

---

# Декодирования сигналов головного мозга в аудиоданные

---

Набиев Мухаммадшариф  
Кафедра интеллектуальных систем  
МФТИ  
nabiev.mf@phystech

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 эмбеддингов 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 668 пар вида ЭЭГ-стимул общей продолжительностью стимулов 9431 минута. В качестве метрики для выбора оптимальной модели используется F1-мера. В данной работе предлагается исследовать передовые методы машинного обучения, 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. В данной области выделена задача декодирования мозговых сигналов в аудиоданные.

Задачу декодирования можно поставить двумя способами: классификация и регрессия. В данной работе мы сконцентрируемся на задаче классификации. Требуется решить задачу классификации в парадигме match-mismatch, когда на вход подается ЭЭГ сигнал и 5 стимулов, из которых только один соответствует сигналу. Под стимулом подразумевается сегмент аудио, который стимулировал сигнал в мозгу субъекта.

Существует базовое решение этой задачи, использующее dilated convolutional network [Accou et al., 2021]. Оно имеет следующую архитектуру (Рис. 1). В начале используется сверточный слой с 8 фильтрами для объединения данных из всех каналов ЭЭГ. Далее  $N$  dilated свёрток с размером ядра  $K$  применяются к выходу первого слоя и к 5 стимулам. Для каждого слоя  $L_n$  dilation factor рассчитывается, как  $K^{n-1}$ , которое взято из статьи [van den Oord et al., 2016], для того, чтобы минимизировать количество параметров. На выходе каждой свертки применяется функция активации ReLU. Затем вычисляется косинусный коэффициент между представлением ЭЭГ и представлениями стимулов. Наконец, линейный слой с сигмоидой используя эти коэффициенты классифицирует стимулы.

Известна модификация базового решения с использованием Multi-Head Attention [Vaswani et al., 2023] и GRU [Cho et al., 2014]. Дополнительно авторы генерируют также и спектрограмму для получения дополнительных признаков, как, например, частота. Также в отличие от базового решения стимулы и спектрограммы проходят через GRU, а уже пото подаются на вход в dilated convolutional блоки.

4, 5

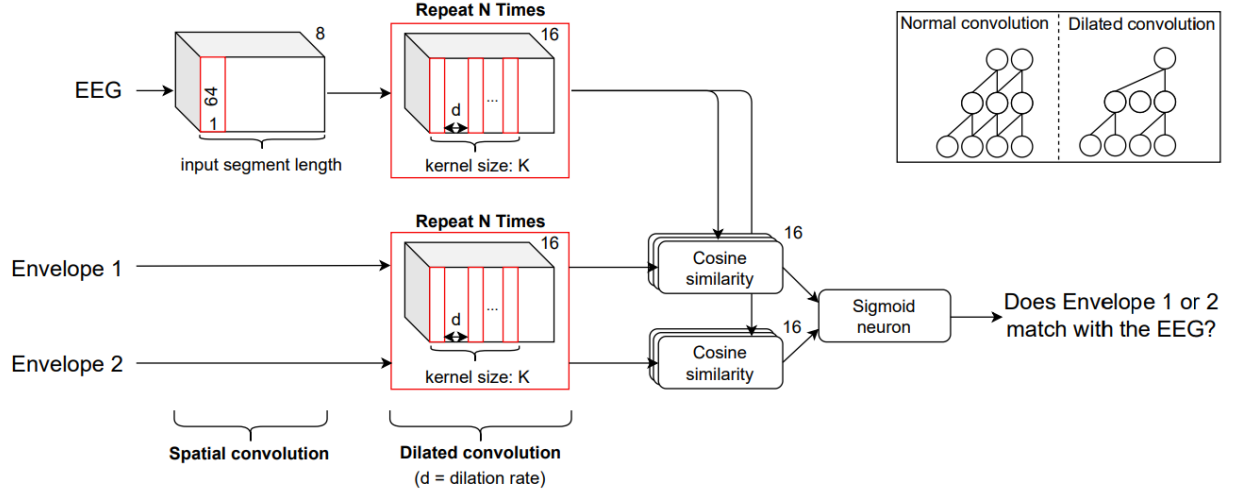


Рис. 1: Dilated convolutional network

Решению задачи декодирования в постановке регрессии посвящена статья Piao et al. [2023]. Авторами была предложена модель Pre-LLN FFT, основанная на модели Feed-Forward Transformer (FFT) network из [Ren et al., 2019]. За счет модификации FFT и добавления global conditioner [van den Oord et al., 2016] и нормализации пред-слоя Xiong et al. [2020], авторы добились улучшения коэффициента корреляции Пирсона по сравнению с базовым решением использовавшим Very Large Augmented Auditory Inference (VLA AI) [Accou et al., 2022].

Предлагается воспользоваться физическими принципами при решении задачи классификации. А именно, по стимулам сгенерировать mel-спектрограммы, а также воспользоваться Self-Attention-ом для учета дополнительных деталей голоса из стимулов и спектрограмм.

## 2 Данные

Были использованы данные SparrKULee [Bollens et al., 2023]. Данные состоят из записей ЭЭГ 85 молодых людей (18 - 30 лет) с хорошим слухом, каждый из которых слушал естественную речь на протяжении 90-150 минут.

## 3 Постановка задачи

Сигналы ЭЭГ представляет собой матрицу  $\mathbf{X} = [\mathbf{x}_1^T, \dots, \mathbf{x}_m^T]^T \in \mathbb{R}^{n \times m}$ , где  $n$ -количество каналов, а  $m$ -время. Обозначим стимулы и их метки, как  $(\mathbf{s}_1, y_1), \dots, (\mathbf{s}_5, y_5) \in \mathbb{R}^{1 \times m} \times \{0, 1\}$ . Требуется по имеющимся  $\mathbf{X}, \mathbf{s}_1, \dots, \mathbf{s}_5$  и  $\mathbf{y} = [y_1, \dots, y_5]^T$  предсказать вероятность для каждого стимула  $\mathbf{s}_k$ . Допустим, что модель из  $\mathbf{F} \subset \mathfrak{F}$ ,  $\mathfrak{F}$ -параметрическое множество моделей. Тогда задача сводится к минимизации Cross-Entropy Loss:

$$CE = - \sum_{k=1}^5 y_k \log(\mathbf{F}(\mathbf{X}, \mathbf{s}_k))$$

### 3.1 Описание модели

На входе сигналы ЭЭГ проходят через 1D сверточный слой с 8 фильтрами и ядром  $1 \times 1$  для пространственной связки каналов и уменьшения размерности.  $\tilde{\mathbf{X}} = \text{Conv}(\mathbf{X}, K_1)$ , где  $K_1$  ядро слоя. Матрица  $\tilde{\mathbf{X}}$  является двумерной и её размерность составляет  $8 \times m$ . Преобразованные сигналы ЭЭГ проходят через N dilated сверточных слоев с ядром  $K_2$  размерностью  $3 \times 3$  и 16 фильтрами. На слое  $L_p$  dilation factor возьмем равным  $K^{p-1}$ . В итоге получим итоговое представление ЭЭГ в виде матрицы  $16 \times m'$ . Через этот же сверточный слой пройдут и стимулы  $\mathbf{s}_1, \dots, \mathbf{s}_5$  и каждый из них также будет отображен

в латентное пространство  $M_{16 \times m'}(\mathbb{R})$  - пространство вещественных матрица размера  $16 \times m'$ . Получив представления в латентном пространстве высчитываются косинусные коэффициенты

$$C_k = \text{CosSim}(\mathbf{X}_{emb}, \mathbf{s}_{emb}),$$

где скалярное произведение производится по столбцам. Каждая матрица  $C_k$  размерностью  $16 \times 16$  подается на вход в линейной слой  $c_k = \text{Linear}(C_k)$ . В итоге по вектору  $[c_1, \dots, c_5]^T$  вычисляется Softmax, по значениям которой и определяется какой стимул является истинным.

## Список литературы

- 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. URL <https://api.semanticscholar.org/CorpusID:229358565>.
- 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.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.
- Zhenyu Piao, Miseul Kim, Hyungchan Yoon, and Hong-Goo Kang. Happyquokka system for icassp 2023 auditory eeg challenge, 2023.
- 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.
- 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.
- 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.
- 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. URL <https://doi.org/10.48804/K3VSND>.

1. Либо пояснить этот термин со ссылкой на литературу, либо убрать
  2. Здесь очень много технических подробностей. Уберите все формулы, буквы и перескажите основную идею. Всегда ли в этой постановке будет 5 стимулов? Если может быть другое число, укажите вместо 5 "некоторое число"
  3. Если текст на русском, лучше найти адекватный перевод этого термина на русский (он встречается в литературе)
  4. опечатка
  5. Аналогично, либо ввести с пояснением термин, либо убрать его вообще. Лучше найти перевод на русский и пояснить
  6. Описание рисунка должно его объяснять, пока объяснения нет. Если картинка не ваша - то вы не имеете права вставлять ее в работу (даже со ссылкой на автора)
  7. Нужен русский термин
  8. Нужен русский термин
  9. Пред-слоями - нетипичный термин, пояснить, если он нужен, или убрать
  10. Не хватает пары предложений о том, как в вашей работе проверяется качество предложенного метода, какие эксперименты проводятся. И я бы немного развил параграф с тем, что вы предлагаете.
  11. Эта секция должна быть позже, в экспериментах
  12. Очень коротко, нужно больше данных. Как представляется выборка? Какая размерность по каналам и т.п.
  13. Здесь и дальше - вместо дефиса должно быть тире, в латехе пишется как три дефиса "---"
  14. Вопрос, аналогичный вопросу 2 - точно здесь всегда 5 стимулов? Если нет, то пятерку заменить на букву.
  15. Лишний пробел
  16. "Допустим, что модель из  $F$ " читается как "у нас есть модель, она из множества  $F$ ", а по смыслу у вас есть модель  $F$ , которая задается из множества  $\mathfrak{F}$ . Если модель у вас работает на векторах, то с маленькой буквы.
  17. Длинное тире
  18. У вас сейчас модель принимает на вход метки объекта  $k$  и ВСЕ объекты выборки. Если это неверно, нужно переправить
  19. одномерный
  20. Предложение не начинают с формул. Объясните словами что за матрицу вы вводите
  21. Нужен русский термин, который вводится ранее
  22. Тире
  23. Пояснить словами что такое CosSim
  24. Матрица пишется жирным
  25. с  $k$  вектор? Тогда жирный.
  26. Linear - вектор-функций? Тогда жирная и пишется не  $\$Linear\$, а  $\text{\textit{Linear}}\$$ .$
- См. презентацию с последнего занятия.