DIAGNOSTICS OF THE AUDITORY SYSTEM USING DEEP-LEARNING-BASED ANALYSIS OF EEG SIGNALS

A Data Management Plan created using DMPonline.be

Creators: Corentin Puffay, n.n. n.n., Hugo Van hamme, Jonas Vanthornhout

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: n.n. n.n., Hugo Van hamme

ID: 199143

Start date: 11-01-2022

End date: 11-01-2026

Project abstract:

When a person listens to sound, various parts of the auditory system are activated, including the brain. We can then measure the brain waves using EEG, decode them and draw conclusions about the auditory system. To that end, we aim to develop a computational model of the auditory system, based on state-of-the-art, deep-neural-network-based systems for automatic speech recognition. The model will be constructed by letting people listen to natural speech signals, and relating the recorded electroencephalogram (EEG) signal to the corresponding acoustic signals. One of the outcomes of this project is brain wave decoders that can be used in novel neuro-steered hearing aids and potentially as a brain-computer interface. Furthermore, by modeling the auditory system, we can pinpoint the origin of hearing disorders which gives invaluable information for the diagnostics of the auditory system. This has applications in clinical diagnostics of hearing and the design of smart hearing aids.

Last modified: 28-04-2023

DIAGNOSTICS OF THE AUDITORY SYSTEM USING DEEP-LEARNING-BASED ANALYSIS OF EEG SIGNALS
Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

Will you reuse existing data, generate new data, or do both ?

- Generate new data
- Reuse existing data

Will you be working with personal data (= any information relating to an identified/identifiable natural person)?

- Yes. From the subjects in general terms, some basic information is gathered (name, age, gender, mother tongue, cochlear implant or not,...) though, strictly limited to the absolute necessary data. Furthermore, the collected data are pseudonymised. The study is being approved by the Ethic Committee "Ethische Commissie Onderzoek UZ/KU Leuven" (S57102).

Type of data	Format	Volume	How created
Personal data: Basic information is gathered (name, age, gender, mother tongue, ...) strictly limited to the absolute necessary data.	Json files PDF	1 GB	Participants are asked to fill in a questionnaire upon arrival.
Audiograms: tonal	paper version -> scanned as pdf	1 GB	Participants undergo a tonal screening before the experiment.
Audiograms: speech	apx and apr files	1 GB	Participants undergo a speech screening before the experiment.
Electroencephalogram (EEG) recordings.	.apr, .apx, .bdf	15 GB per participants, ~200 participants in total.	EEG measurements are performed in a lab environment, in a soundproof cabin. Prerecorded sound files (see next row) are presented to the participants.
Sound stimuli presented to participants during EEG measurements.	.xml, .wav	10 GB	The sound files are pre-recorded and presented to the participants. During the experiments, trigger files are recorded in order to synchronize the sound files with the EEG recording.
Processed data	.npy, .mat, .tfrecords	1000 GB	Processed data of the EEG recordings, ready to be used for creating computational models.
Computational models developed on data	.py, tensorflow files	1000 GB	Models are created in python with the use of tensorflow.

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

1. **Designation of responsible person (If already designated, please fill in his/her name.)**
    1. Prof. Tom Francart, the PI of our research group through highly restricted overwrite and delete authorization. It is also the PI's decision which data are shared with other co-workers and how this will be arranged.
2. **Storage capacity/repository**
    - during the research: The data are stored on KU Leuven servers, and these drives are expandable in blocks, the backup capacity can be extended as much as needed.
    - after the research: idem.

What's the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

In accordance with the applicable law, the data of this study will be retained for 20 years (KB 30 juni 2004). These terms (and conditions) are also mentioned in the Informed Consent that will be signed by the subject, prior to their participation. We thus don't deviate from the minimum preservation term of 5 years.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

We work with the personal data of humans. All data, except for the informed consent, are pseudonymised. The informed consent is stored separately from the other data and is only accessible by the PI.

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

Currently, I don't see other issues.

DIAGNOSTICS OF THE AUDITORY SYSTEM USING DEEP-LEARNING-BASED ANALYSIS OF EEG SIGNALS
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable
- Not applicable

DIAGNOSTICS OF THE AUDITORY SYSTEM USING DEEP-LEARNING-BASED ANALYSIS OF EEG SIGNALS
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

DIAGNOSTICS OF THE AUDITORY SYSTEM USING DEEP-LEARNING-BASED ANALYSIS OF EEG SIGNALS
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	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
A large auditory EEG dataset (Bollens et al. 2023)	Reused	Digital	*Experimental data: EEG recordings, speech stimuli, audiograms*	.apr, .apx, .bdf .pdf .xml, .wav	15 GB for 85 subjects, and 10GB of common stimuli	
A large auditory EEG dataset (Bollens et al. 2023) - extension	Newly collected	Digital	Experimental data: *EEG recordings, speech stimuli, audiograms*	.apr, .apx, .bdf .pdf .xml, .wav	15GB for ~ 115 subjects, and 10GB of common stimuli	

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:

"A large auditory EEG decoding dataset" L. Bollens et al 2023
DOI: https://doi.org/10.48804/K3VSND

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

Yes. The project involves experiments on human subjects and therefore is being approved by the Ethic Committee "Ethische Commissie Onderzoek UZ/KU Leuven" (S57102).

Dataset: "A large auditory EEG decoding dataset"

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

Yes.
Short description of the kind of personal data that will be used: From the subjects in general terms, some basic information is gathered (name, age, gender, mother tongue, cochlear implant or not,...) though, strictly limited to the absolute necessary data. Furthermore, the collected data are pseudonymized. This study is being approved by the Ethic Committee "Ethische Commissie Onderzoek UZ/KU Leuven" (S57102).

Dataset: "A large auditory EEG decoding dataset"

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.

- No

Not at this moment. In case IP possibilities rise, we will contact LRD.

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.

- No

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.

- No

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).

New lab members get a training on how to perform the experiments how and where to store and handle the collected data. Furthermore, standard guidelines and instructions are available as .pdf and/or .readme, stored on network drive. They contain best practices regarding the practical side of your experiment (set-up, parameters,...) as well as policies about how to treat the subjects, how to handle and where to keep sensitive information, etc. In this way, the information given during the training can easily be re-read, refreshed.
This, in combination with the fact that the original collected data, are stored on backed-up drives, in a standard format, should make it possible to understand and reuse the data. The data will be stored based on the BIDS structure (see below).
The code corresponding to each result will be stored on a git repository. This way, reusers will be able to recreate the results. The commentary lines in the code also document how the data is organized and how a result can be obtained, for this we will use regular code and comments but also Jupyter notebooks. Moreover, the code also contains an implicit description of how the collected data is organized, it can be used as an explanation of the data.

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.

- Yes

The data will be stored based on the BIDS structure, an organization structure for neuroimaging and behavioral data (see also the website: bids.neuroimaging.io). This structure will be used to arrange the data in an uniform way.

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

Where will the data be stored?

The data will be stored on KU Leuven administered drives (large volume storage). In order to be able to analyse the data, some files will need (temporarily) to be stored on the encrypted PC hard drive (this since calculations from a non-local source are too slow and lead to computational failures). Once analysed, the raw data are again removed from the local hard drives

How will the data be backed up?

Since the data are stored on KU Leuven storage, the general ICT back-up Policy is applied.

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

Yes. Since the data are stored on KU Leuven servers, and these drives are expandable in blocks, the backup capacity is technically not an issue.

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

The lab policy is that the researchers have only access to the data from the project they are involved in. Furthermore, the data for longer term storage are kept on separate drives with a) limited access (only a limited set of people have access) and b) an overwrite and delete protection (based on read-write access) in order to prevent accidental loss of these data.
Prof. Francart is the only person who has access to the key information for identification of the subjects. His back-up (only to be used in extreme case) is Jan Wouters and Astrid van Wieringen (both PI's within the lab).

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

The estimated cost is around 7.000 Euro (157EUR/TB/year) and will be paid from the C1-C2 project under which this FWO proposal is falling.

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...).

As mentioned in our informed consent, medical data will be retained for at least 30 years and other research data for at least 20 years.

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

The data are stored on a KU Leuven, ICT managed Large Volume Storage, drives especially designed for archive storage. Personal data, will be kept in a similar, secure way. However, these data will be stored at a different location (different server) with other (restricted) access rights (see also above).

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

The estimated cost is around 11.000 Euro (14TB @ 157EUR/TB/year). Given the way projects are financed at KU Leuven, we cannot predict whether funding will be available after the current C2 project ends, unless we are allowed to set aside funding from the C2 project for later use.

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:

To be decided, this will depend on IP issues (if there are "No (closed access)", otherwise "Yes in a restricted access repository".
Dataset: "A large auditory EEG decoding dataset" (and extension)

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

The informed consent contains a section in which the participant can choose whether his/her data is shared. Possibilities are:

- no sharing allowed
- sharing within KU Leuven
- sharing within the European Union
- worldwide sharing
- worldwide sharing and the data can be made available on a public database

We will honour their choice. Furthermore, shared data will always be pseudonymised.

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
- Yes, Intellectual Property Rights

To be decided. Of course the key to identify the subject from which the data were obtained, is not sharable (without the consent of the subject).
It will also depend on IP issues.

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

Dataset: "A large auditory EEG decoding dataset": https://doi.org/10.48804/K3VSND
Dataset: "A large auditory EEG decoding dataset" (extension): not yet applicable

When will the data be made available?

Dataset: already available (https://doi.org/10.48804/K3VSND)
Dataset extension: upon publication of research results.

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

The already available dataset has CC-BY-NC-4.0, so the extension will probably have the same usage license.

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

Dataset: https://doi.org/10.48804/K3VSND
Dataset extension: not yet available

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

Dataset: free
Dataset extension: the research data repository (RDR) is free below 50G, it may not be free anymore as we extend it upon research publication.

6. Responsibilities

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

The researcher, when his contract has ended, the responsibility shifts towards Prof. Francart & Prof. Van hamme to ensure data preservation and reuse.

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

Back-up of the data is arranged by ICTS- KU Leuven (back-up policy dependent on the kind of storage).

Who will manage data preservation and sharing?

The PI through highly restricted overwrite and delete authorisation. It is also the PI's decision which data are shared with other co-workers and how this will be arranged.

Who will update and implement this DMP?

The PIs (Prof. Tom Francart and Prof. Hugo Van hamme) bear the end responsibility of updating & implementing this DMP.