A new (semantic) reflexive brain–computer interface: In search for a suitable classifier
A. Furdea\textsuperscript{a, b, *}, C.A. Ruf\textsuperscript{a}, S. Halder\textsuperscript{a, b}, D. De Massari\textsuperscript{a, c, d}, M. Bogdan\textsuperscript{b, e}, W. Rosenstiel\textsuperscript{b}, T. Matuz\textsuperscript{a}, N. Birbaumer\textsuperscript{a, d}
\textsuperscript{a} Institute of Medical Psychology and Behavioral Neurobiology, University of Tübingen, Germany
\textsuperscript{b} Wilhelm-Schickard Institute for Computer Engineering, University of Tübingen, Germany
\textsuperscript{c} Graduate Training Centre of Neuroscience, International Max Planck Research School, Tübingen, Germany
\textsuperscript{d} IRCCS, Ospedale San Camillo, Venezia-Lido, Italy
\textsuperscript{e} Computer Engineering, University of Leipzig, Germany

\textbf{A B S T R A C T}

The goal of the current study is to find a suitable classifier for electroencephalogram (EEG) data derived from a new learning paradigm which aims at communication in paralysis. A reflexive semantic classical (Pavlovian) conditioning paradigm is explored as an alternative to the operant learning paradigms, currently used in most brain–computer interfaces (BCIs). Comparable with a lie-detection experiment, subjects are presented with true and false statements. The EEG activity following true and false statements was classified with the aim to separate covert ‘yes’ from covert ‘no’ responses.

Four classification algorithms are compared for classifying off-line data collected from a group of 14 healthy participants: (i) stepwise linear discriminant analysis (SWLDA), (ii) shrinkage linear discriminant analysis (SLDA), (iii) linear support vector machine (LIN-SVM) and (iv) radial basis function kernel support vector machine (RBF-SVM).

The results indicate that all classifiers perform at chance level when separating conditioned ‘yes’ from conditioned ‘no’ responses. However, single conditioned reactions could be successfully classified on a single-trial basis (single conditioned reaction against a baseline interval). All of the four investigated classification methods achieve comparable performance, however results with RBF-SVM show the highest single-trial classification accuracy of 68.8%. The results suggest that the proposed paradigm may allow affirmative and negative (disapproving negative) communication in a BCI experiment.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Brain–computer interfaces (BCIs) are devices that allow users to convey messages or commands without using the brain’s motor output pathways. Thus, BCIs provide a non-muscular communication channel for individuals who are no longer able to communicate due to severe physical impairment. Neurological diseases such as amyotrophic lateral sclerosis (ALS), muscular dystrophy, locked-in syndrome (LIS) or high spinal cord injury may lead to severe or complete motor paralysis making communication difficult or even impossible.

In locked-in state (LIS) severely paralyzed patients have residual voluntary control over particular muscles (e.g. eye muscles, lips, fingers) (Birbaumer, 2006b). Patients may, however, develop the completely locked-in state (CLIS) in which all motor control is lost (Birbaumer, 2006b). Over the last years it has been shown that patients with severe motor disability and also patients in the LIS are able to control an electroencephalography (EEG) based BCI (e.g. to select letters and thus to communicate) by regulating slow cortical potentials (SCP), sensori-motor rhythm (SMR) or with the P300 event-related potential (ERP) (Birbaumer et al., 1999; Kübler et al., 2005; Sellers et al., 2006; Neuper et al., 2003; Nijboer et al., 2008).

There are no documented cases of CLIS patients communicating by means of BCI. In their meta-analysis of 29 patients in different stages of physical impairment who were trained with BCIs, Kübler and Birbaumer showed that none of the seven CLIS patients ever achieved any BCI control despite intact passive cognitive functioning, assessed with a battery of cognitive event-related potential-tests (Kübler and Birbaumer, 2008; Kotchoubey et al., 2003a, b). Therefore, one of the most challenging goals of BCI research remains the restoration of communication in CLIS. It has been suggested that a paradigm shift from instrumental-operant learning to classical conditioning is necessary to overcome the failure of CLIS patients to achieve BCI control (Birbaumer, 2006a).
Employment of the auditory modality and of a paradigm which requires less attentional resources and less voluntary effort may provide an alternative for the LIS and CLIS patients to communicate using a BCI.

Pavlovian semantic conditioning refers to conditioning of a response to a meaningful word or sentence irrespective of the particular constituent letters or sounds of the word (Razran, 1961). It has been shown that pairing of unpleasant or painful stimuli with language stimuli leads to conditioning. This has been measured at the level of cortical evoked responses (Montoya et al., 1996).

The goal of the current study is to find a suitable classifier to assess the relative performance characteristics of four classification techniques on EEG data derived from a reflexive semantic classical conditioning paradigm. The paradigm aimed at the conditioning of cortical responses to the truth or falseness of a sentence irrespective of the particular constituent words, letters or sounds of the words. A schematic representation regarding the principle of the classical semantic conditioning is presented in Fig. 1.

We provide a comprehensive comparison of four competitive classification techniques: the stepwise linear discriminant analysis (SWLDA), shrinkage linear discriminant analysis (SLDA), linear support vector machine (LIN-SVM) and the radial basis function support vector machine (RBF-SVM). The SWLDA was selected because it incorporates multivariate statistics and it has been successfully applied to EEG classification, especially in regard to the P300 Speller (Farwell and Donchin, 1988; Donchin et al., 2000). The SLDA, a regularized linear discriminant analysis (LDA), was shown to provide superior classification results compared to classical LDA for single-trial ERP classification (Blankertz et al., 2010). SVMs were considered because they deliver high performance in real-world applications. The LIN-SVM was added to serve as a comparison to the SWLDA, SLDA and RBF-SVM in order to evaluate potential benefits of nonlinear kernel methods (see Section 3.3 for details).

To our knowledge, the study represents a first attempt to classify single-trial conditioned reactions to semantic stimuli. It can provide an estimation of the applicability of classical conditioning as an alternative to operant learning paradigms, commonly used in BCIs, in enabling basic communication.

2. Data collection

2.1. Participants

Fourteen healthy participants (8 women and 6 men, mean age 24.36, SD 5.4, range 21–42) took part in this study. None of the participants had previous experience with the conditioning protocol. The study was approved by the Ethical Review Board of the Medical Faculty, University of Tübingen. Each participant was informed about the purpose of the study and signed informed consent prior to participation.

2.2. Task, procedure and design

All participants took part in two experimental sessions on two consecutive days. Each session was subdivided into three blocks, each block consisted of 50 true and 50 false sentences, i.e. trials. For a schematic representation of the trial, block and session setup see Fig. 2.

Blocks one, two and three will be further referred to as the acquisition phase. The sentences were constructed in such way that the

**Fig. 1.** The principle of the reflexive classical semantic conditioning. Top: every CS is paired with US. As a response to the US, a detectable brain reaction is expected. Bottom: CS are not paired anymore with US. A detectable brain reaction is expected, as a result of the classical semantic conditioning.

**Fig. 2.** The experimental design. (a) Session setup: blocks 1, 2 and 3 from sessions 1 and 2 are referred to as the acquisition phase; block 4 from session 2 is referred to as extinction phase. (b) An example of the conditioned stimuli CS+ and CS−. (c) Number of randomly presented trials per block.
last word determined whether the sentence was true or false. In each block the sentences were presented in random order through pneumatic earphones. The conditioned stimuli (CS) were either ‘yes’ or ‘no’ sentences, according to the type of the sentence (CS1 and CS2). To learn ‘yes’ or ‘no’-thinking two different unconditioned stimuli (US) were used: a pink noise US1 immediately following a true sentence and a white noise US2 immediately following a negative sentence which produced the unconditioned brain response (UR). Both US were set to have the same duration of 500 ms and were presented monaurally at different intensities. That is, US1 was always presented to the right ear with an intensity of 75 dB and US2 was always presented to the left ear with an intensity of 105 dB.

For each session, during the first block, every CS was paired with an US and denoted as CS1+ and CS2+. In the second block 10 CS, at random, were not paired with US1 and US2 and are referred to as CS1− and CS2−. Similarly, in the third block 15 CS, at random, were not paired with US1 and US2. The sentences which are not paired with the US are of a particular importance for on-line communication because in an on-line scenario only unpaired sentences can be used. That is, the subject will be asked a question which is either true or false, however the answer to which is unknown to the experimenter.

At the end of the second session a fourth block of sentences was introduced, further referred to as the extinction phase. Hereby, participants were presented with additional 20 CS1− and 20 CS2− trials, that is, with 20 true and 20 false sentences not paired with the US. The purpose of this phase was to explore the lasting effects of conditioning.

Cortical responses were measured with electroencephalogram (EEG). The participants were seated in a comfortable chair and presented with the auditory stimuli. They were instructed to pay attention to each sentence, attend to the last word and then decide as quickly as possible whether the sentence was true or false. The inter-trial interval (ITI) was set at 5 s. A short break followed each block; its duration was determined by the participant and was typically around 3 min.

2.3. Data acquisition and processing

The EEG was recorded using a cap with 32 Ag/AgCl electrodes (EasyCap GmbH, Germany) that were distributed over the scalp based on the International 10–20 system (Sharbrough et al., 1991). Each EEG channel was referenced to the nose and grounded to the left mastoid. To capture horizontal and vertical electrooculographic (EOG) artifacts four EOG electrodes were used. The EEG/EOG signals were amplified using two 32 channel BrainAmp amplifiers (BrainProducts, Germany), sampled at 500 Hz and notch filtered at 50 Hz. All electrode impedances were kept below 5 kΩ. Aspects of stimulus presentation and experimental control were controlled by the BCI2000 (Schalk et al., 2004) and BCPy2000 (http://bci2000.org/downloads/BCPy2000/BCPy2000.html) programs. Off-line analysis was carried out using Matlab R2010a (www.mathworks.com).

2.4. Preprocessing

Prior to segmentation and classification, the signals were band pass filtered between 0.09 and 30 Hz using a 4th order Butterworth filter. The removal of the eye movement artifacts was performed using the ocular correction with independent component analysis tool (ocular correction with ICA) available in the Brain Vision Analyzer software (Brain Products, Germany). A ten channel subset of the 32 channels was investigated and included the following electrodes: Fz, Cz, C3, P3, Pz, P4, P07, P0z, P08 and Oz.

The critical task-related CR changes in the brain responses were expected to manifest themselves within the first few seconds after the end of a sentence, as the presence of the US1 or US2 had a similar influence as a stimulus in an event-related (ER) setting. Thus, for each channel in the subset 2 s segments, either following the end or preceding the onset of a sentence, were extracted. The reason for extracting the segments which preceded the onset of a trial will be explained in more detail in Section 4. The segments were then moving-average-filtered and decimated by equivalent values. For each trial the resulting segments were then concatenated by channel, creating a single feature vector that was used for training the classifiers. We found that the ten channel nose-referenced subset with a moving average window and decimation factor of 10 resulted in the best general performance. Therefore, in the present study we adopted this channel set and employed the corresponding preprocessing technique. The length of the resulting feature vector was 1000 (1000/10 samples x 10 channels).

3. Classification methods

3.1. Stepwise linear discriminant analysis

The stepwise linear discriminant analysis (SWLDA), an extension of Fisher’s linear discriminant (FLD) is considered a benchmark method for determining the optimal separating hyper-plane in a binary classification problem (Fisher, 1936). SWLDA was originally introduced for classifying the P300 component of evoked responses (Farwell and Donchin, 1988) and has become a well established and successful classification method for EEG data (Krusienski et al., 2006, 2008).

Binary linear classifiers have a decision hyper-plane defined by:

$$w \cdot x + b = 0 \tag{1}$$

where $x$ is the feature vector, $w$ is the projection vector (i.e. a vector of feature weights) and $b$ is the bias term. The classification function assigns a given input $x \in \mathbb{R}^d$, with $d$ representing the number of features, the class label according to $\text{sign}(w \cdot x + b)$.

The SWLDA algorithm seeks an optimal discriminant function by adding spatiotemporal features (amplitude values from particular channel locations and time samples) to a linear equation. The input features are weighted using least-squares regression to predict the type of consensus sentence (true or false). The discriminant function starts with no initial features. In each following step, the algorithm evaluates each input feature and the most significant single feature that has a $p$-value of <0.1 is added to the discriminant function. Every new entry to the discriminant function is followed by a backward stepwise analysis and the features that no longer meet the predetermined criterion, i.e. $p > 0.15$, are discarded from the discriminant function. That is, features that no longer account for a significant amount of unique variance after additional features have entered the model are eliminated. This process is repeated until no features satisfy the inclusion/exclusion criteria or until the discriminant function exceeds a predefined number of features (60 for the present study).

3.2. Shrinkage linear discriminant analysis

LDA is based on the estimation of the covariance matrix using empirical covariance. This estimation may become imprecise if the features used for classification have a high dimension and the number of available class instances is low. A systematic error can be introduced by this imprecision: small eigenvalues of the estimated covariance matrix are estimated to be too small and large eigenvalues are estimated to be too large. This can lead to suboptimal
classification performance of the LDA. A method to counter this bias is shrinkage. The empirical covariance matrix $\Sigma$ is commonly estimated as follows:

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$  \hfill (2)

with $\bar{\mathbf{x}}$ being

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$  \hfill (3)

and $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ the $n$ feature vectors used to train the classifier. The estimation error can be compensated for by introducing a tuning parameter $\gamma \in [0, 1]$. The empirical covariance matrix $\hat{\Sigma}$ is then replaced by

$$\hat{\Sigma}(\gamma) = (1 - \gamma)\hat{\Sigma} + \gamma n I$$  \hfill (4)

with $I$ defined as trace($\hat{\Sigma}$)/$d$ of $\hat{\Sigma}$ in which $d$ is the dimensionality of the feature space. An analytic method to calculate the optimal $\gamma$ was shown in (Schaefe and Strimmer, 2005). With $\mathbf{x}_i$ and $\bar{\mathbf{x}}_i$, denoting the $i$-th element of $\mathbf{x}_i$ and $\bar{\mathbf{x}}$ and $s_{ij}$ denoting the element at the $i$-th row and $j$-th column of $\hat{\Sigma}$, the optimal shrinkage parameter $\gamma^*$ can be calculated as follows:

$$\gamma^* = \frac{n}{(n-1)^2} \frac{\sum_{i,j=1}^{d} \text{var}(z_{ij}(k))}{\sum_{i,j=1}^{d} (s_{ij} - \gamma)^2}$$  \hfill (5)

with $z_{ij}(k)$ defined as:

$$z_{ij}(k) = ((\mathbf{x}_i - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_j))^T.$$

This variation of LDA was suggested for ERP classification in (Blankertz et al., 2010).

3.3. Support vector machine

Support vector machines (SVMs) are designed to determine a hyper-plane that maximizes the separating margin between classes, as opposed to LDA, which constructs a probabilistic model for each class using all data points (Vapnik, 1999).

Given a training set of label pairs $\mathbf{x}_i, y_i = 1, \ldots, l$, where $\mathbf{x}_i \in \mathbb{R}^d$, $d$ represents the number of features and $y \in \{-1, 1\}$, the support vector machines (Cortes and Vapnik, 1995; Boser et al., 1992) require the solution of the following optimization problem:

$$\min_{w, b, \xi_1, \ldots, \xi_l} \frac{1}{2} w^Tw + C \sum_{i=1}^{l} \xi_i$$  \hfill (7)

subject to

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i > 0$$  \hfill (8)

The training vectors $x_i$ are mapped into a higher dimensional space by the function $\phi$. The SVM finds a linear separating plane with the maximal margin in the higher dimensional space. $C > 0$ represents a penalty parameter of the error term and $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ represents the kernel function.

The collected data was classified using two kernel functions (i) a linear (LIN): $K(x_i, x_j) = x_i^T x_j$, and (ii) a radial basis function kernel (RBF): $K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2)$, where $\gamma > 0$ represents a kernel parameter (determines the width of the kernel). The RBF kernel is selected because of the following considerations: (i) it maps the samples into a higher dimensional space so that it can accommodate the cases in which the relation between class labels and attributes is nonlinear, (ii) the number of hyper-parameters which influence the complexity of the model selection is lower and (iii) the RBF kernel has less numerical difficulties. However, in some situations in which the number of features is very large the RBF kernel might not be suitable.

To avoid cases in which attributes in a greater numeric range dominate those in smaller numeric ranges and to improve the performance of the SVM, scaling of the input features is necessary. In addition, a scaling factor is computed so that the features in the test data are scaled relative to those in the training data.

When using an RBF-SVM, $C$ and $\gamma$ can greatly influence the classification results, therefore a parameter search must be performed. For each subject a so called ‘grid-search’ (Chang and Lin, 2001) using 10-fold cross-validation is carried out on the training data to determine the best kernel parameters.

For the classification with SVM, LIBSVM was employed, as described in Chang and Lin (2001).

4. Comparison protocol

The feature vectors that were described in Section 2.4 served as an input to all four classifiers. The parameters for each classification method were optimized as described in Section 3 and the algorithms could freely use any subset of the prescribed features for classification.

To classify the EEG responses eight different off-line classification schemes were investigated (see Table 1), wherein either the number of trials or the type of the segments to be classified were changed. In each scheme off-line performance estimates were computed by 10-fold cross-validation, i.e., the classification accuracy estimate in each scheme was reported as the average of the accuracies at each cross-validation step.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>CS1−</td>
<td>CS3−</td>
<td>Acquisition</td>
</tr>
<tr>
<td>II</td>
<td>CS1−</td>
<td>CS3−</td>
<td>Acquisition and extinction</td>
</tr>
<tr>
<td>III</td>
<td>CS1−</td>
<td>Baseline</td>
<td>Acquisition</td>
</tr>
<tr>
<td>IV</td>
<td>CS1−</td>
<td>Baseline</td>
<td>Acquisition and extinction</td>
</tr>
<tr>
<td>V</td>
<td>CS2−</td>
<td>Baseline</td>
<td>Acquisition</td>
</tr>
<tr>
<td>VI</td>
<td>CS2−</td>
<td>Baseline</td>
<td>Acquisition and extinction</td>
</tr>
<tr>
<td>VII</td>
<td>US1</td>
<td>Baseline</td>
<td>Acquisition</td>
</tr>
<tr>
<td>VIII</td>
<td>US2</td>
<td>Baseline</td>
<td>Acquisition</td>
</tr>
</tbody>
</table>

In schemes I and II we were interested to see whether it is possible to accurately separate between the two conditioned reactions, that is, to distinguish ‘yes’ from ‘no-thinking’. The purpose of schemes III to VI was to assess whether single conditioned cortical reactions (either CS1− corresponding to ‘yes-thinking’ or CS2− corresponding to ‘no-thinking’) can be accurately classified. For that purpose, EEG segments preceding the onset of the CS1− and CS2− sentences were extracted and further referred to as baseline. Schemes VII to VIII were meant to serve as additional information about how well the responses are separable in the presence of the US1 or US2.

Following principles of associative learning, trials of the extinction phase should be separately analyzed since they serve a different function (as trials in the acquisition phase do). Due to insufficient number of available trials we pooled the trials of the acquisition phase without any US with those from the extinction phase. Moreover, acquisition trials with US were analyzed separately.
5. Results

Table two shows the averaged performance results for the classification schemes described in Section 4. The columns provide an averaged estimate (over all participants) and standard error of the percentage of single trials that could be correctly classified off-line (the chance performance is 50% correct). Additionally the range of the accuracies is also reported. For two subjects, the SWLDA algorithm did not generate weights therefore the accuracy at the particular cross-validation step was set to 50%. Such cases occurred in schemes II and III.

In schemes I and II above-chance classification was not possible and all classifiers performed around chance level. For schemes III through VIII, above-chance classification was possible, with means ranging from 55.2% to 68.8% accuracy. For schemes VII and VIII, that is, for the classification of the responses in the presence of US1 or US2, the averaged performance estimates ranged between 67.3 and 80.7%. Amongst all of the classification schemes, on average, the highest classification accuracy was consistently achieved by the RBF-SVM, which was closely followed by the LIN-SVM.

To elucidate whether adding the trials from the extinction phase could influence the classification accuracy, the performance estimates were entered into a two factor repeated measures ANOVA. The factors were classifier type (either SWLDA, SLDA, LIN-SVM or RBF-SVM) and conditioning phase (either the unpaired trials from the acquisition phase or the combination of the unpaired trials from the acquisition and extinction phase). Repeated measures ANOVAs were separately carried out for the scheme pairs: I–II, III–IV and V–VI.

The ANOVA did not reveal any significant differences or interaction effects between factors for the scheme pair I–II.

When analyzing the scheme pair III–IV, Mauchly’s test indicated that the assumption of sphericity had been violated for the main effect of classifier ($\chi^2(5)$ = 14.06; $p < .05$). Therefore degrees of freedom were corrected using Greenhouse–Geisser estimates of sphericity ($\epsilon = .70$). There was a significant main effect of classifier type ($p < .001$), when classifiers were set to distinguish CS1 – segments from their corresponding baseline segments, (F1.97, 25.67 = 19.68). The contrasts revealed that performance estimates of the RBF-SVM, (F1, 13 = 26.16; $p < .001$), LIN-SVM, (F1, 13 = 38.97; $p < .001$) and SLDA, (F1, 13 = 12.33; $p < .05$) were significantly higher than performance estimates of the SWLDA. Similarly the performance estimates of the RBF-SVM, (F1, 13 = 8.68; $p < .05$) and LIN-SVM, (F1, 13 = 16.97; $p < .001$) were significantly higher than the performance estimates of the SLDA. No significant main effect of the conditioning phase or interaction between the two factors (classifier type and conditioning phase) was found.

For the scheme pair V–VI, there was a significant main effect of classifier, $p < .001$, in those classifiers which distinguished the CS2 – segments from their corresponding baseline segments, (F3, 39 = 16.33). The contrasts revealed that performance estimates of the RBF-SVM, F1, 13 = 25.47; $p < .001$); LIN-SVM, (F1, 13 = 33.22; $p < .001$) and SLDA, (F1, 13 = 8.90; $p < .05$) were significantly higher than performance estimates of the SWLDA. Similarly the performance estimates of the RBF-SVM, (F1, 13 = 11.06; $p < .05$) and LIN-SVM, (F1, 13 = 9.57; $p < .05$) were significantly higher than the performance estimates of the SLDA. No significant main effect of the conditioning phase or interaction between the two factors (classifier type and conditioning phase) was found. From these results we conclude that addition of the extinction trials had no effect on the classification accuracy.

To determine whether the trial types that were used to train the classifiers have an influence on the classification accuracy, the performance estimates were again entered into a two factor repeated measures ANOVA. The factors were classifier type (either SWLDA, SLDA, LIN-SVM or RBF-SVM) and segment type (CS1 – vs. CS2 – segments, CS1 – vs. baseline segments and CS2 – vs. baseline segments). Because in the previous statistical analysis no significant main effect of the conditioning phase was shown, for the factor segment type, the unpaired trials from both the acquisition and extinction phase were considered. There was a significant main effect of classifier type, (F3, 39 = 20.9; $p < .001$). Contrasts revealed that the performance estimates of the RBF-SVM (F1, 13 = 42.9; $p < .001$) and LIN-SVM (F1, 13 = 42.2; $p < .001$) were significantly higher than performance estimates of the SWLDA. Similarly, the performance estimates of the RBF-SVM (F1, 13 = 13.6; $p < .01$) and LIN-SVM, (F1, 13 = 20.4; $p < .01$) were significantly higher than the performance estimates of the SLDA. In addition SLDA showed higher estimates than SWLDA (F1, 13 = 6.06; $p < .05$) while no significant differences were found between the LIN-SVM and RBF-SVM performance estimates (F1, 13 = 1.6; $p < .2$).

A significant main effect of segment type was also observed (F2, 26 = 35.8; $p < .001$). Contrasts yielded significantly higher performance estimates when using segments from scheme VI compared to schemes IV and II (F1, 13 = 5.64; $p < .05$; F1, 13 = 55.6; $p < .001$). When using segments from scheme IV performance estimates were higher compared to segments from scheme II (F1, 13 = 29.2; $p < .001$).

Interaction between the two factors (classifier type and segment type) was found significant (F3,36, 43.7 = 5.7; $p < .01$) (degrees of freedom corrected using Greenhouse–Geisser estimates). This effect shows that the classifier type had a different effect depending on which segment type was used for classification. To interpret this interaction repeated contrasts were performed. These revealed significant interactions when comparing SWLDA to LIN-SVM for both segment types from scheme II compared to scheme IV (F1, 13 = 18.1; $p < .01$) and scheme VI compared to scheme II (F1, 13 = 24.1; $p < .001$). Similar results were obtained when SWLDA was compared to RBF-SVM (F1, 13 = 10.8; $p < .01$) for scheme II compared to scheme IV and (F1, 13 = 7.5; $p < .05$) for scheme II compared to scheme VI. When comparing SWLDA to SLDA significant interaction was found for segment types from scheme II vs. scheme IV (F1, 13 = 9.7; $p < .01$). A marginally significant interaction was found for segment types from scheme II vs. scheme IV (F1, 13 = 5.2; $p < .05$) when comparing SLDA to LIN-SVM. Finally, no significant interaction was found when comparing LIN-SVM to RBF-SVM.

The interaction graph in Fig. 3 suggests that SWLDA increased the performance estimates to significantly lower levels when the segment type was either from scheme IV (CS1 – vs. baseline) or scheme VI (CS2 – vs. baseline) compared to the other classifiers. All classifiers performed at chance when segment types were from scheme II (CS1 – vs. CS2 –).

Fig. 4 provides an illustration of the grand averaged responses across all sessions and participants for: (a) CS1 – vs. CS2 –; trials (250 trials/class); (b) CS1 – vs. CS2 – trials from the acquisition and extinction phases (70 trials/class); (c) CS1 – vs. CS2 – from the acquisition phase (50 trials/class); (d) CS1 – vs. CS2 – trials from the extinction phase (20 trials/class).

6. Discussion

In the present study we investigated four classification algorithms with the aim to find which is best suited to distinguish conditioned cortical responses corresponding to right or wrong questions. To learn ‘yes’- or ‘no-thinking’ two different unconditioned auditory stimuli were used. We wanted to make sure that the two stimuli (i) are both unpleasant, (ii) stimulate the same modality (using different modalities e.g., vibrotactile for ‘yes’- and auditory for ‘no-thinking’, might lead to biased conditioned responses
towards the most sensitive modality) and (iii) elicit a detectable brain reaction. The grand averaged responses in Fig. 4a show that the brain responses to both US have similar latencies but different amplitudes (the difference in amplitude is due to the different intensities at which they were presented: US<sub>1</sub> at 75 dB and US<sub>2</sub> at 105 dB). If the difference in amplitude is consistent over the trials the classifiers are expected to separate the two reactions.

The investigated classification algorithms were: stepwise linear discriminant analysis (SWLDA), shrinkage linear discriminant analysis (SLDA), linear support vector machine (LIN-SVM) and radial basis function kernel support vector machine (RBF-SVM).

When discriminating CS<sub>1</sub> – from CS<sub>2</sub> – trials, that is, discriminate between ‘yes’- and ‘no-thinking’, none of the investigated algorithms were able to provide results that were above the chance level (see Table 2, schemes I and II). This was also confirmed by the statistical analysis. Looking at Fig. 4a one can see that the reactions elicited by the US<sub>1</sub> and US<sub>2</sub> are different mainly in what amplitude concerns. This was somehow expected, as US<sub>2</sub> was delivered at a much higher intensity compared to US<sub>1</sub>. Fig. 4b and c shows that in the absence of US, the two conditioned reactions, CS<sub>1</sub> – and CS<sub>2</sub> –, have similar morphologies, i.e., similar amplitudes and latencies. This could explain the classification accuracies at chance level

![Fig. 3](image)
**Fig. 3.** Performance estimates for the classifier and segment type interaction. The factor classifier type, either SWLDA, SLDA, LIN-SVM or RBF-SVM is represented by four lines. The factor segment type is represented for segments from schemes II (CS<sub>1</sub> – vs. CS<sub>2</sub> – trials), IV (CS<sub>1</sub> – vs. baseline) and VI (CS<sub>2</sub> – vs. baseline) by diamonds.

![Fig. 4](image)
**Fig. 4.** Grand average plots of the amplitude change at electrode Pz. The black solid line marks the end of a sentence. (a) US<sub>1</sub> and US<sub>2</sub> trials (b) CS<sub>1</sub> – and CS<sub>2</sub> – trials from the acquisition and extinction phases (c) CS<sub>1</sub> – and CS<sub>2</sub> – trials from the acquisition phase (d) CS<sub>1</sub> – and CS<sub>2</sub> – trials from the extinction phase. The light and dark grey shading indicate normalized standard deviation.

**Table 2**
Averaged performance estimates, standard errors and ranges (across the 14 subjects) for the classification schemes investigated.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Classifier</th>
<th>SWLDA [%]</th>
<th>Range</th>
<th>SLDA [%]</th>
<th>Range</th>
<th>LIN-SVM [%]</th>
<th>Range</th>
<th>RBF-SVM [%]</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean ± SE</td>
<td></td>
<td>Mean ± SE</td>
<td></td>
<td>Mean ± SE</td>
<td></td>
<td>Mean ± SE</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td>51.4 ± 4.5</td>
<td>44.0–62.0</td>
<td>50.4 ± 4.9</td>
<td>42.0–58.0</td>
<td>50.3 ± 4.7</td>
<td>41.0–61.0</td>
<td>48.6 ± 3.8</td>
<td>42.0–65.0</td>
</tr>
<tr>
<td>II</td>
<td></td>
<td>50.3 ± 3.8</td>
<td>44.3–57.9</td>
<td>49.2 ± 3.9</td>
<td>38.6–59.3</td>
<td>49.7 ± 3.7</td>
<td>42.9–58.6</td>
<td>50.5 ± 2.9</td>
<td>43.6–62.9</td>
</tr>
<tr>
<td>III</td>
<td></td>
<td>57.3 ± 4.2</td>
<td>50.0–73.0</td>
<td>61.4 ± 4.4</td>
<td>49.0–73.0</td>
<td>64.7 ± 4.5</td>
<td>53.0–75.0</td>
<td>65.9 ± 4.1</td>
<td>49.0–81.0</td>
</tr>
<tr>
<td>IV</td>
<td></td>
<td>55.2 ± 3.5</td>
<td>47.1–66.4</td>
<td>60.6 ± 4.1</td>
<td>52.1–73.6</td>
<td>65.2 ± 3.8</td>
<td>51.4–78.6</td>
<td>65.4 ± 3.6</td>
<td>50.0–80.7</td>
</tr>
<tr>
<td>V</td>
<td></td>
<td>59.1 ± 4.8</td>
<td>52.0–66.0</td>
<td>64.2 ± 4.6</td>
<td>54.0–74.0</td>
<td>67.0 ± 4.4</td>
<td>54.0–77.0</td>
<td>68.8 ± 4.4</td>
<td>49.0–81.0</td>
</tr>
<tr>
<td>VI</td>
<td></td>
<td>60.5 ± 4.1</td>
<td>50.0–70.7</td>
<td>62.1 ± 4.0</td>
<td>50.7–69.3</td>
<td>66.6 ± 4.1</td>
<td>56.4–80.7</td>
<td>68.8 ± 3.7</td>
<td>49.3–82.9</td>
</tr>
<tr>
<td>VII</td>
<td></td>
<td>70.1 ± 2.5</td>
<td>60.8–82.8</td>
<td>67.3 ± 2.0</td>
<td>55.6–76.0</td>
<td>74.4 ± 1.9</td>
<td>62.2–84.8</td>
<td>75.8 ± 1.7</td>
<td>58.4–85.2</td>
</tr>
<tr>
<td>VIII</td>
<td></td>
<td>74.1 ± 2.7</td>
<td>62.4–86.0</td>
<td>68.7 ± 2.0</td>
<td>58.2–81.0</td>
<td>79.1 ± 1.9</td>
<td>69.4–89.2</td>
<td>80.7 ± 1.7</td>
<td>71.0–90.2</td>
</tr>
</tbody>
</table>
in scheme I and II. The repeated measures ANOVA revealed that the classifier type had different effects on performance estimates depending on which segment type was used. In other words, when the classifiers were trained to separate single conditioned reactions (CS₁ – or CS₂ – from the baseline) the average performance on single trial classification increased above chance level. The statistical analysis revealed that the RBF-SVM and LIN-SVM performed significantly better than SWLDA or SLDA. It has also been shown that SLDA performed significantly better than SWLDA. For schemes III through VI average single trial classification accuracies were ranging from 55.2% to 68.8%. The highest accuracies were achieved consistently by the RBF-SVM.

Based on the results from schemes I and II we conclude that discriminating between ‘yes’- and ‘no-thinking’ is not attainable. The reason is likely due to similar characteristics of both US and that the elicited conditioned reactions are similar.

We obtained encouraging results when the analysis was carried out on the level of single conditioned reactions (schemes III to VI). That is, to distinguish either ‘yes’- or ‘no-thinking’ from the baseline (the time period before the onset of a sentence). In accordance to this, a misclassification of ‘yes-thinking’ does not imply that the subject thought of ‘no’ and vice-versa. At a first glance, extracting features from the segments presented in schemes III to VI might be inconvenient to use in practice for patients with ALS. This inconvenience, however, could be overcome by reformulating the same question from the affirmative to the negative. Even if more time consuming, this type of communication would be extremely valuable for patients in CLS.

Again, the results obtained for schemes III to VI confirm that the detection of the resulting conditioned brain responses (on a single trial basis) is possible and therefore we consider that they provide enough evidence that conditioning occurred.

The repeated measures ANOVAs carried out for the scheme pairs (I–II, III–IV and V–VI) showed that the factor conditioning phase does not make a significant contribution to enhancing the performance. That is, the absence of the extinction phase trials had no effect on the classification accuracy. This was probably due to a not yet diminished conditioned reaction during the presentation of the extinction phase.

Although we showed that cortical reactions corresponding to ‘yes’- or ‘no-thinking’ could be classified in both extinction phase and acquisition phase, more work is needed to address the question of long lasting effects. Future studies should address issues such as (i) long term session to session transfer and (ii) whether prior to on-line use presentation of sentences paired with USs is needed to renew the conditioned reactions. Future work is also needed to determine the minimum number of trials required to consolidate conditioned cortical reactions to the true- ness or falseness and thus to reduce the duration of acquisition phase.

For schemes VII and VIII, that is for the classification of the responses in the presence of US₁ or US₂, the averaged performance estimates ranged between 67.3 and 80.7% accuracy. As for schemes III to VI, the highest classification accuracy was obtained when using the RBF-SVM and LIN-SVM.

The usefulness of the RBF-SVM remains to be investigated on real-time/on-line prediction of brain states activated by conditioned ‘yes’/’no’ responses. However, for the present protocol and selected classification algorithms, on-line classification requirements present a lesser concern because the models are relatively simple and static (the algorithms presume simple transforms and inner products using static weights). In addition, because feedback is usually given after a sentence was presented, moderate processing delay can be tolerated. Thus for an on-line application none of the four algorithms is expected to impose an impractical feedback delay.

To summarize, offline classification above chance level of the two cortical conditioned reactions (‘yes’- vs. ‘no-thinking’ is not attainable with any of the investigated classification algorithms. As mentioned earlier, this is likely due to the nature of the two US that were used, that is, both elicited the same type of cortical reaction (although with different amplitudes). Future studies should elucidate whether using US that elicit different types of reactions lead to a better discriminability of the cortical conditioned reactions.

Acknowledgements

The study was funded by Bundesministerium für Bildung und Forschung (BMBF) as part of Bernstein Focus Neurotechnology (BFNT), by Deutsche Forschungsgemeinschaft (DFG) BI 195/58-1 and by European Research Council (ERC) Grant 227632-BCCI. The authors would like to thank Slavica von Hartlieb for carrying out the measurements and Colleen Dockery for the useful comments and the constructive critique.

References


