Improving the Quality of the TUSZ Corpus

S. Rahman, A. Hamid, D. Ochal, I. Obeid and J. Picone

Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA
{tuh01696, ahmad.hamid, tug48969, iobeid, picone} @temple.edu

The Temple University Hospital Seizure Detection Corpus (TUSZ) [1] has been in distribution since April 2017. It is a subset of the TUH EEG Corpus (TUEG) [2] and the most frequently requested corpus from our 3,000+ subscribers. It was recently featured as the challenge task in the Neureka 2020 Epilepsy Challenge [3]. A summary of the development of the corpus is shown below in Table 1.

Releases	**Patients**	**Sessions**	**Files**	**Seizure Files**	**Total No. Seizure Events**	**Total Duration (Hours)**	**Seizure Duration (Hours)**
v1.0.0 – 04/17/2017	114	510	2,013	291	328	170	4.9
v1.1.0 – 08/04/2017	246	686	2,489	423	3,582	425	28.9
v1.2.0 – 04/15/2018	315	822	3,064	642	1,951	504	36.75
v1.3.0 – 08/16/2018	364	970	4,023	942	2,465	651	52.6
v1.4.0 – 11/14/2018	364	969	4,020	949	2,548	651	53.0
v1.5.0 – 07/22/2019	692	1,661	6,633	1,399	3,591	1,074	74.6
v1.5.1 – 04/23/2020	692	1,575	6,633	1.382	3,554	1,074	73.5
v1.5.2 – 05/09/2020	692	2,608	6,635	1,384	3,561	1,074	73.9
v1.6.0 – 08/31/2020	TBD	TBD	TBD	TBD	TBD	TBD	TBD

Table 1. A summary of the TUSZ release history

The TUSZ Corpus is a fully annotated corpus, which means every seizure event that occurs within its files has been annotated. The data is selected from TUEG using a screening process that identifies files most likely to contain seizures [1]. Approximately 7% of the TUEG data contains a seizure event, so it is important we triage TUEG for high yield data. One hour of EEG data requires approximately one hour of human labor to complete annotation using the pipeline described below, so it is important from a financial standpoint that we accurately triage data.

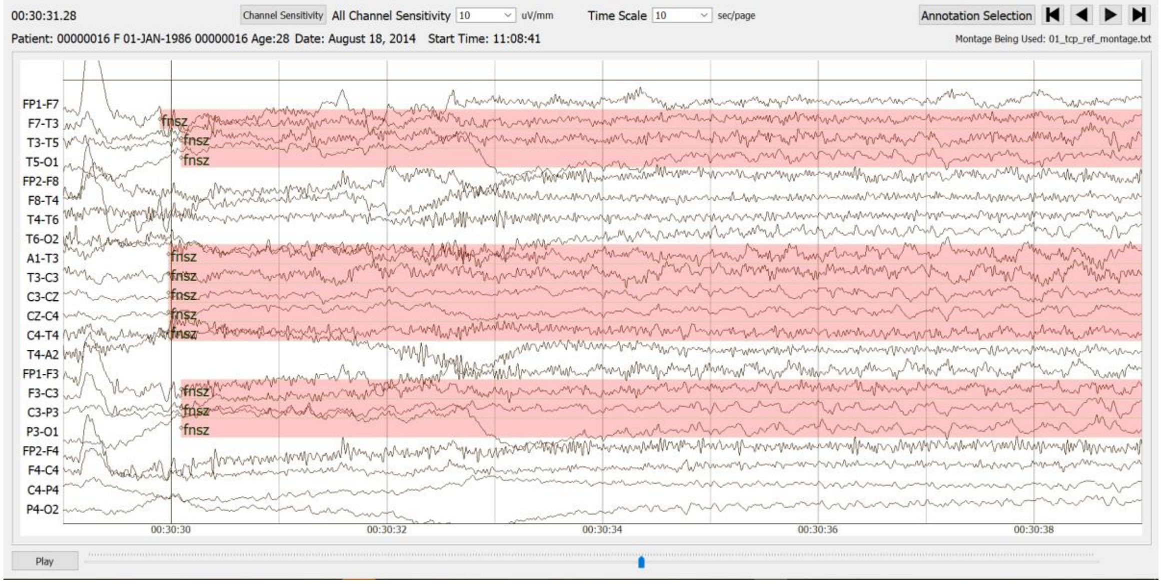
A summary of the labels being used to annotate the data is shown in Table 2. Certain standards are put into place to optimize the annotation process while not sacrificing consistency. Due to the nature of EEG recordings, some records start off with a segment of calibration. This portion of the EEG is instantly recognizable and transitions from what resembles lead artifact to a flat line on all the channels. For the sake of seizure annotation, the calibration is ignored, and no time is wasted on it. During the identification of seizure events, a hard “3 second rule” is used to determine whether two events should be combined into a single larger event. This greatly reduces the time that it takes to annotate a file with multiple events occurring in succession. In addition to the required minimum 3 second gap between seizures, part of our standard dictates that no seizure less than 3 seconds be annotated. Although there is no universally accepted definition for how long a seizure must be, we find that it is difficult to discern with confidence between burst suppression or other morphologically similar impressions when the event is only a couple seconds long. This is due to several reasons, the most notable being the lack of evolution which is oftentimes crucial for the determination of a seizure.

In the first portion of the process, the EEG files are triaged from our TUEG database by an overlapping method of machine learning seizure detection and key word analysis of the respective clinical reports. After the EEG files have been triaged, a team of annotators at NEDC is provided with the files to begin data

Table 2. The labels used to annotate our EEG data are shown.

Index	Label	Description
0	null	An undefined annotation. Should not be seen in the data.
1	spsw	Spike and/or slow wave. A short duration epileptiform event involving an electrographic spike in activity and/or a slow wave (low frequency wave). Usually no more than 1 sec. in duration.
2	gped	Generalized periodic epileptiform discharge. Periodic diffuse spike/sharp wave discharges across multiple regions or hemispheres.
3	pled	Periodic lateral epileptiform discharge. A regular, periodically occurring spike/sharp wave seen in a certain locality of the scalp.
4	eybl	Eyeblink. A specific, sharp, high amplitude eye movement artifact corresponding to blinks.
5	artf	Artifact. Any non-brain activity electrical signal, such as those due to equipment or environmental factors.
6	bckg	All other non-seizure cerebral signals.
7	seiz	Seizure. A basic annotation for seizures.
8	fnsz	Focal nonspecific seizure. A large category of seizures occurring in a specific focality.
9	gnsz	Generalized seizure. A large category of seizures occurring in most if not all of the brain.
10	spsz	Simple partial seizure. Brief seizures that start in one location of the brain (and may spread) where the patient is fully aware and able to interact.
11	cpsz	Complex partial seizure. Same as simple partial seizure but with impaired awareness.
12	absz	Absence seizure. Brief, sudden seizure involving lapse in attention. Usually lasts no more than 5 seconds and commonly seen in children.
13	tnsz	Tonic seizure. A seizure involving the stiffening of the muscles. Usually associated with and annotated as tonic-clonic seizures, but not always (rarely there is no clonic phase).
14	cnsz	Clonic seizure. A seizure involving sustained, rhythmic jerking. Not seen in our datasets, as it is always associated with tonic clonic seizures and is annotated as such.
15	tcsz	Tonic-clonic seizure. A seizure involving loss of consciousness and violent muscle contractions.
16	atsz	Atonic seizure. A seizure involving the loss of tone of muscles in the body. Also never seen as it is always associated with an occasionally occurring phase before a tonic clonic seizure.
17	mysz	Myoclonic seizure. A seizure associated with brief involuntary twitching or myoclonus.
18	nesz	Non-epileptic seizure. Any non-epileptic seizure observed. Contains no electrographic signs.
19	intr	Interesting patterns. Any unusual or interesting patterns observed that don't fit into the above classes.
20	slow	Slowing. A brief decrease in frequency.
21	eyem	Eye movement. A very common frontal/prefrontal artifact seen when the eyes move.
22	chew	Chewing. A specific artifact involving multiple channels that corresponds with patient chewing, “bursty”
23	shiv	Shivers. A specific, sustained sharp artifact that corresponds with patient shivering.
24	musc	Muscle artifact. A very common, high frequency, sharp artifact that corresponds with agitation/nervousness in a patient.
25	elpp	Electrode pop. A short artifact characterized by channels using the same electrode “spiking” with perfect symmetry.
26	elst	Electrostatic artifact. Artifact caused by movement or interference on the electrodes, variety of morphologies.
27	calb	Artifact caused by calibration of the electrodes. Appears as a flattening of the signal in the beginning of files.
28	hphs	A brief period of high amplitude slow waves.
29	trip	Large, three-phase waves frequently caused by an underlying metabolic condition.

annotation. An example of an annotation is shown in Figure 1. A summary of the workflow for our annotation process is shown in Figure 2. Several passes are performed over the data to ensure the annotations are accurate. Each file undergoes three passes to ensure that no seizures were missed or misidentified. This is different than the previous versions of the corpus in which there was only a single

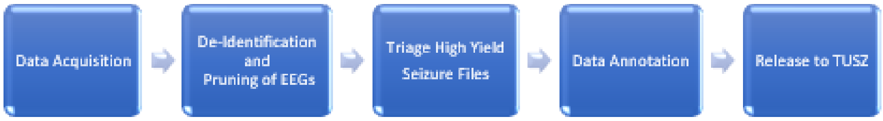


00:30:31.28 Channel Sensitivity All Channel Sensitivity 10 uV/mm Time Scale 10 sec/page Annotation Selection

Patient: 00000016 F 01-JAN-1986 00000016 Age:28 Date: August 18, 2014 Start Time: 11:08:41

Montage Being Used: 01_tcp_ref_montage.txt

Figure 1. An example of an annotated EEG signal



Data Acquisition → De-Identification and Pruning of EEGs → Triage High Yield Seizure Files → Data Annotation → Release to TUSZ

Figure 2. The data preparation pipeline

reviewer for all files. This new workflow results in a greater number of annotators viewing each file and consequently higher quality data. The first pass of TUSZ involves identifying which files contain seizures and annotating them using our annotation tool. The time it takes to fully annotate a file can vary drastically depending on the specific characteristics of each file; however, on average a file containing multiple seizures takes 7 minutes to fully annotate. This includes the time that it takes to read the patient report as well as traverse through the entire file.

Once an event has been identified, the start and stop time for the seizure is stored in our annotation tool. This is done on a channel by channel basis resulting in an accurate representation of the seizure spreading across different parts of the brain. Files that do not contain any seizures take approximately 3 minutes to complete. Even though there is no annotation being made, the file is still carefully examined to make sure that nothing was overlooked. In addition to solely scrolling through a file from start to finish, a file is often examined through different lenses. Depending on the situation, low pass filters are used, as well as increasing the amplitude of certain channels. These techniques are never used in isolation and are meant to further increase our confidence that nothing was missed. Once each file in a given set has been looked at once, the annotators start the review process. The reviewer checks a file and comments any changes that they recommend. This takes about 3 minutes per seizure containing file. After each file has been commented on, the third pass commences. This step takes about 5 minutes per seizure file and requires the reviewer to accept or reject the changes that the second reviewer suggested. Assuming 18% of the files contain seizures, a set of 1,000 files takes roughly 127 work hours to annotate.

Before an annotator contributes to the data interpretation pipeline, they are trained for several weeks on previous datasets. A new annotator is able to be trained using data that resembles what they would see under normal circumstances. An additional benefit of using released data to train is that it serves as a means of constantly checking our work. If a trainee stumbles across an event that was not previously annotated, it is promptly added, and the data release is updated. It takes about three months to train an annotator to a point where their annotations can be trusted. Even though we carefully screen potential annotators during the hiring process, only about 25% of the annotators we hire survive more than one year doing this work. To ensure that the annotators are consistent in their annotations, the team conducts an interrater agreement evaluation periodically to ensure that there is a consensus within the team. The annotation standards are discussed in Ochal et al. [4]. An extended discussion of interrater agreement can be found in Shah et al. [5].

The most recent release of TUSZ, v1.5.2, represents our efforts to review the quality of the annotations for two upcoming challenges we hosted: an internal deep learning challenge at IBM [6] and the Neureka 2020 Epilepsy Challenge [3]. One of the biggest changes that was made to the annotations was the imposition of a stricter standard for determining the start and stop time of a seizure. Although evolution is still included in the annotations, the start times were altered to start when the spike-wave pattern becomes distinct as opposed to merely when the signal starts to shift from background. This cuts down on background that was mislabeled as a seizure. For seizure end times, all post ictal slowing that was included was removed. Only two EEG files were added because, originally, they were corrupted in v1.5.1 but were able to be retrieved for the latest release. The progression from v1.5.0 to v1.5.1 and later to v1.5.2, included the re-annotation of all the EEG files in order to develop a confident dataset regarding seizure identification.

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 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] 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*.

[2] 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*.

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

[5] V. Shah, E. von Weltin, T. Ahsan, I. Obeid, and J. Picone, "On the Use of Non-Experts for Generation of High-Quality Annotations of Seizure Events," *J. Clin. Neurophysiol.*, 2019.

[6] I. Kiral et al., "The Deep Learning Epilepsy Detection Challenge: Design, Implementation, and Test of a New Crowd-Sourced AI Challenge Ecosystem," in Challenges in Machine Learning Competitions for All (CiML), 2019, pp. 1–3.

NEURAL ENGINEERING
DATA CONSORTIUM
www.nedcdata.org

Improving the Quality of the TUSZ Corpus

S. Rahman, A. Hamid, D. Ochal, I. Obeid and J. Picone
The Neural Engineering Data Consortium, Temple University

College of Engineering
Temple University

Abstract

- The Temple University Hospital Seizure Detection Corpus (TUSZ) is a subset of the TUH EEG Corpus (TUEG) and is the most frequently requested corpus from our 3,700+ subscribers.
- It was recently featured as the challenge task in the Neureka 2020 Epilepsy Challenge and an internal challenge conducted by IBM researchers.
- The latest version of TUSZ (v1.5.2) represents our efforts to review the quality of the annotations for the internal deep learning challenge at IBM and the Neureka 2020 Epilepsy Challenge.
- Annotations were reviewed with these goals:
  - ❑ Start times were adjusted to more accurately characterize the evolution of a seizure.
  - ❑ All post-ictal slowing at the end of a seizure was annotated as background.
  - ❑ The locality of a seizure event (the subset of channels on which there was evidence of a seizure) was improved.
  - ❑ All annotations were reviewed by at least three annotators. Disagreements were resolved by committee discussion.
- A new eval set is being prepared as part of v1.5.3 that includes a richer variety of seizure events. Future releases will include all TUH EEG data through mid-2019.

Introduction

- NEDC's historical archive of EEG includes every EEG collected at Temple University Hospital (TUH) since 2012:

  www.isip.piconepress.com/projects/tuh_eeg

  There are several valuable subsets of the data available that facilitate specific research (e.g., seizure detection, artifact detection).
- This corpus now includes over 3,700 active subscribers and has been updated regularly annually since 2012.
- TUSZ is a fully annotated corpus which means that every seizure event that occurs within its files has been annotated.
- This latest version of the corpus uses annotation standards updated to reduce the amount of mislabeled background that was contained in seizure files. This helps machine learning algorithms reduce their false alarm rate.
- New documentation for the corpus includes:
  - ❑ Annotations: This document was updated to describe in great detail the annotation standards used. This document also describes the file formats used to store annotation information.
  - ❑ Electrodes: This document explains how the data was collected (e.g., physical location of the electrodes), visualized (e.g., montages) and stored in EDF files (e.g., channel labels).

TUH EEG Seizure Corpus (TUSZ: v1.5.2)

- Increased the quality of the corpus:

Releases	v1.5.0	v1.5.1	v1.5.2
Patients	692	692	692
Sessions	1,661	1,575	2,608
Files	6,633	6,633	6,635
Seizure Files	1,399	1.382	1,384
Total No. Seizure Events	3,591	3,554	3,561
Total Duration (Hours)	1,074	1,074	1,074
Seizure Duration (Hours)	74.6	73.5	73.9

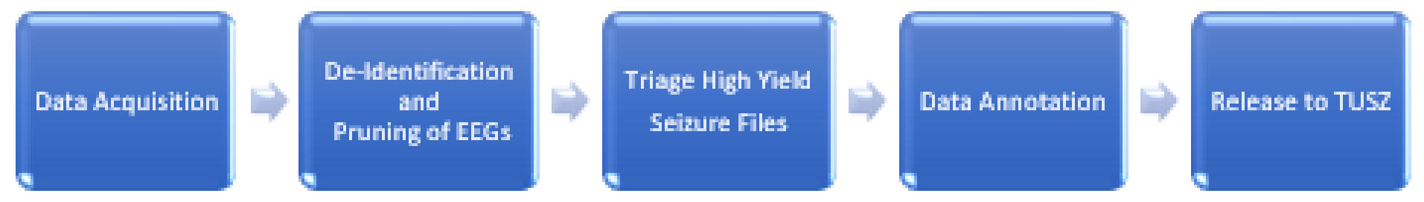
- The corpus (v1.5.2) is divided into three partitions:

Releases	Train	Dev	Eval
Patients	592	50	50
Sessions	1185	238	1185
Files	4599	1012	1023
Seizure Files	869	280	235
Total No. Seizure Events	2377	673	511
Total Duration (Hours)	752.9	170.3	150.9
Seizure Duration (Hours)	47.2	16.2	10.5

- In v1.5.2, the amount of seizure data has been decreased by 0.6 hours. This is in part due to the removal of post-ictal slowing, hypnagogic hypersynchrony, and triphasic waves according to the updated version of our annotation standards.

File Acquisition

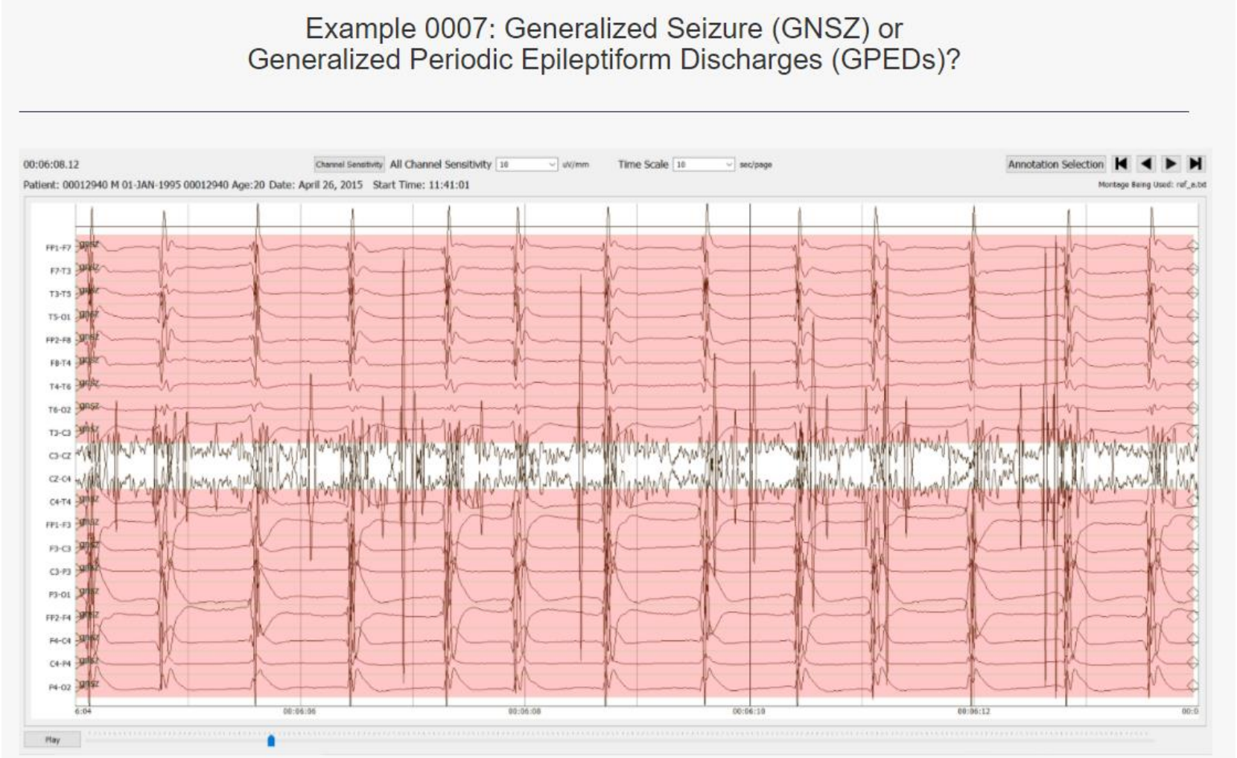
- This is a subset of TUH EEG (TUEG) developed for automatic seizure detection.
- Our database consists of pruned EEGs – a process where technicians discard uninteresting portions of the EEG signal. This saves a significant amount of disk space.
- These files are de-identified and subsequently copied to the TUEG database.


Data Acquisition → De-Identification and Pruning of EEGs → Triage High Yield Seizure Files → Data Annotation → Release to TUSZ

- A majority of the routine sessions are split into files shorter than 30 minutes in duration by the pruning process.
- Files listed as long-term monitoring (LTM) are greater than one hour in duration.
- Within TUSZ, files are divided into the train, dev, and eval sets for training, practice, and final evaluations respectively.
- The eval set is not released to the public, as it remains our standard for scoring all systems.
- The train set is continually developed and augmented.
- We use a two-part triage processing that combines a keyword detection system operating on session reports with a state-of-the-art seizure detection system that uses deep learning.

Annotation Process

- Once the subset of data is acquired from TUEG, our annotation team conducts several rounds of review to annotate the data.
- The files are originally split up amongst the annotators and are annotated individually.
- After each file has been reviewed once, the annotators swap files and examine each others' annotations. This second step is repeated to ensure a minimum of three annotators examines each file.
- Complex files are marked for group review. If the group of annotators do not reach a consensus, the file is posted on our FAQ where experienced members of the community provide their input.

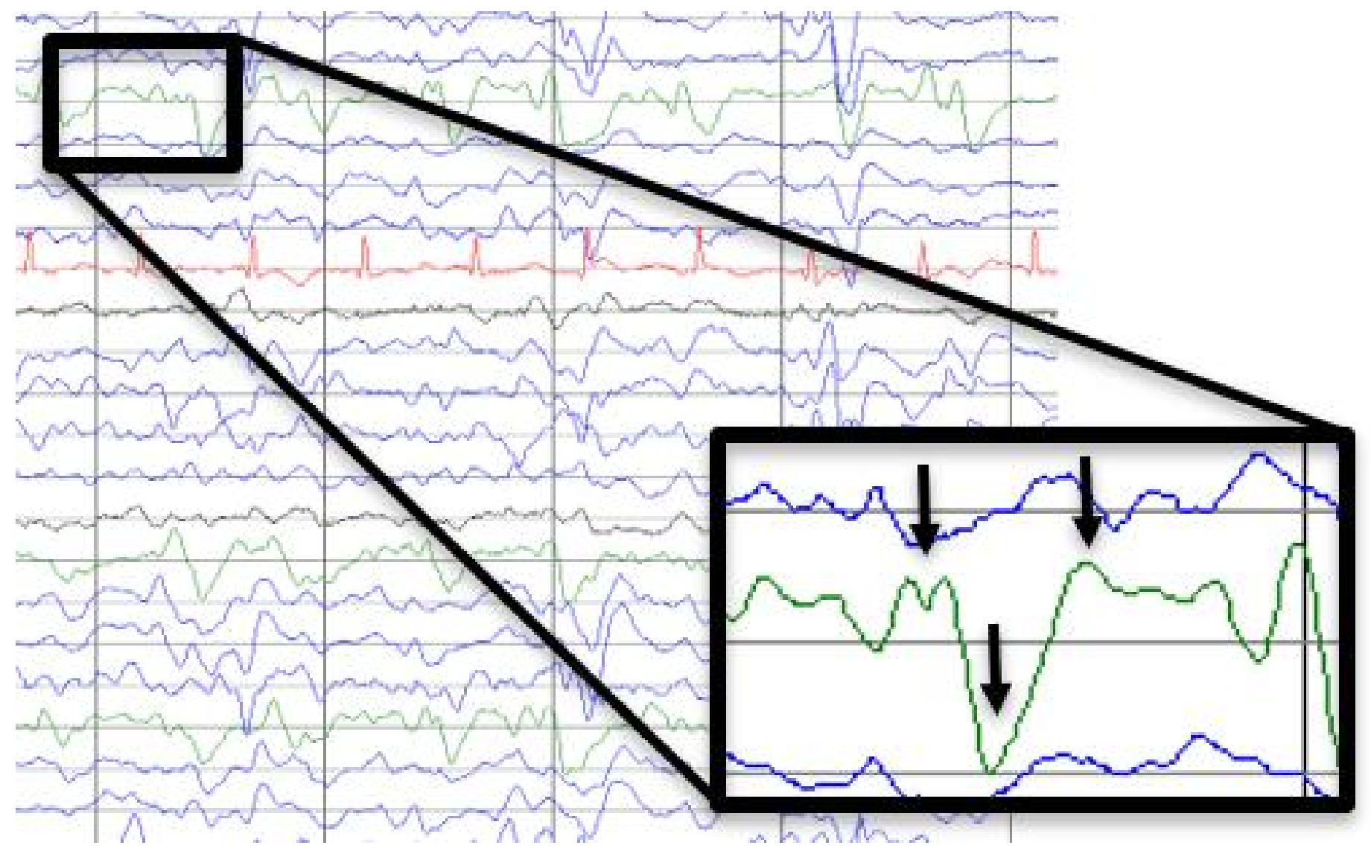

Example 0007: Generalized Seizure (GNSZ) or
Generalized Periodic Epileptiform Discharges (GPEDs)?

00:06:08.12
Patient: 00012940 M 01-JAN-1995 00012940 Age:20 Date: April 26, 2015 Start Time: 11:41:01
All Channel Sensitivity 10 uV/mm
Time Scale 10 sec/page
Montage Being Used: ref_a.txt

- Inter-rater agreement is quite high (kappa > 0.8) for this process.
- Error analysis on machine learning experiments is used to verify the integrity of the data.

Annotation Guidelines

- The official annotation guidelines used by NEDC can be found here: https://www.isip.piconepress.com/ publications/ reports/2020/tuh_eeg/annotations/.
- This is the first corpus released following the publication of the official standards.
- This public documentation allows for clarity and transparency in our work. Easy to follow images and explanations of technical language allow readers unfamiliar with EEG annotation to understand the data generated by our team of annotators.
- The latest version of the corpus, v1.5.2, includes several refinements to the data as outlined in the latest version of the document. Most notably, the removal of postictal slowing previously annotated as seizure, the removal of hypnagogic hypersynchrony, the removal of triphasic waves, and more stringent channel selection during seizure events.



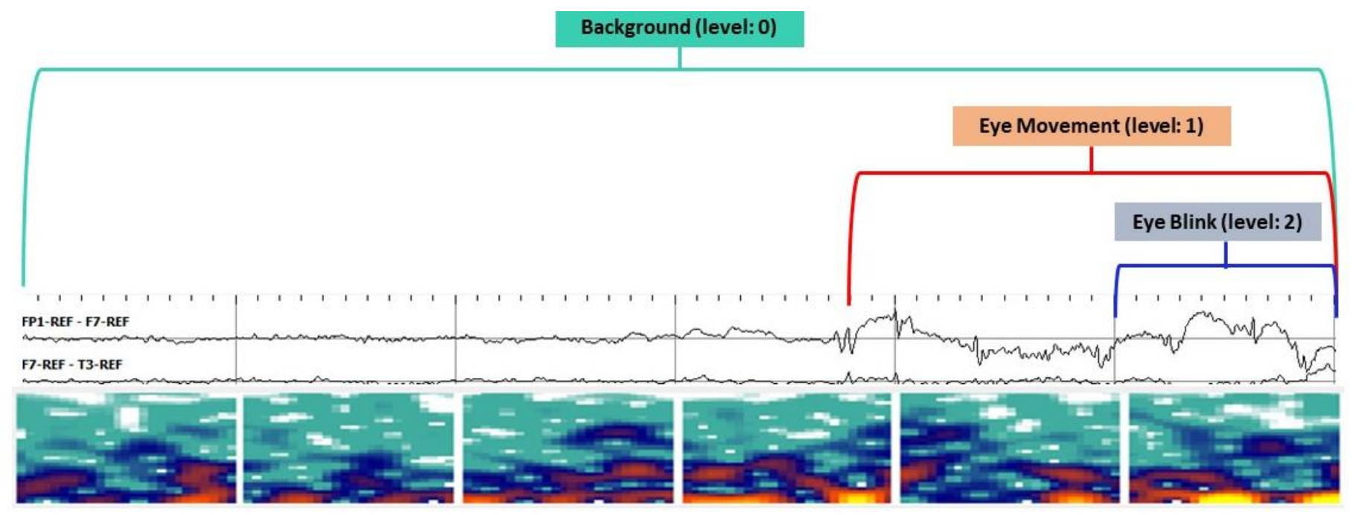
Release Process

- The data is annotated using our annotation tool, which generates ".rec" files – a csv-like format.
- .rec files are a simple way of recording channel-based annotations by using the following fields: start time, stop time, channel, event tag.
- The data that is released is in the form of .tse, .tse_bi, .lbl and .lbl_bi files.
- .tse files support term-based annotations:

```
version = tse_v1.0.0

0.0000 300.0511 bckg 1.0000
300.0511 324.0799 fnsz 1.0000
324.0799 531.9667 bckg 1.0000
531.9667 535.9623 fnsz 1.0000
535.9623 542.9545 bckg 1.0000
542.9545 547.9489 fnsz 1.0000
```

- .lbl files are a more complex, channel-based file type. This file type utilizes a hierarchal annotation format which allows different types of annotations to overlap one another.



Summary

- These corpora and supporting tools are open source and freely available at https://www.isip.piconepress.com/projects/tuh_eeg.
- Future release plans include:

Database	Version	Description	Expected Date
TUSZ	v1.5.3	The replacement of eval and dev patients	January 2020
TUSZ	v1.6.0	The addition of annotated seizure files through 2016	January 2020
TUSZ	v1.7.0	The addition of annotated seizure files from 2017–mid 2019.	May 2020

- We are augmenting our annotation file formats with direct support for csv and xml. Formats .rec and .tse will be obsoleted. We expect this transition to occur by Spring 2020.
- For further information, contact help@nedcdata.org.

Acknowledgements

- Research reported in this publication was most recently supported by the National Science Foundation Partnership for Innovation award no. 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.