Федеральное государственное автономное образовательное учреждение
высшего образования
"Московский физико-технический институт
(национальный исследовательский университет)"
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Северилов Павел Андреевич

# ICASSP 2023 AUDITORY EEG CHALLENGE

**Научный руководитель:**
д. ф.-м. н.
Стрижов Вадим Викторович

Москва
2024 г.

# Содержание

**Аннотация**

In this paper we consider neural network methods for predicting the envelope of a heard sound based on the human electroencephalogram.

# Введение

Создание нейрокомпьютерного интерфейса (НКИ) - одна из важных и популярных задач в 20-21 веках [добавить ссылки]. Под НКИ подразумевается обмен информацией между мозгом и компьютером напрямую, то есть спектр задач, который относится к данном области достаточно широк: синтезировать речь из считанных мозговых сигналов, посылать в мозг сигналы с камеры, имитирую работу глаза и т.п.

Первые известные работы в данной области были написаны ещё в середине 20-го века [найти и проверить автора, в Википедии может быть что угодно написано]. С тех пор был достигнут большой прогресс [написать про инвазивные и неинвазивные методы]. ... В 2023 году компания Meta предоставила исследования, в которых преобразовала в реальном времени МЭГ в изображение.

Решение данной проблемы имеет высокую этическую и социальную значимость, так как может сделать жизнь disabled people лучше.

# 1 Usage of EEG to predict envelope

Глобально хочется научиться преобразовывать показания мозговой активности в речь. Для этого попробуем рассмотреть частный случай такой задачи.

Пусть во время прослушивания стимула (некий источник звука) $y_i$ были сняти показания ЭЭГ $x_i$. Стимул представим в виде огибающей речевого сигнала, т.е. изменений амплитуды речи во времени.

Таким образом, $y_i \in \mathbb{R}^n$, $x_i \in \mathbb{R}^{n*m}$, где n - продолжительность звукозаписи, $m = 64$ - количество каналов для записи ЭЭГ.

Нужно выбрать такую модель из класса нейронных сетей, которая по данным ЭЭГ предсказывает огибающий речевой сигнал $\hat{y}$

$$\{f_k : (w, X) \to \hat{y} | k \in K\},$$

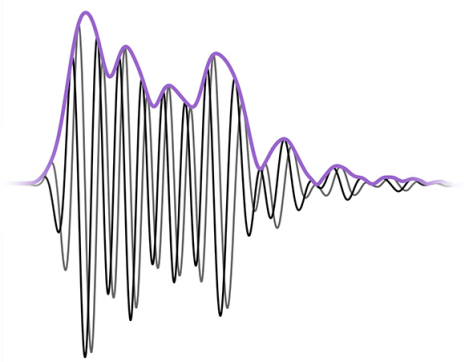

Рис. 1: Огибающая сигнала

где $w \in W$ параметры модели, $X$ - данные ЭЭГ. Причём мы хотим найти такой речевой сигнал, чтобы максимизировать коэффициент корреляции Пирсона между предсказанием и исходным стимулом:

$$\rho_{y,\hat{y},w} = \frac{cov(y, \hat{y})}{\sigma_y \sigma_{\hat{y}}},$$

где $cov(*, *)$ - ковариация, $\sigma_*$ - стандартное отколнение величин $y$ и $\hat{y}$.

То есть имеем оптимизационную задачу:

$$w^* = \arg\max_{w \in W}(\rho(w))$$

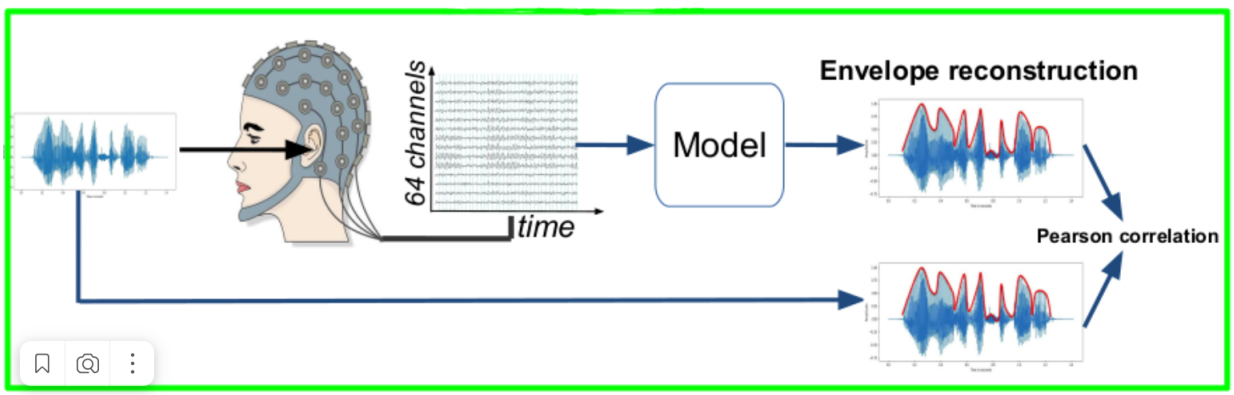


Рис. 2: Постановка задачи регрессии

# 2 Описание данных

Датасет был собран на базе Лёвенского университета - https://rdr.kuleuven.be/dataset.xhtml?persi

Для сбора датаесета были приглашены 85 человек без проблем со слухом и нервной системой, носители бельгийского голландского языка. Измерения производились в звуконепроницаемой лаборатории с помощью высокоточных приборов для снятия ЭЭГ с 64 электродами. Частота дискретезации даных 8192 Гц.

Каждому учатнику предлагалось послушать отрывок подкаста или аудиокниги (случайно) длиной до 15 минут. Всего имеем 668 пар ЭЭГ-стимул (прослушанный отрывок) общей продолжительностью 9431 минута

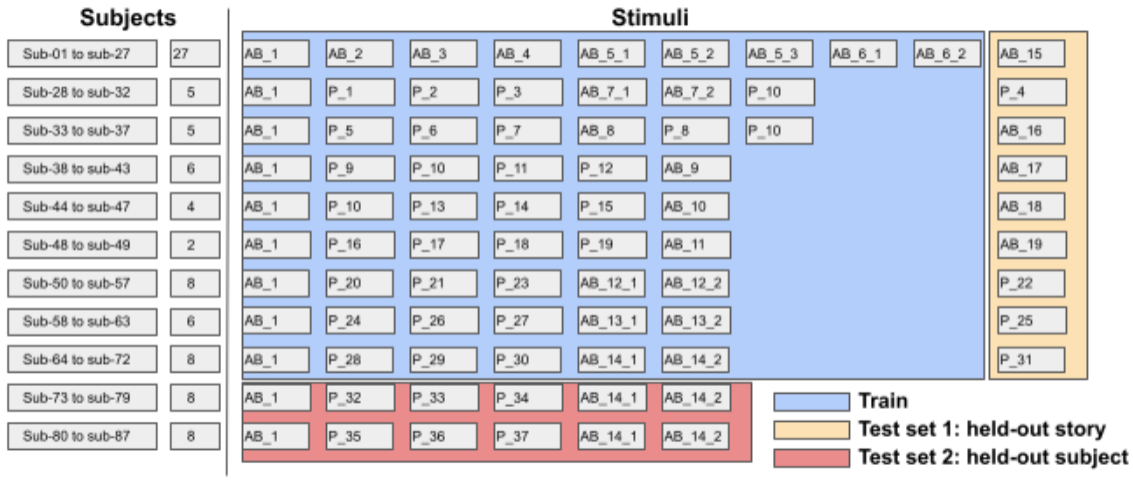


Рис. 3: Представление данных

В качестве метрики для сравнения моделей можно посчитать среднюю корреляцию для Test set 1 и 2, и просуммировать их с весами 2/3 и 1/3 соответственно.

Measure EEG data in a well-controlled lab environment (soundproof and electromagnetically shielded booth), using a high-quality 64- channel Biosemi ActiveTwo EEG recording system with 64 active Ag-AgCl electrodes and two extra electrodes, which serve as the common electrode (CMS) and current return path (DRL). The data is measured at a sampling rate of 8192 Hz. While the temporal resolution is high, the spatial resolution is low, with only 64 electrodes for billions of neurons. All 64 electrodes are placed according to international 10-20 standards.

The dataset contains data from 85 young, normal-hearing subjects (all hearing thresholds <= 25 dB Hl), with Dutch as their native language. Subjects indicating any

neurological or hearing-related medical history were excluded from the study. The study was approved by the Medical Ethics Committee UZ KU Leuven/Research (KU Leuven, Belgium). All identifiable subject information has been removed from the dataset.

Each subject listened to between 8 and 10 trials, each of approximately 15 minutes in length. The order of the trials is randomized between participants. All the stimuli are single-speaker stories spoken in Flemish (Belgian Dutch) by a native Flemish speaker. We vary the stimuli between subjects to have a wide range of unique speech material. The stimuli are either podcast or audiobooks. Some audiobooks are longer than 15 minutes. In this case, they are split into two trials presented consecutively to the subject.

# 3 Обзор литературы

По ЭЭГ получить envelope (в хронологическом порядке):

1. Линейные модели, FCCN (2022) - https://www.researchgate.net/publication/361380348$_{R}obust_{d}ec$

2. VLAAI (2022) - https://www.biorxiv.org/content/10.1101/2022.09.28.509945v2.full

3. Pre-LN FFT (2023) - https://arxiv.org/pdf/2305.06806.pdf

EEG2VEC - https://arxiv.org/pdf/2305.13957.pdf (Свежая статья, посмотреть внимательно. Идейно похоже на wav2vec. Участвовала в соревновании, откуда был взять датасет, там заняли 4 место. Утверждается, что потом смогли улучшить результат и побили 1 место. Кода нет. Кажется, не используют информацию о том, кто слушает, что дале модели из статьи 3. хороший буст метрик. Возможно, можно попробовать объеденить эти решения)

# 4 Вычислительный эксперимент

Сравниваются подходы, основанные на задаче Text to speech. В базовом решении используется модель VLAAI

## 4.1 Базовое решение

В качестве первого базового решения берется линейная модель. Линейная модель восстанавливает речевую огибающую из ЭЭГ, используя линейное преобразование по всем размерностям и времени. Обучаются модели, зависящие от субъекта, то есть у каждого субъекта есть своя модель.

В качестве второго базового решения берется модель Very Large Augmented Auditory Inference (VLAAI). Сеть VLAAI состоит из нескольких (N) блоков, состоящих из трех различных частей. Первая часть - набор сверток (сверточная нейронная сеть CNN). Данная сверточная сеть состоит из M=4 сверточных слоев. Вторая часть - полносвязанный слой размером 64, который пересобирает выходные фильтры стека CNN. Последняя часть - слой контекста. При применении к тренировочному и тестовому наборам вызова получается средний коэффициент корреляции 0,19.

## 4.2 Предлагаемые модели регрессии

Предлагается использовать бейзлайн, основанный на Fastspeech, Tacotron и других современных решениях Text-to-speech задачи

| **Model** | **Pearson correlation** | **Performance improvement (%)** |
|---|---|---|
| Baseline (VLAAI) | 0.1614 | - |
| Proposed | 0.2029 | 25.7 |

Таблица 1: Comparison of Pearson correlation and Performance improvement (%) between Baseline (VLAAI) and Proposed models.