EEG windowed statistical wavelet scoring for evaluation and discrimination of muscular artifacts

François-Benoit Vialatte[1], Jordi Solé-Casals[2] and Andrzej Cichocki[1]

[1] Lab ABSP, RIKEN Brain Science Institute, Wako-Shi, Japan
[2] University of Vic, Vic, Spain

E-mail: fvialatte@brain.riken.jp

Abstract

EEG recordings are usually corrupted by spurious extra-cerebral artifacts, which should be rejected or cleaned up by the practitioner. Since manual screening of human EEGs is inherently error prone and might induce experimental bias, automatic artifact detection is an issue of importance. Automatic artifact detection is the best guarantee for objective and clean results. We present a new approach, based on the time–frequency shape of muscular artifacts, to achieve reliable and automatic scoring. The impact of muscular activity on the signal can be evaluated using this methodology by placing emphasis on the analysis of EEG activity. The method is used to discriminate evoked potentials from several types of recorded muscular artifacts—with a sensitivity of 98.8% and a specificity of 92.2%. Automatic cleaning of EEG data are then successfully realized using this method, combined with independent component analysis. The outcome of the automatic cleaning is then compared with the Slepian multitaper spectrum based technique introduced by Delorme *et al* (2007 *Neuroimage* **34** 1443–9).

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Artifacts in the EEG can be defined as any potential difference due to an extra-cerebral[3] source (Anderer *et al* 1999). In addition to instrument and environmental electrical 50/60 Hz noise

---

[3] Nota bene: retina belongs to the central nervous system. Eye movements, as will be detailed later on, can induce retinal potentials that could be considered as cerebral activity. Nevertheless, electrooculographic artifacts in general will be considered as 'muscular artifacts' for the sake of simplicity in this paper.

and movement artifacts, ocular, electromyographic (EMG), electrodermal, electrovascular and respiratory signals can interfere with the EEG in the form of artifacts. Muscle artifacts are especially problematic, because they can appear in patterns similar to that of true EEG signals; the frequency range of muscle artifacts and the investigated EEG waveforms overlap to a high degree (van de Velde 2000). EEG analysis is therefore greatly impaired by the presence of such muscle artifacts. The importance of artifact detection, either for rejecting corrupted signals or for applying subsequent denoising methods, has already been emphasized. However, since human intervention may be subjective, inconsistent and thus is less reliable, automatic methods are preferable to guide manual rejection. For instance, epoch-by-epoch agreement of the sleep-stage assignment task was extremely inconsistent among five experienced sleep technologists from different laboratories (Norman *et al* 2000)—which means that human recognition consistency, for known EEG patterns[4], is low. In this study, mean agreement was only 73%, and depended on the laboratory.

Automatic detection of artifacts is necessary when EEG activity above 20 Hz is studied (Whitham *et al* 2007) but is also generally necessary to avoid experimenter bias. Automatic methods are usually based on threshold techniques for EEG potentials or power spectra (Durka *et al* 2003), regression-based models (Moretti *et al* 2003), or projection-based methods (Wallstrom *et al* 2004). Independent component analysis (ICA) is especially useful (Ille *et al* 2002), exploiting statistical-independent criteria to separate artifacts (Jung *et al* 2000). Automatic criteria have been proposed (Delorme *et al* 2001) for semi-automatic artifact rejection. However, despite the advantages of ICA, potentially laborious manual identification of components to be removed is still needed to obtain a reliable result. Instead of exploiting time values of frequency spectra, (Zikov *et al* 2002) demonstrated wavelet joint time–frequency representations to be useful for EEG ocular artifact denoising. Using appropriate normalization, the so-called 'z-score' can enhance the precision of wavelet time–frequency maps, resulting in efficient artifact detection (Browne and Cutmore 2004).

However, most of the studies concerning automatic detection of artifacts use either databases screened by humans or databases of EEG signals containing added 'artificial' noise. Unfortunately, the algorithms developed under these conditions use the error-prone human performance as a reference! Human judgment having low reliability, the best way to asses the validity of the database used for testing, would be to use real artifacts generated in a controlled study, instead of artificial data. For instance, neural networks with wavelet preprocessing (Księżyk *et al* 1998) achieved a sensitivity of 80% and a specificity of 75%, ICA and Bayesian classification (LeVan *et al* 2006) achieved a sensitivity of 87.6% and a specificity of 70.2%, and the extraction of time and frequency characteristics (van de Velde *et al* 1998) achieved a specificity of 90% and a sensitivity of 80%. Except for the last study, these rates are rather poor, especially if one considers that the true class in which the signals belong is uncertain (i.e., artifacts were classified by human scorers, whose reliability is low (Norman *et al* 2000)). One must take into account the fact that results usually depend on analysis of the signal by epochs, and that these epoch-length characteristics are not always consistent from one study to another.

An EEG evoked potential and, more generally, event-related potentials are electrical potential recorded in one or more EEG channels following the presentation of a stimulus, as distinct from spontaneous potentials (the background EEG). They can be interpreted as the reorganization of the spontaneous brain oscillations in response to the stimulus (Başar 1980, Başar *et al* 1999). EP (especially visual EP) were observed in several studies to have visible

---

[4] Unfortunately, to the best of our knowledge, no such comparison for artifact rejection is reported. Note that the complexity of EMG artifact detection is comparable to the complexity of sleep staging (especially for low-amplitude artifacts).

outcomes even in single trials (see, e.g., Effern *et al* 2000, Quiroga *et al* 2001), especially observable when using wavelets (Başar *et al* 1999, Quiroga *et al* 2001) who represent the signal with optimal time–frequency resolutions. Despite the usual method to observe EP is based on signal averaging, it does not mean that single trials do not contain the stimulus evoked activity: averaging is only used to enhance an activity that is already present (even if weak as compared to the background EEG) in single trials. Hence, threshold-based methods of artifact rejection can be expected to reject evoked activity. This can be a serious flaw, if one attempt to clean EEG signals using ICA (especially in the high-frequency ranges): if one removes parts or portions of evoked activity in single trials, the resultant averaged signal is not a proper representation of the brain activity. Many reports do not take into account the impact of artifact rejection on evoked potentials, which may dramatically impair the EEG analysis (Stecker 2002).

We present here an approach to analyze corrupted EEG signals. Artifact cleaning methods are usually based on the estimate of a rejection threshold. Whereas rejection thresholds are usually determined using rest EEG, we determined our rejection threshold using EEG recorded during stimulation: our aim was to remove EMG artifacts while preserving event-related potentials. During EEG recordings, muscle artifacts were experimentally evoked and were compared to signals obtained from audio–visual stimulation. The proposed method exploits only EEG signals; it is unnecessary to record the electrooculogram (EOG) signal itself. The approach is developed for a wide variety of muscle artifacts, ranging from eye artifacts to head, jaw or body movements. We exploit time–frequency characteristics of EEG signals to define the optimal length of epochs to be analyzed. A score for each signal is returned, allowing one to either discard noisy signals or to clean them up using an appropriate technique (such as ICA).

2. Description and general observations

EEG signals were acquired using a 64 channel EEG system (sampling rate 1 kHz, gain 1000×). The high sampling rate is necessary for the analysis of high-frequency contents of the signal (the sampling rate should be three to five times the maximal frequency investigated, (Barlow 1993)). The subject was asked to voluntarily produce individual muscle artifacts (ten trials per artifact). Ten different muscle artifacts were produced:

 (i) blink left eye,
 (ii) blink right eye,
 (iii) blink both eyes,
 (iv) look from left to right (eye-movement artifact),
 (v) roll eyes clockwise,
 (vi) speak (say 'kampai'),
 (vii) swallow some water,
 (viii) move head (nodding, first down then up),
 (ix) grind teeth on chewing gum, three times,
 (x) stand up 30% of a full standing position and then sit down again.

Thus, the database consisted of 100 recordings (ten trials for ten artifacts). We recorded transient EP, that can be elicited using either step or impulse functions at the input of the nervous system (Başar *et al* 1999). The subject was exposed simultaneously to a 2000 Hz audio tone (duration: 30 ms) and a pattern offset/onset visual stimulus (a 600×600 pixel checkerboard image displayed on a 32×25 cm LCD screen located 30 cm from the subject's nasion; duration: 10 ms), while EEGs were recorded. Both sound and images are presented

to the subject, so that the EEG will record simultaneously evoked potentials in the visual and auditory areas (usually occipital and temporal areas) together with their multimodal interactions. The EEG signal was recorded before, during and after this stimulation (recording duration = 10 s). Twenty five trials were recorded. We do not intend to study the averaged signal but single trials (the motivation is to clean single trials before averaging, so that evoked potentials would be less distorted by EMG artifacts).

2.1. Different muscle artifacts, different properties

Here, we present general descriptions of the mechanisms that cause EMG artifacts in EEG recordings. A complete review of EEG artifacts is detailed in Benbadis and Rielo (2005).

The eyeball acts as a dipole, with an anteriorly oriented positive pole (cornea) and a posteriorly oriented negative pole (retina). Ocular globe rotation about its axis (for instance, eye movement or eye rolling) generates a large-amplitude alternate current field that is recorded by EEG scalp electrodes. In addition, movements of the eyelids have a shunting effect on the corneal–retinal dipole. Hence, eye blinks also induce EEG perturbations. Myogenic potentials are induced by muscular activity, e.g., during body or head movements. Frontalis and temporalis muscles, which contract during teeth grinding or clenching of jaw muscles, especially induce strong EMG artifacts. Like the eyeball, the tongue is also a dipole, with a negative tip and a positive base that produces a so-called glossokinetic artifact. The tip of the tongue is the most important part of the tongue because it is more mobile than the base of the tongue. Not only chewing, sucking and swallowing can produce EEG artifacts but speaking can also produce artifacts.

See endnote 1

2.2. Different muscle artifacts, different frequency spectra

Figure 1 presents the Fourier spectrum of each kind of artifact (absolute log Fourier power, single 2 s window centered on the artifact averaged for all trials). Even though muscle artifacts are commonly considered to primarily affect high frequencies in the EEG, low-frequency activity of the EEG is also strongly distorted by artifacts (Zimmermann and Sharein 2004, Freeman *et al* 2003). Because high frequencies have lower EEG power than low frequencies, whereas EMG power is of similar order in both high and low frequencies, the 'signal-to-EMG' ratio for high frequencies is poorer than that of low frequencies. Thus, depending on the type of artifact, the spectrum elicited may have specific responses in four frequency bands: $\delta \leftrightarrows \alpha = 1$–10 Hz, $\alpha \leftrightarrows \beta = 10$–35 Hz, $\gamma = 60$–90 Hz, and $\lambda > 90$ Hz.

2.3. Different muscle artifacts, different scalp repartition

Figure 2 presents the QEEG (quantitative EEG) distributions of different muscle EEG artifacts. Each artifact was analyzed by using the dB power of the Fourier transform in the 10–90 Hz frequency range, after baseline removal. We observed specific spatial distributions, which we will use later on to identify the artifacts.

3. Methods

3.1. Wavelet transformation and time–frequency map generation

Wavelets (see Mallat (1997) for details), especially complex Morlet wavelets (Kronland-Martinet *et al* 1988), have already been widely used for time–frequency analysis of EEG signals (Tallon-Baudry *et al* 1996, Düzel *et al* 2003, Caplan *et al* 2001, Li *et al* 2007, Slobounov *et al* 2008). Complex Morlet wavelets $w(t)$ of the Gaussian shape in time (deviation σ) are
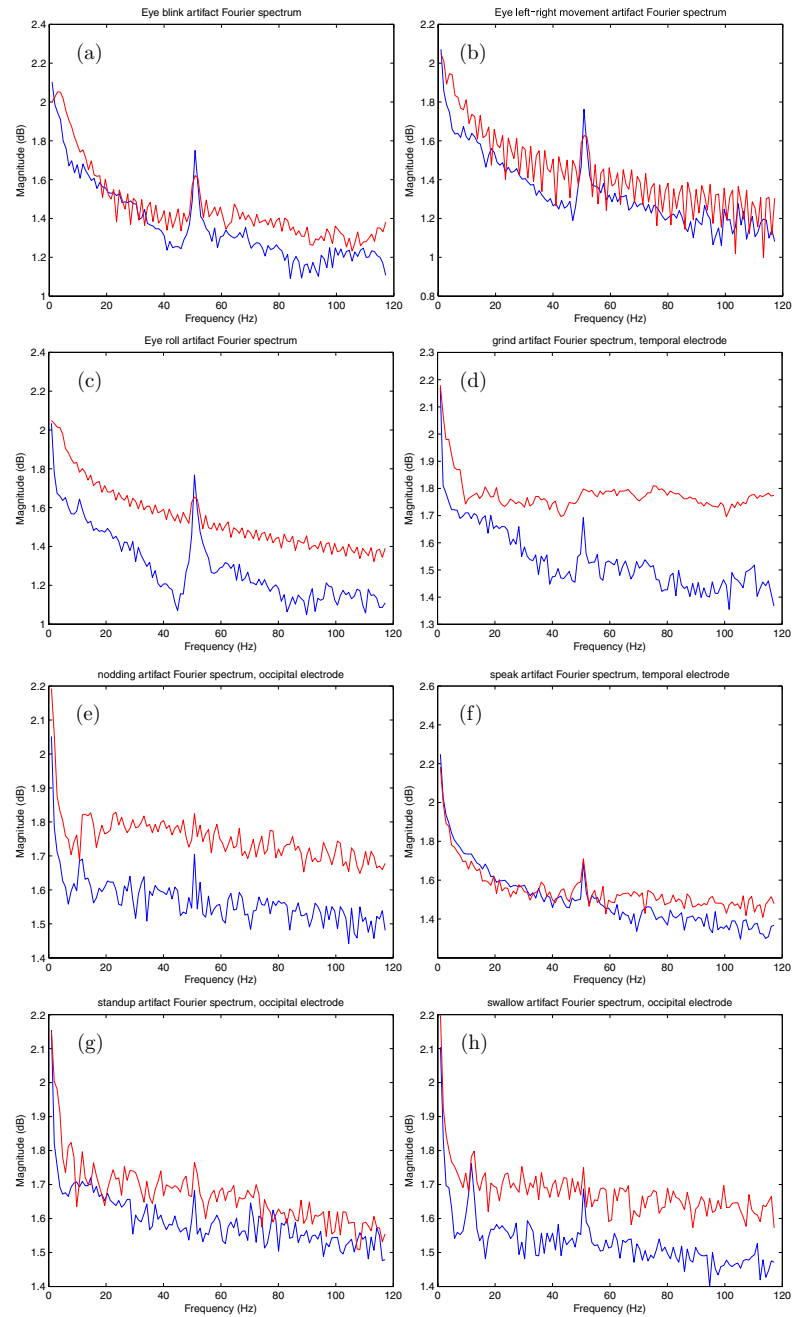
Figure 1. Fourier dB spectrum average for all ten kinds of muscular artifacts (averaged periodogram in the logarithmic scale). (a) Eye blink, (b) gaze left → right, (c) eye roll clockwise, (d) teeth grinding, (e) head nodding up and down, (f) say the word 'Kampai', (g) stand up and down and (h) swallow. The blue curves are obtained by averaging the spectrum of all ($n = 10$) records 5 s after artifact offset, whereas the upper curves are obtained during the artifact presence. Most signals contain a peak around 50 Hz, corresponding to power noise. FFTs were computed at a peak location of the QEEG distribution of each artifact (see 2). Each artifact has a specific effect on the frequency spectrum (for instance, speaking does not disturb low frequencies, but eye rolls do).
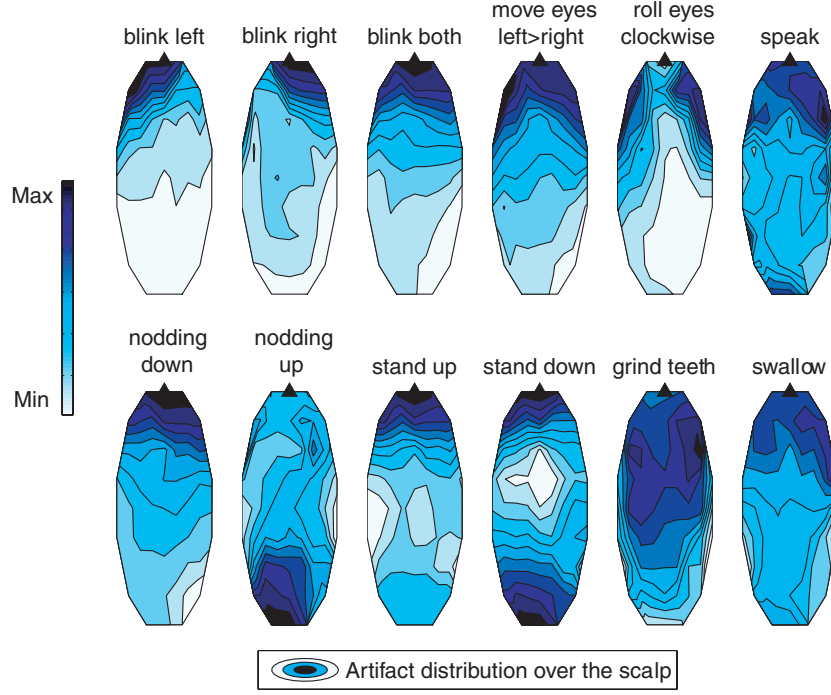
Figure 2. Distribution of the QEEG artifact power (10–90 Hz), after *z*-score normalization compared to the baseline (rest condition). The white zones represent electrode locations where the activity is close to the baseline activity. The darker zones represent specific increases in power. The figure was interpolated from the *z*-score into a location grid, using Matlab®.

defined as

$$w(t) = A \cdot \mathrm{e}^{\frac{-t^2}{2\sigma^2}} \cdot \mathrm{e}^{2\mathrm{i}\pi f t}, \tag{1}$$

where σ and f are interdependent parameters, the constraint $2\pi f t > 5$, and the wavelet family is defined by $2\pi f t = 7$, as described in Tallon-Baudry *et al* (1996). This wavelet has positive and negative values resembling those of an EEG, but also a symmetric Gaussian shape both in the time and frequency domains—i.e. this wavelet locates accurately time–frequency oscillations both in the time and frequency domains. For each time sample t and each frequency bin f, the wavelet transform computes one coefficient c_{ft}. Wavelet representations can be investigated according to baseline activity. To this end, one method typically used is to normalize the time–frequency representation depending on the mean μ_f and standard deviation σ_f of each frequency bin f in the baseline activity (the so-called *z*-score (Browne and Cutmore 2004)). To detect the artifact-corrupted activity, the baseline activity should be a representative of non-noisy signals. However, because EEG signals generally have a low signal-to-noise ratio (SNR), the most reliable method is to repeat the estimation of μ_f and σ_f on several clean signals (with the fewest possible apparent noisy activity). From each signal b (used as baselines), a general normalized score is computed:

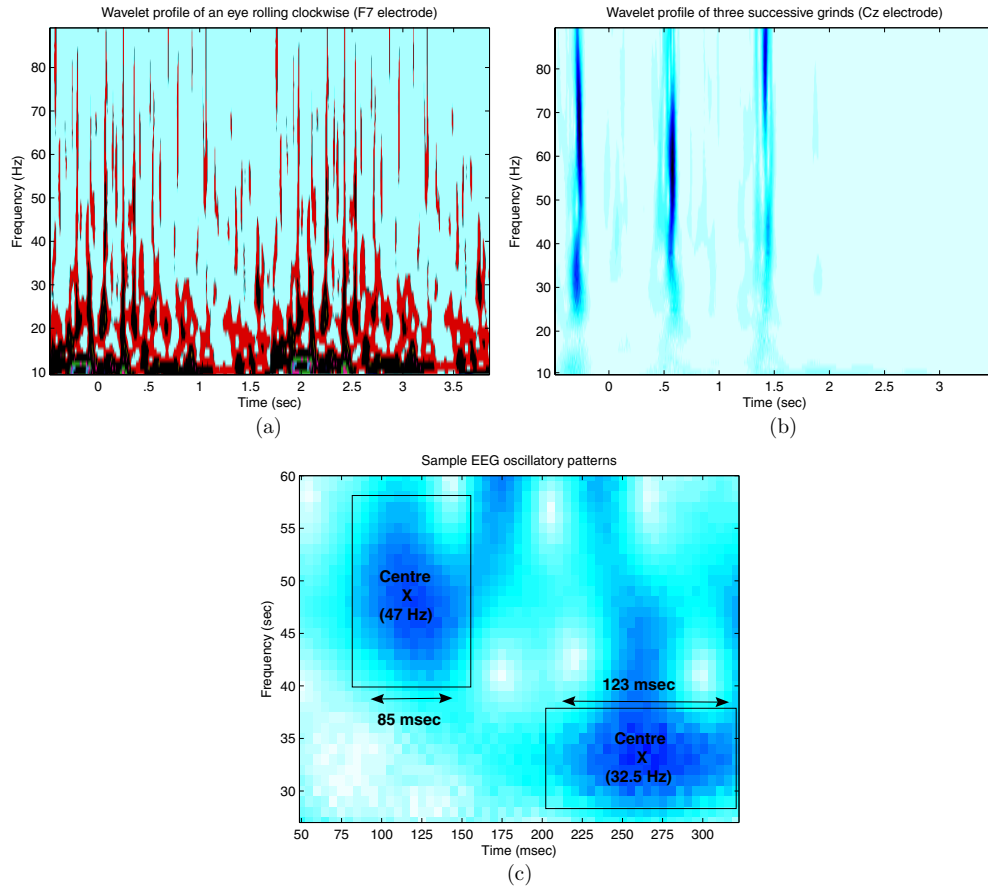$$z_{ft} = \frac{c_{ft} - M_f}{S_f}, \qquad \forall t, \tag{2}$$

Figure 3. Sample wavelet time–frequency profiles for artifact compared to evoked potentials: (a) artifact time–frequency profile, eye rolling clockwise, (b) artifact time–frequency profile, teeth grinding three times and (c) evoked potential time–frequency profile (zoom).

where M_f is the average of the baseline means $\mu_f(b)$ computed for each signal b at the frequency f:

$$M_f = \overline{\mu_f(b)} \tag{3}$$

and S_f is the average of the baseline standard deviation $\sigma_f(b)$ computed for each signal b at the frequency f:

$$S_f = \overline{\sigma_f(b)}. \tag{4}$$

3.2. Why consider time–frequency information?

Each type of artifact has a specific time–frequency shape, with sharp activity in the high-frequency range and/or high amplitudes in the low-frequency range. Evoked potential activity, on the other hand, is less sharp in the high-frequency range and has lower amplitudes in the low-frequency range. The latter also usually has a well-defined duration (more than 3 time periods (Caplan *et al* 2001)), as illustrated in figure 3. Time–frequency joint representations allow the extraction of these characteristics, by defining time–frequency windows of interest, which are more precise than the usual time windows used to define epochs for artifact scoring.

3.3. Windowed z-score

Wavelet coefficients represent a signal with a time–frequency resolution that depends on frequency. For high-frequency activity, the time position of events is very precise, while the frequency content is imprecise and blurred. Conversely, for low-frequency activity, the time position of events is imprecise, while the frequency estimation is very precise and thus is not blurred. Added to this mathematical effect intrinsic to the properties of frequencies, we have shown in the previous section that each artifact has specific time–frequency distributions.

Based on this observation, we define regions of interest in the time–frequency dimension. We will first select frequencies of interest based on the ranges defined by figure 1: $\delta\alpha =$ 1–10 Hz, $\alpha\beta = 10$–35 Hz, $\gamma = 35$–90 Hz, and $\lambda > 90$ Hz. The $\delta\alpha$ band is low frequency; hence, one would need a long-duration recording to efficiently estimate this activity (which is not always feasible). Furthermore, EEG displays strong variations in the α range, which also seems to have the best 'signal-to-EMG' ratio. For λ activity, the sampling rate does not always allow one to study such a high range (it must be three to five times higher than the frequency investigated (Barlow 1993)). Furthermore, technical limitations in clinical scalp recording often prevents reliable recording in high frequencies from scalp EEG. The γ range needs to be restricted even more, to 60–90 Hz, to avoid eventual 50 Hz environmental electrical noise.

Because our goal is to define a general method, we concentrated on the $\alpha\beta$ and γ frequency ranges (10–35 Hz and 60–90 Hz). For these two bands of interest, we defined shifting time windows of four time periods around the central frequency (the period was $T_{\alpha\beta} = 180$ ms for $\alpha\beta$ and $T_\gamma = 53$ ms for γ). Using these windows shifted along the time axis, we computed two scores at five electrode positions $E = [F_{p1}, F_{p2}, T_7, T_8, O_z]$.

The maximal score W_x indicates if the signal contains an artifact. Since high frequencies are the most efficient for this discrimination, we used only the γ range:

$$W_x = \max_e(H_x(e)), \tag{5}$$

where e represents EEG electrodes in E, and H_x represents the score for the signal in γ frequency ranges (from $f_m = 60$ to $f_x = 90$, i.e. $F_\gamma = 31$ frequency bins):

$$H_x(e) = \max_\tau(\sigma_\gamma(\tau)) = \max_\tau \left(\sqrt{\frac{1}{F_\gamma . T_\gamma - 1} \sum_{f=f_m}^{f_x} \sum_{t=\tau}^{\tau+T_\gamma} (z_{ft} - \overline{z_{ft}})^2} \right). \tag{6}$$

The measure is therefore based on the standard deviation and not on amplitude, the latter of which may classify strong EEG activity as an artifact.

On the other hand, the overall score W_s indicates the degree of noisiness within the signal. To assess this general impact, we also took into account low frequencies of the $\alpha\beta$ range:

$$W_s = \max_e(L_s(e), H_s(e)), \tag{7}$$

where e represents EEG electrodes in E, and L_s and H_s represent the scores for the signal in $\alpha\beta$ (from $f_{m\alpha\beta} = 10$ to $f_{x\alpha\beta} = 35$, i.e. $F_{\alpha\beta} = 26$ frequency bins) and γ frequency ranges:

$$L_s(e) = \langle \sigma_{\alpha\beta}(\tau) \rangle_\tau = \left\langle \sqrt{\frac{1}{F_{\alpha\beta} \cdot T_{\alpha\beta} - 1} \sum_{f=f_{m\alpha\beta}}^{f_{x\alpha\beta}} \sum_{t=\tau}^{\tau+T_{\alpha\beta}} (z_{ft} - \overline{z_{ft}})^2)} \right\rangle_\tau \tag{8}$$

and

$$H_s(e) = \langle \sigma_\gamma(\tau) \rangle_\tau \tag{9}$$

with $\sigma_\gamma(\tau)$ as defined in equation 6.

These two indicators are mandatory to accurately identify cases in which many evoked potentials occur (high W_s but low W_x). W_x accounts for the presence of artifacts, while W_s accounts for the quantitative proportion of artifacts in the signal.

3.4. Combination with ICA

Consider the case of multiple EEG channels sampling brain activity over time. If the signals from each channel form the rows of the data matrix D, then each column of D is a time point. Our problem is to identify the original brain sources that were mixed in the EEG channels. This is a typical problem of blind source separation (BSS). ICA is a worthy solution for BSS in the context of EEG recordings (Makeig *et al* 1996, Tang *et al* 2004), and finds the unmixing square matrix W ($n = m =$ the number of channels) such that $W \cdot D = C$ (Brown *et al* 2001). The rows of C are called 'independent components' because they are forced to be as independent as possible; these are the sources for which we were searching. Delorme *et al* (2007) showed that preprocessing EEG data using ICA allow effective artifact detection and that the Slepian multitaper spectrum had the best overall cleaning capability. ICA can effectively separate EEG from EMG background activity. Therefore, if we can detect which of the C components represent artifacts, then it would be possible for us to reconstruct a cleansed matrix D' in which these components are removed. Thus, to clean corrupted signals, we combined the windowed z-score method with ICA. First, the artifact database was mixed with clean EEG signals. Artifacts were combined with the signals, and the ratio $10 \log_{10}(S/N)$ of signal to noise was varied from -4.8 to 4 dB. For signals with good SNR values, artifact detection becomes difficult, even for human eyes (with good SNRs ranging from 0 to 4 dB, medium SNRs ranging from -0.8 to -3 dB and poor SNRs ranging from -3.4 to -4.8 dB). With poor SNR values, however, artifact detection becomes easy, and signals are altered greatly (see figure 4). For all samples, we computed a mean-square error O_N with the original (clean) signal, measuring the original noise.

These corrupted signals were then decomposed using four different algorithms: SOBI, JADE, FastICA and ThinICA (see, e.g., Cichocki and Amari 2002). Sources that contributed to 95% of the total variance were retained, usually leading to a reduction to less than 15 components. These components were then analyzed using either W_x or Slepian multitaper spectrum in the δ or γ ranges. After artifact cleaning, we computed the mean-square error O_E with the original (clean) signal. The cleaning was then evaluated as a ratio $1 - O_E/O_N$ in percentage. A negative value means that the signal deteriorated rather than becoming cleaned.

3.5. Method overview

Here we recapitulate how W_x and W_s indicators are computed, and how to use them.

The computation of W_x depends on the purpose of its application, i.e. artifact rejection or artifact cleaning. For artifact rejection, the method follows five steps:

(i) Compute the reference baseline with appropriate clean signals.
(ii) Transform the EEG signal to be analyzed into a wavelet map.
(iii) Normalize the wavelet map relatively to the baseline.
(iv) Compute W_x.
(v) Test if W_x is above the threshold (reject if above 2).

In the context of artifact cleaning, two steps are added, and the rejection test is replaced by a deletion of corrupted ICA components:
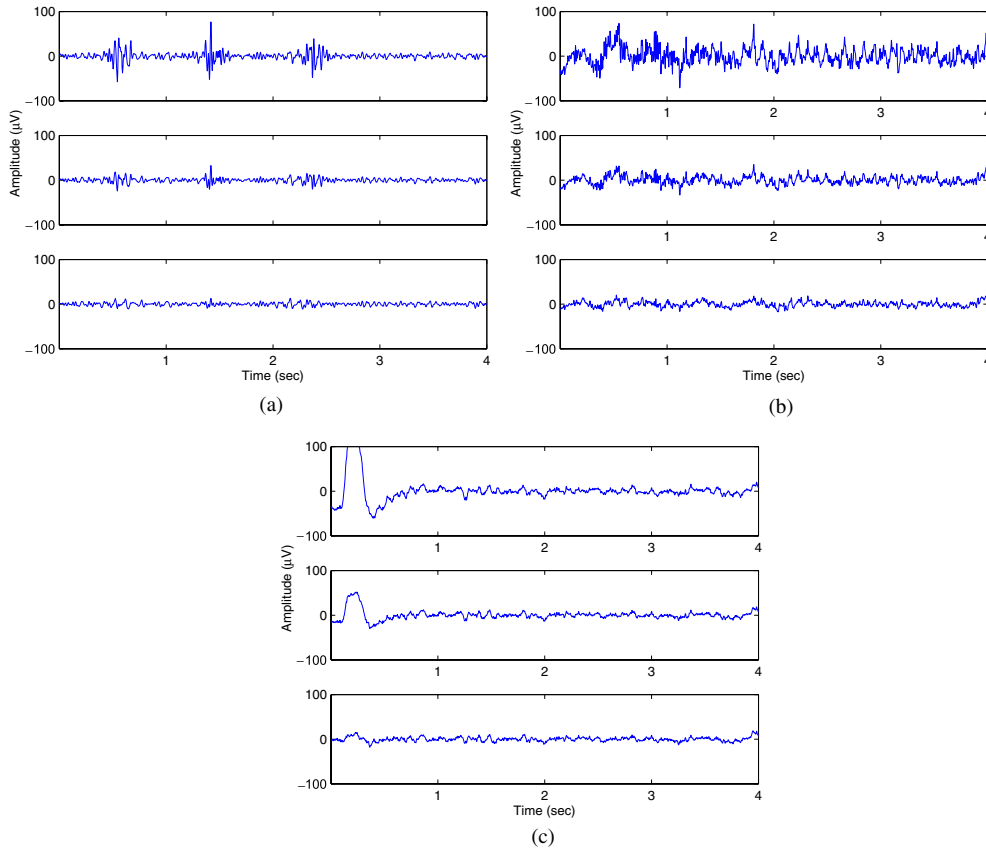
Figure 4. Examples of different artifacts and clean signals having poor, medium and good SNRs. Good SNRs range from 0 to 4 dB, medium SNRs range from −0.8 to −3 dB and poor SNRs range from −3.4 to −4.8 dB. (a) Example of signal corrupted by teeth grinding, having three different SNRs, (b) example of signal corrupted by speech, having three different SNRs and (c) example of signal corrupted by eye blink, having three different SNRs.

(i) Compute the reference baseline with appropriate clean signals.
(ii) Transform the EEG signal to be analyzed into ICA components.
(iii) Transform each ICA component into a wavelet map.
(iv) Normalize the wavelet maps relatively to the baseline.
(v) Compute W_x from each component's normalized map.
(vi) Delete components with high W_x (if above 2)—always remove the worst component even if W_x is below 2 (considering that there is prior knowledge, always keep at least one component (in the worst case, the best W_x) that the signal is corrupted).
(vii) Back project the remaining components to the scalp.

These steps are summarized in the flowchart of figure 5.

There are five steps to follow when implementing database balancing using W_s. In order to balance artifacts from two databases A and B:

(i) Compute the reference baseline with appropriate clean signals.
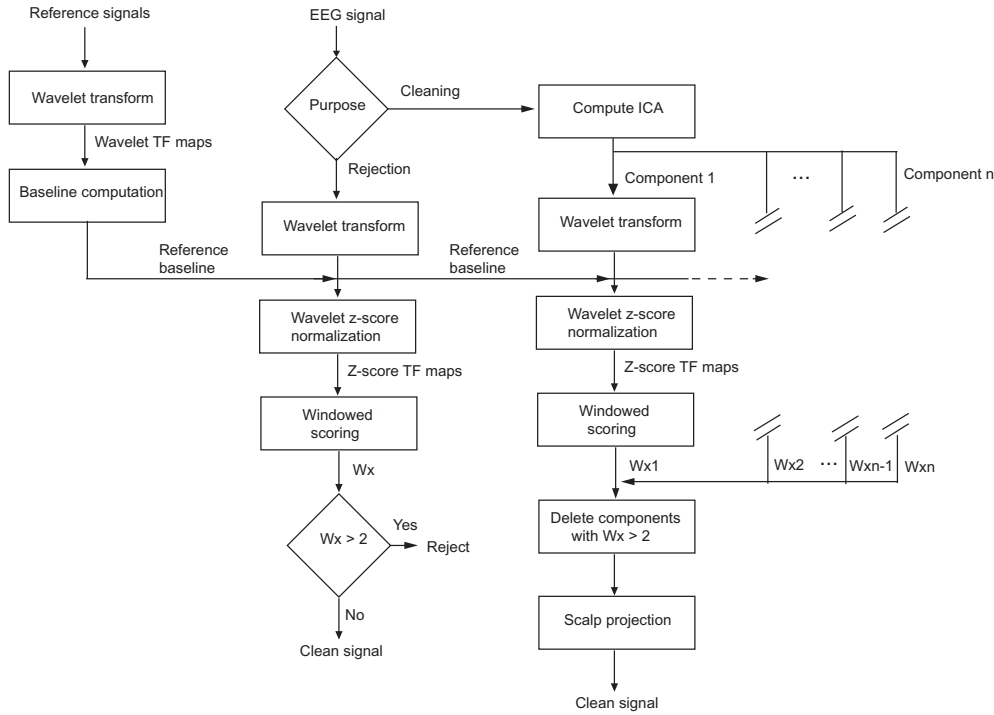(ii) Transform all the EEG signals to be analyzed into wavelet maps.

Figure 5. Flowchart of artifact detection using W_x. Two possibilities are offered: either using W_x to reject corrupted portions of the signal or to clean artifacts in combination with ICA.

(iii) Normalize the wavelet maps relatively to the baseline.

(iv) Compute W_s for each signal.

(v) Compare the distribution of W_s scores for A against B using a Kolmogorov–Smirnov test.

When the test returns a low *p*-value (especially below 0.05), the amount of artifacts in each database is not similar. See Vialatte *et al* (2007) for more details about this procedure. These steps are summarized in the flowchart of figure 6.

4. Results

4.1. Artifact rejection

Figure 7 shows the overall result of the artifact rejection method based on the W_x rejection score. We apply the method on the two databases described above (artifact data as compared to multimodal evoked potentials). Using linear discriminant analysis, we first estimated the generalization capability of this method using a cross-validation method. A two-fold cross-validation error of 4% was obtained, with a sensitivity[5] of 98.8% and a specificity of 92.2%. Eye-movement scores are generally close to the maximal evoked potential activity, with eye blinks displaying the most varying activity. For a simple threshold classification, a cutoff of 2.0 leads to the misclassification of 4.0% of evoked potential signals, while detecting all

---

[5] Here sensitivity is the ratio of correctly rejected artifacts in the test fold, and specificity is the ratio of correctly preserved clean EEG signals in the test fold.
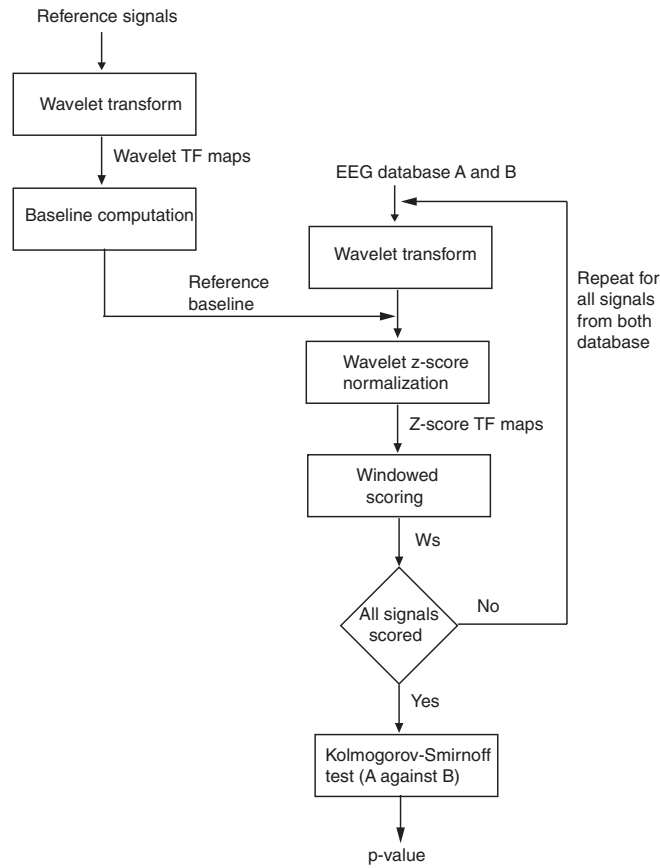
Figure 6. Flowchart of artifact balancing using W_s. Signals from two databases (A and B) are evaluated. Each signal is associated with a W_s score. These scores are compared with a Kolmogorov–Smirnov test. If the *p*-value is below 0.05, the impact of EMG artifact is higher in either A or B (see also Vialatte *et al* (2007)).
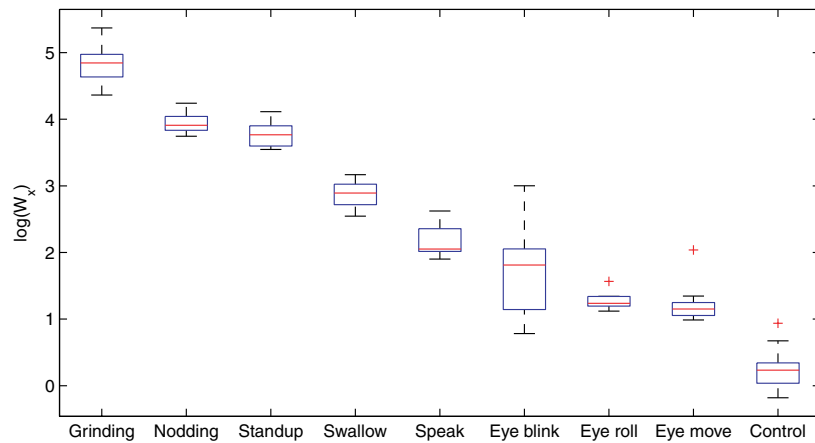

Figure 7. Boxplot of W_s for all artifacts (in the decreasing order of scores) as compared to non-noisy signals containing evoked potentials (labeled control)—see (Benjamini 1988) about boxplots.
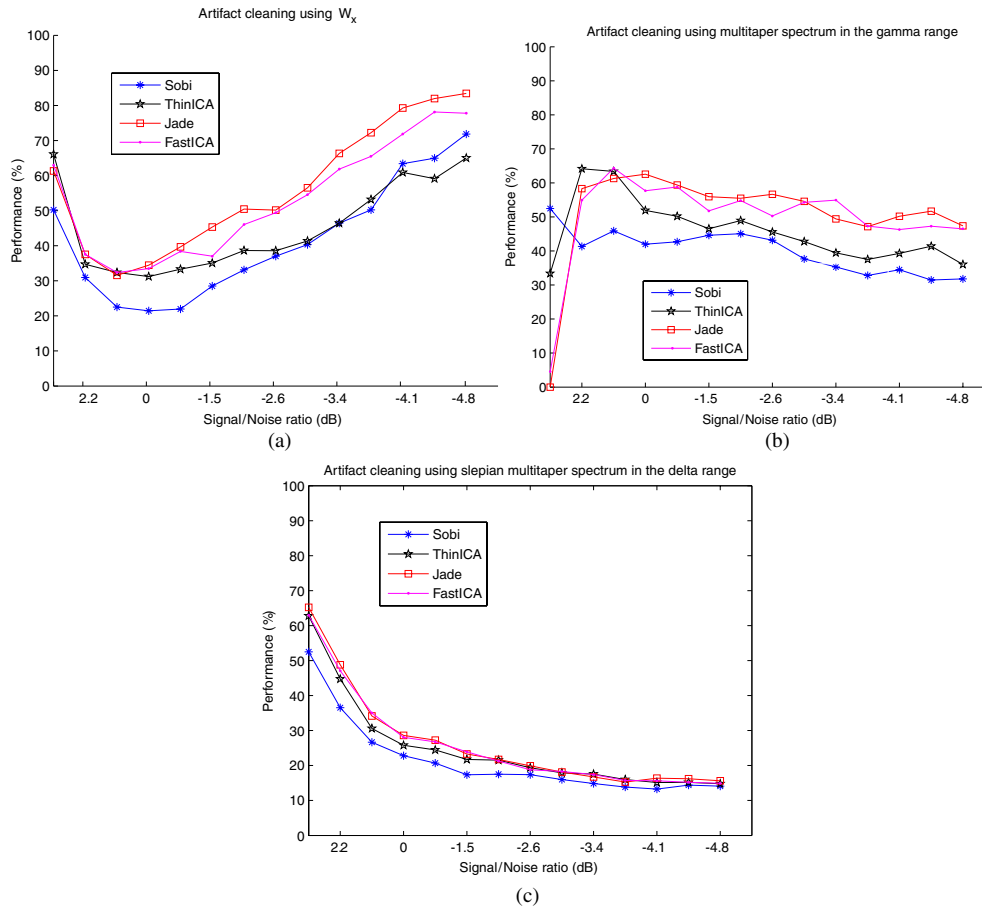
Figure 8. Artifact rejection combined with ICA: (a) using W_x, (b) using Slepian multitaper spectrum in the γ range and (c) using Slepian multitaper spectrum in the δ range.

artifactual signals. A cutoff of 2.6 leads to a misclassification of 2.0% of artifactual data (all from eye blinks, corresponding to 6.7% of the overall eye blinks), while detecting all evoked potentials.

When analyzing false positive occurrences using the 2.0 cutoff, we discovered that some false positives apparently represented an artifact that was not previously spotted (electrode movement artifact) in the evoked potentials database and an artifact that lies outside of the usual distribution of evoked potentials (shown as outlier in figure 7). The last types of artifacts to look for in the EGG are eye-induced artifacts and other muscular artifacts, which are easily discriminable because they clearly differ from one another and from EEGs. Only speech-related artifacts have comparable W_x score values, which range from 6.7 to 13.8.

When combined with ICA, the method's performance varied, depending on the SNR (figure 8). When the SNR is poor, the method is very effective with a rejection rate of up to 80%. When the SNR is moderately good, γ activity becomes more difficult to track and performance declines to 30%. When the SNR is good, low-frequency activity is more easily
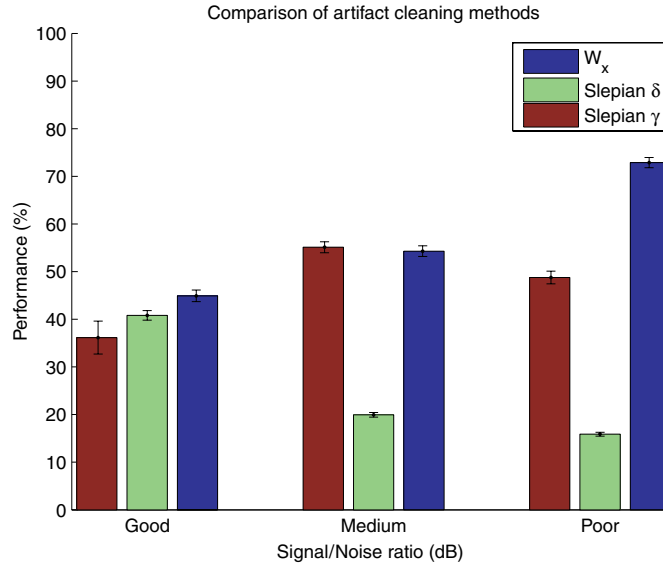
Figure 9. General comparison of ICA artifact cleaning as a function of SNR. The bars represent average of all signals and all ICA algorithms. Error bars represent standard error. Good SNRs range from 0 to 4 dB, medium SNRs range from −0.8 to −3 dB and poor SNRs range from −3.4 to −4.8 dB.

separated by ICA, and the method again becomes efficient with a rejection rate of up to 65%. When compared with the Slepian multitaper spectrum, similar results are obtained: in the γ range, artifact cleaning is more efficient when the SNR is low, whereas in the δ range, artifact cleaning is more efficient when the SNR is high. A systematic comparison (figure 9) shows, however, that W_x is consistently equal or above the multitaper spectrum in all SNR conditions (Mann–Whitney $p < 0.01$ for good an poor SNR conditions, no significant difference for medium condition). With good SNR conditions, the Slepian multitaper spectrum method becomes unstable and frequently yields negative results.

4.2. Artifact scoring

Instead of rejecting all artifacts, one may be interested in keeping signals with a satisfactory 'signal-to-artifact' ratio. For instance, eye blinks cannot be rejected for an 'eyes opened' condition with long duration, which does not elicit the same degree of perturbation within EEG signals. In other words, one may be interested in a quantitative rather than qualitative approach. When W_x represents an artifact, W_s scoring allows such a quantitative evaluation.

Table 1 presents average log W_s scores for each type of artifact (log of 10 averaged W_s for each trial, except for eye blinks, which were grouped in 30 trials, and eye moves, which were grouped in 20 trials). Scores represent the impact of the artifact on the EEG (the stronger the impact the higher the scores). W_s scores were computed for 2.5 s during and after the artifact was triggered. Using this score, one can discard signals depending on the desired quantitative amount of artifacts accepted (for instance, in experiments dealing with the eyes-opened condition, one may accept artifacts having a magnitude of up to 1).

Table 1. Quantitative evaluation of perturbations of EEG signals due to artifacts, using W_s log average scoring. The magnitude is a rounded value.

Artifact type	$\log_{10}\langle W_s \rangle$	Magnitude
Eye blink	0.24	0
Eye move	0.88	1
Speak	1.29	1
Swallow	2.13	2
Grind three times	2.84	3
Nodding	2.95	3
Stand up	3.27	3

This method can be used to balance the artifact corruption of two databases before comparing them. This method was also successfully applied to signals recorded from demented Alzheimer's disease patients in Vialatte *et al* (2007), where guidelines can be found about artifact balancing.

5. Conclusions

We presented a new approach for rejecting artifacts on the basis of the time–frequency properties of artifacts: sharpness of high frequencies and low frequencies. We tested this method using real data. Simple linear discriminant analysis of these data revealed that the artifact likelihood W_x achieved high cross-validation results with a sensitivity of 98.8% and a specificity of 92.2%. W_x allowed satisfactory artifact cleaning when combined with ICA, and outperformed the Slepian multitaper spectrum for this task. This method also allowed the evaluation of the W_s score, the value of which represents the overall impact of muscular artifacts on EEG signals.

With the W_x score, eye-blink artifacts appeared to closely resemble normal EEG high-frequency evoked potentials. To achieve an improved classification, we found that a simple relative power $\frac{H_x(F_{p1})+H_x(F_{p2})}{H_x(O_z)}$ is capable of efficiently discriminating 100% of eye blinks from all other artifacts and from normal EEG, allowing a perfect 0% cross-validation error. However, this result must be interpreted with caution as it may be related to the specific type of evoked potentials being analyzed. For instance, frontal evoked potentials, especially may not be so easily discriminated from eye blinks.

The W_x score is useful for detecting the presence of artifacts within signals. One general application of this method, would be to use W_x scoring on all EEG signals to be analyzed, and then to apply ICA to clean any noisy signals or simply to remove noisy data.

The W_s score can further help the practitioner by providing information about the strength of the artifact encountered. This score can be used, for instance, when several sets of signals are to be compared. Before making a comparison, the practitioner can improve the reliability of his investigation by using W_s scores to determine whether noise levels are approximately the same in all groups.

To compute these scores, one needs to estimate an average baseline from clean signals. We advocate the use of signals in the rest condition, with an averaging on several trials—this means that our method cannot yet be applied online (trials have to be recorded beforehand). Furthermore, the precision of this baseline evaluation is a critical issue, since the artifacts to be detected are activities that significantly differ from this baseline. A preferable solution to dealing with few data or when poorly reliable data are only available is to use the inter-quartile

range as a more stable estimator of standard deviation, instead of estimating the baseline by averaging (Browne and Cutmore 2004). However, as the amplitude distribution of EEG activity is skewed in the presence of evoked potentials, this method may remove some evoked potentials from the distribution. These evoked potentials would be mistakenly considered as outliers, thus achieving a lower specificity.

In the present study, we computed W_s scores using sub-windows within an overall period of 2.5 s. However, a shorter time period can also be considered. As some artifacts elicit long-duration activity in low frequencies, the score would become less reliable for slow muscular artifacts (typically standing up/down, or nodding the head). Longer time periods could also be used, but then information about short transient muscular artifacts (eye blinks, for instance) would become less reliable.

Although only five electrodes—two frontal, two temporal and one occipital—were used in the present study; the proposed method can be extended to other sets of electrodes and be applied optimally to all electrodes. Ideally, the more electrodes can be used, the better result will be obtained. However, the computational demand of such an investigation could become heavy. As muscular artifacts are usually spread (see figure 2), we consider that it is sufficient to use few well-chosen electrodes to detect EMG corruption. This would however not be the valid assumption for other types of artifacts, which can have local effects on isolated electrodes.

This method is well suited for rejecting muscular artifacts. However, one should take into account the effects of other types of artifacts—i.e., electrodermal, electrovascular and respiratory artifacts—that have not yet been evaluated. Furthermore, because epileptic activity may also display sharp waves above the β range and up to the γ range (Worrell *et al* 2004), this type of activity might also be detected as muscular artifacts when using W_x. When one wants to use our method as a first cleaning stage before other artifact or epilepsy detection, it should be combined either with an ad hoc automatic detection for these patterns or with manual detection by an expert.

Finally, other frequency ranges should also be considered when analyzing muscular artifacts, especially those in the $\delta\alpha$ range, which displays dramatic increases for some of them (see figure 1). However, as explained in section 3.3, the $\delta\alpha$ band occurs within a low-frequency range; hence, one would need to record for a long duration to efficiently estimate this activity. Another limitation is that other artifacts may also interfere with this frequency band (such as electrodermal artifacts and electrode movements). However, the extension of the W_x computation to this range remains feasible, with an appropriately chosen time–frequency resolution.

Our results were obtained using a database of 100 real artifact signals, and another of 25 evoked potentials. The processed database was seeded with the reference signal, and with the varying SNR for the artifacts, which leads to hundreds of test signals. Our results on this test set show that our method is stable and consistently performed better than the Slepian Multitaper threshold method. However, the reader should keep in mind that the reproducibility of our results was not *stricto sensus* verified, as we used only one subject. We also encourage the reader to consider figures 1 and 2 as illustrative examples (these distributions might slightly vary with other subjects). This paper has nevertheless a better theoretical validity, and our results are obviously more reproducible, than results obtained with methods solely based on artificial data instead of real EEG signals.

Some scientist might worry about researches intending to replace human intervention by automated methods—a non-supervised method in the field of artifact rejection is not advisable. Nevertheless, this objection should not prevent tentative exploratory researches, as our method is not intended in the present stage to replace human intervention, but instead to help him

in his task: this method would indeed find its best use in a semi-supervised routine, where human intervention confirms the computer's decisions (where the task of the algorithm would be to spot suspect patterns, human confirms or refuse rejection). In conclusion, we suggest that the choice made here is not to try to imitate human experts, but rather to classify known artifacts. This should remain a quality criterion for as long as human scorers' on EEG pattern classification will not have proven to be more consistent (Norman *et al* 2000).

Acknowledgments

References

Anderer P *et al* 1999 Artifact processing in computerized analysis of sleep EEG—a review *Neuropsychobiology* **40** 150–7

Barlow J 1993 *The Electroencephalogram: Its Patterns and Origins* (Cambridge, MA: MIT Press)

Başar E 1980 *EEG-brain dynamics: Relation Between EEG and Brain Evoked Potentials* (Amsterdam: Elsevier)

Başar E, Demilrap T, Schürmann M, Başar-Eroglu C and Ademoglu A 1999 Oscillatory brain dynamics, wavelet analysis, ands cognition *Brain Lang.* **66** 146–83

Benbadis S and Rielo D 2005 Eeg artifacts emedicine, http://www.emedicine.com

Benjamini Y 1988 Opening the Box of a Boxplot *Am. Stat.* **42** 257–62

Brown G, Yamada S and Sejnowski T 2001 Independent component analysis at the neural cocktail party *Trends Neurosci.* **24** 54–63

Browne M and Cutmore T R H 2004 Low-probability event-detection and separation via statistical wavelet thresholding: an application to psychophysiological denoising *Clin. Neurophysiol.* **113** 1403–11

Caplan J B, Madsen J R, Raghavachari S and Kahana M J 2001 Distinct patterns of brain oscillations underlie two basic parameters of human maze learning *J. Neurophysiol.* **86** 368–80

Cichocki A and Amari S 2002 *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications* (New York: Wiley)

Delorme A, Makeig S and Sejnowski T 2001 Rejection For EEG Data Using High-Order Statistics And Independent Component Analysis *Proc. 3rd Int. Conf. on Independant Component Analysis and Signal Separation (ICA'2001') (San Diego, CA)*

Delorme A, Sejnowski T and Makeig S 2007 Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis *Neuroimage* **34** 1443–9

Durka P, Klekowicz H, Cieślak-Blinowska K, Szelenberger W and Niemcewicz S 2003 A simple system for detection of EEG artifacts in polysomnographic recordings *IEEE J. Biomed. Eng.* **50** 526–8

Düzel E, Habib R, Schott B, Schoenfeld A, Lobaugh N, McIntosh A R, Scholz M and Heinze H J 2003 A multivariate, spatiotemporal analysis of electromagnetic time-frequency data of recognition memory *Neuroimage* **18** 185–97

Effern A, Lehnertz K, Schreiber T, David P and Elger C 2000 Nonlinear denoising of transient signals with application to event related potentials *Physica* D **140** 257–66

Freeman W, Holmes M, Burke B and Vanhatalo S 2003 Spatial spectra of scalp EEG and EMG from awake humans *Clin. Neurophysiol.* **114** 1053–68

Ille N, Berg P and Scherg M 2002 Artifact correction of the ongoing EEG using spatial filters based on artifact and brain signal topographies *J. clin. neurophysiol.* **19** 113–24

Jung T, Makeig S, Humphries C, Lee T, McKeown M, Iragui V and Sejnowski T 2000 Removing electroencephalographic artifacts by blind source separation *Psychophysiology* **37** 163–78

Kronland-Martinet R, Morlet J and Grossmann A 1988 Analysis of sound patterns through wavelet transforms *Int. J. Pattern Recognit. Artif. Intell.* **1** 273–301

Ksieżyk R, Blinowska K, Durka P, Szelenberger W and W Androsiuk 1998 Neural Networks with Wavelet Preprocessing in EEG Artifact Recognition *Proc. Medicon'98 Conf. (Lemesos, Cyprus)*

LeVan P, Urrestarazu E and Gotman J 2006 A system for automatic artifact removal in ictal scalp EEG based on independent component analysis and Bayesian classification *Clin. neurophysiol.* **117** 912–27

Li X, Yao X, Fox J and Jefferys J 2007 Interaction dynamics of neuronal oscillations analysed using wavelet transforms *J. Neurosci. Methods* **160** 178–85

Makeig S, Bell A, Jung T and Sejnowski T 1996 Independent component analysis of electroencephalographic data *Advances in Neural Information Processing Systems* vol 8 ed D Touretzky, M Mozer and M Haselmo (Cambridge, MA: MIT Press) pp 145–51

Mallat S 1997 *A Wavelet Tour of Signal Processing* (London: AP Professional)

Moretti D, Babiloni F, Carducci F, Cincotti F, Remondini E, Rossini P, Salinari S and Babiloni C 2003 Computerized processing of EEGEOGEMG artifacts for multicentric studies in EEG oscillations and event-related potentials *Int. J. Psychophysiol.* **47** 199–216

Norman R G, Pal I, Stewart C, Walsleben J A and Rapoport D M 2000 Interobserver agreement among sleep scorers from different centers in a large dataset *Sleep* **23** 901–8

Quiroga R Q, Sakowitz O, Başar E and Schurmann M 2001 Wavelet Transform in the analysis of the frequency composition of evoked potentials *Brain Res. Protocols* **8** 16–24

Slobounov S, Hallett M, Cao C and Newell K 2008 Modulation of cortical activity as a result of voluntary postural sway direction: An EEG study *Neurosci. Lett.* at press   **See endnote 2**

Stecker M 2002 The effects of automatic artifact rejection on evoked potential recordings *Comput. Biol. Med.* **32** 247–59

Tallon-Baudry C, Bertrand O, Delpuech C and Pernier J 1996 Stimulus specificity of phase-locked and non-phase-locked 40 Hz visual responses in human *J. Neurosci.* **16** 4240–9

Tang A, Sutherland M and McKinney C 2004 Validation of SOBI components from high-density EEG *Neuroimage* **25** 539–53

van de Velde M 2000 Signal validation in electroencephalography research *PhD thesis* Technische Universiteit Eindhoven, The Netherlands

van de Velde M, van Erpb G and Cluitmans P J M 1998 Detection of muscle artefact in the normal human awake EEG *Electroencephalogr. Clin. Neurophysiol.* **107** 149–58

Vialatte F, Solé-Casals J and Cichocki A 2007 EEG windowed statistical wavelet deviation for estimation of muscular artifacts *Proc. IEEE 32nd Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'07)* vol 4 (Honolulu, USA) pp 1161–4

Wallstrom G, Kass R, Miller A, Cohn J and Fox N 2004 Automatic correction of ocular artifacts in the EEG *Int. J. Psychophysiol.* **53** 105–19

Whitham E *et al* 2007 Scalp electrical recording during paralysis: quantitative evidence that EEG frequencies above 20 Hz are contaminated by EMG *Clin. Neurophysiol.* **118** 1877–88

Worrell G, Parish L, Cranstoun S L P, Jonas R, Baltuch G and Litt B 2004 High-frequency oscillations and seizure generation in neocortical epilepsy *Brain* **127** 1496–506   **See endnote 3**

Zikov T, Bibian S, Dumont G A and Huzmezan M 2002 A Wavelet Based De-Noising Technique For Ocular Artifact Correction Of The Electroencephalogram *Proc. IEEE 24th Int. Conf. on Engineering in Medicine and Biology Society (EMBS'02)* (Huston, TX) pp 109–16

Zimmermann R and Sharein E 2004 MEG and EEG show different sensitivity to myogenic artifacts *Neurol. Clin. Neurophysiol.* **78**   **See endnote 4**

Endnotes

(1) Author: Please check whether the edit in the sentence 'Not only chewing, sucking and swallowing . . . ' retains its intended sense.
(2) Author: Please update reference 'Slobounov *et al* (2008)' if possible.
(3) Author: Please check whether the name of author in reference 'Worrell *et al* (2004)' is okay as set.
(4) Author: Please provide the page range(s) for reference 'Zimmermann and Sharein (2004)'.
(5) Author: Please be aware that the colour figures in this article will only appear in colour in the web version. If you require colour in the printed journal and have not previously arranged it, please contact the Production Editor now.

Reference linking to the original articles

References with a volume and page number in blue have a clickable link to the original article created from data deposited by its publisher at CrossRef. Any anomalously unlinked references should be checked for accuracy. Pale purple is used for links to e-prints at arXiv.