April 23, 2017





**Temple University**

**ENGR Room 703A**

**1947 North 12th Street**

**Philadelphia, PA 19122**

**Tel: 215-204-4841 (office) / 215-954-7076 (cell)**

**Fax: 215-204-5960**

**Email: joseph.picone@gmail.com**

**URL: http://www.isip.piconepress.com**

To Whom It May Concern:

Please consider the attached paper for publication in your journal. The requested information is as follows:

1. Manuscript title:

**“Objective Evaluation Metrics for Automatic Classification of EEG Events”**

1. Manuscript classification:

**Original Research**

1. Notification of similar or duplicate publication by the author(s):

**This is an original submission.**

1. Statement that the principle/corresponding author has full access to data and has the right to publish such data:

**The authors are members of the Institute for Signal and Information Processing and/or the Neural Engineering Data Consortium at Temple University. We are the developers of the data presented in this paper and distribute it from our web site as open source. We have full access to the data and the right to publish such data.**

1. Statement that all authors participated in a meaningful way in the preparation of the manuscript (in the design or conceptualization of study, interpretation of data or drafting manuscript):

**All authors contributed in a meaningful way to this publication.**

Best regards,



Dr. Joseph Picone

Professor, Department of Electrical and Computer Engineering

**Objective Evaluation Metrics for Automatic Classification of EEG Events**

Seyedeh Saeedeh Khoshgoftar Ziyabari, Meysam Golmohammadi,
Iyad Obeid and Joseph Picone

**Abstract**

*Introduction:* Feedback from critical care clinicians who use automated event detection software has been overwhelmingly emphatic that high false alarm rates are the single most important metric of performance. Traditional measures such as sensitivity and specificity do not completely characterize technology for such applications. Balancing these measures for event detection problems has been extensively studied in other disciplines, but surprisingly these approaches have not been adopted in bioengineering fields. In this paper, we propose the use of Actual Term-Weighted Value (ATWV) as a more appropriate evaluation metric for classification of EEG events such as seizures.

*Methods:* ... now you need to describe **HOW** you proved that ATWV was a good measure. What data did you look at? What experiments did you run? What measures did you use to show it had values? --- about 60 words ---

Manuscript submitted May 1, 2017. The authors are with the Department of Electrical and Computer Engineering at Temple University. The corresponding author is J. Picone, College of Engineering, ENGR 703A, 1947 North 12th Street, Philadelphia, Pennsylvania 19122 USA (phone: 215-204-4841; fax: 215-204-5960; email: joseph.picone@gmail.com).

*Results:* ... your results must be about how you proved ATWV is a good measure ... not about the performance of the technology... did it correlate well with other measures? How well? Did it provide unique insight? (e.g., many people quote sensitivity and specificity, but their systems are unusable... you showed the ATWV for these systems is low and needs to be above 0.5 for the system to be usable). --- about 40 words ---

*Conclusions:* We have introduced a better measure for evaluation difficult tasks such as automated seizure detection. We have shown that state of the art technology still needs improvement, and have established very important performance goals for the technology (e.g., ATWV > 0.5 and FA/24 hrs. ~ 1.0).

*Keywords*— Electroencephalography (EEG), Objective Performance Evaluation Metrics, Average Term-Weighted Value, Spoken Term Detection

# Introduction

Electroencephalograms (EEGs) are used in a wide range of clinical settings to record electrical activity along the scalp. EEGs are the primary means by which physicians diagnose and manage brain-related illnesses such as epilepsy, seizures and sleep disorder (Yamada et al., 2009). Manually interpreting of EEG signals is time consuming and expensive process since highly-trained neurophysiologists are required to read long data stream to identify the key events. The development of a system that can automatically interpret an EEG allows healthcare providers to keep pace with the growing demand for this diagnostic tool and would provide real-time alerting of potentially life-threatening conditions (Obeid and Picone, 2017). Automatic EEG interpretation and identification of critical events can thererefore be expected to significantly increase the quality and efficiency of a neurologist’s diagnostic work. Many methods have been developed over the years with various approaches to interpret the EEG signal automatically such as time–frequency analysis methods (Gotman, 1982; Gotman, 1999; Osorio et al., 1998; Sartoretto and Ermani, 1999), nonlinear techniques (Schad et al., 2008; Schindler et al., 2001; Stam, 2005), mimicking the human observer that reads the EEG (Deburchgraeve et al., 2008; Khamis et al., 2009), neural networks (Ramgopal et al., 2014) and support vector machines (Alotaiby et al., 2014).

Further development of automatic EEG interpretation system is very complicated due to the lake of standard evaluation metric (Temko et al., 2010). Some studies publish the performance of systems based on Good Decision Rate (GDR) and number of False Detection per hour (FD/h) in an event. For instance, a two-class problem such as seizure detection, the GDR is percentage of correctly detected seizure events by system. If system detects seizure event anytime between the start and end of actual reference seizure, it will be considered as a good decision rate. The number of false detection per hour is number of detected seizure in an hour when there is no labeled seizure in reference (Gotmanet al., 1997). The metric ignores the duration of correct and incorrect detected seizure events. Some other researchers report performance in terms of sensitivity and specificity (Japkowicz and Shah, 2011) of epochs in biomedical research applications (Liu et al., 1992, Aarabi et al., 2007, Navakatikyan et al., 2006). Each epoch is considered as a separate testing example even though EEG events can span multiple epochs. The result of classifier is presented in confusion matrix, which gives a very useful overview of performance. For example, for a two-class problem such as seizure detection, a confusion matrix has following categories:

True positives (TP) refer to number of epochs correctly labeled as seizure.

True negatives (TN) are number of epochs correctly labeled as non-seizure.

False positives (FP) are number of epochs incorrectly labeled as seizure.

False negatives (FN) refer to number of epochs incorrectly labeled as none-seizure.

Sensitivity (TP/TP+FN) and specificity (TN/TN+FP) are derived from these quantities. A precision–recall (PR) curve is an alternative method of scoring (Manning et al., 1999) in which precision is percentage of correctly detected seizure divided by predicted seizure epochs (TP/TP+FP), while recall is called sensitivity.

However, sensitivity can often be increased arbitrarily if one is willing to tolerate a poor specificity or a high false alarm rate. Interviews conducted with many clinicians have indicated that the primary reason commercially available technology is not used in clinical settings is due to the high false alarm rate (Obeid and Picone, 2015). This is perhaps the single most important metric today in guiding machine learning research applications in critical care. Critical care units are overwhelmed with the number of false positives that automated event detection equipment generates. To put this in perspective, one false alarm per bed per hour in a 12-bed ICU generates 12 interrupts per hour that must be serviced. This can easily overwhelm healthcare providers. Since many types of automated monitoring equipment are used in an ICU setting, each with significant false alarm issues, the number of false alarms that must be serviced by healthcare providers is overwhelming (Christensenet al., 2014). As a result, clinicians report that in practice they simply ignore these systems (Obeid and Picone, 2017).

Of course, one must balance sensitivity, specificity and false alarms. This has been studied extensively in other communities focused on event-spotting technology such as spoken term detection in voice signals (Mandal et al., 2014). A measure that we have borrowed from this research community is the Term-Weighted Value (TWV) (Doddington et al., 2000), which is based on the notion of a Detection Error Tradeoff (DET) curve (Martin et al., 1997). A DET curve is very similar to a Receiver Operating Characteristic (ROC) originally developed to assess the performance of a communications system (Jacobs &Wozencraft, 1965). TWV essentially assigns an application-dependent reward to each correct detection and a penalty to each incorrect detection.

The primary focus of the paper is to introduce readers to Spoken Term Detection (STD) evaluation metric and how apply it to automatic EEG interpretation systems in section 2. Finally, we present and discuss results of different advanced deep-learning systems based on STD evaluation metric in section 3 and 4.

# Method

* 1. **OVERVIEW**

Process of finding seizure in large corpus of EEG is similar to Keyword Search (KWS) in speech recognition. KWS also known as Spoken Term Detection (STD), is a speech processing task in which the goal is to find all the occurrences of a textual “keyword”, a sequence of one or more words, in a large corpus of speech data. In 2006, the U.S. National Institute of Standards and Technology (NIST) created the STD evaluation initiative to facilitate research and development of technology for retrieving information from archives of speech data. Fig. 1 illustrates the standard framework of an STD system. In this framework, speech signals are first transcribed by a speech recognizer to a certain form of intermediate representation, e.g., word/subword transcripts or lattices; and then a term detector searches the intermediate representation to find putative occurrences of the terms in search; finally, a decision maker judges each putative occurrence and determines if it is a reliable detection or a mistake. We call the speech recognizer in this architecture the ASR subsystem, and the term detector and the decision maker the STD subsystem.

Previous speech retrieval evaluations like TREC’s Spoken Document Retrieval (SDR) (Allan, 2002), and Topic Detection and Tracking (TDT) (Kuhn, 2010) have investigated technologies similar to STD. However, they each addressed different problems. Source data robustness is a key component of STD whereas SDR and TDT focused on the broadcast news domain. The query for STD, a search term, is a markedly smaller unit than SDR’s query definition which was a natural language description of an information need, and more specific than TDT’s topic exemplar documents. A technology similar to STD is keyword spotting (Mandal et al., 2014). Keyword spotting searches keywords by matching the keywords’ templates to the test speech. By sliding the starting frame of the match, potential keyword occurrences were detected using the famous Dynamic Time Warping (DTW) algorithm (Bridle, 1973; Christiansen and Rushforth, 1977; Myers et al., 1980; Higgins and Wohlford, 1985). An obvious weakness of the template-based approach is that templates cannot represent the inherit variation of human speech. This disadvantage arises from the approach itself and cannot be amended even with multiple templates (Christiansen and Rushforth, 1977). A novel approach base on Hidden Markov Model (HMM) was proposed to overcome this problem (Wilpon et al., 1989; Rohlicek et al., 1989). A background model to normalize the acoustic scores of detected keywords, which makes the scores of detections from different utterances comparable, was proposed (Rose and Paul, 1990). STD has shown high performance on rich resource languages, such as English, Chinese and Arabic (Mamou et al., 2007, Vergyrim et al., 2006). The main difference between keyword spotting and STD is the number of words in a search term.

# STD Evaluation Methodology

STD detects all the occurrences of each given term in the reference files. In this evaluation methodology, an estimate is required for the number of trails in the reference. If there are no discrete trials in a continuous reference, a constant will be specified as number of trials. Then, alignment between detected occurrence and reference is needed. This step is done by applying the Hungarian solution to the Bipartite Graph (Kuhn, HW., 2010) matching problem. It uses the kernel functionthat numerically compares the mapping of system and reference occurrences, as well as the missed detections and false alarms.

The kernel function first determines if the reference/system occurrences are mappable by requiring the system occurrence to be within a temporal tolerance collar (ΔT) of the reference occurrence. Specifically, the midpoint of the system occurrence must be within the interval from ΔT before the beginning to ΔT after the end of the reference occurrence as determined by forced alignment of the reference transcript to the audio.

The system generates a list of putative term occurrence. The STD performance is based on false alarms and missed detections. The performance of a system is evaluated base on Detection Error Tradeoff (DET) curves (Martin et al., 1997) and a Term-Weighted Value (TWV) for a specific operating point. The DET curve plots miss probability (PMiss) versus false alarm probability (PFA) which is function of detection threshold θ. The formula for a single term PMiss and PFA are:

$P\_{Miss}\left(term,θ\right)=1-\frac{N\_{correct}(term,θ)}{N\_{true}(term)}$

$P\_{FA}\left(term,θ\right)=1-\frac{N\_{spurious}(term,θ)}{N\_{NT}(term)}$

$θ$ is the detection threshold

$N\_{correct}(term,θ)$ is the number of correct detections of term with a detection score greater than or equal to θ.

$N\_{spurious}(term,θ)$ is the number of incorrect detections of term with a detection score greater than or equal to θ.

$N\_{true}(term)$ is the true number of occurrences of term in the corpus.

$N\_{NT}(term)$ is the number of opportunities for incorrect detection of term in the corpus (= “Non-Target” term trials).

Non-Target” term trials which is proportional to the number of second of speech. The formula for $N\_{NT}(term)$ is:

$N\_{NT}\left(term\right)=n\_{tps}.T\_{speech}-N\_{true}(term)$

Where:

$n\_{tps}$ is is the number of trials per second of speech ($n\_{tps}$ will be set arbitrarily to 1)

$T\_{speech}$ is the total amount of speech in the test data (in seconds).

The overall performance of system will be measured by TWV which assigns a value to each correct detection and a cost to each incorrect detection. It is one minus the weighted sum of the term-weighted probability of missed detection (PMiss(θ)) and the term-weighted probability of false alarms (PFA(θ)).

$TWV(θ )=1-average\_{term}\{P\_{Miss}\left(term,θ\right)+β.P\_{FA}\left(term,θ\right)\}$

Where:

$β=\frac{C}{V}.(Pr\_{term}^{-1}-1)$

Where:

C is the cost of an incorrect detection; defined as 0.1

V is the value of a correct detection; defined as 1

$Pr\_{term}$ is the prior probability of a keyword; defined as $10^{-4}$

A perfect system has the maximum TWV which is one. TWV of system with no output is zero (Fiscus JG et al., 2007). Negative TWV is feasible. The actual TWV (ATWV) is performance measured for a specific decision threshold ­– essentially establishing a specific operating point on the DET curve. This measure is useful when it is preferred to compare two systems based on a single number, though it is always better to compare DET curves over a range of operating characteristics. ATWV and DET curves are our recommended way to evaluate EEG interpretation systems.

# Applying STD evaluation metrice to SEIzure detction applications

The present work addresses the STD task defined by NIST for seizure detection system. The task consists in finding all the seizures matches of a specific query in a corpus of TUH EEG. The corpus is the world’s largest open-source clinical EEG corpus. It comprises 16,986 sessions from 10,874 unique subjects (Obeid and Picone, 2016). A query is the neurology events, background and seizure events. Manual transcripts of the EEG file are used to find true occurrences. The data can be annotated per epoch or per term. To put this in perspective, the term-based and epoch-based annotations (true occurrence) of a system are illustrated in Fig. 2 (a) and Fig. 2 (b) respectively. Also, their corresponding system hypothesis (output occurrence) are depicted in Fig. 3. For evaluating performance of the system, the system hypothesis is judged as correct or not according to whether it is close in time to an annotation of the query retrieved from manual transcripts; it is judged as correct if the midpoint of the system hypothesis is less than or equal to the time span of an annotation of the query. The performance of the system using term-based and epoch-based annotation is presented in Table 1 and Table 2. The system with term-based annotation outperformed the epoch-based annotation in term of specificity and false alarm rate per hour. Therefore, we used term-based annotation in our systems. In this approach, each event such a seizure is denoted by start time and stop time.

We designed a system (A) with 100% detection rate with zero false alarm rate per 24 hours. The hypothesis of system A was perturbed by 10% (5% deletion, 5% insertion), 20% (10% deletion, 10% insertion) and 50% (25% deletion, 25% insertion) to create system B, C and D respectively. System A with ATWV-1 is the best system and the worst one is system D with ATWV- 0.4568. Table 3 shows that a system with less false alarm rate and higher detection rate has better ATWV. The performance of two systems can be compared with a single number, ATWV.

# Results

We have tested the STD evaluation metrics on four advanced deep learning systems. The architecture of the systems is depicted in Fig. 4. An EEG signal is input to the system, typically as a European Data Format (EDF) file. The signal is converted to a sequence of feature vectors. a group of frames are classified into an event on a per-channel basis using combination of deep learning networks. The deep learning essentially looks across multiple epochs, which we refer to as temporal context, and multiple channels, which we refer to as spatial context since each channel is associated with a location of an electrode on a patient’s head. There are a wide variety of algorithms that can be used to produce a decision from these inputs. We have explored three approaches: (1) a seven-layers Convolutional Neural Network (CNN) and Multi-Layer Perceptron (MLP) as our baseline system; (2) a combination of three-layers Convolutional Neural Network (CNN) with Long Short Term Memory (LSTM), CNN/LSTM; (3) a CNN/DLSTM system which is a three-layers Convolutional Neural Network (CNN) with Deep Long Short Term Memory (DLTM) and (3) a CNN/BLSTM which is compose of a three-layers Convolutional Neural Network (CNN) and BidirectionalLong Short Term Memory (BLSTM). Here, we use a subset of TUH-EEG that was specifically labeled using term-based labels for seizures by a series of experts (Golmohammadi, 2017). The data was carefully annotated by a team of students who have been carefully trained to annotate seizures. Their work has been evaluated against expert neurologists who marked a portion of the same data and shown to have good IRA. Each event in the evaluation data has been reviewed by at least five different annotators. The performance of these four approaches using STD evaluation metric is summarized in Table 4. The baseline system (CNN/MLP) has the best ATWV-0.1558 with sensitivity-0.86%; specificity-51.32; false alarms/24hrs-17. The worst system belongs to a CNN/BLSTM. The performance of system is as follow: Sensitivity – 18.03%; Specificity – 62.91%; False Alarms/24hrs – 110; ATWV – -0.2389. The CNN/LSTM based system with ATWV- -0.0116 outperformed CNN/DLSTM based system with ATWV-0.0313.

# Discussion

The ATWV ranges from [-∞,1], with a score greater than 0.5 being indicative of a system that is performing well. Negative ATWV scores are typically indicative of systems with high false alarm rates. As it is shown in Table 4, the ATWV score is extremely poor for these systems largely due to the large emphasis this metric places on false alarms. The fundamental cause of having high false alarm is poorly detection of long events. Suppose there is a five-second seizure event Fig. 5(a) and a system detects one second background event between two two-second seizure events Fig. 5(b). The first seizure is considered as correct detection and the background and second seizure events become false alarms of system. The performance of the system with mentioned reference and system hypothesis is presented in Table 5.

As the actual TWV (ATWV) is performance measured for a specific decision threshold ­– essentially establishing a specific operating point on the DET curve. ATWV and DET curves are our recommended way to evaluate EEG interpretation systems. The DET curve plots miss probability (PMiss) versus false alarm probability (PFA). Miss and false alarm probabilities are functions of the detection threshold, ϴ. This (ϴ) is applied to the system’s detection scores, which are computed separately for each search term, then averaged to generate a DET line trace. The DET curve which is plotted in Fig. 6, shows the CNN/DLSTM system is the best ranked system. The CNN/LSTM, CNN/BLSTM and CNN/MLP systems are placed in the second, third and fourth rank respectively, while the order is not matched with order of ATWVs in Table 4 .

The principle reason is the STD package is tuned for speech recognition applications. The software should be tuned for automatic EEG interpretation applications. Hence, we tweaked two parameters in STD software. One of them is β, the number is set to 999.9 for speech recognition system. The other parameter is ΔT; it represents the maximum time distance between the temporal extent of the reference term and the midpoint of system's detected term for the two to be considered a pair of potentially aligned terms. The ΔT was set to 0.5 in speech recognition system. For automatic EEG interpretation applications, we set β and ΔT to 9.9 and 10 respectively. The performances of the four systems based on new parameters are listed in Table 6. The ATWV of the systems and DET curve are matched. The best system is CNN/DLSTM with ATWV- 0.4489 and the worst one is CNN/MLP with ATWV-0.2634.

# Conclusion

In this paper, we have introduced the current evaluation metrics and explained their problems in automatic EEG interpretation applications. We began by introducing the STD evaluation metric. The basic concept of STD evaluation has been overviewed. We have investigated the performance of four deep learning systems by mean of DET curve and ATWV number.

The STD metric can be an effective means to encourage research and develop state-of-the-art systems. Furthermore, appropriate evaluation methodologies and analysis of results can help illuminate the progress that has been made and identify the factors limiting further advance.

# Acknowledgements

Research reported in this publication was supported by the National Science Foundation under Grant No. IIP-1622765. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The TUH EEG Corpus work was funded by (1) the Defense Advanced Research Projects Agency (DARPA) MTO under the auspices of Dr. Doug Weber through the Contract No. D13AP00065, (2) Temple University’s College of Engineering and (3) Temple University’s Office of the Senior Vice-Provost for Research.

# ReferenceS

Yamada T, Meng E. Practical Guide for Clinical Neurophysiologic Testing: EEG. Philadelphia, Pennsylvania, USA: Lippincott Williams & Wilkins; 2009.

Obeid I, Picone J. Machine Learning Approaches to Automatic Interpretation of EEGS. In: *Automatic Interpretation of EEGs*. ; 2017:1-60.

Gotman J. Automatic recognition of epileptic seizures in the EEG. *ElectroencephalogrClinNeurophysiol*. 1982;54(5):530-540.

Gotman J. Automatic detection of seizures and spikes. *J ClinNeurophysiol*. 1999;16(2):130-140.

Osorio I, Frei MG, Wilkinson SB. Real-time automated detection and quantitative analysis of seizures and short-term prediction of clinical onset. *Epilepsia*. 1998;39(6):615-627.

Sartoretto F, Ermani M. Automatic detection of epileptiform activity by single-level wavelet analysis. *ClinNeurophysiol*. 1999;110(2):239-249.

Schad A, Schindler K, Schelter B, et al. Application of a multivariate seizure detection and prediction method to non-invasive and intracranial long-term EEG recordings. *ClinNeurophysiol*. 2008;119(1):197-211.

Schindler K, Wiest R, Kollar M, Donati F. Using simulated neuronal cell models for detection of epileptic seizures in foramen ovale and scalp EEG. *ClinNeurophysiol*. 2001;112(6):1006-1017. doi:10.1016/S1388-2457(01)00522-3.

Stam CJ. Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field. *ClinNeurophysiol*. 2005;116(10):2266-2301.

Deburchgraeve W, Cherian PJ, De Vos M, et al. Automated neonatal seizure detection mimicking a human observer reading EEG. *ClinNeurophysiol*. 2008;119(11):2447-2454.

Khamis H, Mohamed A, Simpson S. Seizure state detection of temporal lobe seizures by autoregressive spectral analysis of scalp EEG. *ClinNeurophysiol*. 2009;120(8):1479-1488.

Ramgopal S, Thome-Souza S, Jackson M, et al. Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy. *Epilepsy Behav*. 2014;37:291-307.

Alotaiby TN, Alshebeili S a, Alshawi T, Ahmad I, El-samie FEA. EEG seizure detection and prediction algorithms: a survey. *EURASIP J Adv Signal Process*. 2014;2014(183):1-21. http://asp.eurasipjournals.com/content/2014/1/183.

Temko A, Korotchikova I, Marnane W, Lightbody G, Boylan G. Validation of an automated seizure detection system on healthy babies. Proc. INSTICC/IEEE BioSignals 2010b.

Gotman J, Flanagan D, Zhang J, Rosenblatt B. Automatic seizure detection in the newborn: Methods and initial evaluation. *Electroencephalogr Clin Neurophysiol*. 1997. doi:10.1016/S0013-4694(97)00003-9.

Japkowicz N, Shah M. *Evaluating Learning Algorithms*.; 2011.

Liu A, Hahn J, Heldt G, Coen R. Detection of neonatal seizures through computerized EEG analysis. ElectroencephalogrClinNeurophysiol 1992;82:30–7.

Aarabi A, Grebe R, Wallois F. A multistage knowledge-based system for EEG seizure detection in newborn infants. ClinNeurophysiol 2007;118:2781–97.

Navakatikyan MA, Colditz PB, Burke CJ, Inder TE, Richmond J, Williams CE. Seizure detection algorithm for neonates based on wave-sequence analysis. *Clin Neurophysiol*. 2006. doi:10.1016/j.clinph.2006.02.016.

Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*.; 2008.

Obeid I, Picone J. *NSF ICORPS Team: AutoEEG*. Philadelphia, Pennsylvania, USA.;2015. <https://www.isip.piconepress.com/publications/reports/2016/nsf/icorps/report_v01.pdf>.

 Christensen M, Dodds A, Sauer J, Watts N. Alarm setting for the critically ill patient: A descriptive pilot survey of nurses’ perceptions of current practice in an Australian Regional Critical Care Unit. *Intensive Crit Care Nurs*. 2014. doi:10.1016/j.iccn.2014.02.003.

Mandal A, Prasanna Kumar KR, Mitra P. Recent developments in spoken term detection: A survey. *Int J Speech Technol*. 2014. doi:10.1007/s10772-013-9217-1.

Doddington GR, Przybocki MA, Martin AF, Reynolds DA. NIST speaker recognition evaluation - overview, methodology, systems, results, perspective. *Speech Commun*. 2000. doi:10.1016/S0167-6393(99)00080-1.

Martin A, Doddington G, Kamm T, Ordowski M, Przybocki M. The DET Curve in Assessment of Detection Task Performance. *Proc Eurospeech ’97*. 1997. doi:citeulike-article-id:1320711.

Wozencraft JM, Jacobs IM. *Principles of Communication Engineering*. New York: Wiley; 1965.

Allan J. Detection as multi-topic tracking. *Inf Retr Boston*. 2002;5:139-157. [http://www.springerlink.com/index/7AE8B93DKQJCJVMQ.pdf%5Cnpapers2://publication/uuid/AADBEEDF-A4D2-4D0B-82E4-5101CD25FA1B](http://www.springerlink.com/index/7AE8B93DKQJCJVMQ.pdf%5Cnpapers2%3A/publication/uuid/AADBEEDF-A4D2-4D0B-82E4-5101CD25FA1B).

Kuhn HW. The Hungarian method for the assignment problem. In: *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*. Springer Berlin Heidelberg; 2010:29-47.

Bridle J. S. An efficient elastic-template method for detecting given words in running speech. In: *British Acoustical Society Spring Meeting*. ; 1973:1-4.

R. Christiansen and C. Rushforth, “Detecting and locating key words in continuous speech using linear predictive coding,” *IEEE Trans. Acoust.*, 1977.

Myers CS, Rabiner LR, Rosemberg AE. An investigation of the use of dynamic time warping for word spotting and connected speech recognition. *ICASSP ’80 IEEE Int Conf Acoust Speech, Signal Process*. 1980:173-177. doi:10.1109/ICASSP.1980.1171067.

Wilpon J, Lee C, Rabiner L. Application of hidden Markov models for recognition of a limited set of words in unconstrained speech. *Acoust Speech, …*. 1989;(2):254-257. doi:10.1109/ICASSP.1989.266413.

Rohlicek J, Russell W. Continuous hidden Markov modeling for speaker-independent word spotting. *IEEE Int Conf Acoust Speech, Signal Process*. 1989:627-630. doi:10.1109/ICASSP.1989.266505.

Rose RC, Paul DB. A hidden Markov model based keyword recognition system. *Int Conf Acoust Speech, Signal Process*. 1990:129-132. doi:10.1109/ICASSP.1990.115555.

Mamou J, Ramabhadran B, Siohan O. Vocabulary Independent Spoken Term Detection. *Proc 30th Annu Int ACM SIGIR Conf Res Dev Inf Retr*. 2007:615-622. doi:10.1145/1277741.1277847.

Vergyrim D, Stolcke A, Gadde R, Wang W. The SRI 2006 Spoken Term Detection System. *NIST Spok term Detect Work 2006 Work*. 2006:3-6.

Fiscus JG, Ajot J, Garofolo JS, Doddingtion G. Results of the 2006 spoken term detection evaluation. *Proc ACM SIGIRWorkshop Search Spontaneous Conversational Speech*. 2007:51-57. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.132.7776&rep=rep1&type=pdf#page=57>.

Obeid I, Picone J. The Temple University Hospital EEG Data Corpus. *Front Neurosci Sect Neural Technol*. 2016;10:196. doi:http://dx.doi.org/10.3389/fnins.2016.00196.

Golmohammadi M, Shah V, Lopez S, et al. The TUH EEG Seizure Corpus. *J Clin Neurophysiol*. 2017:1.

# LIST OF TABLES

Table 1. System performance based on term-based annotation.

Table 2. System performance based on epoch-based annotation.

Table 3. The performances of a system with various perturbations in hypothesis.

Table 4. The STD-based performances of four different deep-learning systems.

Table 5. System performance based on STD evaluation metric.

Table 6. The STD-based performance of four different deep-learning systems using new parameter value.

Table 1. System performance based on term-based annotation.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Target | #Target | #Correct | #Miss | #FA | Sensitivity | Specificity | FA/h |
| Seizure (s) | 1 | 1 | 0 | 1 | 100% | 75% | 720 |
| Background(b) | 1 | 1 | 0 | 0 | 100% | 100% | 0 |

Table 2. System performance based on epoch-based annotation.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Target | #Target | #Correct | #Miss | #FA | Sensitivity | Specificity | FA/h |
| Seizure (s) | 2 | 2 | 0 | 2 | 100% | 33% | 1440 |
| Background(b) | 3 | 1 | 2 | 0 | 33% | 100% | 0 |

**Table 3. The performances of a system with various perturbations in hypothesis.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| System | A | B | C | D |
| Perturbation | 00% | 10% | 20% | 50% |
| Sensitivity | 100%       | 95.04% | 90.08% | 75.07% |
| Specificity | 100% | 97.55% | 95.29% | 88.58% |
| ATWV | 1 | 0.888 | 0.7814 | 0.4568 |
| Seizure FAs/24 hrs | 0 | 6 | 12 | 30 |
| Background FAs/24 hrs | 0 | 6 | 12 | 31 |
| Total FAs/24 hrs | 0 | 12 | 25 | 62 |

Table 4. The STD-based performances of four different deep-learning systems.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CNN/MLP | CNN/LSTM | CNN/DLSTM | CNN/BLSTM |
| Sensitivity | 0.86%       | 11.02% |  14.45% | 18.03% |
| Specificity | 52.02% | 58.25% | 61.08% | 62.91% |
| ATWV | 0.1558 | 0.0313 | -0.0116 | -0.2389 |
| Seizure FAs/24 hrs | 2 | 23 | 31 | 53 |
| Background FAs/24 hrs | 15 | 31 | 35 | 57 |
| Total FAs/24 hrs | 17 | 54 | 66 | 110 |

Table 5. System performance based on STD evaluation metric.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target | #Target | #Correct | #Miss | #FA |
| Seizure (s) | 1 | 1 | 0 | 1 |
| Background(b) | 1 | 1 | 0 | 1 |

Table 6. The STD-based performance of four different deep-learning systems using new parameter value.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CNN/MLP | CNN/LSTM | CNN/DLSTM | CNN/BLSTM |
| Sensitivity | 0.86%       | 17.88% |  21.95% | 8.87% |
| Specificity | 52.02% | 63.48% | 68.60% | 58.19% |
| ATWV | 0.2634 | 0.4024 | 0.4489 | 0.3323 |
| Seizure FAs/24 hrs | 2 | 36 | 30 | 23 |
| Background FAs/24 hrs | 14 | 38 | 37 | 28 |
| Total FAs/24 hrs | 16 | 75 | 67 | 51 |

# LIST OF FIGURES

Fig. 1. The standard STD architecture (Wang, 2009).

Fig. 2. (a) The term-based annotation. (b) The epoch-based annotation.

Fig. 3. Fig. . (a) The term-based system hypothesis. (b) The epoch-based system hypothesis.

Fig. 4. An overview of the CNN/MLP, CNN/LSTM, CNN/DLSTM and CNN/BLSTM seizure detection system.

Fig. 5. A five-second seizure follows by a one-second background event in an annotation. (b) A one-second background between two two-send seizure events in addition to one second background event in system hypothesis.

Fig. 6. The DET curve of four different deep-learning systems.

# Figures



Fig. . The standard STD architecture (Wang, 2009).



Fig. . (a) The term-based annotation. (b) The epoch-based annotation.



Fig. . (a) The term-based system hypothesis. (b) The epoch-based system hypothesis.



Fig. . An overview of the CNN/MLP, CNN/LSTM, CNN/DLSTM and CNN/BLSTM seizure detection system.



Fig. . (a) A five-second seizure follows by a one-second background event in an annotation. (b) A one-second background between two two-send seizure events in addition to one second background event in system hypothesis.



 Fig. . (a) Reference file. (b) System hypothesis.



Fig. . The DET curve of four different deep-learning systems.