
Декодирования сигналов головного мозга в аудиоданные

Набиев Мухаммадшариф
Кафедра интеллектуальных систем
МФТИ
nabiev.mf@phystech.edu

Севериков Павел
Кафедра интеллектуальных систем
МФТИ
pseverilov@gmail.com

Аннотация

В данной работе исследуется проблема декодирования сигналов головного мозга в аудиосигналы с использованием физически-информированных методов получения эмбедингов сигналов. Предлагается решить задачу классификации стимулов по соответствующим сегментам аудиоданных. Под стимулом понимается аудиосигнал, который вызвал активность мозга, соответствующая ЭЭГ-сигналу. Данные для задачи представляют собой 183 пар вида аудиофрагмент и мозговая активность в виде ЭЭГ, который он вызвал. Продолжительность каждой записи составляет примерно 15 минут. В качестве критерия качества для выбора оптимальной модели используется средняя доля правильных ответов. Предлагается исследовать методы получения скрытых представлений, которые учитывают физические принципы, с целью улучшения качества обработки аудиосигналов и повышения точности их декодирования. Полученные результаты имеют важное значение для развития интерфейсов мозг-компьютер и понимания принципов обработки аудиосигналов человеческим мозгом.

Keywords auditory EEG decoding · natural speech processing · EEG

1 Введение

Слух, одно из наиболее важных человеческих чувств, играет решающую роль в нашем повседневном взаимодействии с окружающим миром. Однако многие люди со всего мира сталкиваются с проблемами слуха, которые могут серьезно ограничить их способность воспринимать звуки окружающей среды. В свете этих проблем возникает интерес к исследованию взаимосвязи между звуком и мозговыми сигналами [6]. В данной области выделена задача декодирования мозговых сигналов в аудиоданные.

Задачу декодирования можно поставить двумя способами: с помощью классификации и регрессии. В данной работе мы сконцентрируемся на задаче классификации. Требуется решить задачу мульти-классовой классификации, когда на вход подается ЭЭГ сигнал и K стимулов, из которых только один соответствует сигналу. Под стимулом подразумевается сегмент аудио, который стимулировал сигнал в мозгу субъекта.

Существует базовое решение этой задачи, использующее расширенную сверточную нейросеть [1]. Оно состоит из трех главных блоков: блок для пространственного преобразования ЭЭГ, энкодер для ЭЭГ и энкодер для стимула. Энкодеры ЭЭГ и стимула получают эмбединги путем свертки со расширенными ядрами. Далее считается близость представлений и определяется стимул.

Известна модификация базового решения с использованием многомерного внимания (Multi-head attention) и управляемого рекуррентного блока (Gated Recurrent Unit) [5]. Также авторы генерируют спектрограмму для получения дополнительных признаков, как, например, частота. Спектрограмма проходит через управляемый рекуррентный блок, а уже потом подаются на вход в энкодер стимула. После получения скрытых представлений, аналогично базовому решению, считается близость.

В постановке классификации наиболее успешные были работы, которые учитывали пол говорящего и особенности речи, такие как фундаментальная частота. В работе [12] авторы показали высокую

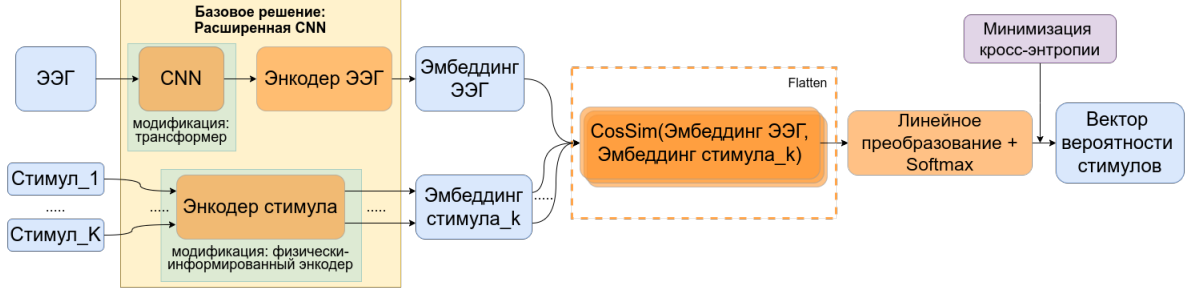


Рис. 1: Архитектура модели. Базовое решение представляет собой расширенную сверточную нейросеть в качестве энкодера ЭЭГ-сигнала и стимулов, а также обычную сверточную нейросеть для пространственного преобразования ЭЭГ-сигнала. Получаются скрытые представления ЭЭГ-сигнала и стимулов, считается их близость и находится истинный стимул. Предлагается использовать трансформера-кодировщика, чтобы уменьшить число каналов ЭЭГ-сигнала до подачи в расширенную сверточную нейросеть и физико-информированные энкодеры для получения эмбедингов стимулов. После получения эмбедингов стимулов аналогично определятся истинный стимул, как наиболее близкий к эмбедингу ЭЭГ-сигнала.

чувствительность ЭЭГ-сигнала от фундаментальной частоты, заменив стимул на его фундаментальную частоту и значительно улучшили качество за счет ансамблирования базового решения. Хотя и выделение фундаментальной частоты повысило качество в целом, было выяснено, что такой подход сильно зависит от пола говорящего [8]. На качество классификации также влияет частота дискретизации стимулов, как это показано в работе [11]. Чем больше частота дискретизации, тем сложнее установить зависимость, и как следствие, снижается качество.

Современные физико-информированные энкодеры аудиосигналов, учитывают разные детали речи, по типу фонов и информацию на уровне слов [13], поэтому решения использующие такие энкодеры показывают хорошие результаты [16]. Такие решения из аудиосигнала получают скрытое представления за счет физико-информированного энкодера, тем самым в латентном векторе инкапсулируется информация о речи.

Решению задачи декодирования в постановке регрессии посвящена статья [7]. Авторами была предложена модель под названием Pre-LLN FFT, основанная на модели прямого распространения с трансформером (Feed-Forward Transformer) из [10]. За счет модификации FFT и добавления информации о субъекте, в качестве внешнего признака [14] и нормализации подготовительного слоя [17], который идет перед FFT, авторы добились улучшения коэффициента корреляции Пирсона по сравнению с базовым решением использовавшим свертки и нормализации слоев [2].

Предлагается проанализировать влияние физико-информированных энкодеров, конкретно модели Wav2Vec2 и Whisper, для стимулов и их спектрограмм, и трансформера-кодировщика для пространственного преобразования ЭЭГ.

2 Постановка задачи

Каждый объект представляет собой кортеж $(\mathbf{X}^i, \mathbf{s}_1^i, \dots, \mathbf{s}_K^i)$, где $\mathbf{X}^i \in \mathbb{R}^{64 \times T}$ — ЭЭГ-сигнал с 64 каналами, $\mathbf{s}_1^i, \dots, \mathbf{s}_K^i \in \mathbb{R}^T$ — стимулы, а K — количество стимулов и T — длина окна. В формальной постановке под стимулом понимается огибающая аудиосегмента длительностью T (рис. 2). Меткой данного объекта будет являться вектор $\mathbf{y}^i \in \{0, 1\}^K$. Метка имеет только одну координату равную единице, которая соответствует стимулу, спровоцировавшему ЭЭГ-сигнал. Требуется по имеющимся $\mathbf{X}^i, \mathbf{s}_1^i, \dots, \mathbf{s}_K^i$ получить распределение вероятностей стимулов $\mathbf{p}^i = [p_1^i, \dots, p_K^i]^T$. Пусть модель представляет собой следующее отображение $\mathbf{f} : \mathbb{R}^{64 \times T} \times (\mathbb{R}^T)^K \rightarrow [0, 1]^K$. Задача сводится к минимизации Cross-Entropy Loss:

$$CE = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K y_k^i \log ([\mathbf{f}(\mathbf{X}^i, \mathbf{S}^i)]_k),$$

где $\mathbf{S}^i = (\mathbf{s}_1^i, \dots, \mathbf{s}_K^i)$. То есть решается задача мультиклассовой классификации.

3 Описание модели

Архитектура модели представлена на рис. 1. Модель получает на вход ЭЭГ-сигнала \mathbf{X} и набор стимулов $\mathbf{s}_1, \dots, \mathbf{s}_K$. В начале снижается размерность ЭЭГ-сигнала от 64 до 8 и подается на вход энкодеру ЭЭГ. Также и стимулы проходят через энкодер и, получив представление ЭЭГ-сигнала \mathbf{E} и представления стимулов $\mathbf{Z}_1, \dots, \mathbf{Z}_K$ в латентном пространстве $\mathbb{R}^{16 \times T}$, считается их близость по формуле $\mathbf{C}_k = \text{CosSim}(\mathbf{E}, \mathbf{Z}_k) = \mathbf{E} \mathbf{Z}_k^T$ для $k \in \{1, \dots, K\}$. Далее для каждой матрицы производится линейное преобразование $c_k = \mathbf{w}^T \mathbf{r}_k + b$, где \mathbf{r}_k — матрица \mathbf{C}_k в виде вектора, $\mathbf{w} \in \mathbb{R}^{256}$ — вектор коэффициентов и b — свободный член. Итоговое распределение вероятностей получается, как $\mathbf{p} = \text{SoftMax}([c_1, \dots, c_K]^T)$.

3.1 Базовая модель

В базовом решении для пространственного преобразования ЭЭГ-сигнала используется сверточная нейронная сеть. Одномерная свертка с ядром 1×1 и 8 фильтрами объединяет информацию по всем 64 каналам и уменьшает размерность до 8.

Структура энкодеров ЭЭГ и стимулов одинаковая, но они имеют разные веса. Эти энкодеры состоят из n блоков расширенной сверточной нейросети. Блоки идентичные и каждый из них имеет 16 фильтров с ядрами 3×3 . В пределах одного блока, в каждом слое ядро имеет свой коэффициент расширения равное 3^{m-1} .

3.2 Предлагаемые решения

Предлагается заменить блок со сверточной нейросетью на трансформер-кодировщик [15], который за счет механизма внимания сможет улавливать долгосрочные зависимости в ЭЭГ-сигнале. Трансформер-кодировщик состоит из двух слоев. Первый представляет собой многомерное внимание, а второй — полносвязная 2х-слойная нейросеть (FFN). После каждого слоя используется сквозная связь и нормализуется выход. Скрытый слой в FFN имеет размерность 32.

Модели автоматического распознавания речи (Automatic Speech Recognition) могут извлечь разные детали речи, по типу фонов и частот. В связи с этим было решено использовать современные модели ASR, такие как Wav2Vec 2.0 [3] и Whisper [9], в качестве энкодера стимула.

- Wav2Vec 2.0. Архитектура модели состоит из сверточной нейронной сети и трансформера. Сверточная нейронная сеть извлекает высокоуровневые признаки из аудио, а трансформер захватывает контекстную информацию.
- Whisper. Модель на основе трансформера с кодировщиком и декодировщиком. Она отображает последовательность признаков спектрограммы речи на последовательность текстовых токенов. Сначала исходные аудиовходы преобразуются в mel-спектрограмму с помощью извлекавателя признаков. Затем трансформер-кодировщик, формирует последовательность скрытых состояний. Наконец, декодер авторегрессивно предсказывает текстовые токены.

4 Вычислительный эксперимент

4.1 Данные

Эксперимент проверялся на данных [4]. Они представляли собой выборку из 85 человек. Все участники прослушали 6, 7, 8 или 10 стимулов на датском языке, каждая из которых имеет примерную продолжительность 15 мин. После прослушивания участников спрашивали про содержания аудиофрагмента. Это было с целью мотивировать участников обращать внимания во время прослушивания.

Стимулы были разделены на следующие категории:

- Справочные аудиокниги
- Аудиокниги для детей и взрослых. Если длина превышала 15 мин, то аудиокнига делилась на части
- Аудиокниги с шумом
- Подкасты про ответы на научные вопросы

- Подкасты с видео

Были использованы обработанные данные аудифрагментов и ЭЭГ-сигналов, а также необработанные аудиофрагменты, с частотами дискретизации равными 64 Гц и 48000 Гц соответственно. Отметим также, что обработанные аудиофрагменты представляют собой огибающую сигнала (рис. 2).

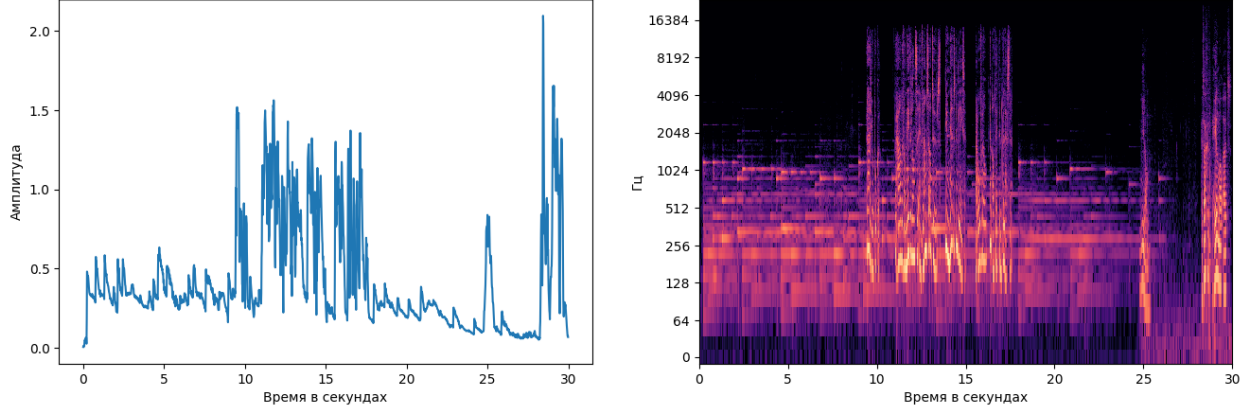


Рис. 2: Пример огибающей(слева) и спектрограммы(справа) одного и того же аудиофрагмента продолжительностью 30 секунд.

4.2 Описание эксперимента

Для эксперимента была выделена случайная подвыборка из 25 участников с равным количеством мужчин и женщин, и аудиофрагменты, которые они слушали, а также соответствующие записи ЭЭГ-сигналов. Всего 183 пар (ЭЭГ-сигнал, стимул).

Так как исходные обработанные ЭЭГ-сигналы и аудиофрагменты слишком велики, они были разбиты на окна фиксированной длины. Для каждой пары окон (ЭЭГ-сигнал, истинный стимул) были сгенерированы $K - 1$ ложных стимулов той же длины, что и длина окна. Далее для обеспечения сбалансированности классов были созданы $K - 1$ копий каждого объекта, со смещением истинного стимула.

С учетом языка аудиосигналов было решено взять предобученную модель wav2vec2-base-960h-phoneme-reco-dutch и whisper-small. Для работы с ASR моделями необработанные аудиофрагменты переводились в подходящий вид. Для каждого аудио с помощью модели создавался эмбединг, который потом выравнивался, по уже готовым данным.

В эксперименте ширина окна была взята равной 5 секунд с шагом 1 секунда. Количество стимулов $K = 5$, то есть один истинный стимул и четыре ложных. Эксперимент проводился в 10 эпох, а размер батча составлял 64 элементов. В качестве метрики взята среднее по участникам долей правильных ответов.

Обозначим множество классов, как $\{0, \dots, K - 1\}$. Учитывая это, метрика качества вычисляется по формуле

$$Score = \frac{1}{22} \sum_{i=1}^{22} \frac{1}{l_i} \sum_{j=1}^{l_i} [y_j^i = \text{pred}_j^i],$$

где $y_j^i \in \{0, \dots, K - 1\}$ — метка объекта, l_i — количество пар ЭЭГ-стимул для i -го участника, а pred_j^i — предсказание модели на объекте j .

4.3 Результаты эксперимента

Результаты экспериментов приведены в таблице 1. Видно, что комбинирование трансформера-кодировщика с физико-информированными моделями Wav2Vec 2.0 и Whisper-small дали наилучшие результаты.

| Model | Score (%) |
|-------------------------------------|------------------------------------|
| Baseline | 47.68 ± 11.75 |
| Transformer Encoder | 48.15 ± 10.33 |
| Wav2Vec2 | 47.92 ± 11.54 |
| Whisper-small | 48.04 ± 9.85 |
| Transformer Encoder + Wav2Vec2 | 48.70 ± 9.44 |
| Transformer Encoder + Whisper-small | 48.36 ± 9.24 |

Таблица 1: Оценка качества на использованных моделях.

По предсказаниям на тестовых данных построим график boxplot 3 для каждой модели. По оси ординат указано доля правильных ответов для каждого участника.

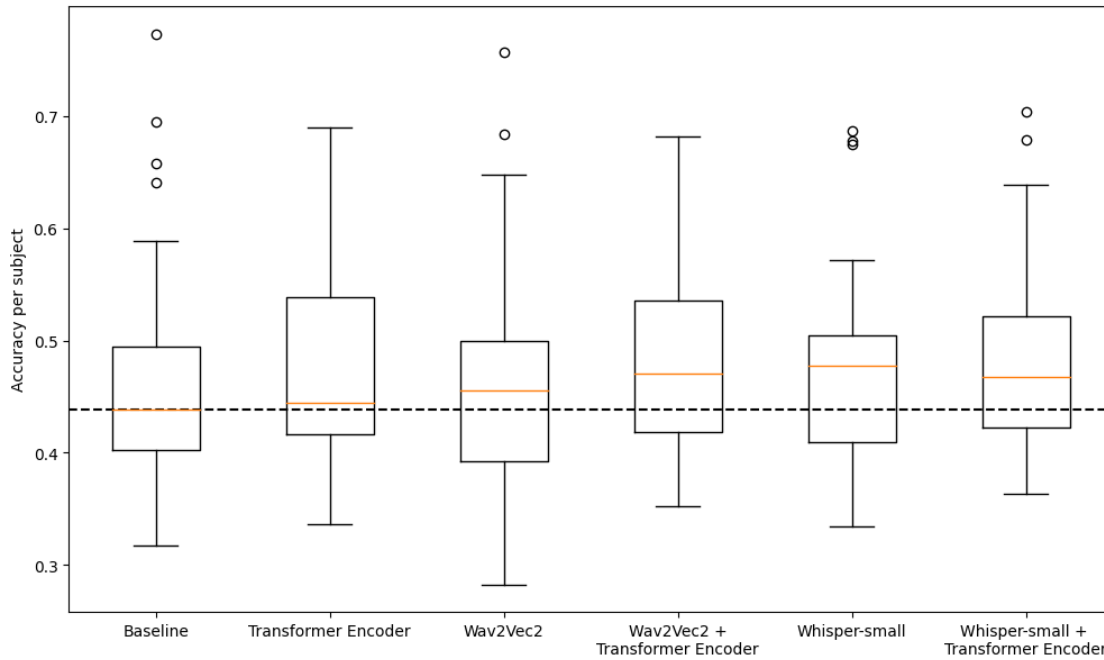


Рис. 3: График boxplot построенный на тестовых данных

Как можно видеть по графику, межквартильный размах расположен выше базового решения у комбинированных моделей и трансформера-кодировщика. Согласно графику, наибольший прирост качества получается за счет трансформера-кодировщика и модели Wav2Vec 2.0. Это объясняется тем, что для построения эмбедингов была взята предобученная модель Wav2Vec 2.0, которая была настроена на датский язык. Использование эмбедингов модели Whisper-small чуть уступает по качеству, так как предобученная модель является общей для всех языков и не настраивалась именно на датский язык.

5 Заключение

Эксперименты показали, что использование трансформеров и физико-информированных моделей в качестве энкодеров приводит к улучшению качества классификации истинного стимула для ЭЭГ-сигнала. Этот результат подтверждает гипотезу о том, что интеграция физических законов может улучшить точность их классификации. Эти результаты имеют важное значение для практических приложений, таких как разработка интерфейсов мозг-компьютер.

В дальнейшем планируется провести эксперименты с разными размерностями скрытого состояния физико-информированных энкодеров. Ожидается найти баланс между информацией инкапсулируемой в эмебеддинге и её размерностью, так как большие размерности требуют огромных затрат памяти.

Список литературы

- [1] Bernd Accou, Mohammad Jalilpour-Monesi, Jair Montoya-Martínez, Hugo Van hamme, and Tom Francart. Modeling the relationship between acoustic stimulus and eeg with a dilated convolutional neural network. 2020 28th European Signal Processing Conference (EUSIPCO), pages 1175–1179, 2021.
- [2] Bernd Accou, Jonas Vanthornhout, Hugo Van hamme, and Tom Francart. Decoding of the speech envelope from eeg using the vlaai deep neural network, 09 2022.
- [3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.
- [4] Lies Bollens, Bernd Accou, Hugo Van hamme, and Tom Francart. SparrKULee: A Speech-evoked Auditory Response Repository of the KU Leuven, containing EEG of 85 participants, 2023.
- [5] Marvin Borsdorf, Saurav Pahuja, Gabriel Ivucic, Siqi Cai, Haizhou Li, and Tanja Schultz. Multi-head attention and gru for improved match-mismatch classification of speech stimulus and eeg response. pages 1–2, 06 2023.
- [6] Steffen Dasenbrock, Sarah Blum, Paul Maanen, Stefan Debener, Volker Hohmann, and Hendrik Kayser. Synchronization of ear-eeg and audio streams in a portable research hearing device. *Frontiers in Neuroscience*, 16, 09 2022.
- [7] Zhenyu Piao, Miseul Kim, Hyungchan Yoon, and Hong-Goo Kang. Happyquokka system for icassp 2023 auditory eeg challenge, 2023.
- [8] Corentin Puffay, Jana Van Canneyt, Jonas Vanthornhout, Hugo Van hamme, and Tom Francart. Relating the fundamental frequency of speech with eeg using a dilated convolutional network. In *Interspeech*, 2022.
- [9] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.
- [10] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech: Fast, robust and controllable text to speech, 2019.
- [11] Mike Thornton, Jonas Auernheimer, Constantin Jehn, Danilo Mandic, and Tobias Reichenbach. Detecting gamma-band responses to the speech envelope for the icassp 2024 auditory eeg decoding signal processing grand challenge. *ArXiv*, abs/2401.17380, 2024.
- [12] Mike Thornton, Danilo P. Mandic, and Tobias Reichenbach. Relating eeg recordings to speech using envelope tracking and the speech-ffr. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2, 2023.
- [13] Aditya R. Vaidya, Shailee Jain, and Alexander G. Huth. Self-supervised models of audio effectively explain human cortical responses to speech, 2022.
- [14] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, page 125, 2016.
- [15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.
- [16] Bo Wang, Xiran Xu, Zechen Zhang, Haolin Zhu, Yujie Yan, Xihong Wu, and Jing Chen. Self-supervised speech representation and contextual text embedding for match-mismatch classification with eeg recording. *ArXiv*, abs/2401.04964, 2024.
- [17] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture, 2020.