Optimization of the Brain Command Dictionary in Silent Speech Recognition Task Based on Statistical Proximity Criterion.

Alexandra Bernadotte
Faculty of Mechanics and Mathematics
Moscow State University. 119991 Moscow, Russia
Dep. of Information Technologies and Computer Sciences
National University of Science and Technology MISIS
119049 Moscow, Russia
Neurosputnik LLC, Russia
Email: bernadotte.alexandra@intsys.msu.ru

Alexandr Mazurin
Faculty of Mechanics and Mathematics
Moscow State University. 119991 Moscow, Russia
Email: ADmMazurin@sberbank.ru

***Abstract*—In our work, we focus on silent speech recognition in electroencephalography (EEG) in order to develop a new brain-computer interface (BCI) with a manipulator which will be capable of assisting people with physical disabilities.**

We have previously shown that the silent speech of some words leads to almost identical distributions of electroencephalographic signal data. In this paper, we continue to develop the previously proposed research topic of algorithmic maximization of the semantic load of a dictionary for BCI. We propose a predictive criterion for the accuracy of classifying words in a dictionary whose aim is to estimate the classifiers' behavior only by measuring statistical criterion (The Kolmogorov-Smirnov method) of the data.

The application of this predictive criterion helps to achieve higher levels of accuracy for neural network classifiers.

The use of the proposed approach makes it possible to form a dictionary for the EEG-based BCIs, taking into account the semantic and phonetic proximity of words with the possibility of increasing the silent speech recognition accuracy.

I. Introduction

In our work, we are developing an EEG-based BCI focused on recognizing mental commands of movement. This interface is especially important for people with damage or underdevelopment of the motor cortex. Bearing this problem in our mind, we decided to develop a BCI that recognizes silent speech; that is, patterns of brain activity in the temporal lobe of the cortex, or rather, in Broca's area (the anterior speech cortex). In our study, silent speech was a set of commands (dictionary) given by the inner voice. To control the manipulator, these commands should not be numerous, but the dictionary should be well recognized by the BCI.

However, we found that some commands pronounced by the inner voice (silent speech) are poorly distinguishable from each other and demonstrate low accuracy of classification by a neural network of BCI. We hypothesized that this silent speech classification problem is due to semantic and phonetic proximity of the words. We have previously shown that the silent speech of some words leads to almost identical distributions of electroencephalographic signal data [1], [2], [3], [4]. Earlier, we showed that this silent speech classification problem can be overcome by selecting a dictionary of well-recognized and accurately classified words. Previously, our group presented a graph dictionary selection algorithm and the applicability of this algorithm to our data [3].

The selection of a dictionary itself is a rather laborious task, and we had an idea to develop a certain criterion that would allow us to predict the behavior of a neural network and at the same time would be quite convenient in terms of time and resource. Thus, we were looking for our criterion in the field of statistical methods that were less expensive in terms of calculating resources.

The new study presented in this paper is based on the analysis of classification behavior of neural network models trained on the set of EEG-data recordings during several sessions corresponding to silent speech pronunciation of words "up", "down", "vira" and "myna". These words are semantically divided into two classes: {"up", "vira"} and {"down", "myna"}. To avoid time- and GPU resource consuming dictionary selection it was interesting for us to consider the existence of a predictive criterion by which we could predict the results of the binary classification of a neural network.

Thinking about the predictive criterion for neural network classification accuracy, we formulated the following hypothesis:

Hypothesis. *The Kolmogorov-Smirnov method on EEG-data of silent speech with reduced dimension can statistically predict the behavior (accuracy) of a neural network classifier on the same EEG-data of a higher dimension.*

Confirming this hypothesis and gaining new knowledge about the patterns of brain activity aims at improving our classification results, not only by trying to tune various neural network architectures but also by intelligently tuning the dictionary with particular attention to statistical proximity of the classes in the dictionary [3].

Moreover, proving Hypothesis directly leads to the fact that there exists a predictive criterion for the classification accuracy of the models based on the overall statistical separability of the classes in the dataset. In particular, we show that the sum of all $p-$values computed pairwise for every two classes in the train dataset can be used to make an estimation of the accuracy rate levels of a given network.

We were looking for our criterion in the field of statistical methods that were less expensive in terms of calculating resources.

II. Methods

A. Dataset

All subjects had reached adulthood, were healthy, and signed voluntary a consent to the study. The subjects could interrupt the study at any time without explanation. The subjects provided their data, which included: gender, age, education, and occupation. The exclusionary criteria for the study were a history of head trauma, alcohol or other intoxication, and epilepsy.

The dataset we used for our study consisted of $32-$channel EEG signal recordings completed at $250Hz$ during several sessions of silent and vocalized speech of 105 subjects. The dry plastic electrodes (Datwyler's SoftPulseTM Medium, brush type electrode) were placed according to the traditional $10-20$ scheme. The 'Afz'-channel was used as a reference electrode. A word presentation signal was also captured via a light sensor and included in data files as a mark.

The data were divided into three sets: train (70%), validation (20%), and test(10%).

B. Data processing

Data preprocessing methods included: 1) Eye noise filtering procedure consisting of morphological selection of eye sensitive channels, Independent Component Analysis (ICA), and further detection and removal of the eye noise component using of Fast Fourier Transform (FFT), Savitskiy-Galey filtering and Inverse FFT; 2) Filtering out the tensors with high noise level by means of applying two filters by the sum of the moduli of the signal amplitudes; 3) Downsampling the sample rate by using index masks on the original EEG data; 4) Separating the electrodes into left and right hemispheres; 5) Reorganizing the elements of the data tensors and combining the described above operations in order to form a two-dimensional matrix from every tensor in the dataset [1].

By applying the described above procedures, we were able to transform any raw signal tensor of size 32×1024 (32 stands for the number of electrodes (channels) used in EEG recording procedure; 1024 is the number of time stamps corresponding to signal discretization procedure) to a square 2D matrix of size 256×256.

C. Predictive criterion

Before using the criterion, it is strongly recommended to use principal component analysis (PCA). After PCA, we looked at distributions of mental words as distributions in $kD-$space. We used $3D-$space. The Kolmogorov-Smirnov test, independent of the nature of the distribution, was applied to the set of distributions in $3D-$space by computing the set of pairwise $p-$values on train and validation sets. Each component (each dimension) gave its own $p-$value, the minimum $p-$value was taken from this set. The existing pairs of synonyms were organized in ascending order of $p-$values.

The predictive criterion was based on statistics of dimension reduced data: on the set of all possible $p-$values of the Kolmogorov-Smirnov test on PCA dimension reduced distributions from validation and train sets, the $p-$value and the network binary classification accuracy are inversely correlated. This absolute accuracy (average or median) depends on the network architecture and has a variance depending on the network architecture.

D. Neural Network Classifiers

Neural network classification models which were tested for multiclass classification of various sets of word commands included Image Transformer and ResNext 18 deep architectures adapted for the preprocessed dataset consisting of square 256×256 tensors. We further briefly describe the mentioned deep learning architectures.

E. ResNet-18 Network Classifier

The main feature that makes ResNext model differ from conventional residual nets is the introduction of the cardinality hyperparameter [1] [5], which separates the channels of the input tensors into several groups, with each one being operated by its own convolutional kernel. This architecture adopts both the strategy of repeating layers, which is a common property of VGGs/ResNets, and the split-transform-merge strategy first applied in Inception model architectures. ResNext models are mainly constructed from building blocks each performing a set of convolutional transformations on a low-dimensional embedding and aggregating their outputs by summation [5]. Model configurations differ by complexity and the number of building blocks in the architecture. We chose ResNet-18 model which consists of 4 modules, each containing 2 convolutional blocks (see Fig. 1.a for the detailed architecture scheme).

F. Vision Transformer Network Classifier

We decided to use the Vision Transformer neural network classifier based on colleagues paper [6] as well as convolution-based architectures. This version of transformer architecture is specifically designed for the task of image classification. Vision Transformer contains a solid number of self-attention blocks, the main purpose of which is computation of pointwise scalar products between the values of all square (16×16) fragments of the preprocessing 256×256 matrix and vectors consisting of trainable weights. Such a solution may allow the network to find hidden patterns in the information provided by the whole input matrix, not only in its neighboring fragments, which grants it a benefit when compared to convolutional neural networks (CNNs). Moreover, the self-attention blocks of Vision Transformer net are connected successively, forming

a chain which makes the process of finding deeper patterns easier.

The architecture of Vision Transformer (see Fig. 1.b) is aimed at dealing with square images by means of initial cropping into a number of square patches of smaller size. These patches serve as inputs to the trainable part of the network, which mainly consists of self-attention blocks responsible for finding connections between them and deciding which of them are most important for completing the process of classification. We use $N = 6$ transformer encoder blocks, $h = 8$ as the number of neurons in the hidden layer of the MLP block and $p = 6$ as the number of heads in the multi-head self-attention layer. The experiments showed that increasing the complexity of the model did not result in higher results for validation and test classification.

III. Results

A. Predictive criterion for Vision Transformer and ResNet-18

The results of the predictive criterion and classification accuracy are listed in the Table I.

words pair	$p-$value	$p-$value	median acc. Vision Transformer	median acc. ResNet-18
	train	val	test	test
'up'/'down'	$4 \times e^{-1}$	$2 \times e^{-1}$	0.49	0.5
'vira'/'myna'	$5 \times e^{-1}$	$1 \times e^{-1}$	0.49	0.5
'up'/'myna'	$2 \times e^{-114}$	$4 \times e^{-10}$	0.91	0.7
'up'/'vira'	$4 \times e^{-117}$	$1 \times e^{-16}$	0.84	0.79
'down' /'vira'	$3 \times e^{-116}$	$1 \times e^{-16}$	0.87	0.6
'down'/'myna'	$4 \times e^{-114}$	$1 \times e^{-41}$	0.92	0.78

TABLE I: Predictive criterion and accuracy for Vision Transformer and ResNet-18.

We see an inverse correlation between the $p-$value and the classification accuracy of the neural network.

B. Predictive criterion for Vision Transformer

Consider a pair of words i, j. Let's denote p_tr as $p-$value on the train sets, p_val as $p-$value on the validation sets, p_test as $p-$value on the test sets. Then the following estimates hold for the median accuracy a of the binary classification of words i, j:

1) if $p\_tr \geq 0.38$, then accuracy $\sim$0.49;

2) if $p\_tr < 0.38$ and $4 * e^{-40} < p\_val \leq 1.11 * e^{-16}$, then accuracy ~ 0.85;

3) if $p\_tr < 0.38$ and $p\_val \leq 4 * e^{-40}$, then accuracy $\sim$0.915.

C. Predictive criterion for ResNet-18

Consider a pair of words i, j. Let's denote p_tr as $p-$value on the train sets, p_val as $p-$value on the validation sets, p_test as $p-$value on the test sets. Then the following estimates hold for the median accuracy a of the binary classification of words i, j:

1) if $p\_tr \geq 0.38$, then accuracy $\sim$0.5;

2) if $p\_val \leq 4 * e^{-40}$, then accuracy ≥ 0.69;

3) if $p\_val > 4 * e^{-40}$ and $p\_tr \leq 4.41 * e^{-117}$, then accuracy $\sim$0.79;

4) if $p\_val > 4 * e^{-40}$ and $4.41 * e^{-117} < p\_tr < 0.38$, then accuracy $\sim$0.59.

IV. Conclusion

In this work, we have shown the fundamental possibility of working with low-dimensional data and reducing resource consumption for selecting the optimal dictionary for BCI based on EEG-data and silent speech recognition. We formulated and confirmed the hypothesis about the existence of a predictive criterion based on distribution statistics of data with reduced dimension and predicting the accuracy of classification by a neural network on the same data of a higher dimension.

REFERENCES

[1] Darya Vorontsova, Ivan Menshikov, Aleksandr Zubov, Kirill Orlov, Peter Rikunov, Ekaterina Zvereva, Lev Flitman, Anton Lanikin, Anna Sokolova, Sergey Markov and Alexandra Bernadotte. Silent EEG-Speech Recognition Using Convolutional and Recurrent Neural Network With 85% Accuracy of 9 Words Classification *Sensors* **2021**, *21.20*, 6744.

[2] Mazurin A., Bernadotte A.. Clustering quality criterion based on the features extraction of a tagged sample with an application in the field of brain-computer interface development // Intelligent systems. Theory and Applications (formerly: Intelligent Systems by 2014, No. 2, ISSN 2075-9460), vol. 25(4), pp. 322-327, 2021. Available: http://intsysjournal.org/pdfs/25-4/MazurinBernadott.pdf

[3] Bernadotte A. The Algorithm that maximizes the accuracy of $k-$classification on the set of representatives of the k equivalence classes // *Mathematics*, vol. 10(15), p. 2810, **2022**, DOI: https://doi.org/10.3390/math10152810

[4] Zubov A. Isaeva M. Bernadotte A.. Neural network classifier of EEG$-$data from people who have undergone COVID-19 and have not // Intelligent systems. Theory and Applications (formerly: Intelligent Systems by 2014, No. 2, ISSN 2075-9460), vol. 25(4), pp.318-322, 2021. Available: http://intsysjournal.org/pdfs/25-4/ZubovIsaevaBernadott.pdf

[5] Saining Xie et al. Aggregated Residual Transformations for Deep Neural Networks *In: CoRR abs/1611.05431 arXiv* **2016**, *1611.05431*, url: http://arxiv.org/abs/ 1611.05431.