Brains in the Wild: Machine learning for naturalistic, long-term neural and video recordings

Nancy (Xin Ru) Wang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Rajesh P.N. Rao, Chair

Bingni W. Brunton, Chair

Ali Farhadi

Program Authorized to Offer Degree:
Computer Science and Engineering

University of Washington

Abstract

Brains in the Wild: Machine learning for naturalistic, long-term neural and video recordings

Nancy (Xin Ru) Wang

Co-Chairs of the Supervisory Committee:
Professor Rajesh P.N. Rao
Computer Science and Engineering

Assistant Professor Bingni W. Brunton
Biology

Developing useful interfaces between brains and machines is a grand challenge of neuroengineering. An effective interface has the capacity to not only interpret neural signals, but it needs to be practical, robust and work without delay in real-world applications. These applications can range from controlling cursors computers to robotic arms for paralyzed patients, all with just a thought. A real world brain computer interface (BCI) needs to be effective outside of well-controlled laboratory experiments. In this thesis. I detail our novel approach to studying long-term naturalistic electrocorticography(ECoG) for behaviour decoding and prediction, as opposed to short experimental data. We develop new machine learning techniques in order to both automatically learn behavioural annotations from the simultaneously recorded video as well as to analyze the neural correlates of behaviour from the ECoG data itself. Firstly, I outline our unsupervised approach to decode high level categories of natural behaviour and functionally map associated areas in each individual brain. My next project centers around applying deep learning in a multimodal fashion to predict spontaneous human arm movements in the future. Finally, I solve the subject transfer learning problem in

ECoG in order to combine data from multiple subjects despite heterogeneity in the electrode input space and apply trained models on unseen patients. Altogether, I show through these projects that a multimodal and multi-subject approach to neural decoding “in the wild” is critical in advancing bioelectronic technologies and human neuroscience.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to University of Washington, where I have had the opportunity to work with amazing professors and students alike. In particular, I would like to thank Raj for seeing the potential in naturalistic ECoG and supporting me and my project ever since. Also, I would like to thank Ali for all of his advice regarding computer vision and deep learning and Jeff for his advice relating to patients and neurosurgery, as well as allowing us to consent patients for our studies. Finally, I would like to especially thank Bing for taking me on as a student and growing the naturalistic ECoG project into the multi-student, multi-grant project it is today. She has worked tirelessly to support me and the project by writing grant applications, listening patiently to (sometimes crazy) ideas and working through each project and paper together with me. I will never forget the weekend writing sessions nor the lunches where I learned ”the sausage-making” of academia. I would also like to thank all of my colleagues from the Brunton, Rao and Ojemann labs for sharing brain scans and the ups and downs of graduate school alike. As well, the undergraduate students who I have mentored over the years continue to remind me that a university is first and foremost a place of education. It has been a pleasure working, teaching and learning from all of them. Finally, I would like to thank the staff at Harborview Medical Center who have been so helpful in the patient consent and data gathering process.

DEDICATION

To my family. Thank you mom and dad for raising me and bringing me to Canada when I was young. The educational and work opportunies available to me now are all from your sacrifice in uprooting your home and career and starting again in a new country. As well, without both of your examples in technical careers, I may not have been interested in computer science and I probably would not have been in a computer science PhD program. I hope that my younger brother and sister will be able to find a happy and fulfilling career path as I have. I also dedicate this thesis to my husband, dog, and daughter. Thank you Jason for moving down to Seattle with me despite your patriotism to our home country, Canada. Thank you for listening to all my complaints and celebrating the successes. Thank you for taking care of me and our family whenever the situation arose, whether it be the primary breadwinner or the primary child taker. Thank you Barkspawn for always being there for me even when I have not been able to pay as much attention to you as I should have been. Finally, to Luna, thank you for being the ”most chill baby ever”. I hope that mom will be able to help you achieve whatever it might be that will make you happy.

Chapter 1

INTRODUCTION

Despite data being ubiquitous in our world today, most scientific work in neural computation have hinged on small sets of carefully curated experimental data. When sources of behavioural and recording variation are actively minimized, it is uncertain how findings or developed techniques would apply to natural behaviour "in the wild". In particular, devices interfacing brains and computers (BCI) need to be robust to changes in the environment and adapt to individual users. If successful, these technologies have great potential in helping individuals with physical and neurological disabilities. As well, BCIs can enhance the daily lives of the general population by augmenting our ability to interact with the world. For instance, one may be able to make a phone call and communicate without moving or making any verbal commands.

In order to tackle neural decoding "in the wild", my research focused on a large long-term multimodal electrocorticography (ECoG) dataset composed of simultaneously recorded video, audio and ECoG. The patients in the dataset have intractable epilepsy and have ECoG electrodes implanted for around a week in order to characterize their seizures. While they are undergoing monitoring, the patients are free to move on the hospital bed, talk with friends and family, eat and interact with electronic devices like laptops and cell phones. My first study centered around unsupervised clustering. I was able to decode human behaviour from ECoG completely automatically without any

manual labels by leveraging the simultaneously recorded video and audio recordings. Next, I was able to use the latest pose recognition deep learning models to extract joint coordinates from the video in order to analyze patient movement. I developed a multimodal deep learning approach that was able to predict whether a subject will initiate a spontaneous movement in the future, a task never before attempted with natural ECoG. Finally, I worked on combining training data across subjects using a virtual grid and deep imputation for missing data in order to train a model that can leverage the power of a larger dataset as well as be able to work without retraining on an unseen patient. Overall, my work combines techniques from the rapidly advancing fields of machine learning, computer vision and speech recognition in order to automatically decode long-term naturalistic human neural recordings.

Below, I introduce fields that are related to my research before outlining my projects in detail. In the conclusion, I provide my vision of what I think the convergence of advancements in machine learning and neural recording will mean for the future of long-term neural decoding and BCI "in the wild".

1.1 Brain Computer Interface (BCI) with Electrocorticography (ECoG)

BCI's are technologies that enable one to directly use neural signals to control some outside interface. This interface can be anywhere from a cursor on a screen, letters on a keyboard to a physical robotic arm. They are in active research as a potential answer to people with disabilities or paralysis in order to regain mobility and communicative functions. In future, they may also be used for human augmentation such that an "able-bodied" person may control robotic components or communication devices directly using their brain, as is currently featured in science fiction video games like Deus

Ex.

The field of BCI is reaching ever closer to real time control of physical devices directly from neural signals in humans. Although local field potential devices such as the Utah array has been shown to be effective in controlling devices in humans, the array implantation is a very invasive procedure performed in very restrictive cases. Alternatively, ECoG, or otherwise known as intracranial EEG, is an array of non-penetrating electrodes implanted underneath the skull. They can be epidural (above the dura) or subdural (below the dura). In most human cases, they are implanted for clinical monitoring of intractable epilepsy, in preparation for ablation surgery. ECoG has been shown to be stable over a long time in monkeys [1], with minimal signal degradation and tissue injury. Thus, ECoG is much more likely to be used by the wider population. In [2], the authors were able to demonstrate for the first time real time online control of a prosthetic hand using high-density ECoG in one patient. They were able to achieve very high accuracy for the detection of the initiation of movement and had moderate performance in classifying which finger is moving. They used gamma power for the classification without prior training in their online decoding. The decoding also seemed to be relying on sensory feedback signals. In offline analysis, they were able to achieve a much higher classification accuracy by using the optimal set of electrodes for decoding.

1.2 Deep Learning with Neural Data

There have been many landmark and breakthrough studies in vision related tasks using deep neural networks in recent years. To the best of my knowledge, there are very few recent work, with the notable exception from Volker et al. [3], tackling ECoG neural decoding with deep learning. This may be due to the limited availability of this type of data. A related signal, electroencephalography

(EEG), is a non-invasive technique that records neural signals outside the skull. The signal-to-noise ratio is much worse than ECoG but the data is much easier to obtain, enabling some studies with deep learning.

For example, Schirrmeister et al. conducted a deep investigation into the optimal deep CNN architecture and training scheme for classification tasks with EEG in an end-to-end training scheme with raw signals ([4]). Overall, they found that convolutional neural network (CNN) marginally outperforms filter bank common spatial patterns (FBCSP), the traditional spectral power modulation based analysis method. Nevertheless, their visualization of learned CNN features showed that the network was indeed learning to use spectral power modulations in the alpha, beta, and high gamma frequencies. Their in-depth study resulted in many findings. Most interestingly, they found that CNN design choices, such as batch normalization and dropout, significantly increased accuracies. As well, they did not find a significant difference in accuracy between shallow and deep neural nets. The two task-based data sets that they performed the experiments on had around 880 and 228 training examples per subject for a 4 way classification task, which is still quite small when compared to the millions of examples in some vision tasks. Since their training set is quite small, this may be why the depth of the architecture does not significantly change the final accuracy. There may not be enough data to learn interesting higher order features to necessitate more layers.

In the study by Tabar et al., the authors were able to improve classification scores of the BCI competition IV dataset 2b by applying a CNN and stacked autoencoder network to EEG data that had been preprocessed with Fourier transformations ([5]). In a study that combines hardware and software improvements, the authors of [6] aimed to demonstrate a proof-of-concept neural network

that can decode EEG activity in real time with the neuromorphic chip, TrueNorth, which uses a small amount of power. They applied a CNN to to a EEG dataset from a hand-squeeze task to classify left or right hand squeezes. They applied the CNN to raw data but did not attempt to improve the architecture or training scheme to achieve higher accuracy. In fact, their final accuracy of 76% was 10% lower than what has been achieved in the past. Despite this, the authors referred to their results as ”state of the art”. Further, the authors did not actually deploy the network in a TrueNorth chip. Nevertheless, it is important to explore architectures that may be deployed in hardware that can be implantable, requiring a low power system.

Chapter 2

DATA

As part of standard clinical care, many hospitals have continuous video monitoring of patients, along with any additional monitoring with medically necessary sensors. Oftentimes, these videos are deleted after the patient's stay as there is no clinical need nor the storage space to keep them. However, with the current advancements in computer vision, it may be possible to study these videos, in conjunction with simultaneously recorded clinical sensors, for better patient care and answer basic scientific questions. With this goal in mind, we were able to obtain approval from the University of Washington Institutional Review Board (IRB) to download video, ECoG and audio data before their deletion from the clinical monitoring of epilepsy patients at Harborview Medical Center. The patients were given informed consent forms to consider participation in the study, which required no active involvement on their part as all data collected were already part of the clinical monitoring process.

Our long-term, naturalistic human movement dataset includes week-long continuous multimodal recordings with invasive (ECoG) electrodes, video, and audio. This opportunistic dataset greatly surpasses all previously analyzed comparable datasets in duration and size. In fact, very few ECoG datasets have been released and all are short (minutes to hours) task-based recordings. Over the course of my PhD, 45 patient data have been collected. This amounts to approximately 7500 hours

of recording, 810 million frames of video (30 frames per second) and more than 20 billion samples of ECoG (1000 Hz). The video was recorded in colour at resolutions of 640X480 for daytime shots and grayscale using the infrared camera for nighttime recordings. The camera was situated in the ceiling with a direct shot of the patient's bed. The camera is controlled at a nurse's station and can be rotated or zoomed in but is generally kept still during the patient's stay.

The patients' implants had various brain coverage that more or less together covered the entire brain. However, the amount and sites of coverage between patients varied greatly. Some patients only had a few strips of electrodes while others also had depth electrodes that can monitor subcortical structures. Because the data was recorded directly from the surface of the brain, the number of motor artifacts was minimal as compared to extra-cranial EEG. Simple frequency based data filtering techniques were employed (detailed within each project) .

During monitoring, these subjects performed no instructed tasks; instead, they simply did as they wished for the duration of their monitoring in the hospital room. The variety of natural behaviors observed was rich and complex, including conversations with friends and family, eating, interacting with electronic devices, and sleeping. The dataset is rich with a large variety of natural behaviour. In fact, some patients are seen arguing with spouses or playing card games with friends. Datasets of this scale are rare in general likely due to the immense task of annotating such long recordings. Additionally, private medical data cannot be shared easily and so public crowd sourcing methods are not possible. This is why our approach centers on using automated techniques to annotate behaviour from videos.

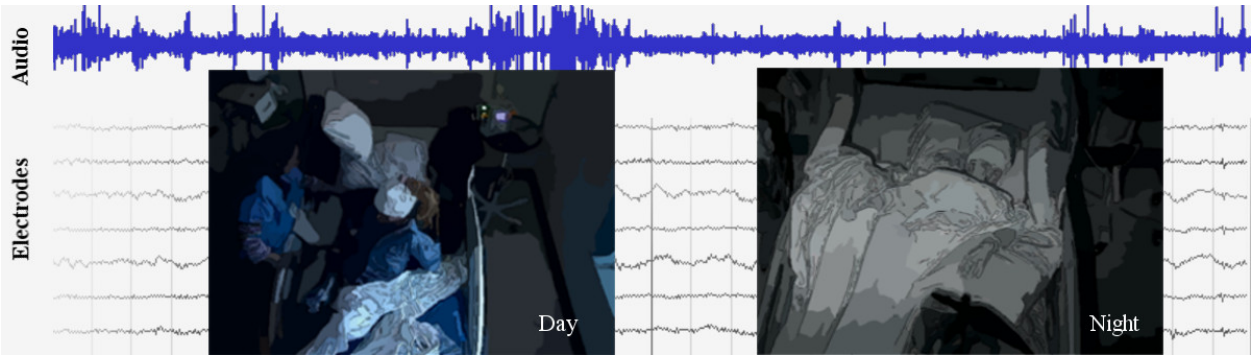


Figure 2.1: An excerpt from the data set, which includes video, audio, and intracranial brain activity (ECoG) continuously recorded for at least one week for six subjects. ECoG recordings from a small subset of the electrodes are shown, along with the simultaneously recorded audio signals in blue. A typical patient has around 100 electrodes. Overlaid are screen shots of the video, which is centered on the patient; on the left is a daytime video of the patient eating, and on right is a nighttime infrared video of the patient sleeping. For patient privacy, faces have been blurred.

Chapter 3

UNSUPERVISED DECODING OF LONG-TERM, NATURALISTIC HUMAN NEURAL RECORDINGS WITH AUTOMATED VIDEO AND AUDIO ANNOTATIONS

Nancy X. R. Wang, Jared D. Olson, Jeffrey G. Ojemann, Rajesh P. N. Rao, Bingni W. Brunton

Sourced from [7]

Fully automated decoding of human activities and intentions from direct neural recordings is a tantalizing challenge in brain-computer interfacing. Most ongoing efforts have focused on training decoders on specific, stereotyped tasks in laboratory settings. Implementing brain-computer interfaces (BCIs) in natural settings requires adaptive strategies and scalable algorithms that require minimal supervision. Here we propose an unsupervised approach to decoding neural states from human brain recordings acquired in a naturalistic context. We demonstrate our approach on continuous long-term electrocorticographic (ECoG) data recorded over many days from the brain surface of subjects in a hospital room, with simultaneous audio and video recordings. We first discovered clusters in high-dimensional ECoG recordings and then annotated coherent clusters using speech and movement labels extracted automatically from audio and video recordings. To our knowledge, this represents the first time techniques from computer vision and speech processing have been used for natural ECoG decoding. Our results show that our unsupervised approach can discover distinct behaviors from ECoG data, including moving, speaking and resting. We verify the accuracy of our

approach by comparing to manual annotations. Projecting the discovered cluster centers back onto the brain, this technique opens the door to automated functional brain mapping in natural settings.

Introduction

Much of our knowledge about neural computation in humans has been informed by data collected through carefully controlled experiments in laboratory conditions. Likewise, the success of Brain-Computer Interfaces (BCIs [8], [9])—controlling robotic prostheses and computer software via brain signals—has hinged on the availability of labeled data collected in controlled conditions. Sources of behavioral and recording variations are actively avoided or minimized. However, it remains unclear to what extent these results generalize to naturalistic behavior. It is known that neuronal responses may differ between experimental and freely behaving natural conditions [10]–[12]. Therefore, developing robust decoding algorithms that can cope with the challenges of naturalistic behavior is critical to deploying BCIs in real-life applications.

One strategy for decoding naturalistic brain data is to leverage external monitoring of behavior and the environment for interpreting neural activity. Previous research that studied naturalistic human brain recordings, including brain surface electrocorticography (ECoG), have required ground truth labels [13]–[15]. These labels were acquired by tedious and time-consuming manual labeling of video and audio. In addition to being laborious, manual labeling is prone to human errors from factors such as loss of attention and fatigue [16]. This problem is exacerbated by very long recordings, when patients are monitored continuously for several days or long. Obtaining labeled data and training algorithms extensively are difficult or even intractable in rapidly changing, naturalistic environments.

In this article, we describe our use of video and audio recordings in conjunction with ECoG data to decode human behaviour in a completely unsupervised manner. Fig. 2.1 illustrates components of the data used in our approach. The data consists of six subjects monitored continuously over at least one week after electrode array implantation; each subject had approximately 100 intracranial ECoG electrodes with wide coverage of cortical areas. Importantly, subjects being monitored had no instructions to perform specific tasks; they were undergoing presurgical epilepsy monitoring and behaved as they wished inside their hospital room. Instead of relying on manual labels, we used computer vision, speech processing, and machine learning techniques to automatically determine the ground truth labels for the subjects' activities. These labels were used to annotate patterns of neural activity discovered using unsupervised clustering on power spectral features of the ECoG data. We demonstrate that this approach can identify salient behavioral categories in the ECoG data, such as movement, speech and rest. Decoding accuracy was verified by comparing the automatically discovered labels against manual labels of behavior in a small subset of the data. Further, projecting the annotated ECoG clusters to electrodes on the brain revealed spatial and power spectral patterns of cortical activation consistent with those characterized during controlled experiments. These results suggest that our unsupervised approach may offer a reliable and scalable way to map functional brain areas in natural settings and enable the deployment of BCI in real-life applications.

Background and Related Work

Intracranial electrocorticography (ECoG) as a technique for observing human neural activity is particularly attractive. Its spatial and temporal resolution offers measurements of temporal dynamics inaccessible by functional magnetic resonance imaging (fMRI) and spatial resolution unavailable to

extracranial electroencephalography (EEG). Cortical surface ECoG is accomplished less invasively than with penetrating electrodes [17], [18] and has much greater signal-to-noise ratio than entirely non-invasive techniques such as EEG **Lal2005**, [19].

Efforts to decode neural activity are typically accomplished by training algorithms on tightly controlled experimental data with repeated trials. Much progress has been made to decode arm trajectories [20]–[22] and finger movements [23], [24], to control robotic arms [25]–[27], and to construct ECoG BCIs [28]–[33]. Speech detection and decoding from ECoG has been studied at the level of voice activity [34], phoneme [29], [35]–[37], vowels and consonants [38], whole words [39], and sentences [40]. Accurate speech reconstruction has also been shown to be possible [41].

The concept of decoding naturalistic brain recordings is related to passive BCIs, a term used to describe BCI systems that decode arbitrary brain activity that are not necessarily under volitional control [42]. Our system, which falls within the class of passive BCIs, may also be considered a type of hybrid BCI combining electrophysiological recordings with other signals [43]. However, past approaches in this domain have not focused on combining alternative monitoring modalities such as video and audio in order to decode natural ECoG signals.

The lack of ground-truth data makes decoding naturalistic neural recordings difficult. Supplementing neural recordings with additional modes of observation, such as video and audio, can make the decoding more feasible. Previous studies exploring this idea have decoded natural speech [13], [44], [45] and natural motions of grasping [14], [15]; however, these studies relied on laborious manual annotations. Entirely unsupervised approaches to decoding have previously targeted sleep stages [46] and seizures [47] rather than long-term natural ECoG recordings.

Our approach to circumvent the need for manually annotated behavioral labels exploits automated techniques developed in computer vision and speech processing. Both of these fields have seen tremendous growth in recent years with increasing processor power and advances in methodology [48], [49]. Computer vision techniques have been developed for a variety of tasks including automated movement estimation [50], [51], pose recognition [52], object recognition [53], [54], and activity classification [55], [56]. In some cases, computer vision techniques have matched or surpassed single-human performance in recognizing arbitrary objects [57]. Voice activity detection has been well studied in speech processing [58]. In this work, we leverage and combine techniques from these rapidly advancing fields to automate and enhance the decoding of naturalistic human neural recordings.

Results

Our general approach to unsupervised decoding of large, long-term human neural recordings is to combine hierarchical clustering of high-dimensional ECoG data with annotations informed by automated video and audio analysis, as illustrated in Fig. 3.1 (further details in the Methods section). Briefly, hierarchical k-means clustering was performed on power spectral features of the ECoG recordings. These clusters are coherent patterns discovered in the neural recordings; video and audio monitoring data was used to interpret these patterns and match them to behaviorally salient categories such as movement, speech and rest. Here we describe results of our analysis on six subjects where we used automated audio and video analysis to annotate clusters of neural activity. The accuracy of the unsupervised decoding method was quantified by comparison to manual labels, and the annotated clusters were mapped back to the brain to enable neurologically relevant

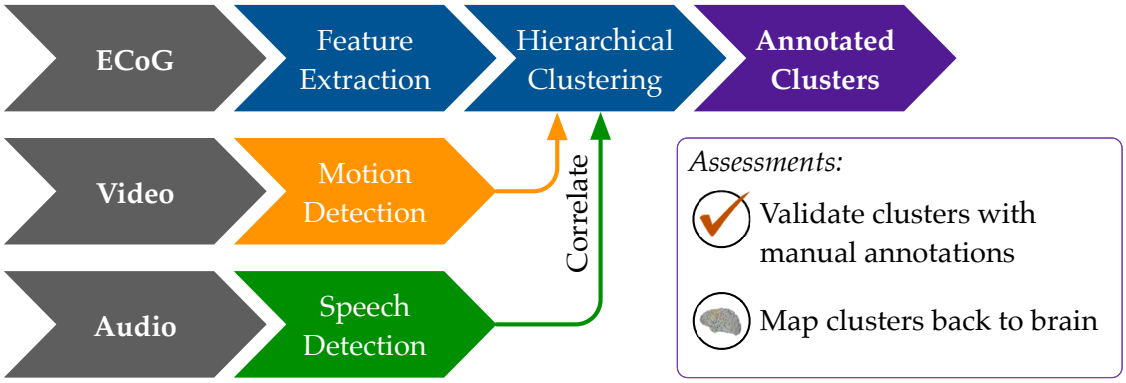


Figure 3.1: An overview of our methods to discover neural decoders by automated clustering and cluster annotations. Briefly, the ECoG recordings was broken into short, non-overlapping windows of 2-seconds. Power spectral features were extracted for each electrode, all electrodes' features were stacked, and the feature space was reduced to the first 50 principal component dimensions. Hierarchical k-means clustering was performed on these 50-dimensional data, and annotation was done by correlations in timing with automated detection of motion and speech levels (see Fig. 3.6). The resultant annotated clusters were validated against manual annotations; cluster centroids mapped to the brain visualize the automatically detected neural patterns.

interpretations.

Actograms: automated motion and speech detection

The automated motion and speech detection methods quantified movement and speech levels from the video and audio recordings, respectively. Fig. 3.2 shows daily "actograms" for all six subjects. Movement levels were quantified by analyzing magnitude of changes at feature points in successive frames of the video. Speech levels were quantified by computing the power in the audio signal in the human speech range.

As expected, Fig. 3.2 shows that subjects were most active during waking hours, generally between 8:00AM and 11:00PM. Also, movement and speech levels are often highly correlated, as the subjects were often moving and speaking at the same time during waking hours. During night

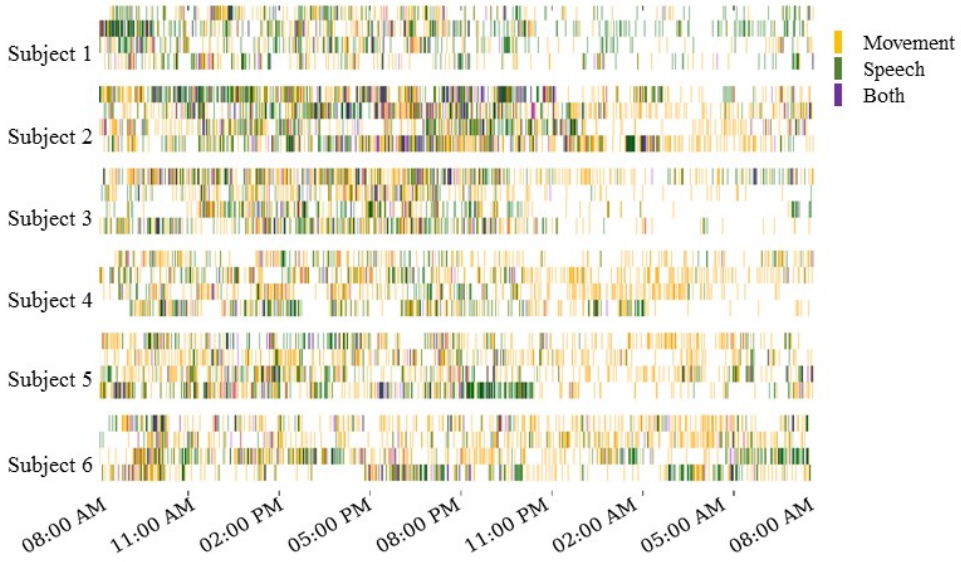


Figure 3.2: Daily actograms for all subjects. Each row shows one day of activity profiles summarized by automated speech and motion recognition algorithms. Days 3–6 post surgical implantation were analyzed. For purposes of this visualization, the activity levels were binned to one-minute resolution. Movement and speech levels were highly correlated on most days, and these were concentrated to the active hours between 8:00 AM and 11:00PM.

time hours, although subjects were generally less active, many instances of movement and speech can still be seen in Fig. 3.2 as the subjects either shifted in their sleep or were visited by hospital staff during the night.

Our automated motion and speech detection algorithms were able to perform with reasonable accuracy when compared to manual annotation of movement and speech. Over all subject days, movement detection was 74% accurate (range of 68% to 90%), while speech detection was 75% accurate (range of 67% to 83%).

Unsupervised decoding of ECoG activity

Unsupervised decoding of neural recordings was performed by hierarchical clustering of power spectral features of the multi-electrode ECoG recordings. Because the subjects' behavior on each day varied widely, both across days for the same subject and across subjects (see actograms in Fig. 3.2), we were agnostic to which specific frequency ranges contained meaningful information and considered all power in frequency bins between 1 and 53 Hz (see Supplemental Information for results considering higher frequency bands). Further, data from each subject day was analyzed separately. Clusters identified by hierarchical k-means clustering were annotated using information from the external monitoring by video and audio. The hierarchical k-means clustering implementation is detailed in the Methods section. Following a tree structure, successive levels of clustering contained larger numbers of clusters (Fig. 3.6).

Fig. 3.3 shows results of the annotated clusters for one subject day (Subject 6 on day 6 post implantation) at clustering levels 1–4 as a function of time of day. At level 1, it is clear that rest is separable from non-rest, and the switch in the dominant cluster occurred around 10:00PM. We presume the timing of the switch to correspond to when the subject falls asleep, as is corroborated by the video monitoring. when the subject is presumed to have fallen asleep as evident in the video monitoring. Video S1 shows an example of the infrared video acquired during night time. The subject is in a consolidated period of rest between 10:00PM and 9:00AM the following day. Interestingly, for a duration of approximately one hour starting at around 11:00AM, the rest cluster dominated the labels (see also red triangle at level 3). This period corresponds to the subject taking

a nap (Video S2).

Starting at level 2, the non-rest behavior separates into movement and speech clusters. These two clusters are generally highly correlated, as moving and talking often co-occur, especially as the movement quantification can detect mouth or face movement. We point out several interesting instances labeled at level 3. First, the inverted triangle points to a period around 11:00AM annotated as rest, when the subject rested during a nap (Video S2). Second, the rectangle marks a period around 1:00PM, annotated as predominantly movement but not speech, when the subject shifted around in their bed but did not engage in conversation (Video S3). Third, the circle marks a period around 5:00PM when the subject engaged in conversation (Video S4); this period was labeled as both movement and speech. As described in the validation analysis in the following section, the accuracy of the automated annotations does not change substantially between levels 3–4 across all subject days (Fig. 3.4 and Fig. S3).

Validation of automated neural decoding by comparison with manual annotations

The automated neural decoding was assessed by comparison with behaviors labeled manually. Manual labels of the video and audio were supplied by two human annotators, who labeled a variety of salient behaviors for at least 40 total minutes (or approximately 3%) of video and audio recordings for each subject day. The labels were acquired for 2-minute segments of data distributed randomly throughout the 24-hour day.

The entirely automated neural decoding performed very well in the validation for all subjects on the categories of movement, speech and rest. Table 3.1 summarizes the accuracy of the annotated clusters averaged over the 4 days analyzed for each subject, comparing the automated labels to

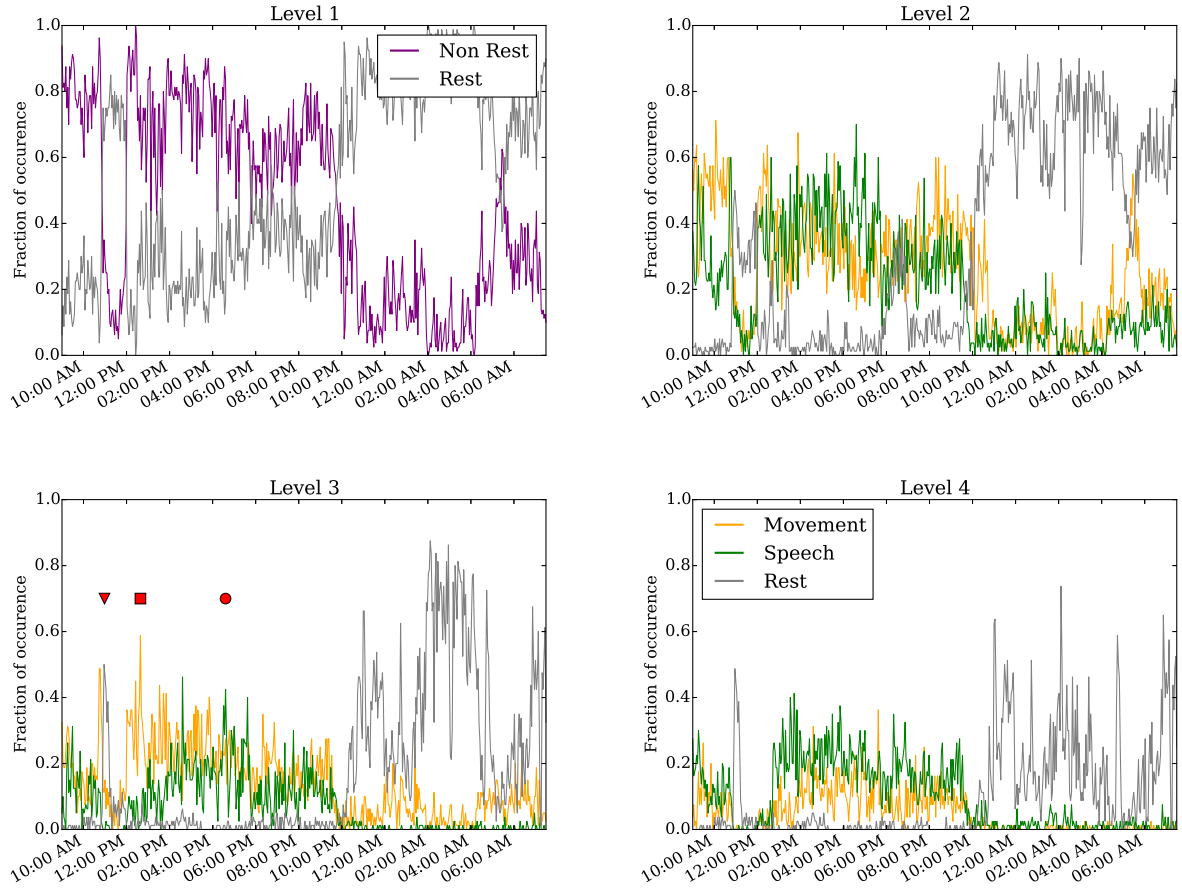


Figure 3.3: Annotated clustering results of one subject day (Subject 6 on day 6 post implant) from hierarchical level 1 to level 4. The vertical axis represents the fraction of time the neural recording is categorized to each annotated cluster. The triangle marks when the subject takes a nap (Video S2), the square marks when the subject is seen to move without speech (Video S3), and the circle marks when the subject spoke more than moved (Video S4). For visualization, the 24-hour day was binned to every 160 seconds.

manual labels during the labeled portions of each day. In addition to computing the accuracy, we also computed the F1 scores of the automated decoding using manual labels as ground truth for each day; the F1 score is a weighted average of precision and recall (Table S1).

To assess the significance of the automated labels' accuracy, we compared the F1 scores on each

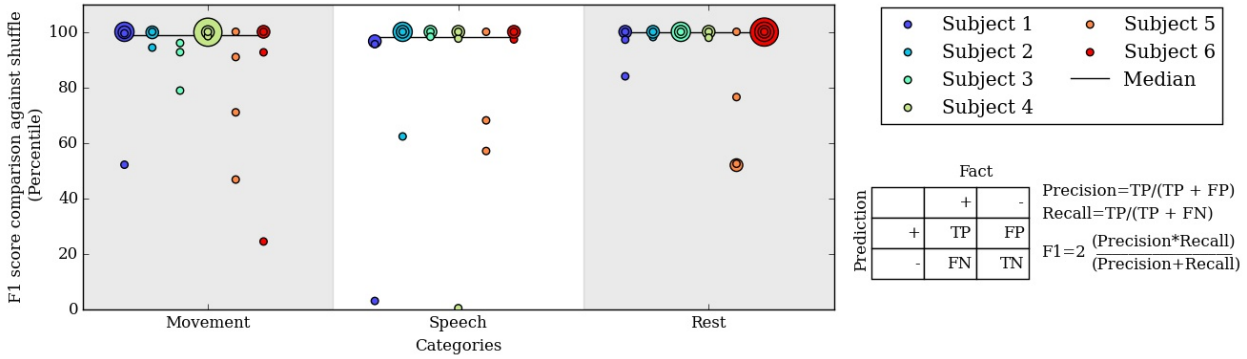


Figure 3.4: Percentile of the F1 score of our algorithm at level 3 compared to F1 scores from randomly shuffled manual labels. Each colored dot corresponds to one day for a subject in each behavioral category.

day to F1 scores of randomly shuffled labels. The shuffled labels preserved the relative occurrence of labels and gave an unbiased estimate of chance performance. Fig. 3.4 shows the percentile of the true F1 scores within the randomly shuffled F1 scores at hierarchical clustering level 3. For each category of movement, speech and rest, the median percentile of the true F1 scores are at or near the 99th percentile; our automatically labeled clusters performed significantly better than chance on most subject days. F1 score percentiles for clustering levels 2 and 4 are shown in Figs S2 and S3. We also repeated the analysis considering spectral frequencies up to 105 Hz, which does not substantially change the performance of the automated decoder (Table S2 and Fig. S4)

Neural correlates of behavior as discovered by unsupervised clustering

Another way to assess the neural decoder discovered through clustering and automated annotation is to examine the neural patterns identified in this unsupervised approach. We mapped these patterns by projecting the centroids of annotated clusters back to feature space. Next, the feature space in electrode coordinates on the brain were averaged within frequency bands, including those typically

	Movement	Speech	Rest
Subject 1	59.06	54.24	64.98
Subject 2	62.88	62.20	64.47
Subject 3	61.22	62.94	64.75
Subject 4	58.57	60.43	65.88
Subject 5	55.61	60.01	58.09
Subject 6	70.08	69.43	57.60

Table 3.1: Percent accuracy as assessed by comparison of level 3 automated cluster annotation to manual annotations averaged over all 4 days for each subject.

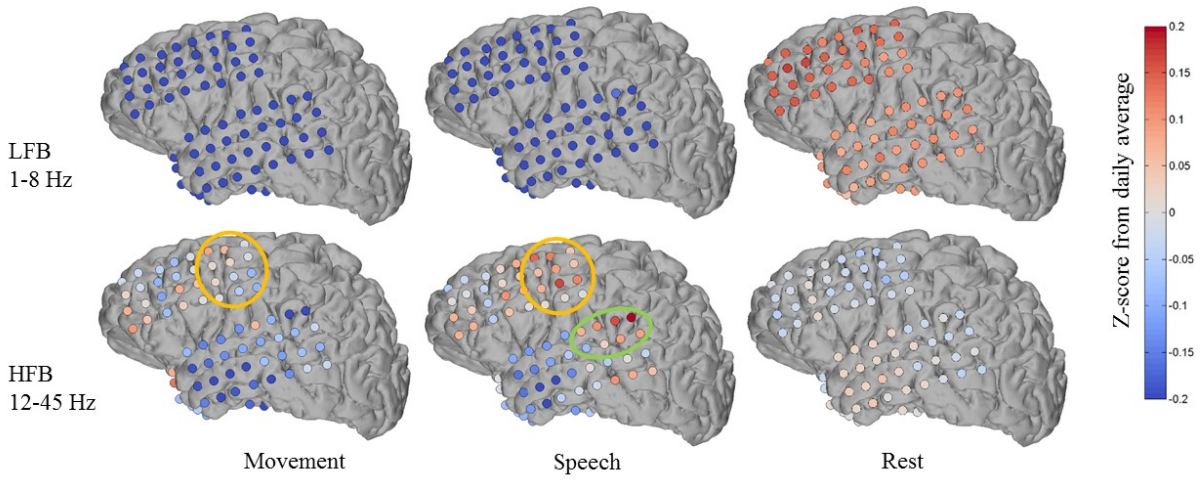


Figure 3.5: Features discovered by automated brain decoding at two different frequency bands are consistent with known functions of cortical areas. Shown for one subject day (Subject 1 on day 6 post implant), the centroids of the movement, speech and rest clusters were back projected to brain-electrode space, and then separately averaged over a low frequency band (LFB, 1–8Hz) and a high frequency band (HFB, 12–45Hz). The orange and green circles mark the approximate extent of locations typically considered to be sensorimotor and auditory regions, respectively. The colormap indicates the Z-Score of the power levels as compared to the daily average.

of interest to studies of human ECoG.

Fig. 3.5 shows an example of one subject day's annotated cluster centroids shown as deviations from the daily average in a low frequency band (LFB, 1–8Hz) and a high frequency band (HFB, 12–45Hz). The LFB was chosen to include activity in the delta and theta range, while the HFB includes beta and low gamma activity. The accuracy of automated decoding on this subject day (Subject 1, day 6 post implant) was 0.56, 0.69 and 0.63 for movement, speech and rest, respectively. In the LFB, there was generalized decrease in power across all recording electrodes during movement and speech, accompanied by a corresponding relative increase in power during rest. In contrast, in the HFB during movement and speech, we observe more spatially specific increase in power that is localized to motor areas (orange circle in Fig. 3.5). There is some overlap in electrodes showing increased HFB power during movement and speech, which may be due to activation of motor areas to produce speech. In addition, during speech but not during movement, there is a localized increase in HFB power at associated auditory region (green circle in Fig. 3.5).

These features are largely consistent with known functions of human cortical areas and ECoG phenomena, as well as the existing ECoG literature on motor activation [23], [59] and speech mapping [60], [61]. We must note that these patterns of frequency band-specific changes in power for different behavioral categories were discovered in an entirely unsupervised approach, using continuously acquired naturalistic data, and without the luxury of subtraction of baseline activation immediately before or after movement. It is important to keep in mind that previous studies typically define rest as the time just before an action, whereas we compare to daily averages as well as to sleep. During non rapid eye movement sleep, the theta and delta bands tend to have high power [62],

a factor that distinguishes our results from those obtained from more controlled experiments. We observed qualitatively similar patterns across the four (4) subjects where anatomic reconstruction of the electrode arrays were available (Figs. S6–S8).

Discussion

Our results represent, to our knowledge, the first demonstration of automated clustering and labeling of human behavior from brain recordings in a naturalistic setting; we achieved annotation without manual labels by leveraging techniques from computer vision and speech processing. Our unsupervised approach discovers clusters for behaviors such as moving, speaking and resting from ECoG data. The discovered cluster labels were verified by comparison to manual labels for a subset of the data. We also demonstrate that projecting the cluster centers back onto the brain provides an avenue for automated functional brain mapping in natural settings.

Our goal was to develop an approach to decode human brain recordings by embracing the richness and variability of complex naturalistic behavior, while avoiding tedious manual annotation of data and fine tuning of parameters. Our current approach has a number of limitations which can be addressed by improving both the available information streams and the algorithmic processing. One limitation of our movement detection algorithm is lack of specificity to the subject when other people enter the frame of the camera. This is particularly challenging when another person overlaps with the subject, for example, when a nurse examines the patient. We are exploring the potential of better subject segmentation using a depth camera. The depth stream information will also allow us to perform much more detailed pose recognition, including obtaining specific movement information from isolated body parts.

A second limitation is our inability to identify the speaker in speech detection. Speech levels include the subject speaking, the subject listening to another person speaking in the room, and the subject listening to the TV or another electronic audio source. We expect that by placing an additional microphone in the room and using algorithms to distinguish speaker voices, it may be possible to more accurately localize the speaker and speech sources.

The temporal aspect of high-dimensional, long-term ECoG data may be better exploited to improve the clusters discovered by unsupervised pattern recognition techniques. For instance, dimensionality reduction by dynamic mode decomposition (DMD [63]) may be able to identify spatio-temporal patterns when repeated trials are not available. Phase synchrony and phase coupling may also serve as important neural correlates of behavior [64].

Overall, these results demonstrate that our method has the practicality and accuracy to passively monitor the brain and decode its state during a variety of activities. In our results, we see some variation in performance and cluster maps across days for the same subject. This variance may be due to changes in brain activity as the patient recovers from surgery, or it may represent natural variation from day to day.

Functional brain mapping acquired by analyzing neural recordings outside instructed tasks has direct relevance to how an individual brain functions in natural conditions. For instance, neural correlates of a subject repeating a series of specific actions may differ from the full range of neural signatures associated with movements in general. Previous attempts to do more “ethological” mapping based on non-cued activities have identified motor [65], [66] and speech [13], [44] related areas. Our approach to ethological functional brain mapping explores the analysis of task-free,

naturalistic neural data augmented by information from external monitoring, which enables us to perform the automated analysis at a much larger scale with long-term data.

We envision our automated passive monitoring and decoding approach with video and audio as a possible strategy to adjust for natural variation and drift in brain activity without the necessity to retrain decoders explicitly. Such an approach may enable deployment of long-term BCI systems, including clinical and consumer applications. More generally, we believe the exploration of large, unstructured, naturalistic neural recordings will improve our understanding of the human brain in action.

Methods

Subjects and recording

All six subjects had a macro-grid and one or more strips of electrocorticography (ECoG) electrodes implanted subdurally for presurgical clinical epilepsy monitoring at Harborview Medical Center. The study was approved by University of Washington's Institutional Review Board's human subject division; all subjects gave their informed consent.

Electrode grids were constructed of 3-mm-diameter platinum pads spaced at 1 cm center-to-center and embedded in silastic (AdTech). Electrode placement and duration of each patient's recording were determined solely based on clinical needs. The number of electrodes ranged from 82 to 106, arranged as grids of 8×8, 8×4, 8×2 or strips of 1×4, 1×6, 1×8. Figs. S5–S9 show the electrode placements of each subject. ECoG was acquired at a sampling rate of 999Hz. All patients had between six and fourteen days of continuous monitoring with video, audio, and ECoG recordings. During days 1 and 2, patients were generally recovering from surgery and spent most of

their time sleeping; in this study, days 3 to 6 post implant were analyzed from each subject.

Video and audio recordings

Video and audio were recorded simultaneously with the ECoG signals and continuously throughout the subjects' clinical monitoring. The video was recorded at 30 frames per second at a resolution of 640×480 pixels. Generally, video was centered on the subject with family members or staff occasionally entering the scene. The camera was also sometimes adjusted throughout the day by hospital staff; for instance, the camera may be centered away during bed pan changes and returned to the patient afterwards. Videos S1–S4 show examples of the video at a few different times of one day. The audio signal was recorded at 48 KHz in stereo. The subject's conversations with people in the hospital room, including people not visible by video monitoring, can be clearly heard, as well as sound from the television or a music player. Some subjects listened to audio using headphones, which were not available to our audio monitoring system. For patient privacy, because voices can be identifiable, we do not make examples of the audio data available in the supplemental materials.

Manual annotation of video and audio

To generate a set of ground-truth labels so that we may assess the performance of our automated algorithms, we performed manual annotation of behavior aided by ANVIL [67] on a small subset of the external monitoring data. Two students were responsible for the annotations, and at least 40 minutes (or 2.78%) of each subject day's recording was manually labeled for a variety of salient behaviors, including the broad categories of movement, speech and rest. Manual labeling was done for 2-minute segments of video and audio, distributed randomly throughout each 24-hour day. For

patient privacy, some small parts of the video (e.g., during bed pan changes) were excluded from manual labeling. These periods were very brief and should not introduce a generalized bias in manual labels. On average, manual labeling of 1 min of monitoring data was accomplished in approximately 5 minutes. At least 10 minutes of each subject day were labeled by both students; agreement between the two labelers was 92.0%, and Cohen's Kappa value for inter-rater agreement is 0.82.

Automated movement and speech detection

For automated video analysis, we first detected salient features for each frame using Speeded Up Robust Features (SURF), which detects and encodes interesting feature points throughout the frame. The amount of motion in each frame was determined by matching the magnitude of change in these feature points across successive frames. Since the subject was the only person in the frame a majority of the time, we are able to determine the subject's approximate movement levels. This approach detected gross motor movements of the arms, torso and head, as well as some finer movements of the face and mouth during speaking. To detect speech, we measured the power of human speech frequency levels (100–3500Hz) from audio data.

We assessed the performance of the automated algorithms by comparison to manual annotations. The manual annotations for each behavior were binary (i.e., either the behavior was present or not in a time window) whereas the automated speech and movement levels were analog values. Therefore, the agreement was computed after applying a threshold to the automated movement and speech levels.

ECoG preprocessing and feature extraction

All ECoG recording was bandpassed filtered between 0.1 and 160Hz to reduced noise. The filtered signal was then converted into a set of power spectral features using short-time Fourier transform using non-overlapping two-second windows. Each 24-hour recording was thus separated into 43200 samples in time.

Because subjects engaged in a variety of activities throughout the day, there is no particular frequency band that would be solely useful for clustering. We considered power at a range of frequencies between 1–52 Hz for each electrode, binning every 1.5 Hz of power for a total of 35 features per electrode per two-second window. At 82 to 106 electrodes per subject, this process resulted in 2870 to 3710 features for each two-second window of recording. To normalize the data, we transformed the binned powers levels at each frequency bin for each channel by computing the Z-Score. The dimensionality of the feature space was then reduced with principal component analysis (PCA), and the cumulative fraction of variance explained as a function of the number of PCA modes for each subject day is shown in Fig. S1. These spectra were highly variable, both within and between subjects. For purposes of unsupervised clustering, we truncated all feature space to the first 50 PC's. The first 50 PC's generally accounted for at least 40% of the variance in daily power spectral features space.

Hierarchical clustering of ECoG features

We used the 50-dimensional principal component power spectral features of the ECoG data as features for our hierarchical k-means clustering. The hierarchical k-means procedure and the

annotation of clusters by correlation with movement and speech levels is shown schematically in Fig. 3.6. For each subject on each day, we first perform k-means with $k = 20$ clusters. Next, we segregate the data points into the single cluster with the most number of data points and all the rest of the clusters. This procedure produces the first level of the hierarchical clustering, which now has two clusters. Next, for level L of the clustering, this procedure is repeated for each cluster from level $L - 1$ using $k = 20/L$ (k floored to the largest previous integer). Again, the single cluster with the most number of data points is separated from the rest of the clusters, so that at level L, we end up with 2^L clusters. This process of recursive k-means clustering and aggregation is stopped when there are fewer than 100 data points in each cluster, or when $L = 10$. In this manuscript, we focused on analyzing annotated clusters in levels 1–4.

Automated annotation of clusters using video and audio recordings

Results from the clustering analysis were automatically annotated using movement and speech levels. Each type of unsupervised analysis produced time series at different temporal resolutions, so they were all first consolidated into a mean analog value for non-overlapping 16-second windows. For ECoG, we counted how many 2-second windows within each 16-seconds were assigned to a particular cluster at the target hierarchy level. For movement and speech detection, we considered what fraction of the 16-second window exceeded a threshold value. These thresholds were determined empirically.

After consolidation into windows of 16 seconds, we computed the Pearson r correlation between each of the ECoG clusters with the movement and speech levels. The “movement” and “speech” labels were assigned to clusters for which the correlation was the highest for each behavior. If

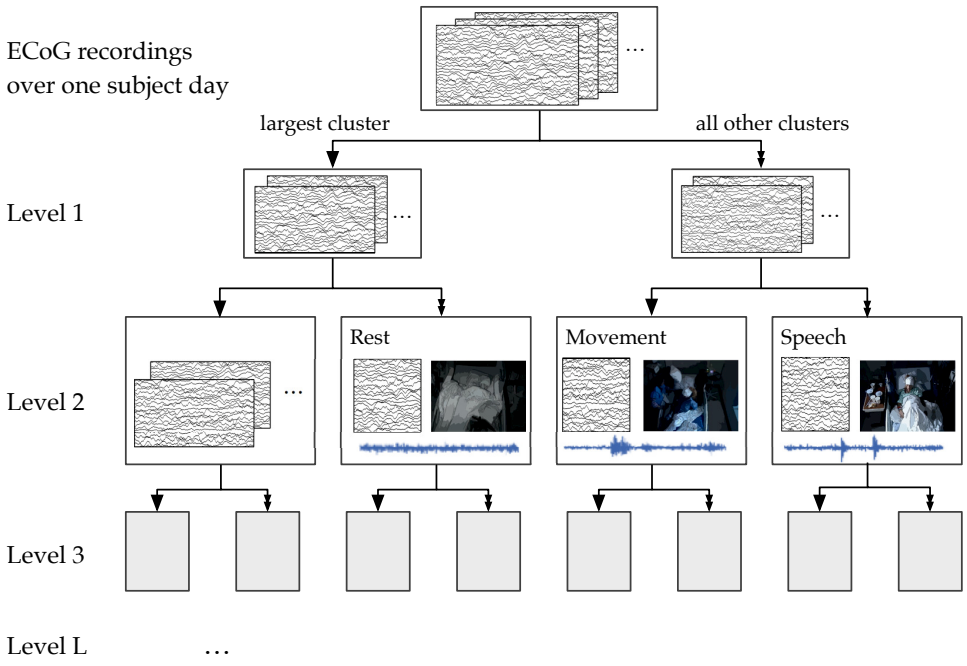


Figure 3.6: A schematic of the hierarchical clustering and annotation method. Features extracted from ECoG recordings of each subject day were recursively clustered and agglomerated at increasing levels. Annotation consisted of finding the cluster within each level whose time course had the highest correlation with automatically extracted movement and speech levels. For illustration purposes, here we show the annotation of clusters at Level 2, which has 4 total clusters.

movement and speech both correlated best with the same cluster, the label was assigned to the second best cluster for the activity type that had a lower correlation. The "rest" label was assigned to the cluster with the largest negative correlation with both movement and speech.

We performed this annotation assignment for clustering levels 1–4 (see Fig. 3.6). At level 1, for which there were only 2 total clusters, labels were simplified to be "rest" and "non-rest" (for movement and speech combined). Level 3, where there are 8 total clusters, appeared to be the most parsimonious level of granularity for the number of categories available automatically.

Validation with ground truth labels

Ground truth labels for random portions of each day were obtained from two students who hand annotated a small random fraction of each subject day (about 40 minutes, or 3% of each day) for visible and audible behaviors. The hand annotations were distributed randomly throughout each day of each patient. The automated results were compared to manual labels using 16 second windows within the manually annotated times. Each 16 second window is determined to contain an activity if the activity is annotated within any point in the window. Since different clusters have different baseline levels, we determined that the cluster detects the activity if its level is at or above the 25th percentile over the day. Using the manual labels as ground truth, accuracy and F1 scores were computed. The F1 score is the harmonic mean of precision and recall.

To determine the statistical significance of the F1 scores compared to chance, we generated shuffled labels by changing the timing of the ground truth labels of each activity, without changing their overall relative frequency. This shuffling was repeated over 1000 random iterations to determine the distribution of F1 scores assuming chance, and the true F1 score was compared against these shuffled F1 scores. We report the percentile of the true F1 scores for all subject days.

Mapping annotated clusters back to the brain

For each annotated cluster, we projected the centroid values of the cluster back to brain coordinates. The centroids values are 50-dimensional vectors in PCA space, reduced from power spectral features of all recording electrodes. The inverse PCA transform using the original PCA basis projects the centroid back to brain coordinates, where the relative power in each frequency bin is available at

each electrode. Note that because of the Z-Score normalization step before computing the original PCA basis, this back-projection reproduces Z-Score values, not voltages. These Z-Scores can be separately averaged according to frequency bins of interesting bands, including a low frequency band (LFB, 1–8 Hz) and a high-frequency band (HBG, 12–45 Hz) as show in Fig. 3.5. Fig. S5–S8 also show results of brain maps at 72–100Hz. Anatomic reconstruction of electrode coordinates on structural imaging of subjects' brains was available for only four of the six subjects, so we were unable to perform this mapping for Subject 2 and Subject 6.

Supplemental Figures

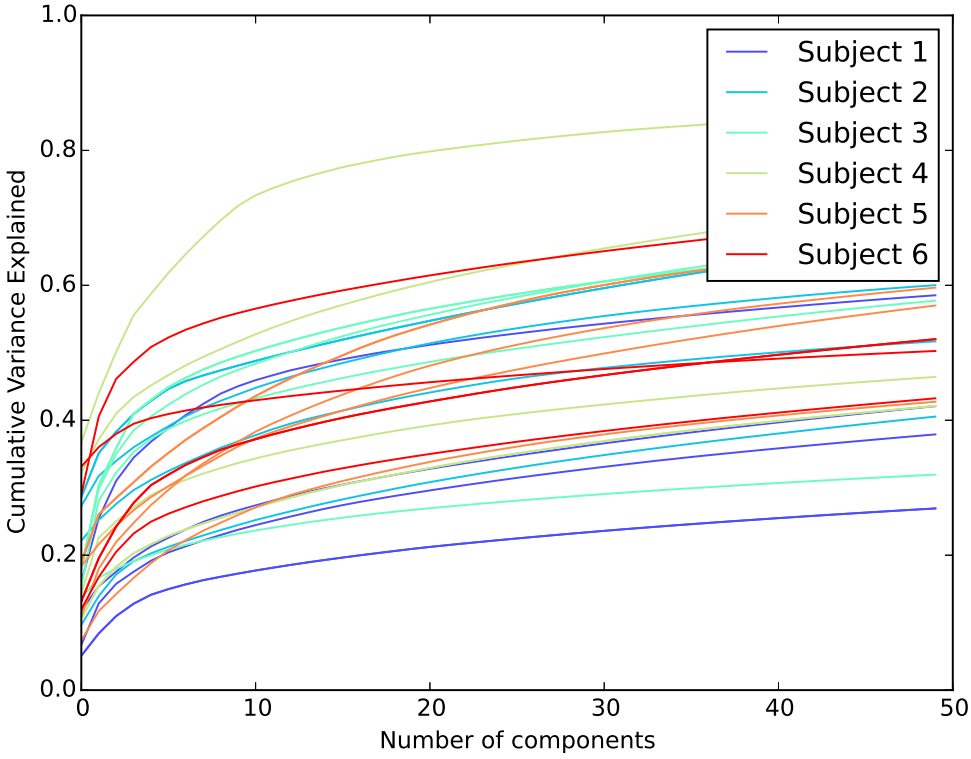


Figure 3.7: Cumulative variance explained by number of principal components (from PCA) shows that there is a wide variation across subjects and days. Although more components may be more accurate in terms of percentage of variance explained, k-means does not work well in very high dimensional space, so we chose to use 50-dimensional PCA features space for clustering. At 50 components, a large majority of days have at least 0.4 accumulated variance.

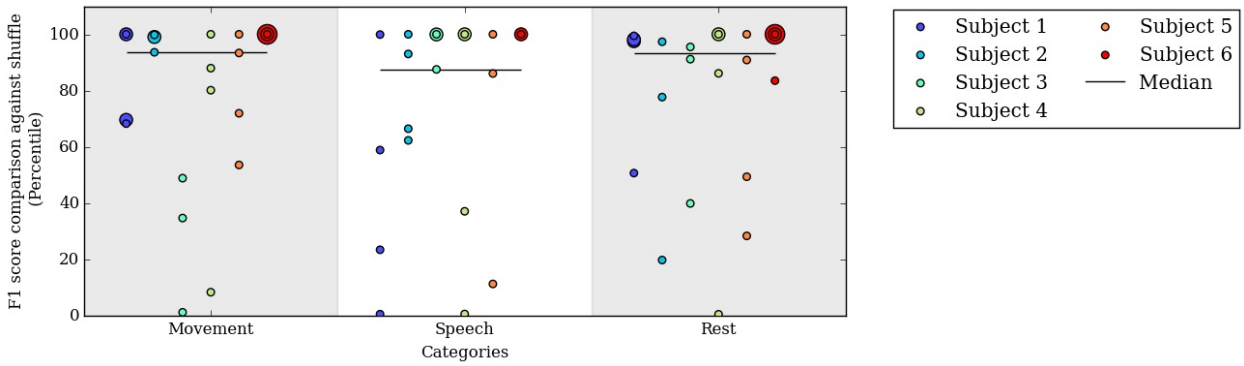


Figure 3.8: Percentile of the F1 score of our algorithm compared to manually annotated labels. Decoding at level 2 shows far worse performance than level 3, which is shown in the main text Figure 5. Each colored dot indicates the percentile of one day for a subject for a behavioural category.

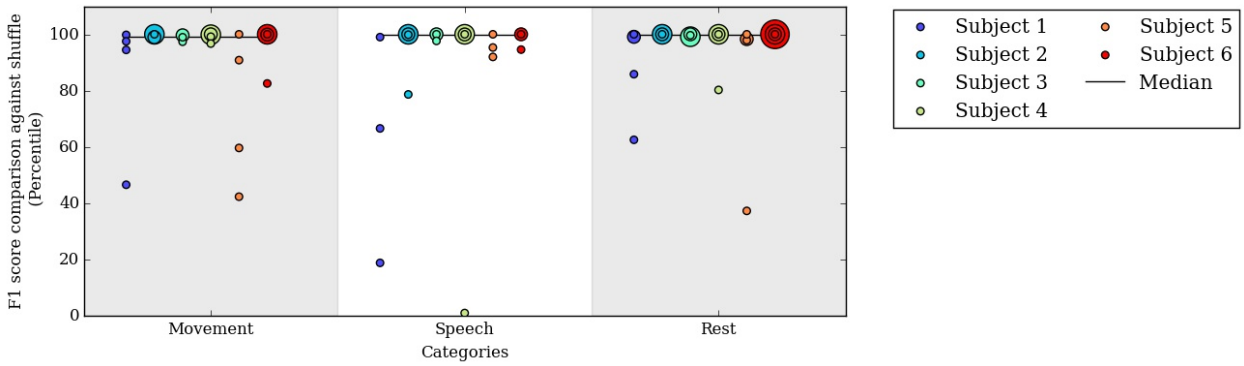


Figure 3.9: Percentile of the F1 score of our algorithm compared to manually annotated labels. Decoding at level 4 shows comparable performance than level 3, which is shown in the main text Figure 5. However, the number of time points belonging to each cluster is rather low, as shown in the level 4 graph of Figure 4 in the main text. Each colored dot indicates the percentile of one day for a subject for a behavioural category.

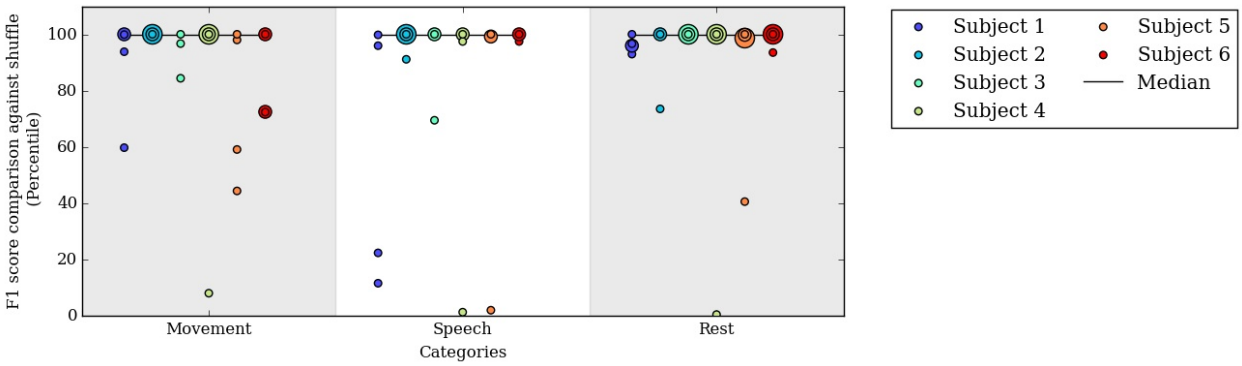


Figure 3.10: Percentile of the F1 score of our algorithm using features containing frequencies up to 105 Hz as compared to manually annotated labels. Decoding at level 3 using spectral frequencies up to 105Hz shows comparable performance as compared to level 3 using power from frequencies below 52Hz as shown in the main text Figure 5. However, as shown in the table S2, the accuracy using the high powers is worse than than without (as shown in the main text Table 1).

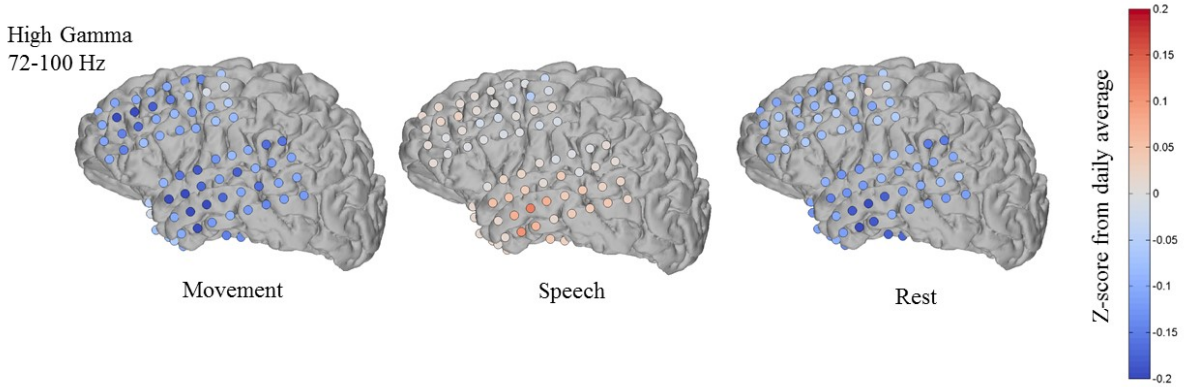


Figure 3.11: Features discovered by automated brain decoding using high gamma power band for Subject 1 day 6 post implant did not show localized high power levels during movement, as may be suggested by motor mapping experiments. However, there is some high power activity in the temporal region that may warrant further studies.

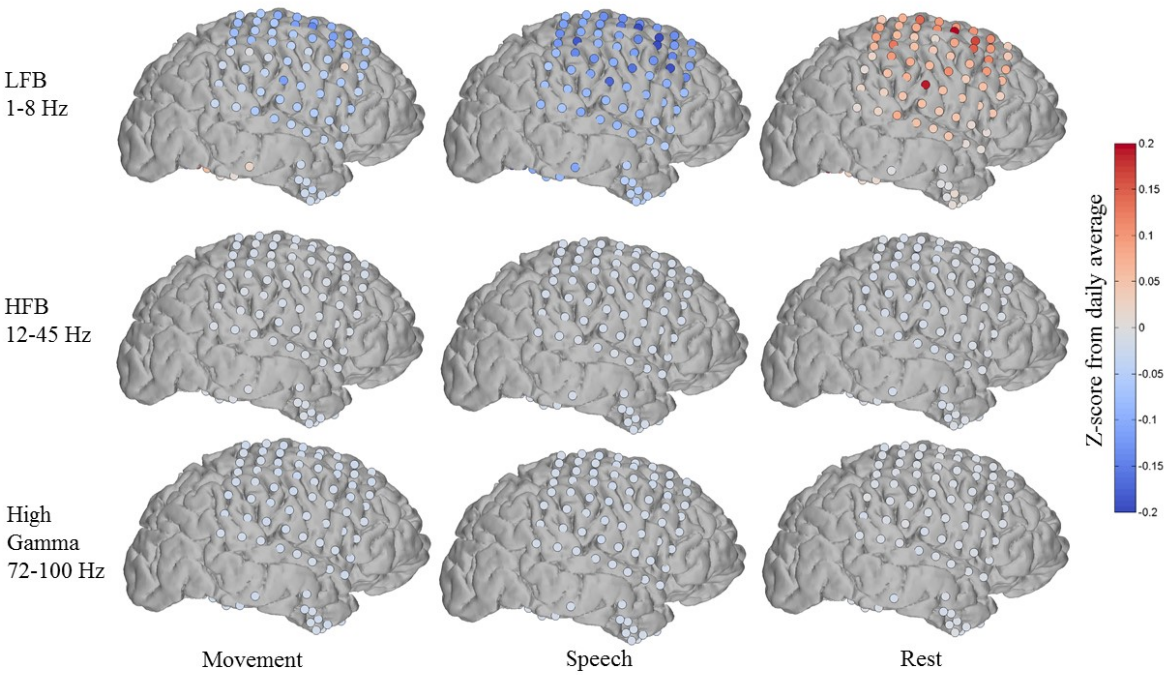


Figure 3.12: Features discovered by automated brain decoding using high gamma power band for a sample day of Subject 3 showing high power during rest in LFB and lower power during movement and speech. However, no localized activity changes can be detected.

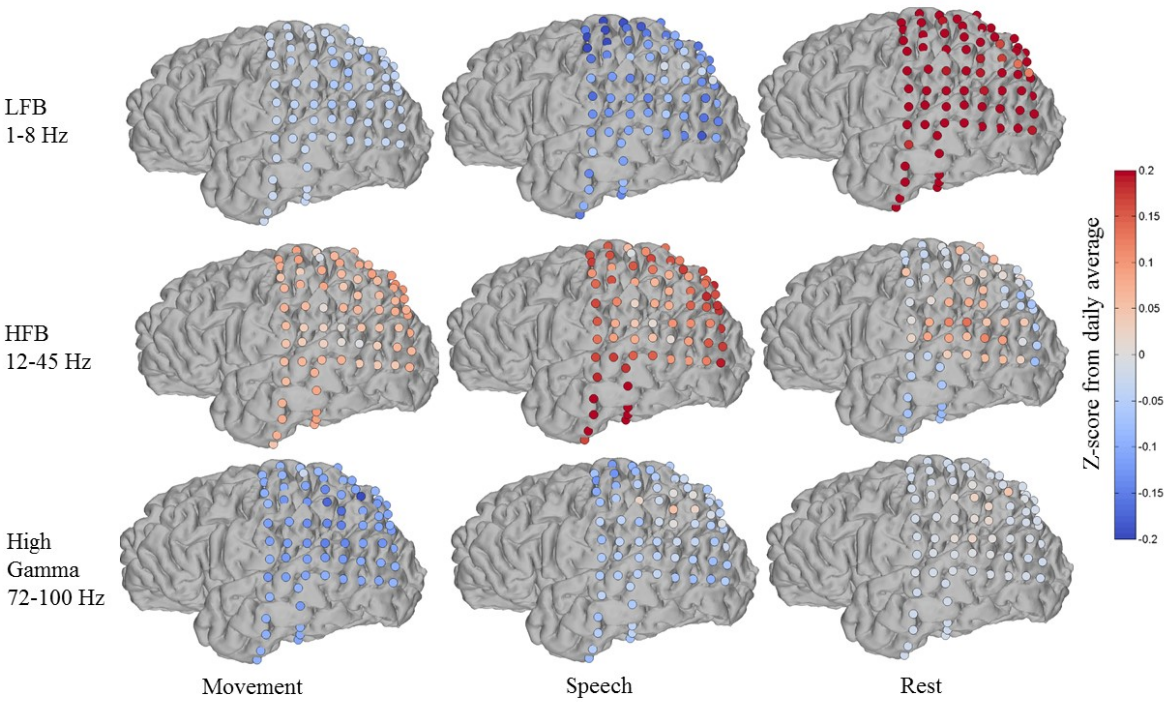


Figure 3.13: Features discovered by automated brain decoding using high gamma power band for a sample day of Subject 4 showing high power during rest in LFB and lower power during movement and speech. This activity level is flipped to high for movement and speech but lowered during rest in HFB. No significant pattern is detected for high gamma.

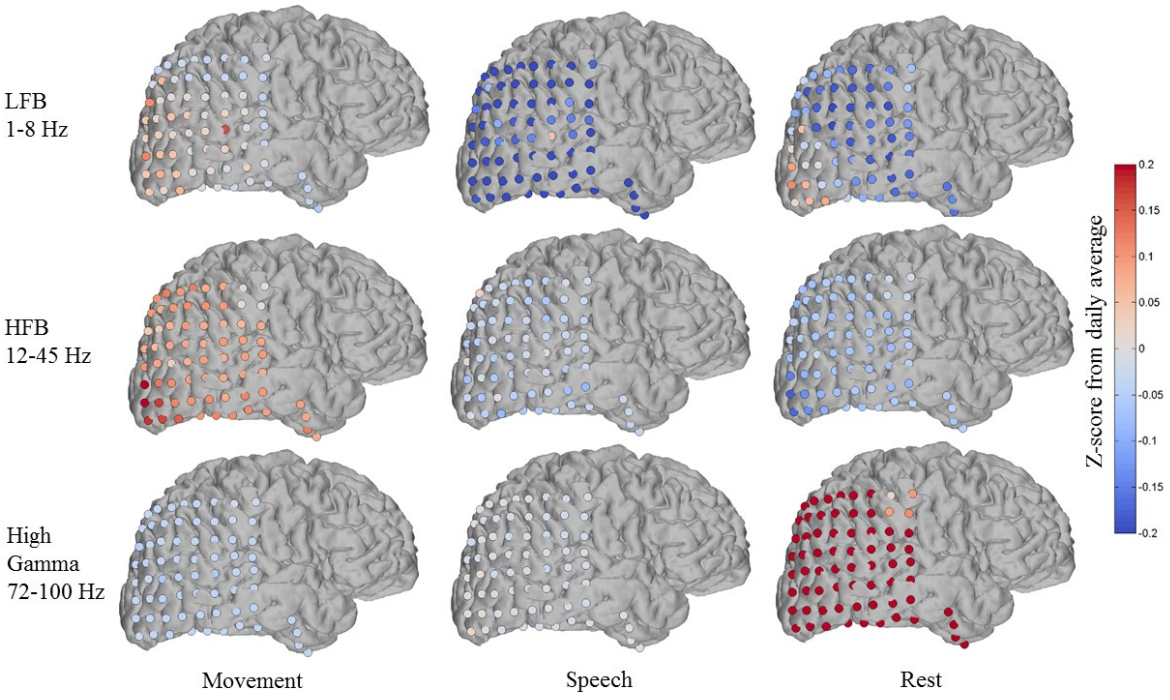


Figure 3.14: Features discovered by automated brain decoding using high gamma power band for a sample day of Subject 5 showing power changes in the visual cortex. This may be due to high correlations between movement, speech and visual sensing.

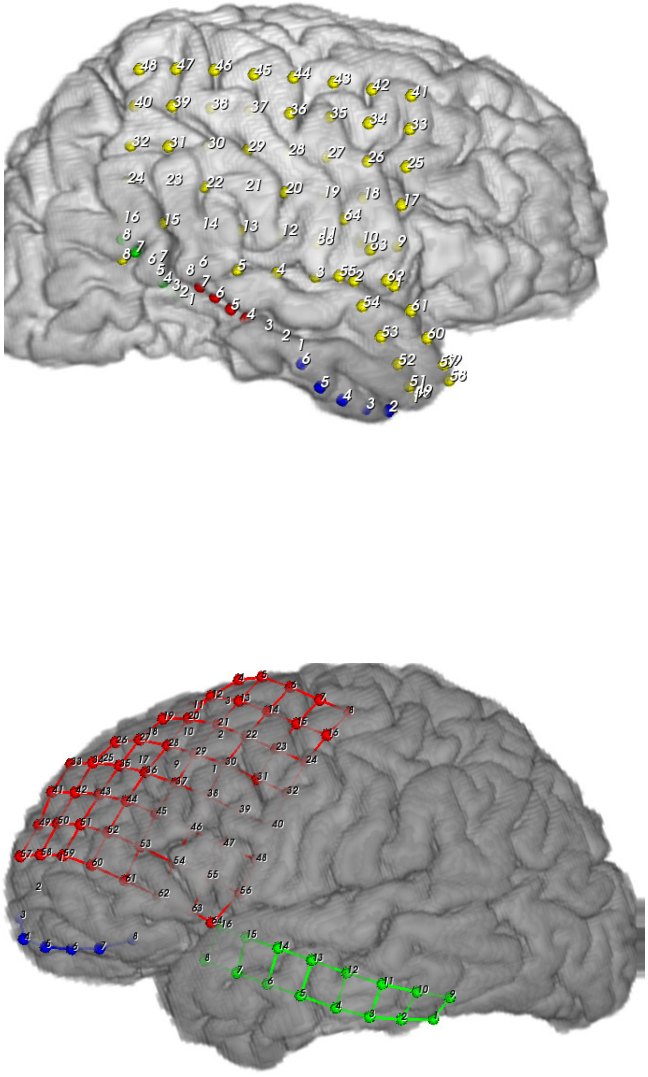


Figure 3.15: Clinical electrode maps are shown for subject 6 (left) and subject 2 (right). Electrode location reconstruction information was not available for these two subjects, so mapping of cluster centroids was not possible.

	Movement				Speech				Rest			
	F1	R_F1	R_std	Pctl	F1	R_F1	R_std	Pctl	F1	R_F1	R_std	Pctl
S1 D3	0.58	0.36	5.67	100.00	0.05	26.28	6.02	0.00	0.68	0.53	3.84	100.00
D4	0.60	0.41	6.42	99.90	0.16	0.31	7.95	2.70	0.82	0.68	3.35	100.00
D5	0.53	0.39	5.28	99.60	0.58	0.51	3.73	96.00	0.40	0.29	5.97	96.40
D6	0.27	0.27	4.88	50.20	0.42	0.31	6.12	96.60	0.55	0.51	3.64	86.25
S2 D3	0.39	0.36	1.55	94.05	0.68	0.56	2.15	100.00	0.35	0.30	2.49	98.65
D4	—-	—-	—-	—-	0.70	0.69	1.44	61.95	—-	—-	—-	—-
D5	0.56	0.47	3.48	99.75	0.56	0.36	4.53	100.00	0.68	0.36	4.68	100.00
D6	0.63	0.51	2.89	100.00	0.71	0.61	2.59	99.90	0.47	0.25	5.52	100.00
S3 D3	—-	—-	—-	—-	—-	—-	—-	—-	—-	—-	—-	—-
D4	0.48	0.39	4.98	96.15	0.79	0.44	4.99	100.00	0.80	0.54	3.89	100.00
D5	0.34	0.26	5.23	92.85	0.67	0.47	3.69	100.00	0.74	0.39	4.01	100.00
D6	0.58	0.55	3.06	80.95	0.52	0.44	4.04	98.05	0.51	0.34	4.93	100.00
S4 D3	0.59	0.51	3.05	99.45	0.58	0.31	4.46	100.00	0.70	0.48	3.22	100.00
D4	0.65	0.57	2.85	99.85	0.40	0.30	4.76	98.25	0.61	0.40	3.94	100.00
D5	0.64	0.57	3.19	98.45	0.34	0.47	3.98	0.00	0.21	0.39	4.55	0.00
D6	0.54	0.30	4.95	100.00	0.67	0.36	5.16	100.00	0.22	0.15	3.38	98.05
S5 D3	0.43	0.44	4.36	45.20	0.33	0.31	3.86	70.00	0.64	0.61	3.11	75.35
D4	0.44	0.34	7.25	90.30	—-	—-	—-	—-	0.78	0.78	2.52	52.65
D5	0.31	0.26	8.69	70.95	0.38	0.36	7.48	57.65	0.59	0.59	4.99	52.60
D6	0.64	0.37	5.12	100.00	0.73	0.36	5.35	100.00	0.83	0.60	3.28	100.00
S6 D3	0.71	0.68	1.78	93.25	0.58	0.49	4.10	97.90	0.57	0.42	4.80	100.00
D4	0.69	0.37	5.02	100.00	—-	—-	—-	—-	0.86	0.47	4.40	100.00
D5	0.29	0.32	3.37	23.95	0.67	0.21	6.91	100.00	0.79	0.66	2.51	100.00
D6	0.71	0.49	3.83	100.00	0.76	0.49	3.81	100.00	0.92	0.34	5.67	100.00

Table S1. F1 scores and percentiles as assessed by comparison of level 3 automated cluster annotation to manual annotations shown for each of the 4 days analyzed for each subject. The true F1 scores are compared to randomly shuffled F1 scores. Acc = Accuracy; R_acc = mean of random shuffle accuracy; R_std = standard deviation of random shuffle accuracy; Pctl = percentile of accuracy score within random shuffles; S = Subject; D = Day; — = Not enough manual labels were collected for . These F1 scores correspond to the same data shown as percentiles against random shuffle in Fig. 5.

Movement	Acc	F1	Spc	Sen/Rec	Prc
Subject 1	60.99	0.52	72.45	49.12	59.94
Subject 2	59.95	0.67	76.37	56.25	85.59
Subject 3	61.57	0.44	72.99	42.64	49.01
Subject 4	57.00	0.59	52.16	61.90	56.05
Subject 5	55.16	0.40	61.91	48.58	50.57
Subject 6	68.96	0.65	68.39	61.86	73.83
Speech	Acc	F1	Spc	Sen/Rec	Prc
Subject 1	50.18	0.42	46.75	57.21	34.40
Subject 2	59.89	0.65	75.52	54.01	85.02
Subject 3	67.78	0.70	59.09	77.50	64.64
Subject 4	62.84	0.54	61.82	62.24	48.43
Subject 5	68.25	0.50	79.17	43.61	53.63
Subject 6	68.33	0.65	70.61	72.88	66.04
Rest	Acc	F1	Spc	Sen/Rec	Prc
Subject 1	58.64	0.61	54.41	67.22	59.88
Subject 2	61.61	0.50	61.61	25.81	46.23
Subject 3	71.80	0.63	85.82	53.22	76.93
Subject 4	57.27	0.46	82.42	36.62	72.39
Subject 5	62.24	0.65	60.65	60.78	77.42
Subject 6	76.91	0.70	83.21	59.33	82.03

Table S2. Performance metrics as assessed by comparison of level 3 annotation (from automated clusters based on frequencies between 1 Hz and 105 Hz) to manual annotations averaged over all 4 days for each subject. Acc = Accuracy; Spc = Specificity; Sen/Rec = Sensitivity/Recall; Prc = Precision. This table shows values also presented in Fig. 3.10. This performance is worse than the accuracy results reported with only frequencies up to 53 Hz as shown in the main text Table 1, particularly in the rest category.

Chapter 4

AJILE MOVEMENT PREDICTION: MULTIMODAL DEEP LEARNING FOR NATURAL HUMAN NEURAL RECORDINGS AND VIDEO

Nancy X. R. Wang, Ali Farhadi, Rajesh P. N. Rao, Bingni W. Brunton

Sourced from [68]

Developing useful interfaces between brains and machines is a grand challenge of neuroengineering. An effective interface has the capacity to not only interpret neural signals, but *predict* the intentions of the human to perform an action in the near future; prediction is made even more challenging outside well-controlled laboratory experiments. This paper describes our approach to detect and to predict *natural* human arm movements in the future, a key challenge in brain computer interfacing that has never before been attempted. We introduce the novel Annotated Joints in Long-term ECoG (AJILE) dataset; AJILE includes automatically annotated poses of 7 upper body joints for four human subjects over 670 total hours (more than 72 million frames), along with the corresponding simultaneously acquired intracranial neural recordings. The size and scope of AJILE greatly exceeds all previous datasets with movements and electrocorticography (ECoG), making it possible to take a deep learning approach to movement prediction. We propose a multimodal model that combines deep convolutional neural networks (CNN) with long short-term memory (LSTM) blocks, leveraging both ECoG and video modalities. We demonstrate that our models are able to detect movements and predict future movements up to 800 msec before movement initiation. Further, our multimodal movement prediction models exhibit resilience to simulated ablation of input neural signals. We believe a multimodal

approach to natural neural decoding that takes context into account is critical in advancing bioelectronic technologies and human neuroscience.

Introduction

Scientists, engineers, and speculative fiction authors have long imagined possible futures when people interact meaningfully with machines directly using thought. Technologies that interpret brain signals to control robotic and virtual devices have tremendous potential to assist individuals with physical and neurological disabilities, to augment engineered systems integrating humans in the loop, and to enhance one's daily life in an increasingly information-rich world.

In recent years, research in brain-computer interfacing (BCI) [8], [9] has been very successful in using decoded neural signals to control robotic prostheses and computer software (for instance, [25], [26], [69]). Even so, these impressive demonstrations have relied on finely tuned models trained on experimentally derived labeled data acquired in well-controlled laboratory conditions. Thus, the remarkable feats of neural decoding to mobilize patients who have lost use of their limbs remain untested outside the laboratory.

One key challenge is how neural decoding may be approached "in the wild," where sources of behavioral and recording variability are significantly larger than what is found in the lab. Further, neural responses are known to differ between experimental and freely behaving conditions [12]. A flexible, scalable approach to detect movement and to predict initiation of natural movement would critically enable technologies to foster seamless collaborations between humans and machines.

In this paper, we present a multimodal deep learning approach that is able to detect whether a subject is initiating a movement and to predict initiation of natural movements in the future. This

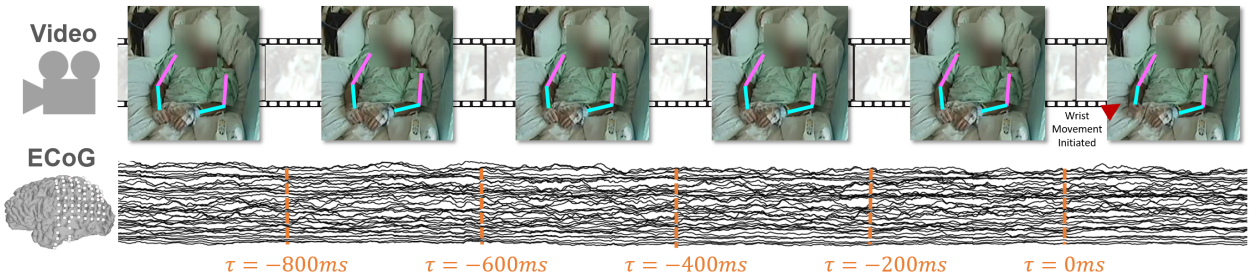


Figure 4.1: An example of a 1-sec window of multimodal data. In each video frame, we show the tracked positions of the upper arm (pink) and forearm (turquoise) for the subject's two arms by the pose recognition algorithm. A right wrist movement was detected in the last frame highlighted. Traces of neural activity acquired by ECoG are shown in black, where deflections represent voltage and a subset of the electrodes are shown stacked vertically.

paper is the first to develop a deep neural network that models naturalistic ECoG signals. The main contributions of the paper are:

- We introduce the AJILE dataset of long-term natural neural recordings with corresponding labeled arm poses.
- Using the AJILE dataset, we show that our proposed multimodal deep neural network can *predict* the intention to move a hand up to 800 msec before the initiation of the movement in naturalistic data, a task never attempted before.
- Our experimental evaluations show the significance of incorporating context provided by the video, using a deep multimodal model to decode and predict movement intention.

Related work

Human intracranial cortical recordings. Neural signals used to monitor brain activity are acquired by a variety of recording technologies that differ by invasiveness and portability. One such technology known as electrocorticography (ECoG) for intracranial recordings is particularly attrac-

tive for studying naturalistic behavior because of its spatial specificity, temporal resolution, and potential longevity of implantation. ECoG offers measurements of temporal dynamics inaccessible by functional magnetic resonance imaging (fMRI) and spatial resolution unavailable to extracranial electroencephalography (EEG). In addition, it is not feasible to use fMRI when one is interested in behaviors over hours and days, especially behaviors that require gross motor movement and meaningful interaction with the surrounding environment. Cortical surface ECoG is accomplished less invasively than with penetrating electrodes [17], [18] and has much greater signal-to-noise ratio than entirely non-invasive techniques such as EEG **Lal2005**, [19].

Decoding movement. Efforts to decode neural activity have been accomplished by training algorithms on tightly controlled experimental data with repeated trials. Recent examples include decoding arm trajectories [20], [22] and finger movements [23], [24]. Decoded neural signals have been used to control robotic arms [25]–[27] and to construct BCIs [28], [29], [31], [32].

Analyzing naturalistic brain recordings. The lack of ground-truth labels makes analyzing and decoding naturalistic neural recordings especially challenging. Labels must be obtained by a separate measurement acquired simultaneously with the neural recordings. Previous studies exploring this idea have decoded natural speech [44], [45] and natural motions of grasping [14], [15]; however, these studies had relied on laborious manual annotations and do not reach the size and comprehensiveness of our AJILE dataset. A few studies have made use of some automation to analyze natural data [7], [70], but none had focused on prediction of future events. Our work takes advantage of recent advances in computer vision to annotate a variety of natural data including automated movement estimation [50], [51] and pose recognition [52], [71].

Deep neural networks for ECoG/EEG data. Deep learning has not been widely applied to ECoG and EEG data, with notable exception in the cases of [72]–[75]. We are inspired by work in related fields that have made use of multimodal data streams. For instance, multimodal networks have been most widely used in the tasks of speaker identification [76] and speech recognition [77]. Our approach also has similarities to [78], where visual data was leveraged to derive sound categories from an otherwise unlabeled large dataset. To our knowledge, there is no previous work combining ECoG with another modality of measurement for developing deep learning models.

Dataset

4.0.1 Long-term, naturalistic neural activity and video

Our long-term, naturalistic human movement dataset includes week-long continuous multimodal recordings with invasive (ECoG) electrodes, video, and audio. This opportunistic dataset greatly surpasses all previously analyzed comparable datasets in duration and size. In total, we have approximately 670 hours of recordings for 4 subjects, which amounts to more than 72 million frames of video and more than 2 billion samples of ECoG at a sampling rate of 1000Hz. Importantly, these subjects performed no instructed tasks; instead, they simply did as they wished for the duration of their monitoring in the hospital room. The variety of natural behaviors observed was rich and complex, including conversations with friends and family, eating, interacting with electronic devices, and sleeping. Understanding the connection between neural activity and naturalistic behavior presents a great data analytic challenge, in part because of the immense task of annotating such long-term recordings. At the same time, making sense of this data presents an unique opportunity to shed light on neural function outside the laboratory.

4.0.2 The AJILE dataset: Annotating joint locations

In this study, we leveraged the latest innovations in computer vision to train a deep neural network to automatically retrieve the patient pose from each frame. We used the YOLO [79] framework for subject detection and caffe-heatmap [71] for pose estimation. Fig. 4.1 shows a schematic of pixel locations of the head, shoulders, elbows, and wrists as they were extracted from the video frames. To improve the performance of the standard trained models on our dataset, we acquired custom manual annotations on a small fraction of the video to retrain YOLO and caffe-heatmap. A small portion of videos taken from 18 subjects was annotated, with one frame manually labeled approximated once every 2 minutes. All four subjects in this study were part of the pose estimation training set. After over 3000 GPU hours of processing, we extracted locations of 7 upper body joints for over 72 million frames. Fig. S1 shows a validation for the accuracy of these joint locations; our pose estimation is extremely accurate for confidence scores above 0.25, so this threshold was chosen as the cutoff for extraction of natural movement annotations. For a typical patient, approximately half of the frames have a confidence score above 0.25.

We plan to make publicly available this Annotated Joints in Long-term ECoG (AJILE) dataset, at the time of publication of this manuscript. AJILE includes raw ECoG voltage recordings, electrode locations, and estimated pose in each video frame. The pose comprises pixel locations of 7 upper body joints, along with an estimated confidence value for each joint. AJILE does not include raw video recordings, a restriction due to patient privacy.

4.0.3 Extracting initiation of natural movements from AJILE

To define movement from estimated poses, we focused on wrist movements of the arm contralateral to the electrode array implant. We first smoothed the joint location results from AJILE using a Savitzky-Golay filter with a 21-frame window. A movement initiation was defined as when movement of the wrist joint averaged over 5 consecutive video frames exceeded an average of 1 pixel per frame, and less than 0.5 pixels of movement was detected when averaged over the previous 10 frames. Times of no movement were selected when there was less than 0.5 pixels of movement in all joints averaged over 30 frames before and after the time point in question.

A small portion of automated movement detection was validated with manual annotations. We found the inter-rater reliability to be 95.3% and the overall accuracy against the raters are 86.9% and 84.2%. After the automated process, the data was curated to discard obviously inaccurate labels. For example, we removed samples where there was another person moving in front of or obstructing the arm, as well as samples during sleep, since neural patterns are known to be drastically different between sleep and wake. Table 4.1 summarizes the number of instances of movement initiation in the dataset for each of the four subjects. The training data includes movements from days 2 to 5 of the clinical monitoring, and day 6 or 7 was used for testing. Each train and test dataset was balanced so that they contained roughly equal numbers of movement and no movement samples.

4.0.4 Data preprocessing

All ECoG recordings were bandpass filtered between 10 and 200Hz. For the neural network models, a 1-second window of high-dimensional time-series data was used as the input (shown schematically

in Fig. 1). Each 1-second window of recording for each electrode was normalized to the mean and standard deviation of its 3-second neighborhood that does not contain any times of movement, then broken into a sequence of five 200 msec chunks. These chunks were used as inputs to the neural network.

For our multimodal model, each 200 msec chunk of data was associated with one video frame, which was extracted from the middle of the 200 msec time window. Video frames were resized from 640×480 down to 341×256. During training, the ECoG data was augmented with noisy perturbations 25% of the time using gaussian random noise of standard deviation 0.001 and temporal shifting of up to 100 msec in either direction. All video frames during training were randomly cropped into 224×224 images for input into the networks. During testing, the images were always cropped at the center of the frame.

In all training schemes, each subject's data is trained and tested separately and independently. Combining data across subjects was not possibly because of large differences in electrode coverage.

	S1	S2	S3	S4
Train	1560	2002	4587	3490
Test	313	575	1952	193

Table 4.1: Number of samples in the dataset for each subject. The test set was chosen to be on a day of recording different from the training set; variations are due to the activeness of each subject.

4.0.5 Clinical data collection details

The subjects in our dataset were patients undergoing pre-surgical clinical epilepsy monitoring. The study was approved by our institute's human subject division; all four (4) subjects gave their informed consent and all methods were carried out in accordance with the approved guidelines. Electrode placement and duration of each subject's recording were determined solely based on clinical needs. Each subject had 80–94 ECoG electrodes implanted subdurally, which is to say, directly on the brain under the skull and dura, a tough membrane surrounding the brain. S1, S2, and S3 had electrode implants in the left brain hemisphere; S4 was implanted in the right hemisphere. The electrodes are arranged as grids of 8×8, 8×4, 8×2 or strips of 1×4, 1×6, 1×8. Electrode grids were constructed of 3-mm-diameter platinum pads spaced at 1 cm center-to-center and embedded in silastic (AdTech). Electrodes that experienced failure during the subject's recording were rejected from the dataset. Fig. S2-S5 show the electrode placements of each subject. All subjects had between 6 and 7 days of continuous monitoring with video, audio and ECoG recordings. Video and audio were recorded simultaneously with the ECoG signals and continuously throughout the subjects' clinical monitoring. Generally, video was centered on the subject with family members or staff occasionally entering the scene. The video was recorded at 30 frames per second at a resolution of 640×480 pixels. The camera was sometimes adjusted throughout the day by hospital staff. For example, the camera may be moved during bed pan changes and returned to the subject afterwards, but not always to exactly the same position. Fig. 4.1 shows example video frames from one subject; the face was blurred to protect their privacy.

Prediction of movement in the future

4.0.6 Defining the problem

The task we address in this paper is the prediction of spontaneously generated arm movements in the future. A viable solution to this problem is critical for the application of brain-computer interfacing "in the wild," where a model must be able to tolerate significant noise and variability in natural, uncontrolled conditions. To our knowledge, this task has never before been attempted.

4.0.7 Our approach

Our approach to this prediction problem is inspired by recent advances in multimodal deep neural networks. We formulated a movement initiation classification problem using training and test data extracted from the AJILE dataset (Table 4.1). We reasoned that such a flexible framework would adapt to the dynamic environment in the data, synthesizing cues from direct recording of brain activity and contextual information provided by the video. Our multimodal model (Fig. 4.2) comprises two parallel towers, one 3-layer 1D CNN for ECoG and one 4-layer 2D CNN for video inputs, which are then merged with a fully connected layer followed by a LSTM layer of 20 units. Input data are fed into the CNN in 5 sequential chunks that includes one second of recordings in total.

Although neural activity is the ultimate director of one's future actions, ECoG is a very incomplete sampling of the brain, so we believe information from the video adds context that may improve accuracy in the prediction problem. In addition, multimodal information should make the model more robust to noise and variability than with a single modality alone. ECoG and video

data are both sequential by nature, so we developed a sequential model to match. However, we know from extensive literature analyzing ECoG signals that power-frequency features are usually more informative than raw voltage, so the convolutional layers act as feature extraction before the sequential layer. Finally, we decided to fuse ECoG and video towers in a late fusion model, hypothesizing that features that we can extract using the CNN would be different for each modality and should have different filtering sizes and layer structure.

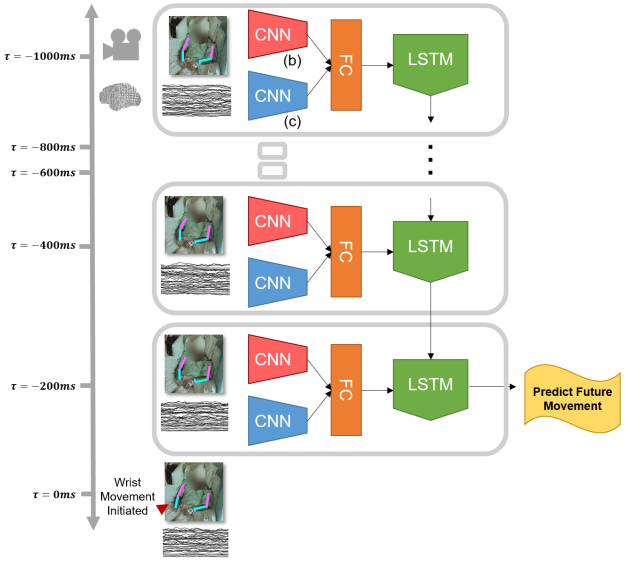
Table 4.2: Multimodal vs. ECoG only vs. Video only

	Multimodal			ECoG only			Video only		
	Pred-b	Pred	Det	Pred-b	Pred	Det	Pred-b	Pred	Det
S1	76.8	81.9	87.0	66.8	58.6	66.4	86.7	75.9	81.3
S2	59.4	61.7	61.5	66.7	65.7	64.6	58.0	49.8	52.1
S3	67.9	71.2	79.8	66.1	68.3	83.2	65.7	64.0	65.4
S4	66.1	62.0	62.5	56.8	54.7	57.3	49.0	54.7	57.8
Average	**67.6**	**69.2**	**72.7**	64.1	61.8	67.9	64.9	61.1	64.2

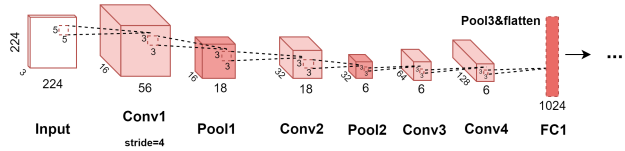
Table 4.3: Late Fusion vs. Early Fusion vs. Naive Averaging

	Late Fusion			Early Fusion			Naive Averaging		
	Pred-b	Pred	Det	Pred-b	Pred	Det	Pred-b	Pred	Det
S1	76.8	81.9	87.0	82.0	67.7	78.1	85.9	80.4	85.4
S2	59.4	61.7	61.5	56.1	51.2	60.6	62.5	57.3	58.0
S3	67.9	71.2	79.8	52.2	55.7	79.4	69.1	68.4	72.4
S4	66.1	62.0	62.5	53.1	69.3	55.2	50.0	60.9	55.7
Average	**67.6**	**69.2**	**72.7**	60.9	61.0	68.3	66.9	66.8	67.9

(a) Multimodal neural network schematic



(b) Video CNN architecture



(c) ECoG CNN architecture

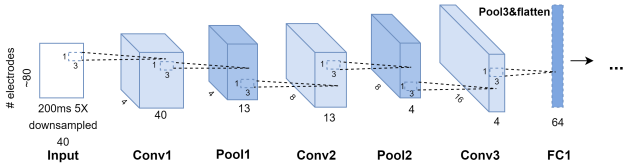


Figure 4.2: A schematic of the multimodal neural network architecture for the prediction of future hand movements using ECoG and video frame data. The ECoG data was separated into five 200-msec chunks. Each ECoG chunk and one frame at the center of the 200-msec time period was extracted from the video to use in the input sequence for the CNN/LSTM neural network.

Experiments

Our proposed multimodal neural network is able to detect and predict future movements. We present results for models trained on three different timing conditions: detection (Det), prediction (Pred),

and further back prediction (Pred-b). For detection, the 1-sec window of data was centered at the time of initiation of movement (i.e., 500 msec before and 500 msec after). For prediction, data was taken 1300 msec to 300 msec before the movement, so that the initiation itself was not included in the data. For further back prediction, data was taken 1800 msec to 800 msec before the movement. Table 4.2 shows that, on average across four subjects, the multimodal model performs well above chance (50%) and out-performs similar models that use a single modality of data in every timing condition.

In the remainder of this section, we describe our analyses of the multimodal model and demonstrate that our proposed approach has advantages over similar, related models. These analyses highlight the significance of combining context from video with direct neural recordings to enhance movement prediction. Further, we report results from synthetic electrode ablation experiments to evaluate the resilience of the multimodal neural network model.

4.0.8 Analyses of the multimodal model

Optimal Fusion Point. Our multimodal model uses video and ECoG streams of data after each stream has been separately fed through their respective CNN's (Fig. 4.2). We refer to this fusion scheme as "late fusion." To compare this strategy with other potential points of fusion for the video and ECoG data, we trained an alternative "early fusion" model that stacks ECoG with video directly in the input, resizing the image as needed. In addition, we compare with a "naive average" model that averages the final sigmoid outputs of the ECoG only and video only models. Table 4.3 shows that "late fusion" outperforms both of the other schemes. Early fusion likely suffers from forcing the two modalities to have the same CNN architecture when many hyperparameters (such

as filter size) should be quite different for the very different types of data. The improvement in accuracy as compared to naive averaging suggests that the fully connected layers and LSTM after the merge layer are learning aspects that are multimodal in nature, beyond a simple combination of probabilities from each modality.

Importance of CNN. Since ECoG is a sequential data type, we investigated the potential of a purely sequential model for the classification task. As shown in Table 4.4, the LSTM-only model performs at around chance. This observation shows the importance of feature extraction from the CNN layers before classification.

ECoG filter dimensions. Each subject in the dataset has at least one large grid of electrodes that is 8×8 in shape. Since this electrode geometry is known, we investigated to the potential of using a 3-dimensional convolutional filter on the ECoG data to take advantage of neighboring electrode positions. In direct comparison with the 1D filters used in our proposed model, which filters the ECoG data only in time, the 3D convolutional filters do not perform as well(Table 4.5). In this comparison, we removed the LSTM portion of the models to more directly compare the CNN filter schemes. We speculate that the time domain is more informative for predicting hand movement than the spatial domain. When our convolutional filters and pooling involve both space and time, this may be reducing the amount of information that can be obtained from the time domain.

Comparison to traditional baseline. To compare our deep learning approach to models more conventionally used to analyze ECoG data, we developed a baseline model using a linear SVM classifier based on power spectral features (see for instance [26], [80]). The power spectral features in two frequency bands (10–30 Hz and 70–100 Hz) were extracted using short-time Fourier transform

using non-overlapping 1-sec windows. Model selection of the baseline model was performed using a validation set drawn from the training days. Table 4.6 shows our deep learning ECoG only model outperforms the traditional spectral feature based SVM model.

Resilience of models after virtual electrode ablation. An important feature of movement detection and prediction models is robustness to disturbances. Here we investigated one type of robustness, namely the resilience of each model to a catastrophic electrode failure. This type of failure, when an electrode becomes entirely not functional, is not uncommon in real-life; the point of failure may be due to the electrode/amplifier interface, the wire connection, or movement/scarring of the brain tissue.

To simulate electrode failure, we systematically ablated each individual electrode in turn, substituting its true signal with a constant set to its mean value over time. Thus, we map each electrode's importance in making a detection or prediction by the impact of its ablation on the overall model accuracy. These maps also allow us to directly compare key electrode locations with known cortical maps from the human neuroscience literature.

Table 4.4: Conv + LSTM vs. LSTM only

	Conv + LSTM			LSTM only		
	Pred-b	Pred	Det	Pred-b	Pred	Det
S1	66.8	58.6	66.4	48.2	49.3	51.2
S2	66.7	65.7	64.6	48.8	53.5	50.0
S3	66.1	68.3	83.2	51.6	50.7	56.8
S4	56.8	54.7	57.3	47.9	54.2	46.9
Average	**64.1**	**61.8**	**67.9**	49.1	51.9	51.2

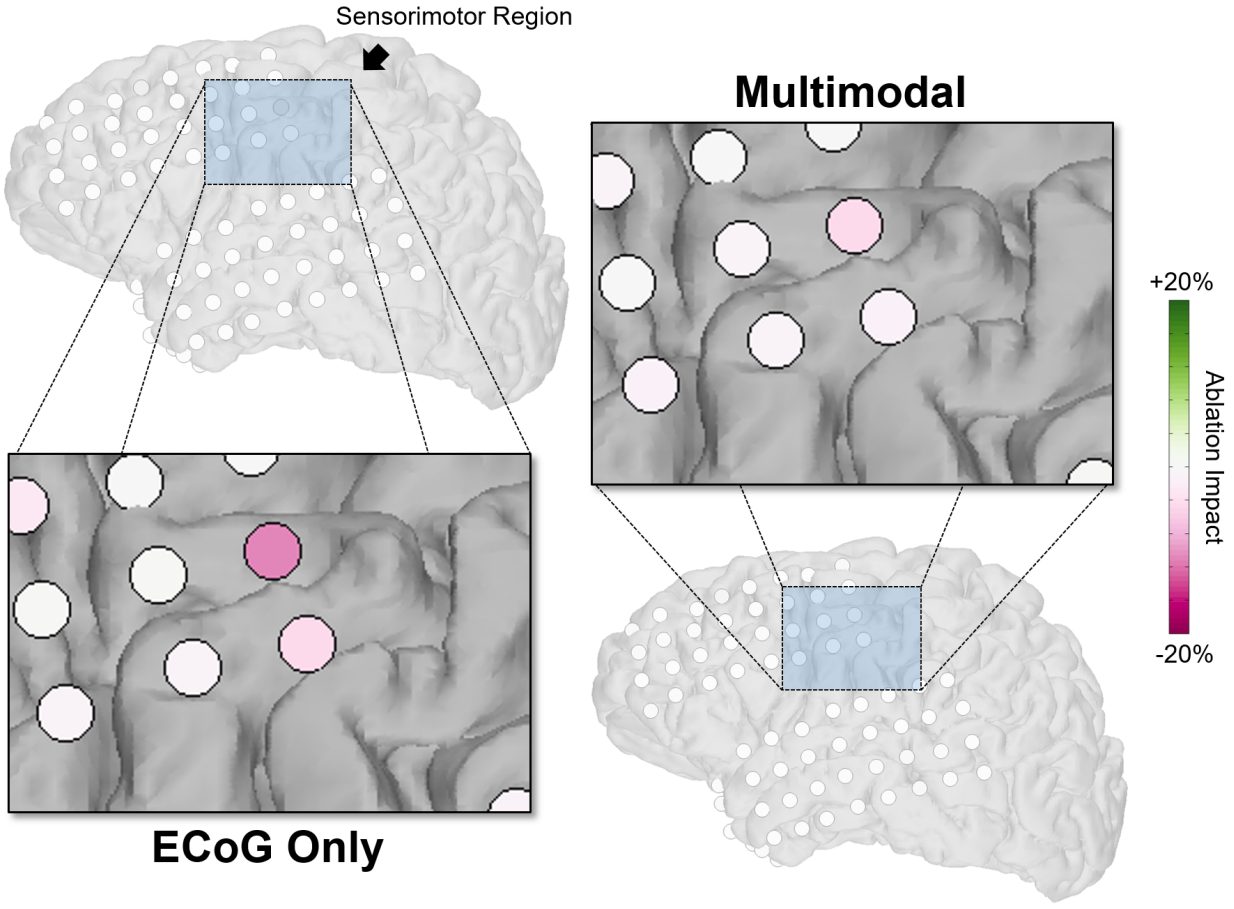


Figure 4.3: Effects of simulated single electrode ablation in the movement detection task for S2 show that the most important electrodes are found in sensorimotor regions. The colormap represents the change in accuracy of the model due to ablation, compared to the intact, original model. The multimodal model is generally more resilient to single electrode ablations.

Table 4.5: 1D convolutional filters vs. 3D convolutional filters

	1D Conv			3D Conv		
	Pred-b	Pred	Det	Pred-b	Pred	Det
S1	58.4	60.6	66.2	51.2	45.4	62.6
S2	63.4	65.9	69.9	57.5	64.5	63.8
S3	64.8	66.3	79.8	66.6	70.7	77.9
S4	52.6	55.7	65.1	47.4	48.4	55.7
Average	**59.8**	**62.1**	**70.3**	55.7	57.3	65.0

Table 4.6: ECoG deep model vs. SVM with spectral features

	Deep			Traditional		
	Pred-b	Pred	Det	Pred-b	Pred	Det
S1	58.4	60.6	66.2	49.5	52.6	63.2
S2	63.4	65.9	69.9	62.7	68.8	67.4
S3	64.8	66.3	79.8	50.2	50.2	62.4
S4	52.6	55.7	65.1	53.1	51.0	50.0
Average	**59.8**	**62.1**	**70.3**	53.9	55.7	60.8

Fig. 4.3 shows ablation maps for one subject (S2) for the detection task, comparing the resilience of the ECoG only model and the multimodal model. The most important electrode for this subject is in the cortical area corresponding to sensorimotor function. Ablation maps for all subjects and experiments are shown in Fig. S2–S5. The ablation analysis revealed that the most important electrodes for the ECoG only model were those in sensorimotor regions, prefrontal regions (implicated in motor planning), and speech regions (e.g. Broca's and Wernicke's areas), likely explained by co-occurrence of speech and movement in this natural dataset.

Our multimodal model, on the other hand, was less impacted by ablation of single electrodes

when compared to ECoG only (Fig. 4.3 and Fig. S2–S5). Table 4.7 shows detailed comparisons of intact vs. ablation experiments. In the worst-case single-electrode ablated experiments, the decrease in accuracy of ECoG model after ablation is always larger than the decrease for the multimodal model except in one case.

In an additional experiment, we ablated all electrodes in the multimodal model, and the accuracy dropped to chance levels. This all-ablation experiment confirms that the multimodal network was not simply ignoring the ECoG input. Instead, the video was able to alleviate dependence on individual electrodes, resulting in a more robust multimodal model.

Table 4.7: Ablation resilience of ECoG only vs. multimodal. **Bold** indicates the model with the higher accuracy post-ablation.

	S1			S2			S3			S4		
	Orig	Ablate	Diff	Orig	Ablate	Diff	Orig	Ablate	Diff	Orig	Ablate	Diff
Detect												
ECoG only	66.4	60.7	5.7	64.6	**62.0**	2.6	83.2	56.3	26.9	57.3	53.1	4.2
Multimodal	87.0	**86.2**	0.8	61.5	59.7	1.8	79.8	**78.0**	1.8	62.5	**59.9**	2.6
Pred												
ECoG only	58.6	55.8	2.8	65.7	55.6	10.1	68.3	58.7	9.6	54.7	49.0	5.7
Multimodal	81.9	**78.6**	3.3	61.7	**57.5**	4.2	71.2	**68.2**	3.0	62.0	**57.8**	4.2
Pred-b												
ECoG only	66.8	63.0	3.8	66.7	**57.3**	9.4	66.1	61.2	4.9	62.0	41.7	15.1
Multimodal	76.8	**74.1**	2.7	59.4	56.3	3.1	67.9	**63.9**	4.0	66.1	**62.5**	3.6

4.0.9 Implementation details

We implemented our networks in Tensorflow [81] with the Keras [82] module. We used a stochastic gradient descent (sgd) optimizer with a learning rate of 0.001, momentum term of 0.9, and decay factor of 0.9 for all experiments. The batch size was 24 in order to ensure that the multimodal

network would fit in memory. The initial weights were generated by the glorot uniform distribution. After every convolution, we used rectified linear activation units (ReLU). Dropout of 0.5 was applied after each fully connected layer. Each network was trained for 200 iterations with an early stopping criteria. Each training procedure was run three times, because on some runs, poor initial weights led to very poor final results. The final model was selected as the model with the best accuracy out of these three runs, as assessed on a validation set randomly sampled from the training days. Optimization speed varied for different subject sets and experiments but typically took a few hours on a TESLA X (Pascal) GPU.

Discussion

In this paper, we introduced the AJILE dataset, which contains over 670 hours (over 72 million frames) of total continuous naturalistic human ECoG data with corresponding upper body joint locations. AJILE greatly surpasses in scope and size all previous datasets of neural recordings of human movements, allowing deep learning approaches to be applied to neural decoding problems. The dataset will be released to coincide with this paper's publication. We also presented the first model that successfully predicts future movement from natural human ECoG data. The ECoG-video multimodal deep neural network models show improved accuracy and robustness beyond using each modality alone.

The current work predicts the initiation of movement of the contralateral hand. This approach can be extended to predict and regress the locations of multiple joints for a more detailed reconstruction of future movements. Because of significant variation across individual subjects, we believe that

deep learning on large quantities of raw data is a more scalable and sustainable approach than models built on hand-crafted features.

4.0.10 Visualizing Filters

To investigate features extracted by layers of the neural network models, we used gradient based input optimization, visualizing ECoG inputs that maximally activated filters at various layers. Fig. 4.4 shows a few example filters from different parts of the neural network, and we make the general observation that features acquire more distinct structure at deeper layers, which is consistent with what has been described in the image realm. In addition, earlier convolutional neural network units tend to show a preference for distinct temporal frequencies in the signal, resembling features represented in a Fourier basis. Deeper network layers tend to prefer more complex temporal features, with dynamic frequencies across space and time.

4.0.11 Implications and connections to neuroscience

Our approach builds custom models tailored to individuals using only raw data, adapting to variations such as individual electrode placement without expert intervention. Dissecting these models revealed several observations that have direct connections to human neuroscience. First, the virtual electrode ablation studies revealed the most important electrode locations lie in an area of cortex known to be sensorimotor cortex. Second, the convolutional filters learned in the ECoG tower have features similar to Fourier bases, which are by far the most common approach to analyzing ECoG data in neuroscience. Moreover, our deep neural networks, especially at deeper layers, learn more complex features that are dynamic in space and time, suggesting that the investigation of our models may

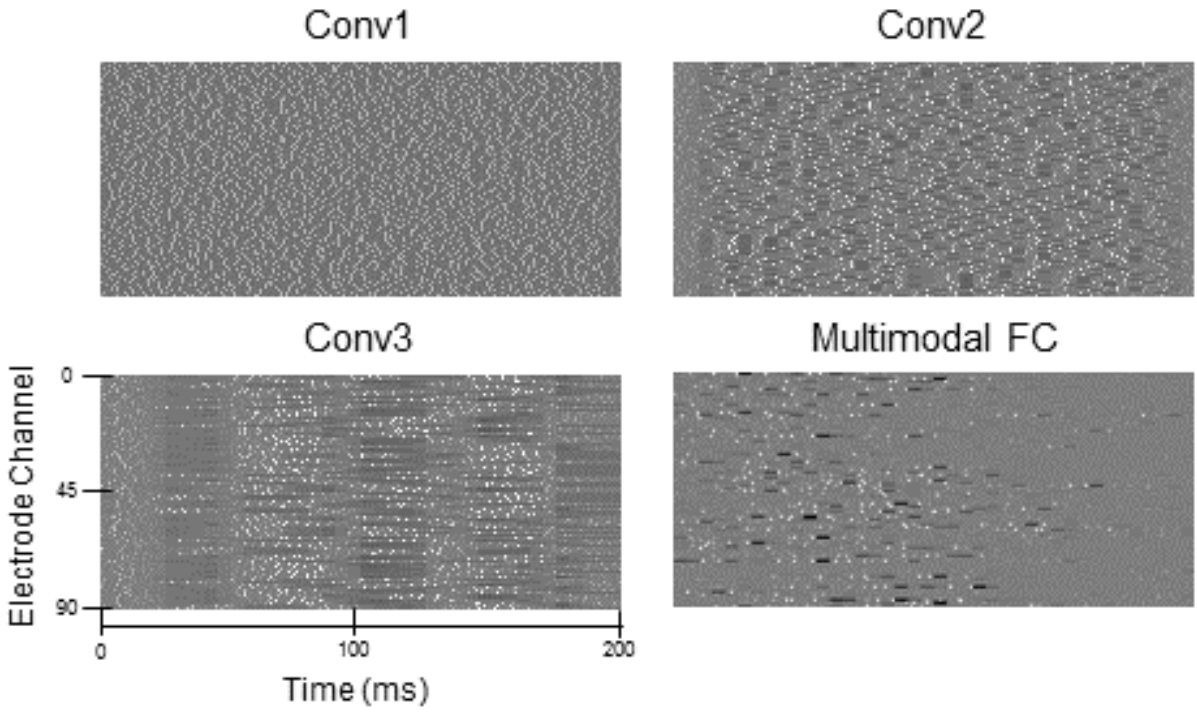


Figure 4.4: Conv1, Conv2 and Conv3 represent input that maximizes activation for sample units across the first three convolutional layers in S1 ECoG only model. Multimodal FC is computed from the S1 multimodal network's first FC. The general pattern is that deeper layers of the network have more unique maximally desired inputs in electrode space and time.

uncover novel and surprising patterns in neural activity underlying naturalistic movements.

4.0.12 Acknowledgements

We thank Maya Felten and Ryan Shean for annotation. We also thank Dr. Jeffery Ojemann and Nile Wilson for aiding in data collection as well as research discussions. This research was support by the Washington Research Foundation (WRF), Moore Foundation, Sloan Foundation and the National Science Foundation (NSF) award 1630178 and EEC-1028725.

Supplemental Figures

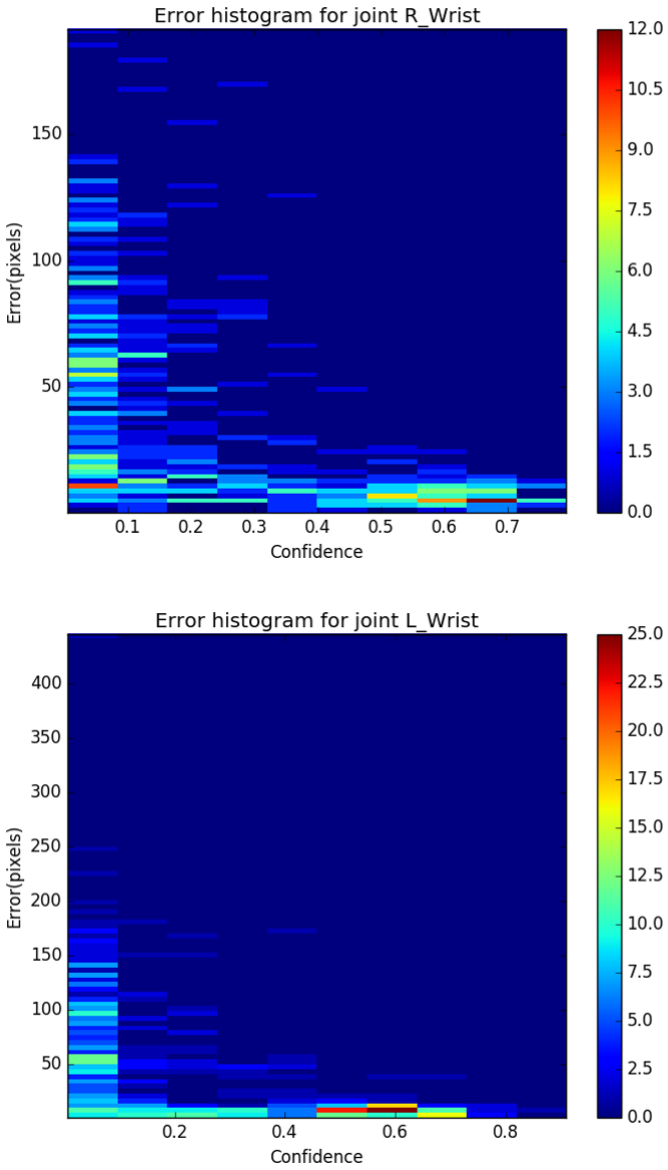


Figure S1: Histograms of accuracy on test data from multiple subjects (not necessarily including subjects in this study) show that when confidence is above 0.25, almost all test instances are quite accurate. For scale, a wrist on the video is approximately 25 pixels wide. Top: Right wrist. Bottom: Left wrist.

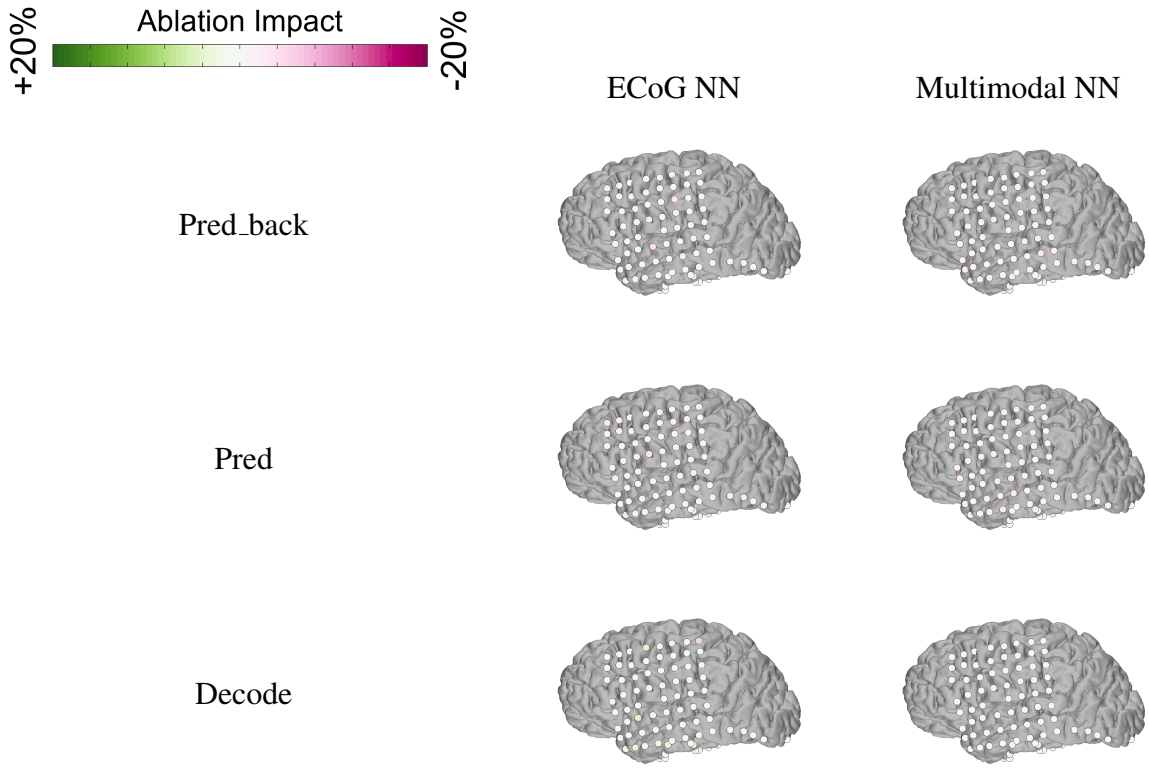


Figure S2: Subject 1 ablation map. Ablation impact references the percentage increase or decrease in accuracy after the particular electrode signal is replaced with its mean over time.

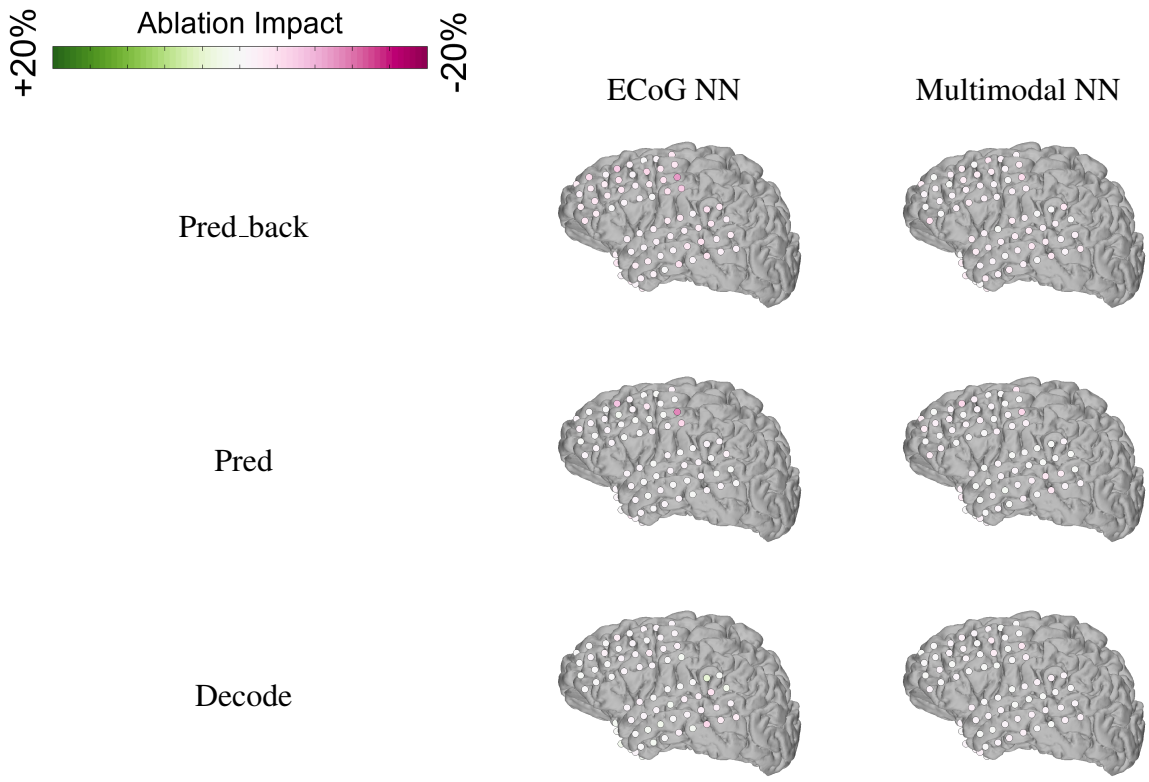


Figure S3: Subject 2 ablation map. Ablation impact references the percentage increase or decrease in accuracy after the particular electrode signal is replaced with its mean over time.

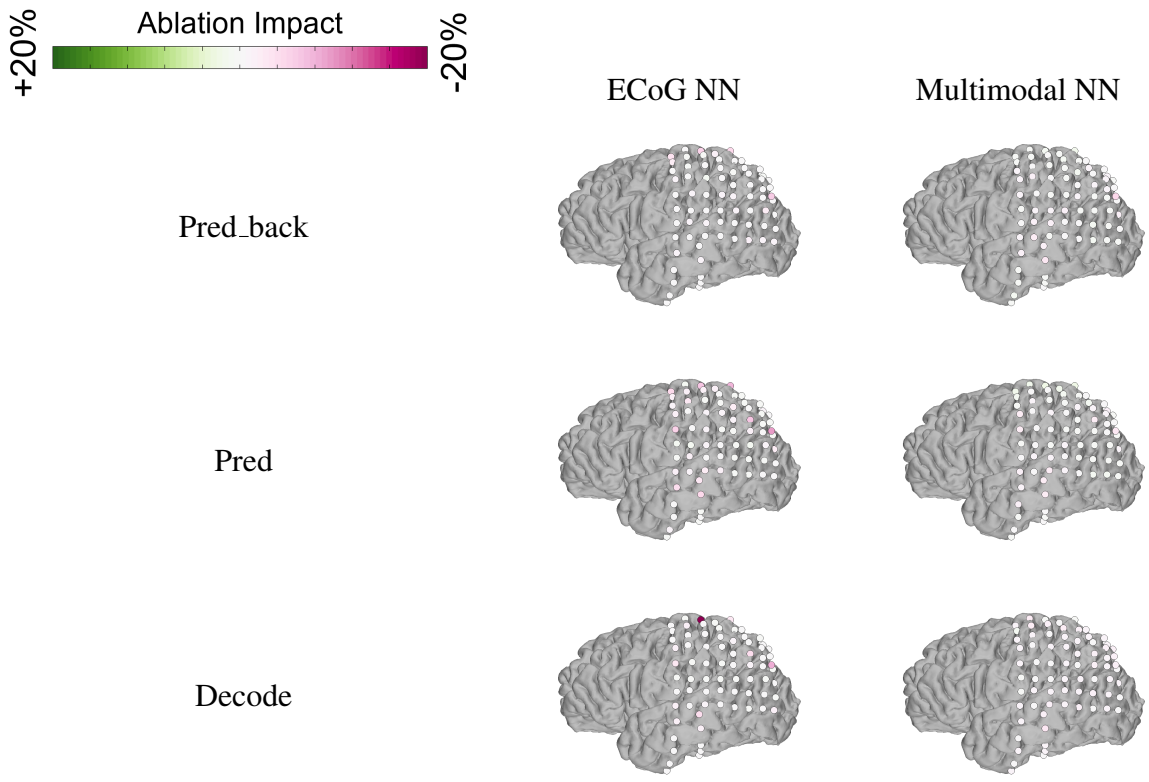


Figure S4: Subject 3 ablation map. Ablation impact references the percentage increase or decrease in accuracy after the particular electrode signal is replaced with its mean over time.

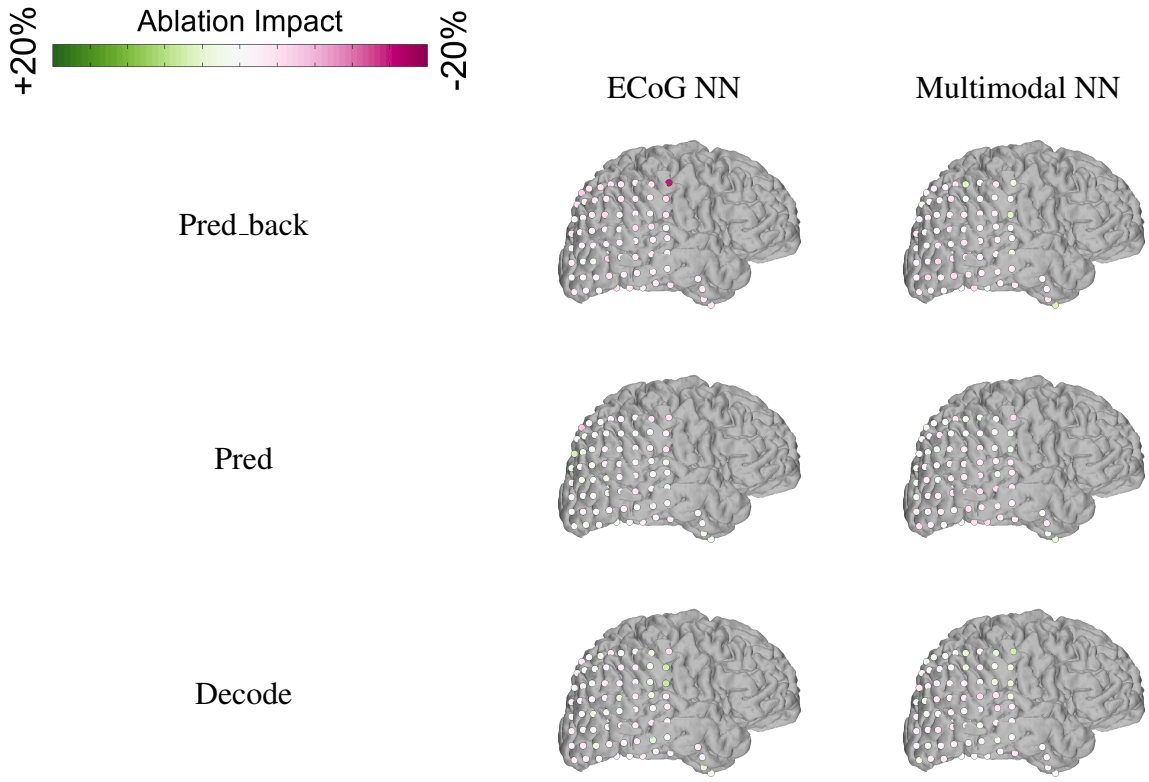


Figure S5: Subject 4 ablation map. Ablation impact references the percentage increase or decrease in accuracy after the particular electrode signal is replaced with its mean over time.

Chapter 5

MIND THE GAP: IMPUTING MISSING NEURAL DATA TO FUSE HETEROGENEOUS MULTI-SUBJECT ELECTROCORTICOGRAPH

Nancy X. R. Wang Steven M. Peterson Rajesh P. N. Rao Bingni W. Brunton

The recent success of deep neural networks in solving previously intractable problems, including those in computer vision and natural language processing, has hinged in part on the sheer size of the available training data. In human neuroscience, data size is typically limited because it is often difficult to combine data from different subjects, due to variability in individual anatomy, neural function, and electrode placement. Training models on subject-specific data decreases the power of the overall dataset, also rendering the models unable to generalize to unseen subjects. This problem is exacerbated for electrocorticography (ECoG), or intracranial EEG, where electrodes are implanted at locations according to clinical needs and the sensor grids may differ in size and location. Here, we develop an approach to fuse heterogenous multi-subject ECoG data using a common, virtual electrode grid such that models trained on one set of subjects can be directly transferred to unseen subjects. In addition, we develop a deep learning approach to impute missing neural data in the virtual grid, filling in gaps where not all subjects had actual recording electrodes. We demonstrate our transfer learning approach on the problem of predicting future movements from naturalistic neural recordings in ou Annotate Joints in Long-Term ECoG (AJILE) dataset [68]. We compare three imputation methods: zero filling, interpolation, and deep learning for imputation.

To our knowledge, this is the first demonstration of transfer learning using a deep neural network trained on multi-subjects ECoG data. By training on merged, heterogeneous data from multiple subjects, we can decode completely unseen subjects without retraining.

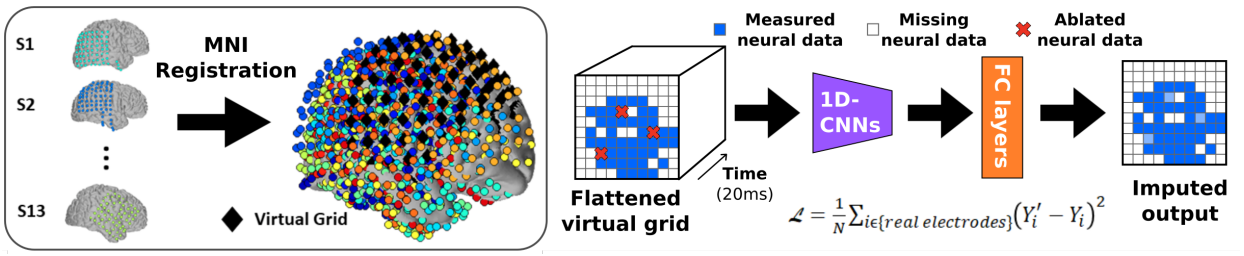


Figure 5.1: (Left) Actual electrodes from 13 different subjects were registered using the standardized MNI human brain coordinate system. The black diamonds represent the 10×10 virtual electrode grid. We develop and train a deep neural network (Right) to impute the missing electrodes. The training involved artificially ablating random electrodes and calculating the loss $\mathcal{L}$ as the reconstruction error of voltage at all virtual electrodes Y' compared to measured neural data Y at the last time point.

Introduction

Modern advanced brain-computer interfaces (BCIs) typically rely on decoders tuned to single subjects, and it is unclear how such models would perform on new, unseen subjects. Training individualized models is challenging because labeled data can be expensive to acquire, training each model may be time-consuming, and this approach is intractable when the subject cannot perform the tasks required for training (e.g., limited mobility due to stroke). Therefore, we are motivated to develop a modeling framework that can leverage large, heterogeneous multi-subject data and generalize to unseen subjects. Deep learning approaches excel at learning features and patterns from large sets of data. This helps remove the need for expertly handcrafted features and can help generalize a model to multiple subjects and even multiple tasks [83]. However, the input to the neural network needs to have the same input size among training samples and any test sample. In

this case, since each subject's electrodes are implanted in different areas and configurations, we cannot train the data from different patients together, nor can we hope to apply any trained model to a new unseen subject. Thus, to fully leverage a neural network model, we need to fill in the gap between subjects' electrodes in order to combine the data together. In our work, we investigate several methods for filling in this gap, including another neural network approach.

Related Works

In ECoG, the electrode placements can vary significantly from subject to subject (see Figure 5.1) as they are placed surgically according to clinical needs. The successes in combining subjects in EEG [84] and fMRI [85] for multi-subject training rely on the similarity of input space from subject to subject. Combining ECoG data from multiple subjects, on the other hand, results in large gaps in input space, where data is missing due to non-overlapping electrode placements. There exists prior work on imputation for time-series data, but the work has focused on when data in some channels are missing only some of the time [86]. Here, we have entirely missing information at locations where an electrode was never implanted.

Methods

In this work, we impute missing electrode values by training a deep neural network (Fig 5.1) with a large ECoG dataset (16 total hours of data segments randomly extracted from 52 days of recordings for 13 subjects). During training, we randomly ablate channels to force the network to learn to reconstruct from missing channels. However, the loss is calculated as the mean squared error of the reconstructed signal from all existing electrodes, not just the ablated ones as this was found to have better quality reconstructions. Fig 5.3.B shows the increase in reconstruction quality with a larger

ECoG Dataset. We assume that there are enough commonalities across different brains that we can project each subject's data onto a standard, virtual electrode grid, with missing values imputed. To demonstrate the effectiveness of this virtual grid with imputed neural recordings, we train another multi-subject deep neural network model on this common input space for predicting future movements (see [68] for network architecture). We compared the deep learning-based imputation approach with two other imputation schemes: 1) zero filling, where we fill in values for all missing electrodes in the virtual grid with zeros and 2) interpolation, where we fill in missing values with the values from the closest existing neighboring electrodes. We use cubic interpolation when nearby electrodes exist and we copy the nearest electrode value when cubic interpolation cannot be conducted. To further increase the size of our dataset, we projected left-side grids to our right sided virtual grid with the assumption that neural activity for contralateral arm movement is similar on both sides of the brain.

We purposely withheld a few electrodes from each subject completely during training. This allows us to assess the quality of reconstruction on unseen channels, as the goal of our imputation task to to infer the signals from these completely unseen brain regions. Figure 5.2 shows that the reconstructed signals from imputation and interpolation methods look qualitatively similar to the ground truth. Quantitatively speaking, the imputation with all 13 subjects has the best mean squared reconstruction error for the withheld channels. However, the interpolation method has the best correlation for both the time series and band power.

We trained the multi-subject neural network to predict natural future wrist movements, a task that is particularly challenging because the movement events are spontaneous and occurred over

	Impute W/ All	Impute W/ AJILE	Interpolate
Correlation	0.428	0.298	**0.503**
Band Power Correlation	0.678	0.640	**0.700**
Mean Squared Error	**0.016**	0.0275	0.0239

Table 5.1: Correlation and reconstruction error of various methods over withheld channels from AJILE subject validation data. The methods are deep imputation trained with all subjects, deep imputation trained with just the AJILE subjects and using cubic interpolation. Results show that the deep imputation with all subjects has the smallest reconstruction error but the interpolation has the best correlation across time and band power correlation to the ground truth.

several days. We extracted these movement events from arm poses labeled automatically from video over at least 4 continuous days of recording for each subject (see details in [68]). In addition, all accuracy results we report are tested on days of recording not used in the training data, including when testing on the same subject.

Results

Using our imputation model, we can predict new unseen patients' future movements above chance (Fig. 5.3). Imputation generalizes to unseen patients, whereas zero-filling does not perform as well above chance. Currently, our deep imputation framework does not perform significantly better than data interpolation. Our ongoing work is exploring different deep imputation architectures and loss functions to improve its performance. Our results suggest that using an imputed virtual grid is a promising approach to train deep learning models that can leverage the power of multi-subject datasets, allowing analysis on new unseen subjects without new training data.

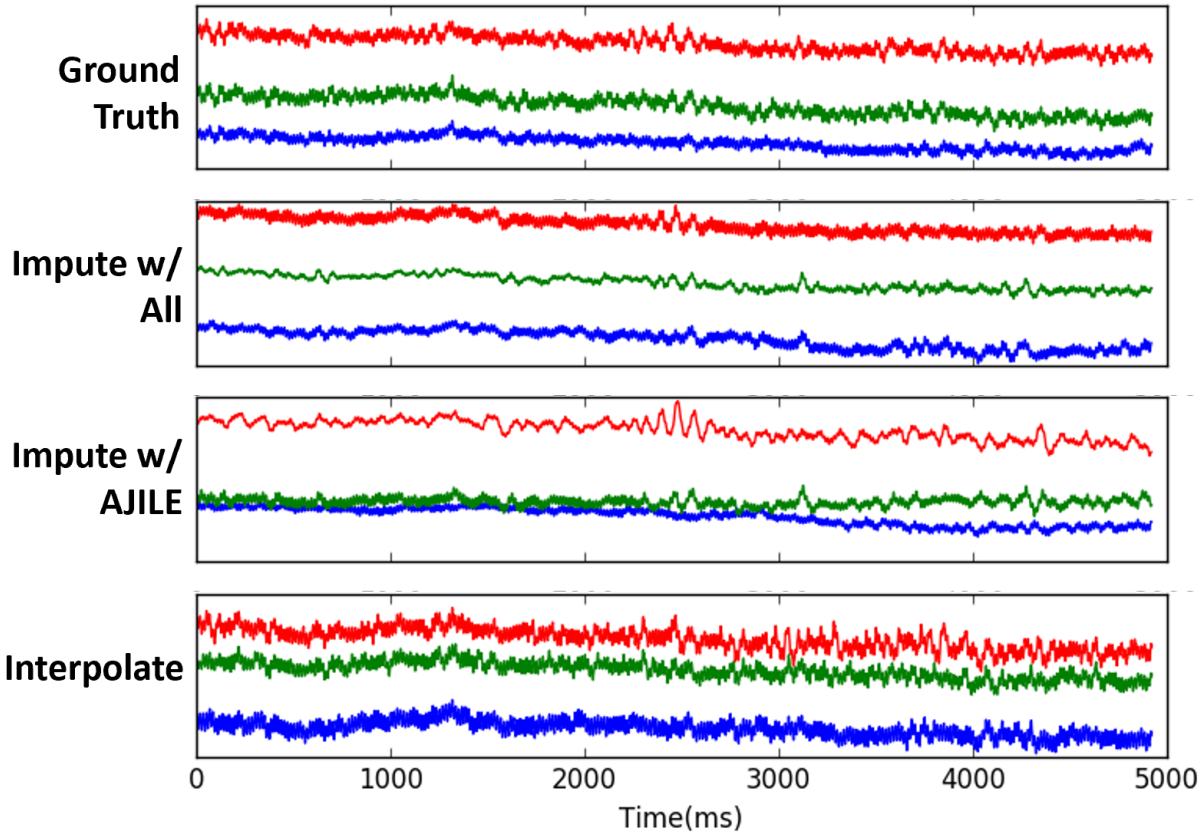


Figure 5.2: Sample snippet from an AJILE subject showing reconstruction quality of three withheld channels using deep imputation trained with all subjects, deep imputation trained with just the AJILE subjects and using cubic interpolation.

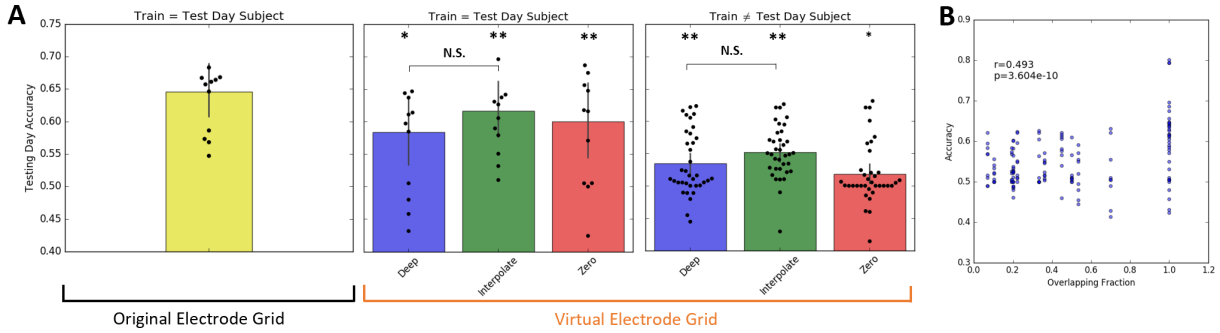


Figure 5.3: Test accuracy of models evaluated on events extract on a test day with-held from the training data, either from the same subject or a different subject. A: Movement prediction performance using the virtual electrode grid is comparable to the original patient electrode grid. Deep learning and interpolate imputation schemes generalize to new unseen patients with test accuracy is doubly significantly above chance (p¡0.01) while zero-filling is less significantly above chance at (p¡0.05). There is no significant difference between our current deep imputation model and interpolation. B: There is a correlation between test accuracy and the number of overlapping electrodes. The electrode overlap is calculated from the number of co-existing virtual electrodes between the train and test subject. In cases where the training and testing subject are the same, the overlapping fraction is 1.

Chapter 6

DISCUSSION AND CONCLUSION

The state of the art in neural recordings vary drastically in spatial and temporal resolution, invasiveness, and long-term monitoring capabilities. In addition, recording devices are at varying stages of approval for animal and human clinical use. Devices that detect hemodynamic changes in the brain, such as functional Magnetic Resonance Imaging (fMRI) and functional Near-Infrared Spectroscopy (fNIRS), in general have a few seconds of delay between neuronal firings and detection. This eliminates these methods for fast real-time monitoring and control. Other non-invasive methods, such as electroencephalograms (EEG), suffer from large motor artifacts due the distance of electrodes from the neural activity that it is attempting to measure. There is also a large amount of high-frequency signal attenuation by the skull that further decreases the signal-to-noise ratio in EEG signals. On the other end of the invasiveness spectrum, local field potential electrodes or single neuron recordings that penetrate through the cortex of the brain have been shown to cause some neural damage and scarring of tissues, which can render the electrode recordings to be no longer useful. In addition, the coverage that such electrodes provides is quite sparse throughout the brain.

There also exists less invasive options such as ECoG and Stentrodes. Currently, Stentrodes have not been approved in human studies whereas ECoG is already used in clinical procedures, like

long-term epilepsy monitoring. This offers a great source of data where techniques for long-term noisy data can be developed, which can then be applied to future electrophysiological devices, like Stentrodes. In order for ECoG to be the recording mechanism of choice outside of clinical situations however, the electrodes must be made wireless. Currently, patients cannot move very far out of their bed because all electrodes are tethered to wire connections into the wall. Thankfully, groups such as Xie and colleagues are already developing implantable, portable ECoG devices that are meant for human use ([87]). Although they have only been demonstrated on animal models so far, it should not be too long before wireless ECoG devices are approved for clinical use. In parallel to the hardware innovations, we should be improving our analysis techniques in order to handle longer recordings (in the order of years) in natural situations which may also have more channels added on as the technology improves.

Partnerships between industry and academia are crucial in enabling useful brain-computer interfaces that a user can take home. Currently, companies are developing state of the art computation hardware (ex. Nvidia GPUs) and hardware that are low-power (ex. IBM TrueNorth). Companies have recently leveraged new machine learning breakthroughs discovered in academia for many computer vision and speech recognition applications. In many instances, models developed in industry have surpassed ones from academia partly due to the sheer amount of data private companies are able to collect. However, in bio-medical applications, academia is still arguably the best investigator. Clinical data is highly personal and private. Patients often do not want their private data in the hands of for-profit organizations. When a useful product has been developed in academia, patients may be more willing to contribute their data to a company that can make a specific bio-medical device that

can improve theirs or someone else's quality of life.

Thus far, few studies in human neural decoding have attempted to combine data from multiple patients. This is due to natural variations between brain structure and function between people and potentially different recording configurations, as is true in the case of different ECoG implantation locations. Unfortunately, techniques that require hand-picked features necessitate some manual tuning of parameters or electrode selection. This is not a scalable approach. Instead, end-to-end systems with deep learning models can train and perform on completely raw data without any manual tuning. In fact, such systems can also learn individual models that are best suited for each subject completely automatically. They can be pretrained on pooled subject data and then fine-tuned with individual patients using a small amount of data when a user first starts the device. The algorithm can also continuously adapt to the user over time. Additionally, to adapt to each individual patient, the algorithm needs to be able to do so without any labelled data. By leveraging automated techniques in other modalities, such as audio and vision, one can automatically annotate and train on patient behaviour in order to decode and predict future intentions for BCI operation.

BIBLIOGRAPHY

[1] Z. C. Chao, Y. Nagasaka, and N. Fujii. "Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkey". In: *Frontiers in neuroengineering* 3 (2010), p. 3.

[2] G. Hotson, D. P. McMullen, M. S. Fifer, et al. "Individual finger control of a modular prosthetic limb using high-density electrocorticography in a human subject". In: *Journal of Neural Engineering* 13.2 (Apr. 2016), p. 026017. ISSN: 1741-2560. DOI: 10.1088/1741-2560/13/2/026017. URL: http://stacks.iop.org/1741-2552/13/i=2/a=026017?key=crossref.2d5ef8bc47143308a3a1ff00aabb53dd.

[3] M. Völker, J. Hammer, R. T. Schirrmeister, J. Behncke, L. D. Fiederer, A. Schulze-Bonhage, P. Marusič, W. Burgard, and T. Ball. "Intracranial Error Detection via Deep Learning". In: *arXiv preprint arXiv:1805.01667* (2018).

[4] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball. "Deep learning with convolutional neural networks for EEG decoding and visualization". In: *Human Brain Mapping* (Aug. 2017). ISSN: 10659471. DOI: 10.1002/hbm.23730. URL: http://www.ncbi.nlm.nih.gov/pubmed/28782865%20http://doi.wiley.com/10.1002/hbm.23730.

[5] Y. R. Tabar and U. Halici. "A novel deep learning approach for classification of EEG motor imagery signals". In: *Journal of Neural Engineering* 14.1 (Feb. 2017), p. 016003. ISSN: 1741-2560. DOI: 10.1088/1741-2560/14/1/016003. URL: http://stacks.iop.org/1741-2552/14/i=1/a=016003?key=crossref.2a406e975c74b8476f2c258fdd5ac841.

[6] E. Nurse, B. S. Mashford, A. J. Yepes, I. Kiral-Kornek, S. Harrer, and D. R. Freestone. "Decoding EEG and LFP Signals using Deep Learning: Heading TrueNorth". In: (2016). DOI: 10.1145/2903150.2903159. URL: http://dx.doi.org/10.1145/2903150.2903159.

[7] N. Wang, J. Olson, J. Ojemann, R. Rao, and B. Brunton. "Unsupervised decoding of long-term, naturalistic human neural recordings with automated video and audio annotations". In: *Front Human Neurosci* 10 (2016).

[8] J. Wolpaw and E. Wolpaw. *Brain-computer interfaces: principles and practice*. London: Oxford University Press, 2012.

[9] R. Rao. *Brain-Computer Interfacing: An Introduction*. New York, NY, USA: Cambridge University Press, 2013. ISBN: 1139032801, 9781139032803.

[10] W. Vinje and J. Gallant. "Sparse coding and decorrelation in primary visual cortex during natural vision". In: *Science* (2000). URL: http://www.sciencemag.org/content/287/5456/1273.short.

[11] G. Felsen and Y. Dan. “A natural approach to studying vision.” In: *Nat Neurosci* 8.12 (2005), pp. 1643–6. ISSN: 1097-6256. DOI: 10.1038/nn1608.

[12] A. Jackson, J. Mavoori, and E. Fetz. “Correlations between the same motor cortex cells and arm muscles during a trained task, free behavior, and natural sleep in the macaque monkey”. In: *J Neurophys* (2007). URL: http://jn.physiology.org/content/97/1/360.short.

[13] J. Derix, O. Iljina, A. Schulze-Bonhage, A. Aertsen, and T. Ball. ““Doctor” or “darling”? Decoding the communication partner from ECoG of the anterior temporal lobe during non-experimental, real-life social interaction.” In: *Front Hum Neurosci* (2012), p. 251. ISSN: 1662-5161.

[14] T. Pistohl, A. Schulze-Bonhage, and A. Aertsen. “Decoding natural grasp types from human ECoG”. In: *Neuroimage* (2012). URL: http://www.sciencedirect.com/science/article/pii/S105381191100749X.

[15] J. Ruescher, O. Iljina, D. Altenmüller, A. Aertsen, A. Schulze-Bonhage, and T. Ball. “Somatotopic mapping of natural upper- and lower-extremity movements and speech production with high gamma electrocorticography.” In: *NeuroImage* 81 (2013), pp. 164–77. ISSN: 1095-9572.

[16] N. Hill, D. Gupta, and P. Brunner. “Recording human electrocorticographic (ECoG) signals for neuroscientific research and real-time functional cortical mapping”. In: *Journal of Visualized Experiments* (2012). URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3471287/.

[17] D. Moran. "Evolution of brain-computer interface: action potentials, local field potentials and electrocorticograms." In: *Curr Opin Neurobiol* 20.6 (Dec. 2010), pp. 741–5. ISSN: 1873-6882. URL: http://www.sciencedirect.com/science/article/pii/S0959438810001789.

[18] J. Williams and J. Hippensteel. "Complex impedance spectroscopy for monitoring tissue responses to inserted neural implants". In: *J Neural Eng* (2007). URL: http://iopscience.iop.org/1741-2552/4/4/007.

[19] T. Ball, M. Kern, I. Mutschler, A. Aertsen, and A. Schulze-Bonhage. "Signal quality of simultaneously recorded invasive and non-invasive EEG." In: *NeuroImage* 46.3 (2009), pp. 708–16. ISSN: 1095-9572. URL: http://www.sciencedirect.com/science/article/pii/S1053811909001827.

[20] Y. Nakanishi, T. Yanagisawa, D. Shin, et al. "Prediction of three-dimensional arm trajectories based on ECoG signals recorded from human sensorimotor cortex." In: *PloS one* 8.8 (2013), e72085. ISSN: 1932-6203.

[21] Z. Wang, A. Gunduz, P. Brunner, A. Ritaccio, Q. Ji, and G. Schalk. "Decoding onset and direction of movements using Electrocorticographic (ECoG) signals in humans." In: *Front in Neuroeng* 5 (2012), p. 15. ISSN: 1662-6443.

[22] P. Wang, E. Puttock, C. E. King, et al. "State and trajectory decoding of upper extremity movements from electrocorticogram". In: *NER*. 2013, pp. 969–972.

[23] K. Miller, S. Zanos, E. Fetz, M. den Nijs, and J. Ojemann. "Decoupling the cortical power spectrum reveals real-time representation of individual finger movements in humans." In: *J Neurosci* 29.10 (Mar. 2009), pp. 3132–7. ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.5506-08.2009. URL: http://www.jneurosci.org/content/29/10/3132.abstract.

[24] Z. Wang, Q. Ji, K. Miller, and G. Schalk. "Decoding Finger Flexion from Electrocorticographic Signals Using a Sparse Gaussian Process". In: *ICPR*. 2010, pp. 3756–3759. ISBN: 978-1-4244-7542-1.

[25] D. McMullen, G. Hotson, K. Katyal, et al. "Demonstration of a semi-autonomous hybrid brain-machine interface using human intracranial EEG, eye tracking, and computer vision to control a robotic upper limb prosthetic." In: *IEEE T Neur Sys Reh* 22.4 (July 2014), pp. 784–96. ISSN: 1558-0210. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4057363%5C&tool=pmcentrez%5C&rendertype=abstract.

[26] T. Yanagisawa, M. Hirata, Y. Saitoh, T. Goto, H. Kishima, R. Fukuma, H. Yokoi, Y. Kamitani, and T. Yoshimine. "Real-time control of a prosthetic hand using human electrocorticography signals." In: *J Neurosurg* 114.6 (2011), pp. 1715–22. ISSN: 1933-0693. DOI: 10.3171/2011.1.JNS101421. URL: http://www.ncbi.nlm.nih.gov/pubmed/21314273.

[27] M. Fifer, G. Hotson, B. Wester, et al. “Simultaneous neural control of simple reaching and grasping with the modular prosthetic limb using intracranial EEG.” In: *IEEE T Neur Sys Reh* 22.3 (2014), pp. 695–705. ISSN: 1558-0210.

[28] W. Wang, J. Collinger, A. Degenhart, et al. “An electrocorticographic brain interface in an individual with tetraplegia.” In: *PloS One* 8.2 (2013), e55344. ISSN: 1932-6203.

[29] E. Leuthardt, C. Gaona, M. Sharma, N. Szrama, J. Roland, Z. Freudenberg, J. Solis, J. Breshears, and G. Schalk. “Using the electrocorticographic speech network to control a brain-computer interface in humans.” In: *J Neural Eng* 8.3 (2011), p. 036004. ISSN: 1741-2552. DOI: 10.1088/1741-2560/8/3/036004. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3701859%5C&tool=pmcentrez%5C&rendertype=abstract.

[30] E. Leuthardt, K. Miller, G. Schalk, R. Rao, and J. Ojemann. “Electrocorticography-Based Brain Computer Interface - The Seattle Experience”. In: *IEEE T Neur Sys Reh* 14.2 (2006), pp. 194–198. ISSN: 1534-4320.

[31] K. Miller, G. Schalk, E. Fetz, M. den Nijs, J. Ojemann, and R. Rao. “Cortical activity during motor execution, motor imagery, and imagery-based online feedback.” In: *PNAS* 107.9 (2010), pp. 4430–5. ISSN: 1091-6490. DOI: 10.1073/pnas.0913697107.

[32] G. Schalk, K. J. Miller, N. R. Anderson, J. A. Wilson, M. D. Smyth, J. G. Ojemann, D. W. Moran, J. R. Wolpaw, and E. C. Leuthardt. “Two-dimensional movement control using electrocorticographic signals in humans.” In: *J Neural Eng* 5.1 (Mar. 2008), pp. 75–84.

ISSN: 1741-2560. DOI: 10.1088/1741-2560/5/1/008. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2744037%5C&tool=pmcentrez%5C&rendertype=abstract.

[33] M. Vansteensel, D. Hermes, E. Aarnoutse, M. Bleichner, G. Schalk, P. van Rijen, F. Leijten, and N. Ramsey. "Brain–computer interfacing based on cognitive control". In: *Ann Neurol* 67.6 (2010), pp. 809–816.

[34] V. G. Kanas, I. Mporas, H. L. Benz, K. N. Sgarbas, A. Bezerianos, and N. E. Crone. "Real-time voice activity detection for ECoG-based speech brain machine interfaces". In: *2014 19th International Conference on Digital Signal Processing*. Aug. 2014, pp. 862–865. ISBN: 978-1-4799-4612-9. DOI: 10.1109/ICDSP.2014.6900790. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6900790.

[35] T. Blakely, K. J. Miller, R. P. Rao, M. D. Holmes, and J. G. Ojemann. "Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids". In: *Conf Proc IEEE Eng Med Biol Soc.* (2008), pp. 4964–7. URL: http://www.ncbi.nlm.nih.gov/pubmed/19163831.

[36] V. G. Kanas, I. Mporas, H. L. Benz, K. N. Sgarbas, A. Bezerianos, and N. E. Crone. "Joint spatial-spectral feature space clustering for speech activity detection from ECoG signals." In: *IEEE Transactions on Bio-medical Engineering* 61.4 (Apr. 2014), pp. 1241–50. ISSN: 1558-2531. DOI: 10.1109/TBME.2014.2298897. URL: http://www.

pubmedcentral.nih.gov/articlerender.fcgi?artid=4005607%5C&tool=pmcentrez%5C&rendertype=abstract.

[37] 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." en. In: *J NEURAL ENG* 11.3 (June 2014), p. 035015. ISSN: 1741-2552. DOI: 10.1088/1741-2560/11/3/035015. URL: http://iopscience.iop.org/article/10.1088/1741-2560/11/3/035015.

[38] 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." In: *Journal of Neural Engineering* 8.4 (Aug. 2011), p. 046028. ISSN: 1741-2552. DOI: 10.1088/1741-2560/8/4/046028. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3772685%5C&tool=pmcentrez%5C&rendertype=abstract.

[39] 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." In: *Journal of Neural Engineering* 7.5 (Oct. 2010), p. 056007. ISSN: 1741-2552. DOI: 10.1088/1741-2560/7/5/056007. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2970568%5C&tool=pmcentrez%5C&rendertype=abstract.

[40] D. Zhang, E. Gong, W. Wu, J. Lin, W. Zhou, and B. Hong. "Spoken sentences decoding based on intracranial high gamma response using dynamic time warping." In: *Annual International*

Conference of the IEEE Engineering in Medicine and Biology Society 2012 (Jan. 2012), pp. 3292–5. ISSN: 1557-170X. DOI: 10.1109/EMBC.2012.6346668. URL: http://www.ncbi.nlm.nih.gov/pubmed/23366629.

[41] 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". English. In: *Frontiers in Neuroscience* 9 (June 2015). ISSN: 1662-453X. DOI: 10.3389/fnins.2015.00217. URL: http://journal.frontiersin.org/article/10.3389/fnins.2015.00217/abstract.

[42] T. O. Zander and C. Kothe. "Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general." en. In: *J of Neural Eng* 8.2 (Apr. 2011), p. 025005. ISSN: 1741-2552. DOI: 10.1088/1741-2560/8/2/025005. URL: http://iopscience.iop.org/article/10.1088/1741-2560/8/2/025005.

[43] G. Muller-Putz, R. Leeb, M. Tangermann, et al. "Towards Noninvasive Hybrid Brain–Computer Interfaces: Framework, Practice, Clinical Application, and Beyond". In: (2015).

[44] J. Derix, O. Iljina, J. Weiske, A. Schulze-Bonhage, A. Aertsen, and T. Ball. "From speech to thought: the neuronal basis of cognitive units in non-experimental, real-life communication investigated using ECoG." In: *Front Hum Neurosci* 8 (2014), p. 383. ISSN: 1662-5161.

[45] M. Dastjerdi, M. Ozker, B. Foster, V. Rangarajan, and J. Parvizi. “Numerical processing in the human parietal cortex during experimental and natural conditions.” In: *Nat Comm* 4 (2013), p. 2528. ISSN: 2041-1723.

[46] M. Längkvist, L. Karlsson, and A. Loutfi. “Sleep stage classification using unsupervised feature learning”. In: *Advances in Artificial Neural Systems* (2012). URL: http://dl.acm.org/citation.cfm?id=2387786.

[47] T. Pluta, R. Bernardo, H. W. Shin, and D. R. Bernardo. “Unsupervised learning of electrocorticography motifs with binary descriptors of wavelet features and hierarchical clustering.” In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 2014 (Jan. 2014), pp. 2657–60. ISSN: 1557-170X. DOI: 10.1109/EMBC.2014.6944169. URL: http://www.ncbi.nlm.nih.gov/pubmed/25570537.

[48] M. I. Jordan and T. M. Mitchell. “Machine learning: Trends, perspectives, and prospects”. In: *Science* 349.6245 (July 2015), pp. 255–260. ISSN: 0036-8075. DOI: 10.1126/science.aaa8415. URL: http://www.sciencemag.org/content/349/6245/255.full.

[49] X. Huang, J. Baker, and R. Reddy. “A historical perspective of speech recognition”. In: *Communications of the ACM* 57.1 (Jan. 2014), pp. 94–103. ISSN: 00010782. DOI: 10.1145/2500887. URL: http://dl.acm.org/ft%5C_gateway.cfm?id=2500887%5C&type=html.

[50] N. Wang, S. Cullis-Suzuki, and A. Albu. "Automated Analysis of Wild Fish Behavior in a Natural Habitat". In: *EMR at ICMR* (2015). URL: http://dl.acm.org/citation.cfm?id=2764875.

[51] R. Poppe. "Vision-based human motion analysis: An overview". In: *Comput Vis Image Und* 108.1-2 (2007), pp. 4–18. ISSN: 10773142.

[52] A. Toshev and C. Szegedy. "DeepPose: Human Pose Estimation via Deep Neural Networks". In: *CVPR*. 2014, pp. 1653–1660. ISBN: 978-1-4799-5118-5. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909610.

[53] R. Girshick, J. Donahue, T. Darrell, and J. Malik. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. June 2014, pp. 580–587. ISBN: 978-1-4799-5118-5. DOI: 10.1109/CVPR.2014.81. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909475.

[54] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. "Scalable Object Detection Using Deep Neural Networks". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. June 2014, pp. 2155–2162. ISBN: 978-1-4799-5118-5. DOI: 10.1109/CVPR.2014.276. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909673.

[55] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. "Large-Scale Video Classification with Convolutional Neural Networks". In: *2014 IEEE Conference on*

Computer Vision and Pattern Recognition. June 2014, pp. 1725–1732. ISBN: 978-1-4799-5118-5. DOI: 10.1109/CVPR.2014.223. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909619.

[56] M. S. Ryoo and L. Matthies. "First-Person Activity Recognition: What Are They Doing to Me?" In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. June 2013, pp. 2730–2737. ISBN: 978-0-7695-4989-7. DOI: 10.1109/CVPR.2013.352. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6619196.

[57] K. He, X. Zhang, S. Ren, and J. Sun. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: (Feb. 2015). arXiv: 1502.01852. URL: http://arxiv.org/abs/1502.01852.

[58] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio. "Efficient voice activity detection algorithms using long-term speech information". In: *Speech Communication* 42.3-4 (Apr. 2004), pp. 271–287. ISSN: 01676393. DOI: 10.1016/j.specom.2003.10.002. URL: http://linkinghub.elsevier.com/retrieve/pii/S0167639303001201.

[59] K. J. Miller, P. Shenoy, J. W. Miller, R. P. Rao, J. G. Ojemann, et al. "Real-time functional brain mapping using electrocorticography". In: *Neuroimage* 37.2 (2007), pp. 504–507.

[60] C. Potes, P. Brunner, A. Gunduz, R. T. Knight, and G. Schalk. "Spatial and temporal relationships of electrocorticographic alpha and gamma activity during auditory processing".

In: *Neuroimage* 97 (2014), pp. 188–195. URL: http://www.sciencedirect.com/science/article/pii/S1053811914003140.

[61] E. F. Chang, J. W. Rieger, K. Johnson, M. S. Berger, N. M. Barbaro, and R. T. Knight. "Categorical speech representation in human superior temporal gyrus". In: *Nature Neuroscience* 13.11 (2010), pp. 1428–1432. URL: http://www.nature.com/neuro/journal/v13/n11/full/nn.2641.html.

[62] C. Cajochen, R. Foy, and D.-J. Dijk. "Frontal predominance of a relative increase in sleep delta and theta EEG activity after sleep loss in humans". In: *Sleep Res Online* 2.3 (1999), pp. 65–69. URL: http://www.ncbi.nlm.nih.gov/pubmed/11382884.

[63] B. W. Brunton, L. A. Johnson, J. G. Ojemann, and J. N. Kutz. "Extracting spatial-temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition". In: *Journal of Neuroscience Methods* (2015). URL: http://arxiv.org/abs/1409.5496.

[64] M. R. Mercier, S. Molholm, I. C. Fiebelkorn, J. S. Butler, T. H. Schwartz, and J. J. Foxe. "Neuro-oscillatory phase alignment drives speeded multisensory response times: An electro-corticographic investigation". In: *The Journal of Neuroscience* 35.22 (2015), pp. 8546–8557. URL: http://www.jneurosci.org/content/35/22/8546.abstract.

[65] M. Vansteensel, M. Bleichner, L. Dintzner, E. Aarnoutse, F. Leijten, D. Hermes, and N. Ramsey. "Task-free electrocorticography frequency mapping of the motor cortex". In: *Clinical*

Neurophysiology 124.6 (2013), pp. 1169–1174. URL: http://www.sciencedirect.com/science/article/pii/S1388245712008115.

[66] J. D. Breshears, C. M. Gaona, J. L. Roland, et al. "Mapping sensorimotor cortex using slow cortical potential resting-state networks while awake and under anesthesia". In: *Neurosurgery* 71.2 (2012), p. 305. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4362698/.

[67] M. Kipp. *Anvil: A universal video research tool. Handbook of Corpus Phonology*. 2012.

[68] N. X. R. Wang, A. Farhadi, R. Rao, and B. Brunton. "AJILE Movement Prediction: Multimodal Deep Learning for Natural Human Neural Recordings and Video". In: *AAAI*. 2018.

[69] L. Hochberg, D. Bacher, B. Jarosiewicz, et al. "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm". In: *Nature* 485.7398 (2012), pp. 372–375.

[70] P. Gabriel, W. Doyle, O. Devinsky, D. Friedman, T. Thesen, and V. Gilja. "Neural correlates to automatic behavior estimations from RGB-D video in epilepsy unit". In: *EMBC*. IEEE. 2016, pp. 3402–3405.

[71] T. Pfister, J. Charles, and A. Zisserman. "Flowing ConvNets for Human Pose Estimation in Videos". In: *ICCV*. 2015.

[72] D. Krug, C. Elger, and K. Lehnertz. "A CNN-based synchronization analysis for epileptic seizure prediction: Inter-and intraindividual generalization properties". In: *CNNA*. IEEE. 2008, pp. 92–95.

[73] Z. Wang, S. Lyu, G. Schalk, and Q. Ji. "Deep Feature Learning Using Target Priors with Applications in ECoG Signal Decoding for BCI." In: *IJCAI*. 2013.

[74] E. Nurse, B. Mashford, A. Yepes, I. Kiral-Kornek, S. Harrer, and D. Freestone. "Decoding EEG and LFP signals using deep learning: heading TrueNorth". In: *ACM Computing Frontiers*. ACM. 2016, pp. 259–266.

[75] R. Schirrmeister, J. Springenberg, L. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball. "Deep learning with convolutional neural networks for brain mapping and decoding of movement-related information from the human EEG". In: *arXiv preprint arXiv:1703.05051* (2017).

[76] J. Ren, Y. Hu, Y. Tai, C. W. L, Xu, W. Sun, and Q. Yan. "Look, Listen and Learn - A Multimodal LSTM for Speaker Identification". In: *AAAI*. 2016.

[77] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng. "Multimodal deep learning". In: *ICML*. 2011, pp. 689–696.

[78] Y. Aytar, C. Vondrick, and A. Torralba. "Soundnet: Learning sound representations from unlabeled video". In: *NIPS*. 2016, pp. 892–900.

[79] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. "You only look once: Unified, real-time object detection". In: *CVPR*. 2016, pp. 779–788.

[80] P. Shenoy, K. J. Miller, J. G. Ojemann, and R. P. Rao. "Generalized Features for Electrocorticographic BCIs". In: *IEEE T Bio-Med Eng* 55.1 (2008), pp. 273–280.

[81] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. URL: http://tensorflow.org/.

[82] F. Chollet et al. *Keras*. https://github.com/fchollet/keras. 2015.

[83] S. Ruder. "An overview of multi-task learning in deep neural networks". In: *arXiv preprint arXiv:1706.05098* (2017).

[84] M. Volker, R. T. Schirrmeister, L. D. J. Fiederer, W. Burgard, and T. Ball. "Deep transfer learning for error decoding from non-invasive EEG". In: *BCI*. IEEE, Jan. 2018, pp. 1–6. ISBN: 978-1-5386-2574-3. DOI: 10.1109/IWW-BCI.2018.8311491. URL: http://ieeexplore.ieee.org/document/8311491/.

[85] A. F. Marquand, M. Brammer, S. C. Williams, and O. M. Doyle. "Bayesian multi-task learning for decoding multi-subject neuroimaging data". In: *NeuroImage* 92 (May 2014), pp. 298–311. ISSN: 1053-8119. DOI: 10.1016/J.NEUROIMAGE.2014.02.008. URL: https://www.sciencedirect.com/science/article/pii/S1053811914000998.

[86] N. Jaques, S. Taylor, A. Sano, and R. Picard. "Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction". In: *ACII*. IEEE, Oct. 2017, pp. 202–208. ISBN: 978-1-5386-0563-9. DOI: 10.1109/ACII.2017.8273601. URL: http://ieeexplore.ieee.org/document/8273601/.

[87] K. Xie, S. Zhang, S. Dong, S. Li, C. Yu, K. Xu, W. Chen, et al. "Portable wireless electrocorticography system with a flexible microelectrodes array for epilepsy treatment". In: *Scientific Reports* 7 (2017).

Chapter 7

ADDENDUM: OTHER PROJECT SUMMARIES

7.1 Visualization of ECoG data

Naturalistic ECoG data is very long, on the order of hours and days. In addition, it is difficult to interpret the time series traces from each electrode by eye. In this project, I sought to come up with visualization tools for this type of data.

First, for passive visualization of how power spectrums change over time, power ratios between bands can be calculated at each timestep over some time window and plotted in a 2D plot (see Figure 7.1). The trace of the relationship between two power ratios can be viewed alongside the video of the patient in order to spot patterns correlating to behaviour as well as how stable the power ratios are locally across time.

A more interactive program was also prototyped. In this application, one can select time intervals of interest and it will automatically cluster small time windows within using spectral power. When a point in a cluster is clicked on, the corresponding video clip is played. As well, each subcluster can be further clustered. This program can help visualize what spectral power clusters correspond to which natural behaviour.

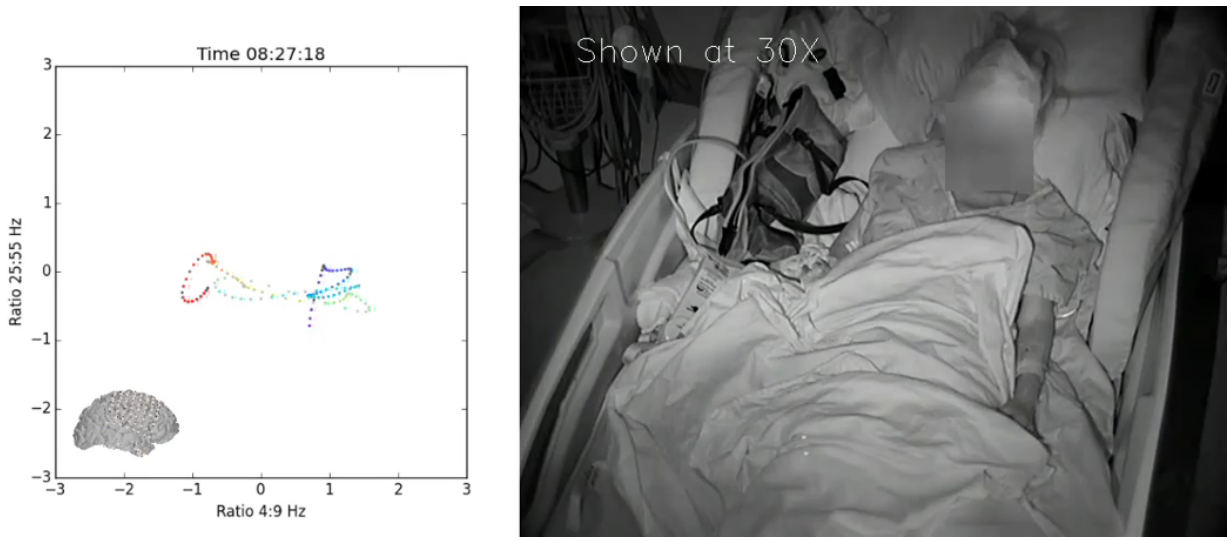


Figure 7.1: Screenshot of the visualization for viewing principal component changes in time alongside the patient video