Journal of Neural Engineering

PAPER

Decoding and interpreting cortical signals with a compact convolutional neural network

To cite this article: Artur Petrosyan *et al* 2021 *J. Neural Eng.* **18** 026019

View the article online for updates and enhancements.

You may also like

- A system identification analysis of optogenetically evoked electrocorticography and cerebral blood flow responses
Rex Chin-Hao Chen, Farid Atry, Thomas Richner et al.
- Wireless opto-electro neural interface for experiments with small freely behaving animals
Yaoyao Jia, Wasif Khan, Byunghun Lee et al.
- Chronic neural interfacing with cerebral cortex using single-walled carbon nanotube-polymer grids
Luigi Pavone, Slavianka Moyanova, Federica Mastroiacovo et al.

This content was downloaded from IP address 83.69.192.100 on 08/08/2022 at 11:27

Journal of Neural Engineering

PAPER

Decoding and interpreting cortical signals with a compact convolutional neural network

RECEIVED
29 September 2020

REVISED
6 January 2021

ACCEPTED FOR PUBLICATION
1 February 2021

PUBLISHED
2 March 2021

Artur Petrosyan[1], **Mikhail Sinkin**[2], **Mikhail Lebedev**[1] **and Alexei Ossadtchi**[1]

[1] Center for Bioelectric Interfaces, Higher School of Economics, Moscow 101000, Russia
[2] Moscow State University of Medicine and Dentistry, Moscow 101000, Russia

E-mail: ossadtchi@gmail.com

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

Abstract

Objective. Brain–computer interfaces (BCIs) decode information from neural activity and send it to external devices. The use of Deep Learning approaches for decoding allows for automatic feature engineering within the specific decoding task. Physiologically plausible interpretation of the network parameters ensures the robustness of the learned decision rules and opens the exciting opportunity for automatic knowledge discovery. *Approach.* We describe a compact convolutional network-based architecture for adaptive decoding of electrocorticographic (ECoG) data into finger kinematics. We also propose a novel theoretically justified approach to interpreting the spatial and temporal weights in the architectures that combine adaptation in both space and time. The obtained spatial and frequency patterns characterizing the neuronal populations pivotal to the specific decoding task can then be interpreted by fitting appropriate spatial and dynamical models. *Main results.* We first tested our solution using realistic Monte-Carlo simulations. Then, when applied to the ECoG data from Berlin BCI competition IV dataset, our architecture performed comparably to the competition winners without requiring explicit feature engineering. Using the proposed approach to the network weights interpretation we could unravel the spatial and the spectral patterns of the neuronal processes underlying the successful decoding of finger kinematics from an ECoG dataset. Finally we have also applied the entire pipeline to the analysis of a 32-channel EEG motor-imagery dataset and observed physiologically plausible patterns specific to the task. *Significance.* We described a compact and interpretable CNN architecture derived from the basic principles and encompassing the knowledge in the field of neural electrophysiology. For the first time in the context of such multibranch architectures with factorized spatial and temporal processing we presented theoretically justified weights interpretation rules. We verified our recipes using simulations and real data and demonstrated that the proposed solution offers a good decoder and a tool for investigating motor control neural mechanisms.

1. Introduction

Brain–computer interfaces (BCIs) link the nervous system to external devices [7] or even other brains [25]. 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 [18]. Clinical uses of BCIs range from assistive devices to neural prostheses that restore functions abolished by neural trauma or disease [5].

BCIs can deal with a variety of neural signals [14, 23] such as, for example, electroencephalographic (EEG) potentials sampled with electrodes placed on the surface of the head [17], or neural activity recorded invasively with the electrodes implanted into the cortex [9] or placed onto the cortical surface [33]. The latter method, which we consider here, is called electrocorticography (ECoG). Accurate decoding of neural signals is key to building efficient BCIs.

Typical BCI signal processing comprises several steps, including signal conditioning, feature extraction, and decoding. In the modern machine-learning algorithms, parameters of the feature extraction and decoding pipelines are jointly optimized within computational architectures called deep neural networks

© 2021 IOP Publishing Ltd

(DNNs) [15]. DNNs derive features automatically when trained to execute 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 or accompanying confounds do not affect the decoding results. DNNs can also be used for knowledge discovery. 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 decoder connects. Such physiologically tractable DNN architectures are likely to 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 [13] 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 technique [36] 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 [34]. Waytowich *et al* [38] describes a compact CNN architecture with separable spatial and temporal convolutions to perform classification of EEG in he SSVEP paradigm. Recent study of Zubarev *et al* [40] reported 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. This feature makes the weights of such architectures readily interpretable with the well-established approaches [8, 11, 26]. This methodology, however, has to be applied taking into account the peculiarities brought about by the separability of the spatial and temporal filtering steps in these architectures.

Here we introduce another simple architecture, developed independently but conceptually similar to those listed above, and use it as a testbed to refine the recipes for the interpretation of the weights in the family of architectures characterized by separated adaptive spatial and temporal processing stages. We refer to this kind of processing as factorized processing. We emphasize that when interpreting the weights in such architectures we have to keep in mind that these architectures tune their weights not only to adapt to the target neuronal population(s) but also to minimize the distraction from the interfering sources in both spatial and frequency domains.

The solutions exercised in [2, 3, 10, 21, 27] and elegantly summarized in [8] take care of this adaptive behavior but is directly applicable only to the regression like models where a single vector of weights is applied to the data(feature) vector. This is not the case with the type of models considered here where filtering in one domain is followed by the application of a filter in another domain. The factorized processing reduces the number of parameters in the architecture but requires a special weights interpretation approach derived here in order to accurately asses spatial patterns of the neuronal sources underlying decision rule learned by the architectures with factorized processing. Also using Wiener filtering arguments we for the first time expand the weights interpretation approach to the analysis of temporal filter weights and show how the learned temporal convolution kernels in combination with the spatially filtered neural activity data give access to the estimates of the power spectral density of the underlying neuronal populations pivotal to the decoding task.

2. Methods

In our description we will use standard notation where bold capitals denote matrices, small bold letters stand for column vectors and small italic symbols for scalars. To refer to the temporal dimension we use discrete index framed into square brackets. Figure 1 illustrates a hypothetical relationship between motor behavior (hand movements), brain activity, and ECoG recordings. The activity, $\mathbf{s}[n] = [s_1[n], \ldots, s_I[n]]^T \in \mathbb{R}^I$, of a set of I neuronal populations, $G_1 - G_I$, engaged in motor control, is converted into a movement trajectory, $z[n]$, through a non-linear transform H: $z[n] = H(\mathbf{e}[n])$ where $\mathbf{e}[n] = [e_1[n], \ldots, e_I[n]]^T$ is the vector of envelopes of $\mathbf{s}[n]$. The activity of another set of J populations $A_1 - A_J$ is unrelated to the movement. The recordings of this activity with a set of L sensors at time instance n are represented by a $L \times 1$ vector of sensor signals $\mathbf{x}[n] \in \mathbb{R}^L$. At each time instance n this vector can be modeled as a linear mixture of signals resulting from application of the forward-model matrices $\mathbf{G} = [\mathbf{g}_1[n], \ldots, \mathbf{g}_I[n]] \in \mathbb{R}^{L \times I}$ and $\mathbf{A} = [\mathbf{a}_1[n], \ldots, \mathbf{a}_J[n]] \in \mathbb{R}^{L \times J}$ to the column vector of activity of task-related sources at the time moment n, $\mathbf{s}[n] = [s_1[n], \ldots, s_I[n]]^T$, and task-unrelated sources, $\mathbf{f}[n] = [f_1[n], \ldots, f_J[n]]^T$, respectively:

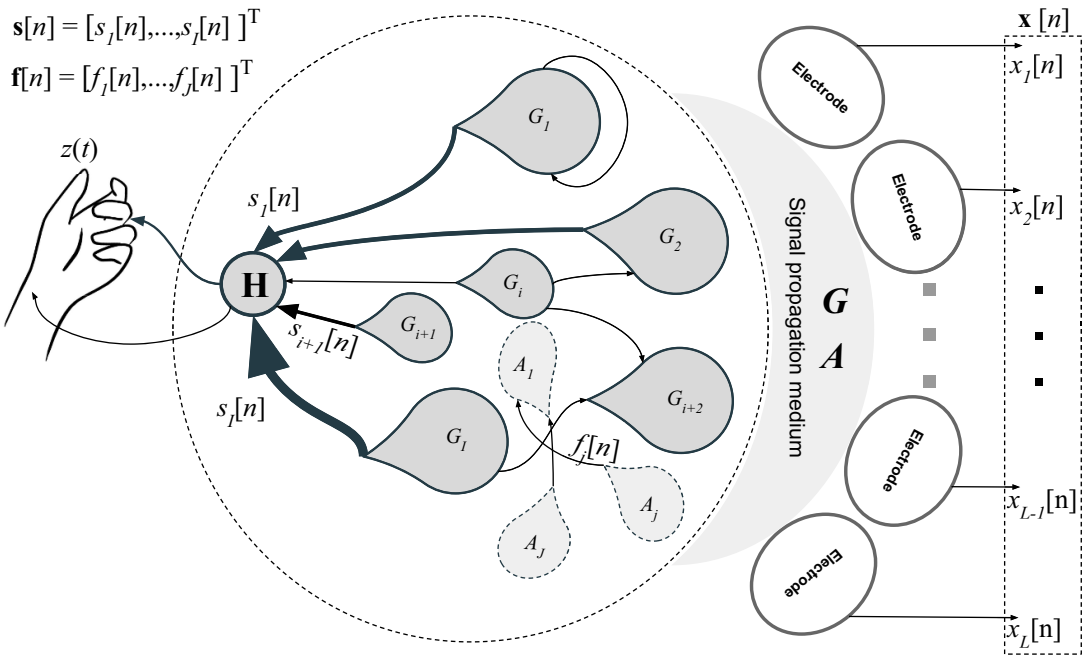

Figure 1. Phenomenological diagram.

$$\mathbf{x}[n] = \mathbf{G}\mathbf{s}[n] + \mathbf{A}\mathbf{f}[n] = \sum_{i=1}^{I} \mathbf{g}_i s_i[n] + \sum_{j=1}^{J} \mathbf{a}_j f_j[n] = \sum_{i=1}^{I} \mathbf{g}_i s_i[n] + \boldsymbol{\eta}[n]. \quad (1)$$

Column vectors $\mathbf{g}_i$, $i = 1,\ldots,I$ and $\mathbf{a}_j$, $j = 1,\ldots,J$ are the topographies of the task related and task-unrelated sources. We refer to the noisy, task-unrelated component of the recording as $\boldsymbol{\eta}[n] = \sum_{j=1}^{J} \mathbf{a}_j f_j[n] \in \mathbb{R}^L$. A similar generative model has been recently described in [32].

Given the linear generative model of electrophysiological data, the inverse mapping used to derive the activity of sources from the sensor signals is also commonly sought in the linear form: $\hat{\mathbf{s}}[n] = \mathbf{W}^T\mathbf{X}[n]$, where columns of $\mathbf{W}$ form a spatial filter that counteracts the volume conduction effect and decreases the contribution from the noisy, task-unrelated sources.

Neuronal correlates of motor planning and execution have been extensively studied [39]. In the cortical-rhythm domain, alpha and beta components of the sensorimotor rhythm desynchronize just prior to the execution of a movement and rebound with a significant overshoot upon the completion of a motor act [19]. The magnitude of these modulations correlates with the person's ability to control a motor-imagery BCI [30]. Additionally, the incidence rate of beta bursts in the primary somatosensory cortex is inversely correlated with the ability to detect tactile stimuli [35] and also affects other motor functions. Intracranial recordings, such as ECoG, allow reliable measurement of the faster gamma band activity, which is temporally and spatially specific to movement patterns [37] and is thought to accompany movement control and execution. Overall, based on the very solid body of research, rhythmic components of brain sources, $\mathbf{s}[n]$, appear to be useful for BCI implementations. Given the linearity of the generative model (1), these rhythmic signals reflecting the activity of specific neuronal populations can be computed as linear combinations of narrow-band filtered sensor data $\mathbf{x}[n]$.

The most straightforward approach for extracting the kinematics, $z[n]$, from brain recordings, $\mathbf{x}[n]$, is to use concurrently recorded data and directly learn the mapping $z[n] = \mathcal{H}(\mathbf{x}[n])$. To practically implement it, 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 observation equation (1) and the neurophysiological description of the observed phenomena illustrated in figure 1, which facilitated our ability to interpret the results.

2.1. Network architecture

The compact and adaptable architecture that we used here is shown in figure 2. As shown the architecture comprises M branches. Each branch is an adaptive envelope detector with its own pair of temporal filters preceded by the branch- specific spatial filter. Our envelope detector approximates the envelope extracted as the absolute value of the analytic signal calculated using Hilbert transform of the input signal. The processing flow we use mimics that of an analog detector receiver and has also been used in other similar compact CNN architectures that employ separate treatment of the spatial and the temporal dimensions

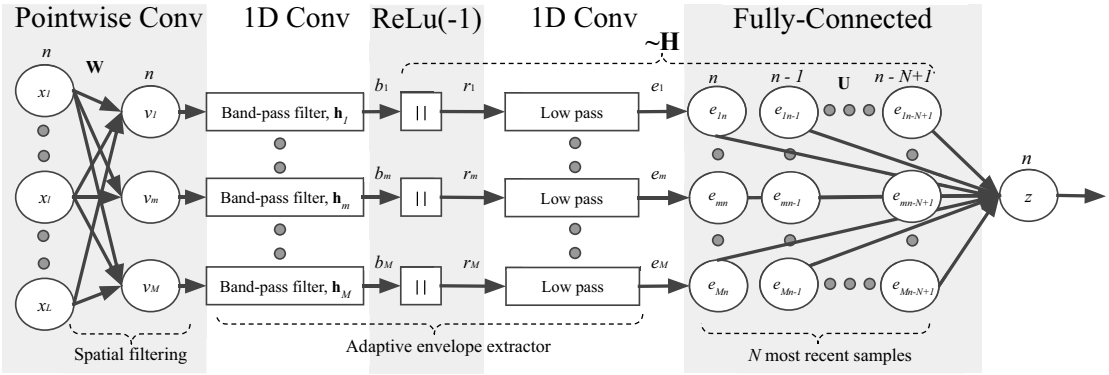

Figure 2. The architecture based on the compact CNN comprises several branches—adaptive envelope detector, receiving spatially unmixed input signals and outputting the envelopes whose N most recent values with indexes $n-N+1,\ldots,n$ are combined in the decoded variable z by the fully connected layer. Note that for compactness we have omitted temporal index in the sequence names.

[34, 40]. Each branch of our network is a parametric pipeline capable of extracting the instantaneous power of the input signal and adapting to the specific neuronal population and frequency band by tuning spatial and temporal filter weights correspondingly.

As shown in the diagram, the envelope detector can be implemented using modern DNN primitives, namely a pair of convolutional operations that perform band-pass and low-pass filtering with a single non-linearity ReLu(-1) in between that corresponds to computing the absolute value of the output of the first 1D convolutional layer. This step rectifies the signal (acts as a full-wave rectifier built using a pair of diodes) and is followed by a low-pass filter that smooths the rectifier output $r_m[n]$ to obtain the approximation of the envelope $e_m[n]$. Note that ReLu(a) is now a standard non-linearity used in the modern neural networks and defined as $\mathrm{ReLu}(x,a)=\{x, x\geqslant 0; ax, x<0\}$. To make the decision rule of this structure tractable, we used non-trainable batch normalization when streaming the data though the structure. This way we can harness the power of the optimization tools implemented within the deep learning approach to tune parameters of our network that uses spatial filters followed by envelope estimation as the feature extraction block.

In our architecture, the envelope detector of the mth branch receives as an input spatially filtered sensor signal $s_m[n]$ calculated by the point-wise convolutional layer. This layer is designed to invert the volume-conduction processes represented by the forward-model matrices $\mathbf{G}$ and $\mathbf{A}$ in our phenomenological model (figure 1). Next, we approximated operator H as a linear combination of the lagged instantaneous power (envelope) of the narrow-band source timeseries $\mathbf{s}(t)=[s_1(t),s_2(t),\ldots,s_I(t)]$ with coefficients from matrix $\mathbf{U}=\{u_{ml}\}$, $m=1,\ldots,M$, $l=1,\ldots,N$. 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 $z[n]=\sum_{m=1}^{M}\sum_{l=1}^{N}e_m[n-l]u_{ml}+u_0$ where u_0 models the DC offset term that may be present in the kinematic profile..

2.2. Two regression problems and DNN weights interpretation

The described architecture processes data in chunks of a prespecified length of N samples. We will first assume that the chunk length is equal to the filter length in the 1D convolution layers. Consider a chunk of input data from L channels observed over the interval of N time moments hat can be represented with a Toeplitz matrix $\mathbf{X}[n]=[\mathbf{x}[n],\mathbf{x}[n-1],\ldots,\mathbf{x}[n-N+1]]\in\mathbb{R}^{L\times N}$. Processing of $\mathbf{X}[n]$ by the first two layers performing spatial and temporal filtering can be described for the mth branch as

$$b_m[n]=\mathbf{w}_m^T\mathbf{X}[n]\mathbf{h}_m \tag{2}$$

where $\mathbf{w}_m\in\mathbb{R}^L$ is spatial weights and $\mathbf{h}_m\in\mathbb{R}^N$ is temporal weights of branch m. 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 [6] and for the original real-valued data, the imaginary part of the analytic signal is uniquely computed via Hilbert transform. Therefore, the original real-valued signal is uniquely mapped to its envelope. Our envelope detector computes a close approximation of the absolute value of the analytic signal and therefore we can state that $e_m[n]$ is uniquely determined by $b_m[n]$. Thus, in order to obtain the proper envelope $e_m[n]$ it suffices to obtain the proper $b_m[n]$ which is achieved by adjusting the spatial and temporal convolution weights of each branch of the compact CNN.

Assume that the training of the adaptive envelope detectors resulted in optimal spatial and temporal convolution weights marked with asterisks, $\mathbf{w}_m^*$ and $\mathbf{h}_m^*$ correspondingly. Let us also assume that

these optimal weights indeed extract the ground-truth population activity signals $b_m^*[n]$ that uniquely determine the envelopes $e_m^*[n]$ that in turn give rise the sought kinematics $z[n]$ when transformed with a non-linear operator $H()$ approximated by the fully connected layer of our network. Imaging that the spatial filter weights are not known but the temporal convolution weights is fixed to its optimal value $\mathbf{h}_m^*$. Then, we can find the optimal spatial weights as the solution to a convex optimization problem formulated over spatial subset of parameters:

$$\begin{aligned}\mathbf{w}_m^* &= \text{argmin}_{\mathbf{w}_m}\{\| b_m^*[n] - \mathbf{w}_m^T\mathbf{X}[n]\mathbf{h}_m^* \|_2^2\} \\ &= \text{argmin}_{\mathbf{w}_m}\{\| b_m^*(n) - \mathbf{w}_m^T\mathbf{y}_m[n] \|_2^2\} \end{aligned} \quad (3)$$

where the temporal weights are fixed at their optimal value, $\mathbf{h}_m^*$, and $\mathbf{y}_m[n] = \mathbf{X}[n]\mathbf{h}_m^*$ is a temporally filtered vector of multichannel data. Similarly, when spatial weights are fixed at the optimal value $\mathbf{w}_m^*$, the temporal weights are expressed by the equation:

$$\begin{aligned}\mathbf{h}_m^* &= \text{argmin}_{\mathbf{h}_m}\{\| b_m^*[n] - \mathbf{w}_m^{*T}\mathbf{X}[n]\mathbf{h}_m \|_2^2\} \\ &= \text{argmin}_{\mathbf{h}_m}\{\| b_m^*[n] - \mathbf{v}_m^T[n]\mathbf{h}_m \|_2^2\} \end{aligned} \quad (4)$$

where $\mathbf{v}_m[n] = [v_m[1], \ldots, v_m[N]]^T = \mathbf{X}^T[n]\mathbf{w}_m^*$ is a spatially filtered chunk of incoming data.

Given the forward model (1) and the regression problem (3) and assuming mutual statistical independence of the rhythmic potentials $s_m[n]$, $m = 1, \ldots, M$, the topographies of the underlying neuronal populations can be found as [8, 11]

$$\mathbf{g}_m = \mathbb{E}\{\mathbf{y}_m[n]\mathbf{y}_m^T[n]\}\mathbf{w}_m^* = \mathbf{R}_m^y\mathbf{w}_m^* \quad (5)$$

where $\mathbf{R}_m^y = \mathbb{E}\{\mathbf{y}_m[n]\mathbf{y}_m^T[n]\}$ is a $L \times L$ spatial covariance matrix of the temporally filtered data, assuming that channel timeseries are zero-mean random processes, L is the number of input channels. Thus, when interpreting individual spatial weights corresponding to each of the M branches of the architecture shown in figure 2 one has to take into account the temporal filter weights $\mathbf{h}_m^*$ this mth branch is tuned for. Therefore, to transform the spatial weights of different branches into spatial patterns, branch-specific spatial covariance matrices $\mathbf{R}_m^y$ should be used that depend on the temporal convolution weights in each particular branch. This becomes obvious if one remembers that the spatial and temporal filtering operations are linear and can be interchanged. However, since each of the branches has its own temporal filter, when depicting the architecture it is more convenient to place the spatial unmixing step first followed by the temporal filter. Considering a single branch, switching the order and first applying the same temporal filter to all data channels followed by the spatial filter makes expression (5) intuitive based on the derivations presented in [8]. Clearly, when interpreting the spatial filter weights of the mth branch one has to use spatial covariance of the data filtered with the temporal filter specific to the mth branch.

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

$$\mathbf{q}_m = \mathbb{E}\{\mathbf{v}_m[n]\mathbf{v}_m^T[n]\}\mathbf{h}_m^* = \mathbf{R}_m^v\mathbf{h}_m^* \quad (6)$$

where $\mathbf{R}_m^v = \mathbb{E}\{\mathbf{v}_m[n]\mathbf{v}_m^T[n]\}$ is an $N \times N$ tap covariance matrix of the spatially filtered data, assuming that channel timeseries are zero-mean random processes, N is the number of taps in the temporal convolution filter and the length of the data chunk processed at a time. 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 $\mathbf{w}_m$ that are used to feed the individual mth branch. To transform the temporal convolution weights of different branches into temporal patterns, branch-specific tap covariance matrices $\mathbf{R}_m^v$ should be used that depend on the spatial point-wise convolution weights of each particular branch. To make sense out of the temporal pattern vector $\mathbf{q}_m = [q_m[0], \ldots q_m[N-1]]^T$, we usually explore it in the frequency domain, i.e. $Q_m[k] = \sum_{n=0}^{N-1} q_m[n]e^{-j2\pi kn/N}$, $k = 0, \ldots, N-1$, which is the discrete Fourier transform (DFT) applied to the finite length temporal filter pattern vector $\mathbf{q}_m$ of the mth branch, n is the discrete time index used to refer to the elements of $\mathbf{q}_m$ and k indexes frequency bins.

It is important to realize that the temporal pattern $Q_m[k]$ of a neuronal population is simply power spectral density of the activity of this neuronal population. Together with the spatial pattern it fully describes both second order dynamical properties and the spatial location of a neuronal population. Similarly to the spatial filter weights that are not equal to the spatial pattern [8], the temporal filter weights and their Fourier representation do not reflect the dynamical properties of the underlying neuronal population.

In the above for illustrative purpose we used temporal embedding to describe operation of the convolutional layer of the CNN. This allowed us to emphasize the formal similarity between the temporal and spatial dimensions and illustrate that the interpretation of temporal patterns requires taking into account the correlation structure of the independent variable in the regression model built on the top of the spatially filtered data.

A more general and practical treatment can be given if we consider operation of this architecture on an extended piece of a signal so that the spatial and temporal filters are applied to the sliding window of data. In this case, operation of the band-pass temporal filter in the mth branch can be expressed as a convolution between the two sequences, i.e. $b_m[n] = h_m[n] \times v_m[n]$, see also figure 2 and compare this expression to the definition of $\mathbf{v}_m[n]$ immediately after expression (4).

The described architecture performs both spatial and temporal filtering to isolate pivotal population activity signals $s_m[n]$. To work with convolution and to derive a general rule for temporal pattern defined as power spectral density of the activity $s_m[n]$ of the task related neuronal population the mth branch is tuned to it is easier to operate in the frequency domain and use the standard Wiener filtering arguments. We assume here that as a result of training the spatial filter is settled at the optimal value $\mathbf{w}^*[n]$ and that the temporal filters weights $\mathbf{h}_m$ are Wiener optimal and provide the best estimate of $s_m[n]$ in the least squares sense. In this case it is possible to connect the power spectral density of the underlying neuronal population activity signal $s_m[n]$ to the spatially filtered sensor data $v_m^*[n]$. Spatially filtered $v_m^*[n] = \mathbf{w}_m^{*T}\mathbf{x}[n]$ represents a noisy mixture of the unknown pivotal population activity signal $s_m[n]$ and the activity $u_m[n] = \mathbf{w}_m^{*T}\boldsymbol{\eta}[n]$ of other task-unrelated populations $A_1, \ldots, A_J$. In other words:

$$v_m^*[n] = \mathbf{w}_m^{*T}\mathbf{x}[n] = s_m[n] + u_m[n]. \tag{7}$$

In the frequency domain, the weights of the Wiener filter $h_m^*[n]$ that is designed to extract signal useful for the posed decoding task from the noisy spatially filtered $v_m^*[n]$ can be expressed as a function of the power spectral density $P_m^{v^*v^*}[k]$ of $v_m^*[n]$ and the cross-spectrum, $P_m^{sv^*}[k]$, between the unknown underlying neuronal population activity signal $s_m[n]$ and the spatially filtered sensor data $v_m^*[n]$ [24]:

$$H_m^*[k] = \frac{P_m^{sv^*}[k]}{P_m^{v^*v^*}[k]}. \tag{8}$$

Following the standard Wiener filter derivation logic, if we assume that $\boldsymbol{\eta}[n]$ and $\mathbf{s}[n]$ in (1) are statistically independent which in turns means that $s_m[n]$ and $v_m[n]$ in (7) are also statistically independent we get that $P^{sv^*}[k] = P_m^{ss}[k]$ and $P_m^{v^*v^*}[k] = P_m^{ss}[k] + P_m^{uu}[k]$ we obtain the following expressions for the optimal Wiener filter transfer function $H_m^*[k]$:

$$H_m^*[k] = \frac{P_m^{ss}[k]}{P_m^{ss}[k] + P_m^{u}[k]} = \frac{P_m^{ss}[k]}{P_m^{vv}[k]}. \tag{9}$$

Therefore, the frequency-domain pattern $Q_m^*[k]$ which corresponds to the power spectral density P_m^{ss} of the neuronal population activity signal the mth branch is tuned to can be computed as

$$Q_m^*[k] = P_m^{ss}[k] = P_m^{vv}[k]H_m^*[k] \tag{10}$$

where $H_m^*[k]$ in (10) is the Fourier transform of the vector $\mathbf{h}_m^*$ containing temporal-convolution weights identified during the adaptation of the envelope detector in the mth branch. Viewing this result as a product of learning, we can say that learning to decode we gain access to the spectral pattern of activity (10) of the neuronal population critical for the particular decoding task. Importantly, the temporal filter weights alone will be informative of the population activity spectral pattern only in case the input timeseries $v_m^*[n]$ are white. We would also like to note that expression (10) is the frequency domain equivalent of equation (6) obtained in the temporal embedding format.

The spatial patterns $\mathbf{g}_m$ of neuronal sources reconstructed from the spatial filtering weights [8] determine spatial location of a neuronal population and are routinely used for dipole fitting to localize functionally important neural sources [20]. The spectral patterns interpreted according to (10) and (6) can be used to fit the models of neural population dynamics, which are relevant to specific decoding tasks, see for example [22].

2.3. Simulations

To explore the performance of the proposed approach, 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 $s_i[n]$. The potentials of these four task-related populations were generated as narrow-band processes in the lower to higher gamma sub-bands (30–80 Hz, 80–120 Hz, 120–170 Hz and 170–220 Hz) obtained from filtering Gaussian pseudo-random sequences with a bank of FIR filters. We then simulated the kinematics $z[n]$, as a linear combination of the four envelopes of these rhythmic signals with randomly generated vector of coefficients. We used 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. In each such band we simulated ten task-unrelated sources resulting into total of 40 task-unrelated sources. To simulate volume conduction effect and the way the task-related and task-unrelated source activity gets measured by the electrodes, we randomly generated a 4×5 dimensional forward matrix $\mathbf{G}$ and a 40×5 dimensional forward matrix $\mathbf{A}$. These matrices mapped the task-related and task-unrelated activity, respectively, onto the sensor space. For each Monte-Carlo trial we generated new mixing matrices $\mathbf{G}$, $\mathbf{A}$ and new source-time series. We have also added $1/f$ noise to the sensor data to simulate spatially uncorrelated brain noise.

We generated 20 min worth of data sampled at 1000 Hz and split them into two equal contiguous parts. We used the first part for training and the second for testing.

3. Experimental datasets

First, in order to compare the compact CNN architecture with the top linear models that rely on preset features, we used publicly available data collected by Kubanek *et al* from the Berlin BCI competition

IV. This dataset contains concurrent multichannel ECOG and finger flexion kinematics 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 sensorimotor cortex. The exact spatial locations and the order of the electrodes were not provided. As a baseline in this comparison, we chose the winning solution offered by Nanying Liang and Laurent Bougrain [16]. 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 next dataset comes from our Center for Bioelectric Interfaces (CBI) 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 synchronized 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, $z[n]$.

The recordings were obtained in two patients with pharmaco-resistant form of epilepsy; ECoG electrodes were implanted for the purpose of pre-surgical localization of epileptic foci and mapping of eloquent cortex. Thus, for these data, unlike in the case of Berlin BCI competition IV data we knew the cortical location of each electrode and were able to visualize spatial patterns of activity. 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.

Finally we apply our architecture to a 32-channel motor-imagery dataset recorded at sampling frequency of 250 Hz from a subject executing motor-imagery (MI) with total of four states: left hand MI, right hand MI, legs MI, rest. Recorded EEG data were split into 2 s long overlapping segments with 0.2 s step, so that each such segment referred to one motor state. The segments extracted from the first contiguous part of the recorded data was used for training and the rest was used for testing.

4. Results for simulated data

4.1. Adaptive envelope detector

As described in the section 2, to interpret optimal temporal convolution weights we need to consider the spectral characteristics of neural recordings. To illustrate this, we first used simplified simulations with one task-related source occupying 50–150 Hz frequency range and one task-unrelated source active within 50–100 Hz band which is a subrange of the task-related signal frequency band. We trained a single-channel ($M = 1$) adaptive envelope detector. 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 as it has a characteristic suppression over the frequency range occupied by the interference. The reduced gain in this frequency range can be understood from the expression of Wiener filter transfer function (9). It shows that the transfer function will have a gain smaller than 1 over the frequency range where an interference is present, i.e. $P_m^{\eta\eta}[k] \neq 0$. For simulations shown in figure 3 the interference occupied 50–100 Hz frequency range. At the same time, the expression in (10) allows us to obtain a proper pattern that matches well the simulated spectral profile. Conversely, using the DFT of the convolutional filter weights yields fundamentally erroneous estimates of the frequency-domain patterns and potentially erroneous interpretation of the underlying neurophysiology.

4.2. Realistic simulations

For the simulated data, we trained the algorithm to predict the kinematic variable $z[n]$. In the noiseless case, the proposed architecture achieved accuracy of 99% measured as correlation coefficient between the true and the estimated kinematics, see 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 shows estimated envelopes $r_m[n]$, superimposed on the true envelopes and the underlying narrow-band process $b_m[n]$ (figure 2). We can see a good agreement between the estimated and true envelope timeseries with pairwise correlation coefficients within 98%–99% range.

As described in the section 2, for spatial weights interpretation, we used the linear estimation theoretic approach [8, 11]. 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,

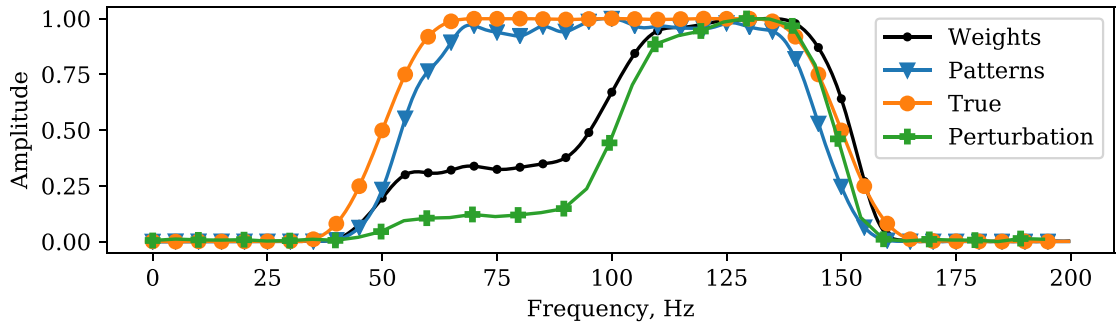

Figure 3. Three possible ways to interpret temporal convolution weights. True pattern of dynamic activity i.e. the power spectral density (PSD) (orange *) of the source. Fourier domain representation of the temporal convolution weights (black •), Ball's method (green +) and the dynamic source activity pattern reconstructed with the proposed approach (blue ▼).

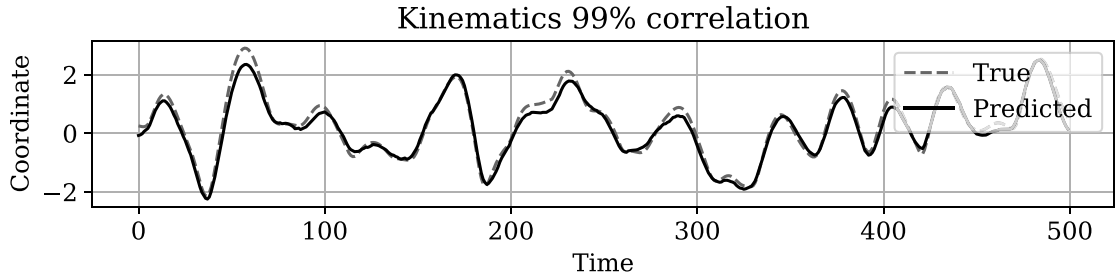

Figure 4. Realistic simulations. Actual and decoded kinematics.

we refer to the patterns determined using this approach as *Patterns naive*. 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* in the subsequent plots.

In the right column of figure 6, we show the results of reconstructing spatial patterns in the noiseless case for all four branches of the network. As expected, the spatial *Patterns naive* and *Patterns* are identical and match the ground truth exactly. The left column shows Fourier representations of the temporal patterns and 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.

In the noisy case demonstrated in figure 7, only *Patterns* match well with the simulated topographies of the underlying sources. Spectral characteristics of the trained temporal filtering weights exhibit characteristic deeps in the bands corresponding to the activity of the interfering sources. After applying expression (10), we obtain the spectral patterns that more closely match the simulated ones and have the deeps compensated.

4.3. Monte-Carlo simulations

In the above plots we showed two specific cases of this architecture operating in the noisy and noiseless cases for a fixed spatial configuration of task-related and task-unrelated sources as modelled by matrices **G** and **A**. To test the proposed approach for weights interpretation we performed Monte-Carlo study with different spatial configuration of sources at each trial and for the four different noise levels. To implement this, matrices **G** and **A** which model the volume conduction effects at each Monte Carlo trial were randomly generated according to $\mathcal{N}(0,1)$ distribution. We created 20 min worth of data sampled at 1000 Hz. For neural network training we use Adam optimiser. We made about 15k steps. At 5k and 10k step we halved the learning rate to get more accurate patterns. In total, we have performed more than 3k simulations. For each realization of the simulated data we have trained the algorithm to predict the kinematic variable $z[n]$ and then computed the patterns of sources the individual branches of our architecture got 'connected' to as a result of training. Figure 8 shows that only the spatial *Patterns* interpreted using branch-specific temporal filters (blue dots) match well the simulated topographies of the true underlying sources. The spectral patterns obtained using the proposed approach also appear to match well with the true spectral profiles of the underlying sources. Directly considering the Fourier coefficients of the temporal convolution weights results into generally erroneous spectral profiles (red triangles). For spatial patterns we also show the results for naively estimated patterns without taking into account branch-specific temporal filtering (green dots). Thus, using the proper spectral patterns of the underlying neuronal population it is now possible to fit biologically

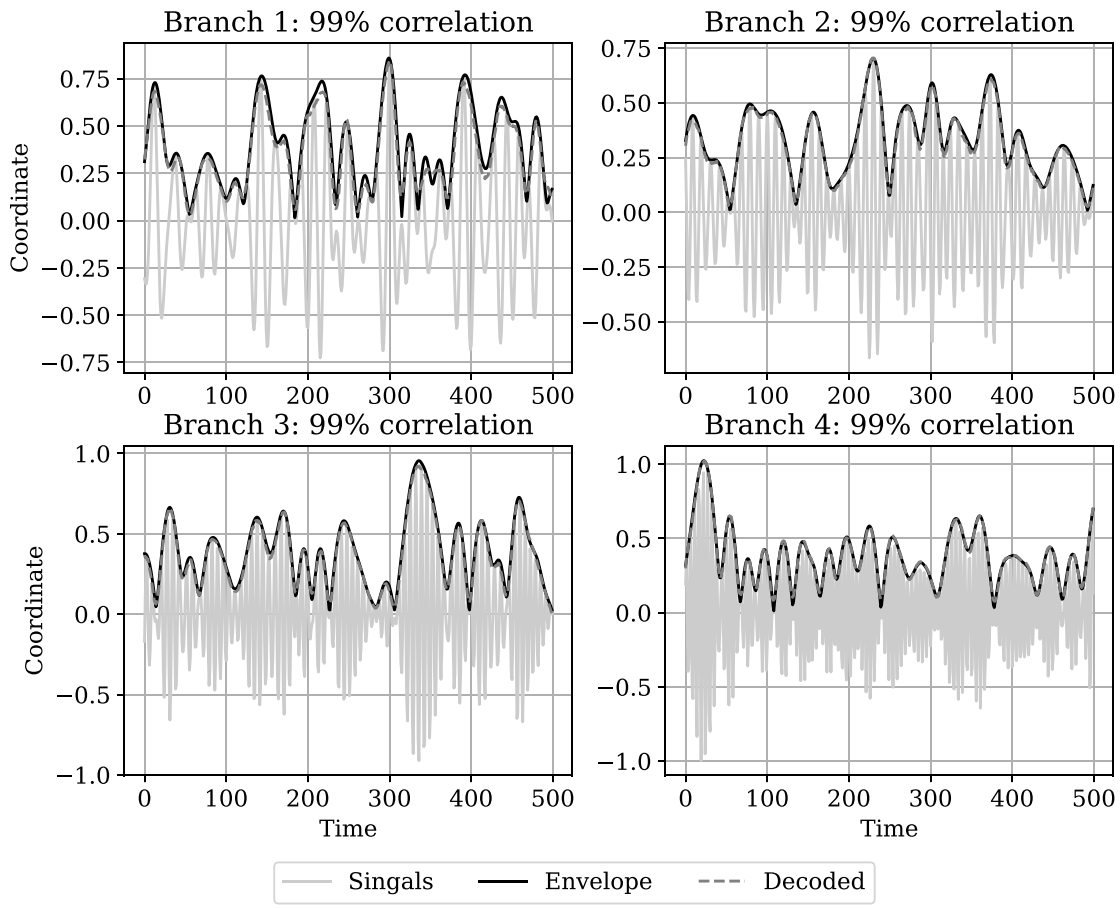

Figure 5. Branch envelopes decoding. Comparison between the true and the decoded envelope in each of the four branches.

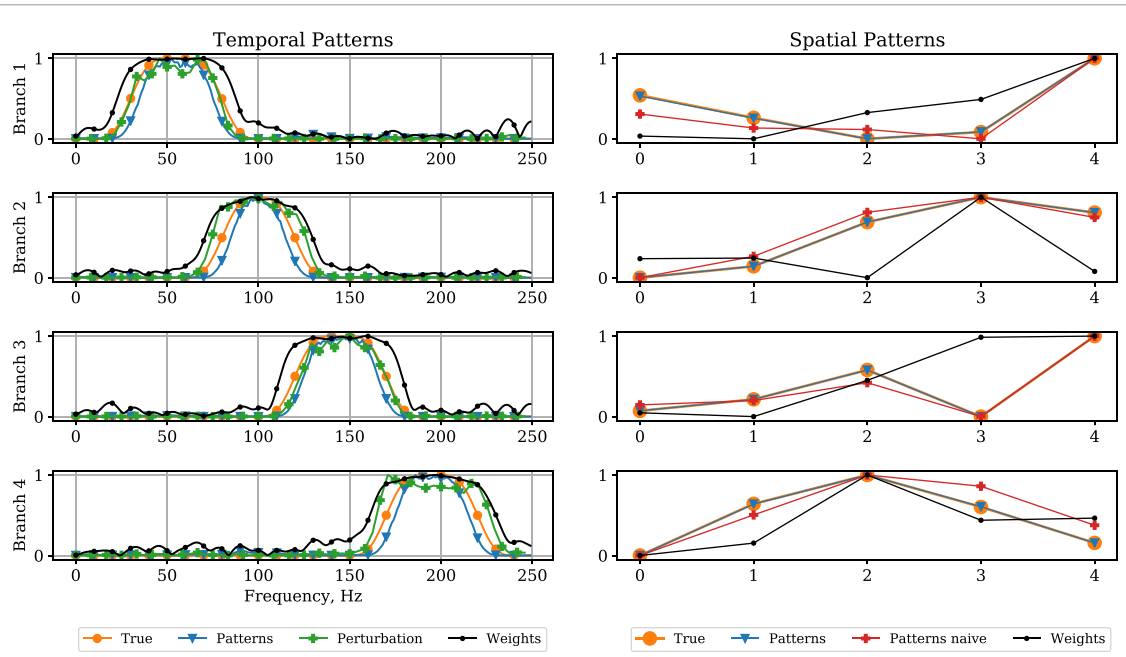

Figure 6. Temporal (left) and spatial (right) patterns obtained for the noiseless case. See the main text for description.

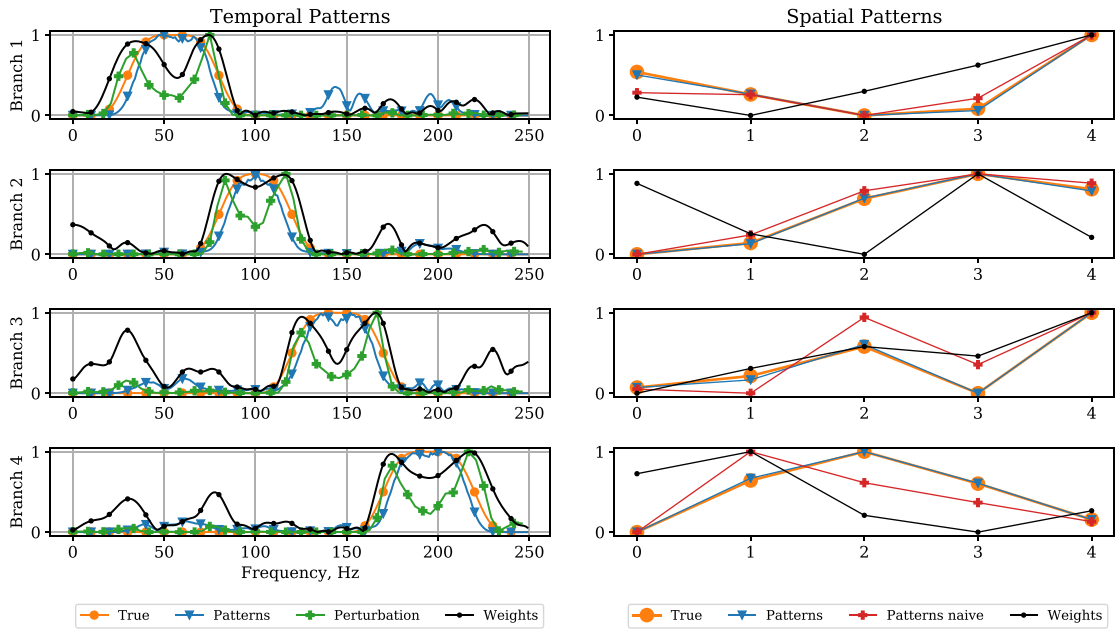

Figure 7. Temporal (left) and spatial (right) patterns obtained for the noisy case, SNR = 1.5. See the main text for description.

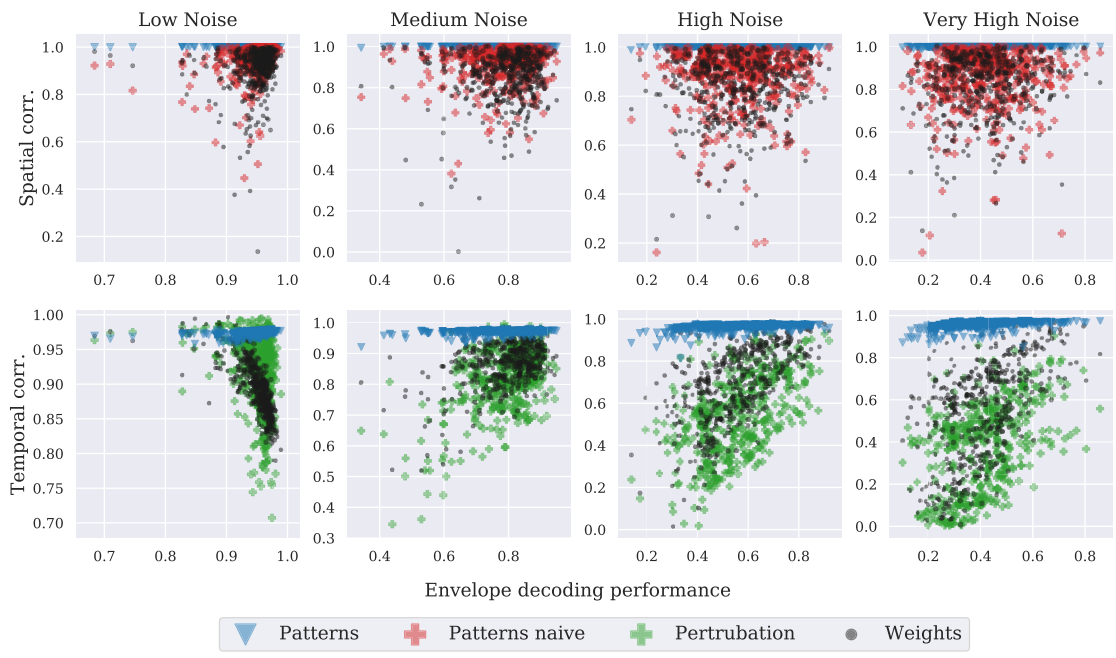

Figure 8. Monte-Carlo simulations. Point coordinates reflect the achieved envelope decoding performance (*x*-axis) and correlation coefficient with the true pattern (*y*-axis) at each Monte Carlo trial. Each point of a specific color corresponds to a single Monte Carlo trial and codes a method used to compute patterns. *Weights*—direct weights interpretation, *Patterns naive*—spatial patterns interpretation without taking branch specific temporal filters into account, *Patterns*—the proposed method.

plausible models, e.g. [22], and unravel true neurophysiological mechanisms underlying the decoded process.

5. Analysis of real 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 [31]. When applied to the data from Berlin BCI competition IV, we did not observe significant differences between the performance of the compact CNN and the winning solution by Lian and Bougrain [16] (Mann–Whitney test, $U = 103.0$, $p = 0.3543$), see table 1. Note, however, that the CNN performed feature engineering automatically.

Table 1. Comparison of the performance of the proposed architecture (NET) and the winning solution (Winner) of Berlin BCI competition IV dataset(4). $<<\pm>>$ denotes the standard deviation in the performance obtained with different launches of training process.

	Subject 1				
	Thumb	Index	Middle	Ring	Little
Winner	0.58	0.71	0.14	0.53	0.29
NET	0.54 ± 0.03	0.7 ± 0.02	0.2 ± 0.07	0.58 ± 0.03	0.25 ± 0.06
	Subject 2				
	Thumb	Index	Middle	Ring	Little
Winner	0.51	0.37	0.24	0.47	0.35
NET	0.5 ± 0.03	0.36 ± 0.04	0.22 ± 0.06	0.4 ± 0.04	0.23 ± 0.06
	Subject 3				
	Thumb	Index	Middle	Ring	Little
Winner	0.69	0.46	0.58	0.58	0.63
NET	0.71 ± 0.02	0.48 ± 0.03	0.5 ± 0.02	0.52 ± 0.02	0.61 ± 0.02

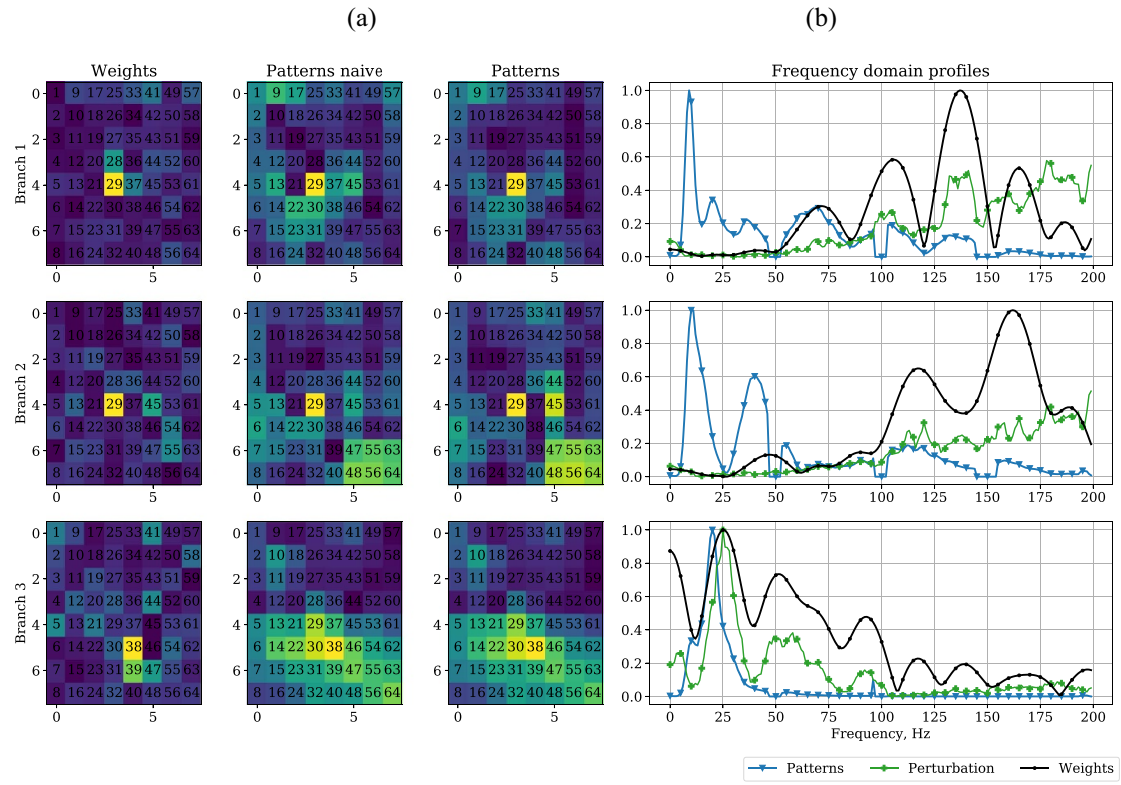

Figure 9. Network weights interpretation for the index finger kinematics decoder in CBI patient 1 (ECOG). Each row of plots corresponds to one of the three branches of the trained decoder. (a) The left most column shows color-coded spatial filter weights, next two columns correspond to naively and properly reconstructed spatial patterns. Blue color corresponds to the minimum absolute activation and yellow to the maximum. (b) Temporal filter weights interpretation in the Fourier domain. FFT of filter weights—(black •), power spectral density (PSD) $Q_m^*[k]$ pattern of the underlying LFP (blue ▼) obtained according to equation (10). Another line (red ◊) is the PSD of the signal at the output of the temporal convolution block. Results of sensitivity analysis using the perturbation approach [4, 34] are shown in (green +). See also Figure 11 for this spatial patterns overlayed on the cortex.

5.2. Dataset 2

We also applied the proposed solutions to the recordings conducted in two patients implanted with 8×8 ECoG microgrids grids placed over the sensorimotor cortex and performing individual finger flexion tasks, see section 3.

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

Figures 9 and 10 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 9, we show the spatial filter weights, naive patterns and proper patterns interpreted using the

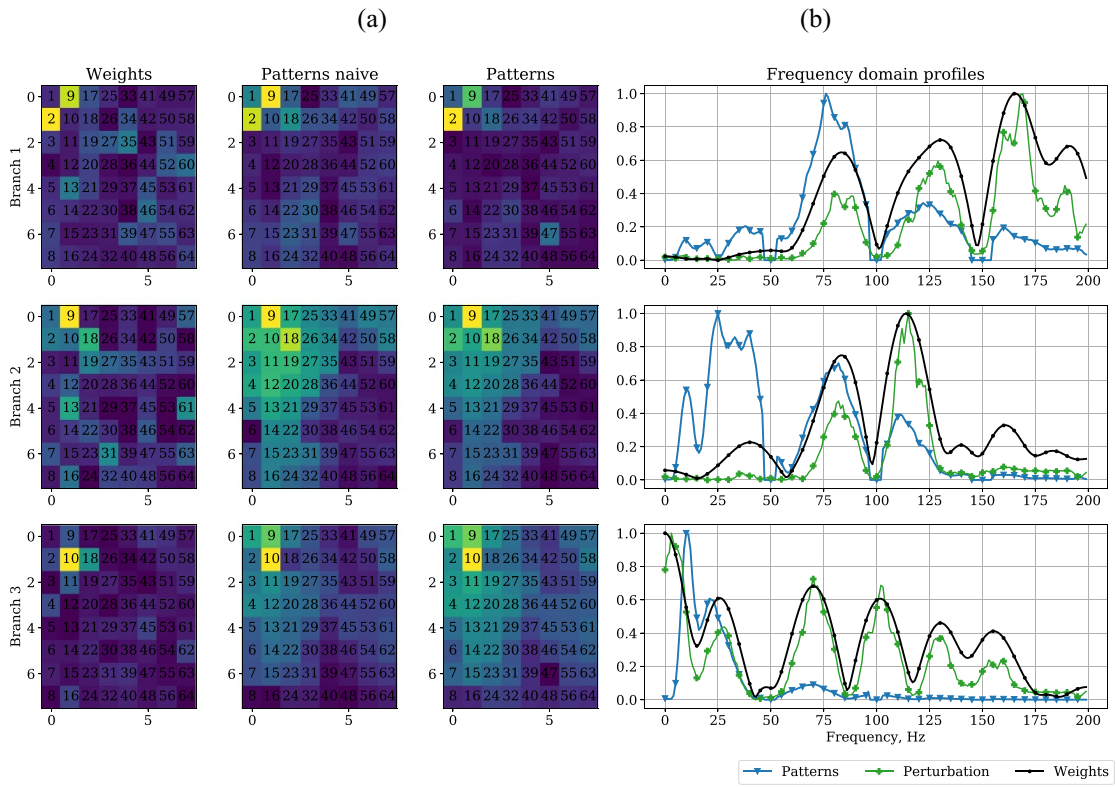

Figure 10. Network weights interpretation for the little finger kinematics decoder in CBI patient 2 (ECOG). Each row of plots corresponds to one of the three branches of the trained decoder. (a) The left most column shows color-coded spatial filter weights, next two columns correspond to naively and properly reconstructed spatial patterns. Blue color corresponds to the minimum absolute activation and yellow to the maximum. (b) Temporal filter weights interpretation in the Fourier domain. FFT of filter weights—(black •), power spectral density (PSD) $Q_m^*[k]$ pattern of the underlying LFP (blue ▼) obtained according to equation (10). Another line (red ◊) is the PSD of the signal at the output of the temporal convolution block. Results of sensitivity analysis using the perturbation approach are shown in (green +).

Table 2. Decoding performance achieved in the two CBI patients. The table shows correlation coefficient between the actual and the decoded finger trajectory for the four fingers in two patients. $<<\pm>>$ denotes the standard deviation in the performance obtained with different launches of training process.

	Thumb	Index	Middle	Ring	Little
Subject 1	0.48 ± 0.04	$\mathbf{0.79} \pm 0.02$	0.77 ± 0.03	0.61 ± 0.03	0.32 ± 0.05
Subject 2	0.73 ± 0.02	0.55 ± 0.03	0.72 ± 0.03	0.78 ± 0.02	$\mathbf{0.79} \pm 0.01$

expression described in the section 2. It can be seen that, while the temporal filter weights (black •) clearly emphasized the frequency range above 100 Hz in the first two branches, the actual spectral pattern of the source (blue ▼) 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. Similar observations can be made from figure 10 that shows to the decoding results for the little finger in patient 2. Interestingly, the second branch frequency domain pattern (blue ▼) appears to be significantly different from that obtained by a simple DFT of the weights vector (black •) and contains contributions from the lower 20–45 Hz frequency range. When fitting dynamical models of population activity to this reconstructed frequency domain pattern, the low-frequency components are likely to significantly affect the parameters of the corresponding dynamical model.

5.3. Dataset 3

Unlike the previous two datasets, which required continuous trajectory decoding from invasive ECoG, the dataset 3 was recorded non-invasively using a discrete state EEG motor-imagery paradigm as described in section 3. The challenge here was to classify the type of performed motor imagery. Given short duration of this data the compact CNN architecture learned the task quite well and yielded on average 0.83 ROC AUC for a single task as measured on the testing segmented represented by the second half of the

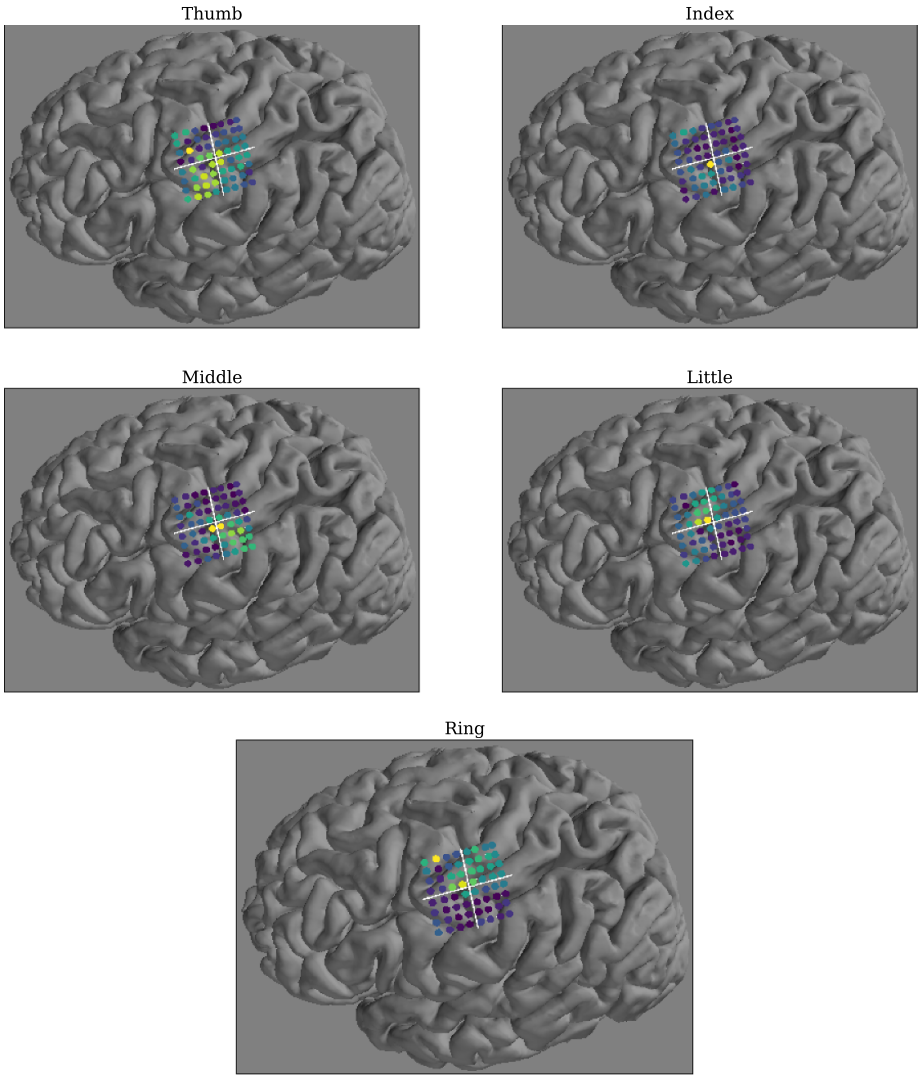

Figure 11. Spatial patterns mapped onto cortex. Blue color corresponds to the minimum absolute activation and yellow to the maximum.

Table 3. Decoding performance achieved for the third CBI patient (EEG). The table shows ROC AUC for different classes $<<\pm>>$ denotes the standard deviation in the performance obtained with different launches of training process.

	Legs	Right	Left	Rest
Subject 3	0.75 ± 0.05	0.87 ± 0.06	0.88 ± 0.04	0.81 ± 0.05

data temporally non-overlapping with the training segment.

The results of weights interpretation are shown in figure 12. The first automatically found pattern reflects the lateralized occipital alpha activity earlier know to be associated with motor tasks [12, 28]. As shown in panel (b) the use of the novel approach for interpreting temporal weights clearly indicates the presence of a well defined alpha-band source. The second branch got tuned to the sensori-motor cortex hand representation area. As evident from the Patterns curves representing the power spectral density of the activity of the second pivotal population the source exhibits well defined and physiologically plausible lower and upper (10, 20 Hz) sensorimotor rhythm components. Exploring the FFT of weights gives a drastically different estimate of the PSD of the pivotal population activity and focuses on rather high 35–40 Hz frequency range. We wold also like to note that the spatial pattern reconstructed using the naive approach (the middle column of panel (a)) appears to be misleading and can not be attributed to hand's sensori-motor cortex. The last pattern appeared to highlight a source located in leg's sensori-motor cortex and active in alpha-band. We emphasize that these patterns (both spatial and temporal) were automatically recovered just using the data and the specific decoding task. As evident from the above description, exploration of these patterns allowed us to get insights into the anatomic and dynamic properties of

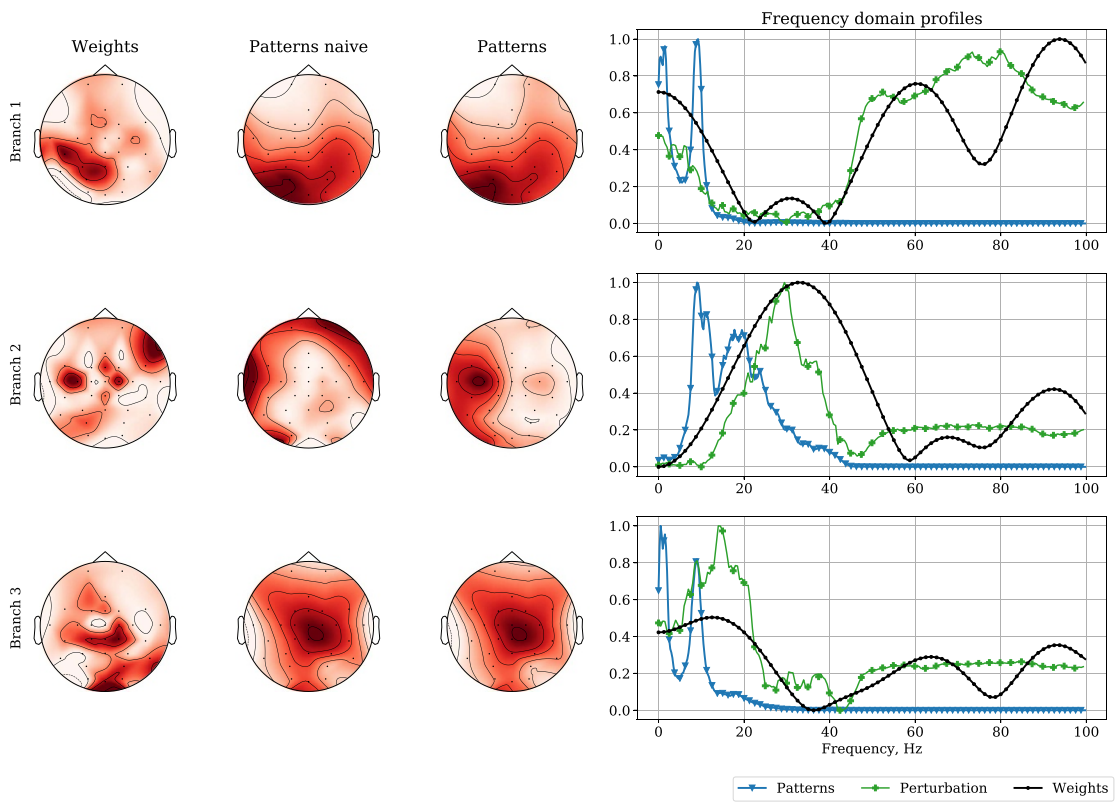

Figure 12. Network weights interpretation for the three branches (three rows of plots) of the decoder trained on a motor-imagery EEG dataset. (a) The left most column shows color-coded spatial filter weights, next two columns correspond to naively and properly reconstructed spatial patterns. White color corresponds to the minimum absolute activation and red to the maximum. (b) Temporal filter weights interpretation in the Fourier domain. FFT of filter weights—(black •), power spectral density (PSD) $Q_m^*[k]$ pattern of the underlying LFP (blue ▼) obtained according to equation (10). Another line (red ◊) is the PSD of the signal at the output of the temporal convolution block. Results of sensitivity analysis using the perturbation approach are shown in (green +).

the neuronal populations subserving motor-imagery mechanisms.

6. Conclusion

We first described a neurophysiologically interpretable architecture based on a compact convolutional network with separable spatial and temporal filtering similar to those reported earlier [13, 34, 38, 40]. Using this architecture, we extended the linear regression weights interpretation approach [8] to the analysis of the temporal convolution weights and adapted it to the architectures with multiple branches with specific spatial and temporal filters in each. We tested the proposed approach using realistically simulated and experimental data. To mimic a real-life scenario our simulated signals comprised activity of both task-related and task-unrelated populations that had to be separated by the decision rule in order to solve the kinematics decoding task. The activity of populations was simulated in the extended gamma range so that the activity of task-related neuronal populations would overlap with electrical signals generated by the task-unrelated sources. The realistically present overlap between the spatial topographies of task-related and task-unrelated sources was ensured by the ten-fold difference in the their quantity.

In the realistically simulated data, the compact CNN reconstructed with high accuracy the simulated neuronal substrate that contributed to the simulated kinematics data. Interestingly, as shown in figure 8, even in cases when the decoding accuracy was low, the spatial and temporal patterns appeared to be accurately estimated.

At the same time, the accuracy of spatial patterns reconstruction was noticeably higher than that for the temporal patterns (interpreted in the frequency domain). Such a behavior may partly stem from the fact that in our simulations the spatial patterns of each branch were encoded using only five coefficients as compared to 100-tap long temporal filters. One possible workaround that would reduce the number of the temporal convolution parameters to be identified is to use sinc-layer described in [29] and applied in a neural network for the analysis of acoustic signals. At the same time, such an approach reduces the flexibility in spectral shaping of individual branches of the network and require more pathways to be added to the network to achieve performance comparable to

that reported here which may complicate the interpretation of the obtained decision rule.

We have also applied the described architecture to Berlin BCI competition IV data. The compact CNN-based architecture delivered similar decoding accuracy compared to the winning solution of this BCI competition [16]. In contrast to the traditional approaches, the compact CNN architecture studied here did not require any additional feature engineering. On the contrary, after the architecture was trained to decode finger kinematics, we were able to interpret its parameters and extracted physiologically meaningful patterns corresponding to both spatial and temporal convolution weights. The latter was demonstrated using our own ECoG data collected from subjects performing a self-paced finger-flexion task for which we knew the spatial locations of ECoG electrodes.

There have been several reports describing compact DNN-based architectures similar to the one presented here. As we demonstrate, the main advantage of such simple architectures with a low count of layers and non-linear elements is the direct and theoretically justified tractability of their parameters in agreement with the linear estimation theory principle [11]. We keep calling these architectures *Deep* neural networks as they retain the distinctive property of the DNNs and use first several layers for feature engineering. In addition, our architecture is confined to derive physiologically meaningful features so that they can be interpreted with the approaches proposed here. In terms of spatial processing, that is combining the data from different sensors with specific weights, each branch of the described architecture, see figure 2, corresponds to the model studied in [8] and therefore as we show using the envelope uniqueness argument [6] this approach for weights interpretation is directly applicable to each of the branches. Our derivations show that when interpreting the spatial weights of each branch one needs to take into account the branch specific temporal filter. This may not be directly obvious when looking at figure 2 since for efficiency reasons the temporal convolution layer is placed past the spatial convolution layer. However, because of the linearity of both operations, the two can be swapped and then the linear estimation theory principle reiterated in [8] directly applies to the spatial weights of each branch receiving appropriately temporally prefiltered data with the temporal filter specific to this branch.

In [8] as one of the considered examples the authors describe the regression task where feature set represents a vector sampled from both spatial and temporal domains. To interpret the regression coefficients vector one needs to use the covariance matrix of the entire LT dimensional feature sampled over $L \times T$ spatial-temporal grid. Unlike in that example the processing preformed by the architecture considered here and in several other recent papers, see [31] for a review, is factorised into both spatial and temporal domains and feature extraction of each branch is served by $L + T$-dimensional vector. In this setting, as we have shown, given the fixed weights in one of the domains (spatial or temporal) the problem of identifying the weights in the other domain is a simple regression task whose weights need to be interpreted according to the linear estimation theory principle. Formally, the spatial and temporal domain weights are identically treated. Therefore, in order to interpret the temporal filter weights and discover temporal (usually visualized in the frequency domain) patterns one also needs to take into account the correlation structure of the input signal. The correlation structure of a signal in time domain is represented by the signal's autocorrelation and equivalently by the power spectral density. As we show and in agreement with the optimal filter theory the frequency domain patterns of activity of the pivotal populations each branch is tuned to can be reconstructed by computing the product between the Fourier representation of temporal filter weights and the power spectral density of the input data filtered with the branch-specific spatial filter.

Traditionally, only the Fourier transform of the temporal filter weights is considered in the context of interpretation of the decision learnt by the adaptive architectures. In the context of motor activity decoding it has been consistently found that gamma band oscillations hallmark the movements and tracking the power in this band can be used to reconstruct the movement trajectory from activity of the sensorimotor cortex. These very observations are also reported in several papers that apply compact architectures for analysis of neural data [34, 40]. However, according to equation (9) it does not in general mean that lower bands are not informative and that the pivotal population is not active in the lower frequency bands and that its activity is not co-modulated with the movement. All this means is that possibly task-related activity of the pivotal population in the lower frequency range is contaminated by the task-unrelated activity of other populations. Here, we for the first time show how using Wiener optimality arguments applied to the interpretation of CNN's temporal convolution weights it is possible to reconstruct the entire power-spectral density (PSD) of the neuronal populations pivotal to the decoding task. The obtained PSD profiles can then be used to fit physiologically justified models such as those described in [22].

Our simulations show that the properly interpreted solutions result into the ground truth activity patterns. However, the uniqueness of the decision rule found by the compact neural architecture described here needs to be studied in further details. For example, a simple analysis shows that in cases when the temporal filters of two branches of our CNN architecture are identical the spatial weights are no longer guaranteed to be unique. Therefore a

rigorous proof of the decision rule uniqueness is an important step that needs to be made in the future in order to take full advantage of the knowledge extraction approach proposed here.

In the present study we did not explicitly address the choice of the number of branches M. In our numerical experiments with simulated data we used the number of branches informed by our simulations setting. In practice we followed a conservative strategy, started from a low number of branches (1–2) and then increased this number until the performance plateau was reached. Note that the resulting number of branches determined this way depends not only on the actual number of populations pivotal to the task but also on the amount of data available for training. This is because increasing the number of branches we also increase the number of parameters to be identified from the same amount of data, which is reflected in the decoding accuracy estimated on an independent data sample. Theoretically, given the ample amount of training data the number of branches is to be equal to the number of distinct neuronal populations subserving the decoded process.

To obtain the results presented we standardized the data, i.e. subtracted the mean and divided by the standard deviation. Therefore, when applied in real-time special care needs to be taken to keep track of the mean and maintain the specific dynamic range of the data. The compact CNN architecture that we have explored here was designed to operate causally, i.e. using only the neuronal activity data from the past to predict the current value of the kinematic parameter or motor imagery state. At the same time, fundamentally, the output of the causal temporal filter layer in this CNN architecture is always delayed with respect to the input signal. Training this architecture to estimate the *current* (not delayed) kinematic parameter value using the data from the immediate past forces the CNN to learn to *predict* the kinematic variable value and thus compensate for the amount of the delay imposed by the temporal filters. Although we did not explicitly study this phenomenon here, the above considerations lead us to think that we can count on the delay-free kinematics decoding when such architectures are used in a real-time processing mode.

Acknowledgment

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

ORCID iDs

Mikhail Sinkin https://orcid.org/0000-0001-5026-0060

Alexei Ossadtchi https://orcid.org/0000-0001-8827-9429

References

[1] Abdulkader S N, Atia A and Mostafa M-S M 2015 Brain computer interfacing: applications and challenges *Egypt. Inform. J.* **16** 213–30

[2] Bießmann F, Murayama Y, Logothetis N K, Müller K-R and Meinecke F C 2012 Improved decoding of neural activity from fMRI signals using non-separable spatiotemporal deconvolutions *Neuroimage* **61** 1031–42

[3] Blankertz B, Lemm S, Treder M, Haufe S and Müller K-R 2011 Single-trial analysis and classification of ERP components—a tutorial *Neuroimage* **56** 814–25

[4] Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32

[5] Chaudhary U, Birbaumer N and Ramos-Murguialday A 2016 Brain–computer interfaces for communication and rehabilitation *Nat. Rev. Neurol.* **12** 513

[6] Hahn S L 2003 On the uniqueness of the definition of the amplitude and phase of the analytic signal *Signal Process.* **83** 1815–20

[7] Hatsopoulos N G and Donoghue J P 2009 The science of neural interface systems *Ann. Rev. Neurosci.* **32** 249–66

[8] Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J-D, Blankertz B and Bießmann F 2014 On the interpretation of weight vectors of linear models in multivariate neuroimaging *Neuroimage* **87** 96–110

[9] Homer M L, Nurmikko A V, Donoghue J P and Hochberg L R 2013 Sensors and decoding for intracortical brain computer interfaces *Annu. Rev. Biomed. Eng.* **15** 383–405

[10] Hyvärinen A, Hurri J and Hoyer P O 2009 *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision* vol 39 (Berlin: Springer)

[11] Kay S M 1997 *Fundamentals of Statistical Signal Processing: Estimation Theory* (Englewood Cliffs, NJ: Prentice-Hall)

[12] Kornhuber H H and Deecke Luder 2016 Brain potential changes in voluntary and passive movements in humans: readiness potential and reafferent potentials *Pflügers Arch.* **468** 1115–24

[13] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2016 EEGNet: a compact convolutional network for EEG-based brain–computer interfaces (arXiv:1611.08024)

[14] Lebedev M A and Nicolelis M A L 2017 Brain-machine interfaces: from basic science to neuroprostheses and neurorehabilitation *Physiol. Rev.* **97** 767–837

[15] Lemm S, Blankertz B, Dickhaus T and Müller K-R 2011 Introduction to machine learning for brain imaging *Neuroimage* **56** 387–99

[16] Liang N and Bougrain L 2012 Decoding finger flexion from band-specific ECoG signals in humans *Frontiers Neurosci.* **6** 91

[17] Machado S *et al* 2010 EEG-based brain–computer interfaces: an overview of basic concepts and clinical applications in neurorehabilitation *Rev. Neurosci.* **21** 451–68

[18] Mak J N and Wolpaw J R 2009 Clinical applications of brain–computer interfaces: current state and future prospects *IEEE Rev. Biomed. Eng.* **2** 187–99

[19] Meirovitch Y, Harris H, Dayan E, Arieli A and Flash T 2015 Alpha and beta band event-related desynchronization reflects kinematic regularities *J. Neurosci.* **35** 1627–37

[20] Mosher J, Leahy R and Lewis P 1999 EEG and MEG: forward solutions for inverse methods *IEEE Trans. Biomed. Eng.* **46** 245–59

[21] Naselaris T, Kay K N, Nishimoto S and Gallant J L 2011 Encoding and decoding in fMRI *Neuroimage* **56** 400–10

[22] Neymotin S A *et al* 2020 Human neocortical neurosolver (HNN), a new software tool for interpreting the cellular and network origin of human MEG/EEG data *eLife* **9** e51214

[23] Nicolas-Alonso L F and Gomez-Gil J 2012 Brain computer interfaces, a review *Sensors* **12** 1211–79

[24] Oppenheim A V and Verghese G C 2010 *Signals, Systems and Inference* 1st Edition (London: Pearson) 608
[25] Pais-Vieira M, Lebedev M, Kunicki C, Wang J and Nicolelis M 2013 A brain-to-brain interface for real-time sharing of sensorimotor information *Sci. Rep.* **3** 1319
[26] Parra L, Alvino C, Tang A, Pearlmutter B, Yeung N, Osman A and Sajda P 2003 Single-trial detection in EEG and MEG: keeping it linear *Neurocomputing* **52** 177–83
[27] Parra L C, Spence C D, Gerson A D and Sajda P 2005 Recipes for the linear analysis of EEG *Neuroimage* **28** 326–41
[28] Pfurtscheller G and Aranibar A 1978 Occipital rhythmic activity within the alpha band during conditioned externally paced movement *Electroencephalogr. Clin. Neurophysiol.* **45** 226–35
[29] Ravanelli M and Bengio Y 2018 Speaker recognition from raw waveform with SincNet *2018 IEEE Spoken Language Technology Workshop (SLT)* (IEEE) pp 1021–8
[30] Reichert J, Kober S, Neuper C and Wood G 2015 Resting-state sensorimotor rhythm (SMR) power predicts the ability to up-regulate SMR in an EEG-instrumental conditioning paradigm *Clin. Neurophysiol.* **126** 2068–77
[31] Roy Y, Banville H J, Albuquerque I, Gramfort A, Falk T H and Faubert J 2019 Deep learning-based electroencephalography analysis: a systematic review *J. Neural Eng.* **16** 051001
[32] Sabbagh D, Ablin P, Varoquaux G, Gramfort A and Engemann D A 2020 Predictive regression modeling with MEG/EEG: from source power to signals and cognitive states *Neuroimage* **222** 116893
[33] Schalk G and Leuthardt E C 2011 Brain–computer interfaces using electrocorticographic signals *IEEE Rev. Biomed. Eng.* **4** 140–54
[34] Schirrmeister R T, Springenberg J T, Fiederer L D J, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W and Ball T 2017 Deep learning with convolutional neural networks for EEG decoding and visualization *Hum. Brain Mapp.* **38** 5391–420
[35] Shin H, Law R, Tsutsui S, Moore C and Jones S 2017 The rate of transient beta frequency events predicts behavior across tasks and species *eLife* **6** e29086
[36] Shrikumar A, Greenside P and Kundaje A 2017 Learning important features through propagating activation differences 1704.02685
[37] Volkova K, Lebedev M A, Kaplan A and Ossadtchi A 2019 Decoding movement from electrocorticographic activity: a review *Front. Neuroinform.* **13** 74
[38] Waytowich N, Lawhern V J, Garcia J O, Cummings J, Faller J, Sajda P and Vettel J M 2018 Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials *J. Neural Eng.* **15** 066031
[39] Wolpert D and Ghahramani Z 2000 Computational principles of movement neuroscience *Nat. Neurosci.* **3** 1212–7
[40] Zubarev I, Zetter R, Halme H-L and Parkkonen L 2019 Adaptive neural network classifier for decoding MEG signals *Neuroimage* **197** 425–34