SIGNAL PROCESSING ALGORITHMS FOR ATTENTION DECODING OF BRAIN RESPONSES TO NATURAL STIMULI IN BRAIN-COMPUTER INTERFACES

A Data Management Plan created using DMPonline.be

Creators: Simon Geirnaert https://orcid.org/0000-0002-4120-4232, Alexander Bertrand https://orcid.org/0000-0002-4827-8568, n.n. n.n.

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: 1242524N

ID: 204448

Start date: 01-10-2023

End date: 30-09-2026

Project abstract:

Brain-computer interfaces (BCI) enable the human brain to interact with machines, opening doors to various high-impact applications. However, most experimental BCI paradigms require the user to concentrate on synthetic and repeated stimuli, inducing fatigue and interfering with natural behavior. This unnatural interaction blocks the widespread usage of BCIs in daily-life situations beyond a few niche clinical applications.

In this project, we envisage ‘passive’ electroencephalography (EEG)-based BCI applications that track the user’s attention to natural audio-visual stimuli, allowing seamless integration with daily-life activities. However, this shift comes with several fundamental signal processing challenges, such as (1) the low signal-to-noise ratio of neural responses to natural speech or video footage, (2) the strong user-specificity of these responses, and (3) the multi-modal integration of audio-visual stimuli. We will tackle these challenges by designing novel algorithms that are inherently unsupervised (avoiding the need for a dedicated training session for each end-user), and exploit side information such as knowledge of the stimuli and data from other users.

We target generic algorithmic tools for EEG-based BCIs with natural stimuli but foresee specific breakthroughs in the context of (1) neuro-steered hearing devices, (2) educational neuroscience, and (3) objective hearing screening in daily-life environments, which act as driver applications.

Last modified: 01-02-2024

SIGNAL PROCESSING ALGORITHMS FOR ATTENTION DECODING OF BRAIN RESPONSES TO NATURAL STIMULI IN BRAIN-COMPUTER INTERFACES

FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Dataset Ia	EEG recordings of 10 subjects listening to audio books, while their attention is modulated over time. Per subject, 71 min of EEG data + audio are available.	Reuse existing data	Digital	Observational (EEG data)	.BDF, .BIDS, .mat, .wav, binary	< 100GB	
Dataset Ib	EEG recordings of 20 subjects watching video lectures, while their attention is modulated over time. Per subject, 60 min of EEG data + video/audio are available.	Reuse existing data	Digital	Observational (EEG data)	.BDF, .BIDS, .mat, .wav	< 1TB	
Dataset II	EEG recordings of 85 subjects listening to speech. Per subject, 90 to 150 min of EEG data + audio are available.	Reuse existing data	Digital	Observational (EEG data)	.BDF, .mat, .wav	< 1TB	

Dataset III	EEG recordings of 16 subjects listening to one out of two competing speech signals. Per subject, 72 min of EEG data + audio are available.	Reuse existing data	Digital	Observational (EEG data)	.mat, .wav	< 100GB	
Dataset IV	EEG recordings of 18 subjects listening to one out of two competing speech signals. Per subject, 138 min of EEG data + audio are available.	Reuse existing data	Digital	Observational (EEG data)	.mat, .wav	< 100GB	
Dataset V	EEG recordings of 28 subjects watching videos. Per subject, 63 to 87 min of EEG data + video are available.	Reuse existing data	Digital	Observational (EEG data)	.mat, .mp4, .fdt, .set	< 100GB	
Software	Various MATLAB toolboxes and experiment files connected to the publications generated by the project (and published)	New data	Digital	Software	.m	< 100 GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

- Dataset Ia: only internally available, stored on the servers of the research unit. Dataset described in *Roebben, A., Heintz, N., Geirnaert, S., Francart, T., & Bertrand, A. (2023). 'Are you even listening?'-EEG-based detection of absolute auditory attention to natural speech with application to neuro-steered hearing devices. bioRxiv, 2023-12.*
- Dataset Ib: currently only internally available on the servers of the research unit and not yet published.
- Dataset II: doi:10.48804/K3VSND, available on https://rdr.kuleuven.be/dataset.xhtml?persistentId=doi:10.48804/K3VSND. Published in *Accou, B., Bollens, L., Gillis, M., Verheijen, W., Van hamme, H., & Francart, T. (2023). SparrKULee: A Speech-evoked Auditory Response Repository of the KU Leuven, containing EEG of 85 participants. bioRxiv, 2023-07.*
- Dataset III: doi: 10.5281/zenodo.3997352, available on https://zenodo.org/records/3997352#.X0fP1sgza. Published in *N. Das, W. Biesmans, A. Bertrand, T. Francart, "The effect of head-related filtering and ear-specific decoding bias on auditory attention detection", Journal of Neural Engineering, vol. 13, 056014,2016.*
- Dataset IV: only internally available, stored on the servers of the research unit. Dataset described in *N. Das et al., "EEG-based auditory attention detection: boundary conditions for background noise and speaker positions," J. Neural Eng., vol. 15, no. 6, 2018, 066017.*
- Dataset V: doi: 10.5281/zenodo.10512414, available on https://zenodo.org/records/10512414. Published in *Yao, Y., Stebner, A., Tuytelaars, T., Geirnaert, S., & Bertrand, A. (2023). Identifying Temporal Correlations Between Natural Single-shot Videos and EEG Signals. J. Neural Eng., 2024.*

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, human subject data

There is no data collection involved in this project, as we have access to existing public and in-house EEG datasets. We thus only re-use existing data, with the informed consent of the participants. Depending on the informed consents, these dataset are only internally used or published. These EEG datasets are pseudo-anonymized using unique subject identifiers (which are stored in separate locations) and stored on secured (encrypted) and password-protected servers of ESAT-STADIUS (Department of Electrical Engineering, KU Leuven) and ExpORL (Department of Neurosciences, KU Leuven).
Ethical approval reference numbers for the own re-used datasets that are not publicly available (no doi) are provided below.

Dataset Ia	G-2023-7126-R2(MAR)
Dataset Ib	G-2023-7126-R2(MAR)
Dataset IV	G-2016-09-633

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

There is no data collection involved in this project, as we have access to existing public and in-house EEG datasets. We thus only re-use existing data, with the informed consent of the participants. These EEG datasets are already pseudo-anonymized using unique subject identifiers (which are stored in separate locations). No explicit personal data regarding participants (name, date of birth, ...) is known at the time or re-usage, on the EEG data is available. We follow KU Leuven's GDPR code for using and processing personal data.
Ethical approval reference numbers for the own re-used datasets that are not publicly available (no doi) are provided below.

Dataset Ia	G-2023-7126-R2(MAR)
Dataset Ib	G-2023-7126-R2(MAR)
Dataset IV	G-2016-09-633

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- Yes

The main newly generated output of the project consists of algorithmic methods and software, which could be of potential interest for commercialization and tech transfer. By default, all algorithmic software will be made publicly available, unless an obvious path for licensing or patenting has been identified, in which case we will collaborate with Leuven Research & Development.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

Dataset IV can only be used by authorized researchers within the research unit.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- Yes

Only when an opportunity for licensing or patenting for software is identified, intellectual property rights will be considered, in collaboration with Leuven Research & Development. However, by default, no legal issues apply.

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable,

for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

We will only re-use existing EEG datasets, for which documentation (README-files, documentation of data loading and preprocessing scripts, protocol descriptions, ...) is already available. For the published algorithmic software and experimental software, a README-file will be written, describing how the toolbox or experimental software can be used. Additionally, per written MATLAB-function or -script (.m), a standardized code comment snippet is written.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

A tailored README-file is written per toolbox/software package, which will be provided together with the software. Functions in the software are described following the MATLAB-standard:

- function-name
- short description
- inputs (type + description)
- outputs (type + description)

3. Data storage & back-up during the research project

Where will the data be stored?

Large Volume Storage: The in-house EEG data used in this project is stored in the research unit's secured (encrypted) and password-protected storage facilities of ESAT-STADIUS.The publicly available datasets are stored on research data platforms such as Zenodo.org, rdr.kuleuven.be (see before).
During development, software is stored on an encrypted personal drive, and version control software (e.g., Github) is used as a backup. After publication, the software is made available on GitHub.

How will the data be backed up?

Internal EEG data is backed up daily on the database server of the research division ESAT-STADIUS. The backup process is managed by the IT division of the department.
Software is backed up daily using version control systems.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The data stored in the ESAT servers has access regulated by an access control list (ACL) that grants read-write access to the project owner, and read-only access to specific users. The ACL is managed by the project owner and the STADIUS data manager. Client computers can access the data using SMB2 (or higher) from specific IP ranges, NFSv4 from specific (IT-managed) systems.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

The (maintenance) costs are covered by the research divisions of the PIs.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All data will be preserved for 10 years according to KU Leuven RDM policy.

Where will these data be archived (stored and curated for the long-term)?

After a period of 10 years if the data generated during the project are no longer in use locally, they will be removed from the servers. Note that some EEG datasets or software data remain publicly available via public online repositories (see above).

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

The (maintenance) costs are covered by the research divisions of the PIs.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Other, please specify:

See above: no new EEG data are collected in this project, only existing datasets are re-used. These datasets are available in Open-Access repositories, or closed-access (only available in-house), as specified above.
Software data will be, by default, made publicly available in Open Access via GitHub.

If access is restricted, please specify who will be able to access the data and under what conditions.

For the in-house datasets stored on the servers of the research division, only researchers from the research division who are involved in the corresponding projects have access to the EEG data.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Privacy aspects

The re-used in-house-only EEG data cannot be shared beyond project collaborators because these are (pseudo-anonymized) personal data (special category data: health data), and this has been agreed to by the participants in the Informed Consent Form.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

All software will be made available on GitHub.

When will the data be made available?

Upon publication of the research results.

Which data usage licenses are you going to provide? If none, please explain why.

Software code will be published under the MIT License, allowing freedom for reuse, modification, and distribution of the software.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

For the re-used datasets, see PIDs (DOIs) above.

What are the expected costs for data sharing? How will these costs be covered?

No specific costs for data sharing.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

The researcher working on the project.

Who will manage data storage and backup during the research project?

The data storage of the in-house EEG datasets and backup is the responsibility of the IT support of ESAT(-STADIUS). Software backups are managed by the researcher.

Who will manage data preservation and sharing?

The PIs (long-term), and the researcher working on the project.

Who will update and implement this DMP?

The researcher, Simon Geirnaert, The end responsibility is borne by the PIs, Alexander Bertrand and Tom Francart.