

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 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 form of artifact reduction technology to suppress these events.

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 [5], 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.

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/.

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

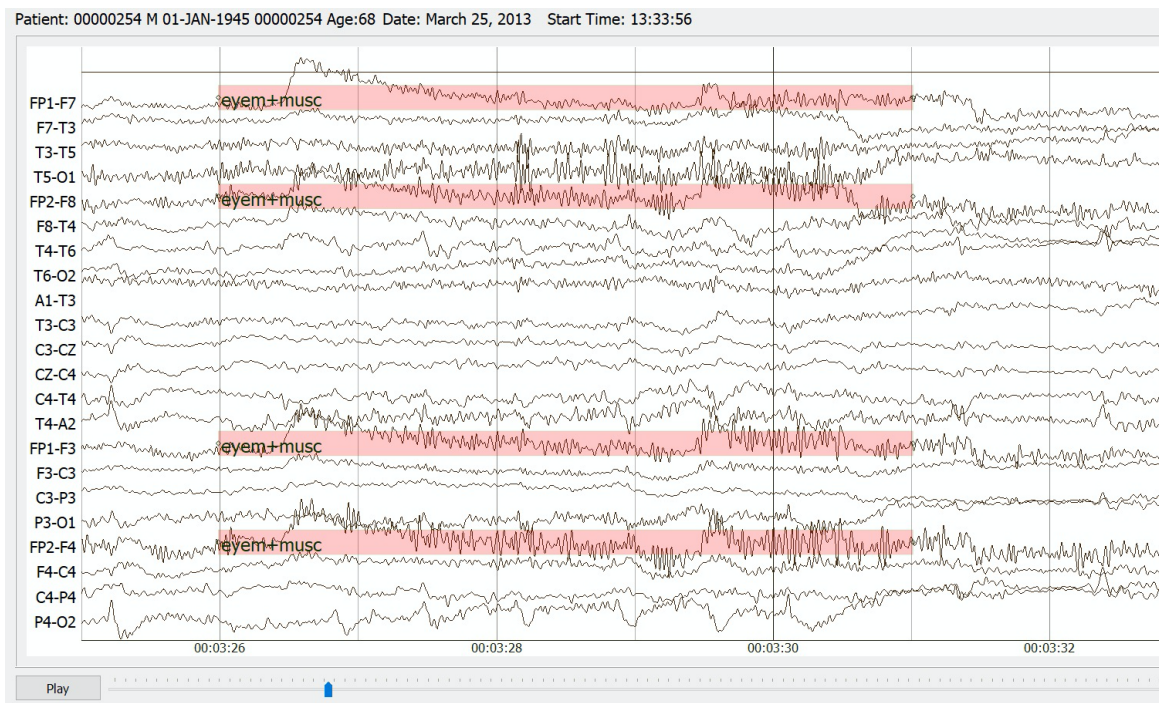


Figure 1. An annotated file depicting an overlapping annotation with eye movement (EYEM) and muscle (MUSC) artifacts

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] 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>.