DECODING NEURAL SIGNALS WITH A COMPACT AND INTERPRETABLE CONVOLUTIONAL NEURAL NETWORK

A PREPRINT

Arthur Petrosyan, Mikhail Lebedev and Alexei Ossadtchi

June 2, 2020

ABSTRACT

Brain-computer interfaces (BCIs) decode information from neural activity and send it to external devices. In recent years, we have seen an emergence of new algorithms for BCI decoding. Here we propose a compact architecture for adaptive decoding of electrocorticographic (ECoG) data into finger kinematics. We also describe a theoretically justified approach to interpreting the spatial and temporal weights in the architectures that combine adaptation in both space and time, such as ours. In hese architectures the weights are optimized not only for decoding of target signals but also for tuning away from the interfering sources, in both the spatial and the frequency domains. When applied to a dataset taken from the repository of Berlin BCI IV competition, our architecture outperformed the competition winners without the need for feature selection. Moreover, by looking at the architecture weights we could explain in physiological terms how our algorithm decodes spatial and temporal parameters of finger kinematics. As such, the proposed architecture offers a good decoder and a tool for investigating neural mechanisms of motor control.

Keywords: *ECoG, limb kinematics decoding, deep learning, machine learning, weights interpretation, spatial filter, temporal filter*

1 Introduction

Brain-computer interfaces (BCIs) link the nervous system to external devices [4] or even the other brains [15]. While there exist many applications of BCIs [1], clinically relevant BCIs have received most attention that aid in rehabilitation of patients with sensory, motor, and cognitive disabilities [12]. Clinical uses of BCIs range from assistive devices to neural prostheses that restore functions abolished by neural trauma or disease [2].

BCIs can deal with a variety of neural signals [14, 8] such as, for example, electroencephalographic (EEG) potentials sampled with electrodes placed on the surface of the head [11], or neural activity recorded invasively with the electrodes implanted in the cortex [6] or places onto the cortical surface [18]. The latter method, which we consider here, is called electrocorticography (ECoG). Accurate decoding of neural signals is key to building efficient BCIs.

BCI signal processing comprises several steps, including signal conditioning, feature extraction, and decoding. In the modern machine-learning algorithms, feature extraction and decoding are not separate but rather simultaneous computations performed with the computational architectures called Deep Neural Networks (DNN) [9]. DNNs derive features automatically when executing regression or classification tasks. While it is often difficult to interpret the computations performed by a DNN, such interpretations are essential to gain understanding of the properties of brain activity contributing to decoding, and to ensure that artifacts do not affect the decoding results. In particular, interpretation of features computed by the first several layers of a DNN could shed light on the neurophysiological mechanisms underlying the behavior being studied. Ideally, by examining DNN weights, one should be able to match the algorithm's operation to the functions and properties of the neural circuitry to which the BCI connects. Moreover, we suggest that physiologically tractable DNN architectures could facilitate the development of efficient and versatile BCIs.

Several useful and compact architectures have been developed for processing EEG and ECoG data. The operation of some blocks of these architectures can be straightforwardly interpreted. Thus, EEGNet [7] contains explicitly delineated spatial and temporal convolutional blocks. This architecture yields high decoding accuracy with a minimal number of parameters. However, due to the cross-filter-map connectivity between any two layers, a straightforward interpretation of the weights is difficult. Some insight regarding the decision rule can be gained using DeepLIFT combined with the analysis of the hidden unit activation patterns. Schirrmeister et al. describe two architectures: DeepConvNet and its compact version ShallowConvNet. The latter architecture consists of just two convolutional layers that perform temporal and spatial filtering, respectively [19].

Here we propose several novel approaches for making the operation of deep architectures tractable and interpretable neurophysiologically. Our approaches bear a resemblance with the recent study of Zubarev et al. [23] reporting two compact neural network architectures, LF-CNN and VAR-CNN, that outperformed the other decoders of MEG data, including linear models and more complex neural networks such as ShallowFBCSP-CNN, EEGNet-8 and VGG19. LF-CNN and VAR-CNN contain only a single non-linearity, which distinguishes them from most other DNNs. Because of this feature, the weights of these architectures are readily interpretable with the well-established approaches for interpreting the weights in linear models. Specifically, the spatial weights can be interpreted based on the principles of the estimation theory [5] combined with several additional assumptions, like the network training provides a Wiener-optimal solution and the subsequent temporal filtering can be disregarded. As to the temporal convolution weights, they can be interpreted by considering the Fourier-domain representations (with a caveat that the input data spectral characteristics are not taken into account).

While the compact architecture described here is conceptually similar to LF-CNN, our goals were different from those of Zubarev et al. [23]. We developed a theoretically justified approach to interpreting the spatial and temporal weights. Our method applies the optimal estimation theory to the space of temporally embedded multichannel data, with factorized spatial and temporal processing. This method allows us to consider two factors: (1) the weights optimizing the output correlation with the target signal, and (2) the weights minimizing the interference from the sources in both spatial and frequency domains.

2 Methods

Figure 1 illustrates the relationship between motor behavior (hand movements), brain activity, and ECoG recordings. The activity, $\mathbf{e}(t)$, of a set of neuronal populations, $G_1 - G_I$, engaged in motor control, is converted into a movement trajectory, $z(t)$, through a non-linear transform H: $z(t) = H(\mathbf{e}(t))$. The activity of populations $A_1 - A_J$ is unrelated to movement. The recordings of $\mathbf{e}(t)$ with a set of sensors are represented by a K-dimensional vector of sensor signals, $\mathbf{x}(t)$. This vector can be modeled as a linear mixture of signals resulting from the application of forward-model matrices $\mathbf{G}$ and $\mathbf{A}$ to task-related sources, $\mathbf{s}(t)$, and task-unrelated sources, $\mathbf{f}(t)$, respectively:

$$\mathbf{x}(t) = \mathbf{G}\mathbf{s}(t) + \mathbf{A}\mathbf{f}(t) = \sum_{i=1}^{I} \mathbf{g}_i s_i(t) + \sum_{j=1}^{J} \mathbf{a}_j f_j(t) \tag{1}$$

We will refer to the noisy component of the recording as $\eta(t) = \sum_{j=1}^{J} \mathbf{a}_j f_j(t)$.

Linear inverse mapping is commonly used to derive the activity of sources from the sensor signals: $\hat{\mathbf{s}}(t) = \mathbf{W}^T\mathbf{X}(t)$, where columns of $\mathbf{W}$ form a spatial filter that counteracts the volume conduction effect and decreases the effect of noisy sources.

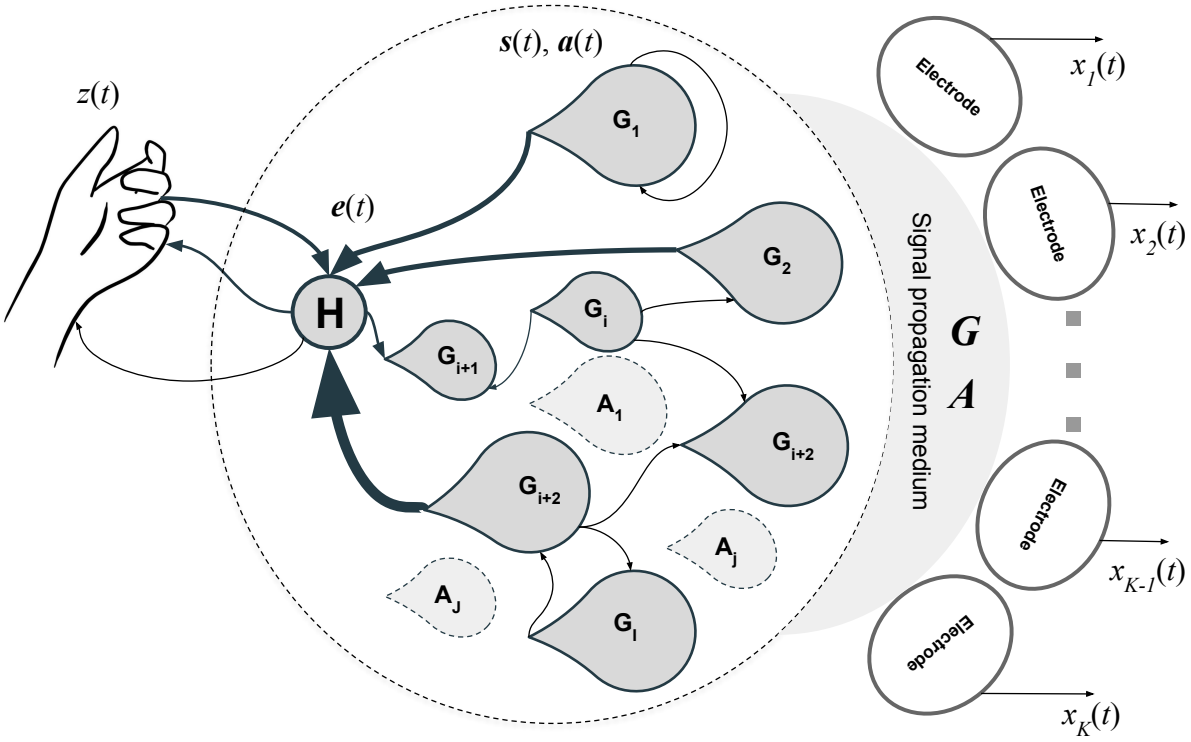

Figure 1: Phenomenological diagram

Neuronal correlates of motor planing and execution have been extensively studied [22]. In the cortical-rhythm domain, alpha and beta components of the sensorimotor rhythm envelope desynchronize just prior to the execution of a movement and rebound with a significant overshoot upon the completion of a motor act [13]. The magnitude of these

modulations correlates with the person's ability to control a motor-imagery BCI [16]. Additionally, the frequency of beta bursts in the primary somatosensory cortex is inversely correlated with the ability to detect tactile stimuli [20]. Intracranial recordings, such as ECoG, allow reliable measurement of the faster gamma band activity, which is temporally and spatially specific to movement patterns [21]. Overall, rhythmic components of brain sources, $\mathbf{s}(t)$, appear to be useful for BCI implementations. These rhythmic signals can be computed as linear combinations of band-passed sensor data, $\mathbf{x}(t)$.

The most straightforward approach for extracting the kinematics, $z(t)$, from brain recordings, $\mathbf{x}(t)$, is to directly learn the mapping $z(t) = \mathcal{F}(\mathbf{x}(t))$. To do so, one needs to parametrically describe this mapping. Here we used a specific network architecture for this purpose. The architecture was constructed in a close correspondence with the neurophysiological description of the observed phenomena, which facilitated our ability to interpret the results.

2.1 Network architecture

The compact and adaptable adaptable architecture that we developed is shown in Figure 2. An adaptive envelope extractor is the key component of this architecture. The envelope extractor, which is a module widely used in signal processing systems, was implemented using modern DNN primitives, namely a pair of convolutional operations that perform band-pass and low-pass filtering and one non-linearity ReLu(-1) that corresponds to computing the absolute value of the output of the first 1-D convolutional layer. To make the decision rule of this structure tractable, we used non-trainable batch normalization when streaming the data though the structure. All input signals were standardized.

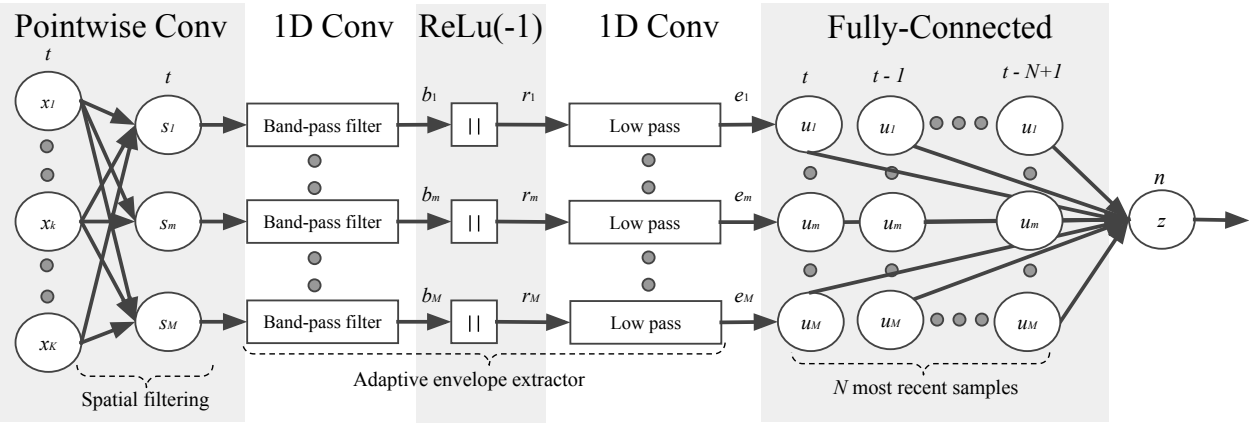

Figure 2: The proposed compact DNN architecture

In our architecture, the envelope detectors received spatially filtered sensor signals $\mathbf{s}_m$ that were calculated by the point-wise convolutional layer. This layer counteracted the volume-conduction processes represented by the forward-model matrix $\mathbf{G}$ in our phenomenological model (Figure 1). Next, we approximated operator H as a function of the lagged instantaneous power of the narrow-band source timeseries. This was done with a fully connected layer that mixed the samples of envelopes, $e_m(n)$, into a single estimate of the kinematic parameter, $z(n)$.

2.2 Two regression problems and DNN weights interpretation

The proposed architecture processes data in chunks of a prespecified length of N samples. The processing of a chunk of input data, $\mathbf{X}(t) = [\mathbf{x}(t), \mathbf{x}(t-1), \ldots \mathbf{x}(t-N+1)]$, by the first two layers performing spatial and temporal filtering can be described for the m-th element as

$$b_m(n) = \mathbf{w}_m^T \mathbf{X}(t) \mathbf{h}_m \tag{2}$$

The non-linearity, $ReLu(-1)$, in combination with the low-pass filtering performed by the second convolutional layer, extracts the envelopes of rhythmic signals.

The analytic signal is mapped one-to-one to its envelope [3]. Additionally, for the original real-valued data, the imaginary part of the analytic signal is uniquely determined as Hilbert transform. Therefore, the adjustment of the spatial and temporal filter weights to obtain some specific envelope $e_m(t)$ is equivalent to the adjustment of the weights to obtain this envelope's generating analytic signal $s_m(t)$. Accordingly, this is a linear regression problem where either spatial or temporal weights are fixed and temporal or spatial weights are sought, correspondingly.

We assume that training of the adaptive envelope detectors results in optimal spatial and temporal convolution weights, $\mathbf{w}_m^*$ and $\mathbf{h}_m^*$, correspondingly. Then, optimal spatial filter weights can be obtained as a solution to a convex optimization problem formulated over spatial or temporal subset of parameters:

$$\mathbf{w}_m^* = argmin_{\mathbf{w}_m} \{ \| \, b_m(n) - \mathbf{w}_m^T \mathbf{X}(t) \mathbf{h}_m^* \, \|_2^2 \} \tag{3}$$

where the temporal weights are fixed at their optimal value, $\mathbf{h}_m^*$. Similarly, when spatial weights are fixed at the optimal value $\mathbf{w}_m^*$, temporal weights are expressed by the equation:

$$\mathbf{h}_m^* = argmin_{\mathbf{h}_m} \{ \| \, b_m(t) - \mathbf{w}_m^{*T} \mathbf{X}(t) \mathbf{h}_m \, \|_2^2 \} \tag{4}$$

Given the forward model 1 and the regression problem 3 and assuming statistical independence of the rhythmic potentials $s_m(t),\ m = 1, \ldots, M$, the topographies of the underlying neuronal populations can be found as [**?** 5]

$$\mathbf{g}_m = E\{\mathbf{Y}(t)\mathbf{Y}^T(t)\}\mathbf{w}_m^* = \mathbf{R}_m^Y \mathbf{w}_m^* \tag{5}$$

where $\mathbf{Y}(t) = \mathbf{X}(t)\mathbf{h}_m$ is a temporally filtered chunk of multichannel data and $\mathbf{R}_m^Y = E\{\mathbf{Y}(t)\mathbf{Y}^T(t)\}$ is a $K \times K$ covariance matrix of the temporally filtered data, assuming that $x_k(t),\ k = 1, ..., K$ are all zero-mean processes. Thus, when interpreting individual spatial weights corresponding to each of the M paths of the architecture shown in Figure 2 one has to take into account the temporal filter weights $\mathbf{h}_m$ the to which the individual $m-$th branch is tuned. Therefore, to transform the spatial weights of different branches into spatial patterns, branch-specific covariance matrices $\mathbf{R}_m^Y$ should be used that depend on the temporal convolution weights of each particular branch.

The temporal weights can be interpreted in a similar way. The temporal pattern is calculated as

$$\mathbf{q}_m = E\{\mathbf{V}(t)\mathbf{V}^T(t)\}\mathbf{h}^*_m = \mathbf{R}^V_m\mathbf{h}^*_m \tag{6}$$

where $\mathbf{V}(t) = \mathbf{X}^T(t)\mathbf{w}^*_m$ is a spatially filtered chunk of incoming data and $\mathbf{R}^V_m = E\{\mathbf{V}(t)\mathbf{V}^T(t)\}$ is an $N \times N$ covariance matrix of the spatially filtered data, assuming that $x_k(t)$, $k = 1, ..., K$ are all zero-mean processes. As with the spatial patterns, when interpreting individual temporal weights corresponding to each of the M branches of the architecture shown in Figure 2, one has to take into account the spatial filter weights $\mathbf{h}_m$ used to filter the individual $m-$th branch. To transform the temporal convolution weights of different branches into temporal patterns, branch-specific covariance matrices $\mathbf{R}^V_m$ should be used that depend on the spatial convolution weights of each particular branch. To assess the temporal pattern, we usually explore it in the frequency domain, i.e. $Q_m(f) = \sum_{t=0}^{t=N-1} q_m(t)e^{-j2\pi ft}$, where $q_m(t)$ is the t-th element of temporal pattern vector $\mathbf{q}_m$.

Hitherto, we assumed that data chunk length N is equal to the length of the filters in the first convolutional layer. In general, this does not have to be the case. Our assumption emphasizes the formal similarity between the spatial and temporal dimensions. Additionally, we emphasize that the approach to the interpretation of temporal patterns requires taking into account the correlation structure of the independent variable in the regression model. When the data chunk is longer than the filter length, equation 2 has to be rewritten using the convolution operation. In this case, instead of a scalar, the equation returns a vector of samples, with the vector length depending on the choice of strategy used to deal with the transient at the edges of the chunk. It is also easier to operate in the frequency domain from the very beginning, and use the standard Wiener filtering arguments. In the frequency domain, the Wiener filter weights can be expressed as a function of the power spectral density $P^{yy}_m(f)$ of the spatially filtered sensor data $y_m(t) = \mathbf{w}^T_m\mathbf{x}(t)$ in the m-th branch and the density of cross-spectrum, $P^{sy}_m(f)$, between $s_m(t)$ and $y_m(t)$:

$$Q^*_m(f) = \frac{P^{sy}_m(f)}{P^{yy}_m(f)} \tag{7}$$

Then, using the assumption that $\eta(t)$ and $\mathbf{s}(t)$ in 1 are statistically independent, we obtain the following expressions:

$$H^*_m(f) = \frac{P^{ss}_m(f)}{P^{ss}_m(f) + P^{\eta\eta}_m(f)} = \frac{P^{ss}_m(f)}{P^{yy}_m(f)} \tag{8}$$

Therefore, the frequency-domain pattern of the signal isolated by the m-th branch spatial filter can be computed as

$$Q^*_m(f) = P^{yy}_m(f)H^*_m(f) \tag{9}$$

where $H^*_m(f)$ in 9 is the Fourier transform of the vector $\mathbf{h}^*_m$ containing temporal-convolution weights identified during the adaptation of the envelope detector in the $m-th$ branch. Viewing this result as a product of learning, it means that the learned vector of temporal convolution weights, $\mathbf{h}_m$, represents the power spectral density of the brain potentials that are important for decoding sensor signals, $\mathbf{x}(t)$, into the kinematics, $z(t)$.

The spatial patterns of neuronal sources recovered from the spatial filtering weights are routinely used for dipole fitting to localize functionally important neural sources. The temporal patterns interpreted according to 9 and 6 can be used to fit the models of neural population dynamics, which are relevant to specific decoding tasks.

2.3 Simulations

To tackle the performance of the proposed architecture, we performed a set of simulations. The simulated data corresponded to the setting shown in the phenomenological diagram (Figure 1). We simulated $I = 4$ task related sources with rhythmic potentials, $s_i(t)$. The potentials of these four task related populations were generated as narrow-band processes (in 30-80 Hz], 80-120 Hz], 120-170 Hz and 170-220 Hz bands) resulting from filtering of Gaussian pseudo-random sequences with a bank of FIR filters. We then simulated the kinematics, $z(t)$, as a linear combination of the four envelopes of these rhythmic signals with randomly generated vector of coefficients. We used $J = 40$ task-unrelated rhythmic sources with activation timeseries obtained similarly to the task-related sources but with filtering within the following four bands: 40-70 Hz, 90-110 Hz, 130-160 Hz, and 180 - 210 Hz bands. As a result, we obtained 10 task unrelated sources active in each of these bands making it a total of $J = 40$ task unrelated sources. To simulate volume conduction effect, we randomly generated a 4×5 dimensional forward matrix $\mathbf{G}$ and a 40×5 dimensional forward matrix $\mathbf{A}$. There matrices mapped the task-related and task-unrelated activity, respectively, onto the sensor space.

We generated 15 minutes worth of data sampled at 1,000 Hz and split them into two equal contiguous parts. We used the first part for training and the second for testing

3 Experimental datasets

To compare the performance of our simple neural network with the top linear models that rely on preset features, we used publicly available data collected by Kubanek et al [] from the BCI Competition IV. This dataset contains concurrent multichannel ECOG and finger flexion measurements collected in three epileptic patients implanted with ECoG electrodes for medical reasons. The database consists of 400 s of training data and 200 s of test data. The recordings were conducted with 64 or 48 electrodes placed over the sensoriomotor cortext. The exact spatial locations and the order of the electrodes are not provided. As a baseline in this comparison, we chose the winning solution offered by Nanying Liang and Laurent Bougrain [10]. This solution employs extracting the amplitudes of the data filtered in 1-60 Hz, 60-100 Hz, and 100-200 Hz band followed by a pairwise feature selection and decoded using Wiener filter with N = 25 taps from the immediate past.

The other dataset comes from our laboratory. The recordings were conducted with a 64-channel Adtech microgrid connected to EB Neuro BE Plus LTM Bioelectric Signals Amplifier System that sampled data at 2048 Hz. The amplifier software streamed data via Lab Streaming Layer protocol. The experimental software supported this protocol, implemented the experimental paradigm (a finger movement task) and synced ECoG and kinematics. Finger kinematics was captured by Perception Neuron system as relative angles for the sensor units attached to finger phalanges, and sampled at 120 Hz . Finger flexion-extension angle was used as kinematics timeseries, $z(t)$.

Recordings were obtained in two patients with pharmaco-resistant form of epilepsy; ECoG electrodes were implanted for the purpose of pre-surgery localization of epileptic foci and mapping of eloquent cortex. Thus,for these data, unlike Kubanek [], we knew cortical location of each electrode and could visualize spatial patterns of activity with high accuracy. The patients performed self-paced flexions of each individual finger for 1 min. The study was conducted according to the ethical standards of the 1964 Declaration of Helsinki. All participants provided written informed

consent prior to the experiments. The ethics research committee of the National Research University, The Higher School of Economics approved the experimental protocol of this study

4 Results for simulated data

4.1 Adaptive envelop detector

As described in Methods, to interpret optimal temporal convolution weight,s we need to consider the spectral characteristics of neural recordings. To illustrate this, we trained a single-channel adaptive envelope detector in the environment with the interference occupying a subrange of the target signal band. As can be seen from Figure 3, the Fourier profile of the identified temporal convolution weights can not be used to assess the power spectral density of the underlying signal. At the same time, the expression in 8 allows us to obtain a proper pattern that matches well the simulated spectral profile. Conversely, using the FFT of the convolutional filter weights yields fundamentally erroneous estimates of the frequency-domain patterns and erroneous interpretation of the underlying neurophysiology.

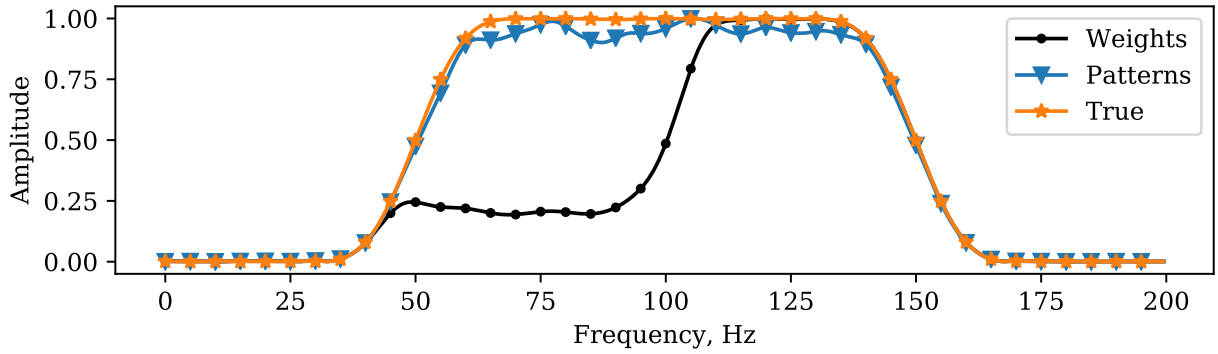

Figure 3: The importance of taking into account the power spectral density of the independent variable when interpreting linear regression weights. The true pattern (*) gets erroneously reproduced as the FFT of the temporal convolution weights (●). Taking into account the power spectral density of the spatially filtered signal allows to fix this situation (▼)

4.2 Realistic simulations

For the simulated data, we trained the algorithm to predict the kinematic variable $z(t)$. In the noiseless case, the proposed architecture achieved accuracy of 99% measured as correlation coefficient between the true and recovered kinematics (Figure 4). We then compared the envelopes at each of the four branches of our architecture and observed that the true latent variable timeseries (in the form of the underlying narrow-band envelopes) matched very well those estimated with our architecture (Figure 5). The correlation between the estimated and true envelope timeseries fell into the 87 - 96 % range.

As described in Methods, for spatial weights interpretation, we used the linear estimation theoretic approach [5] . To warn against its naive implementation in the context of architectures that combine spatial and temporal filtering, we computed spatial patterns where we used the input data covariance, $\mathbf{R}^X$, without taking into account the individual-branch temporal filters. In the corresponding plots, we refer to the patterns determined using this approach as *Patterns*

vanilla. The proper way to apply this estimation approach is to compute spatial covariance, $\mathbf{R}^Y$, for the temporally filtered data 6. These properly determined patterns are labeled as *Patterns vanilla*.

In the right column of Figure 6, we show the results for the noiseless case for all four branches of the network. As expected, the spatial *Patterns vanilla* and *Patterns* are identical and match the ground truth exactly. The left column shows Fourier representations of the temporal weights where we can observe that in the noise-free scenario Fourier representations of the temporal weights matches exactly the power spectral density of the simulated data.

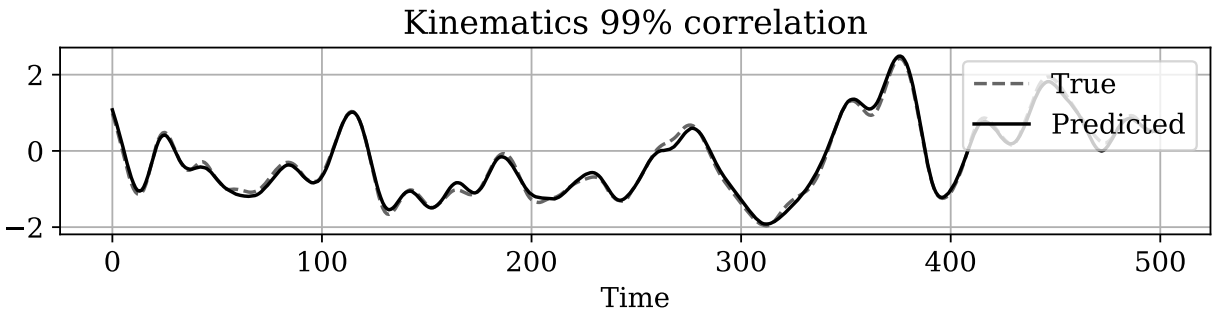

Figure 4: Realistic simulations. Actual and decoded kinematics.

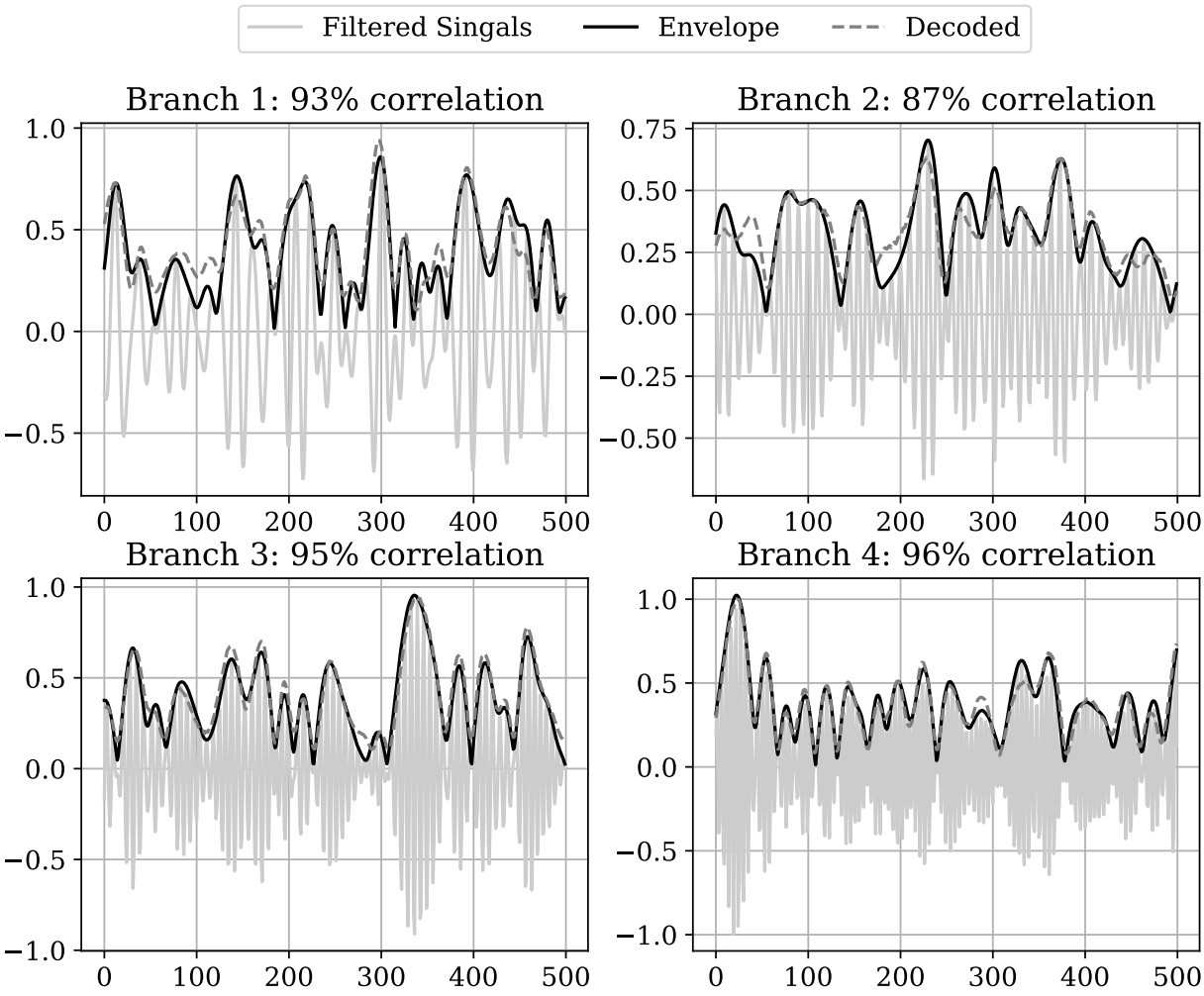

Figure 5: Branch envelopes decoding. Comparison between the true and the decoded envelope.

In the noisy case demonstrated in Figure 7, only *Patterns vanilla* match well with the simulated topographies of the underlying sources. Spectral characteristics of the trained temporal filtering weights exhibit characteristic dips in the bands corresponding to the activity of the interfering sources. After applying the theoretic estimation 9, we obtain the spectral patterns that more closely match the simulated ones and have the dips compensated.

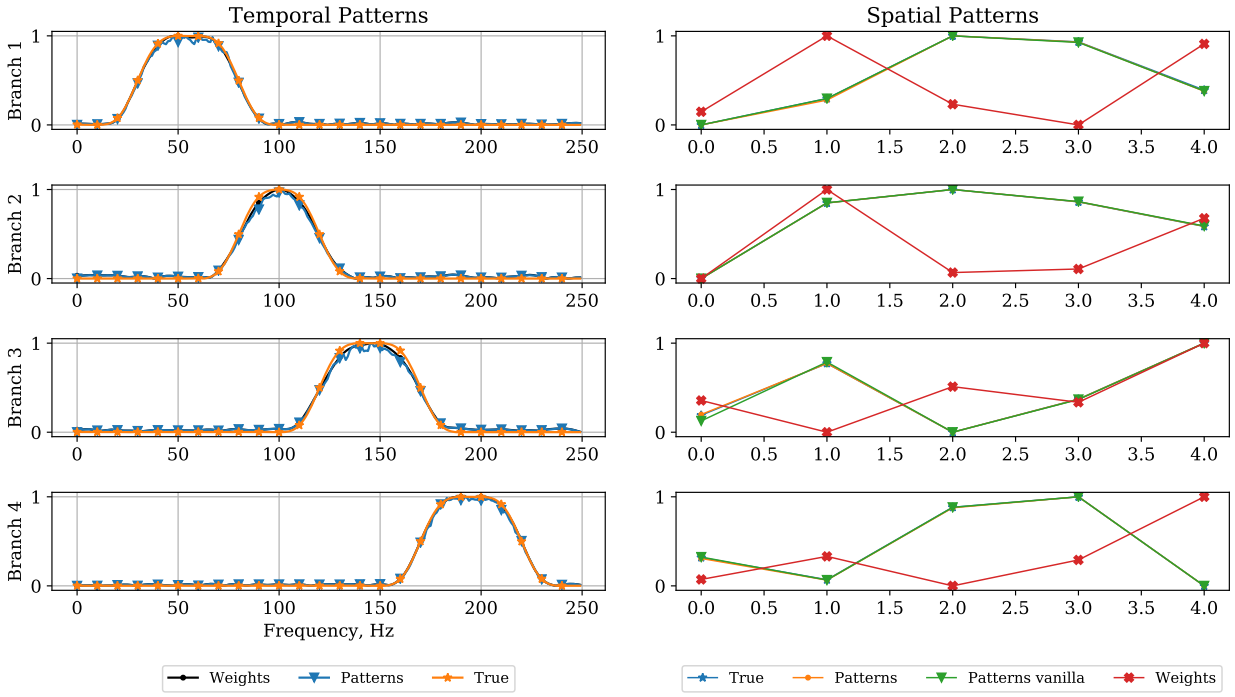

Figure 6: Temporal (left) and spatial(right) patterns obtained for the noiseless case. See the main text for description.

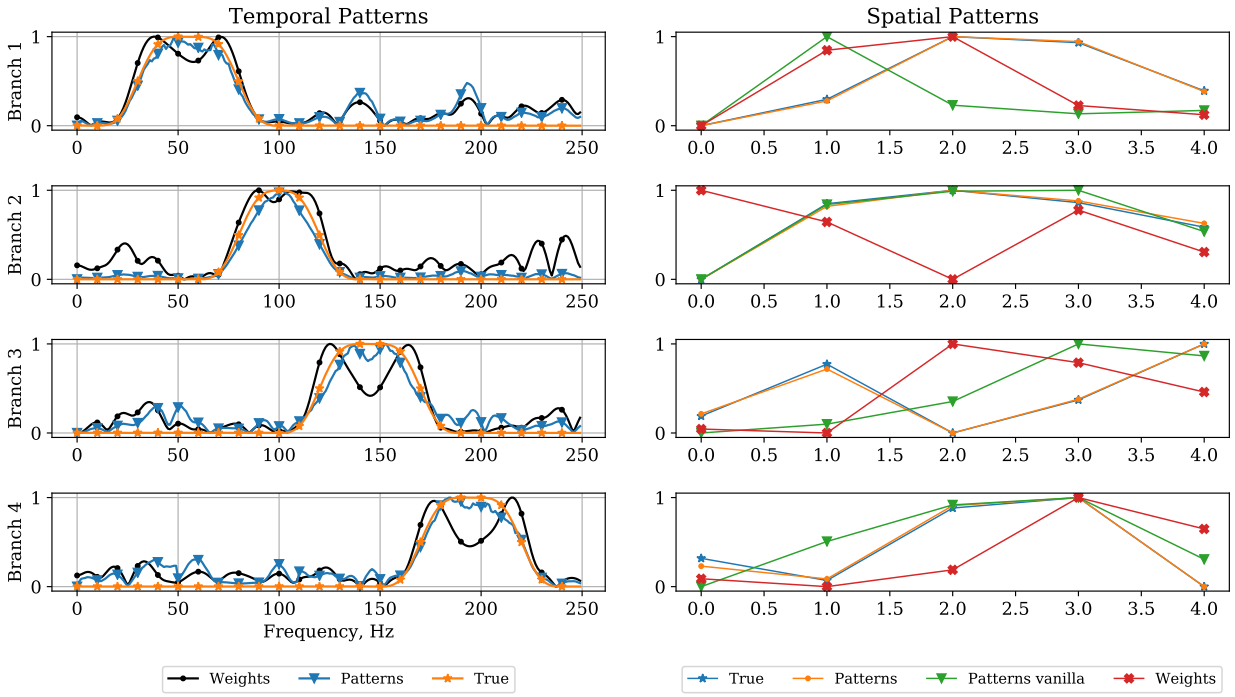

Figure 7: Temporal (left) and spatial(right) patterns obtained for the noisy case, SNR = 1.5. See the main text for description.

5 Analysis of experimental data

5.1 Berlin BCI Competition IV data

In the context of electrophysiological data processing, the major advantage of the architectures inspired by the deep-learning principle is their ability to automatically select features while performing classification or regression tasks [17]. When applied to to the data from Berlin BCI Competition IV, our architecture – based on the adaptive envelope detectors –performed on par or better than the winning solution by Lian and Bougrain [10], see Table 1.

Subject 1

	Thumb	Index	Middle	Ring	Little
Winner	0,58	0,71	0,14	0,53	0,29
NET	0,54	0,7	0,2	0,58	0,25

Subject 2

	Thumb	Index	Middle	Ring	Little
Winner	0,51	0,37	0,24	0,47	0,35
NET	0,5	0,36	0,22	0,4	0,23

Subject 3

	Thumb	Index	Middle	Ring	Little
Winner	0,59	0,51	0,32	0,53	0,2
NET	0,71	0,48	0,5	0,52	0,61

Table 1: Comparison of the performance of the proposed architecture (NET) and the winning solution (Winner) of Berlin BCI IV competition dataset(4)

5.2 The CBI data

We also applied the proposed solutions to the recordings that we conducted in two patients implanted with 8×8 ECoG grids over the sensorimotor cortex.

The following table shows the accuracy achieved with the proposed architecture for the decoding of finger movements.

	Thumb	Index	Ring	Little
Subject 1	0,48	**0,79**	0,61	0,32
Subject 2	0,73	0,55	0,78	**0,79**

Table 2: Decoding performance achieved in the two CBI patients. The table show correlation coefficient between the actual and the decoded finger trajectory for the four fingers in two patients.

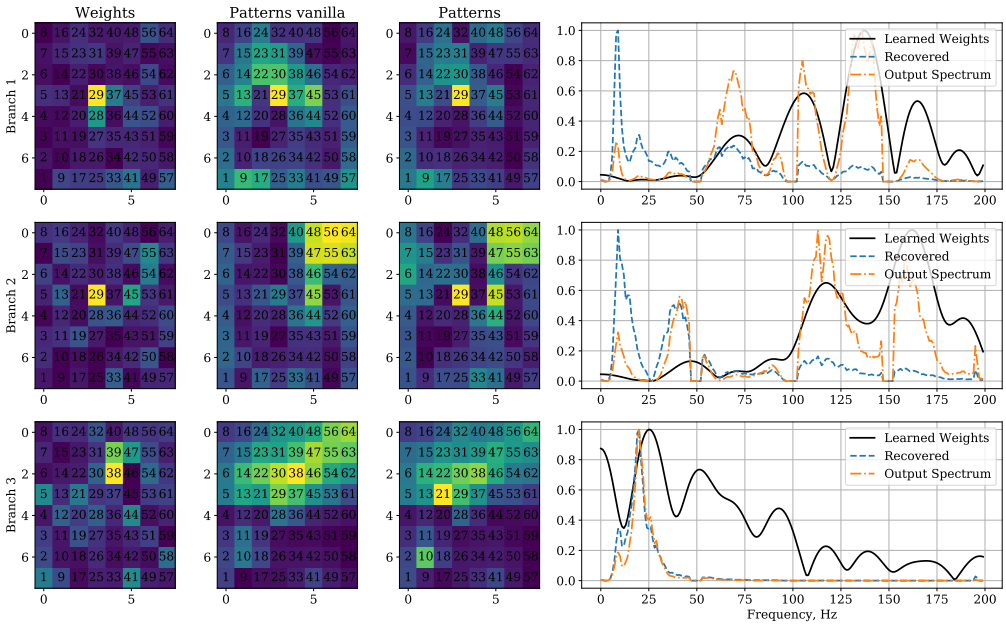

Figure 8: Network weights interpretation for the index finger decoder in CBI patient 1. Each row of plots corresponds to one of the three branches of the trained decoder. The left most column shows color-coded spatial filter weights, next two columns correspond to vanilla and properly rcovered spatial patterns. The fourth column interprets temporal filter weights in the Fourier domain. Filter weights - solid line, power spectral density (PSD) pattern of the underlying LFP - dashed line. Another dashed line more similar to the filter weights Fourier coefficients is the PSD of the signal at the output of the temporal convolution block.

Figures 8 and 9 depict the interpretation of the obtained spatial and temporal weights. The plots are shown for the finger with the highest decoding accuracy (highlighted in bold in Table 2) for two patients.

The decoding architecture for both patients had three branches and each branch was tuned to a source with specific spatial and temporal patterns. In Figure 8, we show the spatial filter weights, vanilla patterns and proper patterns interpreted using the expression described in the the Methods section. It can be seen that, while the temporal filter weights (solid line) clearly emphasized the frequency range above 100 Hz in the first two branches, the actual spectral pattern of the source (dashed line) in addition to the gamma-band content had a peak at around 11 Hz (1st, 2nd branches) and in the 25-50 Hz range (2nd branch). These peaks likely correspond to the sensorimotor rhythm and low-frequency gamma rhythms, respectively.

The third branch appears to capture the lower-frequency range and its spatial pattern is noticeably more diffuse than that in the first two branches that capture the higher-frequency components. This is consistent with the phenomenon that the size and activation frequency of neuronal populations are reciprocally proportional.

Similar observations can be made from Figure 9 that shows to the decoding results for the little finger in patient 2.

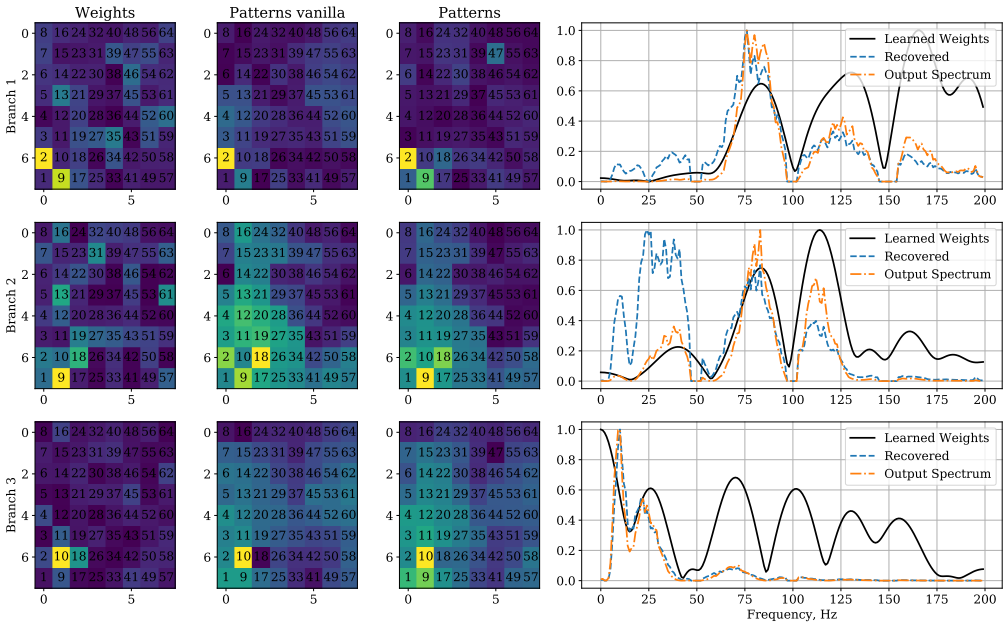

Figure 9: Network weights interpretation for the little finger decoder in CBI patient 2. Each row of plots corresponds to one of the three branches of the trained decoder. The left most column shows color-coded spatial filter weights, next two columns correspond to vanilla and properly rcovered spatial patterns. The fourth column interprets temporal filter weights in the Fourier domain. Filter weights - solid line, power spectral density (PSD) pattern of the underlying LFP - dashed line. Another dashed line more similar to the filter weights Fourier coefficients is the PSD of the signal at the output of the temporal convolution block.

6 Conclusion

We developed a novel compact and neurophysiologically interpretable architecture. Using this architecture, we extended the weights interpretation approach previously applied in [5] to the interpretation of the temporal convolution weights. We tested the proposed approach using simulated and experimental data. In the realistically simulated data, our architecture recovered with high accuracy the neuronal substrate that contributed to the kinematics data.

We also applied the proposed architecture to an experimental dataset taken from the repository of Berlin BCI IV competition. Our architecture delivered similar or better decoding accuracy as compared the winning solution of the BCI competition [10]. In contrast to the traditional approaches, our architecture did not require any preset features. Instead, after the architecture was trained to decode finger kinematics, we could interpret the weights and extracted physiologically meaningful patterns corresponding to both spatial and temporal convolution weights.

7 Acknowledgment

This work is supported by the Center for Bioelectric Interfaces NRU HSE, RF Government grant, ag. No.14.641.31.0003.

References

[1] Sarah N Abdulkader, Ayman Atia, and Mostafa-Sami M Mostafa. Brain computer interfacing: Applications and challenges. *Egyptian Informatics Journal*, 16(2):213–230, 2015.

[2] Ujwal Chaudhary, Niels Birbaumer, and Ander Ramos-Murguialday. Brain–computer interfaces for communication and rehabilitation. *Nature Reviews Neurology*, 12(9):513, 2016.

[3] Stefan L Hahn. On the uniqueness of the definition of the amplitude and phase of the analytic signal. *Signal Processing*, 83(8):1815–1820, 2003.

[4] Nicholas G Hatsopoulos and John P Donoghue. The science of neural interface systems. *Annual review of neuroscience*, 32:249–266, 2009.

[5] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110, 2014.

[6] Mark L Homer, Arto V Nurmikko, John P Donoghue, and Leigh R Hochberg. Sensors and decoding for intracortical brain computer interfaces. *Annual review of biomedical engineering*, 15:383–405, 2013.

[7] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: A compact convolutional network for eeg-based brain-computer interfaces. *arXiv preprint arXiv:1611.08024*, 2016.

[8] Mikhail A Lebedev and Miguel AL Nicolelis. Brain-machine interfaces: From basic science to neuroprostheses and neurorehabilitation. *Physiological reviews*, 97(2):767–837, 2017.

[9] Steven Lemm, Benjamin Blankertz, Thorsten Dickhaus, and Klaus-Robert Müller. Introduction to machine learning for brain imaging. *Neuroimage*, 56(2):387–399, 2011.

[10] Nanying Liang and Laurent Bougrain. Decoding finger flexion from band-specific ecog signals in humans. *Frontiers in neuroscience*, 6:91, 06 2012.

[11] Sergio Machado, Fernanda Araújo, Flávia Paes, Bruna Velasques, Mario Cunha, Henning Budde, Luis F Basile, Renato Anghinah, Oscar Arias-Carrión, Mauricio Cagy, et al. Eeg-based brain-computer interfaces: an overview of basic concepts and clinical applications in neurorehabilitation. *Reviews in the Neurosciences*, 21(6):451–468, 2010.

[12] Joseph N Mak and Jonathan R Wolpaw. Clinical applications of brain-computer interfaces: current state and future prospects. *IEEE reviews in biomedical engineering*, 2:187–199, 2009.

[13] Yaron Meirovitch, Hila Harris, Eran Dayan, Amos Arieli, and Tamar Flash. Alpha and beta band event-related desynchronization reflects kinematic regularities. *Journal of Neuroscience*, 35(4):1627–1637, 2015.

[14] Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain computer interfaces, a review. *Sensors*, 12(2):1211–1279, 2012.

[15] Miguel Pais-Vieira, Mikhail Lebedev, Carolina Kunicki, Jing Wang, and Miguel Nicolelis. A brain-to-brain interface for real-time sharing of sensorimotor information. *Scientific reports*, 3:1319, 02 2013.

[16] Johanna Reichert, Silvia Kober, Christa Neuper, and Guilherme Wood. Resting-state sensorimotor rhythm (smr) power predicts the ability to up-regulate smr in an eeg-instrumental conditioning paradigm. *Clinical Neurophysiology*, 126, 02 2015.

[17] Yannick Roy, Hubert Banville, Isabela Maria Carneiro de Albuquerque, Alexandre Gramfort, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review, 01 2019.

[18] Gerwin Schalk and Eric C Leuthardt. Brain-computer interfaces using electrocorticographic signals. *IEEE reviews in biomedical engineering*, 4:140–154, 2011.

[19] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for brain mapping and decoding of movement-related information from the human eeg. *arXiv preprint arXiv:1703.05051*, 2017.

[20] Hyeyoung Shin, Robert Law, Shawn Tsutsui, Christopher Moore, and Stephanie Jones. The rate of transient beta frequency events predicts behavior across tasks and species. *eLife*, 6, 11 2017.

[21] Ksenia Volkova, Mikhail A Lebedev, Alexander Kaplan, and Alexei Ossadtchi. Decoding movement from electrocorticographic activity: A review. *Frontiers in neuroinformatics*, 13:74, 2019.

[22] Daniel Wolpert and Zoubin Ghahramani. Computational principles of movement neuroscience. *Nature neuroscience*, 3 Suppl:1212–7, 12 2000.

[23] Ivan Zubarev, Rasmus Zetter, Hanna-Leena Halme, and Lauri Parkkonen. Adaptive neural network classifier for decoding meg signals. *NeuroImage*, 197:425–434, 2019.