Neural decoding of spoken vowels from human sensory-motor cortex with high-density electrocorticography

Kristofer E. Bouchard[1,2,3,4] and Edward F. Chang[1,2,3,5*]

Abstract— **We present the first demonstration of single-trial neural decoding of vowel acoustic features during speech production with high performance. The ability to predict trial-by-trial fluctuations in speech production was facilitated by using high-density, large-area electrocorticography (ECoG) combined with an adaptive principal components regression. In experiments from two human neurosurgical patients with a high-density 256-channel ECoG grid implanted over speech cortices, we demonstrate that as much as 81% of the acoustic variability across vowels could be accurately predicted from the spatial patterns of neural activity during speech production. These results demonstrate continuous, single-trial decoding of vowel acoustics.**

I. Introduction

Dysfunction of the central control of speech articulation affects a large number of primary communication disorders including stuttering, aphasia, apraxia of speech, and most devastatingly of all, 'locked-in' syndrome, in which people have lost the ability to communicate through spoken language. Our understanding of the neural processes that generate speech are greatly limited [1]. Studying the neural control of speech presents several distinct challenges. First, speech production is a unique human ability, and therefore can only be studied in humans. Second, the generation of speech requires the precise and coordinated control of several effectors on rapid time scales. Finally, as demonstrated by our previous work, the effectors of the speech plant (i.e. the articulators of the vocal tract, e.g. the lips, tongue, jaw, larynx) are somatotopically represented over ~1300mm^2 of human sensory-motor cortex, and the representation between articulators can transition over spatial scales less then 5mm [2]. Together, the rapid coordination of multiple articulator representations which are spatially localized over large areas of cortex requires high-density, large area recordings with high temporal resolution. Although arrays of penetrating electrodes (e.g. Utah arrays) are capable of recording the spiking activity of 10's of neurons, their limited spatial coverage makes them unsuitable for studying speech. Electrocorticography (ECoG) potentially provides the broad spatial coverage and high temporal resolution, but the standard low-density grids (1cm pitch) lack sufficient spatial resolution to simultaneously monitor the activity of all speech articulator representations. The current state-of-the art brain-machine interfaces for restoring speech show promise [3-4], but are not yet clinically viable.

II. Experimental Methods

A. Subjects and Task

The experimental protocol was approved by the Human Research Protection Program at the University of California, San Francisco. Two native English speaking human subjects underwent chronic implantation of a high-density, subdural electrocortigraphic (ECoG) array over the left hemisphere as part of their clinical treatment of epilepsy [2]. Subjects gave their written informed consent before the day of surgery. All subjects had self-reported normal hearing and underwent neuro-psychological language testing (including the Boston Naming and verbal fluency tests) and were found to be normal.

Each subject read aloud consonant-vowel syllables (CVs) composed of 19 consonants followed by one of three vowels (/a/, /i/ or /u/). Each CV was produced between 15 and 100 times total. Across two subjects, data were taken on 14 different recording sessions. Because we observed that the recorded ECoG signal from a patient could vary from block-to-block, these different recording sessions were used as the samples across which statistical tests were performed.

B. 256 channel high-density electrocorticography

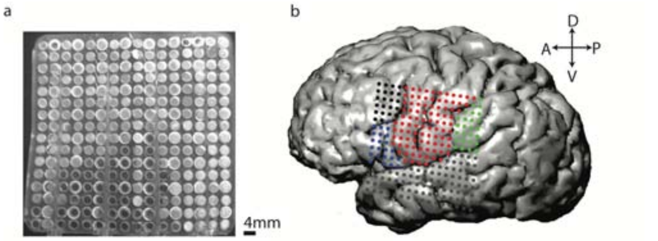

Fig. 1. **High-density electrocorticography from speech areas.** (a) Photograph of high-density (4mm pitch), 256-channel electrocorticography (ECoG) grid. (b) Reconstructed location of ECoG electrode locations over the left hemisphere from one patient [red, ventral sensorimotor cortex; blue, Broca's area; grey, superior and middle temporal gyri; black and green, non-speech frontal (black) and parietal (green) cortices].

[1] Departments of Neurological Surgery and Physiology, University of California, San Francisco

[2] Center for Integrative Neuroscience, University of California, San Francisco

[3] Center for Neural Engineering and Prostheses, University of California, San Francisco and University of California, Berkeley

[4] Computational Research Division, Lawrence-Berkeley National Laboratory, Berkeley

[5] UCSF Epilepsy Center, University of California, San Francisco

E.F.C. was funded by the US National Institutes of Health grants R00-NS065120, DP2-OD00862 and R01-DC012379, and the Ester A. and Joseph Klingenstein Foundation.

We used a customized high-density, large channel count electrocorticography (ECoG) array implanted subdurally to record electrical field potentials directly from the cortical surface. The array had a total of 256 electrode contacts, in a 16 x 16 configuration. Each contact on the array was a 1.5 mm diameter platinum disk with an impedance of ~1-10 Ω (measured in saline). The contacts and connecting wires were embedded in silastic to allow conformability to the cortical surface (Fig.1).

Electrical field potentials were recorded directly from the cortical surface with ECoG arrays and a multi-channel amplifier optically connected to a digital signal processor (Tucker-Davis Technologies [TDT], Alachua, FL). The spoken syllables were recorded with a microphone, digitally amplified, and recorded inline with the ECoG data. ECoG signals were acquired at 3052 Hz. The acoustic signal was acquired at 22kHz. The time series from each channel was visually and quantitatively inspected for artifacts or excessive noise (typically 60 Hz line noise). These channels were excluded from all subsequent analysis and the raw recorded ECoG signal of the remaining channels were then common average referenced and used for spectro-temporal analysis. For each (useable) channel, the time-varying analytic amplitude was extracted from eight bandpass filters (Gaussian filters, logarithmically increasing center frequencies [70-150 Hz] and semi-logarithmically increasing band-widths) with the Hilbert transform. The high-gamma (Hγ) activity was calculated by averaging the analytic amplitude across these eight bands. This signal was down-sampled to 200 Hz and z-scored relative to baseline activity for each channel.

III. Single-trial Decoding of Vowel Acoustics During Speech Production

A. Extraction of acoustic features

Each subject read aloud consonant-vowel syllables (CVs) composed of 18-19 consonants followed by one of the three cardinal vowels (/a/, /i/, or /u/) [2]. /a/, /i/, and /u/ are considered cardinal vowels because they span the acoustic and articulatory space of all vowels, and are found in most of the world's languages [5]. The recorded speech signal was transcribed off-line by a certified speech pathologist using WaveSurfer (http://www.speech.kth.se/wavesurfer/). Vowels are defined by the combination of acoustic features, termed formants. Vowel formants reflect the resonant properties of the vocal tract, which is shaped by the configuration of speech articulators. We measured the vowel formant, F_1-F_4, as a function of time for each utterance of a vowel using an inverse filter method [6]. Briefly, the signal is inverse filtered with an initial estimate of F_2 and then the dominant frequency in the filtered signal is used as an estimate of F_1. The signal is then inverse filtered again, this time with an inverse of the estimate of F_1, and the output is used to refine the estimate of F_2. This procedure is repeated until convergence and is also used to find F_3 and F_4. The inverse filter method converges on very accurate estimates of the vowel formants, without making assumptions inherent in the more widely used linear predictive coding (LPC) method. In Figure 2, we present single-trial time-courses of the F_2/F_1 formant ratio for the cardinal vowels /a/, /i/, and /u/, as well as a scatter plot of the values extracted from the center of each vowel, from one patient.

B. Predicting vowel formants from high-gamma activity

We examined the ability of spatial patterns of neural activity from ventral sensorimotor cortex (vSMC) to predict vowel acoustics on a single-trial basis. We focused on the amplitude of high-gamma band activity from vSMC; ~85 electrodes where located in vSMC for each subject [2]. High-gamma amplitude has been shown to have high spatio-temporal resolution, and correlates well with mid-laminar multi-unit activity [7-8]. We used principle components linear regression combined with a two-stage model optimization procedure decode the acoustics of vowels from vSMC high-gamma activity [9]. Principal components analysis (PCA) was performed on the set of all vSMC electrodes for dimensionality reduction and orthogonalization. This also ensures that the matrices in the calculation of least mean squared error estimators (from multivariate regression below) were well scaled. PCA was performed independently for each non-overlapping 10 ms window preceding the acoustic measurement. First, for each electrode (e_j of which there are n) and syllable utterance (s, of which there are m), we calculated the mean high-gamma activity (Hγ) in 10 ms windows with a non-overlapping two-sample moving average of Hγ with time lag τ. The $H\gamma_j(\tau,s)$ were used as entries in the n x m data matrix **D**, with rows corresponding to channels (of which there are n) and columns corresponding to the number of utterances within a recording session (of which there are m). Each electrode's activity was z-scored across utterances to normalize response variability across electrodes. PCA was performed on the n x n covariance matrix **Z** derived from **D**. The singular-value decomposition of **Z** was used to find the eigenvector matrix **M** and associated eigenvalues. The PCs derived in this way serve as a spatial filter of the electrodes, with each electrode e_j receiving a weighting in PC_i equal to m_{ij}, the i-j^th element of **M**, the matrix of eigenvectors. We included the leading 40 eigenvectors in our analysis. For each utterance (s), we projected the vector Hγ(τ,s) of high-gamma activity across electrodes into the leading 40 eigenvectors ($\mathbf{M}^{40}$):

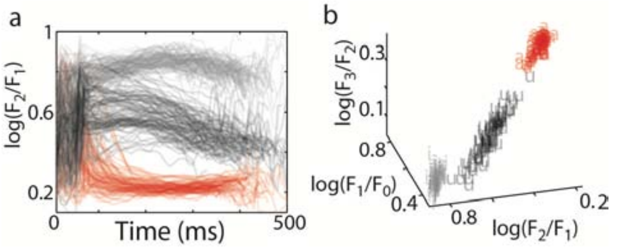

Fig. 2. **Single-trial acoustics of vowels.** (a) Time courses of the F_2/F_1 ratio from more than 100 utterances of the vowels /a/ (red), /i/ (grey), and /u/ (black), from one patient. (b) Scatter plot of formant ratios of the vowels /a/ (red), /i/ (grey), and /u/ (black), extracted from the middle of each utterance. T = 0 is the acoustic onset of the consonant-to-vowel transition.

$$\Psi(\tau,s) = \mathbf{M}^{40}\bullet H\gamma(\tau,s) \qquad (1)$$

It is important to emphasize that the approach described above identifies principal components (spatial filters) derived only from the spatial structure of the data (structure of $H\gamma$ across electrodes); the temporal structure of the $H\gamma$ population does not enter into **M** in any way. Thus, the PC's are completely local in time, up to the autocorrelation of the $H\gamma$ signal itself.

Formant Decoding Model

For each non-overlapping 10ms time window (τ) preceding the behavioral measurement, $\Psi(\tau,s)$ (equation 1) served as the basis for training and testing optimal linear predictors of single-trial vowel formant features. We used a simple linear model to predict the formant features ($F_i(s)$) for a syllable (s) from $\Psi(\tau,s)$:

$$f_i(s) = \beta\bullet\Psi(\tau,s) + \beta_0 \qquad (2)$$

Here, $f_i(s)$ is the best linear estimate of $F_i(s)$ based on the cortical features. The vector of weights (β) that minimized the mean squared error between $f_i(s)$ and $F_i(s)$ was found through a two-stage optimization of multi-linear regression.

Adaptive Threshold Selection-OLS refit decoding

To train linear predictors of produced acoustics from neural data, we innovated a two-stage estimation procedure utilizing adaptive threshold selection of model parameters followed by ordinary least squares refitting of the selected parameters. We thus term this procedure ATS-OLS refit. This approach used cross-validation to train and test separate linear models to predict across vowel acoustic features [9]. Separate models were trained/tested for each time-point (dt = 10ms) and recording block. Predictive performance was calculated as coefficient of determination, R^2. We first verbally describe our decoding approach, and then provide a more formal treatment.

The methodology used here is as follows. First, we derive null distributions of weights (β_{null}) and model performance (R^2_{null}). This was accomplished by randomly permuting (200 times) each vowel formant relative to Ψ on a trial-by-trial basis, yielding randomized data pairings Z_{rnd}. We estimated weights (Equation 2) with OLS minimization on Z_{rnd}, and determined predictive.

We then used 5-fold cross validation to train and test a two-stage estimation of the linear mapping from cortical features to acoustics. The 5-fold cross-validation procedure was included in a 200-iteration bootstrap to arrive at estimates of mean expected models and predictive performance across different data subsets. Specifically, on each iteration, a random 80% subset of the data (Z_{trn}) was used to derive initial estimates of linear weights for the models (Equation 2), and the performance of these models was calculated on the remaining 20% of the data not used in training (Z_{tst}). From this, we arrived at weights (β_{init}) describing the mapping from all cortical activity patterns to each formant features.

We then reduced the dimensionality of the cortical features ('parameter selection') by comparing the observed weights (β_{init}) to the weights derived from randomly permuted data sets (β_{null}) to identify cortical features with weights that were different between the two conditions. Here, cortical features (Ψ_j) were retained if the initial estimate of the weight magnitude ($|\beta^j_{init}|$) was greater than the mean plus one standard deviation of the distribution of weight magnitudes derived from the randomization procedure ($|\beta^j_{rnd}|$). Finally, we re-fit the model based only on this reduced set of cortical features (using the same training data, Z_{trn}), to arrive at optimal weights (β_{opt}), and determined decoding performance (R^2_{opt}) on test data (Z_{tst}).

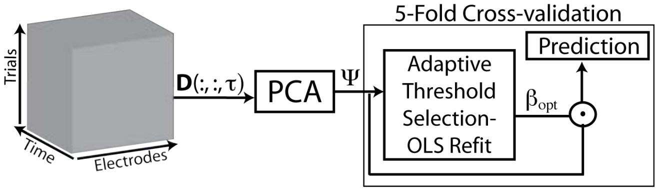

Fig. 3. **Adaptive Threshold Selection-OLS Refit of Principal Components.** At each point in time preceding the acoustic measurements (τ), spatial PCA was performed on the data matrix **D** of high-gamma activity. The data were projected into the leading principal components, yielding $\mathbf{\Psi}$, the projection of the neural activity into the corresponding orthogonal sub-space that maximized the variability amongst the spatial patterns of activity. This projection served as input to a linear decoder, trained and tested with 5-fold cross-validation. The linear weights (β) were selected and by a hard-thresholding procedure based on the distribution of null weights, and the reduced model was refit by ordinary least squares.

The decoding performance for each block and decoding condition was taken as the mean of R^2_{opt} values across random test samples. This quantifies the expected value of predictive decoding performance across randomly selected training and test samples. We confirmed that the expected value of R^2 under the null hypothesis for our data and procedure was 0 by examining the distributions of R^2_{rnd}. Across all blocks, times and conditions, R^2_{rnd} had a median very close to 0 (median < 0.001 for all). Note that, as there are more cortical features in the model used to derive R^2_{rnd} than R^2_{reg}, comparing R^2_{reg} to R^2_{rnd} is a conservative approach for statistical testing. Therefore, we gauged the significance of the across block distributions of R^2_{reg} for each feature and time-window by performing t-tests against the null-hypothesis of 0. The conclusions of significance were insensitive to different statistical tests.

This selection-refitting procedure resulted in improved decoding performance (up to ~10%) on test data. The choice of threshold (mean plus one standard deviation of null distribution) was chosen by visual examination of the weight distributions. As we describe below, an optimization of this threshold may have resulted in better model performance; however, because the chosen threshold resulted in good decoding performance, this optimization was not done to reduce computational run-time. This approach to parameter selection/estimation has the advantage of not imposing a prior over the distribution of weights, as is the case when using either the L_1-norm (i.e. lasso imposes a Laplacian prior) or L_2-norm (i.e. ridge regression imposes a Gaussian prior) to penalize the weight distribution [9].

We now describe the ATS-OLS refit procedure formally, and in full generality. Let $Z_i = (x_i, y_i), i = 1, \dots$m, be the m measurement pairs of output $y \in \mathbb{R}$ and d-dimensional input features $x \in \mathbb{R}^d$. The first step of the ATS-OLS refit procedure, the is to estimate the null distribution of model weights ($\hat{\beta}_{null}$) by, e.g. randomly permuting the relationship between inputs and outputs (Z_{rnd}) multiple times:

$$\hat{\beta}_{null} = E(\underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \mathcal{L}(\beta, Z_{rnd})) \tag{3}$$

Here, we use the typical least-squares loss function:

$$\mathcal{L}(\beta, Z) = \sum_{i=1}^{m} (y_i - \beta x_i)^2 \tag{4}$$

Then, divide the m measurements of input-output pairings $Z_i = (x_i, y_i), i = 1, \dots$m, into non-overlapping train (Z_{trn}), select (Z_{slct}), and test (Z_{tst}) sets for model training, selection, and testing (i.e. cross-validation). Next, derive an initial estimate of model parameters from Z_{trn}:

$$\beta_{init} = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \mathcal{L}(\beta, Z_{trn}) \tag{5}$$

Next, for a range of thresholds, set model parameters to zero if the magnitude of β^j_{init} is less than a multiple (λ) of the expected value of the null magnitudes ($\hat{\beta}^j_{null}$):

$$if \ |\beta^j_{init}| < |\hat{\beta}^j_{null}| \times \lambda, \text{then } \beta^j_{sel}(\lambda) = 0 \tag{6}$$

and re-fit the d-n $\beta^j_{sel} \neq 0$ on Z_{trn} using ordinary least-squares:

$$\beta_{sel}(\lambda) = \underset{\beta \in \mathbb{R}^{d-n}}{\operatorname{argmin}} \mathcal{L}(\beta, Z_{trn}) \tag{7}$$

Finally, select the optimal model parameters (β_{opt}) as the β_{sel} that minimize the expected loss on out-of-sample data (Z_{sel}) across thresholds (λ):

$$\beta_{opt} = \underset{\lambda}{\operatorname{argmin}} \ \mathcal{L}(\beta_{sel}(\lambda), Z_{sel}) \tag{8}$$

We calculated expected predictive performance (R^2) on data (Z_{tst}) not used in parameter training or selection. The specific decoding approach applied in this study is a special case of the ATS-OLS refit method.

IV. Results

We found that spatial patterns of high-gamma activity could be linearly decoded to predict formant features across the cardinal vowels with high fidelity. Decoding performance was greatest ~150ms before the onset of the acoustic measurement. The scatter plot in Figure 4 presents results from a single recording session from one subject in which 361 CV syllables were spoken. For this subject, on average across sessions, 81% of the variability in F_2/F_1 across the vowels could be accurately predicted from the vSMC population neural activity.

We observed that the ability to predict across vowel variability from linear decoders of vSMC spatial patterns of activity varied across different acoustic features. In Figure 5, we plot mean performance of linear decoders for predicting different acoustic features from single-trial activity, averaged across several recording sessions in two patients. On average, predictive performance was highest for the F_2/F_1 ratio.

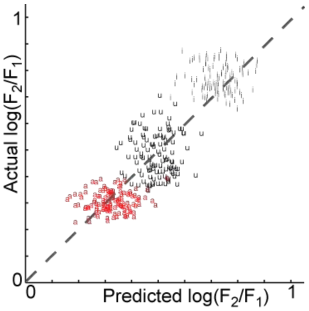

Fig. 4. **Example single-trial acoustic decoding from ECoG.** Scatter plot of the predicted $\log(F_2/F_1)$ ratio vs. the actual $\log(F_2/F_1)$ ratio from one recording session. Red: /a/; Black: /u/; Grey: /i/. Dashed grey line is unity

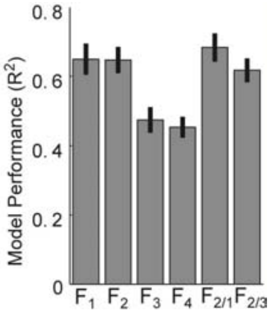

Fig. 5.**Performance of decoder across different acoustic features and recording sessions.** Data are presented as mean ± s.e. from 14 recording sessions in two subjects.

V. Conclusion

We have shown that linear decoders of high-gamma activity recorded from high-density ECOG over the ventral sensorimotor cortex of speaking humans can predict the produced acoustic features of the cardinals vowels with high-performance. These results suggest that continuous decoding of speech maybe a viable approach to speech prosthetics.

References

[1] Conant,D., Bouchard, K.E., Chang, E.F. "Speech map in the human ventral sensory-motor cortex", *Current Opinion in Neurobiology* 24, 63-67, 2014.

[2] Bouchard, K.E., Mesgarani, N., Johnson, K.E., Chang, E.F. "Functional organization of human sensorimotor cortex for speech articulation", Nature 495, 327-332, 2013.

[3] Guenther, F. H. et al., "A Wireless Brain-Machine Interface for Real-Time Speech Synthesis", PLoS One, 2009.

[4] Leuthardt , E.C., et al, "Using the electrocorticographic speech network to control a brain–computer interface in humans", J. Neural Eng., 8, 2011.

[5] Ladefoged, P, Johnson, K. "A course in phonetics", Cengage learning, 2014.

[6] Ueda, Y., et al. "A real-time formant tracker based on the inverse filter control method", Acoustical Science and Technology of the Acoustical Science of Japan 28(4), 271-4, 2007.

[7] Crone, N.E. et al., "Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis- II. Event-related synchronization in the gamma band", Brain 121, 2301-2315,1998.

[8] Ray, S. and Maunsell, J.H.R., "Different origins of gamma and high-gamma activity in Macaque visual cortex", PLoS Biol 9, 2011.

[9] Hastie, T., Tibshirani, R., Friedman, J., "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer, 2009.