On the statistical significance of event-related EEG desynchronization and synchronization in the time-frequency plane

Piotr J. Durka, Jarosław Żygierewicz, Hubert Klekowicz, Józef Ginter, Katarzyna J. Blinowska

IEEE Transactions on Biomedical Engineering, 2003. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Abstract

We propose and discuss a complete framework for estimating significant changes in the average time-frequency density of energy of event-related signals. Addressed issues include estimation of time-frequency energy density (matching pursuit and spectrogram), choice of resampling statistics to test the hypothesis of change in one small region (resel) and correction for multiplicity (False Discovery Rate). We present estimation of the significance of event-related EEG desynchronization and synchronization (ERD/ERS) in the time-frequency plane.

Complete software implementing all the discussed steps is freely available from the Internet at `http://brain.fuw.edu.pl/~durka/tfstat`.

Keywords

time-frequency, matching pursuit, spectrogram, STFT, event-related activity, false discovery rate, event related desynchronization and synchronization, ERD/ERS

I. Introduction

EVENT-RELATED changes of energy in different EEG frequency bands are an important indicator of the underlying brain processes. Sensory processing and motor behavior are connected with the localized decrease of power in certain frequency bands, particularly in the alpha band. This phenomenon was called event related desynchronisation (ERD). The following increase in power after the event was named event related synchronisation (ERS). These effects, particularly in the context of movement planning, were extensively studied by Pfurtscheller [1], who also defined the ERD/ERS as the change in power in particular frequency band relative to the pre-movement epoch. The investigation of the ERD/ERS phenomena, apart from the clinical and scientific merits, has also technological implications: it opens the possibility of a brain-computer interface design [2]. Therefore, the quantification of the EEG reactivity to the externally or internally paced events and the assessment of the significance of changes have broad implications.

The classical method of quantification of event-related desynchronization and synchronization (ERD/ERS) [1] consists of time averaging the squared values of the samples of single trials, band-pass filtered in a priori selected bands. In 1993 S. Makeig [3] presented event-related spectral perturbation, that is time-varying spectra, estimating the time-frequency dynamics of underlying processes. Such an exploratory approach, free of a priori assumptions on fixed frequency bands, is widely used in last years, c.f. [4][5][6][7] to quote only few related works.

Different methods were used for estimating the time-frequency density of signals energy—apart from the short-time Fourier transform (STFT), e.g. smoothed Wigner-Ville distributions [5][6] and continuous wavelet transforms [8][4]. All these functions estimate the same quantity—time-frequency energy density of the signal. However, results may vary significantly not only depending on the chosen method, but also on its parameters (chosen wavelet, smothing of Wigner-Ville distribution, etc.). This produces an additional—apart from the unavoidable inter-subject variability—noise in the published results and makes difficult comparison of different findings, also reducing the value of neurophysiological conclusions drawn from these works. On the other hand, it is hard to judge a priori which of the estimators is "better" when analyzing signals of unknown content, in the absence of well defined criteria.

In 2001 the application of adaptive approximations, implemented the matching pursuit (MP) algorithm, was proposed for ERS/ERD estimation in the time-frequency space [9]. It proved to be a robust estimator, offering the best time-frequency resolution and high sensitivity in the investigation of the microstructure of ERD/ERS. It revealed the detailed structure of the gamma activity and the dependence between beta and gamma components [9]. In [10] the spatio-temporal behavior of different components of mu and beta rhythms was studied—using MP estimates—in the context of their functional role in movement preparation. Finally, this paper provides a framework for comparing the performance

Laboratory of Medical Physics, Institute of Experimental Physics, Warsaw University, ul. Hoża 69, 00-681 Warszawa, Poland, http://brain.fuw.edu.pl.

of different time-frequency estimators in the task of detecting event-related changes of energy (accompanying software implements MP and STFT, extensions of the system are being developped).

Apart from the discrepancies in the approach to the very issue of *estimation* of the time-frequency energy density of signals, the question of *significance* of changes, indicated by these estimators, remains an open issue. It seems to be a crucial point, essential to any procedure which is supposed to bring clinical or research conclusions.

Previous approaches include e.g. [3], [4] and [5], where selected time-frequency regions were compared between different experimental conditions. But the most common need is to evaluate the significance of a given "burst" (time-frequency region of visible change) in relation to a reference period. When we see e.g. "some increase" in a given epoch and frequency, we must ask, before drawing any conclusions, whether this is a statistically significant effect or just a fluctuation. In spite of this, up to now only few works addressed this issue.

E.g. in [6] the significance of an occurence of a burst of activity was estimated and displayed for a single frequency band; the issue of multiplicity in case of several frequency bands was mentioned, but not investigated. In [7] resampling tests were performed for each frequency band separately, and their results were displayed together on a common time-frequency plane. A direct interpretation of such a map of "significant changes" may lead to neglecting the influence of multiple comparisons [11]. For testing the hypothesis of no change in *one* frequency band, we assume a significance level α, corresponding to the probability of type I error (rejecting a true hypothesis). This is valid for a single hypothesis, which includes a priori choice of the frequency band of interest. But in the exploratory analysis, when statistically significant results are displayed for an array of frequencies, the probability of type I error for the whole family of hypotheses can increase dramatically above the significance level assumed for a single frequency band. A region picked from such a map should not be interpreted as significant at the claimed significance level α.

In this paper we present and discuss a complete framework for estimating significant changes in the average time-frequency density of energy of event-related signals. The method is presented in the context of ERD and ERS of the brain electrical activity. It consists of the following steps:

1. Estimation of the time-frequency density of energy of single trials (we implement matching pursuit (MP) and compare its results with the spectrogram (STFT)).
2. Division of the time-frequency plane into *resels* (from res*olution* el*ement*s), for which the average energy density is calculated.
3. Choosing statistics for the null hypothesis of no change in the given resel compared to the reference epoch in the same frequency (we implement two resampling methods).
4. Selecting a threshold for the null hypothesis corrected by multiple comparisons (we propose the use of False Discovery Rate and compare it to the stepdown Bonferroni–Holmes procedure).
5. Display of the energy changes in significant resels.

The following section gives a general discussion of applied methods, and section III presents their application.

II. Methods

A. Estimation of signal's energy density

The two methods discussed below represent the extrema of the wide spectrum of possible time-frequency estimators: spectrogram offers a fast algorithm and well established properties, while MP has the highest presently available resolution at a cost of a high computational complexity. Both these methods offer uniform time-frequency resolution, as opposed to the time-scale approaches (wavelets).

A.1 Short time Fourier transform

Short time Fourier transform (STFT or spectrogram) divides the signal into overlapping epochs. Each of these epochs is multiplied by a window function and then subjected to the Fast Fourier Transform [12], providing spectrum with resolution dependent on the epoch's length. We used Hanning windows with overlap 1/2 of its length.

A.2 Matching pursuit

Matching pursuit (MP) is an algorithm for a sub-optimal solution of the NP-hard problem of an optimal approximation of a function in a redundant dictionary D. It was proposed in [13] for the adaptive time-frequency approximations of signals. In each of the steps the waveform g_{γ_n} is matched to the signal $R^n f$, which is the residual left after subtracting results of previous iterations:

$$\begin{cases} R^0 f = f \\ R^n f = < R^n f, g_{\gamma_n} > g_{\gamma_n} + R^{n+1} f \\ g_{\gamma_n} = \arg\max_{g_{\gamma_i} \in D} | < R^n f, g_{\gamma_i} > | \end{cases} \tag{1}$$

where $\arg\max_{g_{\gamma_i} \in D}$ means the g_{γ_i} giving the largest value of the product $| < R^n f, g_{\gamma_i} > |$.

Dictionaries (D) for time-frequency analysis of real signals are constructed from real Gabor functions:

$$g_\gamma(t) = K(\gamma) e^{-\pi\left(\frac{t-u}{s}\right)^2} \sin\left(2\pi\frac{\omega}{N}(t-u)+\phi\right) \tag{2}$$

N is the size of the signal, $K(\gamma)$ is such that $||g_\gamma|| = 1$, $\gamma = \{u, \omega, s, \phi\}$ denotes parameters of the dictionary's functions. For these parameters no particular sampling is *a priori* defined. In practical implementations we use subsets of the infinite space of possible dictionary's functions. However, any fixed scheme of subsampling this space introduces a statistical bias in the resulting parametrization. In [14] we proposed a solution in terms of MP with stochastic dictionaries, where the parameters of a dictionary's atoms are randomized before each decomposition, or drawn from flat distributions. Results presented in this paper were obtained with such a bias-free implementation.

For a complete dictionary the procedure converges to f, but in practice we use finite sums:

$$f \approx \sum_{n=0}^{M} < R^n f,\ g_{\gamma_n} > g_{\gamma_n} \tag{3}$$

From this decomposition we can derive an estimate $Ef(t,\omega)$ of the time-frequency energy density of signal f, by choosing only auto-terms from the Wigner distribution

$$\mathcal{W}f(t,\omega) = \int f\left(t+\frac{\tau}{2}\right)\overline{f\left(t-\frac{\tau}{2}\right)}\, e^{-i\omega\tau} d\tau \tag{4}$$

calculated for the expansion (3). This representation will be a priori free of cross-terms:

$$Ef(t,\omega) = \sum_{n=0}^{M} |< R^n f,\ g_{\gamma_n} >|^2\, \mathcal{W}g_{\gamma_n}(t,\omega) \tag{5}$$

B. Time-frequency resolution

Time-frequency resolution of signal's representation depends on a multitude of factors, which are even more complicated in the case of the averaged estimates of energy. A general lower bound is given by the uncertainty principle, which states (c.f. [15]) that the product of the time and frequency variances exceeds a constant, which for the frequency defined as inverse of the period (Hz)[1] equals to $\frac{1}{16\pi^2}$:

$$\sigma_t^2 \sigma_f^2 \geq \frac{1}{16\pi^2} \tag{6}$$

It can be proved that equality in this equation is achieved by complex Gabor functions; other functions give higher values of this product. Since the time and frequency spreads are proportional to the square root of the corresponding variances, minimum of their product reaches $\frac{1}{4\pi}$.

However, our attempts to estimate the statistical significance in small resels of area given by (6) resulted in increased "noise", i.e., detections of isolated changes in positions inconsistent across varying other parameters. Therefore we fixed the area of resels at $\frac{1}{2}$, which at least has certain statistical justification: standard, generally used sampling of the spectrogram gives $\frac{1}{2}$ as the product of the localization in time (window length less overlap) and in frequency (interval between estimated frequencies). This sampling is based upon statistically optimal properties, namely independent samples for a periodogram of a Gaussian random process (c.f. [16]). Other values of this parameter can be of course considered in practical applications: the software accompanying this paper allows to investigate the impact of changing this parameter.

C. Integration of MP maps

Unlike the spectrogram (STFT, section II-A.1), MP decomposition (section II-A.2) generates a continuous map of the energy density (5). From this decomposition a discrete map must be calculated with a finite resolution. The simplest solution is to sum for each resel the values of all the functions from the expansion[2] (3) in the center of each resel (t_i, ω_i):

$$E_{\text{point}}(t_i,\omega_i) = \sum_{n} |< R^n f,\ g_{\gamma_n} >|^2\, \mathcal{W}g_{\gamma_n}(t_i,\omega_i) \tag{7}$$

[1] For the angular frequency this constant would be equal to $\frac{1}{4}$

[2] In practice this sum will be of course finite; in this study we used 2 different stopping criteria: for the Dataset I iterations were stopped when 99.9% of signal's energy was explained, which resulted in about 400-500 waveforms per trial. For the Dataset II number of iterations was fixed at 1300. These criterions were chosen to ensure proper repreentation of high frequency structures at a minimum computational cost (number of iterations. Details of the MP procedure can be found in [14]

However, for certain structures or relatively large resels, (7) may not be representative for the amount of energy contained in given resel. Therefore we use the exact solution, obtained by integrating for each resel power of all the functions from expansion (3) within the ranges corresponding to the resel's boundaries:

$$E_{\mathrm{int}}(t_i,\omega_i)=\sum_n |< R^n f, g_{\gamma_n} >|^2 \int_{t_i-\frac{\Delta t}{2}}^{t_i+\frac{\Delta t}{2}} \int_{\omega_i-\frac{\Delta\omega}{2}}^{\omega_i+\frac{\Delta\omega}{2}} \mathcal{W} g_{\gamma_n}(t,\omega)dtd\omega \tag{8}$$

The difference between (7) and (8) is most significant for structures narrow in time or frequency relative to the dimensions of resels. In this study we used relatively large resel's area and the average difference in energy did not exceed 5-7%.

D. Statistics

D.1 Reference epoch

To quantify changes of the energy density we first define a reference epoch. Properties of the signal in this epoch should reflect the "normal" state, i.e. "stationary" properties of the signal, in the sense that the measured changes will be relative to these properties.

Strict assumption of stationarity of the signals in the reference epoch would make an elegant derivation of the applied statistics: the repetitions could be then treated as realizations of an ergodic process. Indeed, epochs of EEG up to 10 sec duration (recorded under constant behavioral conditions) are usually considered stationary [17]. However, the assumption of "constant behavioral conditions" can be probably challenged in some cases. We cannot test this assumption directly, since the usual length of the reference epoch is too short for a standard test of stationarity.[3] Nevertheless, bootstrapping the available data across the indexes corresponding to time and repetition (number of the trial) simultaneously does not require a strict assumption of ergodicity from the purely statistical point of view. But we must be aware that this fact does not diminish our responsability in the choice of the reference epoch, which in general should be long enough to represent the "reference" properties of the signal to which the changes will be related, and at the same time it should be distant enough from the beginning of the recorded epoch (to avoid the influence of border conditions in analysis) and the investigated phenomenon (not to include some event-related properties in the "reference").

D.2 Permutation test on the difference of means

The values of energy of all the N repetitions in each questioned resel will be compared to energies of resels within the corresponding frequency of the reference epoch. Lets denote the time indexes t_i of resels belonging to the reference epoch as $\{t_i, i \in \mathrm{ref}\}$ and their number contained in each frequency slice (defined by the frequency width of a resel) as $\#\{\mathrm{ref}\}$[4]. We have a total of N repetitions of signals and their time-frequency maps. So for each frequency we have a total of $N \cdot \#\{\mathrm{ref}\}$ values of energy in the corresponding resels. For each resel at coordinates $\{t_i, \omega_i\}$ we'll compare its energy averaged over N repetitions with the energy averaged over repetitions in resels from the reference epoch in the same frequency. Their difference can be written as:

$$\begin{gathered}\Delta E_{\mathrm{int}}(t_i,\omega_i)= \\ =\tfrac{1}{N}\sum_{k=1}^{N} E^k_{\mathrm{int}}(t_i,\omega_i)-\tfrac{1}{N\cdot\#\{\mathrm{ref}\}}\sum_{k=1}^{N}\sum_{j\in\mathrm{ref}} E^k_{\mathrm{int}}(t_j,\omega_i)= \\ =\overline{E(t_i,\omega_i)}-\overline{E(t_{\mathrm{ref}},\omega_i)}\end{gathered} \tag{9}$$

where the superscript "k" denotes the k-th repetition (out of N).

The values of $E^k_{\mathrm{int}}(t_i,\omega_i)$ tend to have (especially for the MP estimates) a highly non-normal distributions. It is caused mainly by the adaptivity and high resolution of the MP approximation. Namely, the areas where no coherent signal's structures are encountered by the algorithm are left completely "blank", which causes the appearance of peaks around zero in histograms presented in Figure 1. On the contrary, other estimates of energy density provide a more uniform filling of the time-frequency plane, because of lower resolution and cross-terms.

[3] Standard test for stationarity relies on dividing the questioned epoch into subepochs of the length exceeding the period of the lowest frequency present in the signal, and then applying a non-parametric test (e.g. sign test) to statistical descriptors of these subepochs [18]. Usual length of the reference epoch does not exceed seconds, so considreing the presence of low EEG frequencies (order of few Hz) we would have too few subepochs for a reasonable application of a low-power non-parametric test

[4] $\#\{\mathrm{ref}\}$ is the time length of the reference period measured in chosen time widths of resels.

Given these distributions, we cannot justify an application of a parametric test based upon the assumption of normality. In order to preserve high power of the test we estimate the distributions from the data themselves by means of resampling methods (c.f. [19]).

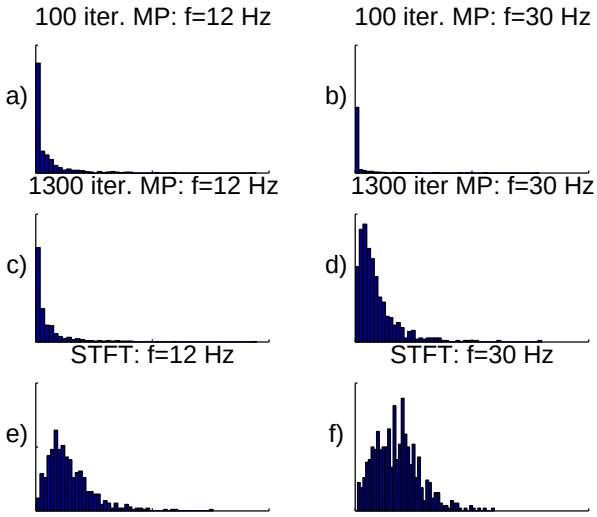

Fig. 1. Histograms of powers in the resels from the reference epoch of the analyzed signals (Dataset II) for two frequencies: 12 Hz (a, c and e) and 30 Hz (b, d and f). First four plots (a–d) are for the MP estimates: distribution is highly skewed toward the zero power, especially in cases when less waveforms are used for the representation (a and b). Lower panels (e and f) present histograms of powers estimated by STFT for the same frequency slices. Vertical scales (number of cases) differ, horizontal (energy) are kept constant across the subplots.

We are testing the null hypothesis of no difference in means between the values $\{E^k_{\rm int}(t_i,\omega_i), k \in 1\dots N\}$ of the i-th resel and all those in corresponding reference region $\{E^k_{\rm int}(t_j,\omega_i), k \in 1\dots N, j \in {\rm ref}\}$, that is energy values for the same frequency band for the resels falling into the reference epochs in all the repetitions (trials). A straightforward application of the idea of the permutation tests for each resel consists of:
1. mixing the values of $E_{\rm int}$ for all the repetitions of the resel under investigation and all the resels of the corresponding reference epoch,
2. drawing from this ensemble (without replacement) one of the $\binom{N\cdot(\#\{\rm ref\}+1)}{N}$ possible combinations (i.e. permutations not distinguishing different ordering), subdividing it into two sets containing N and $N\cdot\#\{\rm ref\}$ elements
3. calculating for the above sets the difference of means (9)
4. repeating the two above steps $N_{\rm rep}$ times and estimating from the obtained values distribution of the statistics (9) under the null hypothesis
5. based upon this distribution, calculating the p value for the actual difference of means

The number of permutations giving values of (9) exceeding the observed value has a binomial distribution for $N_{\rm rep}$ repetitions with probability α.[5] Its variance equals $N_{\rm rep}\alpha(1-\alpha)$. The relative error of α will be then (c.f. [19])

$$\frac{\sigma_\alpha}{\alpha} = \sqrt{\frac{(1-\alpha)}{\alpha N_{\rm rep}}} \tag{10}$$

To keep this relative error at 10% for a significance level α=5%, $N_{\rm rep} = 2000$ is enough. Unfortunately, due to the problem of multiple comparisons discussed in Section II-D.5, we need to work with much smaller values of α. In this study $N_{\rm rep}$ was set to $2\cdot10^5$ or $2\cdot10^6$, which resulted in large computation times. Therefore we implemented also a less computationally demanding approach, described in the following section.

D.3 Bootstrap with pseudo-t statistics

For a large number of points in each frequency of the reference region $\#\{\rm ref\}$, the statistics estimated in the previous section is dominated mainly by the values from the reference region, with little influence from the current resel. Estimation of its distribution only from the resels in the reference region would save a lot of computations. However, we still want to account for the different variances of $E^k_{\rm int}$ revealing the variability of the N repetitions. Therefore we replace the simple difference of means by the pseudo-t statistics:

[5]For the brevity we omit the distinction between the exact value α which would be estimated from all the possible repetitions, and the actually calculated

$$t = \frac{\Delta E_{\mathrm{int}}(t_i, \omega_i)}{s_\Delta} \tag{11}$$

Where ΔE_{int} is defined as in Eq. (9), and s_Δ is the pooled variance of the reference epoch and the investigated resel. We estimate the distribution of this magnitude from the data in the reference epoch (for each frequency $N \cdot \#\{\mathrm{ref}\}$ values) by drawing with replacement two samples of sizes N and $N \cdot \#\{\mathrm{ref}\}$ and calculating for each such replication statistics (11). This distribution is approximated once for each frequency. Then for each resel the actual value of (11) is compared to this distribution yielding p for the null hypothesis.

D.4 Computational complexity

Both presented above methods are computer-intensive, which nowadays causes no problems in standard applications. However, corrections for multiple comparisons imply much lower effective values of cutoff probabilities p_k (see the next section II-D.5). For the analysis presented in this study both the FDR (False Discovery Rate, see next section) and Bonferroni-Holmes adjustments gave critical values of the order of $5 \cdot 10^{-5}$ for the Dataset II (larger number of investigated resels, see section II-E "Experimental Data") and $5 \cdot 10^{-4}$ for Dataset I. If we set this for α in eq. (10), we obtain minimum of $N_{\mathrm{rep}} = 2 \cdot 10^6$ or $2 \cdot 10^5$ bootstrap or resampling repetitions to achieve 10% relative error for α.

Comparisons on presented datasets confirmed very similar results for both these procedures, as expected from the discussion in section II-D.3. This served as a justification to use in practice the pseudo-t bootstrapping, which is significantly faster.[6]

D.5 Adjustment for multiplicity

In the preceding two sections we estimated the achieved significance levels p for a null hypotheses of no change of the average energy in a single resel compared to the reference region in the same frequency. Statistics of these tests for different resels are not independent—neither in the time nor in the frequency dimension, and the structure of their correlation is hard to guess a *priori*.

Adjusting results for multiplicity is a very important issue in case of such a large amount of potentially correlated tests. We chose for this step the procedure assessing the False Discovery Rate (FDR, proposed in [20]). Unlike the adjustments for the family-wise error rate (FWE), it controls the ratio q of the number of the true null hypotheses rejected to all the rejected hypotheses. In our case this is the ratio of the number of resels to which significant changes are wrongly attributed to the total number of resels revealing changes.

The main result presented in [20] requires the test statistics to have positive regression dependency on each of the test statistics corresponding to the true null hypothesis. We use a slightly more conservative version, which controls FDR for all other forms of dependency. Let's denote the total number of performed tests, equal to the number of questioned resels, as m. If for m_0 of them the null hypothesis of no change is true, [20] proves that the following procedure controls the FDR at the level $q\frac{m_0}{m} \leq q$:

1. Order the achieved significance levels p_i, approximated in the previous section for all the resels separately, in an ascending series: $p_1 \leq p_2 \leq \cdots \leq p_m$
2. Find

$$k = \max\{i : p_i \leq \frac{i}{m \sum_{j=1}^{m} \frac{1}{j}} q\} \tag{12}$$

3. Reject all hypotheses for which $p \leq p_k$

Another, more conservative correction for multiplicity is the step-down Bonferroni-Holmes adjustment [11]. It relies on comparing the p_i values, ordered as in point (1.) above, to $\frac{\alpha}{m - i + 1}$, where α controls the family-wise error. We used $\alpha = q = 0.05$, which for the FDR relates to the possibility of erroneous detection of changes in one out of 20 resels indicated as significant.

According to the comparisons performed on the two discussed datasets, false discovery rate (FDR) resulted—as expected—to be less conservative that the stepdown Bonferroni-Holmes procedure. It has also a more appealing interpretation than the family-wise error (FWE) of Bonferroni-like procedures. Both these procedures take negligible amounts of computations.

E. Experimental data

We present results for two datasets: what we shall refer as the "Dataset I" comes from another study, described in [10]. Dataset II was collected especially for the purpose of this study. The main reason of this was the need for longer

[6] In the pseudo-t approach we estimate the distribution of the statistics for the null hypothesis only once for each frequency ($\mathcal{O}(N_{\mathrm{rep}})$), and then the test for each resel is only $\mathcal{O}(1)$. The full resampling estimating distributions for each resel is $\mathcal{O}(\mathcal{N}_{\mathrm{rep}}) \times$ the number of investigated resels.

than usual epochs of EEG between the stimuli repetitions to confirm the absence of false positives detections in neutral time epochs.

E.1 Dataset I

Twenty-year old right-handed subject was lying in a dim room with open eyes. Movements of index finger were performed approximately 5 seconds after a quiet sound generated every 10 to 14 seconds and detected by a micro-switch. The experiment was divided into 5 minutes long alternating sessions for right and left index finger movements, with 3 minutes breaks in between.

EEG was registered from electrodes placed at positions selected from the extended 10-20 system, sampled at 256 Hz, analog bandpass filtered in the 0.5–100 Hz range and sub-sampled offline to 128 Hz. For data processing, EEG was divided into 8 seconds long epochs, with movement onset in the 5th second. We present analysis of 57 artifact-free trials of right hand finger movement from the electrode C1 referenced to Cp1, Fc1, Cz and C3 (local average reference).

E.2 Dataset II

Thirty one-year old right-handed subject was half lying in a dim room with open eyes. Movement of the thumb, detected by a micro-switch, were performed approximately 5 seconds (at a subject's choice) after a quiet sound generated approximately every 20 seconds. Experiment was divided into 15-minutes sessions, and recorded EEG into 20-sec long epochs with the movement occurring in the 12th second. After artifacts rejection, 124 epochs were left for the analysis presented hereby.

EEG was registered from electrodes at positions selected from the 10-20 system. For the analysis we choose the C4 electrode (contra-lateral to the hand performing movements) in the local average reference. Analog filters were 100-Hz low-pass and 50-Hz notch. Signal was sampled at 250 Hz and down-sampled offline to 125 Hz.

III. Results and Discussion

Example applications of the proposed methodology to the datasets described in sections II-E.1 and II-E.2 are presented in Figures 2–4. These plots were constructed only for some of the possible parameters; results and graphics for other parameters can be easily reproduced using the software and data available via Internet.

Figures 2 and 3 present results for the Dataset I (section II-E.1) in the frequency range 0-40Hz. Normally it is enough to investigate only the frequency range of interest, e.g. from 5 Hz up, but we wanted to show that the applied statistical procedures are robust also in the low frequencies. Since the signals were not detrended before decomposition, we have most of the energy concentrated in low frequencies. This deteriorates significantly the possibility of presentation of the whole energy spectrum at once, so for the display (panels a) we used the logarithmic scale (for all the further computations the actual values of energy were used). Statistically significant regions in Figure 2f clearly relate to the known phenomena: μ desynchronization (marked as A), desynchonization of the μ harmonic (B), post-movement β synchronization (C) and desynchronization of the harmonic of β (D). We observe that the low-frequency non-stationarities present e.g. around the 5th second (probably movement artifact) do not show up as a statistically significant effects.

Similarly, Figures 4 and 5 present results for the Dataset II (section II-E.2) in the same frequency range. This dataset was collected with longer inter-movement intervals so we could analyze longer epochs. As expected, we have no significant effects more than 1-2 seconds away from the movement onset, except for the two resels present in the STFT results (Figure 5b)—these can be attributed to the 5% of false discoveries (section II-D.5).

A. Time-frequency resolution

Due to the considerations from the section II-B, we calculated the significant changes in resels relating to the same time-frequency resolution for both MP and STFT. However, it by no means implies that the resolution of MP and STFT are leveled by this approach. Within the significant resels of size equivalent to the resolution of the STFT, we can display the fine microstructure revealed by the MP estimator. Also the energy estimated by MP within the resels of the same size as STFT gives higher values of maximum ERD/ERS. Both these effects are clearly visible in Figures 2 and 4, as compared to Figures 3 and 5. For the Dataset I ERD/ERS estimated by STFT reach -51/65%, while MP gives estimates between -90 and 409%. Similarly for Dataset II (Figures 4 and 5) we got -32/44% for STFT and -68/209% for MP.

Nevertheless, in spite of the generally better sensitivity and resolution of MP, we observe that those two methods give similar and consistent results. Taking into account the high computational cost of the MP procedure, we may consider the STFT estimator as an alternative for cases when speed is more important than sensitivity and resolution.

B. Statistics

In an exploratory approach to the delimination of significant "bursts" of energy, statistical tests for different frequency bands cannot be treated separately if we want to talk about some significance level of the whole procedure. On the other hand, dramatic loss of power incurred by the Bonferroni correction in this setup has led to neglect the issue of multiplicity,

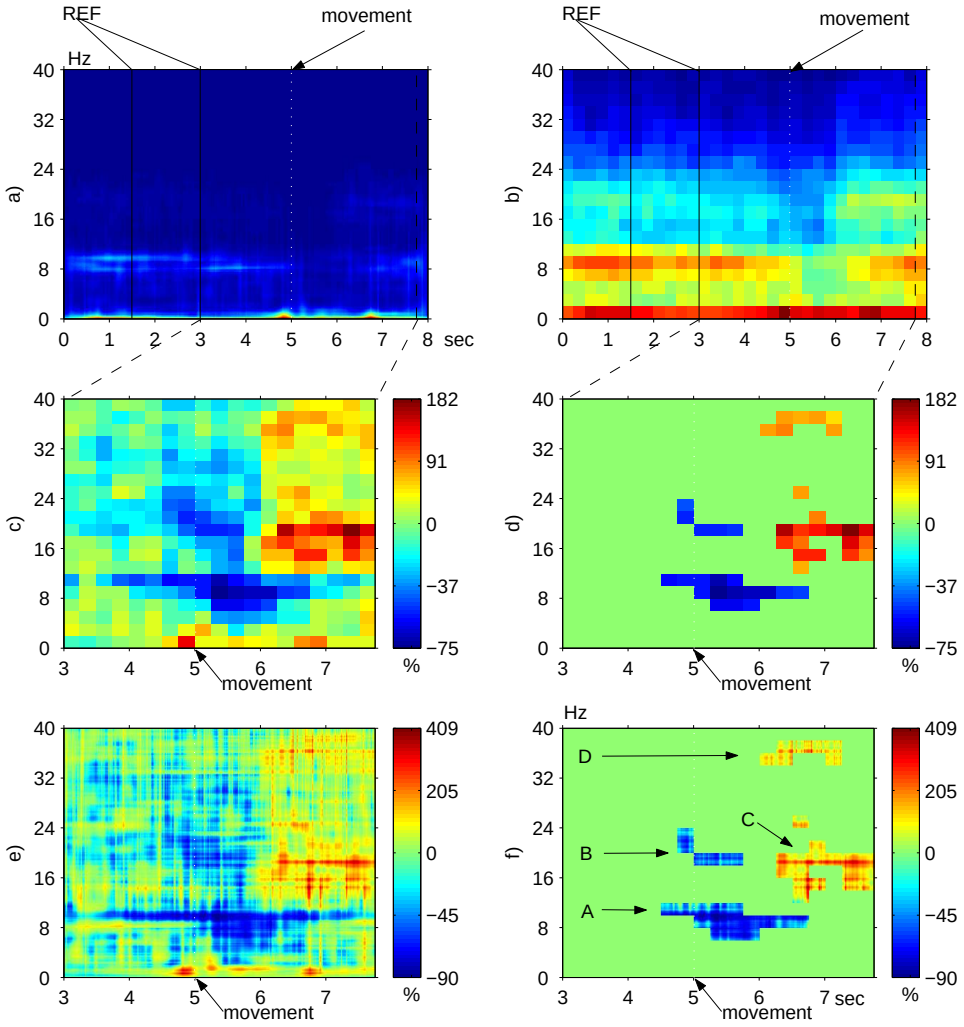

Fig. 2. Calculating the high resolution ERD/ERS from the MP decomposition in statistically significant regions (Dataset I, sec. II-E.1). **a)** average time-frequency energy density approximated from the MP decomposition (eq. 5), for clarity presented in the logarithmic scale (in further computations the actual values of energy are used). Reference epoch marked by black vertical lines, movement onset in the fifth second marked by white dashed line. **b)** energy from (a) integrated in resels 0.25s×2Hz **c)** average values of ERD/ERS calculated for the time from the end of the reference epoch to the end of the recorded epoch (black dashed vertical lines in a and b, the last resel dropped to avoid border conditions) **d)** ERD/ERS from (c) indicated as statistically different from the reference epoch by the pseudo-t bootstrap procedure (sec II-D.3) corrected by a 5% FDR (sec. II-D.5) **e)** high resolution map of ERD/ERS calculated from (a) **f)** high resolution ERD/ERS in statistically significant regions from (d): A—μ desynchronization, B—desynchonization of the μ harmonic, C—post-movement β synchronization, D—harmonic of β. Horizontal scales in seconds, vertical in Hz.

and hence the lack of a statistically correct way to delimit the significant changes over the entire time-frequequency range of interest.

Results in Figures 2–3 suggest that application of nonparametric statistics combined with properly chosen correction for multiplicity (FDR or Bonferroni-Holmes) preserves the power needed to properly detect significant changes even in the case of a low number of repetitions (57).

Figures 4–5 present the performance of these statistics in a case designed especially to contain large (over 80%) time epochs where no activation was expected. This increases artificially the size of the problem which would make the Bonferroni correction unusable, but FDR still seems to lead to perfectly reasonable results.

Among the proposed and tested methods, bootstrap estimation of the pseudo-t statistics in the reference region (section II-D.3) and FDR correction for multiple comparisons (section II-D.5) seem to be the methods of choice, offering good accuracy at a reasonable computational cost. As expected, FDR proved to be less conservative than Bonferroni-Holmes correction. It provided significances in the area coherent with other studies for the MP estimates. When applied to STFT it usually left out some significances in isolated resels (c.f. Figure 5) unrelated to known physiological phenomena, which can be accounted for by the allowed 5% of false discoveries. Application of the Bonferroni-Holmes correction cleared the dubious resels from Figure 5, but this should not be interpreted as suggesting the use of this correction for the STFT in general.

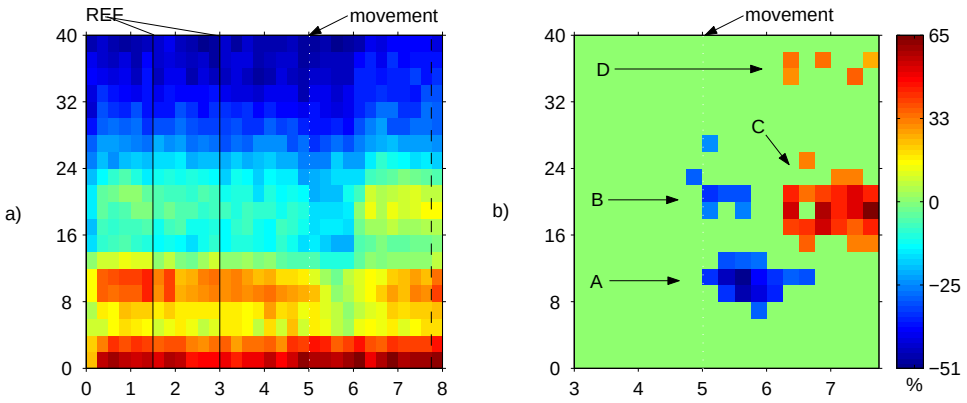

Fig. 3. (a) STFT estimate of power displayed in the logarithmic scale for the Dataset I (same as in Figure 2). (b) ERD/ERS calculated for the same epoch as in Figure 2), displayed for resels revealing significant change (Section II-D.2) corrected by a 5% FDR (Section II-D.5). Epochs and areas marked as in Figure 2. Horizontal scales in seconds, vertical in Hz.

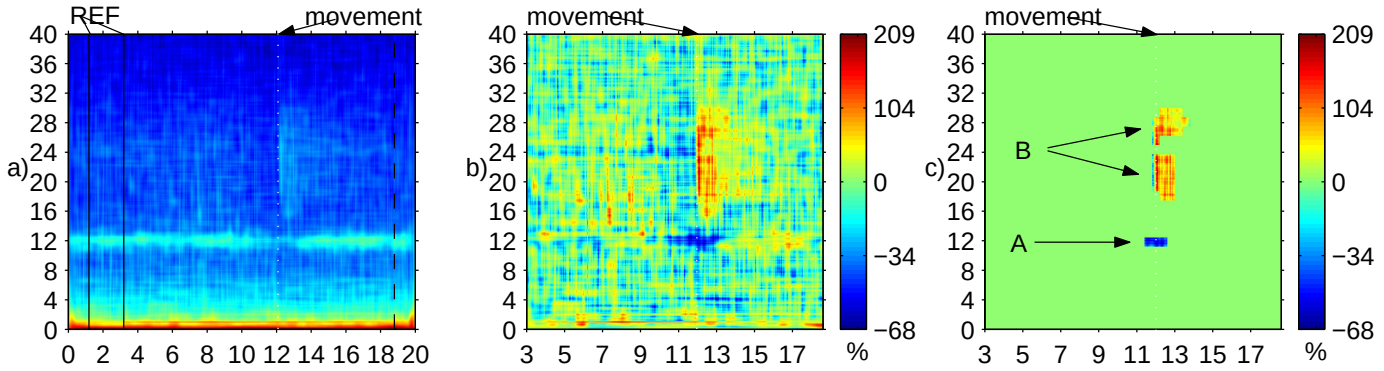

Fig. 4. MP results for the Dataset II (sec. II-E.2), where long epochs of EEG were recorded prior to the movement to test for the absence of false positive detecions in the stationary pre-movement epoch. **a)** average time-frequency energy density approximated from the MP decomposition (eq. 5), for clarity presented in logarithmic scale (in further computations the actual values of energy are used). Reference epoch marked by vertical lines. **b)** High resolution map of ERD/ERS **c)** High resolution ERD/ERS in statistically significant regions from (b), resel size 0.4s × 1.25Hz: we observe the α desynchronization (A) and synchronization of β in the 18-30Hz band (B), divided in two by the desynchronization of α harmonic in 24 Hz. Horizontal scales in seconds, vertical in Hz.

C. Reference epoch

The experiment providing the Dataset II (sec. II-E.2) was designed especially to allow different settings of the reference epoch, owing to the long pre-movement epoch of recorded EEG. Figures 4 and 5 present results for 2-seconds long reference epoch, positioned far away from the movement onset. This indicates the robustness of presented methodology, which gives no false positive detections in the long pre-movement epoch.

Availability of such a long pre-movement EEG allows also to test different choices of the reference epoch. We found that all the settings consistent with the general considerations from section II-D.1 (including a 11 sec long reference) give similar results, i.e. resulting statistics designates similar time-frequency area of significant changes. These figures are not presented, but experiments with different setting of this and other parameters on the datasets used in this study can be easily reproduced using the software and datasets freely available via Internet.[7]

IV. Conclusions

The above results indicate that statistically correct delimination of significant time-frequency regions of ERD/ERS is possible over an arbitrary time and frequency ranges (e.g. no high-pass filtering or a priori choice of the frequency band of interest are needed). Careful choice of the statistics and correction for multiplicity retains the power of the procedure also in cases of a low number of repetitions (57) or an artificially increased size of the problem (calculated for an area where over 80% of the time was expected to exhibit no ERD/ERS).

In estimating the time-frequency energy density, MP offers better sensitivity and higher resolution than STFT. However, both these methods give similar and consistent results in the context of delimination of time-frequency regions

[7]Playing with different settings on the two presented datasets involves only changing designated parameters in configuration files; adapting the system to a different dataset requires writing a new configuration file. One has to be also aware that the applied statistical procedures are computer-intensive, therefore in some cases computations can take up to hours.

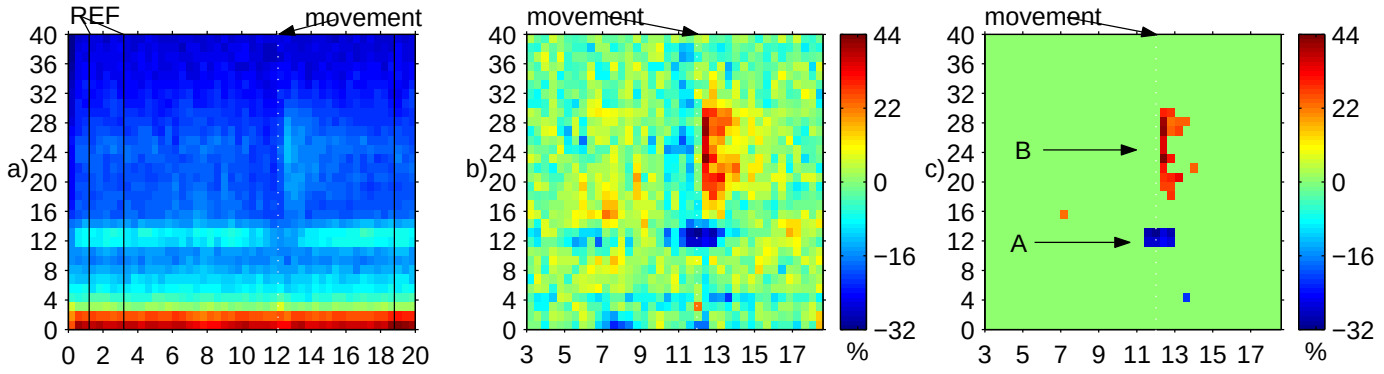

Fig. 5. STFT results for the Dataset II (sec. II-E.2, movement in the 12th second), the same as analyzed in Figure 4 using MP estimates. The harmonic of α is hardly visible in (a) and (b) and its effect is absent in (c). Two isolated resels indicated as significant can be accounted to the allowed 5% of false discoveries (section II-D.5). Horizontal scales in seconds, vertical in Hz.

of significant changes of energy. Taking into account the high computational cost of the MP procedure, we may consider the STFT estimator as an alternative for cases when speed is more important than sensitivity.

Presented methodology and software gives the missing criterion for the significance of event-related changes in the time-frequency energy density of signals. It introduces objective criteria to related neurophysiological research and allows to investigate quantitatively e.g. the following issues:

- minimal number of experiment's repetitions and duration of each trial, required to correctly delimit the significant ERD/ERS area,
- application of other than MP and STFT estimators of the time-frequency energy density,
- minimal number of MP iterations (accuracy) in decomposition of each trial, needed to achieve statistically significant results,
- choice of the time-frequency resolution for calculating statistics,
- influence of the (fixed) time position of the reference epoch.

V. Acknowledgments

This work was supported by the grant of Committee for Scientific Research (Poland) 4T11E 028023.

A. Reproducible Research

Complete software and datasets used for all the calculations and figures presented in this study is available from `http://brain.fuw.edu.pl/~durka/tfstat`. Calculation of time-frequency energy density from the MP parametrization, statistics and display are written in Matlab and available as m-files. Software for the matching pursuit decomposition in stochastic dictionaries is written in C and available with complete source code and executables for GNU/Linux and MS Windows at `http://brain.fuw.edu.pl/mp`.

References

[1] G. Pfurtscheller, "EEG event-related desynchronization (ERD) and event-related synchronization (ERS)," in *Electroencephalography: Basic Principles, Clinical Applications and Related Fields*, E. Niedermayer and F. Lopes Da Silva, Eds., pp. 958–965. Williams & Wilkins, fourth edition, 1999.

[2] J. R. Wolpaw, N. Birbaumer, D. McFarland, G. Pfurtscheller, and T.M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, pp. 767–791, 2002.

[3] S. Makeig, "Auditory event-related dynamics of the EEG spectrum and effects of exposure to tones," *Electroencephalogr Clin Neurophysiol*, vol. 86, pp. 283–293, 1993.

[4] C. Tallon-Baudry, A. Kreiter, and O. Bertrand, "Sustained and transient oscillatory responses in the gamma and beta bands in a visual short-term memory task in humans," *Visual Neuroscience*, vol. 16, pp. 449–459, 1999.

[5] E. Rodriguez, N. George, J.-P. Lachaux, J. Martinerie, B. Renault, and F. J. Varela, "Perception's shadow: long-distance synchronization of human brain activity," *Nature*, vol. 397, pp. 430–433, 1999.

[6] J.-P. Lachaux, E. Rodriguez, J. Martinerie, C. Adam, D. Hasboun, and F. J. Varela, "A quantitative study of gamma-band activity in human intracranial recordings triggered by visual simuli," *European Journal of Neuroscience*, vol. 12, pp. 2608–2622, 2000.

[7] B. Graimann, J. E. Huggins, S. P. Levine, and G. Pfurtscheller, "Visualization of significant ERD/ERS patterns in multichannel EEG and ECoG data," *Clinical Neurophysiology*, vol. 113, pp. 43–47, 2002.

[8] O. Bertrand and C. Tallon-Baudry, "Oscillatory gamma activity in humans: a possible role for object representation," *International Journal of Psychophysiology*, vol. 38, pp. 211–223, 2000.

[9] P. J. Durka, D. Ircha, Ch. Neuper, and G. Pfurtscheller, "Time-frequency microstructure of event-related desynchronization and synchronization," *Med Biol Eng Comput*, vol. 39, no. 3, pp. 315–321, May 2001.

[10] J. Ginter Jr, K. J. Blinowska, M. Kamiński, and P. J. Durka, "Phase and amplitude analysis in time-frequency space—application to voluntary finger movement," *Journal of Neuroscience Methods*, vol. 110, pp. 113–124, 2001.

[11] P. H. Westfall and S. S. Young, *Resampling-Based Multiple Testing*, John Wiley & Sons, 1993.

[12] *Signal Processing User's Guide for use with MATLAB*, The MathWorks Inc., 2002.

[13] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Tran Signal Process*, vol. 41, pp. 3397–3415, Dec 1993.
[14] P. J. Durka, D. Ircha, and K. J. Blinowska, "Stochastic time-frequency dictionaries for matching pursuit," *IEEE Tran Signal Process*, vol. 49, no. 3, pp. 507–510, March 2001.
[15] L. Cohen, *Time-Frequency Analysis*, Prentice Hall, 1995.
[16] M. B. Priestley, *Spectral Analysis and Time Series*, Academic Press, 1981.
[17] E. Niedermayer and F. Lopes Da Silva, *Electroencephalography: Basic Principles, Clinical Applications and Related Fields*, Williams & Wilkins, fourth edition, 1999, p. 1154.
[18] J. S. Bendat and A. G. Piersol, *Random data: Analysis and measurement procedures*, John Wiley & Sons, 1971.
[19] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, 1993.
[20] Y. Benjamini and Y. Yekutieli, "The control of the false discovery rate under dependency," *Ann. Stat.*, vol. 29, pp. 1165–1188, 2001.

Piotr J. Durka received MSc and PhD in physics from the Warsaw University, where he is currently assistant professor in the Laboratory of Medical Physics. His research includes time-frequency signal analysis and a pursuit to make the clinical electroencephalography a science rather than an art. Methods rely on designing correct and relevant EEG parametrization. Its verification requires reproducible research and open exchange of data, algorithms and publications. More information, papers and software are available at `http://durka.info`.

Jarosław Żygierewicz was born in 1971. He received the M.Sc. degree in physics in 1995 and the Ph.D. in medical physics in 2000 from the Warsaw University. His research interests are the analysis of biological signals and modeling of bioelectric phenomena. Currently he is an Assistant Professor in Physics Faculty of Warsaw University.

Hubert Klekowicz received the M.Sc. degree in physics from the Warsaw University in 2000. His research interests include EEG signal analysis and it's clinical applications. Currently he concentrates on investigating methods of detecting artifacts in sleep EEG and sleep stages classification.

Józef Ginter graduated in physics (1999) at Warsaw University. In his studies dealing with the analysis of EEG signals that accompany real and imagined movements, he links the mathematical approach with his own experiments.

Katarzyna J. Blinowska received M. Sc. and in 1969 Ph. D. degree in experimental physics from Faculty of Physics, Warsaw University, where she has been employed since graduation till now. In 1979 she received Sc.D. (doctor habilitatus) degree for her research on biological signal analysis. During 1982-1983 she was invited professor at the University of Southern California. Since 1984 she is a head of Medical Physics Laboratory at Warsaw University. In 1994 she became Full Professor. Her main research interests concern investigation of biophysical basis of electrical activity of central nervous system by modeling and signal analysis.