A Comparison of Common Spatial Patterns with Complex Band Power Features in a Four-Class BCI Experiment.

George Townsend, Bernhard Graimann, Gert Pfurtscheller, Member, IEEE

Abstract—We report on the offline analysis of four-class BCI data recordings. Although the analysis is done within defined time windows (cue-based BCI), our goal is to work towards an approach which classifies on-going EEG signals without the use of such windows (un-cued BCI). To that end, we provide some elements of that analysis related to timing issues that will become important as we pursue this goal in the future. A new set of features called Complex Band Power (CBP) features which make explicit use of phase are introduced and are shown to produce good results. As reference methods we used traditional band power features and the method of Common Spatial Patterns (CSP). We consider also for the first time in the context of a four-class problem the issue of variability of the features over time and how much data is required to give good classification results. We do this in a practical way where training data precedes testing data in time.

Index Terms — Brain-computer interface (BCI), electroencephalogram (EEG) classification, Phase, Common Spatial Filters, Multi-class BCI.

I. INTRODUCTION

There is a growing interest in the field of Brain Computer Interfaces (BCI) as evidenced by the sudden emergence of more and more research groups focusing on this area in recent years [1]. As interest increases in this field, researchers begin to consider more seriously the limitations of this technology and how these limits can be pushed further. One of the most difficult problems in BCI development is the restricted data throughput that can be achieved [2]. This information transfer rate needs to be improved, and often a balance between accuracy and speed must be maintained. One way to help to improve this problem is to consider the use of multiple classes. Obermaier et. al. [3] and Andersen et. al. [4] discuss this problem in the context of a five class interface. Although the addition of classes has the potential to increase speed, the level of difficulty in accurately recognizing higher numbers of classes can counterbalance this gain. In the present work, four-class data involving motor imagery is investigated, and it will be shown that for some subjects, a high degree of accuracy is possible with four classes.

One of the most successful approaches to the accurate detection and recognition of brain patterns in multichannel EEG data associated with motor imagery is the method of Common Spatial Patterns (CSP) outlined by Müller-Gerking et. al. [5] and by Ramoser et. al.[6] and more recently used by Dornhege et al. [7]. Given its success, it seemed reasonable to use it as a basis for comparison against a new phase-related feature that we will introduce here. Although in some cases CSP produces comparable results, it has the disadvantage of requiring a large number of electrodes to get good results [8]. Furthermore, it does not work well in the presence of artifacts [6]. The use of features that require a smaller number of electrodes can be more robust and allow quicker, easier preparation of subjects.

We report on the offline analysis of four-class BCI data recordings. Although the analysis is done within defined time windows, that is cue-based (synchronous), our goal is to work towards an un-cued (asynchronous) approach. To that end, we provide some elements of that analysis related to timing issues that will become important as we pursue this goal in the future. Band power measurements are often applied to the analysis of EEG data, and we use that feature here, however since we derive these measurements in the frequency domain, as a side effect, phase features are also made available by the method. Phase information contained in EEG has not been neglected completely in the past, appearing implicitly in some other kind of feature [9], however we present here an explicit variation of this in the form of a new feature called delta-phase which will be discussed. We consider also for the first time in the context of a four-class problem the issue of how much data is required to give good classification results by providing an examination of training with not only differing amounts of training data, but with data from various times in the past.

The results we present for each analysis are specified in terms of a parameter “kappa” related to the confusion matrix [10]. For the reader who is unfamiliar with kappa, a brief explanation will be given of what it measures and why it is an appropriate parameter by which to gauge classification performance in this context.

Finally, feature selection will be discussed, and we will note an interesting result that the electrodes most often selected tend to be selected on the basis of having good phase features or good amplitude features, but usually not both.
II. METHODS

A. Experimental Paradigm

At the start of the paradigm, the presentation screen was black. At second 2, a cross appeared at the center of the screen and at the same time, an audio tone sounded. By second 3-4, a red arrow appeared, which pointed either left, right, downwards or upwards. In each case, this indicated the imagination of left hand movement, right hand movement, foot movement, or tongue movement respectively.

The test subject was instructed that as soon as the arrow appeared on the screen, they should imagine the specified movement as often as possible (during the following 3s period), until the fixation cross disappeared from the screen. This sequence of events is captured in Fig. 1.

The cross was to be fixated on by the subject during the imagined movement. During a session, each arrow appeared 10 times in randomized sequence.

![Figure 1](image)

Fig. 1. The experimental paradigm begins with a blank screen. After two seconds a fixation cross appears and an audio tone warns the subject to prepare. At second three, where the reference trigger for the data occurs, an arrow appears on the screen, the direction of which indicates which movement imagery the subject should imagine as outlined above.

Sixty EEG channels were recorded, referenced to the left mastoid with ground at the right mastoid. The recording was done by a 64 channel Synamp Neuroscan machine. The recordings were band pass filtered with a broadband antialiasing filter from 1Hz to 50Hz and notch filtered at 50Hz. The sampling rate was 250 Hz.

Between six and nine runs per subject were recorded during a single session. Subjects were given a three to five minute break between each run. Every run consisted of ten trials each of the four classes of imagery described above. There were then 40 trials in total per run. The data from four subjects, namely subjects s1, s2, s3 and s4, was analyzed, and the results are presented and discussed here. The ages of these subjects were 34, 26, 22, and 26 respectively. All the subjects were right handed.

B. Determination of Optimal Training/Classification Time

Our goal is to develop a system for processing this data in an un-cued way, and although this paper focuses on a cued approach, there was one issue that arose early in this research which has consequences in the context of both cued and un-cued paradigms. This issue is related to the time points used for training and testing of the classifier.

Let us discuss this first in the context of an un-cued approach. In such an approach, the timing of the events is clearly defined during the training of the classifier, however there is no time frame of reference during testing of the classifier. As a result, it is interesting and useful to ask the question of whether there is a possibility that some combination of testing and training time points might yield better results than when testing and timing are symmetric (done at the same time point). Even in an un-cued application, training is done in a cued-based manner, since it has to be in order to evaluate the progress of the training and to adjust the training parameters appropriately. Of course this is also true of training in a cued-based application. In other words, both approaches are very similar during training. If we assume that we should test at the same time that we train, there is a possibility that we could miss some alternative set of features which, although they might yield inferior results when tested at the same time point as when training was done, could in fact yield superior performance with some other combination of testing and training times.

For example, perhaps some feature set which produces dismal results when used for training at second four and testing at second four could produce the best possible results of all feature sets and timing combinations if we train at second three, and test at second five. Since there is no time frame of reference concerning when testing will take place in an un-cued approach, this is important information to know in that context; i.e. in an un-cued application, asymmetric testing/training times are a given by definition. This is not a matter simply dismissed, and we felt we ought not to rely on our intuition, but rather that the matter should investigated by considering asymmetric testing and training times for some cases to determine whether it was an issue.

C. Feature Extraction

To test the hypothesis that phase information is important to obtain improved classification results, and to determine how important it is, it was felt that a method of feature extraction that produces explicit phase information should be developed, and that the same algorithms for all stages of the classification process should be tested with and without phase features. To do this, phase and band power information was extracted from Laplacian filtered EEG signals [11] using a sliding hamming window of 0.25s to which an FFT was applied. All of the electrodes involved were suitable for Laplacian filtering, each having four surrounding neighbor electrodes. The FFT yields complex coefficients Xf for each frequency range f. The length of the FFT window used was approximately 250 msec. The subset of these coefficients, representing the frequency bands of interest mentioned below are then considered further in (1) and (2). These equations denote how amplitude and phase information is separated from the complex coefficients. The magnitude and phase of the frequency bands corresponding to eight equally spaced frequency bands were extracted directly as raw features. These bands were 4-8Hz, 8-12Hz, 12-16Hz, ... 31-35Hz. There is a tradeoff between selecting narrow bands that give more specific information about smaller frequency ranges and limiting the number of bands used to reduce the number of features generated. After previous analysis with another data set using the same paradigm, we
determined that these eight bands were reasonable to use and gave good results.

\[ a_f = |X_f| \quad f \in S_f \]  

(1)

\[ \phi_f = \tan^{-1} \frac{\text{Im}(X_f)}{\text{Re}(X_f)} \]  

(2)

Raw phase and amplitude features were derived from the values of the coefficients of the discrete Fourier transform of the raw EEG. From the FFT performed, the instantaneous phase of the signal in each band is taken by calculating the angle of the complex vector for the corresponding frequency band in the FFT. Since we are interested in how quickly the phase angle is changing and in what direction and how quickly, we differentiate this result. The differentiated signal forms a signal representing the change in phase. This is done because the absolute phase without respect to some reference is not meaningful. Equation (3) describes this process. Here, is the time corresponding to one sample. As a final processing step, the phase and amplitude features were then smoothed using a one second moving average FIR filter.

\[ \hat{\phi}_f = \frac{\Delta \phi_f}{\Delta t} = \phi_f - \phi_{f-1} \]  

(3)

Such features were produced for each of fifteen electrodes. Feature selection is a very time consuming process, and we found that in tests where all sixty channels were incorporated, the most often selected electrodes with the most useful features were located in a smaller subset of electrodes. Since there was no gain in incorporating so many electrodes, a decision was made to limit the number of electrodes to fifteen. It is known that the signals from electrodes C3, C4, and Cz reveal the most important signals [3] for the tasks in the experiment, and we confirmed experimentally that these electrodes, and the four immediately surrounding electrodes in each case (see Fig. 2) could be used to provide reasonable results. To incorporate temporal information capturing the changes in the features over time and so that an alternate timing of the features would be available for feature selection, the entire set of features was duplicated and delayed by 0.5 s to augment the feature set. With eight bands, fifteen electrodes, two timeframes, and the two main features of delta phase and amplitude, a total of 480 features were then made available for feature selection. The amplitude component corresponds to traditional bandpower features, however our derivation comes direct from the FFT rather than the more typical approach of bandpass filtering and squaring of the samples. Since both phase and amplitude information appear together as complex variables when applying FFT and are used together, these features were termed “complex band power” (CBP) features.

D. Feature Selection and Classification

In order to perform feature selection to reduce the dimensionality of the feature space down to an appropriate size for the training data available, the Sequential Floating Forward Selection (SFFS) method was used [12]. This method begins with the empty set as the selected features. At each stage, it selects the unselected feature that improves the results the best, then discards previously selected features until the performance stops increasing. These steps are repeated iteratively to a specified maximum number of selected features. Of course such procedures are inherently time consuming, but of the many choices of feature selection algorithms available, SFFS provides a good tradeoff between speed and performance. We found that as more features were selected, the classification accuracy increased up to eight features, then began, in general, to decrease. From the 480 features, the best subset of (up to) eight was chosen based on tenfold cross validation testing of all prospective feature sets tested by the SFFS algorithm. Feature selection using cross validation was performed only on the training data, and the selected features were then blindly applied to the test data to generate most of the results reported. The exception to this is the time courses which appear in Fig. 6 which were generated using ten-by-two-fold cross validation, however in every case features were selected only on the basis of the training partition and applied blindly to the testing partition.

Once the feature set to be used was determined, a minimum distance classifier based on the Mahalanobis distance [13] was trained with the runs from the training data. This classifier was also used during feature selection to determine the fitness of the choices made by the SFFS algorithm. Each point was classified as belonging to the class having the class mean to which it was closest to as measured by the Mahalanobis distance. This measure “scales” the normal Euclidean distance
from a test point to a class mean according to the variance of that class. Equation (4) describes the calculation of the Mahalanobis distance from a point to a class mean in terms of this scaling. Here, $d_i$ is the Mahalanobis distance from the point $x$ to the class mean of class $i$, $m_i$, and the covariance matrix for the class data is $C_i^{-1}$.

$$d_i = (x - m_i)^T C_i^{-1} (x - m_i)$$  \hspace{1cm} (4)

E. Determination of Suitable Training Data

During the initial stages of analyzing the data, only single runs were used, and the results were exceptionally good having been evaluated in tenfold and leave-one-out cross validation schemes. Unfortunately, when the classification parameters obtained during training were applied to the next run, the results became dismal. This suggested that the features selected during training were not completely consistent between runs. If this was true, one possible solution for this problem would be to incorporate more runs in the training data. This hypothesis was tested by using a pair of runs for training and an unseen run for testing, and the hypothesis was supported by a significant increase in performance on the unseen test data. This led to the further question of just how much data is required to achieve good results, and since the original hypothesis was that the problem was related to run to run variations in the features, whether or not older data was as useful for training as newer data. To answer this question, the following analysis was performed. All possible contiguous (in time) sets of training runs were tested against all possible subsequent test runs. The choice for limiting the model this way, as opposed to simply testing all possible subsets of training runs against all test runs was two fold. First, we felt that it was important to model a real application where training data can only come from the past, and second, there is no clear justification for omitting any particular run or sequence of runs, since the combinations that result do not address either of the elements of the concerns. That is, they don’t represent combinations that address the question of the “age” of the data, nor the amount of data. It was felt that insisting that the data be taken from contiguous runs of different ages, on the other hand, was a reasonable approach.

F. Classification using Common Spatial Patterns

As a basis for comparison, the CSP method was also employed on the same broadband filtered EEG data used by the CBP method. (The EEG was also filtered from 8 to 30Hz before applying CSP with little effect on the results.) All 60 channels depicted in Fig. 2 were used. To calculate the covariance matrices, a CSP window of 1500 msec extending backwards from the classification point was used [5]. Window sizes from 250 msec to 2500 msec were tested and 1500 msec was found to give the best general results. This matrix produced in the CSP method represents the averaged variances in this time window. Since the first and last columns are used, one could say that the “first and last features” are used there.

Since the CSP method is designed for differentiating between only two classes, it was necessary to set up CSP filters based on the trials for each class versus the trials for all other classes. Therefore, four individual CSP filters were produced. Since the goal was to compare the methods, an effort was made to maintain as much similarity as possible in the implementations. To that end, eight features were also produced here. These were derived from the first and last rows of the spatial filters for each class. The actual CSP features were produced by calculating the normalized variances for time windows of length 1500 ms (as suggested by Mueller Gerking et. al.[14]). Again, for consistency, the same distance based classifier was used in the same way to produce classification results, but this time using these eight features. The strategy is very appropriate, since the classifier combines features which distinguish between one classes and the others in a way that allows identification of a particular class from many. The outputs of the each of the four CSP filters represent appropriate dimensions to use in this context. In the context of a two class problem, Ramoser showed that it is generally best to use the first two and last two rows of the filters, however that was for two-class problems. In the context of this four-class problem, the results did not improve by doing this. Topographical maps for the CSP method revealed that it generally focused on electrodes nearby those selected by CBP/SFFS, but often outside the fifteen electrodes available to CBP.

G. Using Kappa as a Performance Measure

To aid in interpreting Cohen’s Kappa statistic [15], Fig. 3 shows the relationship between kappa and average classification accuracy for two class and four class problems. Note that this relationship holds only in the event that the number of items to be classified in each class is equal (or balanced). As there are an equal number of trials of each class in every run of the data we discuss in this paper, this is always the case. Since a kappa of zero represents random performance, clearly the relationship between kappa and percentage accuracy depends on the number of classes present. In general, kappa can be between -1 and +1 with values below zero corresponding to “worse than random” performance. Kappa scales from 0 to 1 linearly onto the range between random performance and perfect classification (for balanced classes only). A classification accuracy of 50% in a two class problem is equivalent to an accuracy of 25% in a four class problem, making it difficult to do a fair comparison of multi class problems using classification accuracies. In both cases, the kappa value is the same, specifically zero, indicating random performance. Readers familiar with two class problems should be aware that, for example, 75% average accuracy in the case of a two class problem is equivalent to a performance of 62.5% average accuracy in a four case problem. For both examples, however, kappa is 0.5 representing the midpoint between random performance and perfect classification. Equations (5) through (9) describe the
derivation of the kappa statistic. Despite the complexity of its derivation, it reduces to the linear scaling described above, whenever the number of items in each class is balanced. One can verify, depending on the orientation of the confusion matrix (and we prefer the orientation in which the columns add to the correct number of items in each class), that either all the rows or all the columns will total to the same constant. Under such conditions, it can be shown that the proportion of overall agreement \( P_o \) in (5) and (6) is equivalent to the average classification accuracy rate over all the classes, and that the proportion of chance expected agreement \( P_e \) is 0.25 for four balanced classes. Since all our confusion matrices contain four balanced classes, kappa reduces to \( K_{b4} \) in (10) below. The later term represents kappa for four balanced classes. In the derivation below, \( P_c \) and \( P_r \) represent row vectors containing as elements the sums of the columns and the sums of the rows of the confusion matrix respectively, while \( \Sigma_d \) and \( \Sigma_m \) are the sum of the diagonal elements and the sum of all the elements in the matrix respectively. The mean classification accuracy rate is represented by \( \overline{C}_r \) in (10).

\[
K = (P_o - P_e) / (1 - P_e) \quad (5)
\]
\[
P_o = \frac{\Sigma_d}{\Sigma_m} \quad (6)
\]
\[
P_e = P_c P_r \left( \Sigma_m \right)^2 \quad (7)
\]
\[
P_c = \left[ \Sigma_{c1}, \Sigma_{c2}, \ldots, \Sigma_{cN} \right] \quad (8)
\]
\[
P_r = \left[ \Sigma_{r1}, \Sigma_{r2}, \ldots, \Sigma_{rN} \right] \quad (9)
\]
\[
K_{b4} = \frac{\left( \overline{C}_r - 0.25 \right)}{0.75} \quad (10)
\]

**III. RESULTS**

A. Determination of Testing/Training Time

A typical example of the output generated for investigating training/testing times is shown in Fig. 4. This particular example is for subject s2, run 1. Similar results were generated for many subjects and runs with consistent results, but at lower time resolutions. Since it was very time consuming to generate these at higher time resolutions, we only produced a limit number and picked an exemplary one to show here. The kappa value is shown from zero to one as shades of grey from black to white respectively. Note the presence of the diagonal band in the figure. This indicates that for any given testing time, the best kappa is achieved at the same training (or classification) time, and vice versa. Closer inspection reveals that this band becomes wider as we move further into the trial indicating that this issue becomes less critical further into the trial. For the better performing subjects, s1 and s4, this widening was much more pronounced.

B. Complex Band Power Features

Since we found that the best results are obtained by using the maximum available training data, such data was used to select features and train the classifier both with and without phase information available to gauge the importance of phase. The results appear in Table I which compares the performance of the classifier with only amplitude features (band power) as opposed to including both amplitude and delta-phase features. (We will reference this table again, since it also includes information concerning CSP performance which we will discuss later.) The results in the table show an average performance of Kappa=0.51 without phase features, and Kappa=0.67 when phase features were included.
The labels of the classes appear above while the labels for the final run with training data taken from all previous runs.

I. The abbreviations LH, RH, F, and T stand for left hand, right hand, foot, and tongue movement imagery respectively. The columns of each run were used for both testing and training was the same as described earlier in Table I.

<table>
<thead>
<tr>
<th>Sub.</th>
<th>Train runs</th>
<th>Test run</th>
<th>No phase</th>
<th>CBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>1-8</td>
<td>9</td>
<td>0.83</td>
<td>0.85</td>
</tr>
<tr>
<td>s2</td>
<td>1-5</td>
<td>6</td>
<td>0.17</td>
<td>0.47</td>
</tr>
<tr>
<td>s3</td>
<td>1-5</td>
<td>6</td>
<td>0.51</td>
<td>0.57</td>
</tr>
<tr>
<td>s4</td>
<td>1-7</td>
<td>8</td>
<td>0.51</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Results on kappa of including and excluding phase information appear in last two columns of table. In all cases, testing occurred on the last available run while all previous runs were used for training.

The confusion matrices are shown in Table II for testing of the final run with training data taken from all previous runs. The labels of the classes appear above while the labels for the classification results appear on the left side.

<table>
<thead>
<tr>
<th>Sub.</th>
<th>Train runs</th>
<th>Test run</th>
<th>No phase</th>
<th>CBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>1-8</td>
<td>9</td>
<td>0.83</td>
<td>0.85</td>
</tr>
<tr>
<td>s2</td>
<td>1-5</td>
<td>6</td>
<td>0.17</td>
<td>0.47</td>
</tr>
<tr>
<td>s3</td>
<td>1-5</td>
<td>6</td>
<td>0.51</td>
<td>0.57</td>
</tr>
<tr>
<td>s4</td>
<td>1-7</td>
<td>8</td>
<td>0.51</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Confusion matrices for the four subjects using the CBP method. Data used for testing and training was the same as described earlier in Table I. The abbreviations LH, RH, F, and T stand for left hand, right hand, foot, and tongue movement imagery respectively. The columns of each matrix indicate the actual imagery while the rows capture the classified imagery.

C. Determination of Suitable Training Data

The results generated were of course dependent on the evaluation scheme used. When a single run was used for both training and testing, and even though ten-fold and leave-one-out cross validation schemes were used, the results were much higher than when the more practical and realistic model of using separate runs for training and testing purposes. We felt that using leave-one-out and ten-fold cross validation produced results that were too optimistic. Caution should be exercised when comparing the results here with results based on only single-run cross validated data. (The only results we present based on cross validated data are in the time courses presented later in Fig. 6, however these are each comprised of several runs for testing and several runs for training.) This is a particular pitfall when feature selection is required since limiting testing and training to a single run can result in features being picked that work well with that run but do not generalize well, and without adequate data, it is difficult to work around this problem. For that reason, we decided to present only the more realistic results generated by the use of separate runs for testing and training. The most favorable argument in support of this is that it represents what would be done in a real BCI application. Furthermore, the data evaluated and displayed this way gives us the opportunity to investigate the issue of how much data is required to achieve good results.

In Fig. 5, we see a typical example of the averaged results of all possible training sets as described earlier against all possible test sets. The reader may first choose to view this simply as a randomly organized table of classification accuracies to get a feel for the range of performances over many different training/testing scenarios. A more thorough examination reveals that the rows indicate how many runs were used in the testing data, while the columns indicate which run was the first in the contiguous sequence of runs used. Effectively, the latter captures how far in the past the training data is coming from. Each square represents a particular group of runs used for training data. The number of runs in that set is denoted by the row number on the left. The first run in the set is denoted by the number at the bottom of the column. So, for example, the column labeled “7” captures the two cases where run 7 alone and runs 7 and 8 together are used for training. All runs which follow the last run in the training set are tested separately as test runs to produce a classification result. The averaged result appears in the square in question.

There does not appear to be any clear trend concerning the age of the data, that is, the results are spread out horizontally with no clear increase or decrease in performance accuracy to the left or right. On the other hand, there is a very clear trend concerning the amount of training data used. The results in the columns tend to increase towards the top of the columns. This indicates that the more training data that is used, the better the result. Of course this is neither surprising, nor new, however in this chart we can discern that in the context of this particular four class data and paradigm, minimally three or four training runs (i.e. thirty to forty trials) should be used to obtain good results. Since the age of the data does not appear to be an issue, we can draw the conclusion that it does not really matter which three or four training runs from the past are used. Ironically, the lack of any trend in the rows is a valuable piece of information that serves to inform us that the age does not matter in the context of our paradigm and with the amount of
data we have available. We can speculate that if older data was not as useful, then this would lead to the conclusion that more data would not continue to improve the results, since in the context of this analysis, more data implies older data.

Fig. 5. Averaged results of all possible combinations of contiguous training runs against subsequent test runs. Each row denotes the number of runs used for training. The columns indicate which run was first in the training set. The corresponding kappa value is denoted by the shade of grey of the square along with the numeric value present there. For example, the 0.93 black square in row 5, column 3 is the average classification result of using runs 3 to 7 as training data against runs 8 and 9 as test data.

D. Common Spatial Patterns

Figure 6 shows a comparison of the time courses for classification from 0.5 s before the trigger to 6 s after the trigger. In each course, the black line represents the classification accuracy using CBP features while the grey line is the corresponding classification accuracy for the CSP method. For both methods, the classification accuracy at the start is near kappa=0, representing random performance. Although the windows used by the methods differ, their widths and positions have been optimized for each method providing a fairer comparison than using common widths and positions as a compromise to each method. The CSP method appears to react slightly faster, however the CBP method reaches higher results.

Figure 7 captures a comparison of the CSP method against using CBP features. Notice that largest difference occurs where there are only one or two training sets. CSP is known to require much training data to derive a reliable estimation of the covariance matrix and thus obtain good results.

The artifacts were visually identified by an EEG recording expert. Most of the artifacts were EOG artifacts and would not have played a significant role given the oscillatory features being considered. EMG artifacts were present too but in a smaller proportion. Moreover, the bordering electrodes were rarely selected by the SFFS algorithm, which would not have been the case if artifacts were significant for classification. It is important to note that nearly 30% of the trials used in this analysis were contaminated with artifacts. None of that data was excluded from the study. Despite this, we saw above that very good results were achieved for subject s1 and s4 with the use of complex band power features. For those subjects, the CSP method produced inferior to comparable results. For the other two subjects, s3 and s2, however, CBP performed notably better than CSP. All four results comparing CSP to CBP are shown in Fig. 7. The numbered bars indicate the number of runs used for training, and here sets of runs always beginning with the first available run are used. In each case, the run immediately following the last training run is used for testing.
whose sizes correspond to the ranks are indicated on the values of all the kappas were ranked, and circles and squares most often selected frequency band(s) and electrode(s). The SFFS algorithm, a diagram showing all the characteristics of possible concerning the nature of the features selected by the feature.

Finally, phase information is shown in black or grey circles, diagram (the lower the kappa, the smaller the circle or square). The corresponding main electrode. The two grid lines indicate the bottom axis indicates which electrode was used. The main electrodes (C3, C4, and Cz) are indicated, and in each case, the following four electrodes are, in order, the electrodes immediately to the right, left, behind, and in front of the corresponding main electrode. The two grid lines indicate the most often selected frequency band(s) and electrode(s). The values of all the kappas were ranked, and circles and squares whose sizes correspond to the rank are indicated on the diagram (the lower the kappa, the smaller the circle or square). Finally, phase information is shown in black or grey circles, and amplitude information in black or grey squares. In both cases, the grey color represents the delayed instance of the feature while the black figures represent the non-delayed feature.

IV. DISCUSSION

A. Determination of Suitable Training Data

Fig. 5 indicates that having only one or two training runs available is not sufficient to produce improved results. When three or four runs minimum are used, the results improve considerably, and although a further increase in training data continues to improve the results, the improvement levels off. Therefore, there are some run-to-run differences in the patterns being recognized which require three or four runs to average out. The evidence to support this conclusion is that any contiguous set of runs containing three or four runs total gives similar good results. Using additional training runs continues to average out these differences or inconsistencies. Clearly in a larger study, the age of the data would become an issue, especially when feedback is used. In that context, the subject learns to control and modify more and more suitable brain patterns for the task, and older data begins to become unsuitable for training purposes. In a larger study, we would have to revisit this issue to determine how far back to go when collecting training data.

B. Determination of Testing/Training Time

The clear diagonal trend shown in Fig. 4 gives evidence that symmetric testing and training times yield the best results. This was consistent over all subjects, however this diagonal widened towards the top right corner of the figure particularly in subjects s1 and s2, who were the best subjects. This suggests that for very good subjects, the timing for testing and training is less critical, particular further into the trial, and that later time points in the trial generally yield better results since the brain patterns being recognized are well established as the trial continues. The conclusion is that in the absence of any information to suggest otherwise, it is only necessary to examine symmetric testing and training times when attempting to determine the optimal testing/training time. The implication of this is that in an asynchronous application, the detection accuracy will peak at a point in the ongoing EEG which corresponds to the same time point in the trial to that which was used for testing.

In the context of the current investigation, this was an important result, since feature selection is very time consuming, and it was useful to know whether or not it was worthwhile performing feature selection for all possible test-train timing pairs. If we had deemed it to be a worthwhile exercise, it would have increased the number of feature selection iterations exponentially. For example, to investigate test times in a resolution of 0.25 s from say 1.0 s to 5 s past the trigger, the best features must be found 20 times to determine the best time point to use. If, on the other hand, there was an advantage in testing asymmetric training and testing times, this process would have to be performed 400 times instead.

C. Complex Band Power Features

The results showed that the method of using feature selection with complex band power features produced very good classification results. Although Fig. 8 shows variability across the subjects in the features picked, C3, C4, and often Cz (or an adjacent neighbor of each) were predominantly picked. These results were obtained using only three or four electrodes selected from fifteen channels as compared with the sixty channels required to generate comparable (and in some cases inferior) results using CSP. More importantly, the explicit delta-phase information introduced in this paper is an important feature for achieving improved classification results. Despite that, it does not replace amplitude information. We can

![Image](https://via.placeholder.com/150)
Fig. 8. Consolidated view of all features selected during ten by two fold cross validation of all the runs available for each subject. Electrode numbers appear on the bottom and frequency bands on the side. The grid lines represent the most often selected channel(s) and frequency band(s). Squares represent band power, while circles represent phase information. Dark circles and squares denote current features, while those in grey show the delayed versions of the features.

see this in Fig. 8 by the way that phase and amplitude features are selected in similar proportions to one another. It does, however, augment the information contained in the amplitude features as can be seen in Table 1 where exclusion of the phase information had a significant negative effect on the classification results.

D. Feature Selection

The success of using complex band power features hinges on uncovering specific features that contain the most information about the classes we are attempting to classify. Not all 480 features generated contain useful information and an examination of Fig. 8 shows that the majority of the features available remain completely unused. Some caution must be exercised here, however, since it is probably the case that artificially eliminating some “important” feature by removing it from the feature space might not necessarily have a significant effect on the results. For example, in subject s3, if channel 39 were not available, we can speculate that the arrangement of the many concentric circles that we see at the intersection of the two crossed lines would move over to another one of CH30, CH28, CH19, or C3 in the 12-16Hz band. This also explains why C3, despite its importance, was not selected – all the information provided by C3 was probably present at CH39 as well, however CH39 produced comparable, but perhaps slightly superior classification results. This does not mean that the information at C3 was not important, but merely that it was also sufficiently represented elsewhere. Finally, we can see another interesting result concerning phase and amplitude. Although we can find some minor exceptions, the band-electrode pairs selected appear to “specialize” in either phase or amplitude features. The best example of this can be found in subject s3 at CH39/12-16Hz and C4/16-20Hz. Here, we see a large number of concentric circles at C4 and “concentric” squares at CH39 representing
primarily phase and primarily amplitude features respectively. This requires further investigation and we do not have a good explanation for it at the moment; however, it is an interesting observation.

The frequency band 12-16 Hz includes sensorimotor rhythms (mu rhythms, central beta) originating in the hand representation area (electrode positions C3/C4 and surrounding electrodes). These sensorimotor rhythms are reactive not only during execution of movement as reported by Pfurtscheller et al. 2000 [16], but also during imagination of movement and are therefore also important for discrimination between different types of motor imagery in EEG data.

It was reported recently by Pfurtscheller et al. (submitted) 2005 [17], that frequency components particularly around 12/13 Hz contribute to a high classification accuracy and kappa value close to one.

V. CONCLUSION

The explicit phase information contained in complex band power features has been shown to contain important information for achieving improved classification results. The use of CBF requires fewer electrodes than CSP, and unlike CSP, it produced improved results in our study with far less training data and in the presence of artifacts. These features can be produced efficiently in an online application in a causal way. A further increase in performance can be realized by incorporating a minimum of three to four runs as training data.

We have found that symmetric training/testing times produce the best results, and that this knowledge is important since the amount of computation required to test asymmetric times grows exponentially compared with symmetric timings. The implications of this in the context of an un-cued paradigm are also important since asymmetry is a given there.

ACKNOWLEDGMENT

G. Townsend would like to thank his coauthors for their guidance and advice during the preparation of this paper. Prof. Pfurtscheller’s has continued to be a source of valuable insight and an understanding and patient supervisor. Dr. Graumann has been a kind and generous friend whose firm but gentle criticism has always helped to improve the quality of my work.

REFERENCES


Gert Pfurtscheller (M’00) received the M.S. and the Ph.D. degrees in electrical engineering from the Graz University of Technology, Graz, Austria. He is currently a Professor of medical informatics and the Director of the Institute of Biomedical Engineering at the University of Technology, Graz. He is involved in a number of research programs including the design of a “brain-computer interface” and studies in functional brain topography using event-related desynchronization.