Real-time Brain-Computer Interfacing: a preliminary study using Bayesian learning

Stephen J. Roberts & William D. Penny

Robotics Research Group
Department of Engineering Science
University of Oxford, UK
sjrob@robots.ox.ac.uk

Abstract

We present preliminary results from real-time 'brain-computer interface' experiments. Our analysis is based on autoregressive modelling of a single EEG channel coupled with classification and temporal smoothing under a Bayesian paradigm. We show that uncertainty in decisions is taken into account under such a formalism and that this may be used to reject uncertain samples thus dramatically improving system performance. Using the strictest rejection method a classification performance of $86.5 \pm 7\%$ is achieved over a set of seven subjects in two-way cursor movement experiments.

Keywords: Brain-computer interfacing, real-time EEG analysis, biosignal analysis, Bayesian learning.

1 Introduction

Several previous publications have detailed the use of measures extracted from the human electroencephalogram (EEG) to form rudimentary computer interfaces (Pfurtscheller *et al.*, 1993; Wolpaw & McFarland. 1994; Roberts *et al.*, 1998). These are generally referred to as brain-computer interfaces (BCIs). There are two main paradigms under which BCIs have been researched; off-line and on-line. The former has the typical protocol of an externally-cued prompt which requires the user to make real or imagined movements (of the fingers typically) at intervals of order 10-15 seconds. Event-related

EEG changes, in particular the 10-12Hz rhythm over the primary motor cortex (the mu rhythm), are then analysed off-line from the EEG recordings (Peltoranta *et al.*, 1994). Imagined finger movements can be distinguished using such a protocol with an accuracy of around 70% (Penny & Roberts, 1998; Roberts *et al.*, 1998) with distinction between left and right-hand imagined movements at a similar level (Pfurtscheller *et al.*, 1994). The second paradigm, on-line or real-time, is probably the more exciting of the two. In Pfurtscheller *et al.*(1993) the development of an analysis system is reported in which real finger movements are used for training and subsequent imagined movements are assessed in real-time in the presence of bio-feedback (i.e. the subject is given an indication of the analysis state at each time-frame). Whilst initial results were similar in accuracy to that of off-line experiments this rose in some subjects due to the presence of bio-feedback. Wolpaw *et al.* (1994) also comment on a similar system which has the advantage of being self-paced (i.e. no external cues are given to the subject during the experimental block). Such systems, however, are reported as having the drawback of very long (several weeks) training times during which some subjects became acclimatised to the system (MacFarland *et al.*, 1993).

If such BCIs are to be used in the evaluation and re-habilitation of, for example, severely disabled subjects, then accuracies of order 70%, the requirement for real movements to be performed for training the systems and long subject acclimatisation times must all be improved upon. In this paper we present preliminary results which appear to indicate that a rapid and accurate real-time BCI system is achievable with little or no subject acclimatisation. The system has the added advantages of simplicity and is based on a single, differential EEG signal only.

2 Experimental Protocol

A single differential channel of EEG is recorded using silver-silver chloride electrodes between positions C3’ and C4’ (3cm posterior to the standard 10-20 positions C3 and C4) which lie over the primary motor cortex. The ground was placed on the head just behind the right ear (the right mastoid). The signal was amplified using an ‘ISO-DAM’ (World Precision Instruments Inc. n.d.) isolation amplifier with gain 10^4 and analogue bandpass filtered with 3dB points set at 0.1Hz and 100Hz. The signal was subsequently digitised at 384Hz. All data were both stored and processed in real-time using a Pentium 266MHz PC.

The subject is seated before the computer screen on which is a task window, in which are two icons, a ‘goal’ icon which is fixed and lies at top or bottom of the screen and an icon which indicates the current position of the BCI cursor. The latter is confined to move in the vertical direction. The experiments were performed with the subjects sitting in front of the computer screen in daylight. The room contained only the experimenter and the subject. The experimental protocol was completely automated via custom-written software.

The subject was asked to attempt a series of cursor-movement tasks which alternate between the goal at the top and bottom of the screen. Each up or down movement task lasted for some 10-15 seconds. In all the results presented here upwards movement of the cursor was associated with simple mental arithmetic by the subject and downwards motion by imagined movement of the dominant hand. The choice of mental artihmetic and imagined movements as the two cognitive tasks was based upon the results of a separate study (Penny & Roberts,1999a,b) which indicated that the combination of these tasks was significantly better than using either task in association with a baseline period, as reported in Stam *et al* (1996), Fernandez *et al.* (1995), Pfurtscheller *et al.* (1998) and McFarland *et al.* (1997). The significance was at $p < 0.003$ based on two-way ANOVA analysis.

In these experiments re-training of the classification system was performed after each experiment using, as a training data set, the previous experimental data. Training of the analyser for classification during the first experimental block was achieved via an initial recording of 10 seconds each of ‘up’ & ‘down’ tasks performed without biofeedback.

3 Data Analysis

3.1 Autoregressive modelling

We adopt a standard parameterisation of the EEG signal based on parametric modelling methods. An 8th-order autoregressive (AR) model is fitted to 1/3 second blocks of data (128 samples) which slide 32 samples (1/12 second) from one processing time step to the next. The choice of an 8th-order model was based on model-order estimation from a section (some 20 seconds) of EEG data. Residual errors for models of order one to twenty were calculated over blocks of 128 samples as detailed above. Also estimated were the errors due to parameter uncertainty (for a fixed size block the number of samples available, on average, to fit each parameter decreases with the number of parameters) estimated using a Bayesian scheme (see O' Ruanaidth and Fitzgerald, 1996, for example). This may be used to penalise the residual error, which is monotonically decreasing with model order. Figure 1 shows the resultant penalised error function. The solid curve shows the error functional with model order averaged over all the data and the dashed lines are at ± 1 S.D. We note that this curve is relatively flat from model orders of 3 to 10 and our choice of 8 is for computational convenience (8 and 128, our block size in samples, being powers of 2).

The coefficients, $\{\theta_1, ..., \theta_8\}$, from the AR model are rapidly estimated using a lattice-filter approach which adapts the coefficients so as to minimise the square error between the set of actual samples, $\{s[n]\}$ and the set of samples, $\{\hat{s}[n]\}$ predicted via linear combination of the 8 (the model order) past samples, i.e.

$$\hat{s}[n] = -\sum_{i=1}^{8} \theta_i s[n-i] \tag{1}$$

Our choice of a parametric model for EEG is based upon prior publication in this area (Pardey *et al.*, 1996, for example) and also the fact that it is computationally efficient which is important in a real-time application. The AR coefficients code information regarding resonances in the signal and we utilise the coefficient sets produced over 1/3 second data blocks as input to the classification stage detailed next.

3.2 Bayes logistic classifiers

The logistic classifier is a well known methodology for data classification. In previous analysis of off-line BCI data it was found that simple linear classification was only marginally worse than considerably more computationally intensive methods, such as flexible models (i.e. 'neural' networks; see Penny & Roberts, 1998). The choice of a linear model in the present context is driven in part by its speed of computation and minimal loss in performance. The general logistic classifier has the following form for a two class (C_1, C_2) problem,

$$y(\mathbf{x}; \mathbf{w}, D) = P(C_1|\mathbf{x}) = g\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}\right) \tag{2}$$

in which $\mathbf{w}$ is a vector of the classifier parameters or weights, $\mathbf{x}$ is the corresponding input to the classifier and D is the training data set. Note that in this formulation we regard the *bias weight* of the linear classifier as an element of the vector $\mathbf{w}$ hence $\mathbf{x}$ represents the vector of input variables to the classifier augmented by a 'one'. The function $g(\cdot)$ is the logistic function given as

$$g\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}\right) = g(a) = \frac{1}{1 + \exp(-a)} \tag{3}$$

where a is referred to as the 'latent variable'. We approach the issue of weight estimation under a Bayesian paradigm in which the explicit dependence of Equation 2 on the parameters $\mathbf{w}$ is removed by integration,

$$y(\mathbf{x}; D) = \iint g(a)P(a|\mathbf{x}; \mathbf{w}, D)P(\mathbf{w}|D)d\mathbf{w}da \tag{4}$$

By Bayes theorem we may rewrite the above (ignoring $P(D)$ as it is constant given the training data)

$$y(\mathbf{x}; D) \propto \iint g(a)P(a|\mathbf{x}; \mathbf{w}, D)P(D|\mathbf{w})P(\mathbf{w})d\mathbf{w}da \tag{5}$$

in which $P(\mathbf{w})$ is the prior parameter distribution (more on this later). We take the popular quadratic-approximation approach to solving the integral over $\mathbf{w}$ in which the density over the variable a (the so-called latent variable in the logistic classifier) is approximated by a Gaussian located at the value of a specified by the maximum-likelihood value of the weights, $a^*(\mathbf{x}) = a(\mathbf{x}; \mathbf{w}^*)$. This approach, popularised in the neural network literature by MacKay (1992) is often referred to as the Laplace approximation or, more recently, as the 'evidence scheme'. Full (and readable) details of the methodology may be also found in Bishop's textbook (1995). Using this scheme we obtain,

$$P(a|\mathbf{x}, D) \approx N[a^*(\mathbf{x}), \sigma^2(\mathbf{x})] \tag{6}$$

where $N[m, v]$ denotes a Gaussian (normal) distribution with mean m and variance v. The variance, $\sigma^2(\mathbf{x})$, of Equation 6 is given as (Bishop, 1995):

$$\sigma^2(\mathbf{x}) = \mathbf{g}^{\mathsf{T}}(\mathbf{x})\mathbf{H}^{-1}\mathbf{g}(\mathbf{x}) \tag{7}$$

where $\mathbf{g} = \partial a/\partial \mathbf{w}$ evaluated at $\mathbf{w} = \mathbf{w}^*$ and $\mathbf{H}$ is the *Hessian* matrix, defined as

$$H_{i,j} = \frac{\partial^2 E}{\partial w_i \partial w_j} \tag{8}$$

in which E is the error function of the classifier, which may be given as:

$$E = -\ln P(\mathbf{w}|D) \propto -\ln P(D|\mathbf{w}) - \ln P(\mathbf{w}) \tag{9}$$

By taking a zero-mean Gaussian prior with precision[1] α (the appropriate reference prior in this case, as detailed in Lee (1994), for example) over the weights,

$$P(\mathbf{w}) = N[\mathbf{0}, \alpha^{-1}\mathbf{I}] \tag{10}$$

so the total error is a combination of the prior over weights and the *cross-entropy* error function (being the negative logarithm of the target distribution which is Bernoulli distributed). Denoting the target (desired) outputs t_n associated with each input $\mathbf{x}_n$ (so $D = \{\mathbf{x}_n, t_n\}$), this error function may be written as (Bishop,1995):

$$E = -\sum_{n=1}^{N} \left(t_n \ln y_n + (1 - t_n)\ln(1 - y_n)\right) + \frac{1}{2}\alpha\mathbf{w}^{\mathsf{T}}\mathbf{w} \tag{11}$$

which is simply a *regularised* error functional[2]. The *hyper-parameter*, or regularisation constant, α, is automatically evaluated as part of the Bayesian-learning paradigm using a second level of Bayesian inference (MacKay, 1992; Bishop, 1995) in which the likelihood of α is estimated as

$$P(\alpha|D) = \int P(\alpha|\mathbf{w}, D)P(\mathbf{w}|D)d\mathbf{w} \approx P(\alpha|\mathbf{w}^*, D) \tag{12}$$

where $\mathbf{w}^*$ is the most-probable parameter set at any instant in the learning process. As $\mathbf{w}^*$ changes during learning, so α is periodically re-estimated during the learning process by choosing the value at the mode of the above distribution. This is known as the maximum-likelihood II formalism.

The Hessian, Equation 8, is approximated using the outer product method (Bishop, 1995):

$$\mathbf{H} = \sum_{n=1}^{N} \frac{\partial y(\mathbf{x}_n)}{\partial a(\mathbf{x}_n)}\mathbf{g}(\mathbf{x}_n)\mathbf{g}^{\mathsf{T}}(\mathbf{x}_n) + \alpha\mathbf{I} \tag{13}$$

The gradient term, for a linear classifier, is simply $\mathbf{g}(\mathbf{x}) = \mathbf{x}$ hence

$$\mathbf{H} = \sum_{n=1}^{N} y(\mathbf{x}_n)[1 - y(\mathbf{x}_n)]\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}} + \alpha\mathbf{I} \tag{14}$$

[1] Inverse variance

[2] From a non-Bayesian viewpoint this represents a *weight-decay* regulariser governed by the factor α

As the integral of Equation 4 is non-analytic we make use of the approximation popularised in the neural network literature by MacKay (1992) and detailed by Spiegelhalter and Lauritzen (1990).

$$P(C_1 \mid \mathbf{x}) \approx g\left(\kappa\left(\sigma^2(\mathbf{x})\right) a^*(\mathbf{x})\right) \tag{15}$$

where

$$\kappa(\sigma^2) = \left(1 + \frac{\pi\sigma^2}{8}\right)^{-1/2} \tag{16}$$

This process of *moderation*, whereby uncertainty in the latent variable (a) propagates to adjust the posterior estimates, has the elegant effect of moving the posterior probability estimates closer to the class priors in the presence of increasing σ^2, thus indicating a reduced certainty in the resultant decision (we re-consider this again in section 3.4).

3.3 Latent-space smoothing

In a real-time BCI context we wish to perform a smoothing such that low certainty decisions may be rejected or 'over-ruled' by higher certainty decisions from the recent past (in an off-line situation, decision smoothing may also take place using future information and such an analysis lends itself to the use of hidden Markov models, for example). Such a paradigm is also suitable for adaptive-learning approaches, based for example on the Kalman filter algorithm (Puskorius & Feldkamp, 1994; Penny & Roberts, 1999). We do not take such an approach in this paper but use a computationally simpler methodology whereby temporal smoothing over the latent space variable is performed using a two-second window with $p = 24$ taps (remember that our features are obtained with an effective resolution of 32[samples]/384[Hz] = 1/12 second). The choice of a two-second window may be argued to be pragmatic but represents our prior belief that BCI-related EEG changes do not occur over a significantly smaller timescale than this. We perform a fusion of information from each two-second window by regarding the set of p latent variable distributions as forming a Gaussian mixture density. It is noted that this is the same approach as is taken in Bayesian interpretation of a *committee* of classifiers (Roberts & Penny, 1997). The resultant density is hence,

$$P(\tilde{a}_n|\{\mathbf{x}_n, \mathbf{x}_{n-1}, ..., \mathbf{x}_{n-p+1}\}) = \sum_{i=0}^{p-1} \gamma_i P(a_{n-i}|\mathbf{x}_{n-i}) \tag{17}$$

where γ_i are the set of mixture coefficients. This set can, in principle, be re-estimated from samples of the data. This requires, however, extensive calculation of cross-error correlation matrices which in turn require large data sets for accurate estimation. We choose the more pragmatic approach, given the relative sparsity of data in each time window ($p = 24$ exemplars), of simply setting $\gamma_i = 1/p$.

The mixture density of Equation 17 may be represented as a single density with mean

$$\tilde{a}_n^* = \sum_{i=0}^{p-1} \gamma_i a_{n-i}^* \tag{18}$$

where a^* is at the mode (and mean for Gaussians) of the relevant latent distribution. The variance of the density over $\tilde{a}$ is given by

$$\tilde{\sigma}_n^2 = \text{Var}\left[\tilde{a}_n\right] = \sum_{i=0}^{p-1} \gamma_i \sigma^2(\mathbf{x}_{n-i}) + \sum_{i=0}^{p-1} \gamma_i (\tilde{a}_n^* - a_{n-i}^*)^2 \tag{19}$$

The above mean and variance estimates are thence used with the approximation of Equation 15 to obtain the moderated posterior estimate over the p-tap window,

$$\tilde{P}(C_1 \mid \{\mathbf{x}_n, ..., \mathbf{x}_{n-p+1}\}) \approx g\left(\kappa\left(\tilde{\sigma}_n^2\right)\tilde{a}_n^*\right) \tag{20}$$

This moderatated posterior probability takes into account uncertainty due to imprecision in the parameters of each constituent member of the committee (the first term in Equation 19) and also uncertainty due to disagreement between committee members (the second term in Equation 19). It is noted that committees are provably better in performance than the average performance of their members (see Bishop, 1995, for example). In this context we guarantee by using a set of results from a temporal window, therefore, to outperform the mean performance of partial decisions within the window.

3.4 Bayes decision theory for rejection

From a Bayesian decision-theoretic viewpoint decision uncertainty (or confidence in a decision) is uniquely taken into account via the moderated posterior probabilities. We consider an optimal classifier, which provably operates by assigning an unknown datum $\mathbf{x}$ to class C_1 if and only if

$$\tilde{P}(C_1|\mathbf{x}) = \max_k\{\tilde{P}(C_k|\mathbf{x})\} \tag{21}$$

If a decision is made, on this basis, to classify $\mathbf{x}$ to C_1 then a strict measure of the loss or uncertainty associated with the decision to $1 - \tilde{P}(C_1|\mathbf{x})$. Our inherent confidence in a decision is given by this quantity. Note that, if equal penalties are accrued for misclassification from all classes (the so-called *loss matrix* is isotropic) the same *decision* will be made for $\tilde{P}(C_1|\mathbf{x}) = 0.51$ or 0.99 but our confidence in the decision is dramatically different. Indeed, it is common practice to include a 'reject' class such that $\mathbf{x}$ is rejected if $\max_k\{\tilde{P}(C_k|x)\} < 1 - d$ where $d \in [0, 1/2]$ is a measure of the cost associated with falsely rejecting the sample $\mathbf{x}$.

In all cases reported in this paper, we use the above Bayesian formalism to estimate moderated posterior probabilities which are thence classified to one of three classes; cursor movement up, cursor movement down and reject (if $\max\{\tilde{P}(C_{up}|\mathbf{x}), \tilde{P}(C_{down}|\mathbf{x})\} < 1 - d$). For the results presented in this paper we reject using a threshold of $1 - d = 0.6$ on the maximum posterior. This setting, equivalent to $d = 0.4$, means that we are assuming (arguably arbitrarily, however) that the relative cost of misclassification is 1/0.4 = 2.5 times the cost of falsely rejecting the sample. These threholds can, of course, be adjusted. By decreasing the cost d, so the rejection threshold, $1 - d$, rises and more samples are rejected. We require a tradeoff between reducing the 'cost' of decision making and maximising the fraction of data on which we make a decision (cursor up or down). From an initial study on a separate data set (not included in the results of this paper) we expected a value of $d = 0.4$ to result in a rejection of some 30-40% of the data. The lower bound for reject-sample 'gaps' in the decision making process (assuming the rejections are temporally uncorrelated) is only about 1/7 second. The upper bound (assuming the rejections all occur in a highly correlated block) is around 8 seconds (being 40% of the experimental block length of approximately 20s). This issue is further discussed in the results section later in the paper.

3.4.1 Information gain

If input $\mathbf{x}$ is classified to class C^* then we may also measure the certainty of the decision in terms of the number of bits of information gain by observing $\mathbf{x}$. For the two-class problem we consider here, the information gain, $\delta I(\mathbf{x})$ say, is bounded above by 1 bit and below by 0 bits (being the information entropy of priors set to 1/2). Letting, hence, $\pi(C^*) = 1/2$ be the prior we may write,

$$\delta I(\mathbf{x}) = \log_2\left(\frac{\tilde{P}(C^*|\mathbf{x})}{\pi(C^*)}\right) \qquad (22)$$

4 Results

We present results from a a set of seven volunteer subjects (not including the authors) each of whom performed a set of 12 experimental blocks. Each experimental block consisted of a single attempt to move the cursor upwards followed by an attempt to move it downwards by a similar amount thus giving a set of six ‘up/down’ runs. This protocol was itself repeated four times per subject. The cursor movement was proportional to the dominant posterior probability and in the direction of that class (up or down). If $D(t)$ represents the cursor displacement with $D(0) = 0$ then, save for an arbitrary constant,

$$D(t) = \sum_{t'=0}^{t} \left(\tilde{P}(C_{up}|\mathbf{x}_{t'}) - \frac{1}{2} \right) \tag{23}$$

For each experimental block, as we know (from the location of the goal icon) the desired reponse, we may calculate an accuracy measure on a sample-by-sample basis.

From the set of six ‘up/down’ runs our training and testing process (as discussed in section 2) gives rise to a set of five performance measures per subject per block and thus a total of $5 \times 7 \times 4 = 140$ results.

Figure 2 shows this cursor-movement trace for one cursor-up/down experimental run. The overall performance (‘correct’ classification into up or down classes) was 82%. This value is calculated over the 84% of the data which was not rejected. Note that the rejected samples (shown in the graph as smaller markers) are not randomly scattered but consist of several blocks of duration order one second.

Sample rejection occurs in regions where discriminatory information is low. As detailed in the previous section, we may quantify the information content (in bits) over each section of data using the mutual information measure between the class posteriors and priors (Equation 22). Figure 3 shows the information gain for the first two subjects (which are typical) averaged over all experimental blocks these subjects performed (the solid line). The dashed lines are at ± 1 S.D. from this average. The first two seconds are not plotted to ensure that no initialisation effects due to temporal smoothing are present in the plots. We note that there is a statistically significant decrease in the information measure over the first few seconds. This significance is at the $p < 0.05$ level as evaluated using rank statistics.

4.1 Classification performance

As the cursor movement is continuous rather than discrete, i.e. cursor movements are proportional to the estimated posteriors rather than moving a fixed distance in some direction, some continuous error measure (such as the cross-entropy error itself) is perhaps more appropriate than a simple count of correct classifications. The latter measure is, however, a stricter one, penalising poor classifiers more heavily, and also has the benefit of easy interpretation. We present such classification results for three scenarios:

Hard rejection: Temporal smoothing and moderated posteriors are used along with a reject option. If, however, more than 50% of an experimental block is rejected then the entire block is removed from the data set as a 'corrupted' data epoch.

Soft rejection: Smoothing and moderation is once more applied but no removal of experimental blocks is performed.

Baseline: No smoothing, moderation or rejection is performed and classification is made on a sample-by-sample (each 1/12 second) basis.

Figure 4 shows the mean results on a subject-by-subject basis for these three protocols. The error bars are at ± 1 S.D. Using the hard rejection scheme (circles) we see that mean performance, for all subjects, is high (an average of 0.8648 ± 0.0694 over subjects). This high performance is offset by the fact that 21% of the data blocks were entirely rejected and of the remaining data an average of 28% sample rejection still took place. When no blocks are excluded from the data set (soft rejection) the performance reduces (crosses) but is still at an average of 0.7595 ± 0.0667 correct classification. Table 6 shows the subject-by-subject fraction of data classified (i.e. not rejected). On average 66% of the data samples are classified. The final results, the baseline measures, show that with no temporal smoothing or rejection the sample-by-sample classification performance is poor, with an average success rate of only 0.5318 ± 0.1153 (stars).

4.2 Significance on block basis

We may also consider the overall significance of each up/down experimental block. This may be quantified by considering a χ^2 analysis of the 2×2 confusion matrix (whose elements are the numbers of true positive and false positive decisions for each of the two classes) (Press *et al.*, 1991). The null-hypothesis in this case is that of a stochastic (random) classifier. For the soft rejection scheme the average $\log_{10}(p)$ value from this statistic is -6.29 ± 4.45 thus indicating highly significant results. All subjects showed an average $p < 10^{-3}$.

4.3 Control blocks

We also recorded, for each subject, a set of control blocks in which the subject was asked to relax and not perform either of the two mental tasks. In 86% of these runs performance was indistinguishable from that of a random classifier to the $p < 0.05$ level (using a χ^2 test on the confusion matrix). Figure 5 shows the cursor-up probability from one such run. The up/down decision boundary is marked at 1/2 (dashed line) as are the sample rejection boundaries (at 0.6 and 0.4, dotted lines).

5 Conclusions & Future Work

In this paper we have considered the performance of a real-time ‘brain-computer interface’. Surprisingly good results are obtained using only a single EEG channel and no subject training. The analysis is based on an AR model of the signal coupled with a logistic classification scheme. Both these approaches are rapid and well-suited to real-time application. The classifier is trained under a Bayesian paradigm which enables estimates of the latent-variable density to be made as well as automatically regularising the system. A Bayesian committee decision over two-second windows is thence used to perform temporal smoothing of the latent variable. This enables uncertainty due to temporal variance to be taken into account as well as uncertainty due to parameter imprecision. Dramatic improvements in results are achieved by thence using a reject option on the ‘moderated’ posterior probabilities estimated. We would argue that the reject option offers a principled methodology for data exclusion which avoids attempted classification in areas of poor data reliability. Our results are shown to be highly significant against the null-hypothesis of a random classifier.

Future work must extend the database on volunteer subjects and also evaluate the performance of the system on patients with neuro-disability. We intend, furthermore, to extend this analysis to two-dimensional (up/down & left/right) cursor movement.

6 Acknowledgements

Thanks are first due to Prof. Maria Stokes, whose support and advice are much appreciated. Will Penny is funded via the UK EPSRC (grants GR/K79062 and GR/M41339) whose support we gratefully acknowledge. We would also like to acknowledge the equipment support of the Living Again Trust.

References

Bishop, C. (1995). *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford.

Fernandez, T. (1995). EEG activitation patterns during the performance of tasks involving different components of mental calculation, *Electroencephalography and Clinical Neurophysiology* **94**: 175–182.

Lee, P. (1994). *Bayesian Statistics : An Introduction*, Edward Arnold.

MacKay, D. (1992). The Evidence Framework applied to Classification Networks, *Neural Computation* **4**: 720–736.

McFarland, D., McCane, L., Miner, L., Vaughan, T. and Wolpaw, J. (1997). EEG Mu and Beta Rhythm Topographies with Movement Imagery and Actual Movement, *Society for Neuroscience Abstracts*, p. 1277.

McFarland, D., Neat, G., Read, R. and Wolpaw, J. (1993). An EEG-based method for graded cursor control, *Psychobiology* **21**(1): 77–81.

O' Ruanaidth, J. and Fitzgerald, W. (1996). *Numerical Bayesian Methods Applied to Signal Processing.*, Springer.

Pardey, J., Roberts, S. and Tarassenko, L. (1996). A Review of Parametric Modelling Techniques for EEG Analysis, *Med. Eng. Phys* **18**(1): 2–11.

Peltoranta, M. and Pfurtscheller, G. (1994). Neural network based classification of non-averaged event-related EEG responses, *Medical & Biological Engineering & Computing* **32**: 189–196.

Penny, W. and Roberts, S. (1998). Imagined Hand Movements Identified from the EEG Mu-Rhythm, *Technical report*, Imperial College, University of London. Available via `http://www.ee.ic.ac.uk`.

Penny, W. and Roberts, S. (1999a). Dynamic models for nonstationary signal segmentation, *Computers and Biomedical Research*. To appear.

Penny, W. and Roberts, S. (1999b). Experiments with an EEG-based computer interface, *Technical report*, Imperial College, University of London. Available via `http://www.ee.ic.ac.uk`.

Penny, W., Roberts, S. and Stokes, M. (1999). EEG-based communication: a pattern recognition approach, *IEEE Transactions on Rehabilitation Engineering*. To appear.

Pfurtscheller, G., Flotzinger, D. and Kalcher, J. (1993). Brain-Computer Interface – a new communication device for handicapped people, *Journal of Microcomputer Applications* **16**: 293–299.

Pfurtscheller, G., Flotzinger, D. and Neuper, C. (1994). Differentiation between finger, toe and tongue movement in man based on 40 Hz EEG, *Electroenceph. and Clin. Neur.* **90**: 456–460.

Pfurtscheller, G., Neuper, C., Schloegl, A. and Lugger, K. (1998). Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters, *IEEE Transactions on Rehabilitation Engineering* **6**: 316–325.

Press, W., Flannery, B., Teukolsky, S. and Vetterling, W. (1991). *Numerical Recipes in C*, Cambridge University Press.

Puskorius, G. V. and Feldkamp, L. A. (1994). Neurocontrol of nonlinear dynamical systems with Kalman filter-trained recurrent networks, *IEEE Transactions on Neural Networks* **5**(2): 279–297.

Roberts, S. and Penny, W. (1997). A Maximum Certainty Approach to Feedforward Neural Networks, *Electronics Letters* **33**(4): 306–307.

Roberts, S., Penny, W. and Rezek, I. (1998). Temporal and Spatial Complexity measures for EEG-based Brain-Computer Interfacing, *Medical and Biological Engineering & Computing* **37**(1): 93–99.

Spiegelhalter, D. and Lauritzen, S. (1990). Sequential Updating of Conditional Probabilities on Directed Graphical Structures, *Networks* **20**: 579–605.

Stam, C., van Woerkom, T. and Pritchard, W. (1996). Use of non-linear EEG measures to characterize EEG changes during mental activity, *Electroencephalography and Clinical Neurophysiology* **99**: 214–224.

Wolpaw, J. and McFarland, D. (1994). Multichannel EEG-based brain-computer communication, *Electroencephalography and Clinical Neurophysiology* **90**: 444–449.

World Precision Instruments Inc. (n.d.). *ISO-DAM: Isolated low-noise pre-amplifier. Instruction manual*, Sarasota, FL, USA.

mean	0.7552	0.5321	0.6991	0.5743	0.6355	0.7275	0.6914
S.D.	0.1157	0.1799	0.1234	0.2170	0.1699	0.1281	0.2203

Table 1: *Mean and one S.D. fraction of data classified over all experiments for the seven subjects including all rejected runs (soft rejection). Overall mean classified fraction is* 0.6593 ± 0.1777.

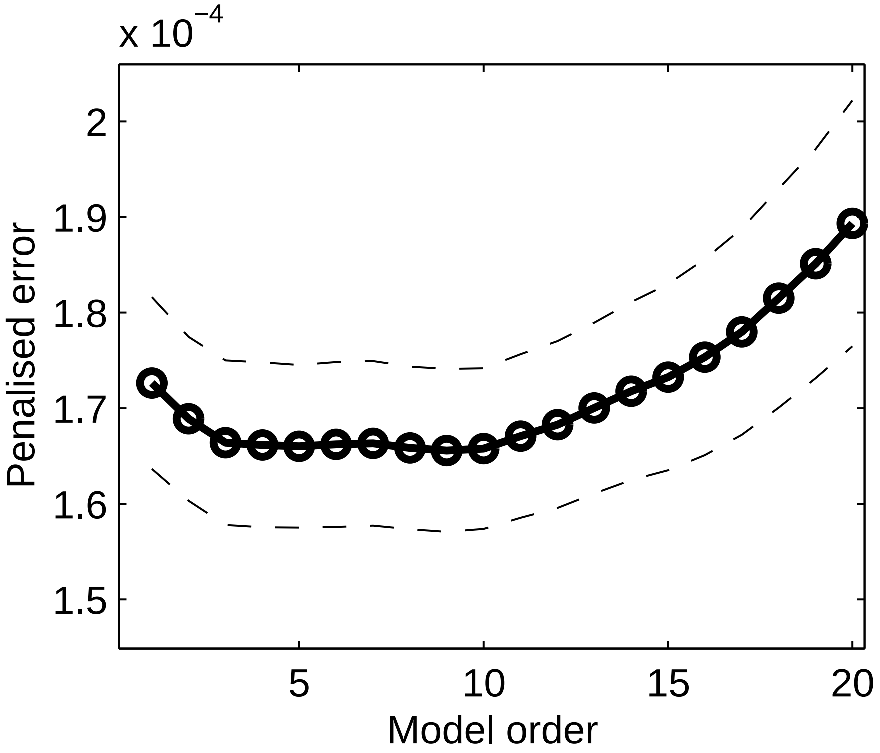

Figure 1: *Model-order estimation for AR modelling. The penalised error functional is fairly flat with model order from 3 to 10.*

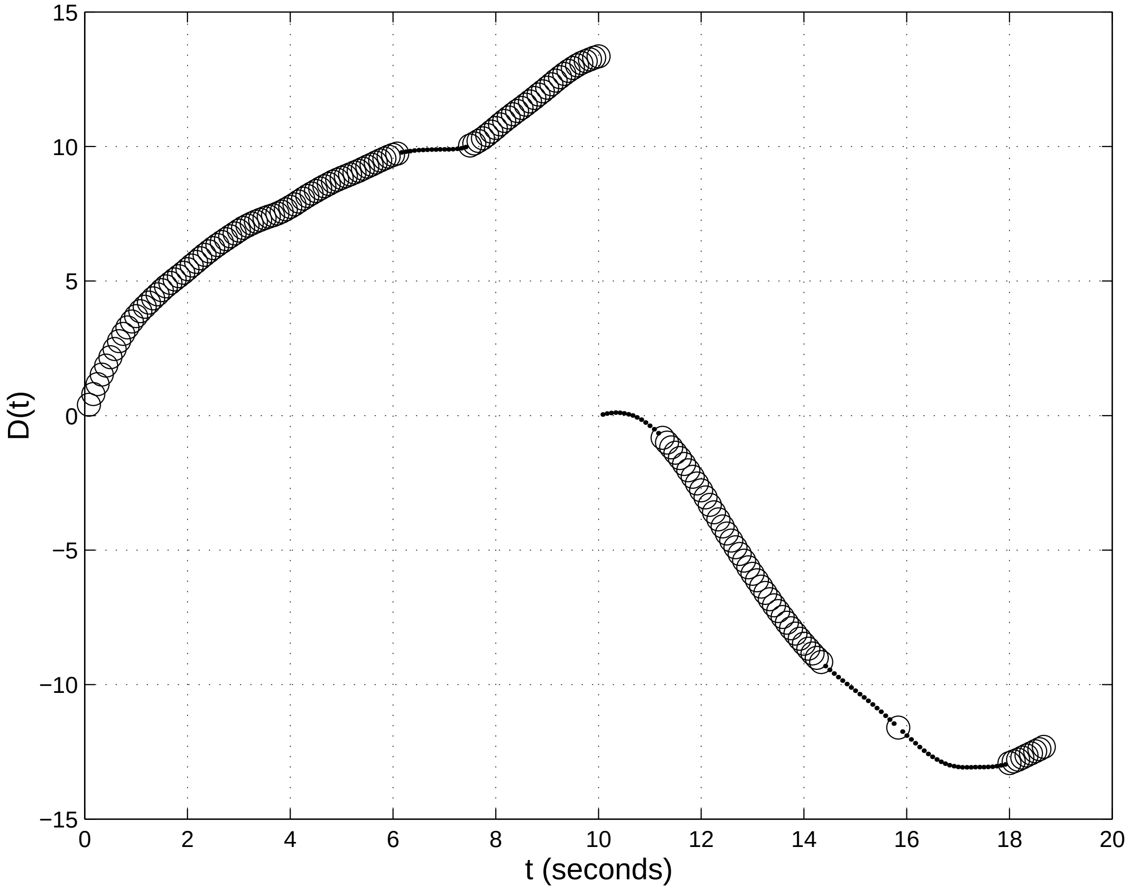

Figure 2: *Cursor displacement measure resulting in classification performance of 82%. A total of 16% of the data was rejected. These samples are shown as smaller points on the graph. Note that these are contiguous blocks.*

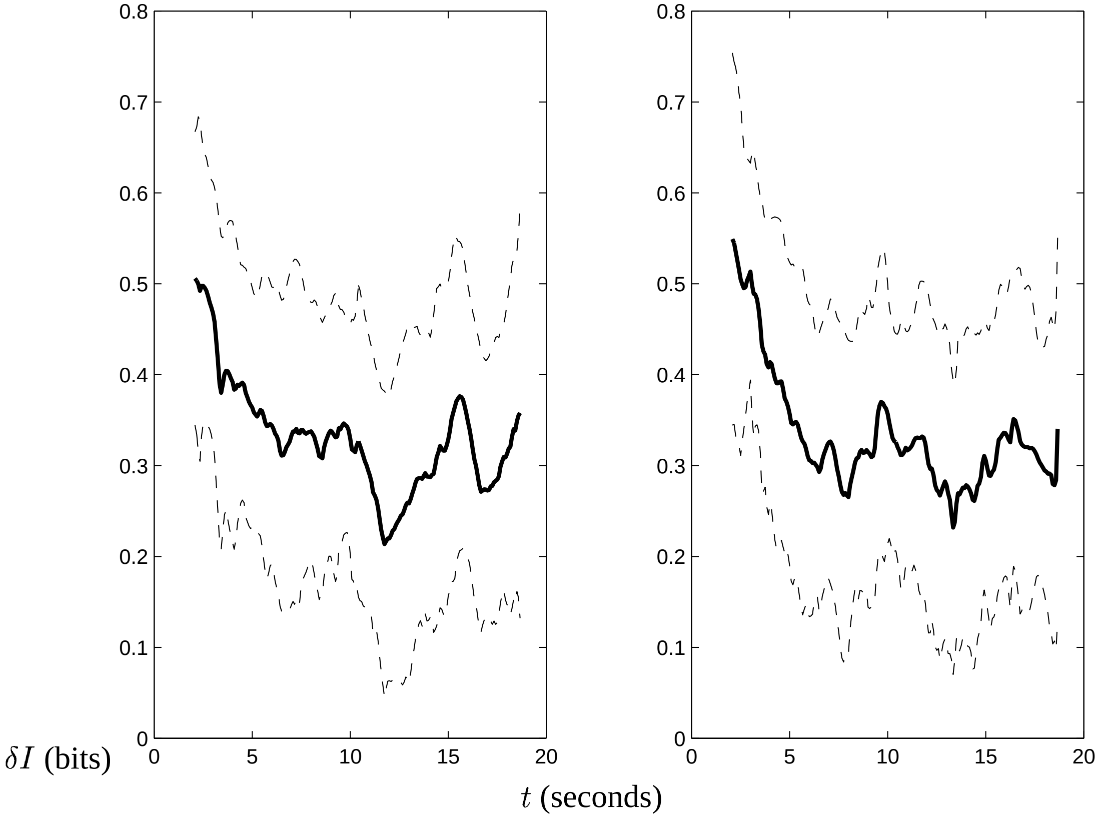

Figure 3: *Mean information gain (bits) with time over all experiments for the first two subjects in the data set (solid line). The dashed lines are at* ±1 *S.D. Note the reduction in information over the first few seconds. The first two seconds are not plotted to avoid any initialisation effects of temporal smoothing.*

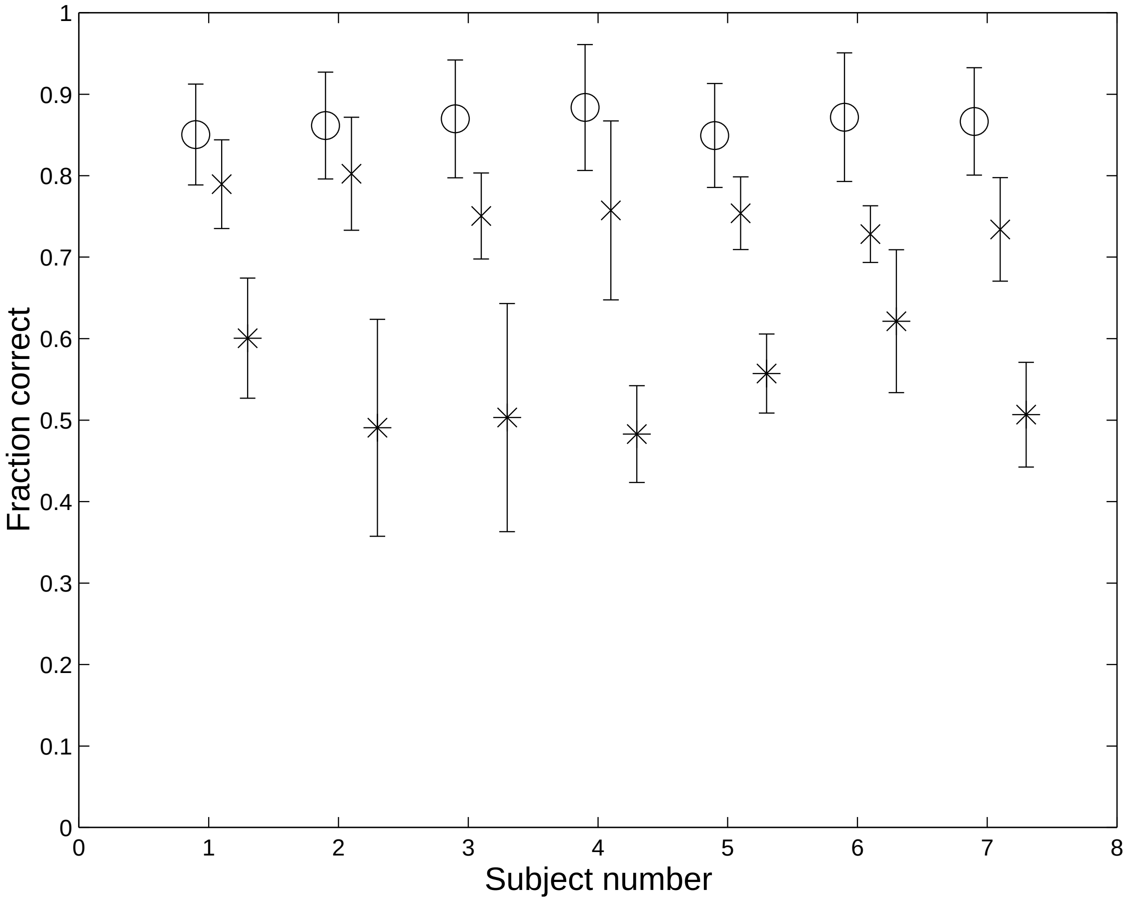

Figure 4: *Mean classification results,* $\pm$ *1 S.D. for all seven subjects. Three sets of results are depicted in this figure, each slightly offset from the corresponding subject number for clarity. Performance for three paradigms are presented; rejection of 'poor' experimental runs (o), intra-run reject option (*$\times$*) and no latent smoothing, posterior moderation or reject option (*). Performance averaged over all subjects is* 0.8648 ± 0.0694 *(o),* 0.7595 ± 0.0667 *(*$\times$*) and* 0.5318 ± 0.1153 *(*).*

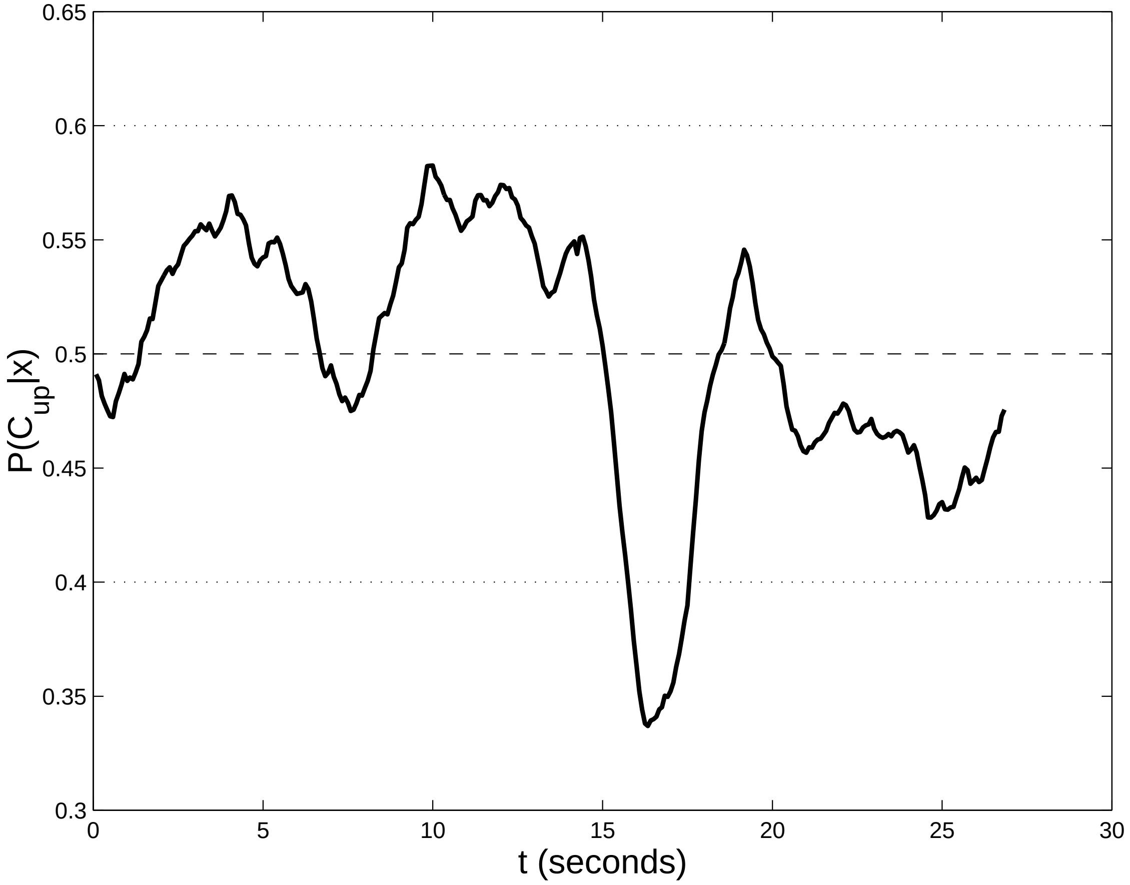

Figure 5: *Control block in which the subject is not attempting to make cursor movements. The solid line represents the moderated posterior probability for cursor up movements. The dashed line is the up/down decision boundary and the dotted lines the bounds of the reject option (as detailed earlier). Note that, save for a small section of data, no cursor-movement classification is attempted and the data is classified to the reject class.*

Figure 1: *Model-order estimation for AR modelling. The penalised error functional is fairly flat with model order from 3 to 10.*

Figure 2: *Cursor displacement measure resulting in classification performance of 82%. A total of 16% of the data was rejected. These samples are shown as smaller points on the graph. Note that these are contiguous blocks.*

Figure 3: *Mean information gain (bits) with time over all experiments for the first two subjects in the data set (solid line). The dashed lines are at* ±1 *S.D. Note the reduction in information over the first few seconds. The first two seconds are not plotted to avoid any initialisation effects of temporal smoothing.*

Figure 4: *Mean classification results,* ± *1 S.D. for all seven subjects. Three sets of results are depicted in this figure, each slightly offset from the corresponding subject number for clarity. Performance for three paradigms are presented; rejection of 'poor' experimental runs (o), intra-run reject option (×) and no latent smoothing, posterior moderation or reject option (*). Performance averaged over all subjects is* 0.8648 ± 0.0694 *(o),* 0.7595 ± 0.0667 *(×) and* 0.5318 ± 0.1153 *(*).*

Figure 5: *Control block in which the subject is not attempting to make cursor movements. The solid line represents the moderated posterior probability for cursor up movements. The dashed line is the up/down decision boundary and the dotted lines the bounds of the reject option (as detailed earlier). Note that, save for a small section of data, no cursor-movement classification is attempted and the data is classified to the reject class.*