Audio-Based Epileptic Seizure Detection

M.N. Istiaq Ahsan[1], Csaba Kertesz[2], Annamaria Mesaros[1], Toni Heittola[1], Andrew Knight[2], Tuomas Virtanen[1]
[1]*Faculty of Information Technology and Communication Sciences, Tampere University, Finland*
email: {istiaq.ahsan, annamaria.mesaros, toni.heittola, tuomas.virtanen}@tuni.fi
[2] *Neuroeventlabs, Tampere, Finland*
email: {csaba.kertesz, andrew.knight}@neuroeventlabs.com

***Abstract*—This paper investigates automatic epileptic seizure detection from audio recordings using convolutional neural networks. The labeling and analysis of seizure events are necessary in the medical field for patient monitoring, but the manual annotation by expert annotators is time-consuming and extremely monotonous. The proposed method treats all seizure vocalizations as a single target event class, and models the seizure detection problem in terms of detecting the target vs non-target classes. For detection, the method employs a convolutional neural network trained to detect the seizure events in short time segments, based on mel-energies as feature representation. Experiments carried out with different seizure types on 900 hours of audio recordings from 40 patients show that the proposed approach can detect seizures with over 80% accuracy, with a 13% false positive rate and a 22.8% false negative rate.**

***Index Terms*—Epileptic seizure detection, convolutional neural network (CNN), sound event detection, audio processing and analysis.**

I. Introduction

The epilepsy is an incurable disease, but the patients can achieve seizure free state by taking medicines. To find the optimal drugs, the doctors need to know the epilepsy type and the impact of certain drug combinations on the patient. The nurses and family members write a diary about the seizure events, but the emerging modern technologies can help to automate this process. The epileptic seizure detection has many challenges because this neurological malfunction can affect varying brain regions and cause different limb movements, facial expressions or screams. The medical devices worn by the patients are called invasive instruments (e.g. EEG, bracelet with accelerometer). Non-invasive methods include video surveillance or audio recordings. The invasive devices can monitor vital parameters precisely, they are uncomfortable for the patients and they can be disconnected accidentally or by motor seizures. The audio and video recordings are less accurate than the invasive methods though they are always available. An advantage of the non-invasive devices that they can be used in home setting while the patients must stay in a hospital for invasive monitoring what is very expensive in a long-term and it involves long waiting lists.

The golden standard for the epilepsy diagnosis is the video electroencephalography (VEEG) and generalized tonic-clonic seizures (GTCS) can be detected reliable with this approach. During a VEEG session, the patient is constantly monitored by camera and EEG for several days to know the brain activity if a seizure happens. Geertsema et al [1] used optical flow record on VEEG data. They selected the parameters with 72 seizure videos for their algorithm and the evaluation was done on 24 full nights of 12 new subjects. They detected all GTCS events in the testing set and the median false detection rate (FDR) was 0.78 per night. Leijten et al [2] and Beniczky [3] reviewed non-EEG based methods for multiple seizure types. The review included accelerometers, heart rate monitors, electrodermal activity sensors and electromyograph sensors. They found that non-EEG methods need multimodal seizure detection for clinical use. Although they can detect GTCS with high sensitivity (90%) and low FDR (0.2/day), there is limited evidence for detection of other seizures than GTCS. Many false detections (50-216/day) and low sensitivity (19-74%) limit the applications of these wearable devices.

The audio modality of the epileptic seizures (vocalization) got little attention in the research community. Al-Hammadi [4] analyzed if the audio recordings of epileptic and psychogenic non-epileptic (PNE) seizures can be classified with linear, quadratic or support vector machine (SVM) classifiers. The SVM classifier with 4 mel-bands achieved 76% accuracy in cross-validation, but the sample size was limited to 16 epileptic and 20 PNE seizures.

Arends et al [5] studied epilepsy patients with intellectual disability. After they selected ten adults with major seizures and recognizable sounds out of the initial 17 individuals, each person was monitored with audio and video recording for 4 weeks. The audio detection was reliable at detecting either the initial vocalization or noise in a major seizure, but the produced sounds were specific for each patient, therefore, major seizures could be detected for the half of the participants.

Bruijne et al [6] observed vocalization during 61 out of 95 seizures of 17 patients. The study used simple audio statistics from 25 msec-long sliding windows and classifiers were trained to look for e.g. scream, smacking of lips, moans or heavy breathing. Their experimental results showed precisions of 66-77% with 10-fold cross-validation, but precision decreased quickly with added noise, and the classifiers were not evaluated with unseen data.

This paper proposes a novel approach to epileptic seizure detection using audio, by treating the various seizure vocalizations as a single target event class and modeling the seizure detection problem in terms of target vs non-target classes. The method uses a convolutional neural network (CNN) to learn the acoustic characteristics of the two classes, using a large dataset of annotated patient data. At detection stage, the input

consists of long audio recordings (10-14 hours) on which the system performs the seizure detection in short time segments.

This paper is organized as follows: Section II introduces the proposed method for analysis of seizures in audio, including feature representation, problem modeling, and system architecture. Section III describes the dataset, the data preparation process and the cross-validation setup, Section IV presents and discusses the results obtained using the proposed method, and finally Section V presents conclusions and future work.

II. Target sound event detection in audio

Sound event detection usually refers to identifying the onset (starting point) and offset (ending point) of a certain target sound events and labeling them by analyzing the acoustic features. In the proposed method, the epileptic seizure is considered as a target event to be detected in the audio recording. We chose a CNN architecture for detection, as CNNs are often applied and perform well for different audio classification and detection tasks, being one of the top recent approaches for acoustic scene classification [7], [8], audio tagging [9]–[11], or detection of target sounds such as bird sounds [12], [13]. In particular, CNN is among state of the art methods for rare sound event detection [14], which represents a similar type of problem as the seizure detection task.

For detecting a seizure event, the input audio is divided into short segments from which the features are extracted. The CNN is trained using these features; in the test phase, the CNN provides decisions on target event being present in the test audio, in segments of the same length as in training. The segment length for the audio analysis and seizure detection was selected as 10 seconds as it provides enough audible content about the target event in each segment that is needed for training the neural network. This method provides detection of the target event with a 10 s temporal resolution, instead of exact onset and offset. Figure 1 presents the block diagram for detection, with training and testing branches illustrated separately.

A. Feature Extraction

Mel-band energies provide a coarse representation of the spectral information and are proven to be a suitable feature representation for many audio classification/detection problems [15] [16] [17]. In addition, they are compact and easy to calculate. For these reasons, they were selected as feature representation in this work. Specifically, mel-band energy was calculated for each 10 s segment of the input audio, using 40 mel bands, and a window length of 40 ms with a 20 ms hop length.

B. Target event detection using CNN

The extracted features are fed to the CNN as input sequence, which are $I \times 40 \times 500$ dimensional. Here, I indicates the number of filters, 40 is the number of mel bands and 500 is sequence length (= 10 s). These features are learned using the convolutional layers, and the kernels that spread over both frequency and time axes enable the CNN to learn the relevant temporal characteristics of the target class. During the training process, in each segment, the target output is 1 while a seizure event is active and 0 while inactive. A max-pooling operation along both axes is performed after each CNN layer to reduce the dimensionality of the data. Batch normalization [18] is done after both of the convolutional layers for normalizing layer inputs, and the activation function used here is rectified linear unit (ReLu). To prevent overfitting, a 30% dropout is used after each layer. The output layer contains one sigmoid unit, and the final output is binary, indicating target class active or inactive. We use the binary cross-entropy as the loss function, and Adam optimizer [19] during the training. The CNN layers are presented in Fig. 2.

In the test phase, binary classification by the trained CNN is obtained for each 10-s segment.

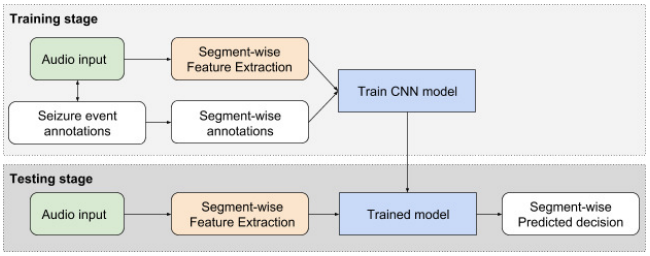

Fig. 1. Training and testing the system for seizure detection

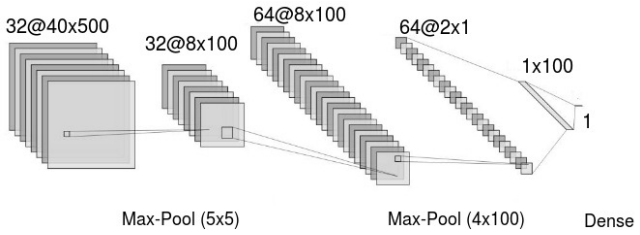

Fig. 2. Different layers of the CNN model architecture

III. Dataset

The dataset used in this paper consists of audio recordings of 40 seizure patients. The monitoring of the patient involved video recording during the night inside their residence; the average recording length is about 10-14 hours. The audio was extracted from the video recordings and then prepared for analysis and detection of seizures.

The data was annotated by certified nurses who watched the entire video recordings on high speed and labeled seizure events based on visual cues only. As the audio clips are extracted directly from the video, these have the same annotations. For preparing the training dataset, the target class (seizure) events were clipped from the audio based on these

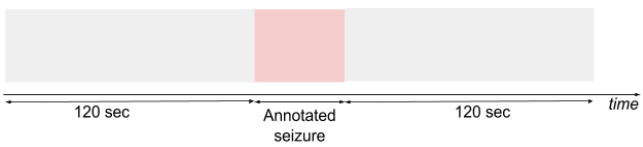

Fig. 3. Extraction of target seizure events from video

TABLE I
NUMBER OF EXAMPLES IN TRAIN AND TEST SET OF DIFFERENT SETUPS; SRCONT TEST SET CONTAINS 74 WHOLE NIGHT RECORDINGS

Experiment setup	Test set	No. Train examples	No. of Test examples
SRSeg	patient- A, B, C, D	765	199
SRCont	patient- D, E, F	841	74

annotations, with a 120-s additional segment of non-target (non-seizure) material both before and after the target one. The role of this padding is to serve as example of non-target class and provide context information; considering that usually seizures are rather short (10-30 s), we considered that 120-s padding provides a sufficient degree of data imbalance to the model. In consequence, each audio file has the content as illustrated in Fig. 3. For testing, also continuous audio as recorded throughout the night was used. The audio material is mono, with a sampling rate of 48 kHz, and 80 kbps bit rate which is attained using the Ogg Vorbis compression format.

Two different experimental setups have been prepared. First, a total of 954 audio examples containing target events and non-target audio padding were prepared according to the procedure mentioned above; among these, we manually labeled 199 of them based on different characteristics of the acoustic content using the following categories:

- **Sound seizure:** containing typical seizure vocalizations, no background noise.
- **Noisy seizure:** containing lots of background noise (other people, TV, etc).
- **Silent seizure:** no vocalization, mostly silent.

In addition, whole night recordings are used as test data; and some may not contain any seizure events. There are in total 74 nights of full-length recordings, estimated around 850-900 hours.

The two experimental setups are constructed as follows:

- **Seizure Recordings Segmented (SRSeg):** containing the 954 audio clips that were prepared with a 120 s padding. The test set contains the 199 clips that have detailed annotation, while the training set contains the other 765 clips. Training data contains data from 18 patients, while test data contains data from 4 patients which are not present in the training set. For added anonymity and ease of future reference, we denote individual patients alphabetically (A, B, C and D).
- **Seizure Recordings Continuous (SRCont):** containing the same type of segmented data as SRSeg for training, and whole night recordings (continuous) for testing. The training data of SRCont contains data from 18 patients (not exactly the same as SRSeg), while the test data contains data from 3 patients (D, E, and F).

Table I shows the details of these two setups, including the number patients in the test set, and number of examples in train and test set.

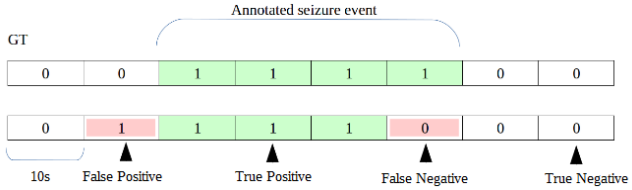

Fig. 4. Error types and correct detection output when comparing system output and ground truth using 10-s length segments

TABLE II
DETECTION RESULTS ON SEGMENTED AUDIO WITH VARIABLE THRESHOLD VALUES

Threshold	FPR	FNR	TPR	TNR	Accuracy
0.5	23.75	47.86	52.14	76.25	71.42
0.1	39.13	22.08	77.92	60.87	64.28
0.05	65.26	9.74	90.26	34.74	45.86

IV. EXPERIMENTAL RESULTS

The experiments were conducted using the setup described in the previous section. Results are analyzed overall and by considering the different characteristics of seizures in SRSeg and the different patient characteristics in SRCont.

A. Evaluation Metrics

We measure five different metrics to evaluate the system: false positive rate (FPR), false negative rate (FNR), true positive rate (TPR), true negative rate (TNR), and detection accuracy. Evaluation is performed in segments of length 10 s according to the modeling and system output, following the methodology in [20] for sound event detection. Fig. 4 illustrates the evaluation procedure by comparing the the system output (SO) with the ground truth (GT) of a test audio examples containing a seizure event. The illustration contains one false positive–no seizure annotated in GT, but detected as active in system output and one false negative–segment annotated as seizure in GT, but not detected in system output.

B. Seizure detection in segmented audio data

Overall results using the SRSeg dataset are presented in Table II, using different thresholds for the decision making. Because of the nature of the data and target application, it is important to detect all seizure events, therefore the emphasis is on minimizing the number of false negatives: reducing the number of times the system fails to detect a seizure event whereas there is one. As a consequence, accepting false positives is tolerable to some extent. From Table II, we can see that with the decision threshold set to 0.5, the FNR is about 47% whereas the FPR is 23%, with 71% detection accuracy. When the detection threshold reduced, FNR tends to decrease; the increase of FPR is however not at the same rate.

Detection performance evaluated separately for the three seizure types described in Section III is presented in Table III. We observe that the sound seizures are detected with 73% accuracy with 48.2% false negative rate and 22.7% false positive rate which is similar to the overall results for

TABLE III
DETECTION RESULTS ON SRSEG FOR DIFFERENT SEIZURE TYPES, USING THRESHOLD 0.5

Seizure Type	FPR	FNR	TPR	TNR	Accuracy
Sound seizure	22.71	48.28	51.72	77.29	73.31
Noisy seizure	29.35	44.14	55.86	70.65	65.95
Silent seizure	18.53	62.96	37.04	81.47	73.94

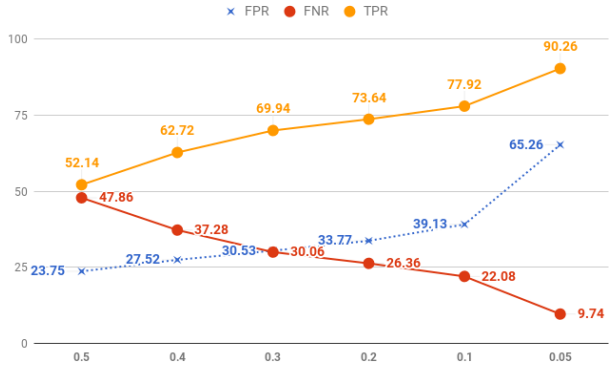

Fig. 5. FNR, FPR and TPR values against threshold for different values

the same threshold 0.5, mentioned in Table II. However, the noisy seizures have slightly less amount of false negatives but increased FPR at about 29% which happens to be the highest among others because of the extra background noises. The false negative rate is the highest for the silent seizures, which is expected, as the amount of audible audio content present in this type of seizure is significantly lesser than the others. That is why the false negative rate is about 63% and the true positive rate is also lowest, at 37%.

Figure 5 illustrates the behaviour of FPR, FNR and TPR for different threshold values from 0.05 to 0.5. When the threshold is reduced, FNR decreases while FPR and TPR increase; FPR has a sharp increase at the low detection thresholds. When the threshold is set to 0.3, FNR and FPR are quite similar, at about 30%. If the operating point is reduced as low as 0.05, FNR is reduced to about 9%, while FPR is at its highest at 65%. Finding an optimal operating point to balance between the FNR and FPR is the key to obtaining the optimal detection performance for the task.

C. Seizure detection in continuous recordings

The system was tested with whole-night recordings from three different patients, to understand its behavior in a real-life detection situation. Experiments on SRCont were conducted with the same parameter set and system setup as for SRSeg. The evaluation results for continuous recordings of three patients are presented in the table IV. For this experiment too we varied the detection threshold to observe the behaviour of the system. Overall, as the threshold decreases, FNR reduces to as low as 1.8%, but with the very high FPR of 85%. A more balanced performance is achieved for the 0.5 threshold, 82% accuracy with a false negative rate of about 22% and the false positive rate about 13%.

We also evaluated the system separately for each patient, using the threshold of 0.5. Results are shown in Table V. We

TABLE IV
DETECTION RESULTS ON CONTINUOUS RECORDINGS FOR DIFFERENT THRESHOLD VALUES

Threshold	FPR	FNR	TPR	TNR	Accuracy
0.5	13.01	22.81	77.93	86.99	82.14
0.3	63.35	11.38	88.62	36.65	36.95
0.2	85.37	1.85	98.15	14.63	15.12

TABLE V
PATIENT-WISE-SRCONT

Patient	FPR	FNR	TPR	TNR	Accuracy
D	4.92	40.72	59.23	95.09	77.18
E	5.59	18.81	83.39	94.40	87.92
F	28.50	8.90	91.09	71.49	81.29

observe that FNR varies for different patients although the decision threshold is the same for all of them. This suggests that a different operating point could provide optimal detection for different patients. In this case, FNR is lowest for patient F but at the same time FPR is the highest. For patient E, FNR is about 18% whereas FPR is as low as 5%, with 87% detection accuracy. A TNR of about 94% suggests that the system is able to distinguish between the target and non-target events quite successfully. Tuning the operating point for each patient individually might help detecting the target events better.

According to table V, the best detection performance was achieved for patient F. An inspection of the data reveals that the recordings of patient F contain clear seizure-specific vocalizations and almost no background noise, while for example data of patient E contains lots of continuous talking and other background noise, while the annotated seizures do not have prominent seizure-specific vocalizations. Some of the particularly difficult cases also contain continuous snoring, heavy breathing, television on, or loud conversation.

V. CONCLUSIONS AND FUTURE WORK

This paper presented a novel approach for epileptic seizure detection in audio, using a CNN trained using mel-energies. The method was tested on a dataset of over 900 h of patient monitoring data. By treating all seizure vocalizations as a target event, we showed that a CNN is capable of learning and detecting seizures in unseen audio data, containing recordings of patients not encountered in training. The performance of the system is insufficient for reliable independent monitoring, but the system provides a tool that reduces significantly the monitoring effort. By tuning the seizure detection system such that it produces very small amount of false positives, the human annotation effort is directed towards verification of the events detected by the system in order to discard the false positives.

Future work includes investigation of different methods for the audio-based detection, with different neural network architectures such as convolutional recurrent neural network (CRNN) could be exploited in order to obtain a better balance between error types. In addition, combining the audio and video modalities is likely to provide additional information which could improve the overall detection performance.

REFERENCES

[1] E. E. Geertsema, R. D. Thijs, T. Gutter, B. Vledder, J. B. Arends, F. S. Leijten, G. H. Visser, and S. N. Kalitzin, "Automated video-based detection of nocturnal convulsive seizures in a residential care setting," *Epilepsia*, 2018.

[2] F. S. Leijten, D. T. Consortium, J. van Andel, C. Ungureanu, J. Arends, F. Tan, J. van Dijk, G. Petkov, S. Kalitzin, T. Gutter *et al.*, "Multimodal seizure detection: A review," *Epilepsia*, vol. 59, pp. 42–47, 2018.

[3] S. Beniczky and J. Jeppesen, "Non-electroencephalography-based seizure detection," *Current Opinion in Neurology*, Jan. 2019.

[4] F. M. Al-Hammadi, "The impact of audio classification on detecting seizures and psychogenic non-epileptic seizures," 2015.

[5] J. Arends, J. van Dorp, D. van Hoek, N. Kramer, P. van Mierlo, D. van der Vorst, and F. I. Y. Tan, "Diagnostic accuracy of audio-based seizure detection in patients with severe epilepsy and an intellectual disability," *Epilepsy Behavior*, vol. 62, pp. 180–185, 2016.

[6] G. R. de Bruijne, P. C. W. Sommen, and R. M. Aarts, "Detection of epileptic seizures through audio classification," in *4th European Conference of the International Federation for Medical and Biological Engineering*, J. Vander Sloten, P. Verdonck, M. Nyssen, and J. Haueisen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1450–1454.

[7] T. Nguyen and F. Pernkopf, "Acoustic scene classification using a convolutional neural network ensemble and nearest neighbor filters," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 34–38.

[8] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13.

[9] I.-Y. Jeong and H. Lim, "Audio tagging system using densely connected convolutional networks," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 197–201.

[10] K. Wilkinghoff, "General-purpose audio tagging by ensembling convolutional neural networks based on multiple features," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 44–48.

[11] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: task description, dataset, and baseline," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 69–73.

[12] M. Lasseck, "Acoustic bird detection with deep convolutional neural networks," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 143–147.

[13] Q. Kong, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "DCASE 2018 challenge surrey cross-task convolutional neural network baseline," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 217–221.

[14] H. Phan, M. Krawczyk-Becker, T. Gerkmann, and A. Mertins, "DNN and CNN with weighted and multi-task loss functions for audio event detection," DCASE2017 Challenge, Tech. Rep., September 2017.

[15] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

[16] S. Adavanne and T. Virtanen, "A report on sound event detection with different binaural features," *arXiv preprint arXiv:1710.02997*, 2017.

[17] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification: an overview of DCASE 2017 challenge entries," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 411–415.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[20] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016. [Online]. Available: http://www.mdpi.com/2076-3417/6/6/162