Московский Государственный Университет имени М.В. Ломоносова Факультет вычислительной математики и кибернетики

Распознавание мысленно произносимых фонем по данным ЭЭГ и ЭМГ

ДОКЛАД

Выполнила: Поиленкова Анна,
621 группа

2024

Введение

В современном мире техника стала неотъемлемой частью нашей жизни. Вычислительные устройства помогают нам в работе, учебе и быту, расширяя наши возможности и экономя время. Способы взаимодействия с вычислительными ресурсами разнообразны. Например, клавиатуры заменили перфокарты для ввода текста, а "мышь" стала популярным устройством ввода.

Современная мобильность устройств привела к широкому использованию голосового ввода, что требует новых технологий для его обработки и анализа. Одним из способов общения с машинами на естественном языке является система автоматического распознавания речи. Однако такие системы ненадежны в условиях окружающего шума и могут нарушить конфиденциальность диалога. Предлагается подход, где устройство ввода использует электрофизиологический сигнал для управления устройством или компьютером. Эти сигналы могут быть получены из различных источников, таких как мозг, мышцы или глаза.

Устройства ввода на основе электрофизиологии представляют собой многообещающую технологию, которая может иметь широкий спектр применений в медицине, науке и быту, а также помогать людям с ограниченными возможностями. У некоторых людей вследствие медицинских причин нет возможности общаться не только с людьми, но и компьютерами с помощью классических устройств ввода.

Наиболее распространенными устройствами ввода на основе электрофизиологии является электроэнцефалография (ЭЭГ) и электромиография (ЭМГ). ЭЭГ отражает малейшие изменения функции коры головного мозга за счет регистрации электрической активности в мозге. Построенный по такому принципу, интерфейс мозг-компьютер (Brain-Computer Interface, BCI) может быть использован для взаимодействия с вычислительным устройством.

Одна из областей применения ЭЭГ и ЭМГ — это распознавание произнесенных слов, фраз или фонем. Особенный интерес и развитие направлено в область распознавания внутренней речи.

Внутренняя речь — беззвучная, мысленная речь, которая возникает в тот момент, когда мы думаем о чем-либо или по-другому, когда произносим что-то про себя. Эта речь обладает своими свойствами и является производной от внешней речи. Исследования показали, что во время мысленной речи возникает отчетливая речедвигательная импульсная активность либо в форме повышения общего тонуса речевой мускулатуры, либо в форме кратковременных вспышек. Однако импульсы могут ослабевать или даже исчезнуть по мере автоматизации действий, сопровождаемых внутренней речью [1].

Сейчас выделяется несколько подходов к созданию интересов распознавания внутренней речи: на основе полных слов, слогов или фонем. В первом случае существуют проблемы, поскольку количество слов для классификации резко возрастает. Таким образом, существующие системы сводят поиск к очень ограниченному набору фраз или целых предложений. Другой подход заключается в распознавании слогов. Такой способ сохраняет информацию о звуке и ритме речи. Однако в большинстве исследований по реконструкции речи стимульным материалом выступают фонемы, поскольку они являются более мелкими элементами речи, чем слоги. Как из фонем, так из слогов можно воссоздать речь, и в зависимости от языка и особенностей человека будет проще одно или другое. Фонемы и слоги позволяют в дальнейшем построить и интегрировать классификатор в систему ввода.

В работе рассматривается задача распознавания мысленно произносимых фонем русского языка на основании данных, получаемых с устройства электроэнцефалографии с учетом данных электромиографии в области, отвечающей за речь.

Задача распознавания разбивается на несколько подзадач: задачу извлечения и сбора данных, их предобработки, и построение алгоритма классификации фонем.

Целью является разработка алгоритма для распознавания мысленно проговариваемых фонем русского языка на основе данных, получаемых с устройства электроэнцефалографии и данных электромиографа, снятых в области отвечающей за речь.

Новизна работы обусловлена добавлением данных ЭМГ, которые будут участвовать в классификации, а также использованием данных с мысленным произношением русских фонем. Существующие работы основаны на англоязычных или других иностранных фонемах.

Обзор литературы

Задачу распознавания мысленного проговаривания решают, как с помощью классического машинного обучения, так и с помощью нейронных сетей. В качестве исходных данных могут выступать сигналы ЭЭГ или ЭМГ, функционально магнитно-резонансная томография (фМРТ), а также электрокортикография (ЭКоГ) — метод прямой регистрации биоэлектрической активности коры головного мозга непосредственно на поверхности коры. В основном нейроинтерфейсы для задачи распознавания безмолвной речи строятся либо на данных снятых с лицевых мышц, то есть сигналах ЭМГ, или только на данных снятых с головы — ЭЭГ.

В работе [2] по данным электромиографии, снятых в области шеи, классифицировались мысленно-проговариваемые фонемы хинди с помощью

метода опорных векторов (Support Vector Machine, SVM). Из данных предварительно извлекали признаки с помощью вейвлет-преобразования. Полученные результаты для бинарной классификации — 75–80%.

Авторы работы [3] использовали SVM и решающие деревья (Decision Tree, DT). Была получена точность для многоклассовой классификации 19.69% и 20.8% соответственно для каждого метода. Использовалось мел-частотное кепстральное преобразование для извлечения признаков из данных ЭЭГ.

В статье [4] была проделана большая исследовательская работа, которая выделяется большой выборкой данных из 270 испытуемых. Авторами с помощью сверточной нейронной сети была получена точность в 85% при классификации 9 слов, а при бинарной классификации точность в 88% в среднем.

В работе [5] на открытых данных [6] (4 испанские слова: вверх, вниз, налево, направо) были получены следующие результаты по метрике accuracy: SVM - 26.2%, XGBoos - 27,9%, BiLSTM - 36,1%. Для глубокого обучения использовались как необработанные данные со всех каналов, так и каналы, связанные с наиболее важными функциями, извлеченными с помощью XGBoost. Также выделение признаков производилось с помощью спектральной плотности мощности (Power spectral density, PSD) по методу Уэлча на основе относительной мощности в определенных диапазонах частот: альфа (8–13 Гц), бета (13–30 Гц) и гамма (30–100 Гц). Помимо этого авторы использовали открытые данные [7] испанских гласных (/a/, /e/, /i /, /o/, /u/) и слов (вверх, вниз, налево, направо). Наилучшая точность в 25.1% была получена на BiLSTM при уровне случайности в 16.6%. Сами же авторы данные [7] использовали для классификации SVM и DT, а для извлечения признаков дискретное вейвлет-преобразование (Discrete Wavelet Transform, DWT) с использованием вельвета Добеши. Получения средняя точность распознавания 6 гласных составила для RF 22.72%, а для SVM 21.94%.

В работе [8] используется вышеупомянутый набор данных [7]. Авторы провели подробный анализ параметров обучений и получили среднюю точность от 28,95% до 30,25% на модели CNN.

Некоторые используют комбинации разных моделей нейронных сетей. Например, в работе [9] модель состоит из трех сетей: свёрточной нейронной сети (Convolutional Neural Network, CNN), сети с длительным и коротким периодом (Long Short-Term Memory, LSTM) и глубокого автоэнкодера. Достигнутая точность для бинарной классификации - 77,9%.

По результату обзора можно подчеркнуть, что для извлечения признаков из сырых данных широко используются статистики (среднее, стандартное отклонение и тд.), нахождение спектральной плотности, а также вейвлет-преобразования. В работах часто используются традиционные методы машинного обучения (SVM, LDA), прежде всего из-за ограничения

объема данных, доступных для обучения. Среди нейросетевых подходов часто встречается CNN[]. В качестве метрики качества в работах используется accuracy — доля объектов, для которых правильно предсказан класс.

Данные

Для сбора данных и их извлечения использовался неинвазивный метод на основе ЭЭГ и ЭМГ. Такой способ позволяет получать данные без хирургических вмешательств, что делает этот способ доступным и более простым в реализации. Однако стоит отметить, что неинвазивные методы содержат больше шумов и имеют низкое разрешение, что означает невозможность точного определения местоположения источников электрических сигналов внутри головного мозга.

Регистрация электрической активности мозга проводилась с помощью 19-канального электроэнцефалографа, датчики которого крепились на поверхность головы. Для сбора данных ЭМГ используются области речевых мышц, которые связаны с артикуляционной активностью. Один датчик клеился на гортань, второй над губой. На Рисунке 1 и 2 представлены схемы расположения электродов для снятия ЭЭГ И ЭМГ соответственно. Данные в таком случае представляют временные ряды по каждому каналу для каждого испытуемого.

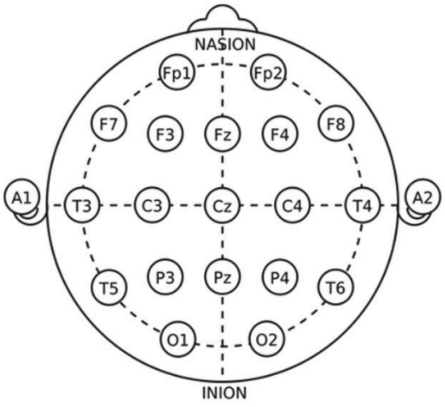

Рис 1. Схема расположения электродов электроэнцефалографа

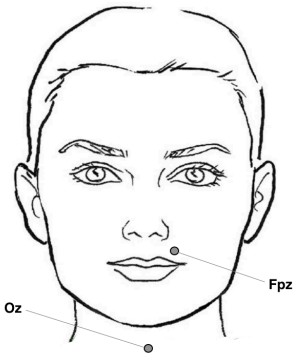

Рис 2. Схема расположения электродов электромиографа

Рассматривались 6 испытуемых в возрасте от 17 до 28 лет. В качестве предъявляемых стимулов были выбраны семь фонем русского языка: А — [а], Б — [б], Ф — [ф], Г — [г], М — [м], Р — [р], У — [у], где первое фонемы, а в скобках обозначение звука. Схема эксперимента состоит из четырех этапов: предъявления стимула, пауза, мысленное проговаривание фонем, пауза. Все этапы четко разделены командами, информация о времени которых заносится в виде меток для каждого испытуемого в отдельный канал записи. Стимул используется

визуальный, то есть демонстрируется изображение фонемы, которую нужно будет проговорить. Порядок предъявления фонем используется случайный.

В исследованиях [11, 12] в качестве результата были выделены области, которые порождают наибольшую активность в процессе мысленного проговаривания фонем — это области Брока (Broca) и Вернике (Wernicke), которым соответствуют электроды F7, F3, T3, C3 на Рисунке 1.

Сигналы с выбранных каналов предварительно очищаются от излишних шумов путём отсечения сигнала вне частотного диапазона от 3 до 30 Гц с целью удаления низко- и высокочастотных шумов. Очистка производится с использованием фильтра Баттерворта.

Запись эксперимента для каждого испытуемого сопровождалась фиксацией следующих меток: 1N - начало предъявления стимула, N - команда для начала мысленного проговаривания фонемы, где N — номер фонемы(А – 1, Б – 2, Ф – 3, Г – 4, М – 5, Р – 6, У – 7).

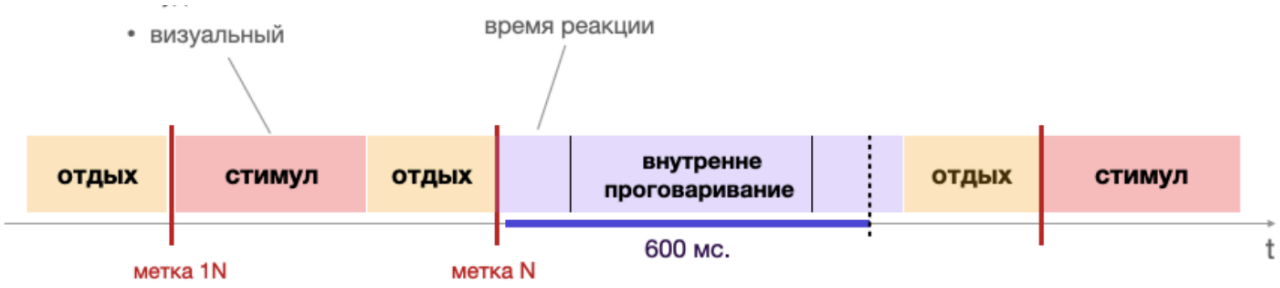

Рисунок 3. Разметка данных.

Из очищенные данных для каждого испытуемого отбираются сегменты длиной T = 600 миллисекунд, начинающиеся с метки N. Визуально это отражено на Рисунке 3. Такой временной интервал выбран поскольку считается, что за это время испытуемый успевает среагировать на команду и мысленно проговорить фонему.

Количество сегментов для каждого испытуемого представлено на Рисунке 4. Посколько количество данных мало для обучения разделение выборки будет проводится только на тренировочную и тестовую в соотношение 4 к 1. Классы меток сбалансированны для каждого испытуемого.

В качестве **метрики качества** используют метрику числа верно классифицированных временных рядов — accuracy.

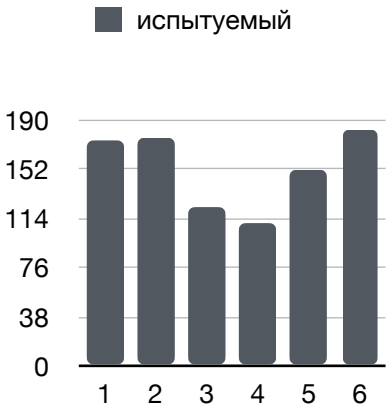

Рисунок 4. Количество сегментов для каждого испытуемого.

Методология

В качестве входных данных используются значения, полученные с электродов электроэнцефалографа, расположенных на поверхности головы, и значения сигналов электромиографа, снятых в области отвечающей за речь. На основе этих данных необходимо разработать алгоритм для распознавания (классификации) мысленно проговариваемых фонем русского языка.

Классическое машинное обучение

Рассмотрим SVM для многоклассовой классификации на семи классах фонем. В качестве метода выделения признаков лучше всего себя показало четырехуровневое дискретное вейвлет-преобразование (Discrete Wavelet Transform, DWT) с использованием вельвета Добеши и с последующим извлечением коэффициентов авторегрессии из низкочастотных сигналов. На Рисунке 5 продемонстрировано преобразование исходного сигнала и его разложение. Двойной рамкой обозначены сигналы, для которых находят авторегрессионные коэффициенты, а также медианное абсолютное отклонение (Median absolute deviation, MAD) и среднеквадратичное отклонение (Standard deviation, STD).

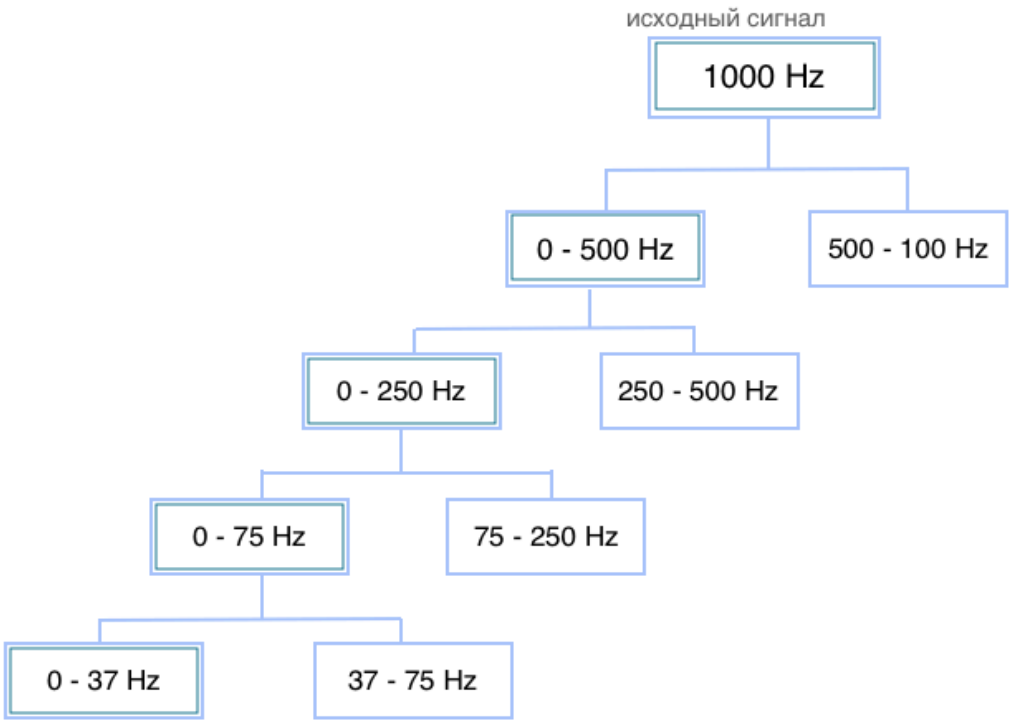

Рисунок 5. Четырехуровневое дискретное вейвлет-преобразование для исходного сигнала.

Таким образом число признаков для одного канала равно $n_{level} * (n_{AR} + 2)$, где $n_{level} = 4$, а n_{AR} - порядок авторегрессии. Будем использовать авторегрессионную модель 4, 6 и 8 порядка (в дальнейшем будем обозначать AR4, AR6, AR8).

В экспериментах использовался SVC с различным типом ядер (linear, poly, rbf, sigmoid) и разным количеством авторегрессионных коэффициентов. Обучение производилось для каждого испытуемого отдельно. Распределение метрики качества было получено с помощью метода Bootstrap. В результате модели с полиномиальным ядром (poly) показали наибольшую точность. Разделение сегментов на тестовую и обучающее выборку закреплено и использовалась во всех дальнейших экспериментах.

Дальнейшие эксперименты проводились для подбора параметров SVC модели с полиномиальным ядром. С помощью оптимизатора Optuna[13] при подборе значений для каждого испытуемого было замечено, что лучшее качество дает 3 степень полинома (degree), параметр регуляризации (C) нужно выбирать в пределах от 0.005 до 0.01, а коэффициент ядра не влиял на точность.

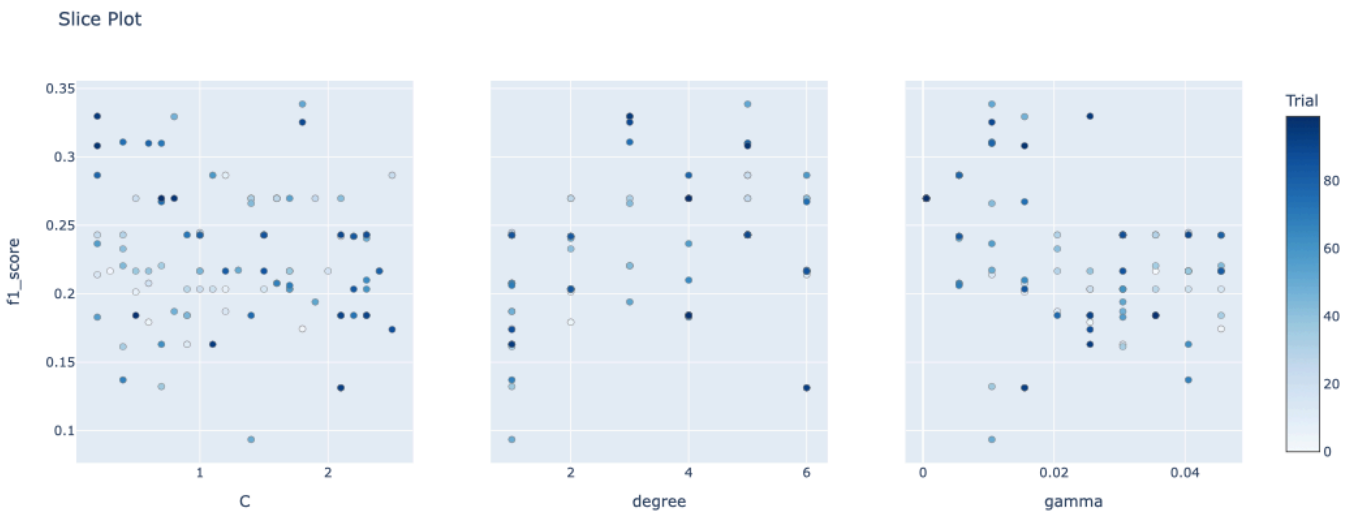

Рисунок 6. Подбор гиперпараметров для SVC для одного из испытуемых с помощью инструмента Optuna.

Приведем таблицы 1 и 2 усредненных значений точности для моделей с выбранными параметрами C = 2, degree = 3, gamma = 0.01. В первом случае обучение производилось на 4 выбранных каналах ЭЭГ и таком случае количество признаков для одного сегмента равно $16(n_{AR} + 2)$. Во втором данные объединялись с каналами ЭМГ и тогда количество признаков

составляет $24(n_{AR}+2)$. На Рисунке 7 показан один из графиков значений точности с доверительным интервалом. Значения получены с помощью метода Bootstrap.

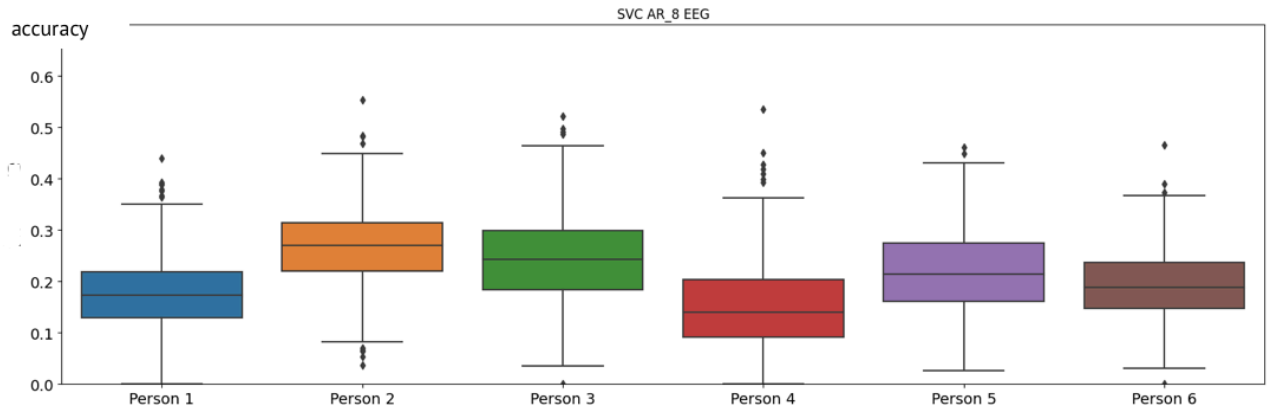

Рисунок 7. Пример графиков полученных метрик качества.

Таблица 1. Усредненные значения точности для данных ЭЭГ.

	Испытуемый					
	1	2	3	4	5	6
AR4	26.8%	13.6%	22 %	23.4%	15.3%	22 %
AR6	24.3%	24.1%	32.5%	11.1%	27.8%	23.6%
AR8	17.5%	26.8%	24.3%	14.7%	21.7%	19.2%

Таблица 2. Усредненные значения точности для объединённых данных ЭЭГ и ЭМГ.

	Испытуемый					
	1	2	3	4	5	6
AR4	21.6%	12.7%	20 %	23.1%	13.4%	10.4%
AR6	23.5%	18.2%	20.3%	16.9%	18.8%	15.1%
AR8	15 %	15.4%	18.4%	17.2%	18.3%	13.4%

Результаты

Результаты проведенного экспериментального исследования показали, что задача многоклассовой классификации решается более эффективно с помощью классического машинного обучение, что может быть связано с ограниченным объемом данных. На модели с SVC с подобранными гиперпараметрами на данных ЭЭГ точность в среднем 23.9% по всем испытуемым, а при объединении данных ЭЭГ и ЭМГ —18.8%.

Литература

1. Соколов, Александр Николаевич. Внутренняя речь и мышление / Александр Николаевич Соколов. — "Просвещение, 1968.
2. Khan, Munna. Classification of myoelectric signal for sub-vocal Hindi phoneme speech recognition / Munna Khan, Mosarrat Jahan // Journal of Intelligent & Fuzzy Systems. — 2018. — Vol. 35, no. 5. — Pp. 5585–5592.
3. Cooney, Ciaran. Mel frequency cepstral coefficients enhance imagined speech decoding accuracy from EEG / Ciaran Cooney, Rafaella Folli, Damien Coyle // 2018 29th Irish Signals and Systems Conference (ISSC) / IEEE. — 2018. — Pp. 1–7.
4. Silent eeg-speech recognition using convolutional and recurrent neural network with 85% accuracy of 9 words classification / Darya Vorontsova, Ivan Menshikov, Aleksandr Zubov et al. // Sensors. — 2021. — Vol. 21, no. 20. — P. 6744.
5. Gasparini F., Cazzaniga E., Saibene A. Inner speech recognition through electroencephalographic signals //arXiv preprint arXiv:2210.06472. – 2022.
6. Nieto N. et al. Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition //Scientific Data. – 2022. – Т. 9. – №. 1. – С. 52.
7. Coretto G. A. P., Gareis I. E., Rufiner H. L. Open access database of EEG signals recorded during imagined speech //12th International Symposium on Medical Information Processing and Analysis. – SPIE, 2017. – Т. 10160. – С. 1016002.
8. Cooney C. et al. Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG //Sensors. – 2020. – Т. 20. – №. 16. – С. 4629.
9. Saha, Pramit. Deep learning the EEG manifold for phonological categorization from active thoughts / Pramit Saha, Sidney Fels, Muhammad Abdul-Mageed // ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) / IEEE. — 2019. — Pp. 2762–2766.
10. Panachakel J. T., Ramakrishnan A. Decoding covert speech from EEG-a comprehensive review. Front. NeuroSci.(2021). – 2021.
11. Single-sweep EEG analysis of neural processes underlying perception and production of vowels / Daniel E Callan, Akiko M Callan, Kiyoshi Honda, Shinobu Masaki // Cognitive brain research. — 2000. — Vol. 10, no. 1-2. — Pp. 173–176.
12. Suppes, Patrick. Brain wave recognition of words / Patrick Suppes, Zhong-Lin Lu, Bing Han // Proceedings of the National Academy of Sciences. — 1997. — Vol. 94, no. 26. — Pp. 14965–14969.
13. [Электронный ресурс] URL: https://optuna.org