IOP PUBLISHING PHYSIOLOGICAL MEASUREMENT

Physiol. Meas. **28** (2007) 335–347 doi:10.1088/0967-3334/28/4/001

Techniques for early detection of Alzheimer's disease using spontaneous EEG recordings

W L Woon[1,2], A Cichocki[1], F Vialatte[1] and T Musha[3]

[1] Laboratory for Advanced Brain Signal Processing, BSI, RIKEN, 2-1, Hirosawa, Wako Saitama 351-0198, Japan
[2] Malaysia University of Science and Technology, Unit GL33, Block C, Kelana Square, 17 Jln.SS7/26, 47301 PJ, Malaysia
[3] Brain Functions Laboratory, Inc., 1-1 JFE Keihin Bldg 8th Floor, Minami-Watarida-cho, Kawasaki-ku, Kawasaki 210-0855, Japan

E-mail: wlwoon@brain.riken.jp, cia@brain.riken.jp, fvialatte@brain.riken.jp and musha@bfl.co.jp

Received 11 October 2006, accepted for publication 12 February 2007
Published 7 March 2007
Online at stacks.iop.org/PM/28/335

Abstract
Alzheimer's disease (AD) is a degenerative disease which causes serious cognitive decline. Studies suggest that effective treatments for AD may be aided by the detection of the disease in its early stages, prior to extensive neuronal degeneration. In this paper, we propose a set of novel techniques which could help to perform this task, and present the results of experiments conducted to evaluate these approaches. The challenge is to discriminate between spontaneous EEG recordings from two groups of subjects: one afflicted with mild cognitive impairment and eventual AD and the other an age-matched control group. The classification results obtained indicate that the proposed methods are promising additions to the existing tools for detection of AD, though further research and experimentation with larger datasets is required to verify their effectiveness.

Keywords: MCI, Alzheimer's disease, EEG, CSP, entropy

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative condition which brings severe cognitive decline. To date, conclusive diagnosis of AD is achievable only by direct examination of affected brain tissues, which unfortunately is only possible posthumously (Turner 2003). In practical scenarios, diagnosis is commonly achieved either by the careful observation of clinical symptoms, such as cognitive decline, or via the detection of AD biomarkers, for

example the presence of certain metabolites or genes. A more recent approach is to look for markers of AD using medical imaging methods, such as SPECT and EEG. Unfortunately, while a wide selection of such biomarkers are available, the reliable detection of AD remains hugely challenging as none of these methods offer near-definite diagnoses. At present, clinical diagnosis of probable or possible AD has accuracy rates of about 90% and 50%, respectively (Turner 2003).

The low diagnosis rate with possible AD is unfortunate given the importance of early detection, especially with respect to treatment prospects. Studies have indicated that the biological changes leading to AD begin long before the appearance of clinical symptoms, by which time the neurological degeneration is likely to be fairly advanced and resistant to medical treatment. In this context it is noteworthy that AD has an early, pre-clinical stage during which neural degeneration is underway (Turner 2003, Blennow *et al* 2006). This condition is a subset of a more common ailment known as mild cognitive impairment (MCI), from which there is an expected 1–25% yearly conversion to AD (van der Hiele *et al* 2006). This strongly motivates the development of an effective yet practical method for screening individuals with MCI; given its low cost and widespread availability, EEG is an outstanding candidate for this purpose.

The rest of this paper is organized as follows. In section 2, the novelty and objectives of the research effort are discussed while section 3 describes the procedures observed during data collection, feature extraction and classification. Section 4 presents the outcome of the experiments. Finally, section 5 summarizes the findings of the study and proposes avenues for further research.

2. Novelty and objectives

Ideally we would like to develop methods which discriminate between MCI-afflicted patients who subsequently develop AD and patients who do not. Unfortunately, as such data are not currently available to us, this paper will instead focus on the classification of steady state EEG recordings from two classes: MCI-afflicted patients who later develop AD and age-matched controls. While this is likely to be a simpler problem it is felt that the results presented here will still be of interest to researchers seeking methods for predicting AD, though more extensive testing will be needed to verify the usefulness of these methods.

While the use of EEG to predict MCI to AD conversion is still relatively new, there are already a number of studies dealing with this problem (van der Hiele *et al* 2006). A variety of markers have been investigated, for example band power (Huang *et al* 2000, Cichocki *et al* 2005, Chapman *et al* 2007), dipolarity (Musha *et al* 2002), coherence or synchronization (Hogan *et al* 2003, Koenig *et al* 2005, Vecchio *et al* 2006) and ERP-based features (Chapman *et al* 2007).

In the case of full-blown AD (as opposed to MCI) there is a somewhat larger body of research, providing a broader range of features and techniques to choose from, though of course the validity of using these features with MCI data must be established. One approach which has been widely applied to the diagnosis of AD, but which has hitherto been largely overlooked for prediction of MCI-AD progression, is the estimation of signal complexity, as well as the related concept of entropy. Examples of studies which focus on the former include Pritchard *et al* (1994) and Jeong *et al* (2001) while Abásolo *et al* (2005, 2006) focused on measures which reflect the entropy or *irregularity* of the EEG time series; this issue is discussed in a little more detail below (in section 3.3.3) but interested readers are referred to Jeong (2004), which provides a good review. In addition, there is considerable variability

amongst the data used in different studies, for example in terms of the severity of AD, the length of the EEG recordings and even the sampling frequency.

Finally, a third aspect considered here is the selection of appropriate signal processing techniques for noise rejection. EEG signals are notoriously prone to noise contamination, while computational analysis is hampered by the high dimensionality of the data. In general, there is a need for a more systematic approach for extracting components of interest from the multi-channel data. Some studies have resorted to analyzing individual channels selected based on some measure of optimality (e.g. Abásolo *et al* (2005)). Clearly, using such an approach is time consuming and fails to exploit all available information. A more principled approach is to use blind source separation (BSS) techniques (as done in Cichocki *et al* (2005), Vialatte *et al* (2005), for example). However, this also has its problems and will be discussed in more detail in section 3.

Hence, despite the variety of solutions that have been proposed, the problem is still not solved and there is a need for alternative classification techniques, as well as the verification of existing approaches. The research presented here aims to address some of the shortcomings mentioned above. Specifically, this paper will make the following three important contributions:

(i) to propose a method for finding sub-optimal spatial filters for improved detection of AD;
(ii) to test the effect of varying window sizes when extracting AD biomarkers;
(iii) to evaluate a measure of entropy as a marker of early AD.

3. Proposed method

3.1. Data

The data used here have been analyzed in previous studies concerning early detection of AD (Musha *et al* 2002, Cichocki *et al* 2005, Vialatte *et al* 2005). They consist of steady-state EEG data recorded from 21 sites on the scalp based on the 10–20 system. The sampling frequency was 200 Hz, which allowed for signals of up to 100 Hz to be represented.

The subjects comprised two study groups. The first consisted of a group of 25 patients who had complained of memory problems. These subjects were then diagnosed as suffering from MCI and subsequently developed mild AD. The criteria for inclusion into the MCI group were a mini mental state exam (MMSE) score of $\geqslant$24, though the average score in the MCI group was 26 (SD of 1.8). The other group was a control set consisting of 56 age-matched, healthy subjects who had no memory or other cognitive impairments. The average MMSE of this control group was 28.5 (SD of 1.6). The ages of the two groups were 71.9 ± 10.2 and 71.7 ± 8.3, respectively. Finally, it should be noted that the MMSE scores of the MCI subjects studied here are quite high compared to a number of other studies. For example, in Hogan *et al* (2003) the inclusion criterion was MMSE $\geqslant$ 20, with a mean value of 23.7, while in Chapman *et al* (2007), the criterion was MMSE $\geqslant$ 22 (the mean value was not provided); thus, the disparity in cognitive ability between the MCI and control subjects was comparatively small, making the present classification task relatively difficult.

All recording sessions were conducted with the subjects in an awake but resting state with eyes closed and under vigilance control. In addition, pre-selection was conducted to ensure that the data were of a high quality, as determined by the presence of at least 20 s of artifact free data. Based on this requirement, the number of subjects in the two groups described above was further reduced to 22 and 38, respectively.

3.2. Spatial filtering for EEG classification

As EEG data are very noisy, the first problem is to find a method of separating significant components from other background mental activity and noise. To date a number of signal-processing methods have been applied to EEG, one of the most popular of which is the diverse class of algorithms known collectively as blind source separation (BSS) (see Cichocki and Amari (2003) for an in-depth review).

However, while BSS methods are effective at exploiting statistical properties inherent in the data, additional information such as class labels are not used. It is reasonable to suggest that better signal-noise separation could be achieved if this information could be incorporated in some way. An additional property that would be useful is the capability to perform dimensionality reduction, which is difficult using the popular ICA (independent component analysis) class of BSS algorithms, for example. Since EEG datasets are invariably high dimensional, the ability to rank or order the separated components and subsequently separate relevant or interesting components would be an added advantage.

Fortuitously, an algorithm that provides both of these properties is already available and is in fact widely used for the classification of EEG signals, though hitherto this has been in the context of brain–computer interfaces (BCI (Wolpaw *et al* 2002)). This technique, known as common spatial patterns (CSP), works by finding spatial filters which maximize the difference in signal power between the two classes to be discriminated (Ramoser *et al* 1998). It is interesting to note that, while most of the recent interest in CSP has been due to its excellent performance in BCI (Blanchard and Blankertz 2004, Shenoy *et al* 2006), one of the earliest uses of the method was in the study of cognitive disorders such as schizophrenia and depression (Koles *et al* 1994). This broad spectrum of applicability suggests that CSP might also be useful for performing signal separation in the AD context.

Briefly, the CSP filters are found as follows.

(i) Partition the full data matrix $\mathbf{X}$ into the two class-specific matrices $\mathbf{X}_A$ and $\mathbf{X}_B$ corresponding to the two classes to be discriminated.

(ii) Calculate the corresponding covariance matrices $\mathbf{C}_A$ and $\mathbf{C}_B$, as well the sum $\mathbf{C} = \mathbf{C}_A + \mathbf{C}_B$.

(iii) Find the whitening matrix $\mathbf{W}$ such that $\mathbf{W}^T\mathbf{C}\mathbf{W} = \mathbf{I}$, where $\mathbf{W}$ may be found via the eigenvector decomposition:

$$\mathbf{C} = \mathbf{U}^T\Sigma\mathbf{U}$$

then setting $\mathbf{W} = \mathbf{U}\Sigma^{-1/2}$. Hence,

$$\mathbf{W}^T\mathbf{C}\mathbf{W} = \mathbf{I}$$

$$\Rightarrow \mathbf{W}^T\mathbf{C}_A\mathbf{W} + \mathbf{W}^T\mathbf{C}_B\mathbf{W} = \mathbf{I}. \quad (1)$$

(iv) Apply a rotation matrix $\mathbf{Y}$ to both sides of (1)

$$\mathbf{Y}^T(\mathbf{W}^T\mathbf{C}_A\mathbf{W} + \mathbf{W}^T\mathbf{C}_B\mathbf{W})\mathbf{Y} = \mathbf{I}. \quad (2)$$

(v) Choose $\mathbf{Y}$ such that it diagonalizes $\mathbf{W}^T\mathbf{C}_A\mathbf{W}$. Hence, from (2) it follows that $\mathbf{Y}^T[\mathbf{W}^T\mathbf{C}_B\mathbf{W}]\mathbf{Y}$ will also be diagonal, and the sum of corresponding diagonal elements will be 1.

(vi) Hence, to create a spatial filter that maximizes the variance of class A trials while minimizing the variance of class B trials, set $\mathbf{Y}$ to be the eigenvectors of $\mathbf{W}^T\mathbf{C}_A\mathbf{W}$. Then, the columns of the matrix $\mathbf{WY}$ provide the CSP spatial filters and may be sorted on the basis of the eigenvalues.

3.3. Feature extraction

Three potential biomarkers of AD were considered:

(i) theta band power;
(ii) windowed signal power;
(iii) sample entropy.

These will now be briefly described and their inclusion in this study justified.

3.3.1. Signal power (PF). For steady-state EEG data, the analysis of frequency components is the most common method of signal analysis. A variety of frequency bands have been used for AD detection, and it has been observed that the theta band (approximately 4–8 Hz) provides good performance in the detection of AD (van der Hiele *et al* 2006, Vialatte *et al* 2005). To restrict the scope of the paper, we will focus exclusively on EEG activity in this band for now.

For this study, both the absolute and relative theta band power were studied. For brevity, this is henceforth referred to as the 'power feature' (PF), which is defined thus:

$$\mathrm{PF} = \sum_{i\in[4,8]} F_i, \tag{3}$$

$$\widetilde{\mathrm{PF}} = \frac{\sum_{i\in[4,8]} F_i}{\sum_{i\in[1.5,25]} F_i}, \tag{4}$$

where PF and $\widetilde{\mathrm{PF}}$ are the absolute and relative theta powers respectively, and F_i is the Fourier coefficient of the signal at frequency bin i. To evaluate this feature, the power spectrum of the relevant time series (either raw data or CSP-filtered) was calculated using the fast Fourier transform, where the ordinates of the power spectrum falling within the theta band were extracted then added up.

3.3.2. Windowed signal power (WP). As discussed in section 2, there is a lot of variability amongst the data used in AD studies, which complicates the task of comparing different methodologies. While there are many factors involved, we hope to address one aspect of this problem via the use of the windowed signal power feature, which for future convenience we will denote as WP.

The recordings used in this study are quite long ($t = 20$ s), which allows us to investigate the use of a variety of window sizes. In addition, it is known that transient features present only for short periods of time can be an extremely effective biomarker (Vialatte *et al* 2005). To address both these issues, we modify the PF feature by dividing the signal from each channel into a set of shorter segments, which are used to derive an alternative feature pair as follows:

$$\mathrm{WP}_n = \max_{i=[i,n]} \mathrm{PF}_i, \tag{5}$$

$$\widetilde{\mathrm{WP}}_n = \max_{i=[1,n]} \widetilde{\mathrm{PF}}_i, \tag{6}$$

where n is the total number of segments and PF_i and $\widetilde{\mathrm{PF}}_i$ are the PF and $\widetilde{\mathrm{PF}}$ features calculated using the ith segment, but which otherwise are evaluated in the same fashion as in subsection 3.3.1.

The choice of segment with the maximum power content is based on an earlier study (Vialatte *et al* 2005) in which a method known as 'bump modeling' is used to detect prominent

features in a time-frequency representation of the signal generated using Morlet wavelets. In order to focus the analysis on segments of the time-frequency spectrum which contain significant activity, the segments are ranked according to power before being extracted for modeling.

3.3.3. Sample entropy (SE). The use of nonlinear characteristics of the EEG time series in the detection and diagnosis of Alzheimer's has been around for some time. An early example is Pritchard *et al* (1994), where the correlation dimension (D2) of the EEG time series is used to augment the classification rate of standard power features. A more recent example further includes the use of the first Lyapunov exponent (L1) to distinguish between probable AD and control subjects (Jeong *et al* 2001). In the context of MCI data, however, the use of such measures have largely been ignored.

To help address this shortcoming, we wish to include a nonlinear measure as one of the features investigated. However, we note that despite the early enthusiasm in D2, its effectiveness depends on the assumption that the underlying system is intrinsically low-dimensional and chaotic. The validity of this line of reasoning has been questioned due to the significant complications caused by noise and nonstationarity, as well as the requirement for large amounts of data (see, e.g., Palus (1996, 1999)). Similar to D2, L1 is also based on the notion of an underlying low-dimensional chaotic system, and its use should be discouraged for similar reasons.

The measure which we have selected for use here is sample entropy (SE) (Richman and Moorman 2000). In contrast to D2, which measures the system's innate dimensionality, and L1 which measures the rate of divergence of neighboring points, SE is derived from approximate entropy (ApEn), which measures the extent to which a time series tends to repeat itself (though without any implications of periodicity). SE (and ApEn) hence serve to provide an indication of the level of *regularity* or entropy (randomness) of a time series without making any assumptions regarding the presence of chaos. ApEn has already been used for the diagnosis of AD (but not MCI-AD conversion) (Abásolo *et al* 2005) while in a more recent study, SE has also been used in the context of AD (Abásolo *et al* 2006). The results of the latter indicate that SE is a useful feature for discriminating between AD-afflicted subjects and controls and while the experiments do not deal with the case of MCI, they provide us with some degree of confidence in the usefulness of SE. A detailed discussion of SE is not possible here but the basic premise is that SE provides several improvements over the standard ApEn procedure (Richman and Moorman 2000). SE has also been used for AD detection in Vialatte and Cichocki (2006) though the approach used was very different. Hence, there is good reason to believe that SE represents a promising feature for the prediction of AD. The basic idea is to approximate the following statistic:

$$SE(m,r) = -\ln\left[\frac{\Pr(d[\mathbf{x}_{m+1}(i),\mathbf{x}_{m+1}(j)] \leqslant r)}{\Pr(d[\mathbf{x}_m(i),\mathbf{x}_m(j)] \leqslant r)}\right], \tag{7}$$

where $\mathbf{x}_m(i) \in \mathcal{R}^{1\times m}$ is the vector extracted from the time series $x(t)$ as follows:

$$\mathbf{x}_m(i) = [x(i), x(i+1), \ldots, x(i+m-1)], \tag{8}$$

and $d[\mathbf{x}_m(i), \mathbf{x}_m(j)]$ is defined as

$$d[\mathbf{x}_m(i), \mathbf{x}_m(j)] = \max_{k=[1,m]}(|x(i+k-1) - x(j+k-1|). \tag{9}$$

In a practical application, the probability values in equation (7) are approximated by averaging over the entire time series. As can be seen, SE provides a measure of how quickly the predictability of the time series drops as the length of the segments considered is increased.

For the purpose of this project, the Matlab tools for calculating sample entropy provided at http://www.physionet.org/physiotools/sampen/matlab/ were used.

3.4. Feature classification

For feature classification, we adopt a simple model in which the likelihood function of the feature is modeled using a Gaussian distribution. To train the models, the selected features were extracted from the training data and grouped according to their class labels; as only a limited amount of data is available, leave-one-out cross validation is used, where a single instance from the data set is isolated to use as the test data, while the remaining data are used as the training set. This procedure is repeated for each instance in the data set and the test errors thus obtained are averaged to produce the final results.

For each class $c \in \{1, 2\}$, we fit the features extracted using a Gaussian distribution. This is similar to the more common approach of using linear discriminant analysis (LDA), but allows a separate covariance to be used for each class.

To classify a test vector $\mathbf{f}$ as either an AD or a control subject, the maximum *a posteriori* (MAP) decision rule is used:

$$c_{\mathrm{MAP}}(\mathbf{f}) = \mathrm{argmax}_c\, P(c|\mathbf{f}).$$

$P(c|\mathbf{f})$ can be found via Bayes' theorem. Also, in this case, we have uniform prior and constant evidence terms, hence

$$\begin{aligned} P(c|\mathbf{f}) &= \frac{P(\mathbf{f}|c)\,p(c)}{P(\mathbf{f})} \\ &= k\,P(\mathbf{f}|c) \\ &\propto \exp\left[-(\mathbf{f}-\mu_c)^T\mathbf{C}_c^{-1}(\mathbf{f}-\mu_c)\right], \end{aligned}$$

where μ_c and $\mathbf{C}_c$ are the mean and covariance matrix of the Gaussian for class c.

The effectiveness of the features can now be evaluated in terms of the classification rates, which are calculated as follows:

$$\mathrm{accuracy}(\%) = 100 \times \left[\sum_{i=1}^{n} \delta(c_{\mathrm{MAP}}(\mathbf{f}_i) - c(i))\right] \Big/ n, \tag{10}$$

where i is the subject index and n is the number of subjects. $\mathbf{f}_i$ and $c(i)$ denote the feature vector and class labels for trial i and $\delta(.)$ is Dirac's delta function.

4. Procedures and results

4.1. Basic methodology

In simulations which required CSP pre-processing, a fifth-order Butterworth filter with the passband 4–8 Hz was used to extract signal components in the Theta band before these were submitted to CSP processing as described in section 3.2. To create CSPs tuned to the relative theta power, each channel was first normalized based on the overall power content in the frequency band 1.5–25 Hz (Cichocki *et al* 2005), before being bandpass filtered and analyzed using CSP. Finally, for all SE calculations, we used the parameter settings $m = 1$ and $s = 0.2$ times the SD of the data, as recommended in Abásolo *et al* (2005).

In addition, leave-one-out cross validation or jack-knifing was used to prevent problems like over-fitting and bias. Briefly, this technique works by training the models using all available data except for data from one subject; this excluded set is then used to test the performance of the models. The procedure is repeated for all subjects to obtain a more reliable

Table 1. Classification rates using the various feature extraction methods. Entries in the leftmost column are of the form *X–Y*, where *X* represents the feature extraction method used and *Y* is the channel. Entries in bold indicate the best result(s) in the respective columns.

Markers/ Channels	Misclassified (%)		Correctly classified (%)		
	MCI	Controls	MCI	Controls	All
PF-CSP	**22.7**	**15.8**	**77.3**	**84.2**	**81.7**
PF-P4	40.9	21.1	59.1	78.9	71.7
$\widetilde{\text{PF}}$-CSP	**22.7**	18.4	**77.3**	81.6	80.0
$\widetilde{\text{PF}}$-P4	31.8	28.9	68.2	71.1	70.0
WP-CSP	**22.7**	18.4	**77.3**	81.6	80.0
WP-P4	36.4	23.7	63.6	76.3	71.7
$\widetilde{\text{WP}}$-CSP	27.3	21.1	72.7	78.9	76.7
$\widetilde{\text{WP}}$-P4	31.8	26.3	68.2	73.7	71.7
SE-CSP	36.4	44.7	63.6	55.3	58.3
SE-P4	68.2	42.1	31.8	57.9	48.3

estimate of classification performance. The benefit of this method is that it allows a larger training set to be used while still calculating the test error based on unseen data. This is very beneficial in situations where there is a limited amount of data available.

4.2. Comparison of signal separation with single channel

In the first set of tests, all five potential markers (PF, $\widetilde{\text{PF}}$, WP, $\widetilde{\text{WP}}$ and SE) are extracted, first from the MCI-specific CSP channel, then the unprocessed EEG channel with the highest difference between class means. Results of these tests are given in table 1. From this table, some observations are as follows.

(i) With all five marker types, the CSP filtering produces results that are consistently superior to the data obtained using the single channel approach. We also used receiver operating characteristic (ROC) curves to study the results in more detail and the curves for PF-CSP and PF-P4, which had the best classification rates for CSP and single channel approaches respectively, are presented in figure 1.

(ii) Surprisingly, the absolute theta power based features produced better performance compared to the relative theta based features. This was true for both PF and WP. A variety of different normalizing bands were experimented to verify this result but the outcome was invariably the same. At present, it is unclear why this is the case as we would expect that relative theta should instead be the more reliable feature (as it allows for trial-to-trial variability in the power levels). One possible explanation is that some form of over-fitting has occurred but this cannot be confirmed without having access to more data. The ROC curves for PF-CSP and $\widetilde{\text{PF}}$-CSP are presented in figure 2 from where it can be seen that, while not as clear as with figure 1, the area under the PF-CSP curve (also referred to as the 'AUC') is slightly greater compared to the AUC of the $\widetilde{\text{PF}}$-CSP curve (0.826 versus 0.817). The AUC of a ROC curve can be interpreted as the probability of a particular model returning the correct classification.

(iii) Though we have not compared CSP directly to BSS yet, we note that the results for PF-CSP are better than earlier results published in Cichocki *et al* (2005) which also use frequency band power, but in combination with the AMUSE BSS technique (Cichocki and Amari 2003).

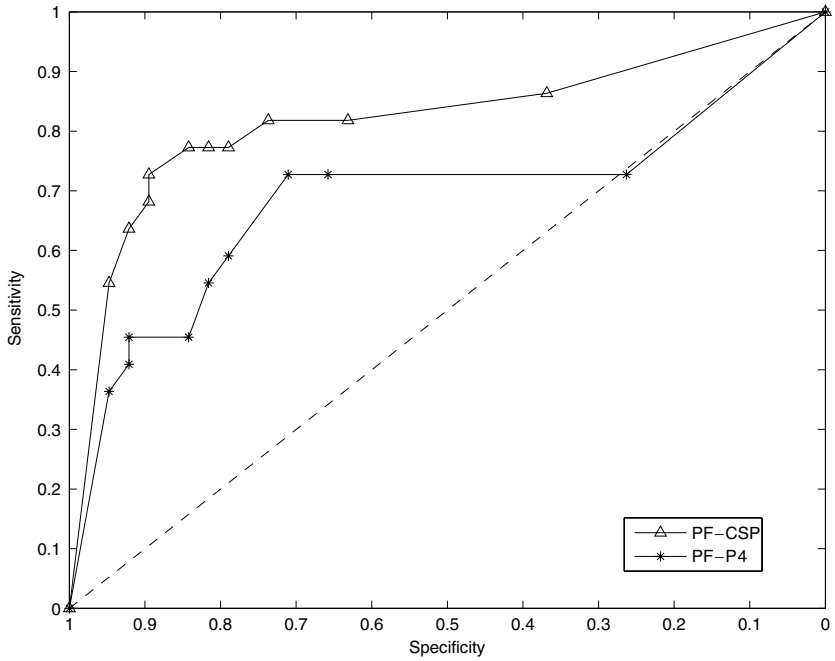

Figure 1. ROC curve comparing the classification performance of PF-CSP and PF-P4. The diagonal dotted line is known as the 'no-discrimination line' and corresponds to a classifier which returns random guesses. The AUCs for PF-CSP and PF-P4 were 0.820 and 0.6884 respectively.

(iv) The results obtained using SE are disappointing but it might be too early to dismiss the technique entirely. In the case of SE-P4, the performance is really poor (approximately random. The misclassification rate for the MCI case is curiously rather high but we discount this as a coincidence), but some discriminatory ability was present when applied to the CSP channel. This is encouraging considering that the criteria for CSP is the maximization of *variance*.

(v) The classification rates using WP are very similar to PF. Though this seems to imply that there is no further information to be gained using the WP features, we feel that, similar to the previous item, this result is still encouraging given that the CSP filters are optimized with respect to total variance. This issue is pursued further when we study the classification performance using multiple CSP channels later in section 4.4.

4.3. Evaluation of window length

We now focus on the WP-CSP feature and conduct a review across a range of window lengths. Consecutive segments extracted from the time series were positioned to overlap by 50% so as not to miss any interesting features. The results of this analysis are shown in table 2. From the table it can be seen that the results do not vary a lot over the entire range of window lengths. At the lower end (0.5 s) the classification rate dropped off but at larger window sizes it appears to have stabilized with only minor fluctuations.

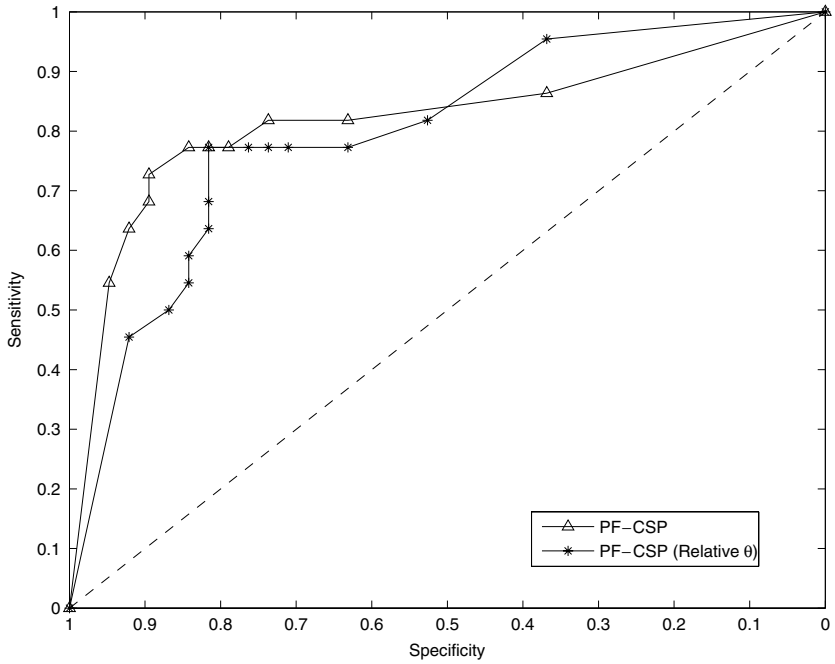

Figure 2. ROC curve comparing the classification performance of PF-CSP and $\widetilde{\text{PF}}$-CSP. The AUCs for PF-CSP and $\widetilde{\text{PF}}$-CSP were 0.820 and 0.799 respectively.

Table 2. Classification rates using different window lengths.

Window lengths	Misclassified (%)		Correctly classified (%)		
	MCI	Controls	MCI	Controls	All
0.5 s	27.3	21.1	72.7	78.9	76.7
1.5 s	22.7	13.2	77.3	86.8	83.3
2.5 s	22.7	13.2	77.3	86.8	83.3
5.0 s	22.7	15.8	77.3	84.2	81.7
10.0 s	22.7	15.8	77.3	84.2	81.7
15.0 s	22.7	13.2	77.3	86.8	83.3
20.0 s	22.7	16.2	77.3	83.8	81.4

4.4. Classification using multiple CSP channels

Finally, we would like to test the performance of the method when applied to two or more CSP channels. For the following experiments, we used leave-one-out cross-validation when fitting the classification models but due to computational constraints used a common CSP filter derived from the entire data set. To confirm that this was a reasonable simplification, CSP vectors were calculated using the entire data set, then compared to a variety of CSP vectors calculated after individual instances had been removed. Our observation was that the vectors were practically identical in all these cases.

Table 3. Classification rates over a variety of feature-channel combinations. The feature class is represented by an *X*–*Y* combination, where *X* is the number of CSP channels used, and *Y* is the feature type. The row in bold indicates best overall performance. All results were obtained using leave-one-out cross validation over the classification models.

Feature class	Misclassified (%)		Correctly classified (%)		
	MCI	Controls	MCI	Controls	All
2-PF	31.8	7.9	68.2	92.1	83.3
2-WP	45.5	5.3	54.5	94.7	80.0
2-SE	63.6	18.4	36.4	81.6	65.0
4-PF	31.8	7.9	68.2	92.1	83.3
4-WP	**27.3**	**2.6**	**72.7**	**97.4**	**88.3**
4-SE	45.5	15.8	54.5	84.2	73.3

We present here a selection of the various parameter combinations possible. For features, we will review the PF, WP and SE methods, while each method will be applied to two different sets of time series:

- two MCI-specific CSP channels;
- two MCI-specific and two control-specific CSP channels.

Also, for the WP experiments a window size of around 7 s was used. The experiments were run and the results are presented in table 3. Based on this table, we made the following observations.

(i) The highest score was obtained using **4-WP**, i.e. the windowed theta power classifier applied to four CSP channels. To analyze the results of this classifier in more detail, the ROC curve was plotted and compared to the ROC curves of 4-PF and 2-PF, which had the next best classification rates. These are presented in figure 3. The AUCs for 4-WP, 4-PF and 2-PF are 0.9025, 0.8732 and 0.8200 respectively, indicating that the 4-WP classifier gives the best all-round performance based on the estimated AUC. However, the differences between the curves are not that large and this is validated experimentally where there does not appear to be any clear trend or relationship between window size and classification rate.

(ii) The performance of SE showed great improvement when used with the set of four CSP channels. This indicates that information relevant to the detection of early AD is present in the entropy of the time series. However, it is possible that CSP is not a suitable pre-processing method when dealing with SE.

(iii) Overall, the sensitivity of all the classifiers was quite poor when compared to the specificity. It is not immediately clear why this was the case but this same effect was observed in other studies which used this data set (Cichocki *et al* 2005, Vialatte *et al* 2005), indicating that it was probably a property of the sample groups.

5. Discussions

We have presented the results of a set of experiments intended to extend and improve existing methods for the early detection of AD. While the results described here are still preliminary and will need to be verified on larger data sets, they are encouraging and suggest that the described methods could be useful in the prediction of MCI-AD progression. We concede, however,

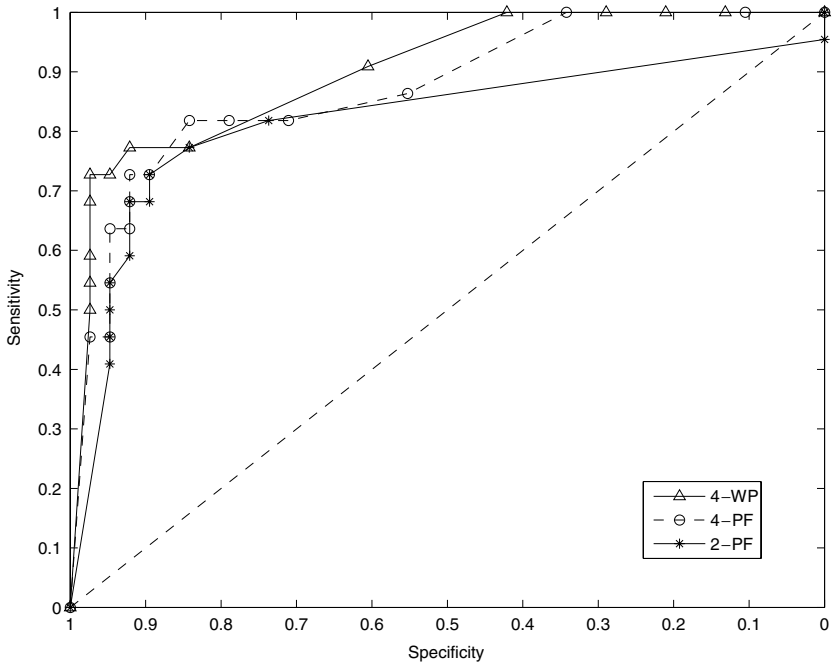

Figure 3. ROC curve comparing the classification performances of 4-WP, 4-PF and 2-PF. The AUCs are 0.9025, 0.8732 and 0.8200 for 4-WP, 4-PF and 2-PF, respectively.

that the present set of results only establish that the proposed methods can distinguish between MCI-afflicted patients who develop AD, and normal controls who suffer from no cognitive defects. As such, it is possible that the features tested are markers of MCI and not of the future AD. To exclude this possibility, more extensive data would have to be collected which include recordings from MCI-afflicted patients who *do not* subsequently develop AD.

Overall, the results obtained using CSP were very promising and represent a clear improvement over the results published in Cichocki *et al* (2005), which used BSS in combination with a similar band power feature. However, we also note that the results obtained in Vialatte *et al* (2005) are better, where the best classification rate obtained was 93.3%. However, the difference is reasonable (accounting for 3 misclassified subjects out of a total of 60). More importantly, the results in Vialatte *et al* (2005) were obtained using 'bump-modeling', a totally different feature extraction method. It would be interesting to see if results could be improved still further using a combination of CSP and bump modeling.

Hence, many problems remain which need to be addressed, though these should be viewed as exciting avenues for future research rather than hindrances. In particular, in its current form CSP may only be used to find spatial filters which are optimal with respect to the variance (signal power) in specified bands. This has probably biased the results in favor of the power-based features. As can be seen in table 3, however, sample entropy is also likely to be useful for the prediction of AD, provided appropriate filtering mechanisms could be devised. Further work in the above mentioned directions is now being planned and we hope to report the findings in future publications.

References

Abásolo D *et al* 2005 Analysis of regularity in the EEG background activity of Alzheimer's disease patients with approximate entropy *Clin. Neurophys.* **116** 1826–34

Abásolo D *et al* 2006 Entropy analysis of the EEG background activity in Alzheimer's disease patients *Physiol. Meas.* **27** 241–53

Blanchard G and Blankertz B 2004 BCI Competition 2003: Data set: IIa. Spatial patterns of self-controlled brain rhythm modulations *IEEE Trans. Biomed. Eng.* **51** 1062–6

Blennow K *et al* 2006 Alzheimer's disease *The Lancet* **368** 387–403

Chapman R *et al* 2007 Brain event-related potentials: diagnosing early-stage Alzheimer's disease *Neurobiol. Aging* **28** 194–201

Cichocki A and Amari S 2003 *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications* (New York: Wiley)

Cichocki A *et al* 2005 EEG filtering based on blind source spearation (BSS) for early detection fo Alzheimer's disease *Clin. Neurophys.* **116** 729–37

Hogan M *et al* 2003 Memory-related EEG power and coherence reductions in mild Alzheimer's disease *Int. J. Psychophysiol.* **49**

Huang C *et al* 2000 Discrimination of Alzheimer's disease and mild cognitive impairment by equivalent EEG sources: a cross-sectional and longitudinal study *Clin. Neurophys.* **111** 1961–7

Jeong J 2004 EEG dynamics in patients with Alzheimer's disease *Clin. Neurophys.* **115** 1490–505

Jeong J *et al* 2001 Nonlinear dynamic analysis of the EEG in patients with Alzheimer's disease and vascular dementia *J. Clin. Neurophysiol.* **18** 58–67

Koenig T *et al* 2005 Decreased EEG synchronization in Alzheimer's disease and mild cognitive impairment *Neurobiol. Aging* **26** 165–71

Koles Z *et al* 1994 Spatial patterns in the background EEG underlying mental disease in man *Electroenceph. Clin. Neurophys.* **91** 319–28

Musha T *et al* 2002 A new EEG method for estimating cortical neuronal impairment that is sensitive to early stage Alzheimer's disease *Clin. Neurophys.* **113** 1052–8

Palus M 1996 Nonlinearity in normal human EEG: cycles, temporal asymmetry, nonstationarity and randomness, not chaos *Biol. Cybern.* **75** 389–96

Palus M 1999 Is nonlinearity relevant for detecting changes in EEG? *Theory Biosci.* **118** 179–88

Pritchard W *et al* 1994 EEG-based, neural-net predictive classification of Alzheimer's disease versus control subjects is augmented by non-linear EEG measures *Electroenceph. Clin. Neurophysiol.* **92** 118–30

Ramoser H *et al* 1998 Optimal spatial filtering of single trial EEG during imagined hand movement *IEEE Trans. Rehab. Eng.* **8** 441–6

Richman J and Moorman J 2000 Physiological time-series aalysis using approximate entropy and sample entropy *Am. J. Physiol. Heart Circ. Physiol.* **278** H2039–49

Shenoy P *et al* 2006 Towards adaptive classification for BCI *J. Neural Eng.* **3** R13–23

Turner R 2003 Biomarkers of Alzheimer's disease and mild cognitive impairment: are we there yet? *Exp. Neurol.* **183** 7–10

van der Hiele K *et al* 2006 EEG and MRI correlates of mild cognitive impairment and Alzheimer's disease *Neurobiol. Aging* available online DOI:10.1016/j.neurobiolaging.2006.06.006

Vecchio F *et al* 2006 Sources and coherence of cortical EEG rhythms could predict progression from mild cognitive impairment to Alzheimer disease *Clin. Neurophysiol.* **117** (Suppl. 1) 1–2

Vialatte F and Cichocki A 2006 Sparse bump sonification: a new tool for multichannel EEG diagnosis of mental disorders; application to the detection of the early stage of Alzheimer's disease *ICONIP 2006* pp 92–101

Vialatte F *et al* 2005 Blind source separation and sparse bump modelling of time frequency representation of EEG signals: New tools for early detection of Alzheimer's disease *IEEE Workshop on Machine Learning for Signal Processing* pp 27–32

Wolpaw J *et al* 2002 Brain computer interfaces for communication and control *Clin. Neurophysiol.* **113** 767–91