The Temple University Artifact Corpus: An Annotated Corpus of EEG Artifacts

A. Hamid, K. Gagliano, S. Rahman, N. Tulin, V. Tchiong, I. Obeid and J. Picone

The Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA
{ahmad.hamid, tuk18614, tuh01696, tug47034, tug94380}@temple.edu

The Neural Engineering Data Consortium has recently developed a new subset of its popular open source electroencephalogram (EEG) corpus – TUH EEG (TUEG) [1]. The TUEG Corpus is the world's largest open source corpus of EEG data and currently has over 3,300 subscribers. There are several valuable subsets of this data, including the TUH Seizure Detection Corpus (TUSZ) [2], which was featured in the Neureka 2020 Epilepsy Challenge [3]. In this poster, we present a new subset of the TUEG Corpus – the TU Artifact Corpus. This corpus contains 310 EEG files in which every artifact has been annotated. This data can be used to evaluate artifact reduction technology. Since TUEG is comprised of actual clinical data, the set of artifacts appearing in the data is rich and challenging.

EEG artifacts are defined as waveforms that are not of cerebral origin and may be affected by numerous external and or physiological factors. These extraneous signals are often mistaken for seizures due to their morphological similarity in amplitude and frequency [4]. Artifacts often lead to raised false alarm rates in machine learning systems, which poses a major challenge for machine learning research. Most state-of-the-art systems use some forms of artifact reduction technology to suppress these events [5].

The corpus was annotated using a five-way classification that was developed to meet the needs of our constituents. Brief descriptions of each form of the artifact are provided in Ochal et al. [4]. The five basic tags are:

- **Chewing (CHEW):** An artifact resulting from the tensing and relaxing of the jaw muscles. Chewing is a subset of the muscle artifact class. Chewing has the same characteristic high frequency sharp waves with 0.5 sec baseline periods between bursts. This artifact is generally diffuse throughout the different regions of the brain. However, it might have a higher level of activity in one hemisphere. Classification of a muscle artifact as chewing often depends on whether the accompanying patient report mentions any chewing, since other muscle artifacts can appear superficially similar to chewing artifact.
- **Electrode (ELEC):** An electrode artifact encompasses various electrode related artifacts. Electrode pop is an artifact characterized by channels using the same electrode "spiking" with an electrographic phase reversal. Electrostatic is an artifact caused by movement or interference of electrodes and or the presence of dissimilar metals. A lead artifact is caused by the movement of electrodes from the patient's head and or poor connection of electrodes. This results in disorganized and high amplitude slow waves.
- **Eye Movement (EYEM):** A spike-like waveform created during patient eye movement. This artifact is usually found on all of the frontal polar electrodes with occasional echoing on the frontal electrodes.
- **Muscle (MUSC):** A common artifact with high frequency, sharp waves corresponding to patient movement. These waveforms tend to have a frequency above 30 Hz with no specific pattern, often occurring because of agitation in the patient.
- **Shiver (SHIV):** A specific and sustained sharp wave artifact that occurs when a patient shivers, usually seen on all or most channels. Shivering is a relatively rare subset of the muscle artifact class.

Since these artifacts can overlap in time, a concatenated label format was implemented as a compromise between the limitations of our annotation tool and the complexity needed in an annotation data structure used to represent these overlapping events. We distribute an XML format that easily handles overlapping events. Our annotation tool [6], like most annotation tools of this type, is limited to displaying and

manipulating a flat or linear annotation. Therefore, we encode overlapping events as a series of concatenated names using symbols such as:

- **EYEM+CHEW:** eye movement and chewing
- **EYEM+SHIV:** eye movement and shivering
- **CHEW+SHIV:** chewing and shivering

An example of an overlapping annotation is shown below in Figure 1.

The files are annotated by students who have been trained in artifact recognition and have conducted an inter-rater agreement in order to ensure uniformity between annotations.

This release is an update of TUAR v1.0.0, which was a partially annotated database. In v1.0.0, a similar five-way system was used as well as an additional "null" tag. The "null" tag covers anything that was not annotated, including instances of artifact. Only a limited number of artifacts were annotated in v1.0.0. In this updated version, every instance of an artifact is annotated; ultimately, this provides the user with confidence that any part of the record that is not annotated with one of the five classes does not contain an artifact. No new files, patients, or sessions were added in v2.0.0. However, the data was reannotated with these standards. The total number of files remains the same, but the number of artifact events increases significantly. Complete statistics will be provided on the corpus once annotation is complete and the data is released. This is expected to occur in early July – just after the IEEE SPMB submission deadline.

The TUAR Corpus is an open-source database that is currently available for use by any registered member of our consortium. To register and receive access, please follow the instructions provided at this web page: *https://www.isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml*. The data is located here: *https://www.isip.piconepress.com/projects/tuh_eeg/downloads/tuh_eeg_artifact/v2.0.0/*.

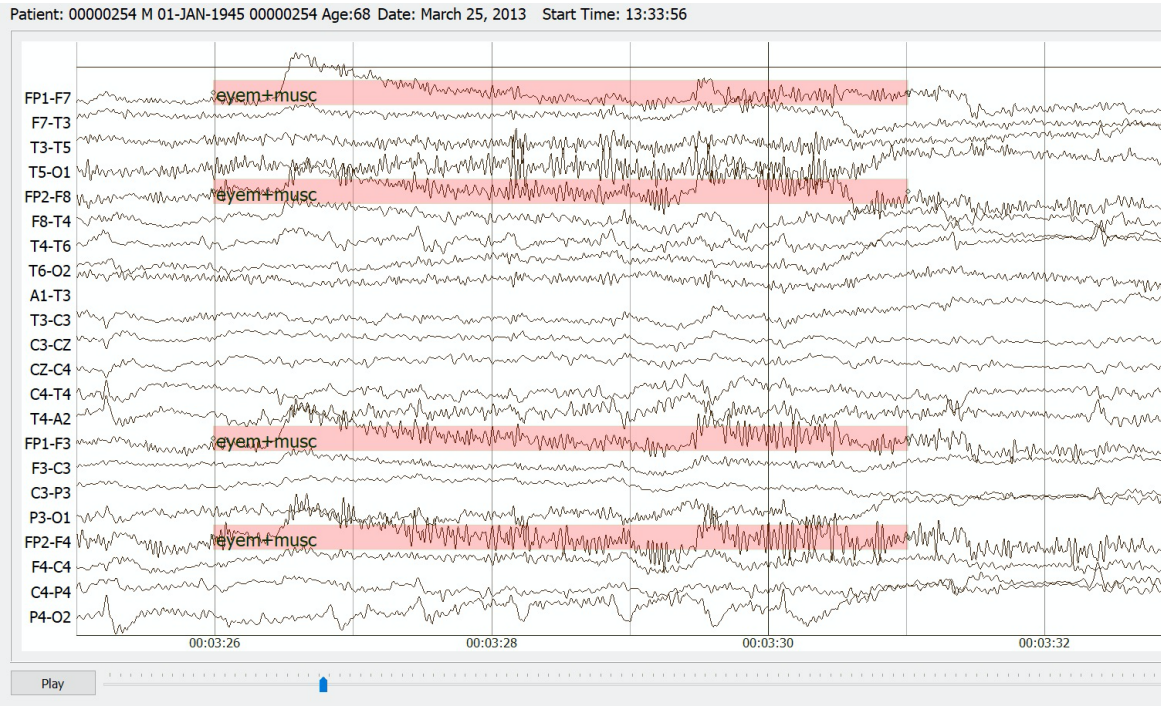


Figure 1. An annotated file depicting an overlapping annotation with eye movement (EYEM) and muscle (MUSC) artifacts

ACKNOWLEDGMENTS

Research reported in this publication was most recently supported by the National Science Foundation Partnership for Innovation award number IIP-1827565 and the Pennsylvania Commonwealth Universal Research Enhancement Program (PA CURE). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the official views of any of these organizations.

REFERENCES

[1] I. Obeid and J. Picone, "The Temple University Hospital EEG Data Corpus," in *Augmentation of Brain Function: Facts, Fiction and Controversy. Volume I: Brain-Machine Interfaces*, 1st ed., vol. 10, M. A. Lebedev, Ed. Lausanne, Switzerland: Frontiers Media S.A., 2016, pp. 394-398. *https://doi.org/10.3389/fnins.2016.00196*.

[2] V. Shah et al., "The Temple University Hospital Seizure Detection Corpus," *Front. Neuroinform.*, vol. 12, pp. 1–6, 2018. *https://doi.org/10.3389/fninf.2018.00083*.

[3] Y. Roy, R. Iskander, and J. Picone, "The Neureka™ 2020 Epilepsy Challenge," NeuroTechX, 2020. [Online]. Available: *https://neureka-challenge.com*. [Accessed: 16-Apr-2020].

[4] Ochal, D., Rahman, S., Ferrell, S., Elseify, T., Obeid, I., & Picone, J. (2020). The Temple University Hospital EEG Corpus: Annotation Guidelines. Philadelphia, Pennsylvania, USA. *https://www.isip.piconepress.com/publications/reports/2020/tuh_eeg/annotations*.

[5] Bertrand, A., Mihajlović, V., Grundlehner, B., Van Hoof, C., & Moonen, M. (2013). Motion artifact reduction in EEG recordings using multi-channel contact impedance measurements. 2013 IEEE Biomedical Circuits and Systems Conference, BioCAS 2013, 258–261. *https://doi.org/10.1109/BioCAS.2013.6679688*

[6] N. Capp, E. Krome, I. Obeid, and J. Picone, "Facilitating the Annotation of Seizure Events Through an Extensible Visualization Tool," in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 2017, p. 1. *https://doi.org/10.1109/SPMB.2017.8257043.*

NEURAL ENGINEERING DATA CONSORTIUM
www.nedcdata.org

The Temple University Artifact Corpus: An Annotated Corpus of EEG Artifacts

A. Hamid, K. Gagliano, S. Rahman, N. Tulin, V. Tchiong, I. Obeid, and J. Picone

The Neural Engineering Data Consortium, Temple University

College of Engineering
Temple University

Abstract

A new subset of the popular open source electroencephalogram (EEG) corpus – TUH EEG:

- The Temple University Artifact Corpus (TUAR) consists of high yield artifact files annotated using a five-way classification system:
  - ❑ Chewing (CHEW): An artifact resulting from the tensing and relaxing of the jaw muscles.
  - ❑ Electrode (ELEC): An artifact that encompasses various electrode related phenomena.
  - ❑ Eye Movement (EYEM): A spike-like waveform created during patient eye movement.
  - ❑ Muscle (MUSC): A common artifact with high frequency, sharp waves corresponding to patient movement.
  - ❑ Shiver (SHIV): A specific and sustained sharp wave artifact that occurs when a patient shivers.
- EEG artifacts are waveforms that are not of cerebral origin and may have been affected by several external and physiological factors.
- These artifacts cause false alarms in seizure prediction machine learning systems.

This corpus was developed to support research and evaluation of artifact suppression technology.

TUAR v2.0 Statistics

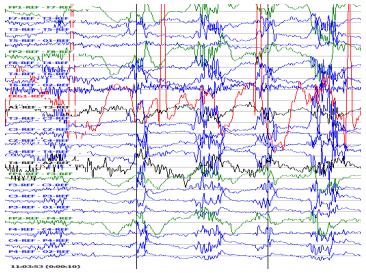
- v2.0 is a major upgrade over v1.0 because every artifact event was labeled:

Item	v1.0	v2.0
Patients	213	213
Sessions	259	259
Files	310	310
Events:		
CHEW	180	350
ELEC	464	2,499
EYEM	1,117	7,182
MUSC	374	5,644
SHIV	289	24
TOTAL	2,424	15,699
Total Signal Dur. (Hrs)	99.64	99.64
Total Events Dur. (Hrs)	5.18	42.17

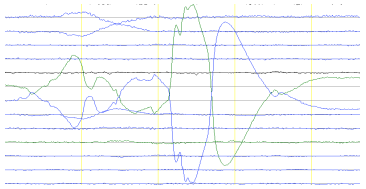
- TUAR files have an average duration of 20 minutes in comparison to 9 minute TUSZ files. Certain files can extend to greater than 3 hours.
- Specific Artifact Morphology:
  - ❑ CHEW is a subset of the muscle artifact class that has the same characteristic high frequency sharp waves with 0.5 sec baseline periods between bursts.
  - ❑ ELEC is an artifact that includes various electrode related artifacts such as pop, electrostatic discharge, lead movement and poor conductivity.
  - ❑ EYEM is created during patient eye movement and is usually found on all the frontal polar electrodes.
  - ❑ MUSC often occurs because of agitation in the patient and tends to have energy above 30 Hz.
  - ❑ SHIV occurs when a patient shivers, usually seen on all or most channels.

Five-Way Classification

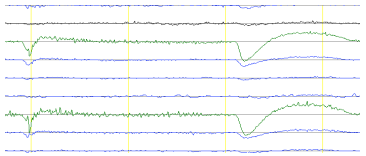
- Chewing (CHEW): results from the tensing and relaxing of the jaw muscles.



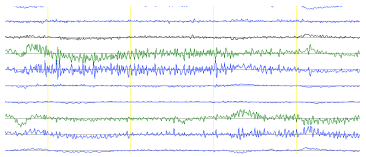
- Electrode (ELEC): encompasses electrode related phenomena such s poor conductivity and electrostatic discharges.



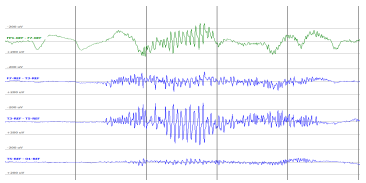
- Eye Movement (EYEM): caused by patient eye movement.



- Muscle (MUSC): high frequency waves caused by patient movement during an EEG recording.



- Shiver (SHIV): Sharp wave seen as the patient shivers during an EEG recording.



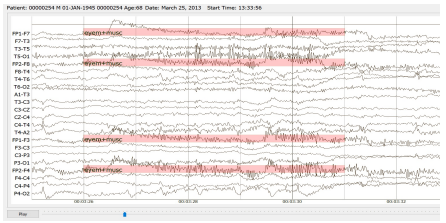
The Annotation Process

- Annotated by a team of five student workers who have been trained to annotate seizures, artifacts, slowing and other common phenomena.
- Inter-rater agreement for this team is very high (κ > 0.8) and continuously monitored to ensure uniformity of the annotations.
- Each file is annotated by a minimum of two annotators. Difficult cases are reviewed by the group. Weekly review sessions are conducted.
- Patient reports are used as references in order to develop a common interpretation between student annotators as well as neurologists.

Annotation Tool

- Our open-source visualization tool, written in PyQt, supports visualization and interpretation of EEGs.
- Annotations are made directly onto this visualization tool using a mouse click and drag function. Associated tags are chosen from drop-down menus.

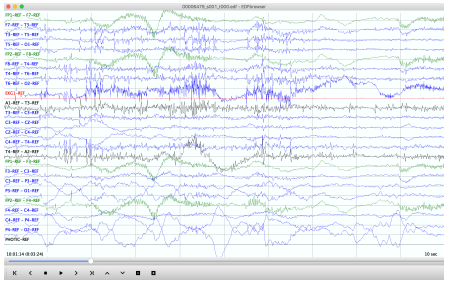
- Annotations are overlaid on the waveform and can be interactively edited.



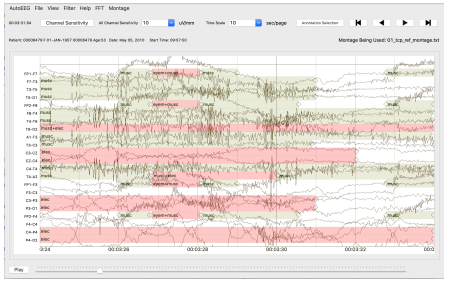
- The annotations consist of artifacts that overlap in time which require overlapping annotations.
- Such annotations are completed using a concatenated labeling format (e.g., EYEM+ELEC is used to label a region where both artifacts occur).
- This compensates for the limitations of the annotation tool, as well as the complexity needed in annotation data structure to represent such overlapping events.
- A new format based on XML will soon be released that will handle this in a more elegant manner. Tags can overlap in time and be arranged in a hierarchy.

Annotation Challenges

- When overlapping events on several channels obscure our view, it takes significantly longer to annotate a file.



- Annotations that show alternating artifact events throughout the course of the file also take a significant amount of time to annotate.



- The average time spent annotating a file in TUAR v2.0 is over 30 mins compared to 15 mins for v1.0 and 10 mins for TUSZ.

Summary

- The TUAR v2.0.0 release is a fully annotated database using a five-way classification system.
- The TUAR v2.0.0 consists of 310 files, 259 sessions, 213 patients and 15699 artifact events.
- Overlapping events in time are annotated using a concatenated labeling system.
- The data compiled in TUAR can be used to evaluate artifact reduction technology.
- The data can be downloaded from *isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml* which includes instructions for anonymous rsync.

Acknowledgements

- Research reported in this publication was most recently supported by the National Science Foundation Partnership Program (PA CURE). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the official views of any of these organizations.