MACHINE LEARNING FOR SIGNAL PROCESSING (MLSP 2005)

BLIND SOURCE SEPARATION AND SPARSE BUMP MODELLING OF TIME FREQUENCY REPRESENTATION OF EEG SIGNALS: NEW TOOLS FOR EARLY DETECTION OF ALZHEIMER'S DISEASE

François Vialatte[1], *Andrzej Cichocki*[2], *Gerard Dreyfus*[1],
Toshimitsu Musha[3], *Tomasz M. Rutkowski*[2], *Rémi Gervais*[4]

[1] ESPCI (ParisTech), Laboratoire d'Electronique (CNRS UMR 7084)
10 rue Vauquelin, 75005 Paris, France
(francois.vialatte@espci.fr, gerard.dreyfus@espci.fr)
[2] RIKEN BSI, Laboratory for Advanced Brain Signal Processing
2-1 Hirosawa, Wako, Saitama, 351-0198, Japan
(cia@brain.riken.jp, tomek@brain.riken.jp)
[3] Brain Functions Laboratory Inc., KSP Building E211
Sakado, Takatsu Kawasaki-shi, Kanagawa, 213-0012, Japan
(musha@bfl.co.jp)
[4] Equipe Neurobiologie de la Mémoire Olfactive Institut des Sciences Cognitives
(UMR 5015 CNRS UCB), 67 Boulevard Pinel, 69675 Bron Cedex, France
(gervais@isc.cnrs.fr)

ABSTRACT

The early detection of Alzheimer's disease (AD) is an important challenge. In this paper, we propose a novel method for early detection of AD using only electroencephalographic (EEG) recordings for patients with Mild Cognitive Impairment (MCI) without any clinical symptoms of the disease who later developed AD. In our method, first a blind source separation algorithm is applied to extract the most significant spatiotemporal uncorrelated components; afterward these components are wavelet transformed; subsequently the wavelets or more generally time frequency representation (TFR) is approximated with sparse bump modeling approach. Finally, reliable and discriminant features are selected and reduced with orthogonal forward regression and the random probe methods. The proposed features were finally fed to a simple neural network classifier. The presented method leads to a substantially improved performance (93% correctly classified - improved sensitivity and specificity) over classification results previously published on the same set of data. We hope that the new computational and machine learning tools provide some new insights in a wide range of clinical settings, both diagnostic and predictive.

1. INTRODUCTION

Alzheimers disease (AD) is the most common neurodegenerative disorder [1, 2]. Since the number of individuals with AD is expected to increase in the near future due to societies aging phenomenon, early diagnosis and effective treatment of such brain degenerative disease are challenging issues for neurophysiological research [1]. Physiological or clinic studies have both showed that AD is characterized by a presymptomatic phase, usually lasting a few years, during which neuronal degeneration is occurring prior to the clinical symptoms appearance. This poses both a challenge: how do we identify individuals during this preclinical period; as well as an opportunity: can preventive therapy be started during the preclinical period before disease symptoms appear? Therefore, a major goal is to improve the performance of early detection of this disease by developing advanced computational and machine learning tools, especially for analysis of EEG data. Since an early detection method should be inexpensive, in order to allow simple and possibly mass screening of elderly patients electroencephalography (EEG) the most promising candidate in that respect [1, 2, 3].

Due to high complexity and variability of EEG signals, early detection of AD based on EEG depends on development of advanced computational tools [2]. In [3], Blind Source Separation (BSS) approach was first applied for the above purposes, while standard methods were used for feature extraction and classification.

In the present paper, we propose a multistage procedure (see Fig. 1) employing BSS for filtering/enhancement

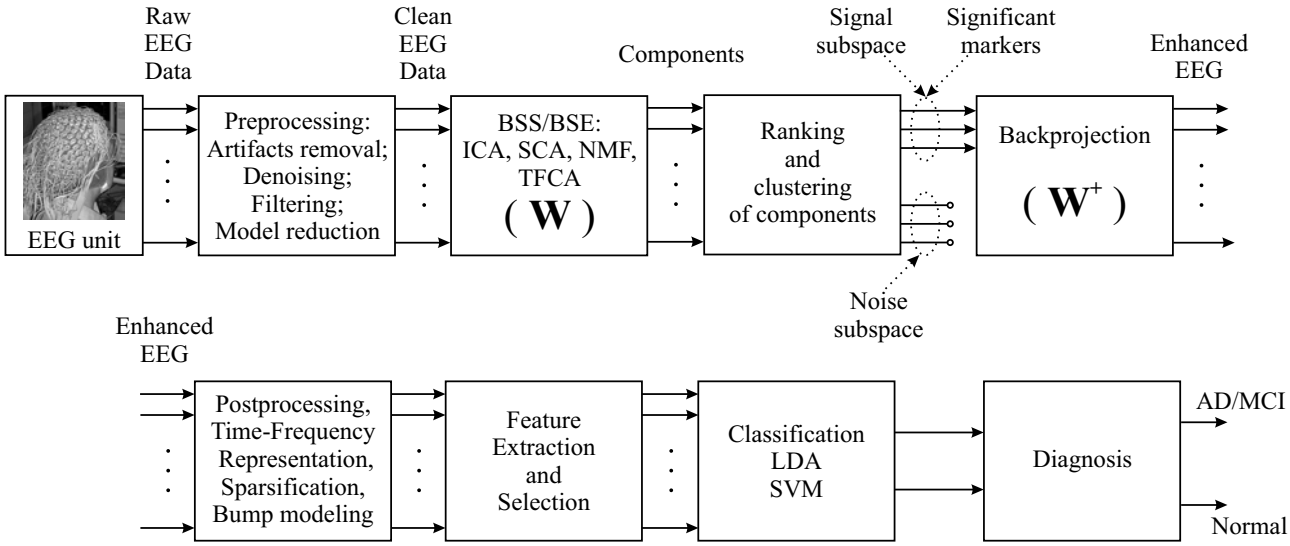

Fig. 1. Schema of the new method applied to obtain reliable features from raw EEG signals.

of EEG signals, time frequency representation, subsequent "bump modeling" for feature generation and dimensionality reduction (see Fig. 2), and for statistical feature selection. We show that such multistage approach provides a substantial improvement in discrimination between AD cases and healthy subjects, as compared to similar classification results obtained previously [1, 3] with the same data set.

2. METHODS

In this section several necessary steps of proposed approach, are illustrated in Fig. 1. The most important steps are: EEG signals preprocessing which remove some noise and artifacts, optional blind sources separation for EEG denoising and enhancement, wavelets or time frequency representation (TFR) with sparse bumps modeling and feature extraction. Since the number of features is relative large model reduction/selection plays also a key role. The EEG signals were recorded for three age matched groups: Healthy control, Mild Cognitive Impaired (MCI) patients and mild/severe AD diagnosed patients. In that study, patients who complained only for memory impairment, but had no apparent loss in general cognitive, behavioral, or functional status, were recruited. Fifty-three patients of this group met the following criteria for MCI: MMSE score 24 or higher, Clinical Dementia Rating (CDR) scale score of 0.5 with memory performance less than one standard deviation below the normal reference (Wechsler Logical Memory Scale and Paired Associates Learning subtests, IV and VII, ≤ 9, and/or ≤ 5 on the 30 min delayed recall of the Rey-Osterreith figure test). These patients were followed clinically for $12-18$ months. Twenty-five of them developed probable or possible AD according to NINDS-ADRDA criteria. Normal age-matched controls were recruited from family members of the patients (mainly spouses) participated in the study as control group. Both patients and controls underwent general medical, neurological, psychiatric, and neuroimaging (SPECT, CT and MRI) investigation for making the diagnosis more precise. EEG was recorded within one month after entering the study from all patients and controls, but only EEG recorded from the patients who progressed to AD ($n = 22$; below: MCI group) and age-matched controls $(n = 38)$ was used for the analysis. No patient or control subject received psychotropic medication at the period when EEG was recorded. Mean MMSE score was 26 ± 1.8 in MCI group and 28.5 ± 1.6 in control group; age 71.9 ± 10.2 and 71.7 ± 8.3, respectively. EEG recording was done in an awake resting state with eyes closed, under vigilance control. Ag/AgCl electrodes (disks of diameter 8 mm) were placed on 21 sites according to 10/20 international system, with the reference electrode on the right ear-lobe. EEG was recorded with Biotop 6R12 (NEC Sanei, Tokyo, Japan) using sampling frequency of 200Hz (see [1, 3] for more detail).

2.1. Blind source separation for signal denoising

According to the currently prevailing view of EEG signal processing, those signals can be modeled as a linear mixture of a finite number of sources with additive noise [4]. Therefore, blind source separation techniques can be used advantageously for decomposing recorded EEG into brain signal related subspace and noise subspace. The AMUSE (Algorithm for Multiple Unknown Signals Extraction [5]) algorithm, a blind source separation technique that relies on second-order statistics for spatiotemporal decorrelation,

was previously used in order to select five significant components of the signal that had the best linear predictability. This algorithm uses simple principles that the estimated components should be spatiotemporally decorrelated and less complex (i.e., have better linear predictability) than any mixture of those sources. The components are ordered according to decreasing values of singular values of a time-delayed covariance matrix. As in PCA (Principal Component Analysis) and unlike in many ICA algorithms, all components estimated by AMUSE are uniquely defined (i.e., any run of algorithms on the same data will always produce the same components) and consistently ranked. Mathematically AMUSE algorithm is based on the following two stage procedure: In the first step we apply a standard or robust prewhitening (sphering) as linear transformation
$\mathbf{x}_1(k) = \mathbf{Q}\mathbf{x}(k)$ with

$$\mathbf{Q} = \mathbf{R}_x^{-1/2} = (\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T)^{-1/2} = \mathbf{V}(\mathbf{\Lambda})^{-1/2}\mathbf{V}^T \qquad (1)$$

where $\mathbf{R}_{xx} = E\{\mathbf{x}(k)\mathbf{x}^T(k)\}$ is a standard covariance matrix and $\mathbf{x}(k)$ is a vector of observed data for time instant k. In the second step, for pre-whitened data, the SVD (Singular Value Decomposition) is applied for time-delayed covariance matrix

$$\mathbf{R}_{x1x1} = E\{\mathbf{x}_1(k)\mathbf{x}_1^T(k-1)\} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \qquad (2)$$

where $\mathbf{\Sigma}$ is diagonal matrix with decreasing singular values and $\mathbf{U}$, $\mathbf{V}$ are orthogonal matrices of left and right singular vectors. Then, an unmixing (separating) matrix is estimated as $\mathbf{W} = \mathbf{U}^T\mathbf{Q}$.
AMUSE algorithm is much faster than the vast majority of BSS algorithms (its processing speed is mainly defined by the PCA processing within it) and is very easy to use, because no parameters are required. It is implemented as a part of package *ICALAB for Signal Processing* [6] freely available on authors' web site.

In the present paper, the same pre-processing method is used for collected 21 EEG channels, as a baseline for assessing the efficiency of sources detection. In order to remove components carrying noise, only first six components are kept. After such preprocessing procedure we obtain three databases with 21 deflated signals for each of the cases: MCI, Control (healthy subjects), and Severe set. We will name D the database featuring 60 recordings: 22 from MCI set and 38 from Control sets; and S the database featuring 45 recordings : 22 from MCI set and 23 from Severe AD sets.

2.2. Time-frequency maps and bump modeling for feature generation

In order to obtain a compact representation of the signals of D database suitable for automatic discrimination of MCI patients from control individuals, the signals are first analyzed in the time-frequency domain by wavelet transformation, and the resulting time-frequency maps were modeled by bumps [7], as described following subsections.

2.2.1. Wavelet transformation and time-frequency map generation

EEG signals are first transformed to time-frequency domain maps using wavelets. Complex Morlet wavelets are appropriate for time-frequency analysis of EEG signals [8]. Complex Morlet wavelets $w(t)$ of Gaussian shape in time (deviation σ_t) are defined as:

$$w(t) = A\exp\left(-t^2/2\sigma_t^2\right)\exp\left(2i\pi ft\right), \qquad (3)$$

where σ_t and f are appropriately chosen parameters, which cannot be chosen independently, since the product $\sigma_t f$ determines the number of periods that are present in the wavelet). In the present investigation, the wavelet family defined by $2\pi\sigma_t f = 7$ was chosen, as described in [8]. The signals present in the D database were wavelet transformed in two different frequency discrete ranges with steps of 0.25Hz frequency bins:

- 1.5 to 16.5Hz: in order to monitor *Theta range* of EEG (3.5 to 7.5Hz);
- 9 to 31.75Hz: in order to study *Beta range* of EEG (12.5 to 25Hz).

These ranges take into account borders needed for the bump modeling procedure. We obtained two databases of time–frequency maps, D_1 for *Beta range* and D_2 for *Theta range*. S database is wavelet transformed only in the *Theta range*.

2.2.2. Bump Modeling

The bump modeling [7] technique is a two-dimensional generalization of the Gaussian mesa function modeling technique that was initially designed for one-dimensional signals (electrocardiogram analysis - ECG) [9]. In the present study, it is used for extracting information from the time-frequency maps. This method of bump modeling was initially successfully applied to the analysis of local field potential signals, gathered from electrophysiological (invasive) measurements [10, 7]. The present paper reports an application of bump modeling to EEG signals. The main idea of this method is to approximate a time-frequency map with a set of predefined elementary parameterized (adaptive) functions called bumps (non-overlapped or overlapped). Therefore, the map is represented by the rather limited set of parameters of the bumps, which is a very sparse encoding, resulting in information compression rates ranging from one hundred to one thousand (further details are given in [7, 9]).

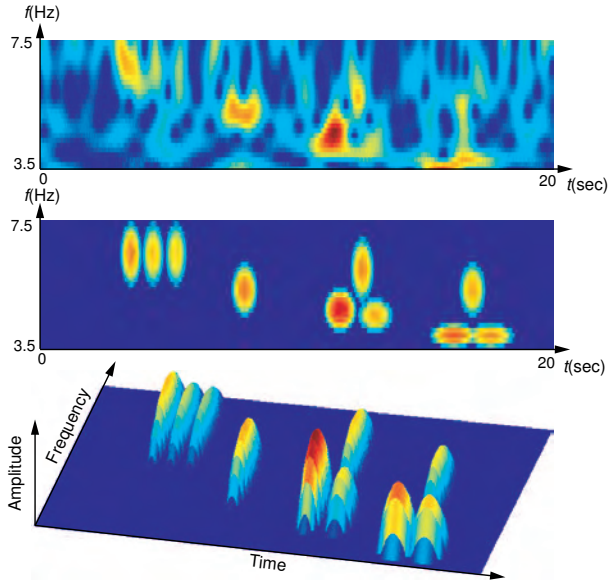

Fig. 2. Bump modeling of a recording, in the *Theta range*. Top panel: wavelet time-frequency map computed using complex Morlet wavelets. Middle panel: bump decomposition of the time-frequency map. Bottom panel: 3D view of the bump decomposition.

The algorithm is divided into following steps of normalized time-frequency maps analysis:

(i) Window the map in order to define the zones to be modeled (those windows forms a set of overlapped time-frequency sub-areas of the map);

(ii) Find the window that contains the maximum amount of energy;

(iii) Adapt a bump $\phi_b(f,t)$ to the selected zone, and withdraw it from the original map, the parameters of the bumps are computed in order to minimize the cost function C defined by:

$$C = \frac{1}{2} \sum_{t,f \in W} \left(z_{ft} - \phi_b(f,t)\right)^2, \qquad (4)$$

where the summation collects values of all the pixels within the window W. z_{ft} are the time-frequency coefficients at time t and frequency f, and $\phi_b(f,t)$ is the bump function value at time t and frequency f.

(iv) If the amount of energy of the modeled bumps reaches a threshold, stop; else return to (iii).

Half ellipsoids were found to be the most appropriate bump functions for the present application (Fig. 2 shows a typical example of bump modeling of the time-frequency map of an EEG recording). The principal axes of the bumps are restricted to be parallel to the time frequency axis in order to have the less possible parameters - each bump is described by five parameters: its coordinates on the map (two parameters), its amplitude (one parameter) and the lengths of its axes (two parameters). Half ellipsoids are defined by:

$$\phi_b(f,t) = \begin{cases} a\sqrt{1-\nu} & \text{for} \quad 0 \leq \nu \leq 1 \\ 0 & \text{for} \quad \nu > 1 \end{cases} \qquad (5)$$

where $\nu = (e_f^2 + e_t^2)$ with $e_f = (f - \mu_f)/l_f$ and $e_t = (t - \mu_t)/l_t$; μ_f and μ_t are the coordinates of the center of the ellipsoid, l_f and l_t are the half lengths of the principal axes, a is the amplitude of the function, t is the time and f the frequency. We applied bump modeling to D_1 and D_2 databases. The parameters of the bumps obtained are candidate features for classification. Although the model is sparse, feature selection is necessary because of the small size of the data sets.

2.3. Feature selection

After bump modeling, the signals under investigation are represented by the set of parameters that describe the bumps. The following features were defined and computed:

F_1**:** the number of bumps;

F_2**:** the sum of the amplitudes of the bumps present.

For D_1, bumps are studied in two separate frequency ranges: β_1 (12.5 - 17.5Hz) and β_2 (17.5 - 25Hz). For D_2 and S, the *Theta range* (3.5-7.5 Hz) is used.

Thus four features are computed per signal for D_1, and two for D_2 or S. Therefore, for each recording (21 time-frequency maps), the number of candidate features f_i is 84 for D_1 and 42 for D_2 or S. Since the number of candidate features remains still large, given the number of examples in the data bases, feature selection could be performed based on the Gram-Schmidt orthogonal forward regression (OFR) algorithm [10]. OFR consist the following steps:

(i) Select the candidate feature x_j that is most correlated to the quantity to be modeled: $x_j = \arg\max_i \cos^2 f_i$;

(ii) Project the output vector and all other candidate features onto the null space of the selected feature and compute the parameter pertaining to that feature;

(iii) Iterate in that subspace, return to (i).

This method of selection is applied three times, to obtain the best features in order to separate MCI sets from Controls ones, and MCI set for Severe set: first for D_1 set, then for D_2, and finally for S. Only a few features relevant to the classification are selected using OFR. Subsequently, in order to choose optimally the reduced number of features, the

random probe method [11] is applied. One hundred probes, i.e., realizations of random variables, are computed and appended to the feature set. A risk level is defined, which corresponds to the risk that a feature might be kept although, given the available data, it might be less relevant than the probe. At each step of the selection procedure, the following steps are performed iteratively:

(i) Obtain a candidate feature from OFR;

(ii) Compute the value of the cumulative distribution function of the rank of the probe for the rank of the candidate feature;

(iii) If that value is smaller than the risk, select the feature and return to (i);

(iv) Else, discard the feature under consideration and terminate.

Since the database is small compared to the number of features to be tested, the ensemble feature ranking method was used: E subsets are built by iteratively removing one example from the database (thus E = number of recordings in the database). OFR is then applied to these subsets. The overall distribution of features, and the average number N_k of selected features is computed; finally, the N_k overall best features are selected. Accepting a percentage P of false positive variables (i.e. irrelevant variables that are wrongly selected), we obtained finally F features, ranked as the most significant ones. P was set to 15%.

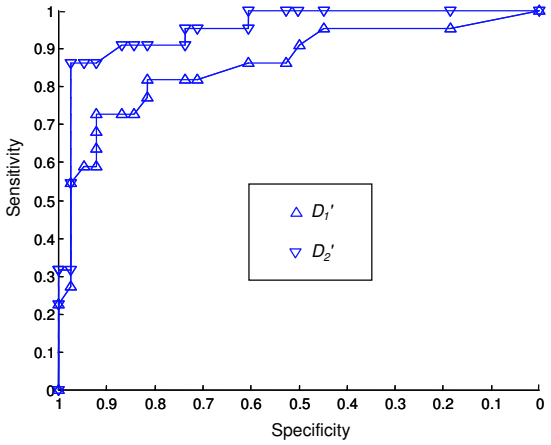

Fig. 3. R.O.C. curves of classification for D_1' (upward triangles) and for D_2' (downward triangles).

3. RESULTS

Each data set was used for training and validating a neural network classifier (multi-layer perceptron model, see for instance [12]). The generalization performance was estimated using the leave-one-out ("jackknifing") cross-validation method [13] which was also used in the previous study of the data set [3]. The best results, shown in Table 1 (D_1' and D_2' are obtained without denoising by AMUSE), are obtained with linear classifiers (no hidden layer).

In case of database D, the non-preprocessed signals are better classified than denoised ones, however when the method was applied to S, we observed that denoised signals show better performance. This emphasizes the question of how bump modelling and denoising algorithm may interfere, which still needs further investigation. ROC curves confirm the overall superiority of the *Theta range* for such classification task (see Fig. 3). The six best features for discrimination of MCI cases from Controls using D_1' data set are: in β_1 range, the number of bumps for signals corresponding to electrode F3 and the sum of bump amplitudes for signals corresponding to electrodes Cz, F7 and F3; and in β_2 range the number of bumps in electrodes FP2 and P4. Consideration of only these features leads to a validation error of 18.3%. The six best features for discimination of MCI cases from Controls using D_2' data set are: the number of bumps in signals of electrodes F8, 02, T7, C3, and FPz and the sum of bump amplitudes of signals from the electrode P4. The use of only these features leads to a validation error of 11.7%.

4. DISCUSSION

In the present paper, we reported a novel application of blind source separation combined with time frequency representation and sparse bump modeling to the automatic classification of EEG data for early detection of Alzheimers disease. The developed method was applied to EEG recordings that had been analyzed previously [3] with standard feature extraction and classification methods. Compared to that previous analysis, some further improvement was achieved, the overall correct classification rate was raised from 80% to 93% (sensitivity 86.4% and specificity 97.4%).

The task was to discriminate the EEG recordings of normal individuals from those of patients who developed Alzheimers disease in the period about eighteen months later. Therefore, the present study provides exciting prospects for early mass detection of the disease. The method is very cheap as compared to PET, SPECT and fMRI scans, requiring only a $21-$channel EEG apparatus. Note that short intervals (20 seconds) of artifact-free recording of spontaneous EEG is already sufficient for high accuracy of classification.

Sparse bump modeling appeared to be a useful tool for compressing information contained in EEG time-frequency maps. Amplitude variations and bursts of EEG oscillations are highly related to the brain state dynamics. Bump modeling can be a good approximation of bursts and sufficiently well follow important features of amplitude modulation of

Table 1. Number of subjects correctly and incorrectly classified by neural network models (SEN-sensitivity, SPE-specificity).

Data set	Misclassified	Correctly classified
D_1, β-*wave* range $(P = 15\%, F = 7)$	MCI = 8 / 22 Controls = 5 / 38	SEN = 63.6% SPE = 92.1% All = 78.3%
D_2, *Theta range* $(P = 15\%, F = 9)$	MCI = 5 / 22 Controls = 5 / 38	SEN = 77.2% SPE = 86.8% All = 83.3%
D_1', β *wave* range without denoising $(P = 15\%, F = 8)$	MCI = 6 / 22 Controls = 3 / 38	SEN = 72.7% SPE = 92.1% All = 85.0%
D_2' *Theta range* without denoising $(P = 15\%, F = 8)$	MCI = 3 / 22 Controls = 1 / 38	SEN = 86.4% SPE = 97.4% All = 93.3%
MCI Vs Severe AD denoised data $(P = 15\%, F = 5)$	Severe = 5 / 23 MCI = 4 / 22	SEN = 78.3% SPE = 81.8% All = 80.0%
MCI Vs Severe without denoising $(P = 15\%, F = 2)$	Severe = 11 / 23 MCI = 10 / 38	SEN = 52.3% SPE = 54.5% All = 53.3%
Previous study [3] (without bumps modeling)	MCI = 6 / 22 Controls = 6 / 38	SEN = 72.7% SPE = 84.2% All = 80.0%

EEG oscillations, therefore it can become a promising way of compressing information contained in EEG and be widely used for its analysis. Although our preliminary results are quite promising, a full validation of the method requires investigating more extensive databases. Furthermore, there is presumably a lot of information present in the recordings that is not yet exploited, such as the dynamics of the bumps and the brain functional connectivity. This will be the subject of our future reports.

5. ACKNOWLEDGMENTS

This work was partially supported by *Centre de Microelectronique de Paris-Ile de France*. François Vialatte is supported by a *MENESR* grant.

6. REFERENCES

[1] T. Musha, T. Asada, F. Yamashita, T. Kinoshita, Z. Chen, H. Matsuda, U. Masatake, and W.R. Shankle, "A new EEG method for estimating cortical neuronal impairment that is sensitive to early stage alzheimers disease," *Clinical Neurophysiology*, vol. 113, no. 7, pp. 1052–1058, 2002.

[2] J. Jeong, "EEG dynamics in patients with alzheimers disease," *Clinical Neurophysiology*, vol. 115, pp. 1490–1505, 2004.

[3] A. Cichocki, S.L. Shishkin, T. Muash, Z. Leonowicz, T. Asada, and T. Kurachi, "EEG filtering based on blind source separation (BSS) for early detection of alzheimers disease," *Clinical Neurophysiology*, vol. 116, no. 3, pp. 729–737, March 2005.

[4] A. Cichocki, "Blind signal processing methods for analyzing multichannel brain signals," *International Journal of Bioelectromagnetism*, vol. 6, no. 1, 2004.

[5] A. Cichocki and S. Amari, *Adaptative blind signal and image processing: learning algorithms and applications*, Wiley, New York, NY, 2003.

[6] A. Cichocki, S. Amari, K. Siwek, and T. Tanaka et al., "ICALAB toolboxes," available online at http://www.bsp.brain.riken.jp/ICALAB/.

[7] F. Vialatte, C. Martin, R. Dubois, B. Quenet, R. Gervais, and G. Dreyfus, "A machine learning approach to the analysis of time-frequency maps, and its application to neural dynamics," *Neural Networks*, (submitted).

[8] C. Tallon-Baudry, O. Bertrand, C. Delpuech, and J. Pernier, "Stimulus specificity of phase-locked and non-phaselocked 40Hz visual responses in human," *Journal of Neuroscience*, vol. 16, pp. 4240–4249, 1996.

[9] R. Dubois, B. Quenet, Y. Faisandier, and G. Dreyfus, "Building meaningful representations in knowledge-driven nonlinear modeling," *Neurocomputing*, p. (in print), 2005.

[10] F. Vialatte, C. Martin, N. Ravel, B. Quenet, G. Dreyfus, and R. Gervais, "Oscillatory activity, behaviour and memory, new approaches for lfp signal analysis," in *Proceedings of 35^{th} Annual General Meeting of the European Brain and Behaviour Society*, Barcelona, Spain, 17–20 September 2003, vol. 63 of *Acta Neurobiologiae Experimentalis*, supplement.

[11] S. Chen, S.A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *International Journal of Control*, vol. 50, pp. 1873–1896, 1989.

[12] H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar, "Ranking a random feature for variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1399–1414, 2003.

[13] S. Haykin, *Neural Networks a Comprehensive Foundation*, Prentice Hall, second edition, 1999.