Classifying spoken syllables from human sensorimotor cortex with deep networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

We investigate deep neural networks (DNNs) for closed vocabulary speech classification of electrocorticography (ECoG) recordings from human sensorimotor cortex. The behavioral data consisted of multiple instances of 57 isolated consonant-vowel syllables. Fully connected and convolutional networks are compared against linear models, (e.g. LDA and HMM-GMM), traditionally used in automatic speech recognition and speech classification from ECoG. We show that DNNs exceed state-of-the-art results on classifying speech from human sensorimotor cortex, with the best performance of 38.7% for syllable classification using a fully connected network (22x over chance). Furthermore, DNNs more efficiently use data, having a steeper performance increase per training example, an important consideration for neural data sets. Finally, we characterize the inherent structure in confusions made by the classifier, revealing the articulatory nature of neural processes involved in speech production. Together, these results demonstrate the power of DNNs for extracting meaningful structure from noisy, single-trial brain signals and suggest that deep networks will be useful for next generation brain-machine interfaces.

1 Introduction

Vocal articulation is a complex task requiring the coordinated and continuous orchestration of several parts of the vocal tract. To study the neural basis of speech requires monitoring cortical activity at high spatio-temporal resolution (on the order of tens of milliseconds) over large areas of sensorimotor cortex ($\sim$1300mm^2) [1]. To achieve the simultaneous high-resolution and broad coverage requirements in humans, electrocorticography (ECoG) is an ideal method for recording neural signals. Using such recordings, there has been a surge of recent efforts to understand the cortical basis of speech production [1, 2, 3, 4, 5]. For example, analyzing mean activity, Bouchard et. al. [1] demonstrated, much in the spirit of Penfield's earlier work [6], that the ventral sensorimotor cortex (vSMC) has a spatial map of articulator representations (i.e. lips, jaw, tongue, and larynx) that are engaged during speech production. Additionally, it was found that spatial patterns of activity across the vSMC network (extracted with principal components analysis (PCA) at specific time points) organized phonemes along phonetic features emphasizing the articulatory requirements of production. Building off of these results, Bouchard and Chang [2, 3] demonstrated high-performance, single-trial continuous prediction of produced vowel acoustics from these recordings.

Many studies have attempted to classify produced speech from ECoG activity. For example, whole word studies have achieved classification performance just under 50% on a ten word corpus (5x over chance, 10%) [7]. However, classifying smaller phonetic parts of speech may better match the internal representations [1] and, in the context of brain machine interfaces for speech prosthetics, would exploit the inherent combinatorial nature of language. Pei et al. [8] used consonant-vowel-consonant (CVC) syllables and classified vowels or consonant pairs achieving 40.7 ± 2.7% for vowels and 40.6 ± 8.3% for consonant pairs (both 1.5x over chance, 25%) with four classes in each task. Herff et al. [9] use techniques from speech recognition and achieve phone accuracies above 50% (11x over chance, 4.7%), but have incorporated a language model into their classifier unlike other studies. To our knowledge Mugler et al. [4] have achieved the best speech classification performance purely from ECoG recordings, with 36% accuracy on 24 consonants (5x over chance, 7.4%) and 24% on 15 vowels in a single subject (2x over chance, 12.9%).

Building on these methods, deep networks have surpassed or become competitive with previous state-of-the-art models in a number of fields including computer vision and speech recognition [10, 11]. These networks can have many architectures including convolution and recurrence that lend themselves to structured ECoG-like data such as images [**?**], or time series [12], in addition to 'flat' data. Deep neural networks (DNNs) have recently been applied as classifiers for other types of physiological data including electromyographic (EMG) and electroencephalographic (EEG) signals [13, 14, 15, 16] and on stimulus reconstruction in sensory regions using ECoG [17]. Our goal was to investigate deep networks as an alternative computing framework for brain machine interfaces/neural prosthetics, with a specific focus on the uniquely human capacity to produce spoken language, and to examine the structure of speech representations discovered by the networks towards understanding brain computations in general.

2 Methods

2.1 Experimental Data

The experimental protocol, collection, and processing of the data examined here have been described in detail previously [1, 2, 3]. Briefly, one native English speaking human subject underwent chronic implantation of a subdural electrocortigraphic (ECoG) array over the left hemisphere as part of their clinical treatment of epilepsy. The subject gave their written informed consent before the day of surgery. The subject read aloud consonant-vowel syllables (CVs) composed of 19 consonants followed by one of three vowels (/a/, /i/ or /u/), for a total of 57 Consonant-Vowel (CV) syllables. All syllables were produced approximately 45 times, except for /θi/, which only had 10 examples. In total, there were 2621 produced CV syllables.

Electrical field potentials were recorded directly from the cortical surface with a high-density (4mm pitch), 256-channel ECoG array and a multi-channel amplifier optically connected to a digital signal processor (Tucker-Davis Technologies [TDT], Alachua, FL). 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 from 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. Baseline is defined as a period of time in which the subjects were silent, the room was silent, and the subject was resting. Based on the results described in [1, 2, 3], we focused on the 85 electrodes in the ventral sensorimotor cortex (vSMC). The Hγ activity for each of the 2621 examples in our data set was aligned to the acoustic onset of the consonant-to-vowel transition. For each example, 1.3 seconds of data was extracted, giving 258 time points per example. The mean of the first and last ten time points was subtracted from the data for each channel and vocalization. Example ECoG data for average activity of three CVs /ba/, /da/ and /ga/, processed in the way described above, is shown in Figure 1. As described previously [1], different syllables are generated by distinct, but partially overlapping spatio-temporal patterns of activity. Furthermore, the Hγ activity of many electrodes begins rising several hundred milliseconds before acoustic onset, emphasizing the motor nature of the recordings. Further details are provided in Supplementary Information.

2.2 Deep Networks

Recently, deep networks have demonstrated state-of-the-art classification performance on a number of tasks [18]. In this work, we evaluate DNNs as classifiers of produced speech from human ECoG.

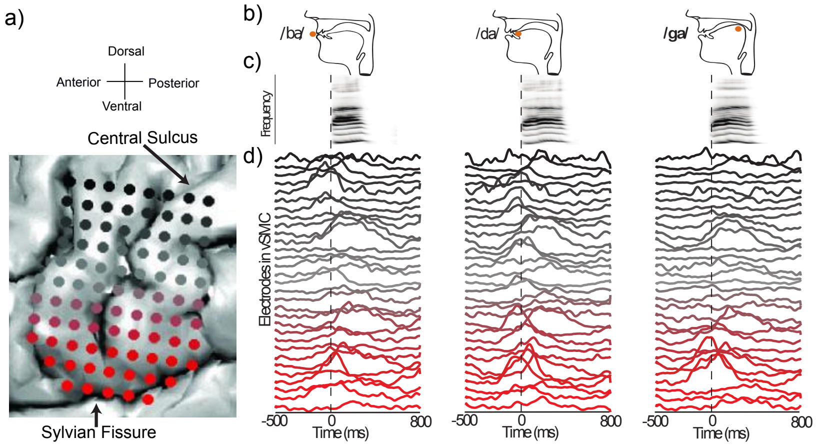

Figure 1: ECoG data from three different CV utterances. **a)** Electrode layout over-layed on the vSMC. **b)** Articulator position and point of constriction for the production of /ba/ (lips), /da/ (coronal tongue), and /ga/ (dorsal tongue). **c)** The audio spectrogram aligned to CV transition (dotted line). **d)** Mean high-gamma activity from electrodes in vSMC aligned to CV transition. Traces are colored red-to-black with increasing distance from the Sylvian fissure. The syllables /ba/, /da/, and /ga/ are generated by spatio-temporal patterns of activity across the vSMC.

In addition to classifying CV syllables, we also consider classifying consonants and vowels as separate entities from the entire time series. Previous ECoG work has primarily used linear models for classifying speech tokens. These models compute linear filters on the data, and are therefore limited in their ability to represent complex nonlinear relationships, which may be present in the data. Multilayer networks can combine features in nonlinear ways when predicting class. This gives them more expressive power in terms of the types of mappings they can learn at the cost of more model parameters to train and more difficult training dynamics.

Each layer in a fully connected network consists of an affine transform and a nonlinearity:

$$y_i = f(w_i \cdot x + b_i), \tag{1}$$

where x is a batch of input vectors, w and b are trainable parameters (weights and biases, respectively), and $f(\cdot)$ is a nonlinearity. Convolutional networks [19] are made up of convolution and maxpooling layers:

$$y_i = \max_{j \in n}(f(w_j * x + b_j)), \tag{2}$$

where x is a batch of structured data, w and b are trainable parameters, $f(\cdot)$ is nonlinearity and $\max_{j \in n}(\cdot)$ takes the maximum value from a group of n neighboring units. Convolutional layers can be followed by fully connected layers. We convolve and maxpool along the time axis as the data is fed in as a time series with 85 channels. Convolutional networks potentially make more efficient use of their parameters by sharing weights across the data dimension that has continuity (time in our case).

To train the networks, the data is randomly organized into 10 subsets (folds) with mutually exclusive test sets and 80-10-10% splits (Training-Validation-Testing). Since the validation and testing sets may have fewer than 10 examples per class it was important to split each class proportionally so that all classes were equally distributed. We also explored data augmentation by temporally shifting each example ±50 ms in 5 ms steps (20× augmentation) to mimic small temporal changes in acoustic-neural alignment [19]. During training, we also scale each example in each new batch by a random scalar factor drawn online from $N(1, .05)$. Training terminated when the validation accuracy did not improve for 10 epochs and typically lasted about 25 epochs. So each training example was rescaled approximately 25 times. Data augmentation in this way makes the classifier insensitive to temporal shifts and scalings, as they carry no class information in the augmented data. Only real data was used in the validation and testing sets.

These networks have a number of hyper-parameters that govern network architecture and optimization, such as the number of layers, the layer nonlinearity, and the optimization parameters (the full list of hyper-parameters is listed in Supplement: Table 2). For all results, hyper-parameters were selected by choosing the model with the best cross-validated classification accuracy on the validation set. Hyper-parameter search was done using a combination of random search, the Spearmint library [20], and the cloud-based optimizer from Whetlab [1]. Deep networks perform best when large amounts of data are used for training. Since our corpus was relatively small for training DNNs, we regularized the models in three ways: dropout, weight decay, and filter norm-clipping in all layers of the model. The dropout rate, activation-rescaling factor, max filter norm, and weight decay coefficient were all optimized hyper-parameters. Pylearn2 [21] was used to train all models.

As baseline models, we trained an HMM-GMM and two linear classifiers on the data. One of the linear classifiers (softmax linear regression) was optimized using the same hyper-parameter selection method as the fully connected networks but was limited to having no hidden layers. The second was one-versus-all linear discriminant analysis (LDA) to compare deep networks with previous results. It is important to note that unsupervised dimensionality reduction is commonly applied before applying LDA. In our case, we have effectively applied 'dimensionality reduction' by choosing the Hγ band and the electrodes in the vSMC. Additionally, our goal was to evaluate methods with as little pre-processing as possible.

2.3 Hidden Markov Models

Hidden-Markov Models (HMMs) have, until recently, been the state-of-the-art generative models in speech recognition, both for their elegant training/prediction algorithms and their generalization capabilities across a range of speakers and speaking styles, given appropriate data [22]. For closed vocabulary prediction, each entry in the lexicon is expanded as a sequence of constituent phonemes, in this case as the consonant and vowel. Each phoneme in turn is modeled as a left-to-right Markov chain of five states per phoneme (two non-emitting initial and final states and three intermediate emitting states). The distribution, $P(x_t|k)$, at state k is modeled as a mixture of Gaussians. Maximum likelihood estimation of the model parameters in training is done using the Baum-Welch algorithm, given sequence data and class labels. The dynamic programming based Viterbi algorithm is used for prediction, where likelihood is assigned to all possible state sequences generating a given unseen sequence of observations, x. The class with the highest likelihood at the final observation vector is chosen as the estimated class label

$$\begin{aligned} V_{1,k} &= P(x_1|k) \cdot \pi_k \\ V_{t,k} &= \max_{n \in S}(P(x_t|k) \cdot a_{n,k} \cdot V_{t-1,n}), \end{aligned} \tag{3}$$

where likelihood $V_{t,k}$ of state k at timepoint t is iteratively computed as the best likelihood at timepoint $t-1$ times the emission probability $P(X_{t-1}|k)$, and transition probability $a_{n,k}$ from state n to state k. π_k is the marginal probability of states. The state sequence giving the best likelihood can be tracked back as the decoded hypothesis. In the experiments reported here, a five state per phoneme model is used, and the emission probabilities modeled as mixtures of eight Gaussians, parameters determined through cross-validation performance over these parameters (see details in Supplement: Section 2).

2.4 Information Transfer Rate

To compare our results with previous speech classification studies, we report estimated information transfer rates (ITR) in addition to classification statistics. ITR is a unified way of calculating the effectiveness of different classifiers, which can have differing numbers of classes, durations, and modalities. To calculate ITR, the symbol rate is multiplied by the information per symbol, defined as the channel capacity, C, between the ground truth class, Y, and predicted class, $\hat{Y}$:

$$C = \sup_{P(Y)} I(\hat{Y};Y) = \sup_{P(Y)} \sum_{\hat{Y}_i} \sum_{Y_j} P(\hat{Y}_i|Y_j)P(Y_j)\log_2\left(\frac{P(\hat{Y}_i|Y_j)}{P(\hat{Y}_i)}\right). \tag{4}$$

For previous work, we must approximate the channel capacity since we do not have access to the details of the classification performance, $P(\hat{Y}|Y)$. Wolpaw et. al. [23] suggest an approximation

[1] www.whetlab.com

that assumes all classes have the same accuracy as the mean accuracy and all errors are distributed equally. To make a fair comparison, we compute this approximate value for our results (and so make fair comparisons) in addition to the exact value. We find that the approximation underestimates the true ITR.

2.5 Performance Trend with Training Data Size

We compared performance scaling of different models by training on variably sized fractions of the training set. For each fraction of the data, samples were randomly chosen so that the training set had roughly equal numbers of examples from each class. The validation and test splits were left the same size. Hyper-parameters were chosen independently for each fraction of the training data using an equal number of hyper-parameter optimization steps.

3 Results

For the models considered, we report overall classification accuracy and peak single phoneme classification accuracy for the best model. All models have classification accuracy significantly above chance ($p < 0.01$, Wilcoxon Signed Rank Test, 10 measurements from folds, chance calculated by training on data with shuffled labels). Additionally, we report consonant and vowel accuracy for multilayer networks. Furthermore, for increasing amounts of training data, we also show how performance scales for different models. Finally, we show that deep networks reveal structure in neural recordings that is not apparent with other methods.

3.1 Deep Networks Achieve State-of-the-Art Classification

Categorical BMIs for speech rely on accurate classification of neural signals for robust and high-throughput communication. To this end, we study the ability of DNNs to classify ECoG signals as one of 57 CV syllables and phonemes (1 of 19 consonants or 1 of 3 vowels). We found that DNNs outperform a set of baseline models on all classification tasks examined here (model performance is reported in Table 1). For fully connected networks without data augmentation, a two layer network with 996 hidden units gave the best performance, achieving 38.2 ± 3.3% accuracy (22x chance, 1.7%). When trained to predict consonants or vowels independently, fully connected networks achieved a prediction rate of 47.0 ± 1.7% (9x chance, 5.0%) and 75.9 ± 2.1% (2x chance, 33%) respectively. The joint model (CV syllable) had a consonant accuracy of 39.7 ± 2.9% (8x chance, 5.0%) and vowel accuracy of 58.7 ± 2.1% (2x chance, 33%). The fact that syllable classification results in higher accuracy compared to consonant or vowel classification (relative to chance) emphasizes the cortical expression of coarticulation [2]. The CV syllable with the best decoding performance is /ha/ with 72.0 ± 16.0% (7x chance, 11%) accuracy. Of the baseline models, softmax linear regression performs best, with 26.6 ± 2.0% (16x chance, 1.7%) accuracy on syllables, while HMM-GMMs get 20.7 ± 3.1% (12x chance, 1.7%) accuracy. To make a close comparison to Mugler et. al. [4], we use LDA and perform consonant classification from the entire time series. Using LDA, Mugler et. al. report 36.1% on a 24 consonant task (5x chance, 7.4%), while LDA on our data gives 26.9 ± 1.8% on a 19 consonant task (5x chance, 5.0%).

Surprisingly, convolutional networks did not perform better than fully connected networks for any network trained (briefly discussed in Supplement: Section 5). A two layer network with one convolutional layer and a fully connected softmax output performed the best with 34.4 ± 2.4% accuracy. The number of filters, filter shape, pool shape, and pool stride were all optimized for with best values of 71, 20 (100 ms), 24, and 19 respectively.

All networks were found to overfit on the training set and almost always achieved 100% training set accuracy. To reduce overfitting, we trained fully connected and convolutional networks on augmented data. A two layer fully connected network achieved 38.7 ± 2.8% (23x change, 1.7%) accuracy, which, to our knowledge, is the state-of-the-art in speech classification from human brain activity (hyperparameters listed in Supplement: Table 3.) We additionally quantified performance by calculating the estimated information transmitted per symbol and per unit time. For the CV syllables, 38.7% classification accuracy would lead to a channel capable of transmitting 1.3 bits per syllable as reported in Table 1. Given typical syllable rates of 4 syllables per second [24], this could transmit up to 5.2 bits per second, which corresponds to 58 words per minute [25]. Figure 2(a)

Table 1: Classification and ITR Results

Model	Classification Accuracy	Folds over Chance	Approximate ITR (bits per second) Exact in parenthesis
LDA [4], 24 cons., single subject	36.1%	5x	0.75
LDA [4], 24 cons., subject average	20.4 ± 9.8%	3x	0.25
LDA, 19 consonants	26.9 ± 1.8%	5x	0.4 (1.2 exact)
Fully Connected Cons., 19 cons.	**47.0 ± 2.7%**	**9x**	**1.0 (2.1 exact)**
HMM-GMM, 57 CV	20.7 ± 3.1%	12x	0.4 (2.4 exact)
Linear Classifier, 57 CV	26.6 ± 2.0%	16x	0.75 (2.5 exact)
Fully Connected, 57 CV	38.2 ± 3.3%	22x	1.3 (3.1 exact)
Convolutional, 57 CV	34.4 ± 2.4%	20x	1.1 (2.8 exact)
Fully Connected Augmented, 57 CV	**38.7 ± 2.8%**	**23x**	**1.3 (3.1 exact)**
Convolutional Augmented, 57 CV	32.7 ± 3.2%	19x	1.0 (2.8 exact)
LDA [4], 15 vowels, single subject	23.9%	2x	0.22
LDA [4], 15 vowels, subject average	19.2 ± 3.7%	2x	0.12
Fully Connected Vowel, 3 vowels	**75.9 ± 2.1%**	**2x**	**0.6 (0.6 exact)**

compares the potential information capacity of our models. Consonant classification performs well, but vowel classification is limited due to the small number of classes. It is also clear that the approximation to channel capacity from Wolpal et al. [23] will generally underestimate the true channel capacity. Together, these results demonstrate that both fully connected DNNs have state-of-the-art performance on classifying produced speech from ECoG recordings. This strongly motivates the use of neural networks in BMIs to improve information transfer rates compared to previous methods.

3.2 Deep Networks Scale Efficiently with Limited Data

Deep networks often out perform other classification methods when the amount of training data is large. In contrast to the data sets to which deep networks are commonly applied, which typically contain $\sim 10^7$ or more examples, most data sets in neuroscience are often limited to $\sim 10^2 - 10^3$ examples. In particular, since human ECoG data is only collected in clinical situations with limited access to patients, knowing how these methods scale with increasing dataset size is important for future data collection, as well as their applicability to BMIs. To assess how classification performance of deep networks depended on the amount of available training data compared to other methods, we trained models on differently sized subsets of the data.

In Figure 2(b), we plot the classification performance for three classifiers as a function of the number of samples included in the data set. While performance improves with increasing dataset size for all models, deep networks have the best scaling. The scaling of the fully connected networks: 11.0 ± 1.2% per thousand examples, is significantly better than linear models: 6.8 ± 1.0% per thousand examples ($p < 0.01$, t-test, 10 measurements from folds). Convolutional models did not have significantly better scaling with 7.4 ± 1.2% per thousand examples. Although, the curves in Figure 2(b) appear to have similar structure between models as a function of dataset size, small dataset size and uncertainty in training and hyper-parameter search likely accounts for this. These results demonstrate that DNNS use limited data more efficiently than other methods, an important consideration when applying to neuroscience data.

3.3 Phonetic Organization of Deep Network Output

Brain computations are non-linear functions operating on spatio-temporal patterns of neural activity. However, most methods used to understand brain computation are linear, inherently limiting the capacity to extract structure from neural recordings. This is an issue not only for optimization of BMIs, but also for understanding brain computations. As DNNs are essentially adaptive bases function, non-linear function approximators, it is possible that they can extract structure from noisy, single-trial neural recordings that reveal important organization of representations.

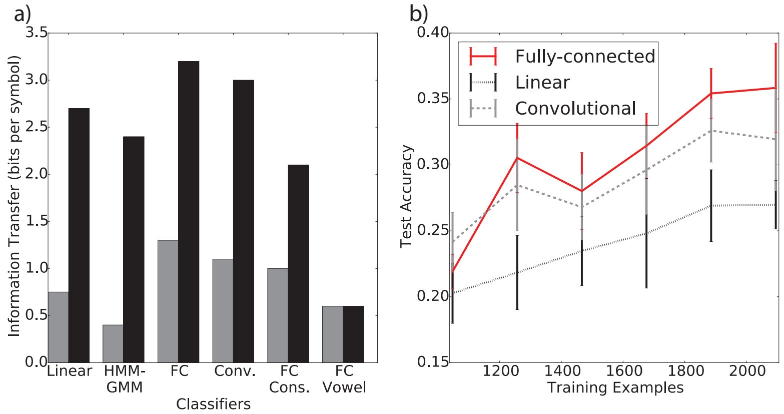

Figure 2: **a)** Comparison of estimated information transfer rate for different decoders. Grey bars are approximate information transfer per syllable. The black bars are exactly calculated, showing the possible underestimation of channel capacity from the approximation in Wolpaw et. al. [23]. **b)** Performance of different models as a function of training set size. Cross validated mean accuracy and standard deviation for different fractions of the training set used.

We examined the structure of network output to more fully understand the organization of syllable representations in vSMC. In Figure 3, we show the average confusion matrix resulting from the output of the softmax layer of the fully connected network (i.e. before binary classification), with target syllables arranged along rows and predicted syllable across columns. The syllables are ordered according to the results of agglomerative hierarchical clustering using Ward's method. To the right is a bar-plot of the mean accuracy with which a specific syllable was correctly classified. Note that the syllable with worst accuracy is the one with the smallest number of examples in the dataset. At the highest level, syllables seem to be confused only within the articulator involved (lips, back tongue, or front tongue) in the syllable. This is followed by a characterization of the place of articulation within each articulator (bilabial, labio-dental, etc.). At the lowest level there seems to be a clustering across the vowel categories that capture the general shape of the vocal tract in producing the syllable. These results are in excellent agreement with the previous analysis of mean spatial patterns of activity at separate consonant and vowel time points [1]. In contrast, we found that the structure of the HMM-GMM syllable confusion matrix had much less phonetically meaningful organization (Supplement: Figure 1). These results demonstrate the capacity of deep networks to reveal important structure in single-trial neural recordings that is not recoverable with other methods.

4 Discussion

This study is the first to use DNNs to classify produced speech from human sensorimotor cortex. Compared to other classification methods (LDA, HMM-GMM), we find that DNNs classify syllables with state-of-the-art accuracy (∼39%) and push the boundary on information transfer rates for consonants, vowels and CV classification (58 words per minute). Future studies are required to understand the factors that influence classification performance across different patients. Fully connected architectures have the highest performance on this dataset, although convolutional networks also surpass previous methods. The success of these simple architectures encourages future work in applying recurrent networks to neural time series data. This could allow classification of variable length utterances which would not need to be hand-aligned. Beyond having the best current performance, deep networks exhibit the best performance scaling as a function of data set size, and are projected to keep improving with larger neural datasets. This is important in their practical application to most biological data sets, which have small sample sizes.

Outside of their utility as classifiers, deep networks can also provide insight into the latent structure of data. Since DNNs can learn continuous mappings to probabilities over classes, their outputs can contain implicit information about the relationship between data points. In fact, when hierarchical clustering was done on the CV confusion matrix from the softmax outputs of the network, we re-

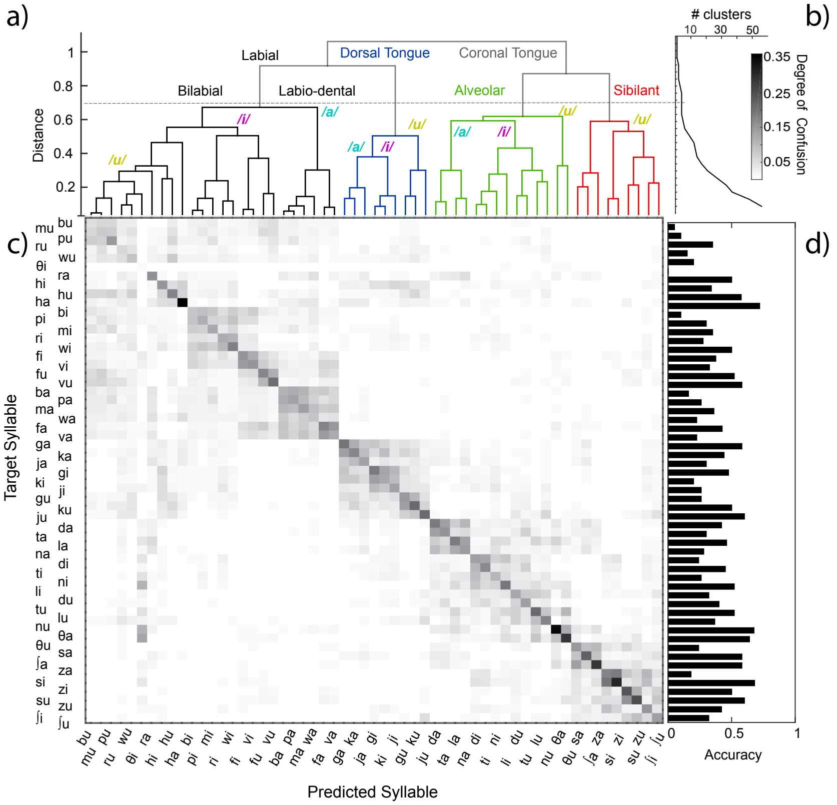

Figure 3: Analysis of the outputs of the best performing network. **a)** Dendrogram showing hierarchical clustering based on the softmax outputs. **b)** The trend of the number of discovered clusters as a function of minimum allowed distance between clusters. The dashed line represents the distance 0.71 used for the clusters identified in panel **a**, which is in a relatively flat region of the number of clusters vs. threshold relationship. **c)** Confusion matrix for network's predictions. Errors show block-diagonal structure when ordered by clustering in **a**. **d)** Mean accuracy per syllable across folds.

cover the phonetic organization of the spoken syllables, which is directly related to the articulators and shape of the vocal tract during speech production. This highlights the close relationship between neural signals in the vSMC and articulators that has also been seen in previous studies [1]. However, we note that the previous study applied dimensionality reduction to trial-averaged data at hand-picked time points. Thus, the ability to reveal this structure from single-trials across all times is noteworthy, as trial-averaging is not possible in the BMI context. While the performance of deep networks on neural data is indeed encouraging, it falls short of the performance of similar networks on acoustic data (∼90% classification accuracy on the acoustic data from this dataset, see Supplement: Section 4). This difference can be reduced by improved understanding and continued investigation into robust feature representations of ECoG data as related to speech production. For example, we found the highest (relative to chance) decoding performance was for the entire consonant-vowel syllable, which agrees with previous descriptions of context dependent phoneme representations (i.e. coarticulation) in vSMC activity [2]. Again, deep networks can enable this by revealing the latent structure of neural data as shown in this study (Figure 3). The ability to extract meaningful structure from single-trial data is important for understanding the nature of brain computations and suggests that DNNs may be useful for this purpose in general.

References

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

[2] K. E. Bouchard and E. F. Chang. Control of Spoken Vowel Acoustics and the Influence of Phonetic

Context in Human Speech Sensorimotor Cortex. *Journal of Neuroscience*, 34(38):12662–12677, 2014.

[3] K. E. Bouchard and E. F. Chang. Neural decoding of spoken vowels from human sensory-motor cortex with high-density electrocorticography. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 6782–6785. IEEE, 2014.

[4] E. M. Mugler, J. L. Patton, R. D. Flint, Z. A Wright, S. U. Schuele, J. Rosenow, J. J. Shih, D. J. Krusienski, and M. W. Slutzky. Direct classification of all American English phonemes using signals from functional speech motor cortex. *Journal of neural engineering*, 11:035015, 2014.

[5] F. Lotte, J. S. Brumberg, P. Brunner, A. Gunduz, A. L. Ritaccio, C. Guan, and G. Schalk. Electrocorticographic representations of segmental features in continuous speech. *Frontiers in Human Neuroscience*, 09(February):1–13, 2015.

[6] W. Penfield and E. Boldrey. Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain: A journal of neurology*, 1937.

[7] S. Kellis, K. Miller, K. Thomson, R. Brown, P. House, and B. Greger. Decoding spoken words using local field potentials recorded from the cortical surface. *Journal of neural engineering*, 7:056007, 2010.

[8] X. Pei, D. L. Barbour, E. C. Leuthardt, and G. Schalk. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal of neural engineering*, 8:046028, 2011.

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

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.

[11] A. Graves, A. R. Mohamed, and G. E. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6645–6649. IEEE, 2013.

[12] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural computation*, 1997.

[13] S. Stober, D. J. Cameron, and J. A. Grahn. Using Convolutional Neural Networks to Recognize Rhythm Stimuli from Electroencephalography Recordings. *Neural Information Processing Systems*, pages 1–9, 2014.

[14] M. Wand and T. Schultz. Pattern Learning with Deep Neural Networks in EMG-based Speech Recognition. *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, 35:4200–4203, 2014.

[15] A. Supratak, L. Li, and Y. Guo. Feature Extraction with Stacked Autoencoders for Epileptic Seizure Detection. *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 4184–4187, 2014.

[16] D. F. Wulsin, J. R. Gupta, R. Mani, J. A. Blanco, and B. Litt. Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement. *Journal of neural engineering*, 8(3):036015, 2011.

[17] M. Yang, S. A. Sheth, C. A. Schevon, G. M. McKhann II, and N. Mesgarani. Speech reconstruction from human auditory cortex with deep neural networks. In *Telluride Neuromorphic Cognition Engineering Workshop*, 2015.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, pages 1–9, 2012.

[20] J. Snoek, H. Larochelle, and R. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. *Neural Information Processing Systems*, pages 1–9, 2012.

[21] I. J. Goodfellow, D. Warde-Farley, P. Lamblin, V. Dumoulin, M. Mirza, R. Pascanu, J. Bergstra, F. Bastien, and Y. Bengio. Pylearn2: a machine learning research library. *arXiv preprint arXiv:1308.4214*, 2013.

[22] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK, 2006.

[23] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain-computer interfaces for communication and control. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 113(6):767–91, 2002.

[24] A. Cruttenden. *Gimson's pronunciation of English*. Routledge, 2014.

[25] C. M. Reed and N. I. Durlach. Note on Information Transfer Rates in Human Communication. *Presence: Teleoperators and Virtual Environments*, 7:509–518, 1998.