National Research University Higher School of Economics

as a manuscript

Artur Petrosyan

Modern Methods Of Machine Learning In The Interpretation Of Electrical Activity Of The Brain

PhD Dissertation Summary

for the purpose of obtaining an academic degree
Doctor of Philosophy in Computer Science

Academic Supervisor:
Professor, PhD
Alexei Ossadtchi

Moscow - 2022

The PhD dissertation was prepared at the National Research University Higher School of Economics.

Academic Supervisor: Alexei Ossadtchi, PhD, Director of the Centre for Bioelectric Interfaces at the Institute for Cognitive Neuroscience and Professor at the School of Data Analysis and Artificial Intelligence at the Faculty of Computer Science at the National Research University Higher School of Economics.

Contents

1 Introduction

1.1 Subject of the study

Brain-computer interfaces (BCIs) directly link the nervous system to external devices [51] or other brains [41]. While there exist many applications of BCIs [34], clinically relevant solutions are of primary interest since they hold promise to rehabilitate patients with sensory, motor, and cognitive disabilities [53],[31].

BCIs can deal with a variety of neural signals [44, 26] such as, for example, electroencephalographic (EEG) potentials sampled with electrodes located on the scalp [49], or neural activity recorded invasively with intracortical electrodes penetrating the cortex [40] or placed directly onto the cortical surface [48]. In general, methods of recording brain activity can be divided into invasive and non-invasive. In the first case, a medical procedure is assumed for implanting electrodes on the surface of the cerebral cortex (subdural or epidural) and subsequent counting and interpretation of signals of neural population activity. At the moment, interfaces that register brain activity in a non-invasive way do not provide the necessary width of the information channel. The amount of information in the invasively recorded signals far outweighs the complexities and ethical issues associated with this technology.

A promising and minimally invasive way to directly access cortical activity is to use stereotactic EEG (sEEG) electrodes inserted via a burr hole made in the skull. Recent advances in implantation techniques including the use of the brain's 3D angiography, MRI, and robot-assisted surgery help to further reduce the risks of such implantation and make sEEG technology an ideal trade-off for BCI applications [9]. ECoG strips are another method to achieve direct electrical contact with cortical tissue with minimal discomfort to a patient [21].

A step towards improving the work of neural interfaces is the use of advanced and advanced machine learning methods. From the arsenal of available methods, both classical models and deep learning methods can be used. The use of neural networks in many mathematical and medical problems shows good results compared to other methods, so an attempt to test them in predictive signal decoding tasks seems quite reasonable and promising [50] [46].

Nevertheless, one of the problems when decoding brain signals using deep learning algorithms is the interpretability of these methods. Typical BCI signal processing comprises several steps, including signal conditioning, feature extraction, and decoding. In modern machine learning algorithms, parameters of the feature extraction and decoding pipelines are jointly optimized within computational architectures called Deep Neural Networks (DNN) [46].

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 an 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, the interpretation of features computed by the first several layers of a DNN could shed light on the neurophysiological mechanisms underlying the behaviour 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.

Based on the foregoing, the main **object of the thesis** is machine learning methods and, in particular, deep learning methods used in the tasks of decoding brain signals, as well as their interpretation and construction of interpreted architectures.

1.2 Objectives

From all of the above, it becomes obvious that improving machine learning methods for decoding data from brain signals is an **actual task** that directly affects the practical applicability of BCI, and the ability to interpret these methods guarantees the reliability of the results obtained and opens up new possibilities for studying the principles of the brain. The main **research objective** is the development of domain-informed architectures of neural networks in combination with the development of algorithms for interpreting the corresponding weight coefficients and the application to the tasks of decoding neuronal activity in ideomotor and speech neurointerfaces.

The dissertation research was carried out on the basis of the Center for Bioelectrical

Interfaces of the National Research University Higher School of Economics, in which work is underway to create invasive neural interfaces to replace motor and speech functions. The following **research objectives** were formulated:

1. To develop the architecture of a compact neural network, consistent with modern scientific data on the origin of electrophysiological activity, the mechanism of its propagation in tissues, and the physical principles of its registration.

2. To carry out a comparative analysis of the quality of decoding from ECoG and stereo-EEG data of kinematics of the finger and parameters of the articulatory tract, using the proposed compact neural network and other competing solutions.

3. To develop methods for interpreting weight coefficients in the proposed neural network architecture in order to identify the geometric characteristics of key populations of neurons and the dynamic properties of their activity.

4. To implement real-time hand movement kinematics decoding.

5. To implement speech decoding based on the minimum number of spatially segregated electrodes.

1.3 Main ideas, results and conclusions of the dissertation

We have described a compact architecture based on a convolutional network for adaptive decoding of electrocorticography (ECoG) data into finger kinematics and sEEG data into speech. We have also proposed a new theoretically grounded approach to the interpretation of spatial and temporal weights in architectures combining adaptation in both space and time. The resulting spatial and frequency patterns characterizing populations of neurons that are crucial for a specific decoding task are subject to further analysis using electromagnetic and dynamic models in order to characterize the localization and activity parameters of key neuron populations.

First, we tested our solution using a realistic Monte Carlo simulation. Then, in relation to ECoG data from the Berlin BCI Competition IV dataset, our architecture worked comparable to the winners of the competition, without requiring any manual data preprocessing. Using the proposed approach to the interpretation of network weights, we were able to reveal the spatial and spectral patterns of neural processes underlying the successful decoding of finger kinematics from the ECoG dataset. Finally, we also applied the method to the analysis of a 32-channel dataset of imaginary EEG movements and observed physiologically plausible patterns characteristic of the task. Also, we applied our real-time architecture on a real patient and achieved high-quality decoding of the kinematics of the patient's fingers solely from brain activity data. Relevant details are described in [11, 12].

We have also expanded the architecture and applied it to the task of decoding speech from invasive ECoG and stereo-EEG data. To do this, we collected 60 minutes of data (from two sessions) for each of the two patients who were implanted with invasive electrodes. We then used only electrodes related to one stereo EEG shaft or one ECoG strip to decode neural activity into 26 words and one class of silence. The interpretation of the network weights gave a physiologically plausible result, which coincided with the results of stimulation mapping.

We achieved an average of 58% accuracy using only 6 data channels recorded with one minimally invasive sEEG electrode in the first patient, and 72% accuracy using only 8 data channels recorded for one ECoG strip in the second patient in the classification of 26+1 spoken words. Our compact architecture did not require the use of pre-selected features, was quickly trained, and led to a stable, interpretable, and physiologically significant decision-making rule. Spatial characteristics of the main populations of neurons confirm the results of mapping of the active and passive speech and demonstrate the inverse spatial-frequency dependence characteristic for neural activity. When compared with other architectures, our compact solution worked at the level or even better than those solutions that were recently mentioned in the literature on neural speech decoding, while using many times fewer minimally invasive electrodes and trained on a compact amount of data.

We also analyzed the influence of intermediate representations of speech on the quality of the final classification and obtained approximately the same results, despite the fact that the intermediate representations themselves are decoded with different accuracy.

This study takes the first step towards minimally invasive speech decoding prostheses and demonstrates the fundamental possibility of their creation based on a minimally invasive technology for recording brain activity. The details of this study are described in [2].

1.4 Theoretical and practical significance

From a theoretical point of view, we could state the following results:

- For the first time, we justified the architecture of a neural network based on the generally accepted model in electrophysiology for monitoring the electrical activity of the brain using a distributed set of electrodes.

- For the first time, we proposed a theoretically substantiated method for interpreting the weights of a compact neural network with factorized space-time processing and conducted the necessary modeling to demonstrate the efficiency of the proposed method.

- We have demonstrated the physiology of the resulting spatial and frequency patterns characterizing key neuronal populations. The information obtained fully coincided with the results of an active study of the cerebral cortex of patients in order to search for the speech cortex. In the motor task, the somatotopy observed in spatial patterns is fully consistent with the established idea of the organization of the motor cortex.

From a practical point of view, we achieved the following:

- We implemented a prototype of an invasive motor neurointerface in real-time.

- We proposed architecture and methodology for interpreting weight coefficients that can be used to build classifiers in neurophysiological studies. The interpretation of the weight coefficients of such classifiers makes it possible to obtain new knowledge about the studied neurophysiological processes.

- In addition, we implemented and tested a system for decoding speech from ECoG data. Our algorithm worked in a causal mode, that is, it used data from the past in relation to the decoding time. This allows us to hope for a successful transfer of the achieved quality of the work of our decoder in a real patient with impaired speech function.

- Also, we explored the possibility of our speech interface working in an asynchronous mode, which is of great practical importance when translating our solution into clinical practice.

1.5 The author's contribution to the study

The author of this study is the developer of the proposed methodology and architecture of the neural network as applied to the analysis of the model and real data. The developed approach to interpreting the weights of a wide family of architectures was studied in detail by the author in Monte Carlo simulation mode. The author has obtained all the results concerning the accuracy of the proposed algorithms as applied to real data. The results of this work are described in two papers published in first-tier international journals and in three conference papers. In all these works, the author is the first and main author.

1.6 Publications and approbation of the work

1.6.1 First-tier publications:

- **Petrosyan A. et al.** Decoding and interpreting cortical signals with a compact convolutional neural network //**Journal of Neural Engineering (Q1)**. – 2021. – T. 18. – №. 2. – C. 026019 [6].

- **Petrosyan A. et al.** Speech Decoding From A Small Set Of Spatially Segregated Minimally Invasive Intracranial EEG Electrodes With A Compact And Interpretable Neural Network //**Journal of Neural Engineering (Q1)**. – 2022. – T.. – №. . – C. [2].

1.6.2 Second-tier publications:

- **Petrosyan A.**, Lebedev M., Ossadtchi A. Linear Systems Theoretic Approach to Interpretation of Spatial and Temporal Weights in Compact CNNs: Monte-Carlo Study //Biologically Inspired Cognitive Architectures Meeting (Q4). – Springer, Cham, 2020. – C. 365-370 [12].

- **Petrosyan A.**, Lebedev M., Ossadtchi A. Decoding neural signals with a compact and interpretable convolutional neural network //International Conference on Neuroinformatics (Q4). – Springer, Cham, 2020. – C. 420-428 [11].

- **Arthur Petrosyan**, Alexey Voskoboinikov, Alexei Ossadtchi, Compact and interpretable architecture for speech decoding from stereotactic EEG // 2021 Third International Conference Neurotechnologies and Neurointerfaces – IEEE, 2021. – C. 79-82 [4].

1.6.3 Other publications:

- **Petrosyan A. et al.** Compact and Interpretable Architecture for Speech Decoding From iEEG //International Journal of Psychophysiology. – 2021. – T. 168. – C. S195 [5].

- Volkova Ksenia, **Arthur Petrosyan**, Dubyshkin Ignatii, Ossadtchi Alexei, «decoding movement time-course from ecog using deep learning and implications for bidirectional brain-computer interfacing» [30].

1.6.4 Conferences and seminars:

- 2020 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence (BICA*AI 2020) “Linear systems theoretic approach to interpretation of spatial and temporal weights incompact CNNs: Monte-Carlo study” (2020).

- XXII International Conference "Neuroinformatics-2020" - «Decoding neural signals with a compact and interpretable convolutional neural network» (2020).

- BCI Samara - «Decoding neural signals with a compact and interpretable convolutional neural network» (2020).

- Report at the forum «Center for Bioelectric Interfaces» (2020).

- BCI Samara - «Compact and interpretable architecture for speech decoding from sEEG» (2021).

- 20th World Congress of Psychophysiology - “Compact and interpretable architecture for speech decoding from sEEG” (2021).
- The Third International Conference «Neurotechnologies and Neurointerfaces» - “Compact and interpretable architecture for speech decoding from sEEG” (2021).

2 Content of the work

2.1 The architecture of a compact neural network reflecting modern scientific ideas about the origin of neuroelectrophysiological activity

This section contains a summary of the article [12].

Contribution of the author: the architecture of the neural network was developed, a method for its interpretation was developed, computer simulations were implemented (including Monte Carlo simulations).

2.1.1 Phenomenological model

Figure 1 illustrates a hypothetical relationship between motor behavior (hand movements), brain activity, and ECoG recordings. The activity, $\mathbf{s}[n] = [s_1[n], \dots, s_I[n]]^T \in \mathbb{R}^I$, 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 $\mathbf{e}[n] = [e_1[n], \dots, e_I[n]]^T$ 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 the application of the forward-model matrices $\mathbf{G} = [\mathbf{g}_1[n], \dots, \mathbf{g}_I[n]] \in \mathbb{R}^{L \times I}$ and $\mathbf{A} = [\mathbf{a}_1[n], \dots, \mathbf{a}_J[n]] \in \mathbb{R}^{L \times J}$ to the column vector of activity of task-related sources at the time moment n, $\mathbf{s}[n] = [s_1[n], \dots, s_I[n]]^T$, and task-unrelated sources, $\mathbf{f}[n] = [f_1[n], \dots, f_J[n]]^T$, respectively:

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

Column vectors $\mathbf{g}_i,\ i = 1, \dots, I$ and $\mathbf{a}_j,\ j = 1, \dots, J$ are the topographies of the task related and task-unrelated sources. We refer to the noisy, task-unrelated component of the recording as $\eta[n] = \sum_{j=1}^{J} \mathbf{a}_j f_j[n] \in \mathbb{R}^L$. A similar generative model has been recently described in [13].

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.

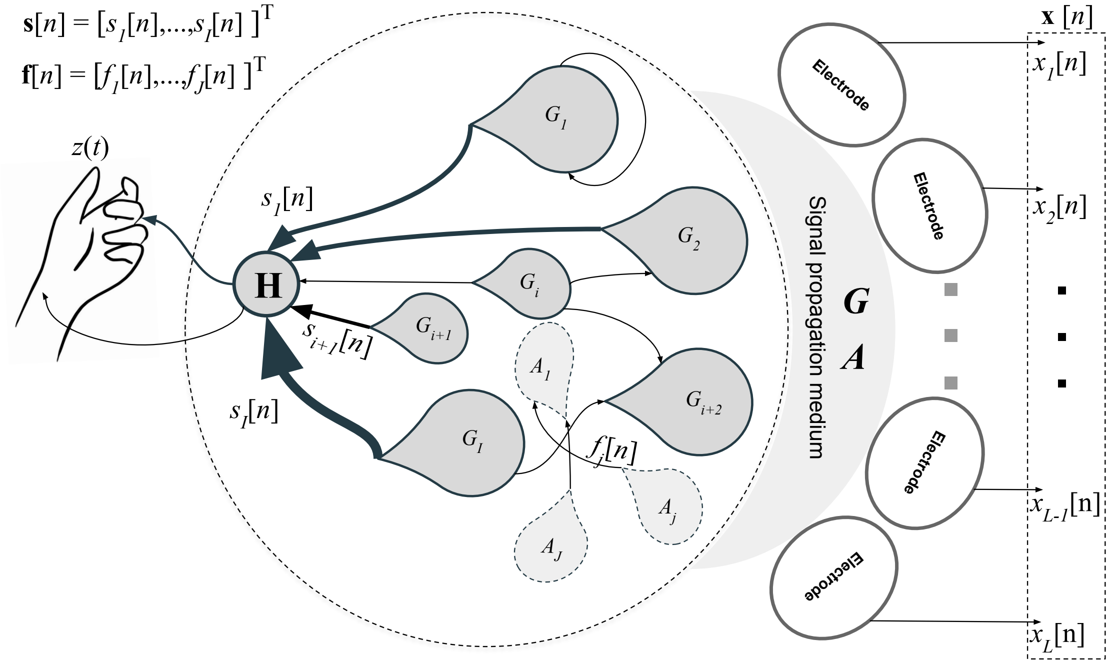

Figure 1: Phenomenological diagram.

Neuronal correlates of motor planning and execution have been extensively studied [59]. 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 [35]. The magnitude of these modulations correlates with the person's ability to control a motor-imagery BCI [36]. Additionally, the incidence rate of beta bursts in the primary somatosensory cortex is inversely correlated with the ability to detect tactile stimuli [28] 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 [19] 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 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.

2.1.2 Network architecture

The compact and adaptable architecture (ED-net) 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 the 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 temporal dimensions [27, 20]. 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 1-D 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 ReLu(x,a) $= \{x, x \geq 0; ax, x < 0\}$. To make the decision rule of this structure tractable, we used non-trainable batch normalization when streaming the data through the structure. This way we can harness the power of the optimization tools implemented within the deep learning approach to tune the parameters of our network that uses spatial filters followed by envelope estimation as the feature extraction block.

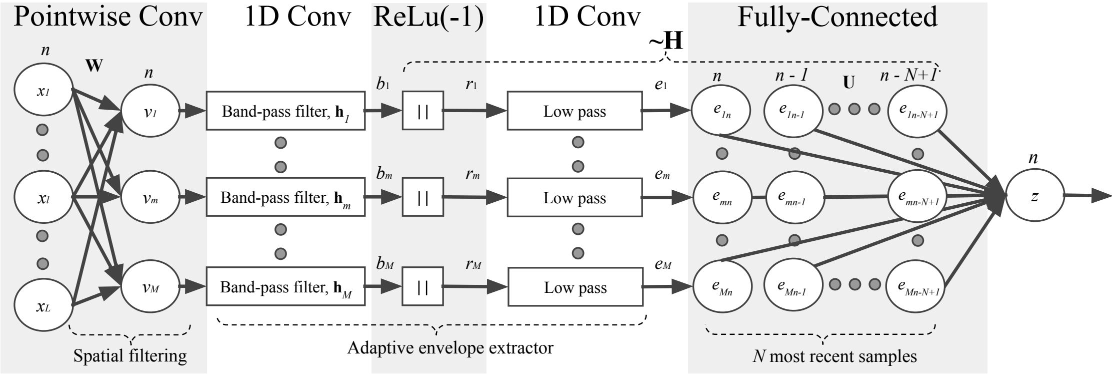

Figure 2: The architecture based on the compact CNN comprises several branches - adaptive envelope detector, receiving spatially unmixed input signals and outputting the envelopes whose N most recent values with indexes $n - N + 1, \ldots, n$ are combined in the decoded variable z by the fully connected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

In our architecture, the envelope detector of the m-th 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 the operator H as a linear combination of the lagged instantaneous power (envelope) of the narrow-band source time series $\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_I(t)]$ with coefficients from the matrix $\mathbf{U} = \{u_{ml}\},\ m = 1, \ldots, M, l = 1, \ldots, N$. This was performed 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.

2.1.3 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 1-D convolution layers. Consider a chunk of input data from L channels observed over the interval of N time moments that can be represented with a Toeplitz matrix $\mathbf{X}[n] = [\mathbf{x}[n],\mathbf{x}[n-1], \dots \mathbf{x}[n-N+1]] \in \mathbb{R}^{L\times N}$. Processing of $\mathbf{X}[n]$ by the first two layers, which perform spatial and temporal filtering, can be described for the m-th branch as follows:

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

where $\mathbf{w}_m \in \mathbb{R}^L$ is the spatial weights and $\mathbf{h}_m \in \mathbb{R}^N$ is the temporal weights of the branch m. The nonlinearity, $ReLu(-1)$, 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 [56] and for the original real-valued data, the imaginary part of the analytic signal is uniquely computed via the 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 to the sought kinematics $z[n]$ when transformed with a non-linear operator $H()$ approximated by the fully connected layer of our network. Suppose that the spatial filter weights are not known but the temporal convolution weights are fixed to their optimal values $\mathbf{h}_m^*$. Then, we can find the optimal spatial weights as the solution to a convex optimization problem formulated over the spatial subset of parameters:

$$\mathbf{w}_m^* = \mathrm{argmin}_{\mathbf{w}_m}\{\parallel b_m^*[n] - \mathbf{w}_m^T\mathbf{X}[n]\mathbf{h}_m^* \parallel_2^2\} = \mathrm{argmin}_{\mathbf{w}_m}\{\parallel b_m^*(n) - \mathbf{w}_m^T\mathbf{y}_m[n] \parallel_2^2\}, \quad (3)$$

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

$$\mathbf{h}_m^* = \mathrm{argmin}_{\mathbf{h}_m}\{\parallel b_m^*[n] - \mathbf{w}_m^{*T}\mathbf{X}[n]\mathbf{h}_m \parallel_2^2\} = \mathrm{argmin}_{\mathbf{h}_m}\{\parallel b_m^*[n] - \mathbf{v}_m^T[n]\mathbf{h}_m \parallel_2^2\}, \quad (4)$$

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 [61, 38]:

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

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 time series are zero-mean random processes, and L is the number of input channels.

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 a $N \times N$ tap covariance matrix of the spatially filtered data, assuming that channel time series 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 shown in [6], if we relax the assumption about the length of the data chunk being equal to the length of the temporal convolution filter we can arrive at the Fourier domain representation of dynamics of a neuronal population as pattern $Q_m(f)$ derived from the power spectral density (PSD) $P_{v_m}(f)$ of the spatially filtered data $v_m[n]$ and the Fourier transform $H_m(f)$ of the temporal weights vector $\mathbf{h_m}(f)$ as in 7:

$$Q_m(f) = P^{v_m}(f)H_m(f). \tag{7}$$

The important distinction that contrasts our weights interpretation approach from the methodology used in the majority of reports utilizing neural networks with separable spatial and temporal filtering operations is that our procedure accounts for the fact that the spatial filter formation is taking place within the context set by the corresponding temporal filter, and vice versa. Also, in [6], the authors for the first time introduced the notion of the frequency domain pattern $Q_m(f)$ of neuronal population's activity.

2.1.4 Realistic simulations

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 the 50-150 Hz frequency range and one task-unrelated source active within the 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. At the same time, the expression in (7) allows us to obtain a proper pattern that matches well the simulated spectral profile.

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 a randomly generated vector of coefficients. We used task-unrelated rhythmic sources with activation time series obtained similarly to the task-related sources but with filtering within the following four bands: 40-70 Hz, 90-110

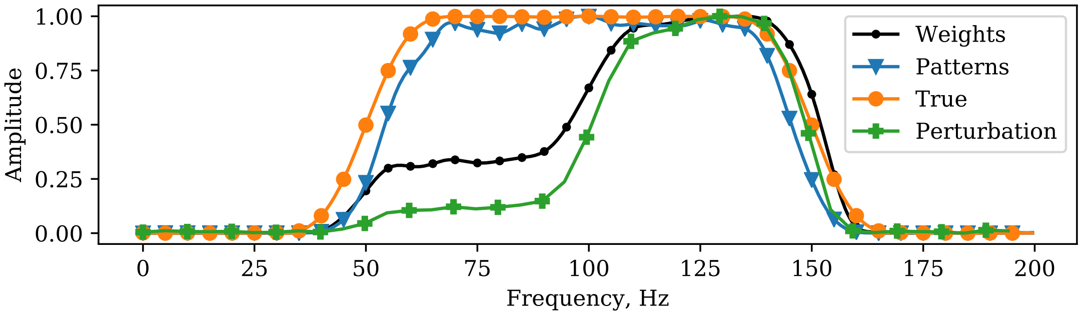

Figure 3: Three possible ways to interpret temporal convolution weights. The true pattern of dynamic activity i.e. the power spectral density (PSD) (orange ●) of the source. Fourier domain representation of the temporal convolution weights (black ●), Ball's method (green +), and the dynamic source activity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

Hz, 130-160 Hz, and 180-210 Hz bands. 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 minutes of data sampled at 1,000 Hz and divided them into two equal contiguous parts.

As a result, in the absence of noise, all interpretation methods coped well (4), but in the presence of noise, as can be seen in the graph 5, 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 (7), we obtain the spectral patterns that more closely match the simulated ones and have the deeps compensated.

Also, in order to obtain reliable results, we applied Monte Carlo simulations with different parameters, which clearly show that the proposed method gives a more correct interpretation (figure 6)

2.2 Decoding and Interpreting Cortical Signals With A Compact Convolutional Neural Network

This section contains a summary of two articles [6] [11].

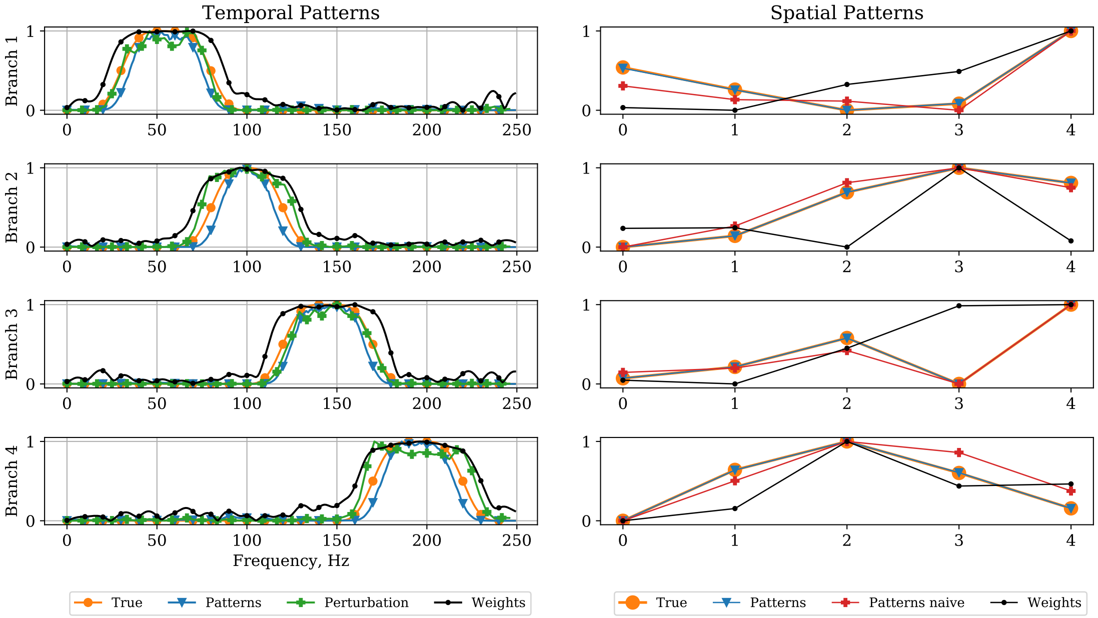

Figure 4: Temporal (left) and spatial (right) patterns were obtained for the noiseless case. See the main text for a description.

Contribution of the author: the architecture of the neural network was developed, a method for its interpretation was developed, computer simulations were implemented (including Monte Carlo simulations), the results of the quality of decoding and interpretation on real patients were obtained.

2.2.1 Introduction and existing methods

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 [33] 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 the DeepLIFT technique [29] combined with the analysis of the hidden unit activation patterns. Schirrmeister et al. describe two architectures: DeepConvNet and its compact version ShallowConvNet. The latter

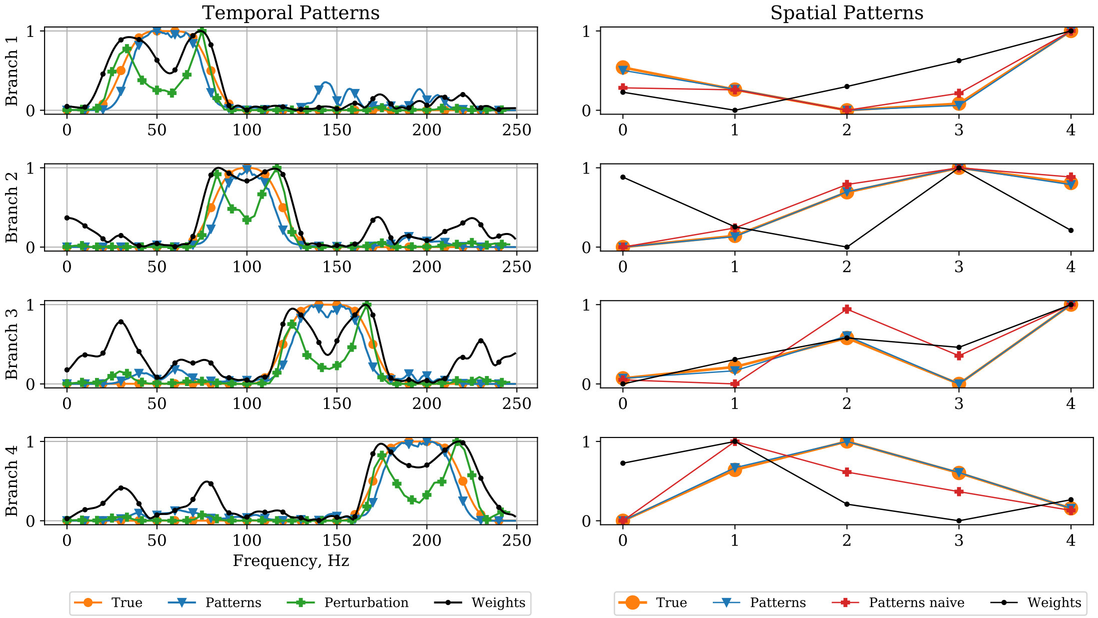

Figure 5: Temporal (left) and spatial (right) patterns were obtained for the noisy case, SNR = 1.5. See the main text for a description.

architecture consists of just two convolutional layers that perform temporal and spatial filtering, respectively [27]. In [23], authors describe a compact CNN architecture with separable spatial and temporal convolutions to perform the classification of EEG in the SSVEP paradigm. A recent study of Zubarev et al. [20] 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 well-established approaches [61, 57, 38]. 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

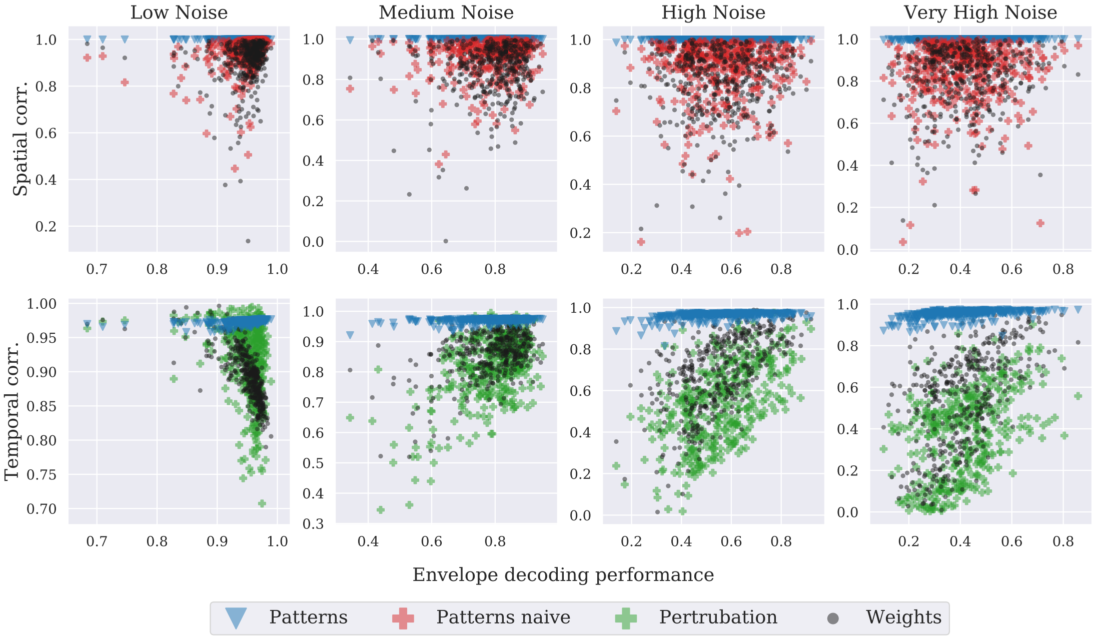

Figure 6: Monte-Carlo simulations. Point coordinates reflect the achieved envelope decoding performance (x-axis) and correlation coefficient with the true pattern (y-axis) at each Monte Carlo trial. Each point of a specific color corresponds to a single Monte Carlo trial and codes a method used to compute patterns. *Weights* - direct weights interpretation, *Patterns naive* - spatial patterns interpretation without taking branch-specific temporal filters into account, *Patterns* - the proposed method.

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 [55, 52, 45, 47, 42] and elegantly summarized in [38] take care of this adaptive behaviour but are 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 assess spatial patterns of the neuronal sources underlying decision rules 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.

To test the work of the developed neural network and the methods of its interpretation, we applied them to three datasets.

2.2.2 Motion decoding at the Berlin BCI competition IV

Firstly, to compare the compact GCN architecture with existing solutions, we used data collected by Kubanek et al and used in the Berlin BCI competition IV (which is publicly available). As a result, we did not observe significant differences between the performance of our algorithm and the winning solution of Lian and Bougrain[43] (Mann-Whitney test, $U = 103.0$, $p = 0.3543$), see table 1. Nevertheless, data on the location of the electrodes in this dataset are not disclosed, so it is not possible to make a complete interpretation.

Subject 1|2|3	Thumb	Index	Middle	Ring	Little
Winner	.58|.51|.69	.71|.37|.46	.14|.24|.58	.53|.47|.58	.29|.35|.63
NET	.54|.50|.71	.70|.36|.48	.20|.22|.50	.58|.40|.52	.25|.23|.61

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

2.2.3 Motion decoding from ECoG

Secondly, we applied the proposed solutions to the data collected in the HSE Center for Bioelectric Interfaces from two patients in the finger movement tasks, who were implanted with 8 × 8 ECoG microgrids placed on top of the sensorimotor cortex of the brain. Based on these data, we knew the locations of the electrodes and as a result, we received an interpretation that is consistent with knowledge from the subject area of neuroscience. An example can be seen in the figure 7

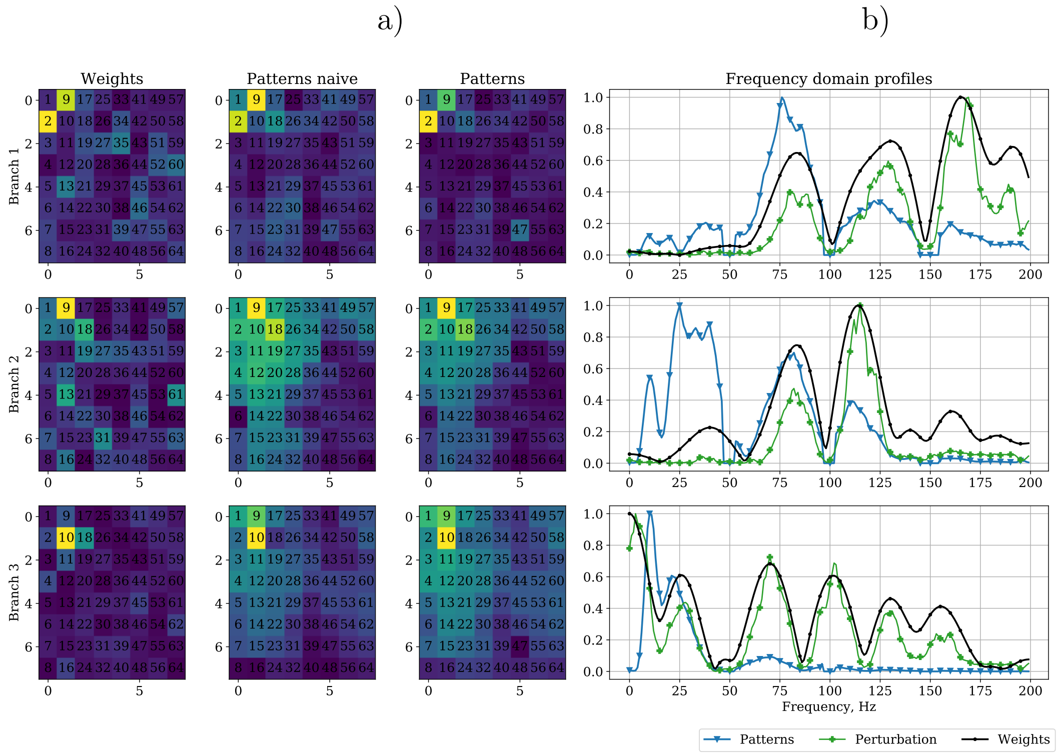

Figure 7: Network weights interpretation for the little finger kinematics decoder in CBI patient 2 (ECOG). Each row of plots corresponds to one of the three branches of the trained decoder. a) The leftmost column shows color-coded spatial filter weights, and the next two columns correspond to naively 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 (7). Another line (red $\Diamond$) 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 +).

2.2.4 Motion classification from EEG

Thirdly, unlike the previous two data sets, which required decoding a continuous trajectory from an invasive ECoG, the third data set was recorded non-invasively within the framework of the paradigm of imaginary EEG movements. The task here was to classify the type of motor actions performed. Given the short duration of these data, the compact CNN architecture solved the problem quite well and gave an average of 0.83 ROC AUC. And based on these data, we also received an interpretation that is consistent with knowledge from the subject area of neuroscience. An example can be seen in the figure 8

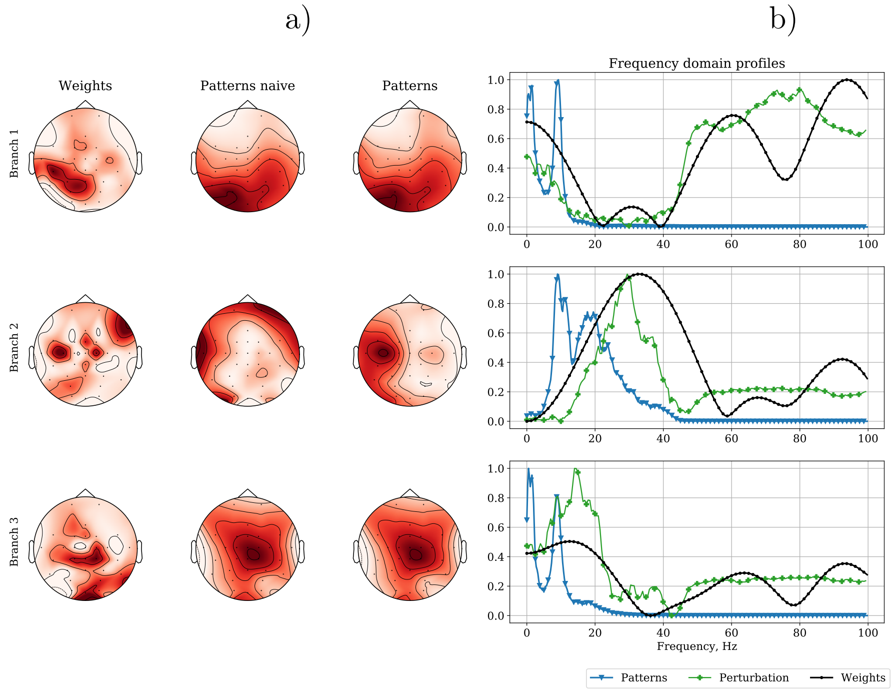

Figure 8: Network weights interpretation for the three branches (three rows of plots) of the decoder trained on a motor-imagery EEG dataset. a) The leftmost column shows color-coded spatial filter weights, and the next two columns correspond to naively and properly reconstructed spatial patterns. White 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 (7). 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 +).

2.3 Speech Decoding From A Small Set Of Spatially Segregated Minimally Invasive Intracranial EEG Electrodes With A Compact And Interpretable Neural Network

This section contains a summary of two articles [2] [4].

Contribution of the author: the architecture of a neural network for speech decoding was developed, the results of the quality of decoding and interpretation on real patients were obtained, the quality of decoding was compared when using different internal speech representations, the data was analyzed for the presence of a microphone effect, an asynchronous mode of operation of the neural network was implemented, an analysis of the mutual information between sound and brain data.

2.3.1 Introduction and existing methods

The ability to communicate is vital to humans and speech is the most natural channel for it. The inability to speak dramatically affects the quality of life. A number of disorders can lead to a loss of this vital function, for example, cerebral palsy and stroke of the brain stem. Also, in some cases, severe speech deficits may occur after a radical brain tissue removal surgery in oncology patients. While several technologies have been proposed to restore communication function they primarily rely on brain-controlled typing or imaginary handwriting [7] and appear to be practical only for severely affected patients. At the same time only in the United States, 50 million people suffer from not being able to use their speech production machinery properly. A significant fraction of them have pathology not amenable by alaryngeal voice prosthesis [25] or "silent speech" devices [54] and require a neurally driven speech restoration solution.

Several successful attempts of BCI-based speech restoration have already been made and significant progress is achieved in decoding phonemes [15, 22, 39], individual words [10, 3, 14], continuous sentences [10, 3, 14], and even acoustic features [18, 14, 17] followed by the speech reconstruction algorithms using either Griffin-Lim or deep neural network algorithms inspired by WaveNet[17].

These solutions employ a broad variety of machine learning approaches for decoding speech from brain activity data. Starting from linear models [15], LDA [8], and metric models [18] to deep neural networks (DNN) [10, 3, 14], which in general do not require manual feature engineering and can be applied directly to the data, however sometimes operating over a set of handcrafted features primarily derived from high-gamma activity. Several different neural network architectures have been tried for the speech decoding task: 1) relatively shallow ones consisting of a few convolutional or LSTM layers, 2) truly deep architectures with inception blocks [14] or with skip connections exploiting residual learning technique [17] as well as those borrowed from the computer vision applications

[24, 37], 3) ensembles of DNN [3] making the final solution more robust. Interestingly, the linear methods demonstrate compatible, or at least close to DNNs, decoding quality. Moreover, the latest studies obtained state-of-the-art decoding accuracy using just a few layers over a set of handcrafted physiologically plausible features [10, 3]

The majority of the existing neural speech decoding studies rely on heavily multichannel brain activity measurements implemented with massive ECoG grids [3, 10, 17, 16] covering the significant cortical area. These solutions for reading off brain activity are not intended for long-term use and are associated with significant risks to a patient [32] and suffer from a rapid loss of signal quality due to the leakage of the cerebrospinal fluid under the ECoG grid even if it is properly perforated. sEEG is a promising alternative whose implantation process is significantly less traumatic as compared to that of the large ECoG grids. The use of sEEG has already been explored for the speech decoding task [8] but the reported decoder again relied on a high count of channels from multiple sEEG shafts distributed over a large part of the left frontal and left superior temporal lobes which reduce the practicality of the proposed solution. A solution capable of decoding speech from the locally sampled brain activity would be an important step toward creating a speech prosthesis device.

The accuracy of neural speech decoding improves with the use of compressed representations encoding speech kinematic or acoustic features as an intermediate representation of the target variable [17] or for regularization [14]. However, it still remains unclear which of the compressed speech representations is optimal for decoding speech from electrophysiological data and how it should be used to yield the best decoding accuracy. In addition to the direct practical benefit, answering this question together with an appropriate interpretation of the decision rule will shed light on the neuronal basis and cortical representation of the speech production processes.

2.3.2 Neural network architecture and its interpretation

Here we explore the possibility of decoding individual words from intracranially recorded brain activity sampled with compact probes whose implantation did not require a full-blown craniotomy. Our study comprises two subjects implanted either with sEEG shafts or ECoG stripes both via compact drill holes. We decode individual words using either 6 channels of data recorded with a single sEEG shaft or the 8 channels sampled using a single ECoG strip. For decoding, we employed our compact and interpretable CNN architecture [6] augmented with the bidirectional LSTM layer [60] to compactly model local temporal dependencies in the internal speech representation that we used as the intermediate decoding target. We also compared the ultimate word decoding accuracy achieved with different internal representations. Our decoder operated causally using only the data from time intervals preceding the decoded time moment and therefore is fully applicable in a real-time decoding setting. Overall our study is the first attempt to achieve acceptable individual word decoding accuracy from cortical activity sampled with compact non-intracortical probes whose implantation is likely to cause minimal discomfort to a patient and can be done even with local anesthesia.

For neural signals to LMSC decoding, we employed the compact and interpretable convolutional network architecture developed earlier for motor BCI purposes [6] and augmented it with a single bidirectional LSTM layer with 30 hidden units to compactly model temporal regularities. The LSTM layer is followed by the fully connected layer with $M = 40$ output neurons each corresponding to a single mel-spectral coefficient whose temporal profile we are aiming to reconstruct from the neural activity data, see Figure 9. The ED during training can potentially adapt to extracting instantaneous power of specific neuronal populations activity pivotal for the downstream task of predicting the LMSCs. In the search for the optimum, the ED weights are not only tuned to such a target source but also tune away from the interfering sources [38, 6]. The proper interpretation of the learned ED's weights allows for the subsequent discovery of the target source's geometric and dynamical properties.

After having trained our compact architecture to decode the ISRs as our intermediate target we used a 2D-convolution network to perform the discrete classification of 26 words and the silent class using the representations developed in the one before the last layer of the compact architecture, see Figure 9.

Here we compared this network to several other architectures. We found that out of several neural networks only Resent-18 offers a comparable, although significantly worse, performance when used instead of the ED block in our architecture, see Figure 2. The LSTM layer also appears to be very useful in capturing the dynamics of features extracted

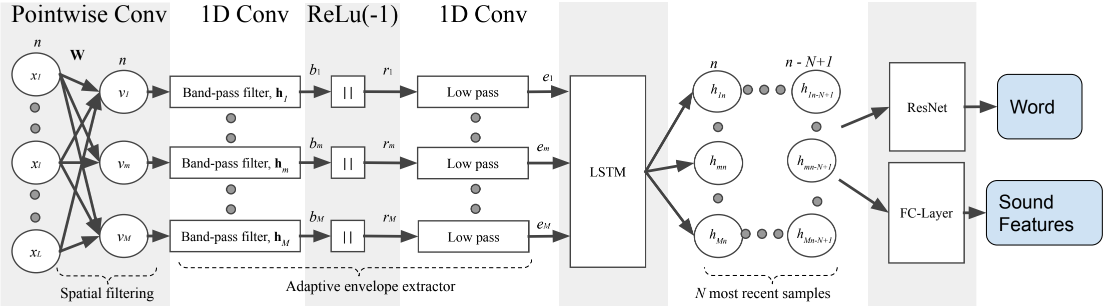

Figure 9: The architecture is based on [6] and adapted for the speech classification task. We used the same envelope detector technique to extract robust and meaningful features from the neuronal data. We then used the LSTM layer to account for the sequential structure of the mel-spectrogram and finally decoded it with a fully connected layer over the LSTM hidden state (h_{ij} on the figure). A separate 2D convolutional network was trained and used to classify separate words from the activity of this pre-trained LSTM.

either with ED or ResNet blocks, see Figure 12.a. We hypothesize that this situation may be caused by the adequate balance in the number of parameters to be tuned for the ED-based network and the amount of data available for training as compared to several other more sophisticated architectures.

2.3.3 Decoding internal speech representation

In this work, we also asked the question of the significance of the internal representation of the language (ISR) for the task of decoding speech. As can be seen from the neural network architecture, it uses an additional output and learns to restore some of the possible internal representations of speech (MELS, LPC, MFCC).

Most of the ISRs are based on modeling speech signals as produced by an excitation sequence passing through a linear time-varying filter [58]. The excitation sequence is the airflow in the larynx and the filter is formed by the articulatory tract elements (pharynx, vocal folds, tongue, lips, teeth) whose mutual geometry changes over time.

Linear predictive coding (LPC) and cepstral analysis are the two principal ways to estimate the parameters of such a filter. LPC analysis is based on a direct estimate of the auto-regressive prediction coefficients (PC) a_i through Burg's method [63]. However,

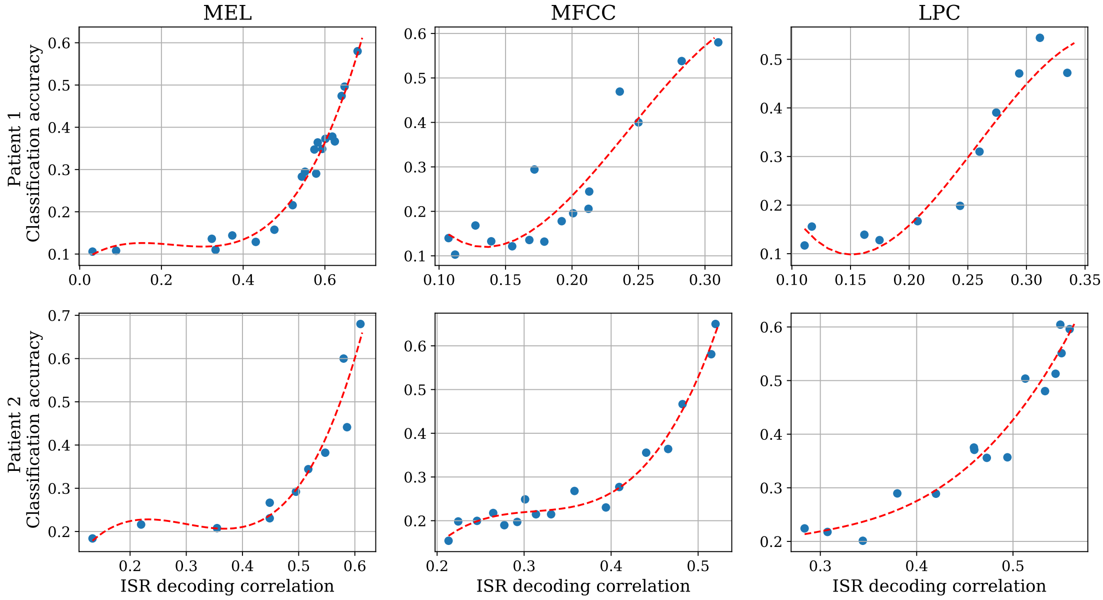

Figure 10: Dependence of the ISR decoding quality on the final word classification accuracy. The Red line is just a third-order trend estimation made for trend visualization.

prediction coefficients themselves are unstable, as their small changes may lead to large variations in the spectrum and possibly unstable filters. In order to decrease such instability the following several equivalent representations are commonly used.

Reflection coefficients (RC) k_i can be computed alongside prediction coefficients through Burg's method and represent the ratio of the amplitudes of the acoustic wave reflected by and the wave passed through a discontinuity.

Another descriptor, log-area ratio (LAR) coefficients, g_i, are equal to the natural logarithm of the ratio of the areas of adjacent sections in a lossless tube equivalent of the vocal tract having the same transfer function and can be computed from the reflection coefficients as $g_i = \ln\left(\frac{1-k_i}{1+k_i}\right)$.

Line spectral frequencies (LSF) is another highly efficient speech data compression technique [62] as errors in representing one coefficient generally result in a spectral change only around that frequency.

In what follows we will present the results of our experiments with several ISRs but our final decoding accuracy results are based on the use of log-mel spectral coefficients

(LMSC).

The results in the figure 12.b. We can see that the first patient log-mel spectrum coefficients (LMSC) target results in the highest word decoding accuracy. Interestingly, in contrast to the actual ISR decoding task displayed in Figure 11 the difference in the word decoding accuracy between various ISRs seems to be significantly less articulated than the differences in the quality of decoding of each of such representations. Nevertheless, for both patients, we observe a similar pattern with PC and LSF yielding relatively worse word decoding accuracy than the other ISRs. In this analysis, LPC reflection coefficients (RC) yield better decoding accuracy as compared to the prediction coefficients. This observation matches the properties of the RC coefficients as informationally equivalent but a more stable version of the original PC.

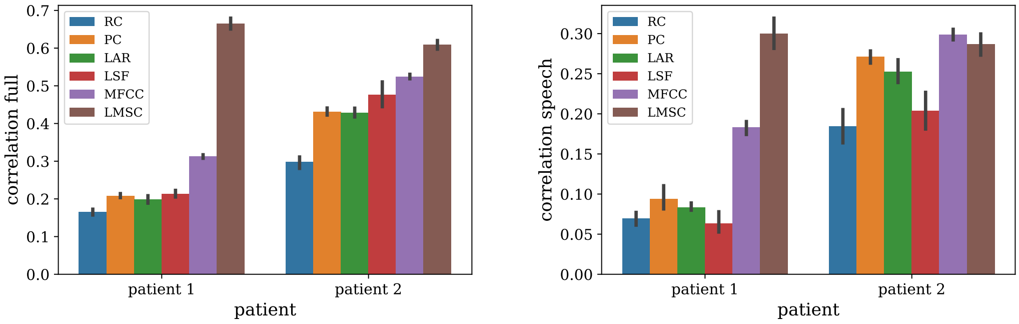

Figure 11: Comparison of the decoding accuracy achieved for different ISRs: PC - autoregressive prediction coefficients, LSF - line spectral frequencies, RC - reflection coefficients, LAR - log-area ratios, LMSCs - log-mel spectrograms, MFCC - mel-frequency cepstral coefficients. The left panel corresponds to the correlation coefficients between the actual and decoded temporal profiles computed over the entire time range of the test data segment. In the right panel, the correlation coefficient is computed only over the time intervals where the actual speech was present.

2.3.4 Synchronous and Asynchronous mode

Traditionally, BCI can be used in two different settings: synchronous and asynchronous. In the synchronous setting, a command is to be issued within a specific time window.

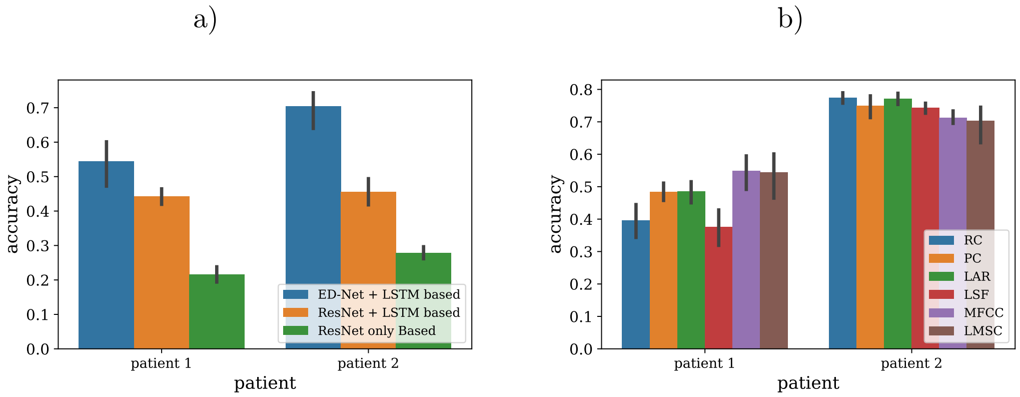

Figure 12: Comparative analysis. a) Comparison of different neural network models, b) Comparison of different possible intermediate sound representation, PC - autoregressive prediction coefficients, LSF - Line Spectral Frequencies, RC - reflection coefficients, LAR - log-area ratios, LMSC - log-mel spectrogram coefficients, MFCC - mel-frequency cepstral coefficients.

Usually, a synchronous BCI user is prompted at the start of such a time window and has to produce a command (alter his or her brain state) within a specified time frame. Therefore, the decoding algorithm is aware of the specific segment of data to process in order to extract the information about the command. In the asynchronous mode, the BCI needs to not only decipher the command but also determine the fact that the command is actually being issued. The delineation between synchronous and asynchronous modes is most clearly pronounced in BCIs with discrete commands implying the use of a categorical decoder.

In BCIs that decode a continuous variable, e.g. hand kinematics,the such delineation between synchronous and asynchronous modes is less clear. The first part of our BCI implements a continuous decoder of the internal speech representation (ISR) features. Should this decoding appear of sufficient accuracy it could have been simply used as an input to a voice synthesis engine. Such a scenario has already been implemented in several reports [17, 16] but these solutions use a large number of electrodes which may explain the better quality of ISR decoding. In our setting, we aimed at building a decoder operating with a small number of ecologically implanted electrodes and decided to focus on decoding individual words. We first used the continuously decoded ISRs to classify 26 discrete words and one silence state in a synchronous manner. To implement this we cut the decoded ISR time series around each word's utterance and use them as data samples for our classification engine.

Figure 13 b) illustrates the performance of our BCI operating in a fully asynchronous mode when the decoder is running over the succession of overlapping time windows of continuously decoded ISRs and the decision about the specific word being uttered is made for each of such windows. To quantify the performance of our asynchronous speech decoder we used precision-recall (see Figure 13 a).)

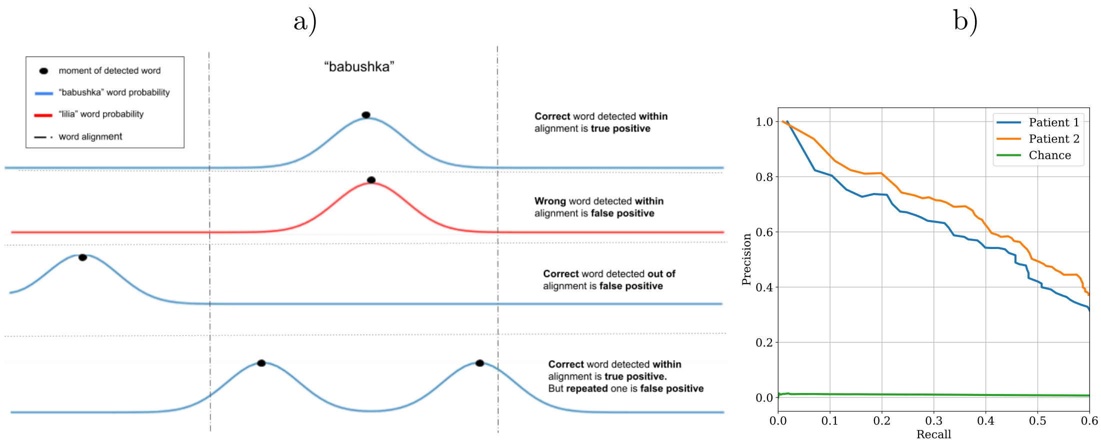

Figure 13: a) For each $i-$th word, we compute smoothed probability profiles $\tilde{p}_i(t)$ for each time instance t. The decision is then made about a word being pronounced only at time points corresponding to the local maximums of $\tilde{p}_i(t)$ that cross the threshold θ. In case the chosen $i-$th word matches the one that is currently being uttered, we mark this event as true positive (TP). If after such a detection $\tilde{p}_i(t)$ remains above the threshold and exhibits another local maximum that exceeds the values of all other smoothed probability profiles, the $i-$th word is "uttered" again, but this event is marked as false positive (FP) even if t belongs to the time range corresponding to the actual $i-$th word. b) PR curves for asynchronous words decoding task. Note that the definition of precision and recall is slightly different from conventional binary classification PR curves (see equation 7). We also show a chance-level PR curve.

Although the observed performance significantly exceeds the chance level, it is not yet sufficient for building a full-blown asynchronous speech interface operating using a small number of minimally invasive electrodes. In our view and based on our experience with motor interfaces, specific protocols to train the patient including those with immediate feedback to the user [1] are likely to significantly improve the decoding accuracy in such systems which will boost the feasibility of minimally invasive speech prosthetic solutions.

3 Conclusion

In this work, two large projects united by a common theme, dedicated to the development and application of modern interpretable neural network models for the analysis and decoding of brain activity, were completed. The work is a completed study, as a result of which a whole range of software and algorithmic means for processing electrophysiological signals was developed, containing at its core a new mathematical method, also proposed by the authors of the work. The solutions obtained were tested in presentations at numerous conferences and their scientific validity was confirmed by a number of publications in leading international scientific journals, including two publications with the main authorship in the Journal of Neural Engineering (Q1 - Scopus, Q2 - WoS). Currently, all the developed tools and algorithms are used in the research activities of the Center for Bioelectrical Interfaces of the National Research University Higher School of Economics.

3.1 List of results submitted for the defense

1. The compact neural network architecture reflects modern scientific ideas about the origin of neuroelectrophysiological activity, the mechanism of its propagation in tissues, and the physical principles of its registration using a distributed set of electrodes.

2. The results of a comparative analysis of the quality of decoding from ECoG and s-EEG of finger kinematics and articulatory tract parameters, demonstrate the superiority of the proposed neural network architecture compared to competing solutions.

3. A theoretically justified method for interpreting weight coefficients in the proposed architecture of a neural network in order to identify the geometric characteristics of

key populations of neurons and the dynamic properties of their activity.

4. The results of the analysis of the dependence of the final classification accuracy on the choice of the intermediate representation of the speech signal.
5. Implementation of real-time hand movement kinematics decoding.
6. Implementation of speech decoding based on the minimum number of spatially segregated electrodes.

References

[1] Miguel Angrick et al. “Towards Closed-Loop Speech Synthesis from Stereotactic EEG: A Unit Selection Approach”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 1296–1300.

[2] Artur Petrosyan et al. “Speech decoding from a small set of spatially segregated minimally invasive intracranial EEG electrodes with a compact and interpretable neural network”. In: *bioRxiv* (2022).

[3] David A Moses et al. “Neuroprosthesis for decoding speech in a paralyzed person with anarthria”. In: *New England Journal of Medicine* 385.3 (2021), pp. 217–227.

[4] Artur Petrosyan, Alexey Voskoboynikov, and Alexei Ossadtchi. “Compact and interpretable architecture for speech decoding from stereotactic EEG”. In: *2021 Third International Conference Neurotechnologies and Neurointerfaces (CNN)*. IEEE. 2021, pp. 79–82.

[5] Artur Petrosyan et al. “Compact and Interpretable Architecture for Speech Decoding From iEEG”. In: *International Journal of Psychophysiology* 168.S (2021), S195.

[6] Artur Petrosyan et al. “Decoding and interpreting cortical signals with a compact convolutional neural network”. In: *Journal of Neural Engineering* 18.2 (2021), p. 026019.

[7] Francis R Willett et al. “High-performance brain-to-text communication via handwriting”. In: *Nature* 593.7858 (2021), pp. 249–254.

[8] Miguel Angrick et al. “Real-time Synthesis of Imagined Speech Processes from Minimally Invasive Recordings of Neural Activity”. In: *bioRxiv* (2020).

[9] Christian Herff, Dean J Krusienski, and Pieter Kubben. “The potential of stereotactic-EEG for brain-computer interfaces: current progress and future directions”. In: *Frontiers in neuroscience* 14 (2020), p. 123.

[10] Joseph G Makin, David A Moses, and Edward F Chang. “Machine translation of cortical activity to text with an encoder–decoder framework”. In: *Nature Neuroscience* 23.4 (2020), pp. 575–582.

[11] Artur Petrosyan, Mikhail Lebedev, and Alexey Ossadtchi. “Decoding neural signals with a compact and interpretable convolutional neural network”. In: *International Conference on Neuroinformatics*. Springer. 2020, pp. 420–428.

[12] Artur Petrosyan, Mikhail Lebedev, and Alexey Ossadtchi. “Linear Systems Theoretic Approach to Interpretation of Spatial and Temporal Weights in Compact CNNs: Monte-Carlo Study”. In: *Biologically Inspired Cognitive Architectures Meeting*. Springer. 2020, pp. 365–370.

[13] David Sabbagh et al. “Predictive regression modeling with MEG/EEG: from source power to signals and cognitive states”. In: *NeuroImage* 222 (2020), p. 116893.

[14] Pengfei Sun, Gopala K Anumanchipalli, and Edward F Chang. “Brain2Char: a deep architecture for decoding text from brain recordings”. In: *Journal of Neural Engineering* 17.6 (2020), p. 066015.

[15] Guy H Wilson et al. “Decoding spoken English from intracortical electrode arrays in dorsal precentral gyrus”. In: *Journal of Neural Engineering* 17.6 (2020), p. 066007.

[16] Hassan Akbari et al. “Towards reconstructing intelligible speech from the human auditory cortex”. In: *Scientific reports* 9.1 (2019), pp. 1–12.

[17] Miguel Angrick et al. “Speech synthesis from ECoG using densely connected 3D convolutional neural networks”. In: *Journal of neural engineering* 16.3 (2019), p. 036019.

[18] Christian Herff et al. “Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices”. In: *Frontiers in neuroscience* 13 (2019), p. 1267.

[19] Ksenia Volkova et al. “Decoding Movement From Electrocorticographic Activity: A Review”. In: *Frontiers in neuroinformatics* 13 (2019), p. 74. ISSN: 1662-5196. DOI: 10.3389/fninf.2019.00074. URL: https://europepmc.org/articles/PMC6901702.

[20] Ivan Zubarev et al. “Adaptive neural network classifier for decoding MEG signals”. In: *NeuroImage* 197 (2019), pp. 425–434.

[21] Abidemi B Ajiboye and Robert F Kirsch. “Invasive Brain–Computer Interfaces for Functional Restoration”. In: *Neuromodulation*. Elsevier, 2018, pp. 379–391.

[22] Nick F Ramsey et al. “Decoding spoken phonemes from sensorimotor cortex with high-density ECoG grids”. In: *Neuroimage* 180 (2018), pp. 301–311.

[23] Nicholas Waytowich et al. “Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials”. In: *Journal of neural engineering* 15.6 (2018), p. 066031.

[24] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[25] Rachel Kaye, Christopher G Tang, and Catherine F Sinclair. “The electrolarynx: voice restoration after total laryngectomy”. In: *Medical Devices (Auckland, NZ)* 10 (2017), p. 133.

[26] Mikhail A Lebedev and Miguel AL Nicolelis. “Brain-machine interfaces: From basic science to neuroprostheses and neurorehabilitation”. In: *Physiological reviews* 97.2 (2017), pp. 767–837.

[27] Robin Tibor Schirrmeister et al. “Deep learning with convolutional neural networks for EEG decoding and visualization”. In: *Human brain mapping* 38.11 (2017), pp. 5391–5420.

[28] Hyeyoung Shin et al. “The rate of transient beta frequency events predicts behavior across tasks and species”. In: *Elife* 6 (2017), e29086.

[29] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. *Learning Important Features Through Propagating Activation Differences*. 2017. arXiv: `1704.02685 [cs.CV]`.

[30] Осадчий et al. “Интерфейс мозг-компьютер: опыт построения, использования и возможные пути повышения рабочих характеристик”. In: *Журнал высшей нервной деятельности им. ИП Павлова* 67.4 (2017), pp. 504–520.

[31] Ujwal Chaudhary, Niels Birbaumer, and Ander Ramos-Murguialday. “Brain–computer interfaces for communication and rehabilitation”. In: *Nature Reviews Neurology* 12.9 (2016), p. 513.

[32] Prasanna Jayakar et al. "Diagnostic utility of invasive EEG for epilepsy surgery: Indications, modalities, and techniques". In: *Epilepsia* 57.11 (2016), pp. 1735–1747. DOI: https://doi.org/10.1111/epi.13515. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/epi.13515. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/epi.13515.

[33] Vernon J Lawhern et al. "Eegnet: A compact convolutional network for eeg-based brain-computer interfaces". In: *arXiv preprint arXiv:1611.08024* (2016).

[34] Sarah N Abdulkader, Ayman Atia, and Mostafa-Sami M Mostafa. "Brain computer interfacing: Applications and challenges". In: *Egyptian Informatics Journal* 16.2 (2015), pp. 213–230.

[35] Yaron Meirovitch et al. "Alpha and Beta Band Event-Related Desynchronization Reflects Kinematic Regularities". In: *Journal of Neuroscience* 35.4 (2015), pp. 1627–1637. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.5371-13.2015. eprint: https://www.jneurosci.org/content/35/4/1627.full.pdf. URL: https://www.jneurosci.org/content/35/4/1627.

[36] Johanna Louise Reichert et al. "Resting-state sensorimotor rhythm (SMR) power predicts the ability to up-regulate SMR in an EEG-instrumental conditioning paradigm". In: *Clinical Neurophysiology* 126.11 (Feb. 2015), pp. 2068–2077. DOI: 10.1016/j.clinph.2014.09.032.

[37] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[38] Stefan Haufe et al. "On the interpretation of weight vectors of linear models in multivariate neuroimaging". In: *Neuroimage* 87 (2014), pp. 96–110.

[39] Emily M Mugler et al. "Direct classification of all American English phonemes using signals from functional speech motor cortex". In: *Journal of neural engineering* 11.3 (2014), p. 035015.

[40] Mark L Homer et al. "Sensors and decoding for intracortical brain computer interfaces". In: *Annual review of biomedical engineering* 15 (2013), pp. 383–405.

[41] Miguel Pais-Vieira et al. “A Brain-to-Brain Interface for Real-Time Sharing of Sensorimotor Information”. In: *Scientific reports* 3 (Feb. 2013), p. 1319. DOI: 10.1038/srep01319.

[42] Felix Bießmann et al. “Improved decoding of neural activity from fMRI signals using non-separable spatiotemporal deconvolutions”. In: *NeuroImage* 61.4 (2012), pp. 1031–1042.

[43] Nanying Liang and Laurent Bougrain. “Decoding Finger Flexion from Band-Specific ECoG Signals in Humans”. In: *Frontiers in neuroscience* 6 (June 2012), p. 91. DOI: 10.3389/fnins.2012.00091.

[44] Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. “Brain computer interfaces, a review”. In: *Sensors* 12.2 (2012), pp. 1211–1279.

[45] Benjamin Blankertz et al. “Single-trial analysis and classification of ERP components—a tutorial”. In: *NeuroImage* 56.2 (2011), pp. 814–825.

[46] Steven Lemm et al. “Introduction to machine learning for brain imaging”. In: *Neuroimage* 56.2 (2011), pp. 387–399.

[47] Thomas Naselaris et al. “Encoding and decoding in fMRI”. In: *Neuroimage* 56.2 (2011), pp. 400–410.

[48] Gerwin Schalk and Eric C Leuthardt. “Brain-computer interfaces using electrocorticographic signals”. In: *IEEE reviews in biomedical engineering* 4 (2011), pp. 140–154.

[49] Sergio Machado et al. “EEG-based brain-computer interfaces: an overview of basic concepts and clinical applications in neurorehabilitation”. In: *Reviews in the Neurosciences* 21.6 (2010), pp. 451–468.

[50] Claudio Castellini and Patrick van der Smagt. “Surface EMG in advanced hand prosthetics”. In: *Biological cybernetics* 100.1 (2009), pp. 35–47.

[51] Nicholas G Hatsopoulos and John P Donoghue. “The science of neural interface systems”. In: *Annual review of neuroscience* 32 (2009), pp. 249–266.

[52] Aapo Hyvärinen, Jarmo Hurri, and Patrick O Hoyer. *Natural image statistics: A probabilistic approach to early computational vision.* Vol. 39. Springer Science & Business Media, 2009.

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

[54] Michael J Fagan et al. “Development of a (silent) speech recognition system for patients following laryngectomy”. In: *Medical engineering & physics* 30.4 (2008), pp. 419–425.

[55] Lucas C Parra et al. “Recipes for the linear analysis of EEG”. In: *Neuroimage* 28.2 (2005), pp. 326–341.

[56] Stefan L Hahn. “On the uniqueness of the definition of the amplitude and phase of the analytic signal”. In: *Signal Processing* 83.8 (2003), pp. 1815–1820.

[57] Lucas Parra et al. “Single-trial detection in EEG and MEG: Keeping it linear”. In: *Neurocomputing* 52 (2003), pp. 177–183.

[58] Xuedong Huang et al. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* 1st. USA: Prentice Hall PTR, 2001. ISBN: 0130226165.

[59] Daniel Wolpert and Zoubin Ghahramani. “Computational Principles of Movement Neuroscience”. In: *Nature neuroscience* 3 Suppl (Dec. 2000), pp. 1212–7. DOI: 10.1038/81497.

[60] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[61] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory.* Prentice Hall, 1997.

[62] Frank Soong and B Juang. “Line spectrum pair (LSP) and speech data compression”. In: *ICASSP’84. IEEE International Conference on Acoustics, Speech, and Signal Processing.* Vol. 9. IEEE. 1984, pp. 37–40.

[63] L. Marple. “A new autoregressive spectrum analysis algorithm”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28 (1980), pp. 441–454.