determined the dimensionality of the neural data by conducting principle component analysis on the average response across stimuli and the pre-stimulus period (Churchland et al., 2010). We found that 5, 7, and 6 dimensions were required to explain 95% of the variance in each subject, respectively. To ensure that we did not overestimate the dimensionality of the data, we modeled the data in a subspace with a dimensionality of 5, the minimum across subjects, which explained 96%, of the variance in the data for S1, 92% for S2, and 93% for S3. FA was conducted independently on each time point. The proportion of variability that was shared was calculated by mean trace  $C^T C$ . This is a proportion because the z-score makes the total variance mean trace  $C^T C + Q = 1$  across time. Qualitatively similar results were observed if no common-average reference is performed.

#### *Relationship between signal and noise correlations.*

To determine whether the reduction in variability was in the directions that benefited stimulus discriminability, we compared signal and noise correlations for each pair of electrodes (Cohen and Maunsell, 2009). “Signal” is the mean  $H\gamma$  during the encoding period, and “noise” is the residual of this activity. Pairwise signal correlations were computed for each pair of electrodes during the encoding period using Pearson correlation. Noise correlations were calculated immediately before and 300 ms after stimulus onset. We calculated the average change in noise correlation. We also calculated the correlation between signal correlation and change in noise correlations before and after stimulus presentation across electrode pairs for each subject.

#### *Decoding.*

As the brain responds to stimuli, the mean response changes, distinguishing between different sounds. The correlation between electrodes also changes, which can have additional effects on the discriminability between sounds. To determine the extent to which noise correlations affect discriminability, we developed an analytical method that imposes the noise correlations of the pre-stimulus period during the stimulus response. The high gamma activity was mean-subtracted for each unique stimulus. Then the response was whitened by pre-multiplying by  $\Sigma_{encoding}^{-\frac{1}{2}}$  and “colored” by the prestimulus noise covariance by premultiplying by  $\Sigma_{pre-stim}^{\frac{1}{2}}$ . The stimulus response is now warped to express the same covariance as the prestimulus activity. The warped activity therefore has pre-stimulus noise correlations, but also the variances and covariance determinant, which can also affect discriminability. To determine the effects solely of correlations, we pre-multiply by an additional diagonal matrix,  $K$ . This matrix scales the warped responses so that they match the relative variances and the determinant of the original stimulus response covariance matrix. Because  $K$  is diagonal, this does not affect the correlations. In summary,

$$\tilde{X}_{warped} = K \Sigma_{pre-stim}^{\frac{1}{2}} \Sigma_{encoding}^{-\frac{1}{2}} \tilde{X}_{encoding} \quad (4)$$

was used, where  $\tilde{X}$  are residual activity matrices,  $\Sigma$  are covariance matrices, and  $K$  is the diagonal matrix that scales the result to impose the relative variances and covariance determinant of the encoding responses. Finally, the means were added back, which resulted in the final warped encoding signal. We examined the response both before and after warping and trained linear classifiers on the warped and original data using linear discriminant analysis.

#### *Analysis across frequency bands.*

We explored the relationship between variability decrease and stimulus representation across frequency bands using the canonical frequency bands theta (4–8 Hz),  $\alpha/\mu$  (8–13 Hz), (13–30

Hz), gamma (30–70 Hz), and high gamma (70–150 Hz) (Mackay, 1997; Canolty et al., 2006; Crone et al., 2011). Analytic amplitude within these bands was calculated analogously to the methods used for  $H\gamma$ .  $\bar{R}^2$ ,  $D$ , and decoding accuracy are calculated across time for each frequency band.

## Results

To understand the role of neural variability in response to speech sounds, we presented subjects with auditory playback of 288 distinct CV syllables. The stimuli consisted of 16 consonants followed by either /a/, /u/ or /i/ (cardinal vowels), spoken by six different speakers, and were chosen to sample the acoustic and phonological space of American English. We recorded neural activity directly from the surface of the STG in the language dominant hemisphere with high-density ECoG arrays. We examined the structure and dynamics of neural responses across multiple presentations of different syllables (across-syllable variability), and compared this with neural responses across multiple presentations of the same syllable (within-syllable variability) to examine how STG encodes different sounds with distinct representations.

### *Mean dependence of response variability*

Figure 1A shows the STG electrodes for one patient (S1), with a single example electrode highlighted in yellow. Figure 1B (top and middle rows) displays the amplitude waveform and spectrogram for the syllables /di/ and /si/, the consonants of which have very different acoustic structure.

We focused on the cortical response in the high-gamma frequency component of field potentials ( $H\gamma$ , 70–150 Hz), which correlates well with multiunit firing rates (Rasch et al., 2008; Ray et al., 2008; Whittingstall and Logothetis, 2009; Ray and Maunsell, 2011; Rey et al., 2014). The  $H\gamma$  responses from the highlighted electrode evoked by these syllables are displayed in Figure 1B.

Responses increased shortly after the onset of both stimuli and returned to baseline by ~500 ms. In this example, /di/ caused greater mean activity than /si/, with the peak difference occurring at 200 ms. Additionally, the response to /di/ had greater variability, as is evident from its broader error bars, which show SD and are compared side-by-side to the right. These example stimuli illustrate the general trend that the variance of  $H\gamma$  responses for a stimulus was positively correlated with the mean. Figure 1C shows the mean and variance of responses across repetitions of each of the 288 unique auditory stimuli for the example electrode. We found a strong positive correlation (Pearson correlation:  $r = 0.6$ ) between mean and variance, as illustrated by the yellow dashed line, which is the best linear fit between the two.

Similar results were observed across all STG electrodes. Figure 1D shows the correlation coefficient calculated in the same manner for all of the STG electrodes across each of the three subjects. We found statistically significant positive correlations between mean and variance of  $H\gamma$  responses across all of the STG electrodes that we recorded ( $r = 0.49 \pm 0.12$ , *mean*  $\pm$  *SD*;  $N = 179$ ; vertical dashed line indicates statistical significance). The positive correlation between mean and variance in His similar to the relationship between mean and variance that has been well studied in the response statistics of single neurons (Tolhurst et al., 1983), and is consistent with the hypothesis that variability of ECoG  $H\gamma$  reflects variability in the activity of underlying populations of neurons.

#### *Reduction in variability following stimulus onset*

We next investigated how the relationship between mean and variance of neural responses changed through time during the representation of a stimulus. We hypothesized that this relationship would decrease, consistent with experiments (Cohen and Maunsell, 2009; Churchland et al., 2010) and models (Litwin-Kumar and Doiron, 2012) of single neuron firing. Figure 2 illustrates the dynamics of the relationship between the mean and variance of neural

responses. We quantified the relationship between the mean to the variance at each time point using linear regression (Eq. 1). The slope of the regression ( $D$ ) is our metric of relative variability and is similar to FF, except that it is calculated with the inclusion of a y-intercept. To illustrate this procedure and metric, five time points are shown in Figure 2A: 200, 0, 200, 400, and 600 ms. Each point of each panel in Figure 2A shows the mean and variance of  $H\gamma$  responses for a stimulus on an electrode of Subject S1. In each panel, a regression was performed, and the best-fit line is shown as a red dashed line. We observed the slope of the best-fit line was  $\sim 1$  before the acoustic onset of the stimulus, decreased following the acoustic onset of the stimulus, and slowly returned to 1. The regression was performed at every sample time and is shown in Figure 2B. Subject S1, the example subject for Figure 2A, is shown in red, and Subjects S2 and S3 are shown in green and blue, respectively. The slope decreased sharply within 100 ms of stimulus onset in each of the three subjects, and this reduction was sustained for 200–300 ms. During the period between 100 and 400 ms after stimulus onset,  $D$  decreased by 70%, 38%, and 42% for S1, S2, and S3, respectively.

The decrease in relative variability is similar to the “quenching” in FF of firing rates after stimulus onset previously observed in animal studies (Cohen and Maunsell, 2009; Churchland et al., 2010).

Additionally, to ensure that the change in  $D$  was not due to a difference in the mean distribution, a version of “mean matching,” (Churchland et al., 2007, 2010) modified for ECoG, was also performed. Here, stimulus-electrode pairs were randomly excluded from each time point until only a subset of stimulus-electrode pairs remained that have the same mean distribution through time. Although mean-matching removed approximately half of stimulus-electrode pairs at each time point (S1: 56%, S2: 66%, S3: 50%), it did not qualitatively change the trend of a sharp decrease in  $D$  (Fig. 2B, light colors). Together, these results indicate that the relative

magnitude of neural variability is dynamically quenched during the presentation of speech stimuli.

#### *Co-occurrence between stimulus encoding and reduction in variability*

In order for changes in variability to affect encoding strength, these changes must coincide with the neural response that discriminates between stimuli, and must be present at electrodes that discriminate stimuli. To ascertain the role of reduced neural variability in the representation of stimuli, we examined its relationship with stimulus encoding. Specifically, we tested the hypotheses that the decrease in D temporally coincides with the encoding of a stimulus, and that electrodes which contribute most to the overall decrease in D are the electrodes that most strongly encode the stimulus. Encoding strength was measured using the coefficient of determination adjusted for degrees of freedom,  $\bar{R}^2$ , which quantifies how much of the variance in the response of an electrode could be explained by which stimulus was presented (Eq. 2).

We found that dynamics of stimulus encoding co-occurred spatially and temporally with the reduction in response variability. The encoding of syllables across the STG for Subject S1 is shown in Figure 3A. To determine the dynamics of encoding strength,  $\bar{R}^2$  was calculated for each STG electrode across the trial, time aligned to the acoustic onset of the stimulus (Fig. 3A, dashed black line). Here, electrodes are ordered by position on the STG, with posterior electrodes on the top and anterior electrodes on the bottom. The ordering of the electrodes was determined by eye. All electrodes had a  $\bar{R}^2$  near 0 before the acoustic onset of the stimulus, which indicates that their variability was not predicted by the upcoming stimulus. Following the stimulus onset,  $\bar{R}^2$  increased first in the posterior electrodes (Hickok and Poeppel, 2000) and peaked in the anterior electrodes over the first 100 ms (DeWitt and Rauschecker, 2012). Similar spatiotemporal structures were observed in the other subjects. Figure 3A (bottom) shows the average  $\bar{R}^2$  across all STG electrodes for each of the three subjects. For each subject, the



mean  $\bar{R}^2$  was maximum at ~100 ms and slowly decreased over the next 900 ms. Qualitatively, the dynamics of average  $\bar{R}^2$  had a similar time course to the decrease in  $D$  shown in Figure 2B. Similar results were obtained with a linear discriminant classifier decoding stimulus identity from  $H\gamma$ .

To quantitatively compare the dynamics of  $\bar{R}^2$  to  $D$ , we used a cross-correlation analysis for each subject (Fig. 3B). The time of the minimum cross-correlation represents the overall lag between decreased variability and increased encoding strength. We found that the time lags were close to 0 for each subject, demonstrating that the decrease in variability was nearly synchronous with stimulus encoding, preceding it only slightly for each subject. The large negative correlations close to 1 indicate that encoding strength and variability reduction have similar overall dynamics (S1:  $r = -0.89$  at 5 ms, S2:  $r = -0.70$  at 25 ms, S3:  $r = -0.97$  at 5 ms). These results show that the representation of stimuli occurs simultaneously with the decrease in variability, which is necessary for the shaping of variability to enhance discriminability.

#### *Colocation of stimulus encoding and change in variability*

To determine the degree to which different speech sounds gave rise to different responses, we calculated the variance of the mean responses to the different stimuli (across-stimulus variance), which would be high if the activity of that electrode had different responses for different acoustic stimuli. The across-stimulus variance of each electrode was calculated during the “encoding period,” 100–400 ms after stimulus onset. Figure 3C (left) shows each electrode of Subject S1 with the color of the electrode indicating the variance of the mean response (darker red electrodes had a greater across-stimulus variance).  $D$  was also calculated separately for each electrode across time, and a change in  $D$  was determined by comparing the encoding period with the baseline period (300 ms before stimulus onset to the moment of stimulus onset). The change in variability is shown in Figure 3C (right). Electrodes that are dark

blue had a strong decrease in the linear relationship between mean and variance, and electrodes that are red had a slight increase. Figure 3D summarizes the relationship between stimulus variance and variability for all three subjects. We found a robust negative correlation between interstimulus variance and D across subjects ( $r = -0.56 \pm 0.09$ ; slope of best fit regression line =  $-5.4 \pm 1.2$ ; both are 95% CI), indicating that the electrodes with a high interstimulus variance tended to also have a greater decrease in variability. The  $\bar{R}^2$  of  $H\gamma$  of each electrode during the encoding period is indicated by color. Electrodes with a high across-stimulus variance and a strong decrease in variability had the strongest encoding strength. Together with the previous analyses, these results demonstrate that the drop in variability and increase in stimulus encoding were nearly synchronous in time, and co-localized to the same electrodes.

#### *Reduction in population variability improves discriminability*

Phonetic features are encoded by a spatially distributed network of activity along the STG (Mesgarani et al., 2014). In distributed representations, noise in the activity of individual sites that is uncorrelated across the sites can be removed through averaging. However, if the variability is correlated across the different sites, averaging will not remove it, and this correlated noise can have a large impact on discriminability (Zohary et al., 1994). We have observed a change in variability coincident with stimulus representation, but the effect of this variability on strength of encoding depends on the shape of the variability. To understand how the decrease in variability was distributed across all of the electrodes, we used factor analysis (Eq. 3) (Churchland et al., 2010). Factor analysis distinguishes between correlated and uncorrelated variability by modeling the data with a generative process that contains a correlated subspace that is “shared” across observations, and corrupted by uncorrelated “private” noise. We hypothesized that, during speech perception, shared variability in the STG neural population would decrease more than the uncorrelated variability, and that the dynamics would be closely

matched to the dynamics of variability reduction at individual sites. This hypothesis is illustrated as a schematic in Figure 4A. Here, factor analysis finds a subspace, depicted by the ellipse that spans mostly Electrode 1 and Electrode 2. The activity is then corrupted by Gaussian noise, which extends activity mostly into the direction of Electrode 3. The arrows illustrated a reduction in shared variability, marked by a shrinking of the ellipse and a relative increase in variability in the direction of Electrode 3, the uncorrelated electrode. The resulting noise structure with less shared noise will be more spherical.

For each subject, we modeled the shared component of  $H\gamma$  variability with a 5-dimensional subspace (see Materials and Methods). Factor analysis was performed separately on z-scored  $H\gamma$  for each sample time of each consonant-vowel stimulus. The overall difference between subjects in the shared variance is explained by the number of electrodes on the STG of each subject. Subject S2 has the fewest electrodes and the most shared variance, and Subject S3 has the most electrodes and the least shared variance. Following stimulus onset, there was a pronounced decrease in the proportion of the population variance that was correlated across electrodes, and the private noise makes up a larger proportion of the variance in each of the three subjects (Fig. 4B). This shows that neural variability is primarily constrained in the dimensions that are most variable during multiple presentations of a stimulus.

Figure 5A shows a schematic of how noise correlations might affect discriminability at the example electrode in Figure 1. In the schematic, “di” causes a higher response than “si” in both electrode 1 and 2 (both 1 and 2 respond like the example electrode in Fig. 1). Because the two electrodes respond similarly to stimuli, they have positive signal correlation. The signal is corrupted by noise, indicated by the ellipses around the mean responses. The orange ellipses show noise that is positively correlated, which lowers the discriminability of the sounds. By reducing the magnitude of positive correlations, the distance between the ellipses increase.

Thus, discriminability improves if the positive correlations are reduced, as shown by the blue ellipses.

To determine how changes in noise correlations affect stimulus discriminability, we examined the relationship between signal and noise correlations for each pair of electrodes before the stimulus and during representation (Cohen and Maunsell, 2009) (Fig. 5B). We find that noise correlations between electrode pairs are uniformly positive, and larger for electrodes with higher signal correlation, both of which are in agreement with neuronal firing rate correlations (Kohn and Smith, 2005; Cohen and Kohn, 2011). The majority of pairs (90%) of electrodes have a positive signal correlation. Noise correlations decreased on average during stimulus encoding for each subject

( $S1 = \Delta r_{noise} = -0.018 \pm 1.7e - 3$ ;  $S2 = \Delta r_{noise} = -0.034 \pm 1.1e - 3$ ;  $S3: \Delta r_{noise} = -0.014 \pm 1.6e - 3$ ; 95% CIs). Moreover, electrode pairs with positive correlations have more strongly reduced noise correlations, although this was only statistically significant in S1 and S3

( $S1: r = 0.28 \pm 0.04$ ;  $S2: r = 0.05 \pm 0.058$ ;  $S3: r = 0.24 \pm 0.04$ ; 95% CIs,  $N = 1711, 1128, \text{ and } 2556$ ). This is shown as an increasing difference between noise correlations before and during representation for higher signal correlations. These results provide another key insight into the shape of variability: the reduction in variability during stimulus representation is stronger in the directions of neural space that benefits discriminability.

Recent theoretical work has demonstrated that noise correlations alone can give a misleading impression of change in strength of representation (Moreno-Bote et al., 2014). To determine the extent to which the observed correlations change the discriminability of speech sounds, we imposed the prestimulus correlations to the neural activity during the encoding period and quantified the difference in the accuracy of a decoder trained to distinguish consonants across time. To evaluate the effect of the change in correlations on discriminability across the STG, we

evaluated the change in decoding performance at 200 ms for all pairs of electrodes with  $\bar{R}^2 > 0.5$  (Fig. 5C). Pairs that had a greater signal correlation tended to discriminate better with the original signal than with the warped signal, whereas electrodes that had anti-correlated responses performed better with the warped signal ( $r = 0.27$ ;  $p < 1e-4$ ). However, warping the signal correlations had no effect on the decode accuracy of a classifier using all electrodes. Although the noise correlations were shaped by the similarity in stimulus response, these changes in the noise did not affect decoding performance across the STG.

#### *Variability reduction was correlated with discriminability across frequency bands*

Although we have emphasized  $H\gamma$ , other frequency bands of the ECoG field potential may play important roles in speech processing. To explore these other bands, we compute the variability and encoding strength for the canonical frequency bands: theta (4 – 8 Hz),  $\alpha/\mu$  (8 – 13 Hz),  $\beta$  (13–30 Hz), gamma (30 –70 Hz), and high gamma (70 –150 Hz) (Mackay, 1997; Canolty et al., 2006; Crone et al., 2011). The time course of encoding and variability across electrodes for Subject S1 are shown in Figure 6, which is representative of the three subjects. Each of the standard frequency bands is represented by a different color. Variability primarily decreases in gamma (purple) and  $H\gamma$  (green) (Fig. 6A). In all bands, mean  $R^2$  was 0 before stimulus onset, rose after the acoustic onset of the stimulus, and decreased over the next 1 s (Fig. 6B).  $H\gamma$  (green) is shown for comparison, and has the strongest encoding. Figure 6C summarizes the across-frequency analysis of the three subjects. In each subject,  $H\gamma$  has the greatest drop in variability and the greatest encoding strength. Moreover, there was a strong negative correlation (Pearson  $r = -0.93$ , slope of line of best fit  $= -5.4e-3 \pm 1.2e-3$ , 95% CI) between the encoding strength of a band and the drop in variability. This suggests that variability shaping is a primary encoding mechanism that distinguishes high-gamma from other frequency bands.

## Discussion

Speech perception relies on the ability to discriminate all of the different sounds of a given language. This requires that the representations within a sound category are similar to each other, and that the representations across sound categories are sufficiently distinct. Here, we examined the human nonprimary auditory cortex (STG) for encodings of speech sounds that underlie their discrimination. High-density ECoG permitted the study of variability on the mesoscopic level, with the spatiotemporal resolution necessary to observe neural dynamics associated with phonetic perception. We found a strong positive correlation between cortical response magnitude and variance that decreased dramatically following onset of an auditory stimulus, similar to neuronal action potentials across several species and brain areas. These results are evidence that  $H\gamma$  trial-by-trial fluctuations are not merely measurement noise but reflect trial-by-trial differences in firing rates.

$H\gamma$  on a single electrode reflects neural activity across an estimated several thousand neurons under that electrode (Miller et al., 2009). Previous studies of firing rate responses have found that neuronal variability decreases following stimulus onset in a shared subspace across neurons (Cohen and Maunsell, 2009; Churchland et al., 2010). However, these previous studies recorded from at most a few hundred neurons at a time. Our study expands upon previous findings, showing that the decrease in shared variability extends to a much larger neural population. The brain is thought to process sensory stimuli with large populations of neurons (Hinton, 1984). Therefore, variability of  $H\gamma$  on individual electrodes may reflect network properties that are important for sensory representation and computational processing.

The change in variability was monitored through the slope of the regression relating neural response mean to variance, which we called  $D$ .  $D$  is an analog of FF for ECoG data analysis; however, care must be taken in its interpretation. A FF of 1 is meaningful because it matches a

Poisson process, but  $D$  should not be interpreted this way; only relative changes in  $D$  are meaningful. Nevertheless, the decrease in  $D$  observed here is consistent with the hypothesis that variability in  $H\gamma$  reflects variability in the activity of the underlying neuronal population.

One concern could be that the relationship between mean and variance was not changing, but instead it was simply a change in the mean distribution that was causing this change in  $D$ . If the variance of response does not change,  $D$  could decrease simply due to normalization by greater mean responses. Another possible concern is that there is an upper bound of activation for each electrode due to physiological limitations and that this upper bound might lower variability of high responses due to a ceiling effect. The mean matching control analysis addresses both of these concerns. Furthermore, our model (Eq. 1) includes a linear offset and a slope, capturing the additive and multiplicative noise components, respectively. Additive noise was captured in the offset term, so the slope,  $D$ , reflects specifically multiplicative noise. If the mean changed but the correlation with variance did not, this would be reflected in the offset term, not  $D$ .

Computational models have provided some explanations for how a decrease in variability could be achieved mechanistically. In models of populations of neurons, networks that are clustered into subgroups exhibit a more stable response to stimuli, lowering the variability of activity (Litwin-Kumar and Doiron, 2012). Such subgrouping organization may also account for the decrease in response variability we observed in field potentials. Alternatively, reduction in ECoG  $H\gamma$  variability could reflect an increase in the synchrony of neural activation of the underlying population (de la Rocha et al., 2007).

To determine the effect of variability on stimulus representation, we compared it with a stimulus-encoding model. We found that the reduction in variability occurred nearly simultaneously with peak stimulus discriminability and was spatially specific to those electrodes that discriminate between different sounds. If variability reduction had been primarily on electrodes that did not

have distinct responses for different sounds, or if it had a very different time course from stimulus representation, the decrease in variability would have had no influence on discriminability. Yet we show that variability decrease does occur together with stimulus representation, which is necessary for changes in variability to influence discriminability.

The effect of changes in variability on the discriminability of stimuli depends on how these changes are correlated across electrodes. We used factor analysis to determine the extent to which changes in variability were correlated across electrodes. The use of factor analysis is similar to a previous work from He and Zempel (2013), where variability was examined using principal components analysis. Here we use factor analysis to model variability explicitly, and apply it to  $H\gamma$  rather than broadband voltage traces. We found that the reduction in variability was primarily shared in the subspace of noise correlations, rather than private to individual electrodes, similar to previous work from Churchland et al. (2010), who studied noise correlations between single neurons. Combined, our findings suggest that a reduction in noise correlations is a multiscale neural phenomenon that may play an important computational role in representation on several spatial scales.

The decorrelation of noise we observe during stimulus representation may be a consequence of feedforward inhibitory input. Computational modeling has demonstrated that variability shaping can be caused by nonlinearities that alter the distribution of neural activity. Electrodes with positive signal correlations are de-correlated during stimulus representation because they receive a common feedforward input that changes the response distribution (Middleton et al., 2012). However, it is unclear how this change in the shape of variance on the scale of neurons would generalize to aggregate neural populations in ECoG.

Noise correlations can either decrease or increase discriminability of stimuli, depending on their relationship with the signal correlations. The “sign rule” (Hu et al., 2014) indicates that noise



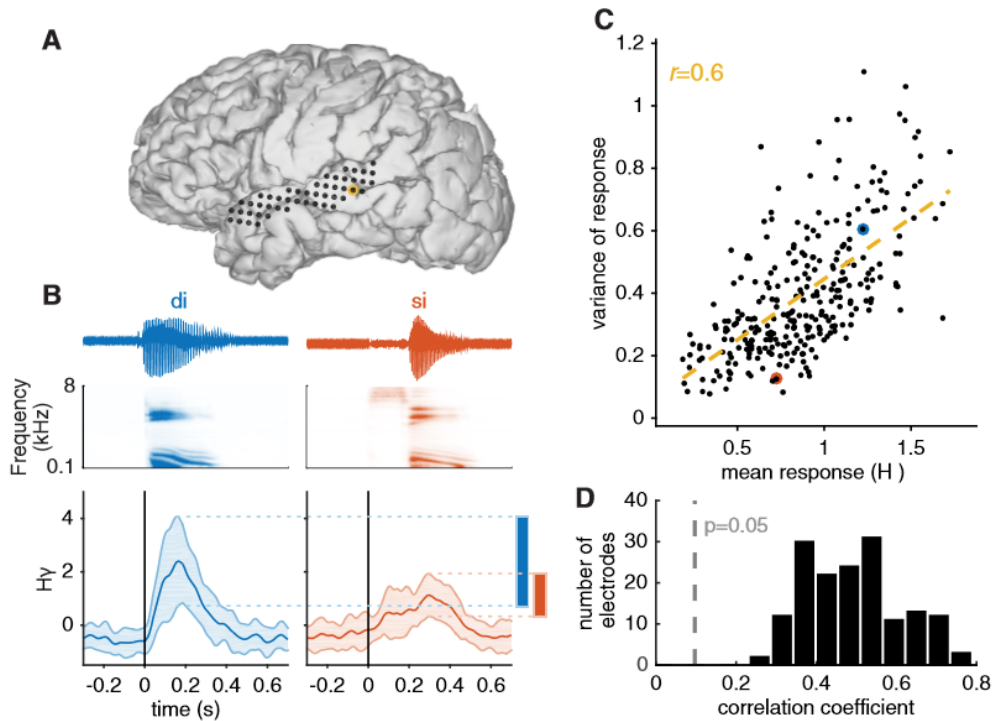
correlations that are correlated with signal correlations decrease stimulus discriminability. On the other hand, noise correlations that are anti-correlated with signal correlations facilitate discriminability (Abbott and Dayan, 1999). We found the largest decrease in noise correlations between electrodes that are similarly tuned. According to the sign rule, these noise correlations lead to less discriminable representations. Therefore, the noise correlations that are mostly strongly diminished during representation are those that most hinder the discrimination of sounds. We used a linear decoder to quantify the change in information due to noise correlations (Moreno-Bote et al., 2014) and found a small but consistent effect for electrodes with high signal correlations.

Our results show changes in the correlation structure that improves the discriminability of stimuli for electrodes with high signal correlations, but the improvement is very slight and only on a subset of electrode pairs. Previous studies have found a larger effect on discriminability of neuron firing rates from multiunit recordings (Cohen and Maunsell, 2009). One possible explanation is that the effect of noise correlations on discriminability is indeed very small for this task, but we think a more likely explanation is that the improvement in discriminability is stronger on the single-neuron level when recording densely from a relatively small volume of cortex. In contrast, ECoG samples over a very large area of cortex, and the average pairwise signal correlation between ECoG electrodes is typically less than single-unit recordings (Downer et al., 2015). Our results imply that decoding activity from ECoG may be less sensitive to the noise structure of neural activity than from recordings of single units.

We compared variability reduction across frequency bands and found that  $H\gamma$  activity has the highest encoding strength and that it exhibits the strongest reduction in variability upon stimulus onset. The gamma band, which also has been shown to modulate with firing rate (Cui et al., 2016), shows a similar but weaker reduction in variability. These results suggest not only that higher bands in particular hold the most information about the acoustic stimuli, but also that they

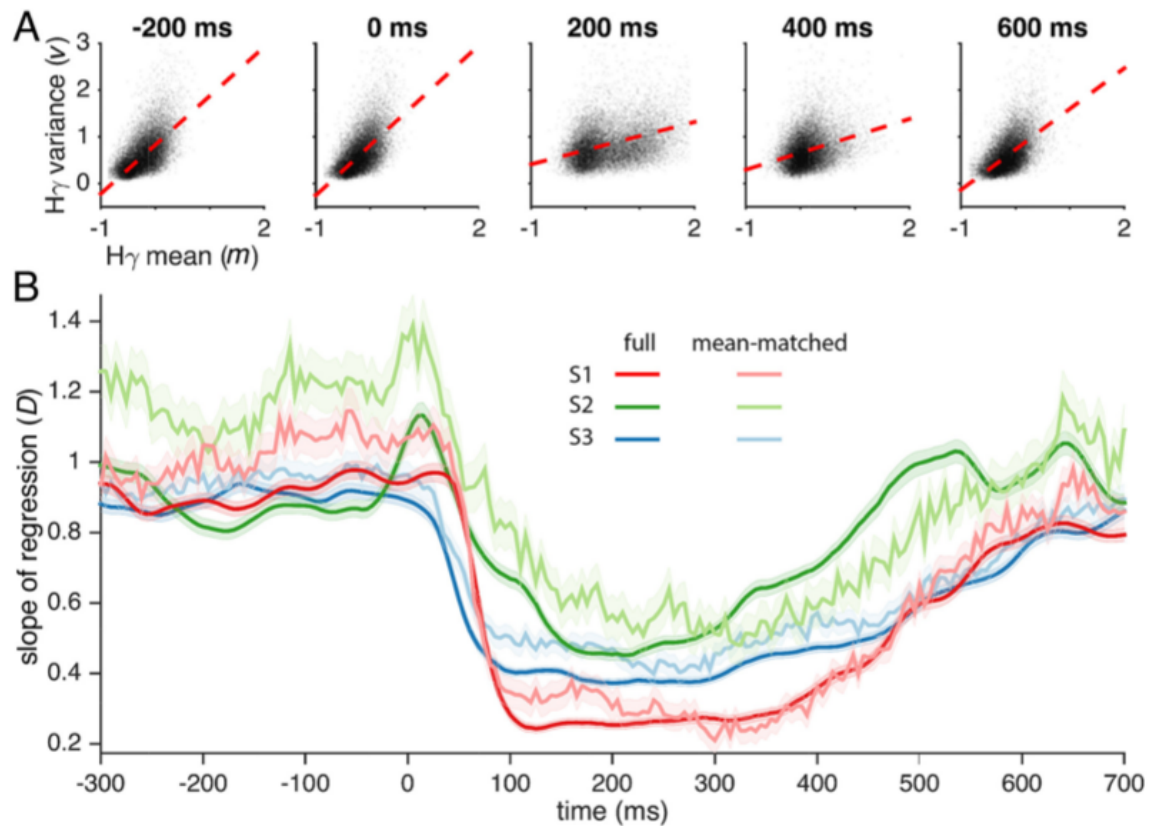
uniquely exhibits variability changes similar to those found in neurons. These results provide further evidence that spectral power modulation in the higher range is distinct from lower bands, reflecting excitation of neuron populations (Rasch et al., 2008; Ray et al., 2008; Crone et al., 2011; Buzsaki et al., 2012).

A complete understanding of neural representation requires not only the mean response to sensory stimuli, but the entire distribution of neural responses. Variability is an integral part of stimulus representation, and changes dynamically as the stimulus is represented. Indeed, we find that neural activity is dynamically shaped in a manner that enhances the discriminability of sounds.

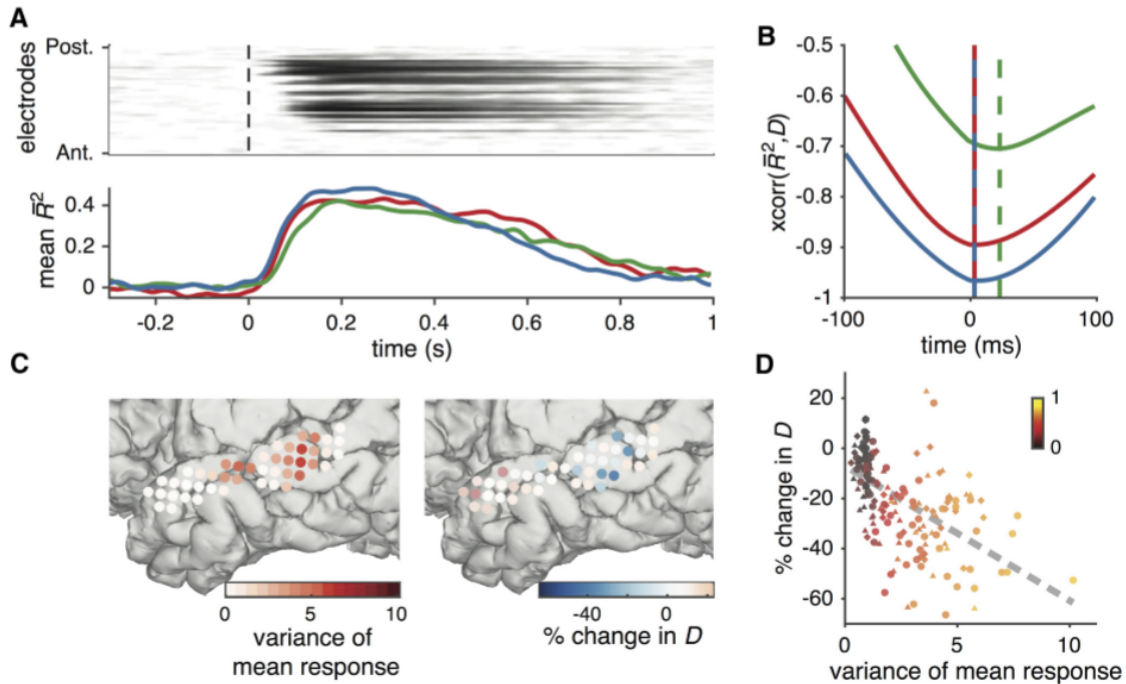


**Figure 1: Example response and correlation between mean and variance of response. A,**

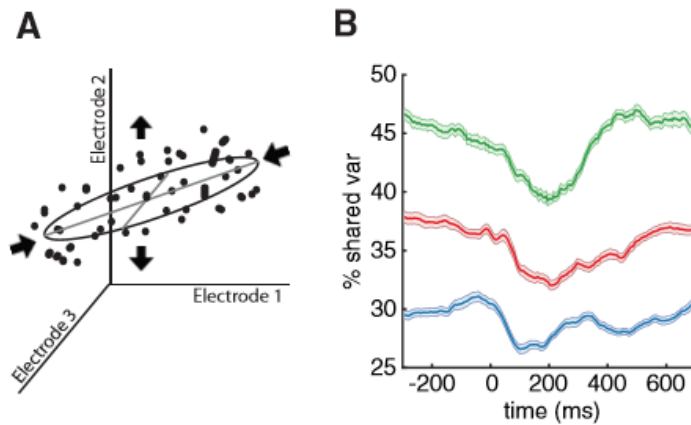
Placement of high-density ECoG electrodes on the cortex of an example human subject. The electrode size is anatomically correct. Yellow electrode is the example electrode used for B and C. **B,** Top, Example consonant-vowel sounds, /di/ and /si/ with acoustic waveform spectrogram. Bottom, Mean and SD of  $H\gamma$  responses across task time. Rectangles to the right compare the SDs of the two response distributions and illustrate that /di/, the sound that elicits a greater mean response, also elicits a greater variance of responses. **C,** Mean and variance of  $H\gamma$  over trials of every stimulus for the example electrode (100–400 ms after stimulus onset). This electrode has a positive Pearson correlation of 0.6, indicating that, for this electrode, stimuli that elicit a greater mean response tend to elicit a greater variance of response. **D,** The Pearson correlation coefficient between mean and variance of  $H\gamma$  responses for each electrode across all three subjects. All electrodes have a correlation greater than chance, which is 0.01 ( $p < 0.05$ ).



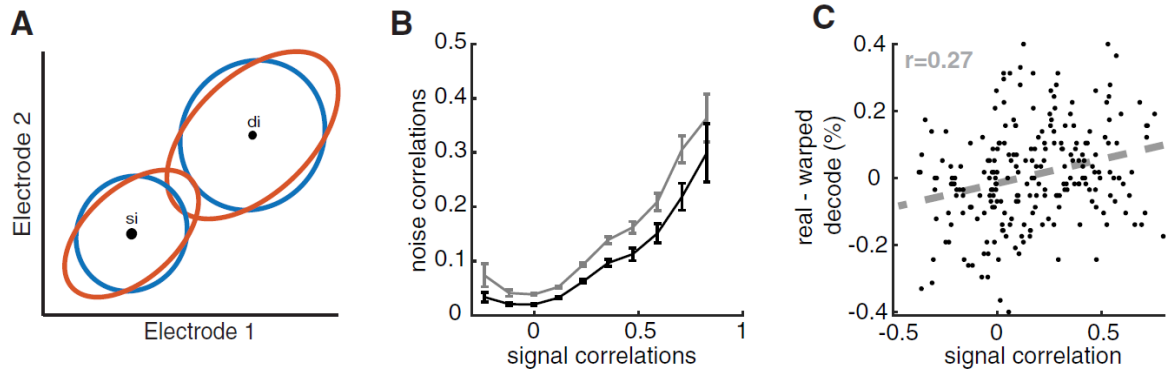
**Figure 2: Reduction in mean dependence of variance.** **A**,  $H\gamma$  for all stimulus-electrode responses for Subject S1. Red lines indicate the result of linear regression with a y-intercept. The regression line has a slope of  $\sim 1$  before and during stimulus onset, decreases 200 and 400 ms after stimulus onset, and returns to 1 by 600 ms. **B**, Slope of regression for each point in time for Subjects S1 (red), S2 (green), and S3 (blue). Lighter shade represents results for mean-matched  $D$ , where regression was performed on a subset of stimulus electrodes with a constant mean distribution through time.  $D$  decreases sharply after acoustic onset for mean-matched and regular regression. Data are mean  $\pm$  SEM.



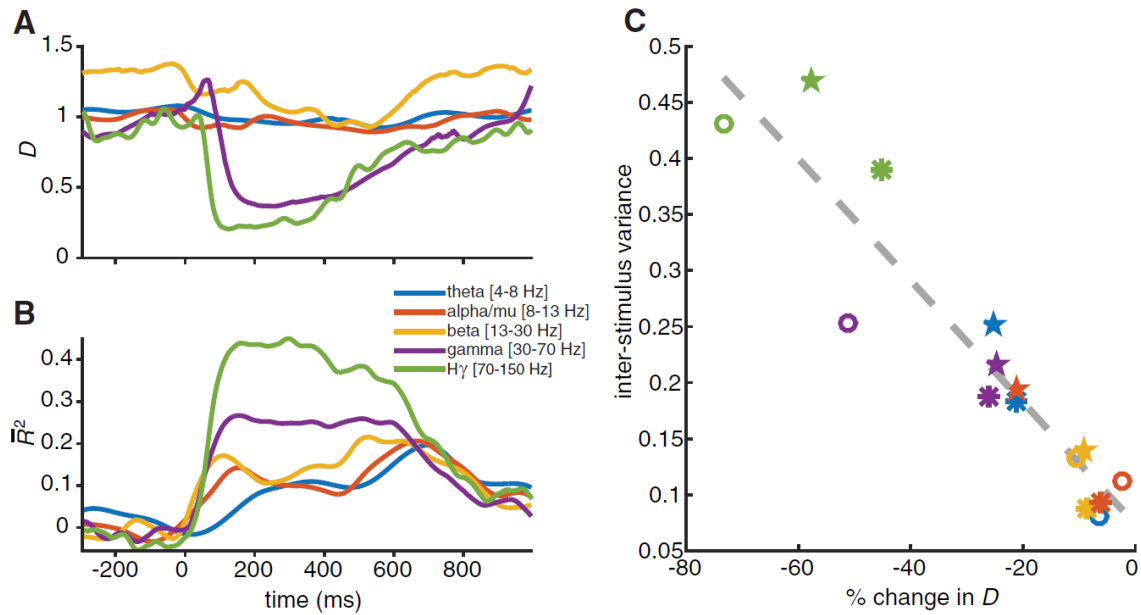
**Figure 3: Temporal and spatial correlation between encoding and drop in variability.** **A**,  $\bar{R}^2$  through time for each electrode on the superior temporal gyrus (STG) of S1 with electrodes ordered from posterior to anterior region of the STG. Mean  $\bar{R}^2$  across electrodes is shown below for S1 (red), S2 (green), and S3 (blue). **B**, Cross-correlation between variability ( $D$ ) and  $\bar{R}^2$  for the three subjects. Minimum cross-correlation is marked as dashed lines colored by subject and was close to 0 in all three subjects. **C**, Left, For S1, the variance of the mean response during the encoding period (100–400 ms) is shown for each electrode. Darker electrodes are those that have more different responses for different stimuli. Right,  $D$  was calculated for each electrode independently. Electrodes are colored according to the change in  $D$  from the average of the pre-stim period (300 to 0 ms) to the average of the encoding period (100–400 ms). **D**, Mean response variance and change in  $D$  are shown for each electrode across all three subjects. Individual electrodes from S1 (triangles) S2 (diamonds), and S3 (circles). Electrodes are also colored according to  $\bar{R}^2$ . Gray dashed line indicates the line of best fit. The negative correlation shows that electrodes that respond differently to different stimuli tend to have a reduced  $D$  during stimulus encoding. Data are mean  $\pm$ SEM.



**Figure 4: Factor analysis.** **A**, Factor analysis schematic illustrating a change in variability from more correlated to less correlated. Shown is the activity on three example electrodes for a single stimulus. Oval represents correlated variability in a 2D subspace of the three electrodes. Dots represent individual data points after the addition of private uncorrelated noise. Arrows indicate a reduction in shared variability and an increase in private noise. The schematic represents a 2D subspace for illustrative purposes, but the data were actually modeled on a 5-dimensional subspace. **B**, Factor analysis results. The proportion of variability that was in the shared subspace is shown over time for Subjects S1 (red), S2 (green), and S3 (blue). Although the portion of variability that was shared across electrodes differs between subjects, it decreases upon stimulus onset in all three subjects. Data are mean  $\pm$ SEM.



**Figure 5: Noise correlations and decoding.** **A**, Schematic of signal and noise correlations. The responses of two hypothetical electrodes are shown for two stimuli, “si” and “di.” Circles in the center of the ellipses represent mean responses for these stimuli. Both of these electrodes have a stronger response to “di” than to “si,” like the real example electrode in Figure 1. Orange ellipses represent response distribution with a positive noise correlation because the major axis of the ellipse has a positive slope. Blue ellipses represent response distributions with reduced noise correlations, which would improve discriminability by increasing the separation between the ellipses. **B**, Noise correlation as a function of signal correlation across all electrode pairs, for example, Subject S1: gray represents noise correlations just before stimulus onset; black represents noise correlations at 200 ms after stimulus onset. Data are mean  $\pm$ SEM. **C**, The effect of prestimulus noise correlations on stimulus discriminability of speech sounds is illustrated on two example electrodes. Leave-one-out linear discriminant analysis is used to test the separability of consonants. The true stimulus response is compared with a warped response that has noise correlations of the prestimulus period at 200 ms. The difference between decoding performance for original and warped activity for each electrode pair for Subject S1. Electrode pairs that had a greater signal correlation had a greater improvement of decoding compared with the warped signal decode (Pearson  $r=0.27$ ;  $p<1e-4$ ).



**Figure 6: Multiband change in variability and encoding.** **A**, The linear relationship between mean and variance of response, for each of the functional bands through time.  $H\gamma$  and gamma have the strongest modulation in variability ( $D$ ), and the lower frequency bands show little or no modulation. **B**, Encoding strength shown as  $\bar{R}^2$  across time for each of the functional bands.  $H\gamma$  shows the greatest increase in  $\bar{R}^2$ . **C**, Encoding strength versus change in  $D$  for all subjects (S1: circles S2: asterisks S3: 5 stars) during the encoding period (100–400 ms after stimulus onset). Gray dashed line indicates line of best fit.



## References

- Abbott LF, Dayan P (1999) The effect of correlated variability on the accuracy of a population code. *Neural Comput* 11:91–101. [CrossRef Medline](#)
- Averbeck BB, Latham PE, Pouget A (2006) Neural correlations, population coding and computation. *Nat Rev Neurosci* 7:358–366. [CrossRef Medline](#)
- Buzsa'ki G, Anastassiou CA, Koch C (2012) The origin of extracellular fields and currents: EEG, ECoG, LFP and spikes. *Nat Rev Neurosci* 13:407–420. [CrossRef Medline](#)
- Canolty RT, Edwards E, Dalal SS, Soltani M, Nagarajan SS, Kirsch HE, Berger MS, Barbaro NM, Knight RT (2006) High gamma power is phase-locked to theta oscillations in human neocortex. *Science* 313:1626–1628. [CrossRef Medline](#)
- Chang EF (2015) Towards large-scale, human-based, mesoscopic neurotechnologies. *Neuron* 86:68–78. [CrossRef Medline](#)
- Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT (2010) Categorical speech representation in human superior temporal gyrus. *Nat Neurosci* 13:1428–1432. [CrossRef Medline](#)
- Churchland AK, Kiani R, Chaudhuri R, Wang XJ, Pouget A, Shadlen MN (2011) Variance as a signature of neural computations during decision making. *Neuron* 69:818–831. [CrossRef Medline](#)
- Churchland MM, Yu BM, Sahani M, Shenoy KV (2007) Techniques for extracting single-trial activity patterns from large-scale neural recordings. *Curr Opin Neurobiol* 17:609–618. [CrossRef Medline](#)

Churchland MM, Yu BM, Cunningham JP, Sugrue LP, Cohen MR, Corrado GS, Newsome WT, Clark AM, Hosseini P, Scott BB, Bradley DC, Smith MA, Kohn A, Movshon JA, Armstrong KM, Moore T, Chang SW, Snyder LH, Lisberger SG, Priebe NJ, et al. (2010) Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nat Neurosci* 13:369–378. CrossRef Medline

Cohen MR, Kohn A (2011) Measuring and interpreting neuronal correlations. *Nat Neurosci* 14:811– 819. CrossRef Medline

Cohen MR, Maunsell JH (2009) Attention improves performance primarily by reducing interneuronal correlations. *Nat Neurosci* 12:1594 –1600. CrossRef Medline

Crone NE, Korzeniewska A, Franaszczuk PJ (2011) Cortical gamma responses: searching high and low. *Int J Psychophysiol* 79:9 –15. CrossRef Medline

Cui Y, Liu LD, McFarland JM, Pack CC, Butts DA (2016) Inferring cortical variability from local field potentials. *J Neurosci* 36:4121– 4135. CrossRef Medline

de la Rocha J, Doiron B, Shea-Brown E, Josiæ K, Reyes A (2007) Correlation between neural spike trains increases with firing rate. *Nature* 448: 802–806. CrossRef Medline

DeWitt I, Rauschecker JP (2012) Phoneme and word recognition in the auditory ventral stream. *Proc Natl Acad Sci U S A* 109:E505–E514. CrossRef Medline

Diehl RL, Lotto AJ, Holt LL (2004) Speech perception. *Annu Rev Psychol* 55:149 –179. CrossRef Medline

Downer JD, Niwa M, Sutter ML (2015) Task engagement selectively modulates neural correlations in primary auditory cortex. *J Neurosci* 35:7565– 7574. CrossRef Medline

- Faisal AA, Selen LP, Wolpert DM (2008) Noise in the nervous system. *Nat Rev Neurosci* 9:292–303. CrossRef Medline
- Formisano E, De Martino F, Bonte M, Goebel R (2008) “Who” decoding “what”? Brain-based and speech of human voice and speech. *Science* 322:970–973. CrossRef Medline
- Heggelund P, Albus K (1978) Response variability and orientation discrimination of single cells in striate cortex of cat. *Exp Brain Res* 32:197–211. Medline
- He BJ, Zempel JM (2013) Average is optimal: an inverted-U relationship between trial-to-trial brain activity and behavioral performance. *PLoS Comput Biol* 9:e1003348. CrossRef Medline
- Hickok G, Poeppel D (2000) Towards a functional neuroanatomy of speech perception. *Trends Cogn Sci* 4:131–138. CrossRef Medline
- Hinton GE (1984) Distributed Representations. Technical Report CMU-CS- 84-157. Carnegie Mellon University, Pittsburgh, PA.
- Hu Y, Zylberberg J, Shea-Brown E (2014) The sign rule and beyond: boundary effects, flexibility, and noise correlations in neural population codes. *PLoS Comput Biol* 10:e1003469. CrossRef Medline
- Kisley MA, Gerstein GL (1999) Trial-to-trial variability and statedependent modulation of auditory-evoked responses in cortex. *J Neurosci* 19:10451–10460. Medline
- Kohn A, Smith MA (2005) Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. *J Neurosci* 25 3661–3673.

Liberman AM, Harris KS, Hoffman HS, Griffith BC (1957) The discrimination of speech sounds within and across phoneme boundaries. *J Exp Psychol* 54:358–368. CrossRef Medline

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. *Psychol Rev* 74:431–461. CrossRef Medline

Litwin-Kumar A, Doiron B (2012) Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nat Neurosci* 15: 1498–1505. CrossRef Medline

Mackay WA (1997) Synchronized neuronal oscillations and their role in motor processes. *Trends Cogn Sci* 1:176–183. CrossRef Medline

Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. *Science* 343:1006–1010. CrossRef Medline

Middleton JW, Omar C, Doiron B, Simons DJ (2012) Neural correlation is stimulus modulated by feedforward inhibitory circuitry. *J Neurosci* 32: 506–518. CrossRef Medline

Miller KJ, Sorensen LB, Ojemann JG, den Nijs M (2009) Power-law scaling in the brain surface electric potential. *PLoS Comput Biol* 5:e1000609. CrossRef Medline

Mitchell JF, Sundberg KA, Reynolds JH (2009) Spatial attention decorrelates intrinsic activity fluctuations in Macaque area V4. *Neuron* 63:879–888. CrossRef Medline

Moreno-Bote R, Beck J, Kanitscheider I, Pitkow X, Latham P, Pouget A (2014) Information-limiting correlations. *Nat Neurosci* 17:1410–1417. CrossRef Medline

Nourski KV, Steinschneider M, Oya H, Kawasaki H, Jones RD, Howard MA (2014) Spectral organization of the human lateral superior temporal gyrus revealed by intracranial recordings. *Cereb Cortex* 24:340–352. CrossRef Medline

Obleser J, Leaver AM, Vanmeter J, Rauschecker JP (2010) Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization. *Front Psychol* 1:232. CrossRef Medline

Perkell JS, Klatt DH (1986) *Invariance and variability in speech processes*. Hillsdale, NJ: Lawrence Erlbaum.

Rasch MJ, Gretton A, Murayama Y, Maass W, Logothetis NK (2008) Inferring spike trains from local field potentials. *J Neurophysiol* 99:1461–1476. CrossRef Medline

Ray S, Crone NE, Niebur E, Franaszczuk PJ, Hsiao SS (2008) Neural correlates of high-gamma oscillations (60–200 Hz) in macaque local field potentials and their potential implications in electrocorticography. *J Neurosci* 28:11526 –11536. CrossRef Medline

Ray S, Maunsell JH (2011) Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biol* 9:e1000610. CrossRef Medline

Rey HG, Fried I, Quiñones Quiroga R (2014) Timing of single-neuron and local field potential responses in the human medial temporal lobe. *Curr Biol* 24:299 –304. CrossRef Medline

Rose D (1979) An analysis of the variability of unit activity in the cat's visual cortex. *Exp Brain Res* 37:595– 604. Medline

Schiller PH, Finlay BL, Volman SF (1976) Short-term response variability of monkey striate neurons. *Brain Res* 105:347–349. CrossRef Medline

Shadlen MN, Newsome WT (1998) The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J Neurosci* 18:3870 –3896. Medline

Steinschneider M (2011) Unlocking the role of the superior temporal gyrus for speech sound categorization. *J Neurophysiol* 105:2631–2633. CrossRef Medline

Theil H (1961) *Economic forecasts and policy*. Amsterdam: North-Holland.

Tolhurst DJ, Movshon JA, Dean AF (1983) The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Res* 23:775–785. CrossRef Medline

Vogels R, Spileers W, Orban GA (1989) The response variability of striate cortical neurons in the behaving monkey. *Exp Brain Res* 77:432–436. CrossRef Medline

Whitsel BL, Schreiner RC, Essick GK (1977) An analysis of variability in somatosensory cortical neuron discharge. *J Neurophysiol* 40:589–607. Medline

Whittingstall K, Logothetis NK (2009) Frequency-band coupling in surface EEG reflects spiking activity in monkey visual cortex. *Neuron* 64:281–289. CrossRef Medline

Zohary E, Shadlen MN, Newsome WT (1994) Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* 370:140–143. CrossRef Medline

# Chapter 3. The control of vocal pitch in human laryngeal motor cortex

**Benjamin K. Dichter, Jonathan D. Breshears, Matthew K. Leonard, and Edward F. Chang**

**Abstract:** The flexible control of vocal pitch during speech production is a fundamental aspect of human oral communication. Intonation patterns created by changing vocal pitch are a rich source of information for conveying meaning. Still, it is currently unknown how the brain generates the complex laryngeal motor commands that allow for prosody and song. Here, we used direct high-density cortical recordings from the human brain to determine the encoding mechanisms of vocal pitch control and voicing during natural speech and song production. We found neural activity at electrodes over the right dorsal laryngeal motor cortex (dLMC) that was highly selective to vocal pitch encoding, but not for other features in speech articulation. Using a model of vocal pitch contours, we found that neural activity at a subset of dLMC electrodes was selective for producing pitch accents, but distinct from those that encoded voicing in dLMC and the separate ventral laryngeal motor cortex (vLMC). The same neural populations showed similar encoding of pitch changes in a non-speech singing task, suggesting a general control mechanism. Finally, we confirmed the causal, feed-forward involvement of dLMC in pitch production by using direct cortical stimulation to evoke laryngeal electromyographic responses and vocalizations. Together, these results have significant implications for understanding the neural basis of larynx-based vocal control in spoken language.

## Introduction

Central to the human ability to speak is the control of the larynx, which gives rise to voicing and modulations of vocal pitch (Ohala, 1983). In English, for example, deliberately controlled changes of vocal pitch are used to convey critical elements of prosody, such as syllable stress (Gay, 1978), word emphasis (Ladd and Morton, 1997), phrase segmentation (Jusczyk et al., 1992), modality (e.g. question vs. statement) (Ohala, 1983), and even mood (Scherer, 1989). In speech, the two dominant functions of the larynx are to generate voicing and modulate pitch. Voicing is created by adduction of the vocal folds (posterior cricoarytenoid muscle), bringing them into close proximity, so that they vibrate when air is passed through. In contrast, pitch is modulated primarily by stretching the vocal folds. (cricothyroid muscle). Greater tension in the vocal folds causes them to vibrate at a higher frequency during voicing, and produce a higher pitch sound (Titze et al., 1989; Hull, 2013). It has been speculated that the human ability to flexibly mimic pitch contours is due to the evolutionary changes in neural control of the larynx, rather than the larynx anatomy itself, and has contributed to the rapid development of language in humans (Brown et al., 2008; Hickok, 2016; Pisanski et al., 2016; Belyk and Brown, 2017).

Previous studies have identified two distinct regions in the human sensorimotor cortex that are correlated with laryngeal movements. The ventral laryngeal motor cortex (vLMC) is at the bottom of the sensorimotor cortex homunculus (Foerster, 1936; Penfield and Boldrey, 1937), and is a likely homologue of LMC in other primate species (Hast and Milojkovic, 1966; Simonyan and Jürgens, 2002; Jürgens, 2009). A completely separate dorsal region (dLMC) has been identified between the cortical representation of the lips and the hand (Brown et al., 2008; Simonyan and Horwitz, 2011a; Belyk and Brown, 2015), and is thought to be unique to humans (Mayer et al., 2002; Belyk and Brown, 2017). The existence of two larynx cortical representations is controversial, in part because it is unknown how and whether each region contributes distinct roles in larynx operation.



Here, we sought to address four fundamental questions about the cortical control of vocal pitch, including: 1) its localization in the sensorimotor cortex, 2) whether encoding differs for functionally-distinct pitch components (accents, phrase, and voicing), 3) whether the same pitch control mechanisms are engaged during speech and non-speech vocalizations like singing, and 4) whether electrical stimulation of the dLMC causes direct and proportional activation of laryngeal muscles. To address these questions, we used high-density intracranial recordings and stimulation of the lateral surface of the brain in participants who were undergoing epilepsy surgery. These high-resolution recordings allowed us to identify the functional roles of both LMC regions during naturalistic vocal production tasks.

## **Results**

To understand how speakers control the pitch of their voices, we designed a lexical emphasis task that required participants to stress specific words in a sentence. Eleven participants spoke the sentence, “I never said she stole my money,” and on each trial, they were cued to change the meaning of the sentence by emphasizing a specific word (Rooth, 1985, 1992) (Fig 1b). For instance, “I never said she stole my money” would imply that someone else had stolen the money. The written forms of the sentences were presented to participants on a computer screen with the target word of emphasis underlined and italicized, and an example audio sentence was played through speakers. In an additional condition, subjects were instructed to say the sentence as a question. This task naturally elicits prosodic differences between conditions, while keeping the lexical and syllabic content for each sentence the same.

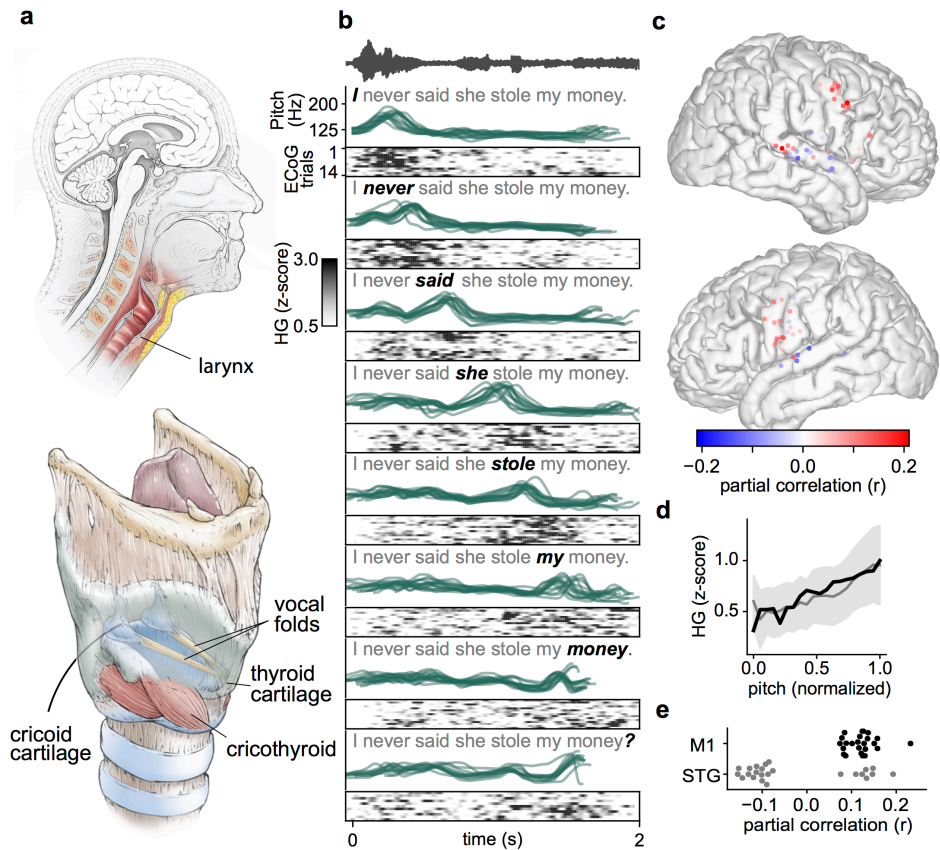
We used an autocorrelation method (Boersma, 1993) to extract the pitch contour (fundamental frequency,  $f_0$ ) from the produced acoustic waveform. On every trial, the pitch contour contained a transient increase in pitch at the time of the emphasized word (Fig 1b, green lines). While participants performed the task, we recorded neural activity from ECoG electrodes on the lateral

cortical surface (Fig S1), and computed the analytic amplitude of the cortical activity in the high-gamma range (HG; 70-150 Hz), which has been found to correlate with multi-unit firing rate (Ray and Maunsell, 2011), and has been shown to reliably track neural activity associated with speech articulation and other movements (Crone et al., 1998; Bouchard et al., 2013).

We found electrodes with increased neural activity that was clearly time-locked to the production of the emphasized word (Fig 1b, single trial raster plots). We next quantified the relationship between vocal pitch and neural activity for every electrode across participants. We also controlled for three potential factors that can correlate with pitch in naturally produced speech. First, we used partial correlation to remove the effect of intensity (amplitude) (Stevens, 1935). Second, to remove the encoding of supralaryngeal articulators (Bouchard et al., 2013), we first used dynamic time warping on the acoustics to temporally align the syllabic sequence across trials, then subtracted the mean activation pattern across trials. Third, to control for natural declination (Ladd, 1984), which causes a correlation between pitch and proximity to the start of the sentence, we used a trial-wise shuffle test, which requires significant electrodes to correlate more strongly with pitch than would be expected from declination alone. (See Fig. S2 and methods for details).

After removing these potential confounds, we found that across participants, the locations of pitch-encoding electrodes were specifically localized to a region of the right precentral gyrus, the dorsal laryngeal motor cortex (dLMC;  $p < 0.001$  using a shuffle test; Fig 1c). Some electrodes were found to correlate with pitch in the left hemisphere, but they were far less confined to a single region. All electrodes in the right dLMC showed a positive monotonic relationship between high-gamma activity and pitch (Fig 1d). We also observed evidence for encoding of the auditory feedback of vocal pitch in electrodes over the bilateral superior temporal gyrus (STG). In contrast to right dLMC, which only had positive correlations, some STG electrodes

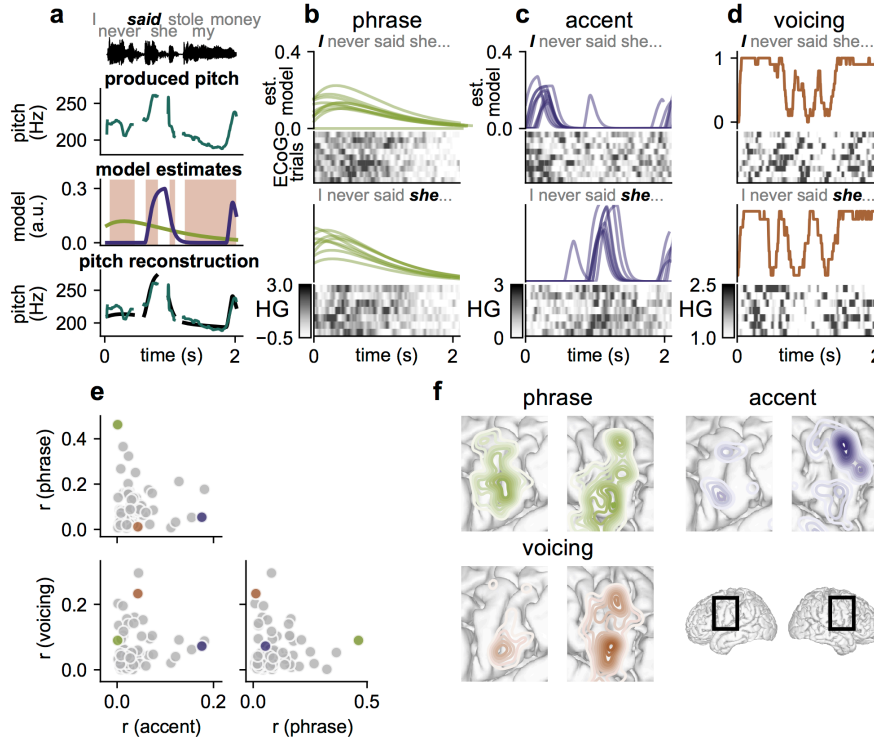
demonstrated positive correlations, while other STG electrodes showed negative correlations (Tang et al., 2017) (Fig 1e).



**Figure 1 | Human cortical encoding of vocal pitch in right dLMC during speech production.** Participants were instructed to emphasize specific words in a sentence. **a**, Diagrams of laryngeal anatomy. The vocal folds are stretched by the cricothyroid muscle, and increased tension in the vocal folds results in a higher produced pitch. **b**, Pitch-correlated neural activity at an example electrode. The speech waveform for one example sentence (emphasis on “I”) is shown at the top. pitch contours (green lines) and high gamma activation for the example electrode (black rasters) for every sentence spoken by a single participant are shown. Trials are grouped by the word of emphasis and co-aligned to the beginning of the emphasized word. On a single trial level, increases in pitch are associated with increased neural activity. **c**, Spatial localization of electrodes that have a significant correlation with vocal pitch, after controlling for intensity and supralaryngeal articulators. Electrodes appear clustered in the right dorsal laryngeal motor cortex (M1) on anterior aspect of the precentral gyrus (dLMC, located lateral to hand and medial to the lip cortical representations), with weak responses in left sensorimotor cortex. The nonprimary auditory cortex on superior temporal gyrus (STG) (feedback responses). **e**, Tuning curves for pitch across electrodes all significant electrodes in M1 (mean and s.d. in grey, example electrode in black) over normalized pitch range. Activation increases monotonically with pitch values (middle 90 percentile range plotted). **d**, Correlation values for significantly tuned electrodes in the M1 and STG regions. Electrodes in M1 were all positively correlated with vocal pitch, whereas activity of electrodes over the STG were both positively and negatively correlated with pitch.

In natural speech, vocal pitch is composed of multiple elements, each with different timescales, and potentially with different encoding mechanisms. To understand the specific sub-processes involved in pitch control, we applied a model-based approach to estimate these distinct components of the pitch contour. We used a well-known mathematical formalization (Fujisaki, 2004), called the Fujisaki model, to explain the neural activity on each electrode in the ventral sensorimotor cortex (vSMC). For each sentence, the components in the model include a fast “accent” component (emphasized words or syllables), and a slow “phrase” component (the declination (Ladd, 1984) in pitch over the course of a phrase). The model is motivated by the physiological mechanisms of pitch control in the larynx, and is capable of parsimoniously modeling pitch contours across many languages (Fujisaki, 2004). We hypothesized that these theoretically distinct components are controlled independently in the brain.

The phrase and accent components, along with whether the segments were voiced or unvoiced, allowed us to reconstruct the produced pitch contours nearly perfectly ( $R^2=0.96$ , Fig 2a). At individual electrodes, high-gamma activity was correlated with these pitch components in a temporally-specific fashion (Fig 2b-d). Crucially, we found a clear and striking dissociation between electrodes that encoded accent, phrase, and voicing (Fig 2e). Although there were electrodes that were significantly tuned to both voicing and pitch, 59% of pitch-tuned electrodes were not tuned to voicing, and 81% of voicing electrodes were not tuned to pitch, suggesting that these components have separable control representations.



**Figure 2 | Cortical representation of pitch contour components in speech: accent, phrase, and voicing.**

**a**, The Fujisaki model decomposes the pitch contour in natural speech into accent, phrase, and voicing components. Inference of the Fujisaki is shown on an example sentence. In order from top to bottom: acoustic waveform of produced sentence; pitch contour extracted from sentence; phrase (green), accent (purple), and voicing (brown) components extracted from the pitch contour; original pitch contour (green) and Fujisaki reconstruction of pitch contour (black). **b**, Single trial high gamma raster for an electrode controlling the phrasal component of the pitch contour. Green curves show the phrase component of the Fujisaki model for each trial, and the grey rasters show the activation of an example “phrase” electrode ( $r=0.45$ ). This electrode responded similarly to sentences with different accents (top and bottom). **c**, Single trial high gamma raster for an electrode controlling pitch accents. Purple lines show the accent component for an example subject separated by sentence style, and the grey raster shows the activation of an example “accent” electrode ( $r=0.17$ ). **d**, Single trial high gamma raster for an electrode controlling voicing. Brown lines show the proportion of sentences that are voiced for each style. This electrode has higher activation when the subject is voicing ( $r=0.2$ ). **e**, The correlation coefficient between activation of the accent and phrase components of the Fujisaki model for each of the electrodes over the sensorimotor cortex. Example electrodes in b-d are marked in their respective colors. Electrodes tend to be predominantly along the axes-- no electrodes in the vSMC correlate with both phrase and accent. **f**, Spatial location of electrodes on the vSMC across all subjects. Accent and voicing electrodes were selected using a trial-wise shuffle test ( $p<.001$ ). Phrase electrodes were selected using a trial-wise shuffle test and a cutoff of  $r<0.15$ . Each brain shows the kernel density estimation illustrating the spatial organization of electrodes on a common brain. Accent electrodes were strongly localized to the dLMC, while voicing and phrase electrodes were in both the dLMC and the vLMC.

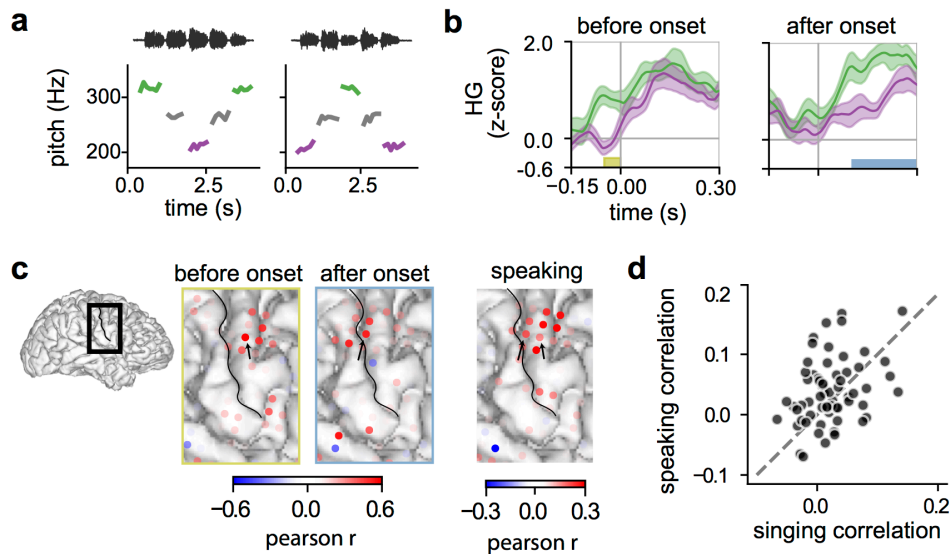
Given the results shown in Figure 1, we hypothesized that pitch encoding in right dLMC more strongly reflects the pitch accent component, consistent with the emphasized word in each sentence. We confirmed that right dLMC electrodes were most strongly associated with pitch

accent (Fig 2f). In contrast, phrase-encoding electrodes were found in bilateral vLMC. Finally, voicing was localized to a distinct subset of right dLMC and bilateral vLMC (Bouchard et al., 2013) electrodes. Together, these results demonstrate a functional-anatomical distinction in the control of pitch phrase, pitch accent, and voicing.

We next asked whether the encoding of vocal pitch was specific to the linguistic context (Mayer et al., 2002), or similar during speaking and singing, a form of non-speech vocal production. In addition to being interesting in its own right, singing provides a method for observing pure vocal pitch control without contamination of other speech features. Participants performed a singing task in which they listened to and then repeated pitch patterns alternating between sol-mi-do-mi-sol (high-middle-low-middle-high) and do-mi-sol-mi-do (low-middle-high-middle-low) on a vowel. Figure 3a shows examples of the two melodies sung by one of the participants (Fig. S4 shows the performance of all subjects). To remove any effects of the sequential order of the produced pitches, the two melodies were interleaved so that the high and low notes occur in the same sequential order, both occurring third in the sequence 50% of the time and fifth in the sequence the other 50% of the time. The first note in each melody was excluded from analysis, so that all analyzed notes were preceded by the same (middle) note. Importantly, this task was specifically designed to avoid some of the potential confounds in interpreting vocal pitch in natural speech described earlier. That is, this singing task did not have pitch declination, correlations between pitch and intensity, correlation between pitch and articulatory gestures, and high autocorrelation of pitch over time that each are common to natural speech.

We examined neural activity time-locked to the onset of each note (Fig 3b). We found electrodes in the anterior part of right dLMC that exhibited pitch-specific activity immediately preceding the acoustic onset of the vocalization (yellow region; Fig 3c). In order to sing the correct pitch at the acoustic onset of each note, a singer must tense the laryngeal muscles, creating the necessary tension in the vocal folds (Fig. 1a). In this brief moment before the

acoustic onset, we observe neural control of the larynx without subjects hearing their own voice. Approximately 100-300ms after acoustic onset, electrodes in the posterior part of the right dLMC were correlated with pitch (blue region; Fig 3c). Both subgroups of electrodes, those that are tuned to pitch before acoustic onset, and those that are tuned to pitch during vocalization, are also correlated with pitch during the speaking task (Fig 3c). Pitch representation is weak in the vLMC both before and during vocalization.



**Figure 3 | Pitch encoding during singing.** **a**, Singing task with two simple melodies. Notes are colored by low, middle, and high target tone. The sound waveforms are shown above, with produced pitch for each note below. **b**, High gamma response for two example electrodes in right dLMC of the example subject for high (green) and low (purple) notes. Time 0 is the acoustic onset of the note. The yellow and blue segments mark time windows used to compute correlations in (c). Error bars are sem across trials. **c**, The Pearson correlation between cortical activation and vocal pitch for low and high notes using 50 ms before acoustic onset (left) and 100 – 300 ms after acoustic onset (middle). right: The Pearson correlation computed between pitch and high gamma activation for the contrastive emphasis task for this subject. Arrows mark the electrodes from (b). **d**, Comparison between pitch encoding in right dLMC electrodes during singing and during speaking for all subjects. There is a strong correlation across electrodes in the two behavioral conditions (Pearson  $r=0.33$ ,  $p$ -value  $< 0.01$ ).

To quantify the similarity of pitch representation during singing and speaking, we compared the continuous correlation between electrode activity and pitch in the two conditions (Fig 3d).

Across all dLMC electrodes from all right hemisphere subjects, there was a strong correlation between how well electrodes encoded pitch in the singing and speaking tasks (Pearson  $r =$

0.33,  $p$ -value=0.01) (Fig 3d). This demonstrates that dLMC activity reflects a task-independent representation of vocal pitch that is not specific to speech or singing.

We have demonstrated that neural activity in dLMC reflects the detailed and temporally-specific features of produced pitch during speaking and singing. To demonstrate definitively that this activity reflects feed-forward control of laryngeal muscles, we used direct focal (bipolar) electrical stimulation during intraoperative clinical brain mapping. In two separate experiments, we examined whether there is a causal link between dLMC activity and laryngeal muscle activation. This approach helps rule out representation that is purely somatosensory feedback (Guenther, 2006), an efference copy signal (Niziolek et al., 2013), or an auditory response to the acoustics of one's own voice (Wilson et al., 2004; Brown et al., 2008; Chang et al., 2013; Cheung et al., 2016).

In the first stimulation experiment, participants undergoing neurosurgical procedures with general anesthesia were intubated with a specialized endotracheal tube with electromyographic (EMG) non-penetrating wire electrodes (Eisele, 1996; Rea and Khan, 1998). These electrodes contacted the left and right vocal folds, and were designed to record laryngeal muscle activations. In 18 participants (5 left), we stimulated cortical sites throughout the sensorimotor cortex (Tate et al., 2013) while recording laryngeal EMG. We found sites that elicited a laryngeal EMG response bilaterally in the dLMC and vLMC. The highest concentration was in the right dLMC, the same cortical region that correlated with vocal pitch during speech and singing (Fig 4a). The dLMC was typically found between areas where stimulation evoked EMG-detected movements from the arm (dorsal), and mouth (ventral) (Fig 4d).

To understand whether there is a causal relationship between the amount of cortical activity and the amount of laryngeal muscle activation, we varied the cortical stimulation amplitude, and found that it caused a proportional increase in the laryngeal EMG response (Fig 4b) with a latency of 11-19 ms (Fig 4g). This demonstrates a monotonic relationship between dLMC neural

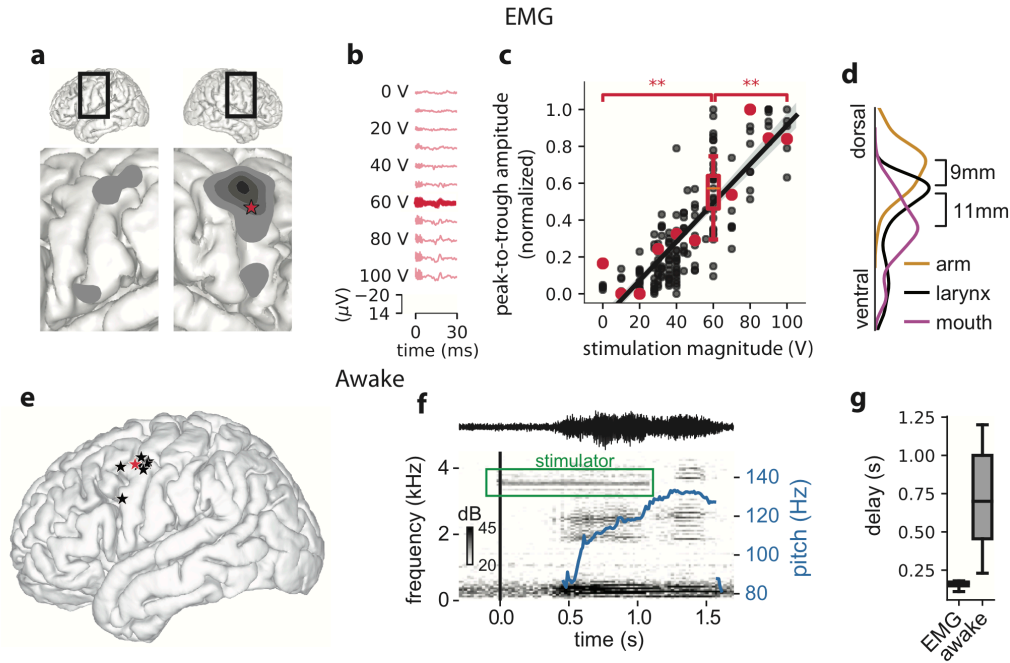


activity and the magnitude of laryngeal muscle activation (Fig 4c). One example subject (red) received 11 cortical stimulations at mid range (60 V), which elicited a distribution of laryngeal responses in between the lowest and highest stimulation magnitude. These findings of proportional responses to graded stimulation are concordant with the monotonic relationship between cortical high gamma activity and vocal pitch, which is determined by tension of the cricothyroid muscle. Furthermore, the fast response is consistent with the timing and representation in the singing task, and suggests involvement of the direct cortical innervation of intrinsic laryngeal motor neurons (Simonyan and Horwitz, 2011b).

In the second stimulation experiment, we asked whether stimulating the specific cortical region of dLMC causes an evoked, involuntary vocalization. In this experiment, stimulation was applied throughout the sensorimotor cortex in 82 neurosurgical patients undergoing awake surgical procedures in which the left hemisphere cortical surface was exposed. While we could not assess the right hemisphere, we were still interested in understanding what effects could be ascribed to dLMC stimulation given that we did find evidence of voicing encoding bilaterally. In 20 participants, we observed that stimulation of dLMC evoked audible vocalizations (Breshears et al., 2015).

We found that the evoked vocalizations were all voiced as demonstrated by energy at the fundamental frequency and voice-related harmonics (Fig 4f). These non-volitional, stimulation-evoked vocalizations were not meaningful or communicative speech sounds, but sounded typically like a prolonged “aaah” that varied in vocal register, including vocal fry (9 subjects, example: Fig. S5a), modal register (10 subjects example: Fig 4f), and falsetto register (1 subject, example: Fig S5b), and lasted 0.5 – 2.9 seconds (mean: 1.1 seconds).

In early descriptions of evoked vocalizations by Penfield (Penfield and Roberts, 1959), similar responses were interpreted at positions spread throughout the ventral sensorimotor cortex. In



**Figure 4 | Electrical stimulation of dLMC.** **a**, Cortical stimulation mapping of larynx responses in the primary sensory and motor cortices for 18 participants. The larynx was monitored using electromyography (EMG) electrodes on a customized endotracheal tube. Other evoked movements not shown. The red star marks the example site that is shown in more detail in (b) and (c). **b**, Laryngeal response for stimulations ranging from 0-100 V. Stimulation was delivered 11 times at 60 V and once at each other magnitude for this patient. **c**, Three other patients also received graded stimulation. Peak-to-trough response amplitude was determined for each stimulation, and is shown for each patient, normalized to the maximum and minimum response for each larynx side of each subject. Laryngeal responses for the example stimulation site of (a) and (b) are shown in red. Stimulation responses to 60V are greater than 0V (p-value < 1e-6, one-sided t-test) and less than 100V (p-value < 1e-3, one-sided t-test). Therefore responses are not an all or none, but rather a graded response where more stimulation yields a greater laryngeal response. Stimulation magnitude is strongly correlated with laryngeal response across subjects (Pearson  $r = 0.85$ , p-value < 1e-52). **d**, Sites that evoked arm movement were dorsal of the larynx sites and sites that evoked mouth movement were ventral of the larynx sites. **e**, Sites that evoked a spontaneous involuntary voiced vocalization during awake stimulation mapping. The vocalization evoked by the red location is shown in (e). **f**, Spectrogram and pitch contour of an example evoked vocalization. Noise from the stimulator created a 3.5 kHz band in the spectrogram. **g**, Delay times between the start of stimulation and the beginning of the response for anesthetized (black) and awake (grey) stimulation. All of the response times for laryngeal response were shorter than times for vocalization response.

fact, a distinct and separate dorsal representation of the larynx was never depicted in the homunculus. Using the precision of an intraoperative stereotactic navigation system and EMG monitoring, however, we found that these responses were well-localized to the dLMC (Fig 4e). Concordant with the EMG results, we found that stimulation at other ventral sensorimotor cortex locations instead evoked contralateral pulling of the face, deviation of tongue, and jaw

movements, or arm movements more dorsally (Breshears et al., 2015). Furthermore, consistent with results from other primates (Jurgens, 1974), stimulating the vLMC did not elicit vocalization.

These results provide definitive evidence that dLMC neural activity reflects the feed-forward encoding of motor commands in the larynx, though they also suggest that the representation might be more complex than control of a single muscle. The vocalization response requires adduction of the vocal folds, and involves precise coordination with respiratory processes in the lungs and diaphragm.

## **Discussion**

In summary, we combined high-resolution cortical physiology and stimulation methods with natural speech and singing to demonstrate that neural signals in human dLMC encode motor commands that allow for the flexible, feed-forward control of vocal pitch. By modeling distinct aspects of the pitch production behavior, we demonstrated a functional-anatomical dissociation of two important dimensions of movement control in the larynx: voicing and pitch. It is known that voicing and pitch activate different laryngeal muscles: voicing is mediated primarily by adduction of the vocal folds, and pitch is mediated by the lengthening and tensing of the vocal folds. However, it was previously unknown whether and how cortical control signals differentiated these two important functions (Belyk and Brown, 2017). Consistent with our previous work (Bouchard et al., 2013), voicing was encoded by both the dLMC and vLMC. Here, we found that a subset of dLMC electrodes was selective for vocal pitch control, and not for other articulatory features, in the context of speaking and singing, demonstrating a distinct circuit for pitch. Furthermore, direct and temporally-precise control of pitch involves independent processes that unfold over short (accent) and long (phrase) timescales in distinct cortical regions.

Humans are unique among primates in our ability to flexibly control vocal pitch, and there is growing evidence that precise control of pitch was one of the first evolutionary developments that ultimately led to human-specific speech abilities (Brown et al., 2008; Hickok, 2016; Pisanski et al., 2016; Belyk and Brown, 2017). The specialization we observed for pitch control in dLMC adds important empirical evidence for the role that vocal pitch plays in species-specific behaviors like speech. These findings support a crucial role for the dLMC in the evolution and development of language specialization in humans.

## **Methods**

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

### *Subjects*

All 11 subjects were native English speaking patients who underwent chronic implantation of a high-density subdural electrocorticography (ECoG) array as part of their surgical treatment of epilepsy. Five of the subjects had ECoG grids on the left hemisphere, and six had ECoG grids on the right hemisphere. Electrode coverage for all patients is shown in Fig. S1. Subjects gave their written informed consent. Each subject reported normal speaking and hearing ability.

### *Neural recordings*

Each subject was unilaterally implanted with a 256-channel lattice array of electrodes, each with an exposed diameter of 1.17 mm and center-to-center spacing of 4 mm. Cortical local field potentials were amplified and quantized using a pre-amplifier (PZ5, Tucker-Davis Technologies), and preprocessed using a digital signal processor (RZ2, Tucker-Davis Technologies).

### *Preprocessing*

The voltage trace of each electrode was visually inspected for artifact and excessive noise, and noisy electrodes were excluded from further analysis. For the remaining electrodes, we used a common average reference across electrode blocks and notch filters at 60, 120 and 180 Hz to remove line noise. For each electrode, we extracted the time-varying high gamma (HG) analytic amplitude using eight Gaussian band pass filters at 73.0, 79.5, 87.8, 96.9, 107.0, 118.1, 130.4, and 144.0 Hz followed by a Hilbert transform (Moses et al., 2016). HG was calculated as the mean of these bands, and z-score was computed relative to the entire experimental block.

### *Acoustic analysis*

We extracted the pitch contour of each sentence using Praat (Boersma, 1993) with pitch min and max bound determined individually for each subject. Voicing was also determined at this point. We then used an 80 ms median filter, then corrected erroneous octave jumps, interpolated through unvoiced regions in log(Hz), and filtered with an 80 ms hanning window. Intensity was also extracted from each trial using Praat, and normalized by recording session. Throughout the text, “pitch” is refers to fundamental frequency.

We found that some articulatory features tended to be correlated with pitch. For instance, nasals tended to have low pitch in all prosodic styles, so electrodes that were strongly tuned to velar movements would appear to be negatively correlated with pitch. To address this potential confound, an acoustic model was used. Although the pitch contours varied, the same syllable sequence was spoken each time, allowing us to examine specifically the control of pitch during natural speech and control for the articulatory movement of the production of the syllables.

First, we used dynamic time warping to temporally align the sentences for each subject. Mel cepstral coefficients (MFCCs) were calculated for each sentence (McFee et al., 2017) and dynamic time warping was used to find the shortest path for the Euclidean distance of the MFCCs. This warp was then applied to the pitch, intensity, and high gamma analytic amplitude

contour of each sentence. By removing the average neural activation across trials in this new timing, we removed the contribution of neural representation of articulatory movements that were consistent across trials.

This acoustic model does not track the articulators directly (Bouchard et al., 2016) or explicitly model the movement of specific articulators from the acoustics (Bouchard et al., 2013), but implicitly models the supra-laryngeal articulators by their effect on the acoustics of the sentence. Our approach has the advantage of being free of modeling assumptions about the relationship between neural activation and articulator movement (e.g. linearity). However, it does not capture trial-to-trial differences in articulation beyond timing differences. For instance, if a subject dropped the “r” of “never” for one trial, an explicit model might capture this but our approach would not. We expect these differences to be relatively rare and small for our task, where the syllabic context is the same across repetitions.

To determine the functional relationship between pitch and neural activation pitch was digitized into 20 bins uniformly spanning the middle 90-percentile range of pitch values for each subject, and the average high-gamma was calculated for each bin and significant electrode.

### *Fujisaki Parameter Estimation*

The Fujisaki model of vocal pitch is a model that separates the pitch contour of an utterance ( $F_0$ ) into three components, the phrase, the accent, and the baseline. The phrase ( $P$ ) is composed of  $I$  individual phrase gestures of amplitude  $A_p$  and shape  $G_p$ , The accent ( $A_c$ ) is composed of  $J$  individual accent gestures of amplitude  $A_a$ . The model is defined by the following equations (Fujisaki, 2004):

$$\ln F_0(t) = \ln F_b + P + A_c$$

$$P = \sum_{i=1}^I A_{pi} G_p(t - T_{0i})$$

$$A_c = \sum_{j=1}^J A_{aj} \{G_a(t - T_{1j}) - G_a(t - T_{2j})\}$$

$$G_p(t) = \begin{cases} \alpha^2 t e^{-\alpha t}, & t \geq 0, \\ 0, & t < 0. \end{cases}$$

$$G_a(t) = \begin{cases} \min[1 - (1 + \beta t)e^{-\beta t}, \gamma], & t \geq 0, \\ 0, & t < 0. \end{cases}$$

The phrase and accent components were estimated for each spoken sentence using FujiParaEditor (Mixdorff, 2000). We used automated inference (Mixdorff, 2009), with manual corrections where necessary.

Correlation of neural activity (z-scored high gamma analytic amplitude) was calculated against  $P$  and  $A_c$  and against the binary voicing metric ( $V$ ) extracted from Praat. For each metric and each of the 2816 electrodes, we conducted a shuffle test similar to the significance test for pitch (nboots = 1000,  $p < 0.001$ ). Since the activation of many electrodes were correlated weakly with phrase due to sentence timing, electrodes were also required to have correlation  $> 0.1$  to be labeled significantly tuned with  $P$ .

### *Singing*

The singing performance was measured quantitatively for each subject. First, the value of each note was determined by the median pitch produced for the duration of the note. To enable comparison between subjects with different vocal ranges, each note converted to a semitone value:

$$s = 12 \log_2(f)$$

where  $f$  is pitch and  $s$  is semitone. Using the semitone values, the performance of each singer was measured by the average interval between “do” and “so” (target = 7.0) and the standard deviation for low and high notes. A single subject was best in both of these metrics (black,

Figure S2), and is used as the example subject in Figure 3. This subject also had approximately the same loudness distribution for low and high notes.

A cross-subject analysis was used on the remaining subjects. Several of these subjects were not able to successfully mimic the melody of the task, but were still able to sing notes that varied in pitch. The median pitch through the duration of each note was used as the note's pitch, and we calculated a timepoint-by-timepoint correlation between high-gamma activation and pitch for each electrode in the right dLMC.

### *Stimulation Mapping*

Intraoperative direct electrical stimulation mapping of the peri-rolandic cortices was performed in 18 subjects (5 left) as a part of their clinical care prior to surgical resection (4 of these subjects also participated in the contrastive emphasis and singing task experiments). After the induction of anesthesia, electromyography needles were placed in the orbicularis oris, tongue, and hand by a certified neuromonitoring specialist. A NIM® endotracheal tube (Medtronic, Minneapolis, MN) was placed under direct visualization with wire electrodes in contact with the vocal folds bilaterally to record laryngeal EMG activity (Eisele, 1996; Rea and Khan, 1998). The time-locked EMG activity and stimulation parameters were recorded on a Cascade® intraoperative neuromonitoring system (Cadwell, Kennewick, WA).

A craniotomy was performed, the dura was opened, and the exposed fronto-temporo-parietal cortical surface was densely mapped. The mapping was performed using a bipolar Ojemann Cortical Stimulator® probe (Integra, Plainsboro, NJ) with 5mm electrode spacing. The stimulator probe was applied sequentially to one cortical site at a time, as the voltage was increased from 0V to 100V, in increments of 5-10V, or until an EMG response was observed at that site. A train of 5-9 biphasic square waves, each with equal positive and negative phases of 75 $\mu$ s duration was used (Tate et al., 2013). For each trial of stimulation, the voltage was held constant, while the current was allowed to vary. EMG activity was simultaneously recorded from orbicularis oris,



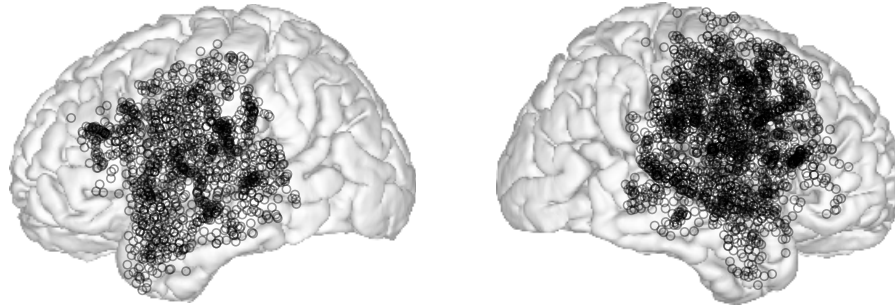
tongue, hand, and larynx as voltage was increased on each trail at each cortical site. Sites of cortical stimulation were spaced approximately 3-5mm apart. If an evoked potential was observed from any of the EMG electrodes, a voltage threshold was identified and the corresponding cortical site was photographed and recorded on the subject's co-registered MRI surface reconstruction using the BrainLab® neuronavigation system. The cortical sites from each subject were then warped into a common space for visualization (see previous description of electrode warping, Hamilton et al, *in submission*). Relative localization of the arm and mouth were determined by normalizing the location of the sites of each subject to the dorsal-most laryngeal site.

In 4 right hemisphere subjects, multiple additional trials of stimulation across a range of voltages was performed at the dLMC site evoking laryngeal EMG activity, in order to characterize the relationship between dLMC stimulation voltage and the magnitude of laryngeal muscle activation. The cortical site was stimulated at voltages ranging from 10-15V below threshold, up to 100V or the plateau of the laryngeal EMG response. All stimulations were performed 5-10 seconds apart to avoid adaptation. EMG voltage responses were filtered with a 8<sup>th</sup> order Butterworth filter with critical frequency 32 Hz. The normalized peak-to-trough amplitude of the motor evoked potentials recorded from the laryngeal EMG was plotted as a function of stimulation voltage. Normalization was relative to the range of peak-to-trough response for each vocal fold of each subject. For the example subject, two one-tailed t-tests were conducted testing difference from the higher and lower extreme values (n=11, p<.01).

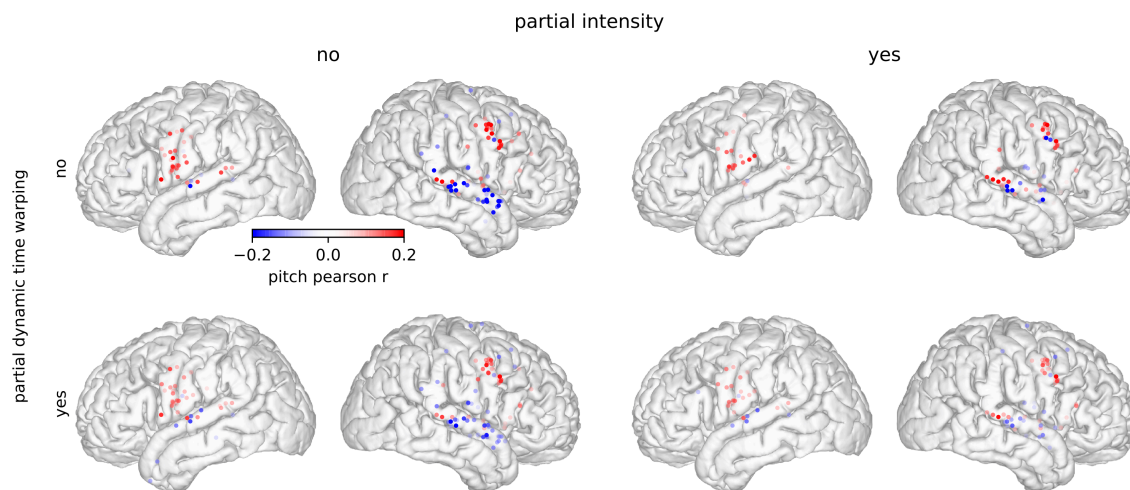
In an independent cohort of patients undergoing craniotomy for surgical resection in the left, dominant hemisphere, stimulation mapping was performed with the patients fully conscious and conversant in order to identify speech areas (see previously published awake mapping protocol) (Breshears et al., 2015; Chang et al., 2016). After exposure of the peri-rolandic cortex and emergence from intravenous sedation (either dexmedetomidine or propofol), intravenous

fentanyl was titrated for optimal balance of pain control with patient arousal during the mapping procedure. The exposed cortex was densely mapped using an Ojemann stimulator (current range: 1 to 3.5 milliamps, pulse frequency 60Hz, pulse width 1ms, stimulus duration: 500 to 1500ms, stimulator electrode spacing: 5 mm). Prior to mapping, the after-discharge threshold was determined; the mapping was conducted at the maximum current that did not result in cortical spread (i.e. after-discharges). This ensured a low false negative rate. Each response or non-response to stimulation was tested for consistency/repeatability with at least 3 non-consecutive stimulations. Responses were considered valid only in the absence of after-discharges or seizure activity on electrocorticography, which was monitored and reported in real-time by an epileptologist. The mapping procedure was recorded simultaneously with 2 video cameras, one with an unobstructed view of the patient's face, and the second with an unobstructed view of the cortical surface. Cortical sites evoking involuntary vocalization responses were documented with a photograph and transferred onto the patient's cortical surface reconstruction from their MR imaging. These were warped into a common space, as described above. Acoustic waveforms of the vocalizations were extracted from the audio files for spectral analysis using librosa (McFee et al., 2017).

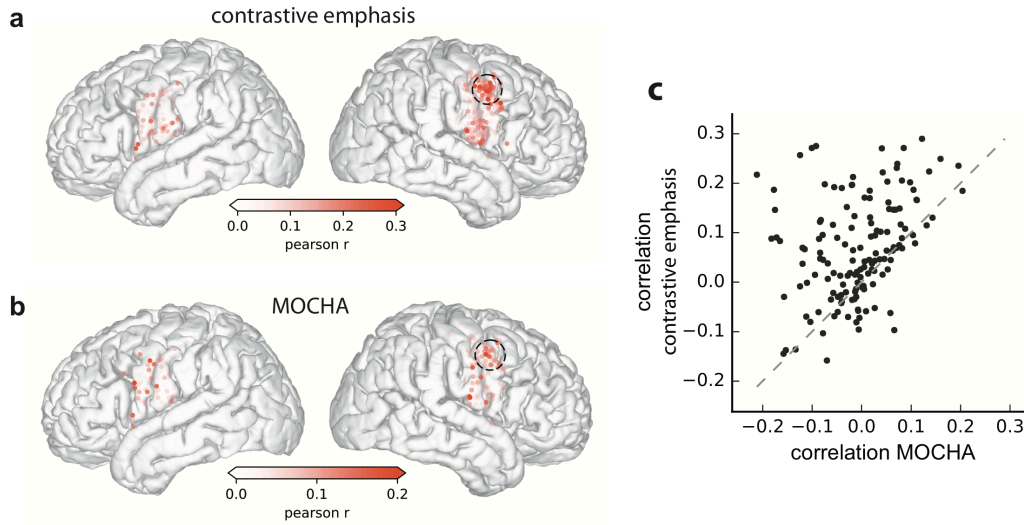
## Supplemental Figures



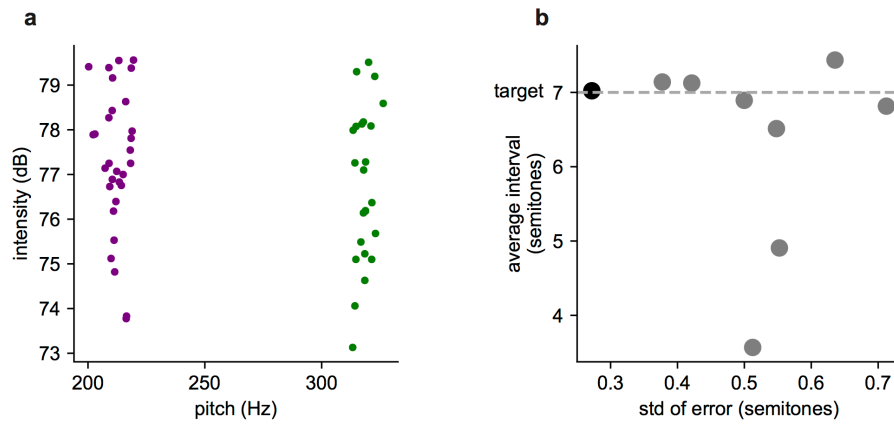
**Figure S1 | Electrode coverage.** Position of electrodes from chronically implanted high-density grids from thirteen subjects. The procedure described in (*Hamilton et al. In Submission*) was used to nonlinearly warp the electrode positions onto an MNI atlas brain (cvs\_avg35\_inMNI152).



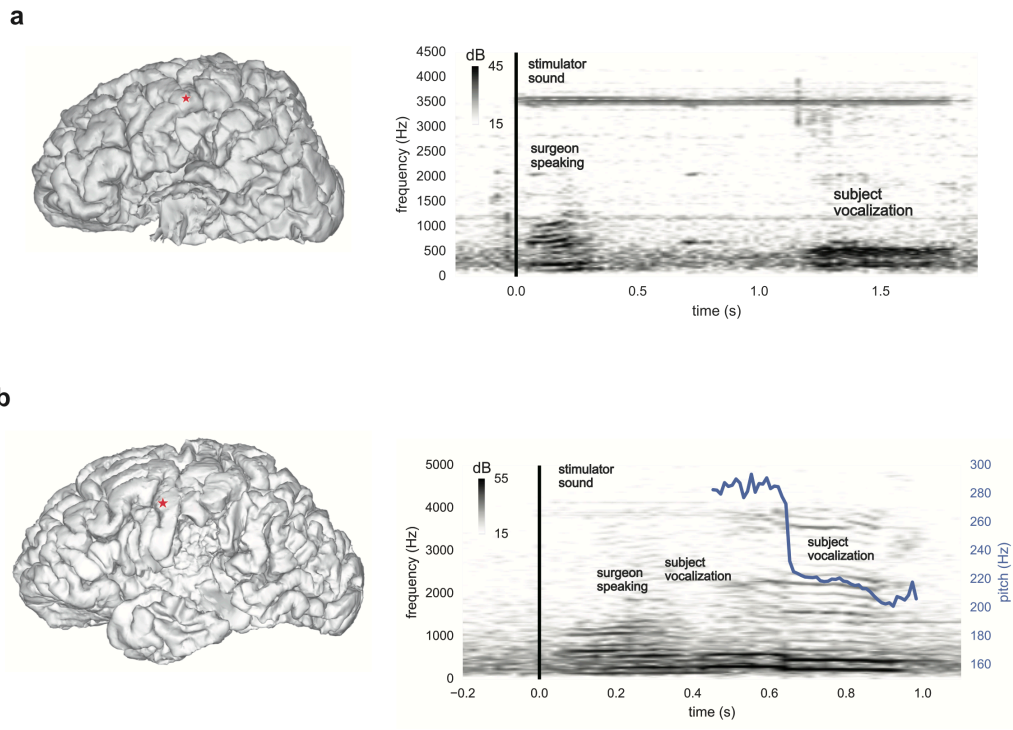
**Figure S2 | Pitch partial correlation analysis.** The left column shows pitch correlation without including partial correlation with intensity in the model, and the right side shows the results after including intensity. The top row shows results that do not include the dynamic time warping, and the bottom shows results that include the dynamic time warping.



**Figure S3 | pitch tuning in MOCHA sentence production.** To test whether pitch tuning in right dLMC generalized to natural speech in general, including speech with natural uninstructed intonation, we conducted an additional experiment on a subset of 10 of the subjects who performed the contrastive emphasis task. In this experiment, subjects read sentences from the MOCHA list out loud as they were presented on a computer screen. MOCHA is a list of semantically meaningful sentences designed to sample the articulatory space of English (Wrench, 1999). These sentences were not designed for pitch production specifically, but did elicit natural variability in pitch during production. This task tests the generalizability of the relationship between dLMC activity and vocal pitch, however each sentence is only spoken once, so we are unable to apply a “pseudo-articulatory model.” Instead, we use a linear model for each electrode in each task to predict the high gamma activation of that electrode from the produced vocal pitch. **a**, Encoding results for electrodes in the vSMC across the 10 subjects. Tuning for pitch was again observed in the right dLMC. **b**, The same analysis performed on the same subset of subjects for the contrastive emphasis task. **c**, Comparison of model fit MOCHA vs. contrastive emphasis for each electrode. There is a positive correlation between the models (Pearson  $r = 0.33$ ;  $p$ -value  $< 1e-4$ ), and the models trained and tested on contrastive emphasis fit better than the models trained and tested on MOCHA.



**Figure S4 | Singing performance.** **a**, Intensity and pitch distribution for the example subject in Figure 3 for low (purple) and high (green) notes. **b**, For each of the nine singers, the performance of the singer is measured by the average interval between the high and low notes and the standard deviation of each note. The black point indicates the best singer by both metrics. This is the singer that is used as the example in Figure 3.



**Figure S5 | Stimulation-evoked vocalizations.** Cortical location and spectrograms of vocalizations and pitch contours are shown for select vocalizations to illustrate the range of vocalization types induced by stimulations to the dLMC. **a**, Example of a vocalization that is voiced but does not have sonorous pitch because it is in the vocal fry register. **b**, Example vocalization that shifted from the falsetto register to the modal register.

## References

- Belyk M, Brown S (2015) Pitch underlies activation of the vocal system during affective vocalization. *Soc Cogn Affect Neurosci*:nsv074- Available at:  
<http://scan.oxfordjournals.org/content/early/2015/06/26/scan.nsv074.full>.
- Belyk M, Brown S (2017) The origins of the vocal brain in humans. *Neurosci Biobehav Rev* 77:177–193 Available at: <http://dx.doi.org/10.1016/j.neubiorev.2017.03.014>.
- Boersma P (1993) Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-To-Noise Ratio of a Sampled Sound. *Proc Inst Phonetic Sci* 17:97–110 Available at: <http://www.cs.northwestern.edu/~pardo/courses/eecs352/papers/boersma-pitchtracking.pdf>.
- Bouchard KE, Conant DF, Anumanchipalli GK, Dichter B, Chaisanguanthum KS, Johnson K, Chang EF (2016) High-resolution, non-invasive imaging of upper vocal tract articulators compatible with human brain recordings. *PLoS One* 11:1–30.
- Bouchard KE, Mesgarani N, Johnson K, Chang EF (2013) Functional organization of human sensorimotor cortex for speech articulation. *Nature* 495:327–332 Available at:  
<http://www.ncbi.nlm.nih.gov/pubmed/23426266> [Accessed February 27, 2013].
- Breshears JD, Molinaro AM, Chang EF (2015) A probabilistic map of the human ventral sensorimotor cortex using electrical stimulation. *J Neurosurg* 123:340–349.
- Brown S, Ngan E, Liotti M (2008) A larynx area in the human motor cortex. *Cereb Cortex* 18:837–845 Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17652461>.
- Chang EF, Breshears JD, Raygor KP, Lau D, Molinaro AM, Berger MS (2016) Stereotactic probability and variability of speech arrest and anomia sites during stimulation mapping of the language dominant hemisphere. *J Neurosurg* 126:1–4.
- Chang EF, Niziolek C a, Knight RT, Nagarajan SS, Houde JF (2013) Human cortical sensorimotor network underlying feedback control of vocal pitch. *Proc Natl Acad Sci U S A*

110:2653–2658 Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23345447>.

Cheung C, Hamilton LS, Johnson K, Chang EF (2016) The auditory representation of speech sounds in human motor cortex. *Elife* 5:1–19 Available at:

<http://elifesciences.org/lookup/doi/10.7554/eLife.12577>.

Crone NE, Miglioretti DL, Gordon B, Lesser RP (1998) Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. *Brain* 121 ( Pt 1:2301–2315.

Eisele DW (1996) Intraoperative electrophysiologic monitoring of the recurrent laryngeal nerve. *Laryngoscope* 106:443–449.

Foerster O (1936) Motorische Felder und Bahnen, *Handbuch Neurologie*, Bumke-Foerster Bd.

Fujisaki H (2004) *Speech Prosody 2004 Information, Prosody, and Modeling*.

Gay T (1978) Physiological and acoustic correlates of perceived stress. *Lang Speech* 21:347–353.

Guenther FH (2006) Cortical interactions underlying the production of speech sounds. *J Commun Disord* 39:350–365.

Hast MH, Milojkovic R (1966) The response of the vocal folds to electrical stimulation of the inferior frontal cortex of the squirrel monkey. *Acta Otolaryngol* 61:196–204 Available at: <http://www.tandfonline.com/doi/full/10.3109/00016486609127056>.

Hickok G (2016) A cortical circuit for voluntary laryngeal control : Implications for the evolution language. *Psychon Bull Rev* Available at: <http://dx.doi.org/10.3758/s13423-016-1100-z>.

Hull DM (2013) Thyroarytenoid and cricothyroid muscular activity in vocal register control.

Jurgens U (1974) On the elicibility of vocalization from the cortical larynx area. *Brain Res* 81:564–566.

Jürgens U (2009) The Neural Control of Vocalization in Mammals: A Review. *J Voice* 23:1–10.

Jusczyk PW, Hirsh-Pasek K, Kemler Nelson DG, Kennedy LJ, Woodward A, Piwoz J (1992) Perception of acoustic correlates of major phrasal units by young infants. *Cogn Psychol*



24:252–293.

Ladd DR (1984) Declination.: a review and some hypotheses. *Phonology* 1:53–74 Available at:  
[http://www.journals.cambridge.org/abstract\\_S0952675700000294](http://www.journals.cambridge.org/abstract_S0952675700000294).

Ladd DR, Morton R (1997) The perception of intonational emphasis: continuous or categorical?  
*J Phon* 25:313–342.

Mayer J, Wildgruber D, Riecker A, Dogil G, Ackermann H, Grodd W (2002) Prosody production  
and perception: converging evidence from fMRI studies. *Proc Int Conf Speech  
Prosody*:487–490.

McFee B et al. (2017) librosa 0.5.0. Available at: <https://doi.org/10.5281/zenodo.293021>.

Mixdorff H (2000) A Novel Approach to the Fully Automatic Extraction of Fujisaki Model  
Parameters. *Proc Int Conf Acoust Speech Signal Process* 1:1281--1284 Available at:  
[http://www.isca-speech.org/archive/eurospeech\\_2003/e03\\_0873.html](http://www.isca-speech.org/archive/eurospeech_2003/e03_0873.html).

Mixdorff H (2009) FujiParaEditor. Available at: [http://public.beuth-  
hochschule.de/~mixdorff/thesis/fujisaki.html](http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html).

Moses DA, Mesgarani N, Leonard MK, Chang EF (2016) Neural speech recognition: continuous  
phoneme decoding using spatiotemporal representations of human cortical activity. *J  
Neural Eng* 13:56004 Available at: [http://stacks.iop.org/1741-  
2552/13/i=5/a=056004?key=crossref.ac3fcc4576d40ed0daf5b1e6181cd472](http://stacks.iop.org/1741-2552/13/i=5/a=056004?key=crossref.ac3fcc4576d40ed0daf5b1e6181cd472).

Niziolek CA, Nagarajan SS, Houde JF (2013) What Does Motor Efference Copy Represent?  
Evidence from Speech Production. *J Neurosci* 33:16110–16116 Available at:  
<http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.2137-13.2013>.

Ohala JJ (1983) Cross-Language Use of Pitch: An Ethological View. *Phonetica* 40:1–18.

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

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

Pisanski K, Cartei V, McGettigan C, Raine J, Reby D (2016) *Voice Modulation: A Window into*

- the Origins of Human Vocal Control? *Trends Cogn Sci* 20:304–318 Available at:  
<http://dx.doi.org/10.1016/j.tics.2016.01.002>.
- Ray S, Maunsell JHR (2011) Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biol* 9.
- Rea JL, Khan A (1998) Clinical evoked electromyography for recurrent laryngeal nerve preservation: use of an endotracheal tube electrode and a postcricoid surface electrode. *Laryngoscope* 108:1418–1420.
- Rooth M (1992) A theory of focus interpretation. *Nat Lang Semant An Int J Semant Its Interfaces Gramm* 1:75–116.
- Rooth ME (1985) Association with focus.
- Scherer KR (1989) Vocal correlates of emotional arousal and affective disturbance. In: *Handbook of psychophysiology: Emotion and social behavior*, pp 165–197. London: John Wiley & Sons.
- Simonyan K, Horwitz B (2011a) Laryngeal Motor Cortex and Control of Speech in Humans. *Neurosci*.
- Simonyan K, Horwitz B (2011b) Laryngeal motor cortex and control of speech in humans. *Neuroscientist* 17:197–208.
- Simonyan K, Jürgens U (2002) Cortico-cortical projections of the motorcortical larynx area in the rhesus monkey. *Brain Res* 949:23–31.
- Stevens SS (1935) The relation of pitch to intensity. *J Acoust Soc Am* 6:150–154.
- Tang C, Hamilton LS, Chang EF (2017) Cortical representation of speech intonation in human non-primary auditory cortex. *Science* (80- ).
- Tate MC, Guo L, McEvoy J, Chang EF (2013) Safety and efficacy of motor mapping utilizing short pulse train direct cortical stimulation. *Stereotact Funct Neurosurg* 91:379–385.
- Titze IR, Luschei ES, Hirano M (1989) Role of the thyroarytenoid muscle in regulation of fundamental frequency. *J Voice* 3:213–224.

Wilson SM, Saygin AP, Sereno MI, Iacoboni M (2004) Listening to speech activates motor areas involved in speech production. *Nat Neurosci* 7:701–702 Available at:  
<http://www.ncbi.nlm.nih.gov/pubmed/15184903>.

Wrench A (1999) The MOCHA-TIMIT articulatory database.

Wu W, Black MJ, Gao Y, Bienenstock E, Serruya M, Shaikhouni A, Donoghue JP (2003) Neural Decoding of Cursor Motion Using a Kalman Filter. *Adv Neural Inf Process Syst* 15 Proc 2002 Conf:133–140.

## Chapter 4. Decoding Prosody

Brain computer interfacing (BCI) enables the control a computer using neural activity directly, and is a promising technology for assistive communication devices of paralyzed individuals (Wolpaw et al., 2002; Hochberg et al., 2006). While many BCIs focus on the movement of cursors (Leuthardt et al, 2004) and prosthetic arms (Collinger et al., 2013), speech BCIs could restore the ability to create speech sounds. Speech BCIs would record from neural activity involved in the speech networks of the brain and output speech as desired by the user. In chapter 3, I outline how prosody has expressive power across many linguistic levels, and showed that the control of vocal pitch is represented in cortical activity. It would thus be beneficial for an assistive communication technology to capture this meaning by estimating the intended intonation of the user. Here, I examine the possibility of incorporating prosody control in a speech BCI. I now develop a framework for using neural activity to decode prosodic information directly from the brain.

Two decoding approaches are developed: pitch decoding and word-of-emphasis decoding. In pitch decoding, pitch is treated as a variable that is continuous both in value and in time, and fluctuations in pitch are extracted from neural activity. In word-of-emphasis decoding, pitch is not estimated, but rather the word-of-emphasis is estimated directly. The word-of-emphasis is treated as a discrete variable, which is classified using neural activation as features.

The methods for the task, recording setup, high gamma extraction, and pitch calculation are described in Chapter 3. A single female subject was used, who had a right hemisphere high-density electrocorticography grid neurosurgically implanted laterally as part of her treatment for epilepsy.

## Pitch Decoding

To reconstruct pitch accurately, I used a dynamical model of pitch combined with a model of the relationship between vocal pitch and neural activity. I used a Kalman Filter, which has been widely, and successfully used in reaching BCIs (Wu et al., 2003) and which implies the following assumptions:

1. **Pitch is a Markov dynamical process.** In other words, the probability of the state at time  $t+1$  given the state at time  $t$  is independent of the previous states. In this case, since I using the value and velocity of pitch, this means that all higher order dynamics (e.g. acceleration) will have a negligible benefit for estimating the value and velocity of pitch at the next time-step.
2. **The dynamics are linear, time-invariant, and Gaussian.** In this case, this means that the state (value and velocity;  $x_t$ ) of pitch at a specific time can be predicted by a matrix multiplication of the previous state, with a normally distributed error and that matrix does not change.
3. **The encoding model is linear and Gaussian.** Here, the high gamma analytic amplitude of each electrode for a specific point in time ( $y_t$ ) can be predicted by a matrix multiplication of  $x_t$ , and that the high gamma analytic amplitude has a normal distribution around that value.

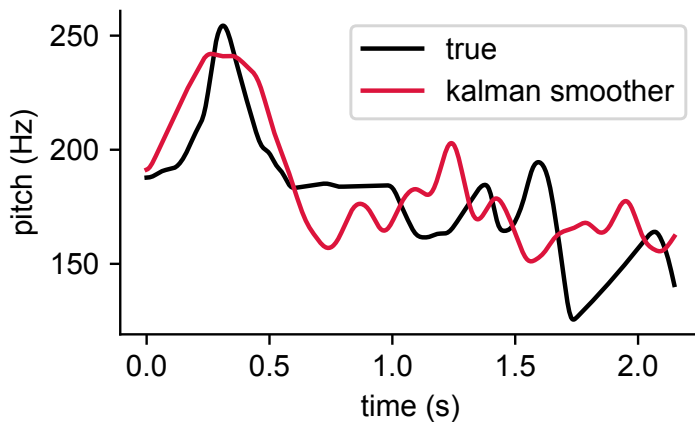
Putting these assumptions together:

$$\begin{aligned}x_{t+1} &= Ax_t + w_t & w_t &\sim \mathcal{N}(0, Q) \\y_t &= Cx_t + v_t & v_t &\sim \mathcal{N}(0, R)\end{aligned}$$

Where  $w$  and  $v$  are noise variables with covariances  $Q$  and  $R$  respectively. Since the pitch states ( $x$ ) are extracted from the acoustics of the subject's speech, and ( $y$ ) are extracted from the neural recordings,  $A$  and  $C$  are learned simply with linear regression. No expectation maximization is required, as if would be for unobserved  $x$ . We use a Kalman smoother approach, which performs inference of the pitch contour using the neural activity recorded

through the entire sentence. This solution could not be performed in truly real-time, since it uses activity later in the sentence to inform estimates of pitch earlier in the sentence. In practice, this inference must be made using data recorded from the entire sentence.

The model was trained for all electrodes on the high density ECoG grid for a single subject. The model was trained on the first 30 sentences, and tested on the remaining 8. There was a correlation of 0.8 between the extracted and inferred pitch of the held-out sentences (Fig. 1).



**Figure 1 | Example pitch decode of a sentence.** The black line shows the extracted pitch for a held-out sentence where the subject emphasized the word “I”. The red line shows the pitch contour inferred from the neural data using a Kalman smoother.

### Decoding Word-of-Emphasis

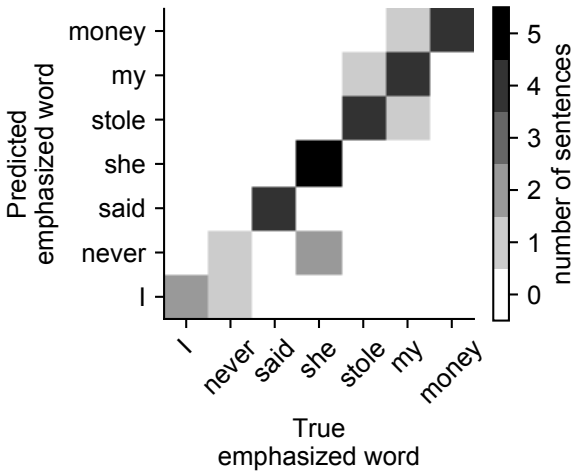
Next, we determined how well the word-of-emphasis could be decoded directly from cortical activity patterns. We removed all of the question condition sentences from the collected data, and were left with only sentences where one of the seven words was emphasized. We used linear discriminant analysis, which models the relationship between neural activity and word of emphasis as following generative model (Fisher, 1940):

$$Y_n | w_n = i \sim \mathcal{N}(\mu_i, \Sigma)$$

where  $n$  is the sentence number,  $i$  is one of 7 words, and  $w_n$  is the word that was emphasized for sentence  $n$ . The word boundaries were determined for each sentence using a phoneme aligner. Then the high gamma analytic amplitude was averaged over the duration of each word in the sentence. This resulted in 256 (number of electrodes) x 7 (number of words) = 1,792

features per sentence.  $Y_n$  is the high gamma analytic amplitude averaged for each word and combined across words for each sentence  $n$ .

We used the same data as in pitch decoding. A leave-one-sentence-out scheme was used so that performance reflected what would be expected for unseen data. We found an accuracy of 80%, and all of the misclassifications were within 2 words of the correct word (Fig. 2).



**Figure 2 | Confusion matrix for word-of-emphasis classification.** Leave-one-out results are shown. Weight on the diagonal reflects accurate decoding. Accuracy is 80% and all errors are within 2 words of the correct word.

Prosody has the potential to increase the richness of communication for assistive communication technology, yet few technologies have the ability to incorporate this feature. Here, we demonstrated successful decoding of spoken prosody from cortical activation in a continuous and discrete modes. This is a proof-of-concept using neural recording technology that is currently only available in neurosurgical patients who are undergoing this procedure as part of their treatment of epilepsy. Neurosurgery is rarely used for assistive technologies, and has not yet been used for a speech prosthetic. These results indicate that cortical signals are able to recover prosodic information that is difficult to achieve by other means.

One caveat of this study is that it is unclear how much prosodic information is due to a feedforward command and how much is due to an acoustic encoding of the subject hearing her own voice. Stimulation results from Chapter 3 would suggest that at least some of the cortical activity is actually representing a feedforward command, however some of the activity,

particularly in known auditory areas, could be due to the subject's own voice, and therefore might not generalize well to use as an assistive technology. Further work is needed to explore how this approach will perform in a subject who is unable to speak.



## References

- Collinger J, et al. (2013) High-performance neuroprosthetic control by an individual with tetraplegia. *The Lancet*. 381:557-564.
- Fisher, Ronald A. (1940) The precision of discriminant functions. *Annals of Human Genetics* 10.1:422-429.
- Leuthardt E, et al. (2004) A brain-computer interface using electrocorticography signals in humans. *Journal of neural engineering*. 1:63-71.
- Wolpaw, J et al. (2002) Brain-computer interfaces for communication and control. *Clinical neurophysiology*. 113:767-91.
- Wu, W, et al. (2003) Neural Decoding of Cursor Motion Using a Kalman Filter. *Advances in Neural Information Processing Systems* 15. 113-140

## Conclusion

Whether monitoring the movement of a limb or understanding speech, making sense of one's surroundings and responding appropriately requires precise encoding of the senses. Here, I have examined encoding of the senses theoretically and experimentally. I have demonstrated how dynamical state estimation can be achieved within the physical limitations of the brain and characterize the representation in a specific sensory modality. I also examined experimental data of the cortical encoding of speech and found a quenching of variability during auditory encoding. Finally, I showed how the brain commands control movement of an effector. I chose to examine the larynx because humans have a unique ability among primates to flexibly mimic pitch contours using their larynx.

The human control of the larynx is a particularly interesting behavior, because humans appear to be specialized in the flexibility of its control, an ability that allows us to produce the pitch fluctuations that carry meaning in speech. Non-human primates do not appear to have this flexible control, and it is hypothesized that the difference in our abilities stems from differences in neural structure and activity. Here, we have shown representation of laryngeal control in a brain region that appears to be unique to humans among primates, the dorsal laryngeal motor cortex (dLMC). These results suggest that the dLMC is crucial for the fine control of pitch present in speech.

In the auditory domain, pitch tends to follow the "receptive field" model outlined in the Introduction, where neurons have a specific preferred pitch, and respond according to the proximity of an incoming signal to that pitch (this is why we see both positive and negative correlations in the auditory response). However, the representation of pitch command we find in the dLMC follows a linear model, where higher pitch correlates with higher activity across the area. This finding illustrates a difference between auditory and motor representations of pitch.

We might be able to use this difference to determine if other speech areas are encoding in the auditory or motor space. This difference might also apply to other sensory and motor representations-- that motor actions are encoded according to the muscle activity necessary for the action and senses are not.

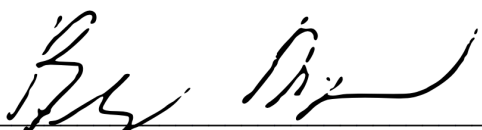
Further work is needed to understand how this area appeared in the evolutionary track of humans, and what precise role it played in the development of spoken language. I finish by demonstrating how these neural signals might be incorporated in a neural prosthetic device to improve the communicative ability of patients with difficulty speaking.

**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

08/23/17  
\_\_\_\_\_  
Date