Database-driven Artifact Detection method for EEG systems with few channels (DAD)

Jamilah Abdur-Rahim*, Takeshi Ogawa*, Jun-ichiro Hirayama* Shin Ishii*†
*Dept. of Dynamic Brain Imaging, Advanced Telecommunications Research Institute International (ATR),
2-2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto, Japan.
†Dept. of System Science, Graduate School of Informatics, Kyoto University, Kyoto, Japan.
E-mail: jamilah@atr.jp, t.ogawa@atr.jp, hirayama@atr.jp, ishii@i.kyoto-u.ac.jp

***Abstract*—We demonstrate a method for identifying and removing electroencephalograph (EEG) artifacts from mobile Brain Computer Interface (BCI) data with few channels. The main components of the method entail; one-time selection of visually inspected "good quality" EEG data from past experiments and outlier detection using statistical methods. The standardly used thresholding and filtering method was compared with the DAD method on 32 datasets of real daily-life EEG data and 660 simulated EEG datasets using sensitivity and specificity measures. Both methods detected artifacts within a similar range, however the DAD method was more accurate at deciphering artifact from brain related activity within both real and simulated datasets. Additionally, the DAD method can identify specific types of artifact signatures that directly relate to behavior. The identified artifacts/behaviors include; bitting, lead artifact, electrode pop-off, electrode artifact, electrode and lead artifact, high frequency artifact, and alpha activity.**

I. Introduction

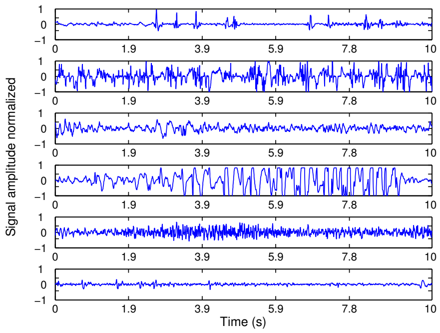
In a daily-life environment, it is challenging to record "good quality" EEG data because of contamination from physiological and/or environmental sources. Non-brain related signal patterns are often present in daily-life EEG data because the user is free to behave in a natural manner within an uncontrolled environment, as opposed to the classical laboratory setting. In addition, portable EEG devices, designed for usage in real-life environments, have limited sensors (1-20 channels) as opposed to the numerous sensors (32, 64, 128 channels) used in normal EEG experiments. Conventional artifact removal methods that employ spatial components can not be used because the number of observable sensors must be large [1]–[3]. We believe that there needs to be a more structured framework for identifying and removing artifact for EEG data collected using mobile BCI measurement devices.

We introduce and evaluate a framework for identifying and removing EEG artifacts for mobile BCI measurement devices. Our DAD method utilizes the same implementation of calculating a z-score to find statistical outliers [3], [4] to reject entire time series (datasets) or segments (epochs) of data. A novel feature of our proposed framework is that it uses previously collected EEG data from the same spatial location and from various subjects to identify erroneous datasets.

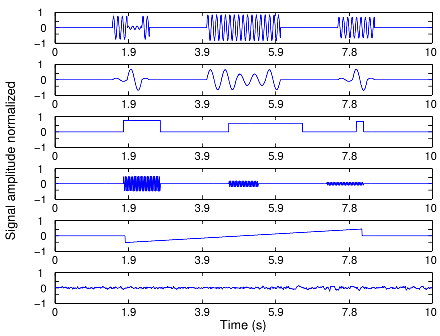
The general concept of our DAD method is to numerically define a certain range (in terms of amplitude, frequency, and signal repeatability) of what is considered "good quality" data based on past datasets. The advantages of this method are: the experimenter needs to visually inspect a hand-full of datasets one-time only such that the algorithm can use these datasets to evaluate signal quality for all future datasets, pre-processing of BCI EEG data can be performed accurately by non-experienced experimenters, the method ensures all future data will be consistent (in the same range) with past data regardless of different subjects. Other advantages of the DAD method that we demonstrate within this paper include: more accurate identification of artifacts in comparison to the standard method, improved clarity of brain-related activity, ability to identify behavior using artifact identification measures.

II. EEG data acquisition methods: real-life and simulated data

Healthy volunteers with normal corrected vision performed various daily-life behaviors over a period of 1-2 hours, after giving informed consent approved by the Ethics Committee of ATR. The participants performed the experiment in a realistic home environment (ATR's Brain Machine Interface (BMI) House) wearing a portable 2-channel EEG and 8-channel near-infrared spectroscopy measurement device (kNIRSEEG, Shimadzu Corp., Kyoto Japan). The 2-channel EEG device was positioned over central brain areas C3 and C4, both EEG channels had semi-dry electrodes that were covered with a gel-sheet for increased conductivity. EEG signals were sampled at 256 Hz and wirelessly sent to a computer for storage. The daily-life behaviors included; walking, sitting, reading a book, turning on/off the television, sweeping the floor, turning on/off the air conditioner, opening/closing the refrigerator, brushing teeth, pouring water into a cup. Sixty datasets were collected over a period of 3 months, 32 datasets were visually inspected. Figure 1(a) shows examples of six prevalent artifacts that were identified within the collected datasets. Artifacts were visually identified based on an EEG pattern atlas [5]. 660 simulated datasets, with varying length and percentage of artifact presence, were created using the existing procedure for creating realistic EEG data with artifacts [4]. The simulated datasets had varying length from 20,480 data points ($\sim$1 min) to 1,249,280 data points ($\sim$81 mins) at intervals of 20,480 data points (comparable to the daily-life datasets). Additionally, for each dataset length we varied the amount of artifacts from



(a) Examples of EEG artifacts present within the daily-life experimental datasets. From top to bottom, the artifacts include: bitting, lead (severe wire movement/broken wire), electrode (poor scalp contact, movement), electrode and lead, high frequency (electrical interference), electrode pop-off.



(b) Examples of EEG artifacts that were numerically simulated. From top to bottom, the artifacts include: transient high-frequency (temporal muscle artifacts), low-frequency (eye blink, pop-off, movement), signal discontinuities, high noise EEG artifacts, linear trends, clean EEG signal that was added to each of the five artifacts above.

Fig. 1. Types of EEG artifacts generated during the daily-life experiments (top) and numerical simulation (bottom).

0% to 100% at intervals of 10%. The type of artifact chosen and the location of each artifact was randomly determined to ensure a similar composition of each dataset. Figure 1(b) shows examples of the five types of simulated artifacts.

III. Standard and DAD methods

A. Standard method

Typical BCI EEG data preprocessing consists of amplitude thresholding and filtering, in order to remove artifact signatures. BCI2000, a standard BCI toolbox, employs filtering, normalization, and other signal processing tools to remove undesirable signal characteristics and perform online classification [2]. EEGLAB, the state of the art EEG data processing toolbox, uses filtering, mean, standard deviation, skewness, kurtosis, median, thresholding, and other signal processing tools [3], [6]. Both whole and 10 second segments of EEG datasets were evaluated in the following order: bandpass filtered between 1 Hz and 50Hz, a threshold was selected by finding 2/3 of the maximum peak for whole and segmented datasets respectively. Advantages of the standard method include: effective for extremely different behavior (resting versus movement), fast processing time (1-10 seconds), easy to implement. The disadvantages include: visual inspection of the data everytime new data is collected in order to identify a threshold, the experimenter needs to have some prior knowledge of signal quality to wisely select an appropriate threshold, not effective for various types of noisy data.

B. DAD method

EEG datasets were first bandpass filtered between 1 Hz and 50Hz. The metric that defines (an outlier) contaminated data is the z-score from ± 2 to ± 3 depending on the parameter. Figure 2 shows a diagram for each of the parameters to evaluate and corresponding z-score threshold for the DAD method.

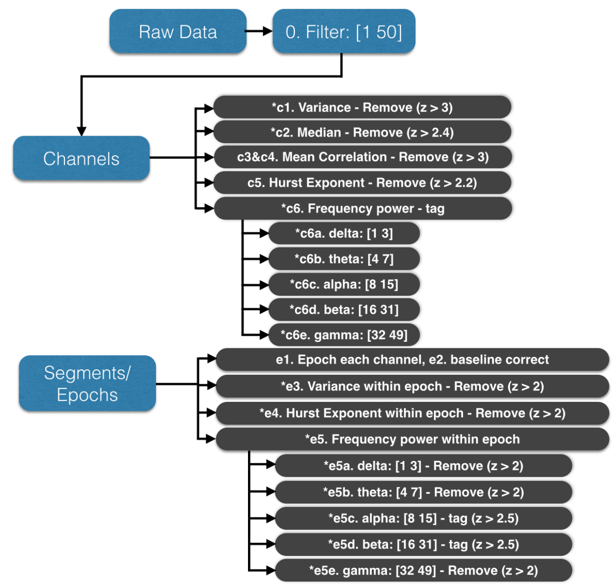


Fig. 2. Flow diagram of the DAD method based on outlier detection. "Remove" denotes that we delete the dataset or epoch if the z-score threshold is violated, "tag" denotes that we output behavioral related information (refer to section IV-B) if the z-score threshold is violated. Threshold values were originally chosen to be ± 3, based on [4], and modified via trial-and-error.

1) Whole dataset artifacts: We find that 8-16 past good quality datasets (database) compared with 1-2 test datasets produce reliable results. Each dataset varied in length, the minimum and maximum length was 4,594 and 1,300,441 data points respectively. The absolute value of the z-score was calculated for each STEP(c1-c6e) using MATLAB's standardized z-score, such that $|z| = \left|\frac{\mathbf{r}_k - \overline{\mathbf{r}_k}}{std(\mathbf{r}_k)}\right| \in \mathbb{R}^{k\times 1}$ where $k = 1, 2, \ldots, N$ and N is the total number of datasets. $\mathbf{r} \in \mathbb{R}^{k\times 1}$ represents a general vector of length k, the overbar represents the mean, and $std(\cdot)$ represents the standard deviation. A vector composed of a specific parameter (c1: variance, c2: median, c4: mean correlation, c5: hurst exponent [4], [7], c6: frequency power) for each dataset was evaluated using $|z| < \alpha$ where any kth entry greater than (α = 3, 2.4, 3, 2.2, 3) was rejected, respectively.

2) Epoched dataset artifacts: STEP e1 entails dividing each dataset into 10 seconds segments, and appending each of the 10 second segments with "good quality" data, such that the test dataset represents 40% and the "good quality" dataset represents 60% of the vector to be evaluated. In STEP e2 each 10 second epoch is baseline corrected (zero mean) to ensure we compare each epoch similarly. A vector composed of a specific parameter (e3: variance, e4: hurst exponent, e5: frequency power) for each dataset's epochs (q) was evaluated using $|z| < \alpha$ where any qth entry greater than ($\alpha = 2$, 2, 3, (2,2,2.5,2.5,2)) was rejected, respectively.

IV. Algorithm validation metrics

A. Quantification of artifact detection

The standard and DAD method were quantitatively compared with visual inspection reports of artifact contaminated epochs. We visually inspected 32 daily-life EEG datasets and reported each 10 second epoch that contained artifact contamination. Four metrics were used to quantify the accuracy of the standard and DAD methods: sensitivity, specificity, accuracy, and a count of the number of epochs that the algorithms falsely identified (over or under predicted). Sensitivity is the percentage of true artifacts that were detected, it was calculated such that Sensitivity (True Positive Rate) = TP/P. Specificity (True Negative Rate) = TN/N and Accuracy = (TP + TN)/(P + N), where P and N are the numbers of positive and negative instances, respectively. TN is the number of correct non-identified epochs, and TP is the number of correct identified epochs. The count of the wrongly identified epochs was obtained by subtracting the number of contaminated epochs reported by visual inspection from the number of identified epochs via each method [4], [8].

B. Behavior identification

Specific user behaviors, such as biting or sleepiness, produce unique EEG artifact signatures. Automatic identification of human behavior from EEG data is important because it can assist in understanding the relationship between brain activity and behavior. During artifact detection, the DAD method tags contaminated datasets or epochs with a behavior based on the combination of parameters. Biting events are tagged during epoch evaluation if the following are present: gamma frequency power with $z > 2$, variance with $z > 2$, detection of 5 or fewer significant peaks. Alpha events are tagged during both whole dataset and epoch evaluation of $z > 2.5$.

V. Results

A. Artifact detection performance in real daily-life data

Of the 32 visually inspected datasets, 14 datasets had a large amount of biologically related artifacts (EMG, user movement) and 18 datasets contained non-biologically related artifacts (lead/electrode malfunction, high frequency). We evaluated both methods using the whole dataset and then epochs. The DAD method was able to reliably identify lead artifacts via the hurst exponent in whole dataset; 50% and 27%of the 18 non-biologically classified datasets were detected using the hurst exponent and median. For epoch detection, Table I (bold faced values in Biological) shows that the standard method overly rejected data in comparison to the DAD method, as a result in many instances brain related activity was removed. In addition, the standard method could not reliably identify non-biological signals, many segments of data were purely lead or electrode artifacts as shown in Table I (bold faced values in Non-Biological).

TABLE I. Averaged artifact detection metrics for real-life data for epoch detection.

	Method	Sensitivity	Specificity	Accuracy	Over	Under
Bio	Standard	0.47	0.99	0.54	**183**	0
	DAD	0.68	0.9	0.79	34	5
Non-Bio	Standard	0.92	0.1	0.82	24	**64**
	DAD	0.96	0.1	0.98	2	0

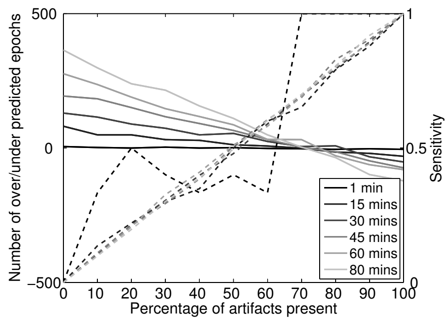
B. Simulated data

In order to clearly understand the advantages and disadvantages of both methods we varied two parameters, length and amount of artifact present, that influence the detection of artifacts. In the context of daily-life data collection, length of datasets and the amount of artifact present are two parameters that often change. Thus, it is of importance to know how these variables influence the detection of artifacts. The DAD method is indirectly dependent on dataset length due to the fact that we evaluate the whole dataset and compare z-score values for all the segments. The standard method is also indirectly dependent on dataset length, in the sense that the longer the dataset the more difficult it becomes to select a threshold that correctly characterizes specific segments of data compared to the whole. Both the DAD and standard method's ability to correctly detect artifact, directly depend on the amount of artifact present.

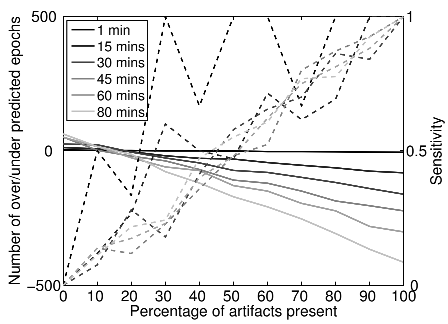
Figures 3(a) and 3(b) shows that for relatively short datasets (1-30 mins) that have 0-40% of artifact present in the data, the DAD method would be more accurate in comparison to the standard method at determining which epochs are contaminated because the standard method over predicts. If the datasets are relatively short datasets (1-30 mins) and there is more than 40% artifact present in the data, the experimenter can decide if they want to be more conservative in terms of cleaning the data and use the standard method over the DAD method (the DAD method under predicts). A disadvantage maybe that the standard method is too conservative and removes brain related data. For long datasets (30-80 mins) that have 0-40% of artifact present, the DAD method will produce more accurate results in comparison to the standard method. And, for long datasets (30-80 mins) that have more than 40% artifact present, the standard method is better at predicting.

C. Behavioral clarity of daily-life data

We verify the integrity, brain related activity, of the data after preprocessing with both methods. Both the DAD and standard method produced data that had comparable alpha activity (one-way ANalysis Of VAriance (ANOVA): $p < 0.65$).



(a) Standard method.



(b) DAD method.

Fig. 3. The epoch sensitivity (dashed lines) and number of over/under predicted (solid lines) epochs are displayed for 66 simulated datasets of varying length and presence of artifact.

The DAD method produced data that was similar to the trusted standard method, in terms of variance and mean. The alpha power amplitude was greater for data cleaned using the DAD method in comparison to the standard method. The change in amplitude from the first half of each dataset compared to the second half of each dataset was greater for data cleaned using the DAD method. This implies that the DAD method is able to remove artifact related data and still preserve brain related information to a greater degree than the standard method. The standard method has difficulty distinguishing artifact from brain activity, thus more data tends to be eliminated and brain related trends are diminished in terms of amplitude. Figure 4 shows that the DAD method removed only artifact relevant data, in comparison to the standard method, and behavior relevant trends were preserved.

VI. Conclusion

We proposed a novel database-driven method for rejecting artifactual channels or epochs for mobile EEG devices. We demonstrate the effectiveness of the framework by showing that our method more accurately predicts non-brain related artifacts than the standardly used thresholding/filtering method, for both real daily-life BCI and simulated data. We highlight advantages and disadvantages of both methods by showing when one method is more effective than another based on the length of the dataset or the amount of artifact present in the dataset. The automatic selection, based on the data within the database, of the threshold for the DAD method was shown

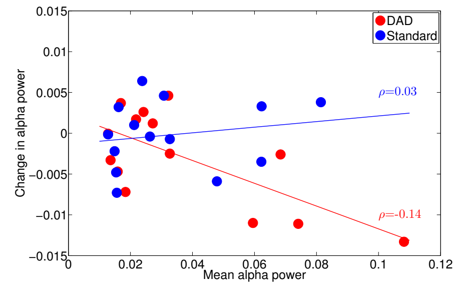


Fig. 4. The change in alpha power from the first half of each dataset to the second half of the dataset in relation to the mean alpha power. The DAD method, denoted by the red line (linearly fit to the red data points) with a slope of -0.14 ($R^2 = 0.36$), shows that several participant's brain activity in alpha band decreased during the experiment. The standard method, denoted by the blue line with a slope of 0.03 ($R^2 = 0.16$) shows no trend in brain activity.

to better identify artifact from brain related data using alpha power. The DAD method is specifically targeted to address the challenges of data collection for mobile EEG devices, such a framework would be advantageous for BCI because real-time artifact detection/removal could improve online classification.

Acknowledgement

This research was supported by a contract with the Ministry of Internal Affairs and Communications entitled, 'Novel and innovative R&D making use of brain structures'.

References

[1] R. Mahajan, D. Bansal, and S. Singh, "A real time set up for retrieval of emotional states from human neural responses," *International Journal of Medical, Health, Pharmaceutical and Biomedical Engineering*, vol. 8, 2014.

[2] G. Schalk, D. McFarland, T. Hinterberger, N. Birbaumer, and R. Wolpaw, "BCI2000: A general-purpose brain-computer interface (BCI) system," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 1034–1043, 2004.

[3] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics," *Journal of Neuroscience Methods*, vol. 134, pp. 9–21, 2004.

[4] H. Nolan, R. Whelan, and R. Reilly, "FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection," *Journal of Neuroscience Methods*, vol. 192, 2010.

[5] J. M. Stern, *Atlas of EEG Patterns*, 1st ed., J. E. Jr., Ed. 530 Walnut Street, Philadelphia, PA 19106 USA: Lippincott Williams and Wilkins, 2005.

[6] C. Brunner, A. Delorme, and S. Makeig, "Eeglab – an open source matlab toolbox for electrophysiological research," in *Biomedical Engineering / Biomedizinische Technik*. ISSN (Online) 1862-278X, 2013.

[7] S. Vorobyov and A. Cichocki, "Blind noise reduction for multisensory signals using ICA and subspace filtering, with application to EEG analysis," *Biological Cybernetics*, vol. 86, pp. 293–303, 2002.

[8] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.