Journal of Neuroscience Methods 258 (2016) 124–133

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Computational Neuroscience

Automatic artifacts and arousals detection in whole-night sleep EEG recordings

Dorothée Coppieters 't Wallant[a,b], Vincenzo Muto[a,d,e], Giulia Gaggioni[a], Mathieu Jaspar[a,d,e], Sarah L. Chellappa[a], Christelle Meyer[a,e], Gilles Vandewalle[a], Pierre Maquet[a,c,e], Christophe Phillips[a,b,*]

[a] Cyclotron Research Centre, University of Liège, Allée du 6 Août 8 B30, B-4000 Sart-Tilman, Belgium
[b] Department of Electrical Engineering and Computer Science, University of Liège, Allée de la découverte 10 B28, B-4000 Liège, Belgium
[c] Department of Neurology, University of Liège Hospital, B35, B-4000 Liège, Belgium
[d] Department of Psychology: Cognition and Behaviour, University of Liège, Place des Orateurs 2, B32B-4000 Liège, Belgium
[e] Walloon Excellence in Lifesciences and Biotechnology (WELBIO), Avenue Pasteur 6, B-1300 Wavre, Belgium

HIGHLIGHTS

- Automatic artifacts and arousals detection in EEG sleep recordings.
- Automatic detection is applied on whole night long raw multichannel data.
- Performance assessed with 60 recordings from 35 subjects, scored by 6 human raters.
- The method is robust against inter-/intra-subjects and raters variability.
- The method is fully automatic, fast and reproducible.

ARTICLE INFO

Article history:
Received 25 June 2015
Received in revised form 3 November 2015
Accepted 5 November 2015
Available online 14 November 2015

Keywords:
Artifact
Arousal
Sleep
Electroencephalography
Automatic
Adapted threshold
Raw data

ABSTRACT

Background: In sleep electroencephalographic (EEG) signals, artifacts and arousals marking are usually part of the processing. This visual inspection by a human expert has two main drawbacks: it is very time consuming and subjective.
New method: To detect artifacts and arousals in a reliable, systematic and reproducible automatic way, we developed an automatic detection based on time and frequency analysis with adapted thresholds derived from data themselves.
Results: The automatic detection performance is assessed using 5 statistic parameters, on 60 whole night sleep recordings coming from 35 healthy volunteers (male and female) aged between 19 and 26. The proposed approach proves its robustness against inter- and intra-, subjects and raters' scorings, variability. The agreement with human raters is rated overall from substantial to excellent and provides a significantly more reliable method than between human raters.
Comparison: Existing methods detect only specific artifacts or only arousals, and/or these methods are validated on short episodes of sleep recordings, making it difficult to compare with our whole night results.
Conclusion: The method works on a whole night recording and is fully automatic, reproducible, and reliable. Furthermore the implementation of the method will be made available online as open source code.

Abbreviations: EEG, electro-encephalographic data; EMG, electro-myographic data; EMG_c, electro-myographic data reconstructed from available EMG data; ECG, electro-cardiographic data; EOG, electro-oculographic data; MAS, mastoid derivation; AD, automatic detection; HR, human rater; PSD, power spectral density; 1s-epoch, epoch of 1 s duration defined from the first bin of the recording.

* Corresponding author. Tel.: +32 43662316; fax: +32 43662946.
E-mail addresses: d.coppieters@ulg.ac.be (D.C. 't Wallant), vincenzo.muto@ulg.ac.be (V. Muto), giulia.gaggioni@ulg.ac.be (G. Gaggioni), mathieu.jaspar@ulg.ac.be (M. Jaspar), sarah.chellappa@gmail.com (S.L. Chellappa), christelle.meyer@ulg.ac.be (C. Meyer), Gilles.Vandewalle@ulg.ac.be (G. Vandewalle), pmaquet@ulg.ac.be (P. Maquet), c.phillips@ulg.ac.be (C. Phillips).

http://dx.doi.org/10.1016/j.jneumeth.2015.11.005
0165-0270/

1. Introduction

One of the first steps in the analysis of a whole night sleep EEG recording consists in detecting and removing artifacts and arousals. The International Federation of Clinical Neurophysiology (IFCN) defines artifacts as non-brain activities included in EEG (Noachtar et al., 1999) whereas arousals are, according to American Academy of Sleep Medicine (AASM) manual (Iber et al., 2007), transient phenomena occurring during sleep. They can either be excluded from further analysis or considered as episodes of interest (e.g., arousal index, Pillar et al., 2002). In any case, artifacts and arousals ought to be detected before proceeding any further. Currently, this detection is typically done manually which is both time consuming and subjective (Anderer et al., 1999). Moreover, considerable disagreement exists between raters and analysis reproducibility is not guaranteed. An automatic detection method that is fast, reproducible and accurate, would usefully address these issues. Many automatic methods have been developed to detect specific artifacts (e.g. ocular artifacts (Nakamura et al., 1996; Betta et al., 2013; Gorji et al., 2013; Li et al., 2013), muscular artifacts (Brunner et al., 1996), arousals (Sugi et al., 2009), heart artifacts (Lanquart et al., 2005) and so forth). Each technique has its specific limitations: decomposition approaches (Jung et al., 2000; Vorobyov and Cichocki, 2002; Delorme et al., 2007; James and Gibson, 2003), rely on strong assumptions about the underlying signal and are extremely slow; supervised learning (Lawhern et al., 2013; Mourão-Miranda et al., 2011) needs an extensive learning data set; and simpler parametric approaches (Nolan et al., 2010) require the ad hoc definition of threshold values. Another kind of detection method uses linear models (Arnold et al., 1998; Schlögl et al., 1999; Rohalova et al., 2001). It consists in predicting EEG signal via an adaptive autoregressive model that parameters are estimated with the Kalman filter. The prediction error of the adaptive autoregressive model is used as an artifact indicator. The limitation of this method is that the model is biased by transient events. For more details, a review of artifacts processing in sleep EEG has been written by Anderer et al. (1999). Here we developed an automatic detection method for artifacts and arousals in whole night raw EEG sleep recordings. The method had to fulfill two main requirements: robustness and accuracy, i.e. regardless of the recording considered the automatic detection should be in agreement with that of manual raters. Artifacts and arousals are considered separately and are subdivided in two categories. A specific detection method has been developed for each type of artifacts considered (Section 2.4). Six recordings were used for the development and optimization of the technique. Then the whole procedure was tested over 60 different recordings under different conditions (Section 3). A key feature of our proposed method is that detection parameters are self-adjusting to the individual sleep recording processed.

2. Method

Data consist of a whole night sleep multichannel EEG recording with electrodes arranged according to the 10–20 system and including at least:

- one electromyogram, EMG, derivation for the detection of movement artifacts and arousal;
- two mastoid channels, MAS, one of which could be the actual recording reference, for finer detection of artifacts in the 'popping artifacts' module.
- one central or parietal derivation as the power spectrum density is computed from them and used as features to detect abnormal activities.

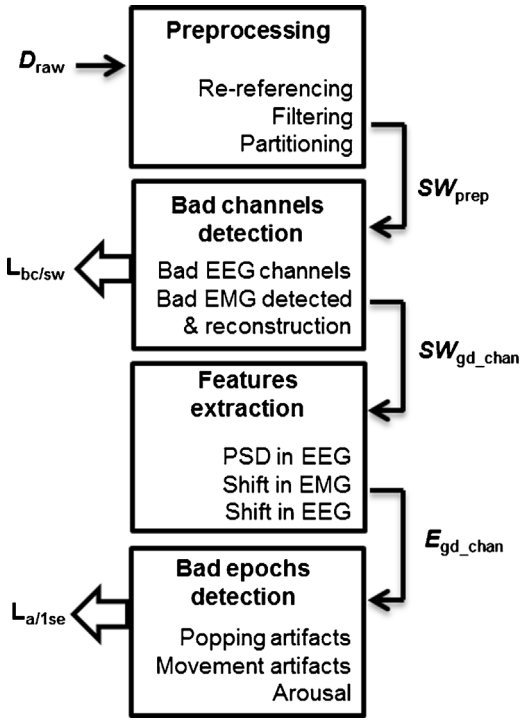

Fig. 1. Workflow of the proposed automatic detection composed of four modules. The whole raw data ('D_{raw}') are preprocessed in 'Preprocessing' and further analyzed by Scoring Window ('SW_{prep}'). Once bad channels are detected and removed from 'Bad channel detection' module, ('SW_{gd_chan}'), features are extracted by epochs of 1s in 'Feature extraction' module ('E_{gd_chan}'). In the last module 'Bad epoch detection', features are processed and bad epochs detected. The final output are lists of: bad channels per scoring window ($L_{bc/sw}$) and, artifact and arousal defined by 1s-epochs ($L_{a/1se}$)

There is no need to use a specific reference channel for the data recording.

With the standard scoring approach, the rater has to reject an entire scoring window when, for example, only a few channels are noisy. To avoid wasting potentially useful data, we decided to let the scorers mark channels or short episodes as bad. These channels (resp. short episodes) are thus not rejected for the entire recording (scoring window) but only where necessary. As suggested by Nolan in 2010 (Nolan et al., 2010), the raw data are processed through different modules that are applied successively. The modules, named 'preprocessing', 'bad channel detection', 'feature extraction' and 'bad epoch detection' are serially connected as illustrated in Fig. 1. The 'preprocessing' module consists in: (1) re-referencing data, (2) filtering data, (3) partitioning data into scoring windows – the duration of which is fixed by the scoring interval used, i.e. 30 s for AASM (Iber et al., 2007) or 20 s for Rechtschaffen and Kales (RK) (Rechtschaffen, 1968), (4) then mean correcting each channel per scoring window.

The 'bad channel detection' module works on each "scoring window". The aim is two folds: the detection of individual EEG channels the data of which are unusable for further quantitative analysis during that time window and the recovery of a good EMG channel from the EMG channels available. From this module, a first output is built and consists of a list $L_{bc/sw}$ of 'bad channels per scoring window'.

In the third module 'feature extraction', a series of features are derived from the EEG, MAS and EMG channels. The EEG features are: the power spectral density (PSD) in four frequency bands – theta (3–7 Hz), alpha (7–13 Hz), sigma (11–16 Hz) and beta (16–30 Hz), the maximal amplitude and slope. From MAS signal, the features are the maximal amplitude and slope. From the EMG, the two features are the maximal and mean amplitude of the absolute rebuilt EMG.

The 'bad epoch detection' module works on short epochs of 1 s duration (named hereafter '1s-epoch') built from all the remaining 'good' channels. The aim is to detect short periods (of one or more '1s-epochs') of artifacted data over all channels simultaneously. In this module, all features extracted in the previous module are used to detect artifacts according to their time and frequential characteristics. The second output consists of a list $L_{a/1se}$ of artifacts and arousal episodes (Fig. 1).

The automatic detection approach developed here capitalizes on previous methods: artifact detection in event-related potential (ERP) data was proposed by Nolan et al. (2010) while artifacts in sleep recordings were studied by Brunner et al. (1996), Durka et al. (2003), and Devuyst (2011). Each of these approaches focused on only a few specific types of artifacts and/or could not handle a whole night sleep recording. Our method usefully combines and adapts these techniques. We also propose a novel method for arousal detection.

By convention and for the sake of clarity, we use the following notation:

- EEG (respectively, EMG and MAS) channels are indexed with k_e (respectively, k_m and k_r);
- the scoring windows are indexed with l_w and the 1s-epochs with l_e;
- when a function is applied on data considered over time (respectively, across channels or epochs) then the subscript 't' (respectively, 'ch' or 'ep') is added to the function. For example, the mean of data is estimated over time with mean_t, across channels with mean_{ch} and across epochs with mean_{ep}.
- Mean and median values are noted respectively 'μ' and 'md'.
- Specific mathematical notation are used to differentiate scalar (s), vector ($\mathbf{V}$), set (S) and function (fct).

The modules introduced in Fig. 1 are described in the next 3 sub-sections.

2.1. Preprocessing module

This module simply prepares data for the next three modules.

2.1.1. Re-referencing

Automatic detection can be directly performed on the raw data, whatever the reference used. Nevertheless, there are two main advantages to re-reference all the channels to the average of both mastoids: it reduces the likelihood of artificially inflating activity in one hemisphere (Teplan, 2002) and electrocardiographic (EKG) artifacts (Berry and Wagner, 2015).

2.1.2. Filtering

The AASM Manual (Iber et al., 2007) suggests specifications for routinely recorded filter settings: EEG (0.3–35 Hz) and EMG (10–100 Hz). Except for the cut-off of the high-pass filter in EEG (which has been increased from 0.3 Hz to 0.5 Hz to remove slow undulations due to sweating and breathing (Devuyst et al., 2008)), we use these settings to fit best to manual scoring. Data are filtered (Butterworth filter of order 3) by a low- and high-pass forward-backward filter.

2.1.3. Partitioning

The detection of artifacts proceeds in a similar way to the manual procedure: the EEG recording is first considered in scoring windows (e.g., 20 or 30 s), then further split into (non-overlapping) 1 s epochs (1s-epoch) (Fig. 2). The 1s-epochs allow a finer detection of transient artifacts (or arousals), based on local signal features. Finally, within each scoring window, the signal of each channel is mean corrected.

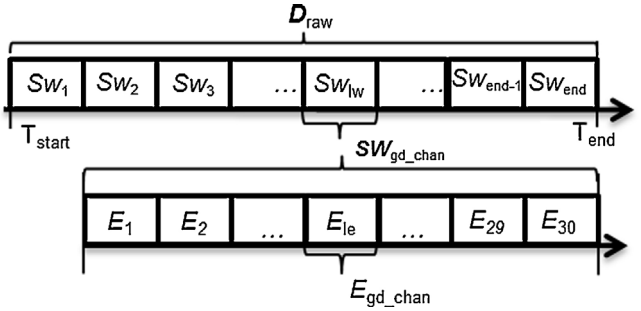

Fig. 2. The entire sleep recording is first partitioned into scoring window ('Sw_{lw}') with length depending on scoring rules used. Then, short epochs of 1 s are defined ('E_{le}') in each scoring window.

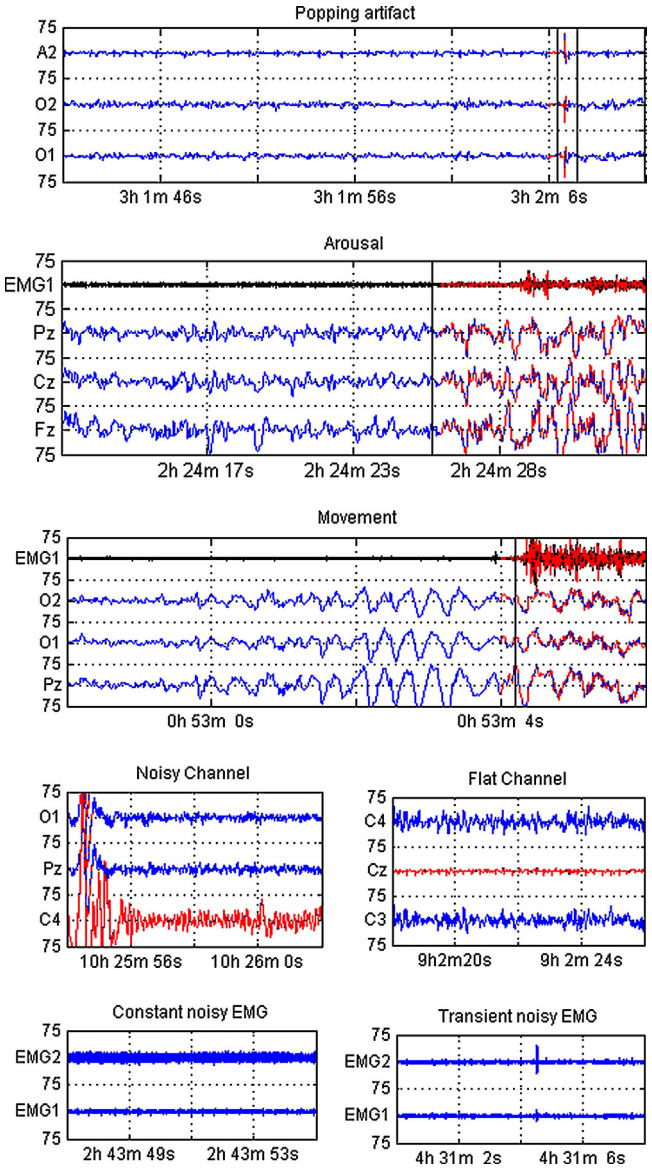

Fig. 3. Illustration of the artifacts detected by the automatic detection.

2.2. Bad channel detection

The 'bad channel detection' module is applied to both EEG and EMG channels. Two types of 'bad channels' are considered: flat and noisy channels, (Fig. 3). The noisy channels are mainly due to movement between the electrode and the skin changing the

Table 1
Values of parameters in the 'bad channel detection' module.

Name	Meaning	Value	Units
tr_n	Noisy channel threshold	6.10^3	μV
tr_{f1}	1st flat channel threshold	1	μV
tr_r	Ratio of deviation	5	–
tr_{f2}	2nd flat channel threshold	10	μV
tr_{tf}	Duration of flatness	0.8	s
tr_{fm}	Flat channel threshold	0.1	μV

voltage offset. They are recognized by their abnormally high amplitude and their high deviation from the other channels (Nolan et al., 2010). The flat channel is due to a loose electrode increasing the impedance between the skin and this electrode. Data entering in this module have been pre-processed as described in Section 2.1, ('SW_{prep}'). For EEG channels, the detection of a bad channel within a scoring window works in 3 steps: first, the detection of "obviously bad" channels, then, the finer detection of noisy and flat channels. The output of this module will thus consist in a list ($L_{bc/sw}$) of "bad channels" per scoring window.

For the detection of bad EMG channels, the procedure is similar but a last step is added: a composite EMG channel is built from the remaining (clean) EMG channels (EMG_c). This composite signal will be more reliable to detect local changes in muscular tone behavior which could be linked to movement artifacts or arousals (Sections 2.4.2 and 2.4.3).

2.2.1. Bad EEG channels

Step 1: "obviously bad" channel detection. Before detecting EEG channels presenting finer artifacts, one should first remove channels presenting clearly abnormal signals in terms of amplitude and variability over a scoring window. Therefore any channel k_e will be considered as 'bad' over the scoring window l_w if standard deviation of its signal is too large (noisy channel),

$$\mathrm{std_t}(SW_{prep}(k_e, l_w)) > \mathrm{tr_n},$$

or too small (flat channel),

$$\mathrm{std_t}(SW_{prep}(k_e, l_w)) < \mathrm{tr_{f1}}.$$

The value of these thresholds, tr_n and tr_{f1}, are summarized in Table 1. As explained in Section 3, these have been fixed empirically during the development phase of the automatic detection (AD). From here the channels tagged as 'bad' for a scoring window l_w, are excluded from any further detection steps.

Step 2: noisy channel detection. Noisy signals are induced by varying impedance between electrodes and skin. This typically leads to a signal with larger amplitude and larger variance compared with the other channels (Fig. 3). Here the standard deviation of the signal over each channel is used to detect artifacted channels. Consider the k_e^{th} channel over the l_w^{th} scoring window and calculate the mean signal of the other (not obvious bad) channels:

$$\mathbf{M}_{oth}(k_e, l_w) = \mathrm{mean_{ch}}(SW_{prep}(k, l_w)_{k=1:M/\{k_e\}}).$$

where $\mathbf{M}$ is the number of 'not obviously bad' EEG channels. The discrepancy between the k_e^{th} channel and other channels can be expressed by the ratio:

$$\mathbf{R}_{bad}(k_e, l_w) = \frac{\mathrm{std_t}(SW_{prep}(k_e, l_w))}{\mathrm{std_t}(\mathbf{M}_{oth}(l_w))}$$

If $\mathbf{R}_{bad}(k_e, l_w) > \mathrm{tr_r}$, then the k_e^{th} channel is marked as bad.

Step 3: flat channel detection. Flat channels are produced by loose electrodes. The signal has typically a very low amplitude compared to that of the other channels (Fig. 3). This is not uncommon following hours of sleep recording. According to Devuyst (2011), two parameters are derived from the data $SW_{prep}(k_e, l_w)$ from the scoring window l_w and channel k_e:

$$\mathbf{S}(k_e, l_w) = \mathrm{std_t}(SW_{prep}(k_e, l_w))$$

and

$$\mathbf{T}(k_e, l_w) = \#_t[|(SW_{prep}(k_e, l_w)| < \mathrm{tr_{amp}}] \times T_s$$

where $\mathrm{tr_{amp}} = 0.1 * \mathbf{S}(k_e, l_w)$, $\#_t[.]$ is the number of time points where the expression between brackets is verified, $|.|$ returns the absolute value of values inside, and T_s is the sampling period of the signal.

If both $\mathbf{S}(k_e, l_w) < \mathrm{tr_{f2}}$ and $\mathbf{T}(k_e, l_w) > \mathrm{tr_{f2}}$ then the k_e^{th} channel in the l_w^{th} scoring window is considered as flat channel.

Eventually, by gathering all 'bad' channels detected along the 3 steps, the list $\mathbf{L}_{bc/sw}$ of bad channels per scoring window is built and the corresponding signal is not used any more for any further processing.

2.2.2. Bad EMG detection and reconstruction

Usually EMG signal is displayed in a bipolar montage but artifacts can occur on only one of the two channels. The initial EMG step attempts to recover usable data from the available EMG channels.

Step 1: detection. From the absolute value of EMG, the median (less sensitive to outliers than the mean (Brunner et al., 1996)) is extracted over each scoring window and for each EMG channel. Flat channels correspond to those in which the median value is smaller than a fixed threshold tr_{fm}, for each EMG channel and scoring window. Following first step, two cases can occur. First, one or both, EMG channels are flat and none other test is performed until reconstruction. Second, no EMG channels is considered as flat and a second test is performed to test their relative behavior and identify potential noisy EMG channel. If the median of one channel is at least twice as big as that of the other, then the former needs further testing to distinguish constant from transient noise (Fig. 3). This test consists in decomposing the scoring window in shorter epochs (1s) and taking the ratio of absolute mean values of both EMG channels over each 1s-epoch. A 1s-epoch is considered as noisy if the ratio (suspicion of noise over clean channel values) is larger than 2 and if more than half the 1s-epochs are noisy then this channel is marked as noisy over the window, otherwise the noise is only transient.

Step 2: reconstruction. In the case where both EMG signals would be flat for more than three successive scoring windows, then the 'Movement artifact' and 'Arousal' modules are skipped. By contrast, if both EMG channels are flat for less than 3 scoring windows, we supposed transient flatness and for these scoring windows, we considered both channels as clean for the construction of the composite EMG channel (see here below). Otherwise, according to the detection done in *Step 1*, a new EMG_c is reconstructed:

- **no bad EMG channel:** EMG_c is the bipolar montage of both EMG channels:

$$\mathrm{EMG_c}(l_w) = \mathrm{diff_{ch}}(SW_{prep}(k_m, l_w));$$

- **one flat/noisy** EMG **channel over the window:** the new EMG_c is simply the only available (non-flat or non-noisy) EMG channel re-referenced to both mastoids;
- **one transiently noisy** EMG **channel:** the new EMG_c is also defined as the bipolar montage of both EMG channels for the non-noisy 1s-epochs, and as 'not available' otherwise.

With a bipolar montage, changes in muscular tone are brought out and easier to detect. With a single EMG channel referenced to the average of mastoids, muscular tone changes appear over a higher constant level activity. This additive signal does not prevent

the detection of phasic myogenic activity however the sensitivity is lower (Section 2.3.3).

2.3. Features extraction

2.3.1. Power spectrum density in EEG

For each scoring window, the PSD is computed from the average of central electrodes, where bad channels are excluded from averaging. This signal averaging leads to a smoother PSD, which provides cleaner signal features needed for further artifact detection. If all central electrodes are marked as bad, then parietal electrodes are used instead. The PSD for the central (or parietal) signal is calculated for each 1s-epoch with a multitaper approach (Thomson, 1982) and a centered rectangular 1.4s-window. Then the averaged power within each main frequency band is estimated for each 1s-epoch and stored in vectors: θ_{pow}, α_{pow}, σ_{pow}, β_{pow} for respectively the theta (3–7 Hz), alpha (7–13 Hz), sigma (11–16 Hz) and beta (16–30 Hz) bands.

2.3.2. Estimation of spindle in EEG

In the arousal's definition, Iber et al. specified that no spindle can appear (Iber et al., 2007). Therefore any 1s-epoch that could contain a spindle should be excluded from the arousal detection. Based on Devuyst et al. (2011), we define the relative power of the sigma band for each 1s-epoch as:

$$\sigma^r_{pow} = \frac{\sigma_{pow}}{\alpha_{pow} + \sigma_{pow} + \beta_{pow}}$$

Then, we estimate that the l_e^{th} 1s-epoch could contain a spindle when $\sigma^r_{pow}(l_e)$ is larger than 85% of the maximal value of σ^r_{pow} evaluated over all epochs.

2.3.3. Shift in EMG

In order to detect arousal and/or movement artifact, phasic myogenic activities are detected. They correspond to transient increases in frequency and/or magnitude in the muscular tone (EMG). Inspired by the arousal definition made by Iber et al. (2007), we used the term 'shift' to describe transient changes.

As Moretti et al. (2003), we proceed in three steps: first, along the whole EMG channel, the highest abnormal activities in EMG are rejected in comparison to an estimated baseline, second, we detect peak of activity in EMG in shorter time window to take into account the EMG background (Brunner et al., 1996), and third, we determine which peaks are meaningful to consider influence in the EEG by checking the intensity and the duration of these latter. The only parameter used in these steps is the maximal amplitude over the absolute values in EMG. This parameter is evaluated for every 1s-epoch along the whole recording and values saved as a vector $\mathbf{mx_{EMG}}$. For the last step, two others parameters are also needed.

The muscular tone amplitude decreases gradually from wakefulness to REM sleep, through the different sleep stages: N1, N2 and N3 (Iber et al., 2007). Based on this, we define the EMG baseline, or the higher muscular tone amplitude without artifact, from the first few minutes of recording. Indeed, we assume that the subject is falling asleep and that the muscular tone is still large compared to the remaining recording.

In order to estimate the EMG baseline, we select within the first four minutes of recording, two minutes of scoring windows with an EMG_c consisting in a bipolar EMG montage (Section 2.2.2). If there are not two minutes of scoring windows with EMG_c built from bipolar montage, then the first 2 min of EMG_c are taken. From these 2 min of recording, a uniform set of $\mathbf{mx_{EMG}}$ is selected using an absolute z-score transform lower than 3. From these selected $\mathbf{mx_{EMG}}$ values, are calculated the mean, μ_{EMG}, and standard deviation, sd_{EMG}. The first threshold, tr_{sup}, is defined as:

$$tr_{sup} = \mu_{EMG} + 2 \times sd_{EMG}(l_w, l_s))$$

Over the whole EMG recording, are considered as high muscular tone, $high_{EMG}$, all 1s-epochs with their $\mathbf{mx_{EMG}}$ value larger than tr_{sup}.

For the second step, we analyze EMG in shorter time window to take into account the EMG background. For each scoring window, 'l_w' we define a specific threshold, $tr_{sw}(l_w)$, as being the output of a median filter (Gallagher and Wise, 1981) applied on a symmetric-centered 3s-scoring window. All 1s-epochs composing a scoring window l_w with its $\mathbf{mx_{EMG}}$ value larger than $tr_{sw}(l_w)$ are noted: $\mathbf{loc_{EMG}}$.

For the last step, all 1s-epochs listed in, $\mathbf{high_{EMG}}$ and $\mathbf{loc_{EMG}}$ are pulled together: $\mathbf{sh_{EMG}}$. The last step consists in determining if the increase in the muscular tone is sufficient to influence the EEG. Two parameters are necessary: the intensity, $\mathbf{i_{EMG}}$ and the duration, $\mathbf{d_{EMG}}$, of the muscular tone change corresponding to the $\mathbf{sh_{EMG}}$ epochs. A specific threshold is defined for each $\mathbf{sh_{EMG}}$ epoch. From both sides of each $\mathbf{sh_{EMG}}$, the first ten available (non-listed) 1s-epochs are selected to compute the mean of their $\mathbf{mx_{EMG}}$ value: tr_{se}. Over the absolute value of $\mathbf{sh_{EMG}}$, we count the number of samples larger than tr_{se} and save them in a vector **nb**. The intensity of the muscular tone change, i_{EMG}, is estimated as the number of samples in **nb** divided by the number of samples separating the first and the last sample in **nb**. The duration, d_{EMG}, is the number of samples in **nb** divided by the sample frequency. Finally, a $\mathbf{sh_{EMG}}$ period is considered as meaningful if its relative i_{EMG} and d_{EMG} are respectively larger than tr_i and tr_d.

2.3.4. Shift in EEG

A 'shift in EEG frequency' is defined in the arousal's definition of Iber et al. (2007) as a phasic electroencephalographic activity corresponding to a transient change in EEG frequency. To consider a shift in EEG, two tests are performed over the power in the three frequency bands: β_{pow}, α_{pow} and θ_{pow}. The first test uses a fixed threshold, to detect abnormal activity in EEG relatively to the whole recording, and the second test, an adapted threshold to take into account the specific background of the EEG in shorter time window (plus 10s on each side to ensure continuity across scoring window).

For the beta band (respectively alpha and theta), the fixed threshold is evaluated as the median value of the whole vector β_{pow} (respectively α_{pow} and θ_{pow}): mdt_β (respectively mdt_α and mdt_θ) across the whole recording.

Then, for each scoring window, a more specific threshold is evaluated. First we select all 1s-epochs without corresponding shift in EMG plus the first ten 1s-epochs without shift in EMG from each side of this scoring window. For the beta band (respectively alpha and theta), the adapted threshold for the k^{th} scoring window is the median value over the β_{pow} (respectively, α_{pow} and θ_{pow}) for the selected 1s-epochs: mde_β (respectively mde_α and mde_θ).

Finally, all the 1s-epochs composing the k^{th} scoring window with their power in the beta band (respectively alpha and theta band) larger than two times mde_β (respectively mde_α and mde_θ), and larger than the median value of the whole vector mdt_β (respectively mdt_α and mdt_θ) are considered as shift in EEG.

2.4. Bad epoch detection

A general description of artifacts could be a modification in EEG coming from extracerebral sources (Noachtar et al., 1999). To precisely describe them, several types of artifacts have been defined. We particularly focus on three of them which have a strong impact on the qualitative and quantitative sleep analysis: 'popping

artifacts', 'Movement artifacts' and 'Arousals' (Fig. 3). In practice, within each scoring window, the data will be analyzed in short 1s-epoch across all channels. The aim is thus to list 1s-epochs that should be discarded from any further analysis. This detection will proceed in three different steps, related to the type of artifact detection 'Popping artifact', 'Movement artifact' and 'Arousal' as described in the next subsections. After these three modules, the artifacted 1s-epochs are pooled together. A short (less than 3s) episode of signal surrounded by artifacted episodes will still be considered artifacted. The reason is two folds: 3 s or less of isolated clean signal are not very useful for sleep analysis and are, in this case, more likely a missed artifact.

2.4.1. Popping artifact

The popping artifacts are rapid transitions that would originate from a sudden change in the voltage offset and are sometimes only visible on a few electrodes. They can be classified as 'technical' artifact because of their origin (due to the recording equipment) (Klass, 1995). This typical artifact is represented in Fig. 3.

The term 'popping' comes from Durka et al. (2003) and Devuyst (2011). The latter relied on the standard deviation, while Durka et al. (2003) focused on fast changes of large amplitude. Here we develop an approach similar to that of Durka et al., using the slope and its relative amplitude over 0.5s windows. To detect sudden changes in EEG and MAS signals, two parameters are used: a slope index (the ratio between the maximal and minimal amplitude and the time in between these extrema) and the maximal amplitude, within non-overlapping 0.5s-epoch. This shorter (than previously used 1s-epoch) time window ensures capturing the fast transient spikes in the signal. In EEG, for each 0.5s-epoch, only the largest slope index amongst slopes of all EEG channels, is kept with its corresponding maximal amplitude. A popping artifact is detected in EEG when the slope $\mathbf{sl}_e$ is larger than tr_{sle} with either a maximal amplitude at tr_{ae} or a maximal time between the extrema at tr_{te}. Similarly to EEG, a popping artifact is detected in MAS when the slope $\mathbf{sl_r}$ is larger than tr_{sr1} with an amplitude at least at tr_{ar1} or if $\mathbf{sl_r}$ is larger than tr_{sr2} with an amplitude at least at tr_{ar2}. All the 1s-epochs that are positively tested by these conditions are considered as popping artifact. The thresholding values are summarized in Table 2.

The EKG artifacts come from cardiac potentials on the body surface and are visible on EEG channels during either short or long periods. These artifacts are similarly to popping artifact defined as sharp deflection in EEG signals. Reduced with the re-referencing on both mastoid (Berry and Wagner, 2015), the remaining EKG artifacts still visible on EEG after re-referencing are detected with the popping detection method.

2.4.2. Movement artifact

The method for movement artifact detection has been mainly inspired by Pilcher and Schulz (1987) and Brunner et al. (1996). We firstly notice in EMG signal, abnormal increase which could influence the EEG (Section 2.3) (Pilcher and Schulz, 1987). Then, we investigate the EEG activity from the power in the beta band: $\boldsymbol{\beta}_{pow}$ (Brunner et al., 1996). To observe even weak increase in $\boldsymbol{\beta}_{pow}$, we use an adapted threshold calculated on a time window centered on the EMG shift.

For each EMG shift event, (consecutive 1s-epochs with EMG shift), we select the first 10s available (non shifted EMG) from both sides of the event (surrounded by an error edge of 3s). From these twenty 1s-epochs, the median of $\boldsymbol{\beta}_{pow}$ in EEG signal is calculated to create the adapted threshold: md_{out}. All 1s-epochs whose its power in the beta band are larger than 3 times md_{out} or more were considered as artifacted by movement.

2.4.3. Arousal

To detect arousal, in contrast to movement detection process, we first detect shift in EEG and then, we look at the EMG signal to note if there is an increase in the muscular tone (Pilcher and Schulz, 1987). The increase in the muscular tone can be subtle and the shift in EEG can be composed of alpha, beta and theta frequencies. From EEG, we select shift events (successive 1s-epochs) which last more than 3s. Then, for each event selected, we look at the muscular tone to observe if there is even a subtle increase. From the maximal amplitude over the absolute values in EMG, $\mathbf{mx_{EMG}}$, we select values corresponding to the shift in EEG. From the latter, the mean muscular tone value: μ_{in} is assessed. Finally, μ_{in} is compared to an adapted threshold μ_{out}. This threshold is calculated as follow: the average of $\mathbf{mx_{EMG}}$ corresponding to the first 10s (non-EMG shift) from both sides of the event (surrounded by an error edge of 3s). The EEG shift events are considered as arousal if their relative μ_{in} value is at least 1.5 times larger than the surrounded muscular tone, μ_{out}.

Table 2
Values of parameters in the 'bad epoch detection' module.

Name	Meaning	Value	Units
tr_{sle}	Minimal slope value in EEG relative to tr_{ae}	3×10^3	μV/s
tr_{ae}	Minimal amplitude associated with $\mathbf{sl_e}$	120	μV
tr_{te}	Maximal time associated with $\mathbf{sl_e}$	0.03	s
tr_{sr1}	Minimal slope coefficient associated to tr_{ar1}	10^3	μV/s
tr_{ar1}	Minimal amplitude associated to tr_{sr1}	20	μV
tr_{sr2}	Minimal slope coefficient associated to tr_{ar2}	200	μV/s
tr_{ar2}	Minimal amplitude associated to tr_{sr2}	80	μV
tr_{time}	Minimal time to consider muscular tone event	0.05	s

3. Application

We start by describing the data we used. Then, we introduce the criteria and methods employed to compare our automatic procedure with the manual detection performed by experts.

3.1. Analyze processing

3.1.1. Data

Four datasets were used to develop ('Data 0') and evaluate ('Data I–III') the artifact automatic detection (AD) (Table 3). Amongst these datasets, four types of sleep recordings (baseline (B), standard (S), extension (E) and recovery (R) nights) were obtained from 35 healthy young volunteers: 6 in 'Data 0' (5 male, 21 ± 1 years old), 22 in 'Data I' (10 male, 21 ± 1 years old), 1 in 'Data II' (1 male aged 26) and 6 in 'Data III' (all male, 22 ± 2 years old). All these young healthy participants followed a strict sleep-wake schedule (regular bedtime and caffeine-free diet) during three weeks preceding the data recordings. Data consist of continuous recording (500 Hz sampling rate) over 9 EEG (F3-Fz-F4-C3-Cz-C4-Pz-O1-O2), 1 MAS ('A1'), 2 EOG and 2 EMG channels. All channels were initially referenced to the right mastoid 'A2' before a re-referencing to the average of both mastoids. All acquisitions were performed at the Cyclotron Research Centre (University of Liège, Belgium) in the framework of a larger study (not yet published) which was approved by the local Ethics committee and for which participants gave their written informed consent. Subjects took part either to

Table 3
Types of nights and human raters of the four datasets used.

Dataset	Sleep recordings	Human raters
Data 0	4B, 5R	VM
Data I	20B, 12R	VM, GG
Data II	1B, 1E, 1S, 1R	GV, CM, SC, VM, MJ, GG
Data III	6B, 6E, 6S, 6R	GV, CM, VM, MJ, GG

B, baseline night; E: extension night; S, standard night; R, recovery night.

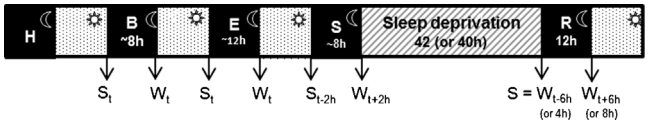

Fig. 4. Protocol 1 and 2 differentiated by the sleep deprived duration (42 h and 40 h). Wt and St are respectively the wake time and the sleep time specific to the volunteer. Abbreviations used are: H, habituation night; B, baseline night; E, extension night; S, standard night; R, recovery night.

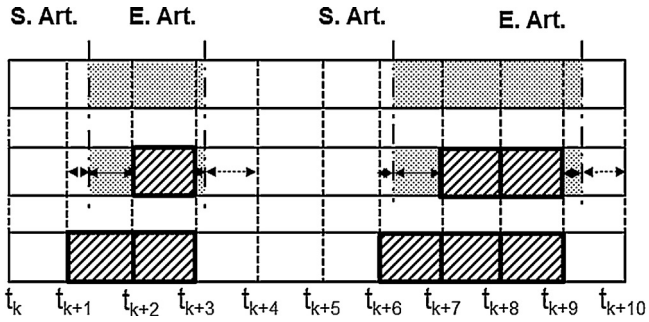

Fig. 5. Conversion from sample-precision to 1s-precision in HR's detection.

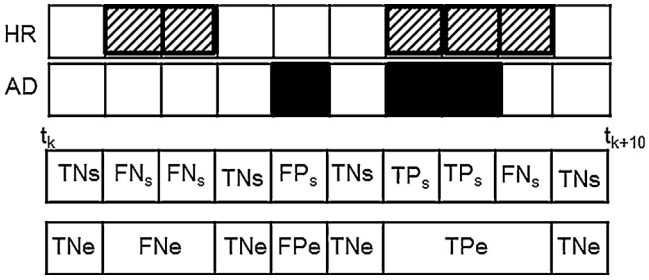

Fig. 6. Two kinds of matches according to the unit of comparison chosen: 1s-epoch or event. The two upper lines represent the scorings of a human rater (HR) and the automatic detection (AD) and the two lower lines, the two possible comparisons. The first comparison matches successive 1s-epochs whereas the second, events formed by successive 1s-epochs.

experimental protocol P1 ('Data 0' and 'Data I') or P2 ('Data II' and 'Data III') (Fig. 4). Each recording was visually inspected for artifact and arousal rejection by 1 to 6 human raters (VM, GG, GV, SC, CM, and MJ). Because artifacts were not removed by human raters in sleep stages W and N1, we compared artifacts detected only on stages: N2, N3 and REM. That means that we limited our evaluation on these stages by discarding approximately 15 min from each recording.

3.1.2. Matching

To evaluate the performance of AD, we compare AD results to artifact rejection by human raters (HR). The comparison consists in matching both sleep recordings to observe agreements and calculate performance statistics. To match results from AD and HR, we converted scorings to the same temporal resolution. HR marks artifacts (and arousals) with the same temporal resolution as the signal, using start/end markers (respectively 'S. Art' and 'E. Art' in Fig. 5). On the other hand, AD detects the artifacts/arousals by 1s-epochs.

To make the detection comparable, we thus convert the HR's marks into 1s-epoch marks. Any 1s-epoch covering more than 0.5 s of HR marked artifacts is marked as artifacted, as illustrated in Fig. 5.

Comparison between HR and AD can be performed in terms of each individual 1s-epoch ('s') but also in terms of 'artifact episode' ('e') i.e. set of consecutively marked 1s-epoch. For each comparison, one of the four matches (with HR as gold standard) is considered (Fig. 6):

Table 4
Cohen's Kappa values and interpretation.

κ	Interpretation
0.00–0.20	Poor
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1.00	Very good

- True positive, TP_s or TP_e: AD and HR agree on the presence of an artifact within the 1s-epoch or 'artifact episode';
- True negative, TN_s or TN_e: AD and HR agree on the absence of artifact for this 1s-epoch or this episode;
- False positive, FP_s or FP_e: AD finds an artifact whereas HR finds none in a given 1s-epoch or episode;
- False negative, FN_s or FN_e: AD finds no artefact whereas HR finds one in a given 1s-epoch or episode.

3.1.3. Statistic parameters

There are two general ways to evaluate the inter-rater reliability: the percent of agreement and the correlation (Feng, 2013). We chose the percentage of agreement because it is an easier concept to visualize concordance, good or bad, between two scorings. Among all coefficients proposed in literature (Scott, 1955; Krippendorff, 1970; Perreault and Leigh, 1989; Gwet, 2002, 2008), we chose the inter-rater agreement S (Bennett et al., 1954) (Eq. (1)) and, the Cohen's Kappa κ, (Cohen, 1960) (Eq. (2)). S does not account for the unbalanced scores to compare (there are many more unartifacted 1s-epochs than artifacted ones) and it generally overestimates the inter-rater reliability. It is nevertheless one of the most commonly used parameter. On the contrary Cohen's Kappa accounts for the expected number of random matches for unbalanced data but is often deemed over conservative.

$$S_s = 2 \times P_0 - 1 \quad \text{with} \quad P_0 = \frac{TP_s + TN_s}{TP_s + TN_s + FP_s + FN} \tag{1}$$

$$\kappa_s = \frac{P_0 - P_r}{1 - P_r} \quad \text{with} \quad P_r = \frac{(TP_s + TN_s) \times (TP_s + FP_s)}{(TP_s + TN_s + FP_s + FN)^2} \tag{2}$$

In order to interpret κ, Landis et al. provided Table 4 (Landis and Koch, 1977).

The proportion of 1s-epochs with artifact, relative, to those without, is always around 5% for all recordings, which is very unbalanced. The inter-rater parameter S_s will thus be overestimated whereas the Kappa of Cohen κ_s will be very strict. S_s and κ_s reveal the agreement between two raters, HR and AD, at a 1s-epoch resolution. The agreement could also be performed at the level of artifacted episodes and we therefore introduce 3 other evaluation criteria: the sensitivity in terms of the number of artifacted episodes (Se_e) (Eq. (3)), the averaged overlap of the events detected (C_s) and the false discovery ratio in terms of events (FDR_e) (Eq. (4)). The false discovery ratio is the proportion of spurious detections to the total number of detections. The FDR is also equal to 1-PPV where PPV is the "Positive Predicted Value" or the proportion of positive results versus the total number of events detected. C_s is the averaged proportion of TP_s contained in an true positive event TP_e (delimited by the HR's detection).

$$Se_e = \frac{TP_e}{TP_e + FN_e} \tag{3}$$

$$FDR_e = \frac{FP_e}{TP_e + FP_e} \tag{4}$$

3.2. Results

As summarized in Table 3, we used four datasets for the development (Data 0) and evaluation of the automatic detection (Data

Table 5
One phase to develop AD ('Phase 0') and three phases to evaluate its performance ('Phase I–Phase III') with five statistic parameters: S_s, κ_s, Se_s, C_s, FDR_e. For the second and the fourth phase, there are two evaluations of these statistic parameters: before and after the review of false detections proceeded by AD.

Phase	GS	S_s Mean ± SD	κ_s Mean ± SD	Se_e Mean ± SD	C_s Mean ± SD	FDR_e Mean ± SD
Phase 0	HR	97 ± 1%	0.70 ± 0.15	87 ± 5%	74 ± 10%	39 ± 17%
Phase Ia	HR	96 ± 2%	0.65 ± 0.10	83 ± 9%	70 ± 7%	43 ± 14%
Phase Ib	HR + review	96 ± 3%	0.72 ± 0.09	85 ± 8%	74 ± 8%	22 ± 10%
Phase II	∪ 6 HR	98 ± 0%	0.50 ± 0.07	61 ± 6%	58 ± 7%	40 ± 7%
Phase IIIa	5 HR	95 ± 4%	0.55 ± 0.12	71 ± 11%	68 ± 14%	39 ± 13%
Phase IIIb	5 HR + review	96 ± 4%	0.63 ± 0.13	76 ± 10%	71 ± 12%	24 ± 13%

GS, gold standard; HR, human rater; ∪, means 'union of'.

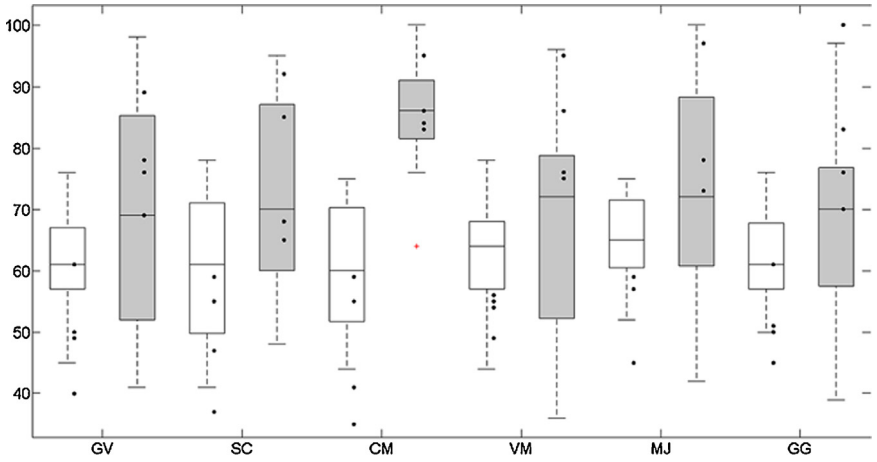

Fig. 7. Phase II: Cohen's Kappa (white boxes) and sensitivity in term of detected artifacted periods (grey boxes) are used to compare the ratings between human raters and AD. Both parameters are evaluated for the automatic detection (black dot) and for all but one human raters (boxplot) with as gold standard the left out human raters. The central mark in boxes indicates the median of statistic parameters for human raters with respectively for the bottom and upper edges' boxes, the 25th and 75th percentiles. The whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually with a plus symbol, '+'.

I, II and III). The AD performance has been evaluated in three phases to evaluate its robustness against: the intra- and inter-subject variability (Phase I), the intra- and inter-rater variability (Phase II) and both intra- and inter-subject and rater, variability (Phase III).

3.2.1. Phase 0: Development

The automatic detection has been developed from the dataset 'Data 0' (Table 3) and the parameters used to define the adaptive thresholds (Tables 1 and 2) have been fixed for all the testing phases. The performance for this dataset is shown on the first line in Table 5. These values were calculated by taking the human rater VM as the gold standard over 9 different sleep recordings.

3.2.2. Phase I: Evaluation of the automatic detection relative to intra- and inter- subject variation

The objective of the first phase was to evaluate the robustness of AD against inter- and intra- subject variabilities. For a same subject and over different sleep contexts (baseline and recovery nights) or for different subjects over a same sleep context, the sleep characteristics vary considerably. The evaluation of AD in this phase has been processed in two steps. Firstly, the artifacts detected by AD were compared with the ones detected by VM considered as the gold standard. Performance obtained was sufficient to confirm that AD was robust against the inter- and intra- subject variability (Table 5, Phase Ia). Secondly, the false detections made by AD have been reviewed by a second rater to check if there were oversights. The false discovery ratio (FDR_e) decreased and the sensitivity increased (Se_e) (Table 5, Phase Ib)

3.2.3. Phase II: Evaluation of the automatic detection relative to intra- and inter- rater variation

The dataset 'Data II' was used to assess the robustness of AD with respect to intra- and inter-rater variability. There is no unique gold standard in this phase. All 7 raters, six humans (Table 3) plus AD, are compared against each other over the 4 sleep recordings, leading to 24 values of S_s, κ_s, Se_e, C_e and FDR_e. The values of Cohen's Kappa, κ_e and sensitivity, Se_e, for each rater are shown in Fig. 7. Overall there is about the same inter-rater variability between humans and AD. The statistic parameters obtained when the union of all the human raters is used as a gold standard are summarized in Table 5. From the observed false detections, we noticed that arousal detection is dependent on the personal interpretation of what is an 'abrupt shift in EEG'. Similarly, sleep patterns (i.e. sleep spindles or vertex sharp) can also be confused with artifacts.

3.2.4. Phase III: Evaluation of the automatic detection relative to intra- and inter- subject and rater variation

Here, AD is assessed against sleep recordings from different subjects and marked by different raters (Table 3). The gold standard in this phase has been considered as the human rater in charge of the sleep recording tested. The mean of the statistic parameters are shown in Table 5.

4. Discussion

We developed a new method for automatic artifact detection in whole night sleep recordings. The automatic detection (AD) provides results from raw data, is not supervised by human raters and does not need training phase. The performance of AD has been

evaluated in three phases to test its robustness against inter- and intra-, subjects and human raters' scorings, variabilities. In the first phase, 'Phase I', the evaluation of AD is performed over 32 sleep recordings from 6 different subjects, ('Data I' in Table 3), with as gold standard (GS) a unique human rater. The aim of this phase was to test the robustness of AD over multiple recordings from several subjects, i.e. in face of inter- and intra-subject variability. As shown in Table 5, the agreement between the human rater and AD is tested in 2 steps. When the 2 ratings are totally independent, they already matched very well overall: the interpretation of Cohen's Kappa coefficient is 'good agreement', sensitivity is high (83% of artifacted episodes detected) and there is a good overlap (70%) between the GS and AD for these episodes. Each of these statistic parameters has a small standard deviation across all recordings proving the reliability and the robustness of AD. Nevertheless, the false discovery ratio, FDR_e, is relatively high. In the 2nd step, the human rater used the AD results to refine his detection: he reviewed the 'false detections' of AD and checked if these were truly false detections or oversights. Since the GS was updated (artifacted episodes were added) the statistic parameters were reevaluated: as expected, the FDR_e decreased (from 43% to 22%) with in parallel an increase in sensitivity (from 83% to 85%) and a better κ_s (from 65% to 72%). 'Phase I' brings out that the automatic detection is robust against inter- and intra- subject variabilities and secondly, human raters made oversights. The stringency of AD to detect artifacted periods justifies the large FDR_e compared to human raters whose performance are neither systematic nor reproducible.

With the second phase of the evaluation, 'Phase II', we demonstrated that AD maintained sufficient performance against several raters' scorings. Despite a moderate Cohen's Kappa value, AD sensitivity is similar to, or better than, that of each human rater against each other (Fig. 7). As all human raters disagree to some extent with AD on what are false detections, we can conclude that each of them made different oversights. These oversights significantly influence the Cohen's Kappa.

Finally, over the last phase of evaluation, 'Phase III', the automatic detection also displayed a good average Cohen's Kappa value and a substantial sensitivity for a mixed dataset with several raters and different sleep contexts. Moreover after manually reviewing the false detections made by AD, all the statistics improved. These results indicate again that AD picks up artifacts that were overlooked by the human raters. In practice there is unfortunately no reliable gold standard to evaluate automatic detection. The presence of oversights in all sleep recordings scored proves the weak reliability and the lack of reproducibility of human raters. This stems from the absence of a clear mathematical definition of artifacts as 'abrupt shift' in EEG. Furthermore, the erroneous detections made by AD, i.e. which would not be considered as artifacts by human raters, are based on abnormal power spectral activities or sudden changes in voltage. These are not easily discernable visually whereas their influence on quantitative analysis could be harmful. Overall the quantity of false detections made by AD and removed from data, compared to the entire recording, is less than a few percent of the whole recording.

Our algorithm is validated for multiple recordings from several subjects sleeping in different contexts (baseline, recovery, extended and standard nights). Admittedly this method is only validated on young healthy people and over scoring windows corresponding to whole night sleep episodes. Wakefulness was not considered by the human raters, we therefore have no comparison points. Applying our algorithm onto different data sets (different population and wakefulness to begin with) will be the topic of further validation. We also hope that making our code open source will encourage other teams to further validate it.

It also would have been interesting to compare our proposed automatic detection method with others but this simply is not possible because of their characteristics and application field: Devuyst et al. (2008) tested short recordings lasting about fifteen minutes and Durka et al. (2003) used 4s-epoch as unit of comparison to evaluate the method's performance. The others automatic detection methods available in the literature (Lawhern et al., 2013; Nolan et al., 2010), are applicable on non-continuous data or on evoked potential responses. With the three phases of evaluation that we described here above, we have been able to show that the automatic detection is robust over recordings from different subjects and environments and its performance is reproducible against different raters.

Nevertheless, we have also stated that with additional information as sleep patterns, we will be able to improve it. Indeed, with sleep patterns (i.e. spindles, alpha waves or slow waves) we could make assumptions on sleep stages states and increase the accuracy for arousals detection which differs from REM to NREM sleep (Iber et al., 2007). Moreover, we know that spindles may sometimes be confused with artifacts by our method and that with an iterative approach, mixing artifact and patterns detection, we could obtain even better results.

5. Conclusion

The objective was to create a fast method to help raters clean their data, in a significantly shorter time framework. Without any recording specific configuration from the users, the automatic detection is able to detect artifacts in 8–12 h continuous sleep recordings over multiple sleep contexts and with scoring window of 20s or 30s. The detection provided by AD can be considered as final or as a first detection step helping different raters to reach an agreement. In general this automatic detection should be applied as the first processing step of polysomnographic data.

Overall, the automatic detection is fast, reliable, systematic and reproducible. It allows removing undulations, to detect bad channels inside each scoring window at a time, and to detect artifacts by epochs of 1s over all channels. With a standard desktop computer with current architecture, all these operations are performed for a single whole night recording ($\pm$8 h) in less than 40 min.

5.1. Practical information

FAST is a M/EEG toolbox developed by researchers from the Cyclotron Research Centre, University of Liège, Belgium, with the financial support of the Fonds de la Recherche Scientifique-FNRS, the Queen Elizabeth's funding, and the University of Liège (Leclercq et al., 2011). The CRC thinks that free accessibility to that software boosts improvements and increases its reliability (Hayden, 2015).

The software with its manual are available at http://www.montefiore.ulg.ac.be/ phillips/FASST.html.

Acknowledgements

Research supported by the FRS-FNRS (Fonds National de la Recherche Scientifique), FRIA, University of Liège, ARC (Action de Recherches Concertée), FMRE (Fondation médicale Reine Elisabeth), WELBIO (Walloon excellence in life sciences and biotechnology), FEDER (Fonds européen de développement économique et régional), fondation Simone et Pierre Clerdent, Fond Léon Fredericq, WBI (Wallonie-Bruxelles International), Bial Foundation. Thanks to the "sleep group" at the Cyclotron Research Centre for their support.

References

Anderer P, Roberts S, Schlögl A, Gruber G, Klösch G, Herrmann W, Rappelsberger P, Filz O, Barbanoj MJ, Dorffner G, Saletu B. Artifact processing in

computerized analysis of sleep EEG - a review. Neuropsychobiology 1999];40:150–7.
Arnold M, Miltner WHR, Witte H, Bauer R, Braun C. Adaptive AR modeling of non-stationary time series by means of Kalman filtering. IEEE Trans Biomed Eng 1998];45:553–62.
Bennett EM, Alpert R, Goldstein AC. Communications through limited response questioning. Public Opin Q 1954];18:303–8.
Berry RB, Wagner MH. Sleep Medicine Pearls. third ed. Elsevier Saunders; 2015].
Betta M, Gemignani A, Landi A, Laurino M, Piaggi P, Menicucci D. Detection and removal of ocular artifacts from EEG signals for an automated REM sleep analysis. In: Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE; 2013]. p. 5079–82.
Brunner DP, Vasko RC, Detka CS, Monahan JP, Reynolds CF, Kupfer DJ. Muscle artifacts in the sleep EEG: automated detection and effect on all-night EEG power spectra. J Sleep Res 1996];5:155–64.
Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measur 1960];20:37–46.
Delorme A, Sejnowski T, Makeig S. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. NeuroImage 2007];34:1443–9.
Devuyst S. (PhD thesis) Analyse automatique de la macrostructure du sommeil (PhD thesis). Université de Mons; 2011].
Devuyst S, Dutoit T, Stenuit P, Kerkhofs M. Automatic sleep spindles detection – overview and development of a standard proposal assessment method. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC; 2011]. p. 1713–6.
Devuyst S, Dutoit T, Stenuit P, Kerkhofs M, Stanus E. Removal of ECG artifacts from EEG using a modified independent component analysis approach. Conf Proc IEEE Eng Med Biol Soc 2008];2008:5204–7.
Durka P, Klekowicz H, Blinowska K, Szelenberger W, Niemcewicz S. A simple system for detection of EEG artifacts in polysomnographic recordings. IEEE Trans Biomed Eng 2003];50:526–8.
Feng G. Factors affecting intercoder reliability: a Monte Carlo experiment. Qual Quant 2013];47:2959–82.
Gallagher N, Wise G. A theoretical analysis of the properties of median filters. IEEE Trans Acoust Speech Signal Process 1981];29:1136–41.
Gorji HT, Koohpayezadeh A, Haddadnia J. Ocular artifact detection and removing from EEG by wavelet families: a comparative study. J Am Water Works Assoc Inform Eng Appl 2013];3:39–48.
Gwet KL. Inter rater reliability: dependency on trait prevalence and marginal homogeneity; 2002]. p. 2.
Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008];61:29–48.
Hayden EC. Rule rewrite aims to clean up scientific software. Nature 2015];520:276–7.
Iber C, Ancoli-Israel S, Chesson A, Quan S. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. American Academy of Sleep Medicine; 2007].
James C, Gibson O. Temporally constrained ICA: an application to artifact rejection in electromagnetic brain signal analysis. IEEE Trans Biomed Eng 2003];50:1108–16.
Jung TP, Makeig S, Humphries C, Lee TW, McKeown MJ, Iragui V, Sejnowski TJ. Removing electroencephalographic artifacts by blind source separation. Psychophysiology 2000];37:163–78.
Klass D. The continuing challenge of a artifact in the EEG. Am J EEG Technol 1995].
Krippendorff K. Bivariate agreement coefficients for reliability of data. Sociol Methodol 1970];2:139–50.
Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977];33:159–74.
Lanquart JP, Dumont M, Linkowski P. QRS artifact elimination on full night sleep EEG. Med Eng Phys 2005].
Lawhern V, Hairston WD, Robbins K. DETECT: a MATLAB toolbox for event detection and identification in time series, with applications to artifact detection in EEG signals. PLOS ONE 2013]:8.
Leclercq Y, Schrouff J, Noirhomme Q, Maquet P, Phillips C. fMRI artefact rejection and sleep scoring toolbox. Comput Intell Neurosci 2011]:11.
Li M, Cui Y, Yang J. Automatic removal of ocular artifact from EEG with DWT and ICA method. Appl Math Inf Sci 2013];7:809–16.
Moretti DV, Babiloni F, Carducci F, Cincotti F, Remondini E, Rossini PM, Salinari S, Babiloni C. Computerized processing of EEG, EOG and EMG artifacts for multi-centric studies in EEG oscillations and event-related potentials. Int J Psychophysiol 2003];47:199–216.
Mourão-Miranda J, Hardoon DR, Hahn T, Marquand AF, Williams SCR, Shawe-Taylor J, Brammer M. Patient classification as an outlier detection problem: an application of the one-class support vector machine. NeuroImage 2011];58:793–804.
Nakamura M, Sugi T, Ikeda A, Kakigi R, Shibasaki H. Clinical application of automatic integrative interpretation of awake background EEG: quantitative interpretation, report making, and detection of artifacts and reduced vigilance level. Electroencephalogr Clin Neurophysiol 1996];98:103–12.
Noachtar S, Binnie C, Ebersole J, Mauguière F, Sakamoto ABW. A glossary of terms most commonly used by clinical electroencephalographers and proposal for the report form for the EEG findings. the international federation of clinical neurophysiology. Electroencephalogr Clin Neurophysiol Suppl 1999];52:21–41.
Nolan H, Whelan R, Reilly RB. FASTER: fully automated statistical thresholding for EEG artifact rejection. J Neurosci Methods 2010];192:152–62.
Perreault WD Jr, Leigh LE. Reliability of nominal data based on qualitative judgments. J Mark Res 1989];26:135–48.
Pilcher JJ, Schulz H. The interaction between EEG and transient muscle activity during sleep in humans. Hum Neurobiol 1987];6:45–9.
Pillar G, Bar A, Shlitner A, Schnall R, Shefy J, Lavie P. Autonomic arousal index: an automated detection based on peripheral arterial tonometry. Sleep 2002];25:543–9.
Rechtschaffen A, Kales A. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Bethesda, MD: US Department of Health, Education and Welfare, Public Health Service; 1968].
Rohalova M, Sykacek P, Koska M, Dorffner G. Detection of the EEG artifacts by the means of the (extended) Kalman filter. Measur Sci Rev 2001];1:59–62.
Schlögl A, Anderer P, Roberts SJ, Pregenzer M, Pfurtscheller G. Artefact detection in sleep EEG by the use of Kalman filtering; 1999].
Scott WA. Reliability of content analysis: the case of nominal scale coding. Public Opin Q 1955];19:321–5.
Sugi T, Kawana F, Nakamura M. Automatic EEG arousal detection for sleep apnea syndrome. Biomed Signal Process Control 2009];4:329–37.
Teplan M. Fundamentals of EEG measurement. Measur Sci Rev 2002]:2.
Thomson D. Spectrum estimation and harmonic analysis. In: Proceedings of the IEEE, vol. 70; 1982]. p. 1055–109.
Vorobyov S, Cichocki A. Blind noise reduction for multisensory signals using ICA and subspace filtering, with application to EEG analysis. Biol Cybern 2002];86:293–303.

Definition

Arousal: Score arousal during sleep stages N1, N2, N3, or R if there is an abrupt shift of EEG frequency including alpha, theta, and/or frequencies greater than 16 Hz (but not spindles) that lasts at least 3 s, with at least 10 s of stable sleep preceding the change. Scoring of arousal during REM requires a concurrent increase in submental EMG lasting at least 1 s (Iber et al., 2007).

Artifact: (1) A potential difference due to an extracerebral source, recorded in EEG tracings. (2) A modification of the EEG caused by extracerebral factors such as alterations of the media surrounding the brain, instrumental distorsion or malfunction, and operational errors (Noachtar et al., 1999).