Декодирование мозговых сигналов в аудиоданные

Набиев Мухаммадшариф Фуркатович

Московский физико-технический институт

Курс: Моя первая научная статья
(практика, В. В. Стрижов)

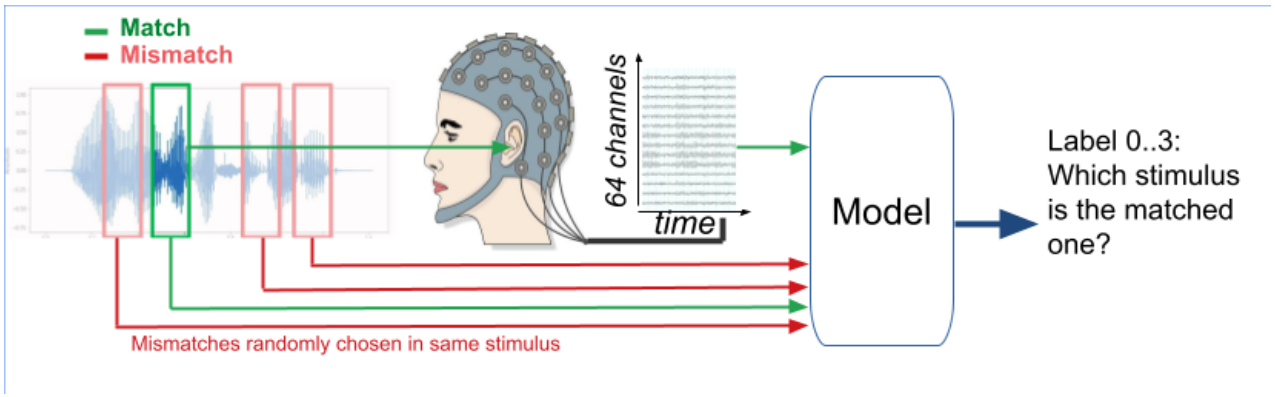
Эксперт: аспирант П. А. Северилов

2024

Цель исследования

Цель: Исследовать влияние физико-информированных энкодеров на качество декодирвование мозговых сигналов в аудиоданные.

Задача: Решить задачу декодирования в постановке классификации, а именно определить, какой сегмент аудио вызвал конкретную мозговую активность.



Постановка задачи

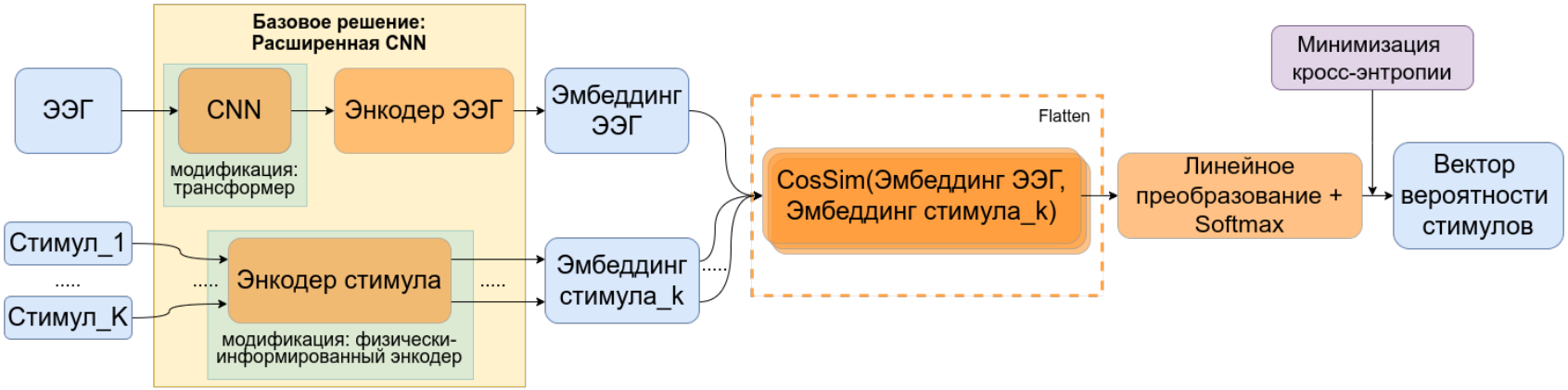
Данные: Кортеж $(\mathbf{X}^i, \mathbf{s}_1^i, \ldots, \mathbf{s}_K^i)$, где $\mathbf{X}^i \in \mathbb{R}^{64\times T}$ — ЭЭГ-сигнал с 64 каналами, $\mathbf{s}_1^i, \ldots, \mathbf{s}_K^i \in \mathbb{R}^{1\times T}$— стимулы, а K — количество стимулов. Меткой данного объекта будет являться вектор $\mathbf{y}^i \in \{0,1\}^K$. Только один стимул является истинным.

Требуется по имеющимся $\mathbf{X}^i, \mathbf{s}_1^i, \ldots, \mathbf{s}_K^i$ получить распределение вероятностей стимулов $\mathbf{p}^i = [p_1^i, \ldots, p_K^i]^T$. Пусть модель представляет собой следующее отображение $\mathbf{F} : \mathbb{R}^{64\times T} \times \left(\mathbb{R}^{1\times T}\right)^K \to \{0,1\}^K$. Задача сводится к минимизации кросс-энтропии:

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

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

Архитектура решения



Базовое решение:
Расширенная CNN — энкодер, который переводит ЭЭГ и стимулы в латентные пространства, где считается их близость (см. [3]).

Предлагаемые улучшения:
Для ЭЭГ заменить CNN на трансформер и использовать физико-информированный энкодер для стимула (см. [2], [4]).

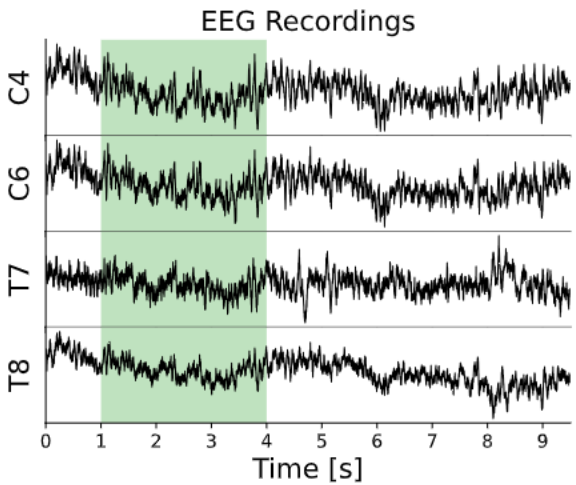
Данные для эксперимента

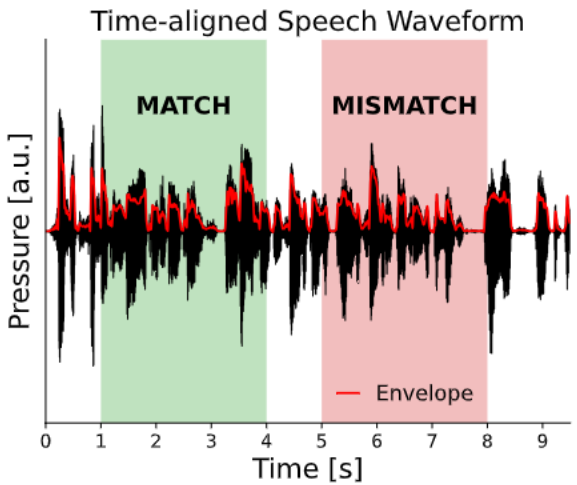
Эксперимент будет проверяться на данных SparrKULee (см. [1]).

- **Участники:** 85 участников.
- **Стимулы:** 6-10 аудиофрагментов разной категории, такие как аудиокниги и подкасты, каждый длительностью ≈ 15 минут.

После обработки, частота дискретизации всех данных была понижена до 64 Гц.

Подготовка данных





Пусть стимул обозначает сегмент аудиофрагмента. ЭЭГ и соответствующий аудиофрагмент делятся на сегменты фиксированной длины и для каждой пары (ЭЭГ, стимул) генерируются ложные стимулы.

Вычислительный эксперимент

	participant_id	age	sex	native_language	handedness	extra_comments
0	sub-001	21 to 23	F	Dutch, Flemish	right	NaN
4	sub-005	18 to 20	F	Dutch, Flemish	right	NaN
10	sub-011	21 to 23	F	Dutch, Flemish	right	NaN
12	sub-013	21 to 23	F	Dutch, Flemish	right	NaN
13	sub-014	21 to 23	M	Dutch, Flemish	right	NaN
16	sub-017	18 to 20	M	Dutch, Flemish	left	NaN
20	sub-021	18 to 20	M	Dutch, Flemish	right	NaN
21	sub-022	21 to 23	F	Dutch, Flemish	right	NaN
23	sub-024	18 to 20	M	Dutch, Flemish	right	NaN
28	sub-029	18 to 20	M	Dutch, Flemish	ambidexter	NaN
29	sub-030	21 to 23	M	Dutch, Flemish	right	NaN
32	sub-033	21 to 23	M	Dutch, Flemish	right	NaN
35	sub-036	24 to 26	F	Dutch, Flemish	right	NaN
41	sub-042	18 to 20	F	Dutch, Flemish	right	NaN
61	sub-062	21 to 23	M	Dutch, Flemish	right	NaN
63	sub-064	21 to 23	F	Dutch, Flemish	right	NaN
71	sub-072	21 to 23	F	Dutch, Flemish	right	NaN
72	sub-073	21 to 23	M	Dutch, Flemish	right	NaN
74	sub-075	24 to 26	M	Dutch, Flemish	right	NaN
75	sub-076	24 to 26	F	Dutch, Flemish	left	NaN
77	sub-078	18 to 20	M	Dutch, Flemish	right	NaN
84	sub-085	21 to 23	F	Dutch, Flemish	right	NaN

Рис.: Выборка, которая использовалась для эксперимента

Эксперимент проводился на подвыборке данных. Были отобраны 22 участника и аудиофрагменты, которые они слушали, а также их записи ЭЭГ.

Для эксперимента были взяты следующие параметры:

- Размер окна - 5 секунд
- Шаг окна - 1 секунда
- Количество ложных стимулов - 4

После разбиения по окнам и генерации ложных стимулов получилось 612500 кортежей.

Результаты эксперимента

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

$$Score = \frac{1}{22} \sum_{i=1}^{22} \frac{1}{l_i} \sum_{j=1}^{l_i} \left[y_j^i = pred_j^i \right],$$

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

Model	Score (%)
Baseline	99.08 ± 0.27
Transformer Encoder	**99.95** ± 0.04
Wav2Vec2	99.43 ± 0.39
Whisper-small	83.31 ± 4.37
Transformer Encoder + Wav2Vec2	99.78 ± 0.16
Transformer Encoder + Whisper-small	95.44 ± 2.50

Источники

[1] 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.

[2] 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.

[3] Bernd Accou et al. Modeling the relationship between acoustic stimulus and eeg with a dilated convolutional neural network. *2020 28th European Signal Processing Conference (EUSIPCO)*.

[4] 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.