Performance Analysis of a Principal Component Analysis Ensemble Classifier for Emotiv Headset P300 Spellers

Amr S. Elsawy, Seif Eldawlatly, Mohamed Taher and Gamal M. Aly

***Abstract*— The current trend to use Brain-Computer Interfaces (BCIs) with mobile devices mandates the development of efficient EEG data processing methods. In this paper, we demonstrate the performance of a Principal Component Analysis (PCA) ensemble classifier for P300-based spellers. We recorded EEG data from multiple subjects using the Emotiv neuroheadset in the context of a classical oddball P300 speller paradigm. We compare the performance of the proposed ensemble classifier to the performance of traditional feature extraction and classifier methods. Our results demonstrate the capability of the PCA ensemble classifier to classify P300 data recorded using the Emotiv neuroheadset with an average accuracy of 86.29% on cross-validation data. In addition, offline testing of the recorded data reveals an average classification accuracy of 73.3% that is significantly higher than that achieved using traditional methods. Finally, we demonstrate the effect of the parameters of the P300 speller paradigm on the performance of the method.**

I. Introduction

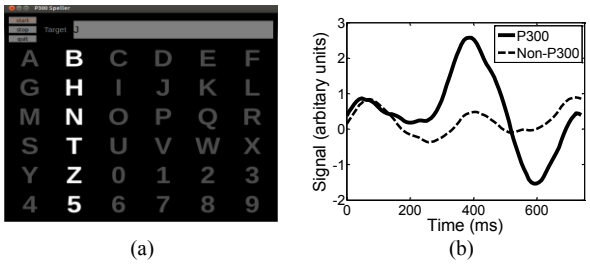
Brain-computer interfaces (BCIs) serve as a communication channel between the brain and the computer to improve disabled people life [1]. For commercial applications, non-invasive BCIs are the most suitable because of their ease of use compared to invasive BCIs. Our study focuses on the P300-based speller BCI using the low cost Emotiv neuroheadset [1]. In this application, a non-intentional signal termed P300 is evoked about 300 ms after the presentation of a rare stimulus. The common mechanism of P300 spellers is to use a grid of characters where its rows and columns are intensified randomly and mutually exclusive as illustrated in Fig. 1a. A P300 signal is evoked when the target character row/column is intensified as illustrated in Fig. 1b. From a machine learning perspective, this problem can be considered as a binary classification problem in which the classifier discriminates among two signal classes: P300 versus non-P300.

Recent advances in electroencephalogram (EEG) recording technology have enabled the production of wireless and portable headsets that could be used for commercial applications [2]. Examples include emotion detection [3], brain-controlled dialing applications for mobile phones [4] and real-time reconstruction of 3D brain activity images [5]. In this paper, we examine the performance of a Principal Component Analysis (PCA) ensemble classifier for P300 speller applications. We have previously introduced this method in [6] where we applied it to the benchmark BCI competition III dataset. However, in this paper we apply this approach to EEG data we recorded using the Emotiv neuroheadset. In addition, we examine the performance of the method with changes in each of the grid row/column inter-intensification interval (ISI), post-stimulus time window used in the analysis and PCA significance threshold. In this approach, PCA is first used to identify the most significant principal components of each individual EEG channel. A classifier is then trained for the i^{th} principal component of all channels combined. Finally, we fuse the outputs of all classifiers where each classifier output is weighted based on the significance of its corresponding principal component. We demonstrate the efficacy of using such method on data recorded using the Emotiv neuroheadset. The results demonstrate an acceptable performance that requires minimal tuning.

II. Methods

A. Datasets Description

Using the wireless Emotiv EPOC neuroheadset - the research edition (Emotiv Systems Inc., San Francisco, USA), we recorded data from three healthy male subjects of different ages. Subjects signed an informed consent approving the use of their data in this study. The Emotiv neuroheadset has 14 electrodes located at positions AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8 and AF4 according to the international 10-20 system. Recorded EEG was sampled at 128 Hz. Subjects were presented with the 6 by 6 grid of characters shown in Fig. 1a where each row/column was intensified for 100 ms [7].



(a) (b)

Figure 1. (a) Our P300 speller interface showing one column intensified while the others are dimmed. (b) Mean response of channel AF3 of the Emotiv neuroheadset for P300 and non-P300 signals of one subject.

Amr S. Elsawy is with the Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University, Cairo, Egypt (e-mail: aselsawy@eng.asu.edu.eg).

Seif Eldawlatly is with the Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University, Cairo, Egypt (phone: 00201005296197; fax: 0020226855582; e-mail: seldawlatly@eng.asu.edu.eg).

Mohamed Taher is with the Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University, Cairo, Egypt (e-mail: mohamed.taher@eng.asu.edu.eg).

Gamal M. Aly is with the Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University, Cairo, Egypt (e-mail: gamal.aly@eng.asu.edu.eg).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

Given that the Emotiv neuroheadset has a limited number of channels and such channels are not the most relevant for P300 applications as reported in [8], it is therefore not expected to achieve good performance in such application. As a result, we examined the use of relatively long ISIs that are expected to enhance the performance as suggested in [9, 10]. We tested different ISI values of 75, 225, and 300 ms.

For each subject and ISI, we recorded two labeled datasets: One that was used as training dataset and the other was used for offline testing. The training dataset consisted of 42 characters and the test dataset consisted of 40 characters. The training and testing sentences were chosen based on English pangrams with digits 1-9 added to span the whole speller grid. Each dataset was recorded in 4 sessions where, in each session, epochs corresponding to 10-12 characters were recorded. For each target character, the user was instructed to focus on the character as the rows/columns were intensified. Each character epoch (i.e. trial) consisted of a sequence of 12 intensifications representing 6 rows and 6 columns repeated 15 times resulting in a total of 180 (i.e. 12*15) intensifications per character epoch [11]. To record the data, we developed our own P300 speller interface using Qt C++ and used the core of the open source smartphone brain scanner project to read the data from the neuroheadset [5]. We adopted such implementation as we intend to use the approach presented here on smartphone devices.

B. Data Preprocessing

Recorded signals were filtered using the common average reference spatial filter to reduce the noise [12]. This was done by subtracting the mean of all channels samples within the same time from each channel sample

$$r_i(j) = s_i(j) - \frac{1}{N}\sum_{k=1}^{N} s_k(j) \quad (1)$$

where $s_i(j)$ represents the raw signal recorded on electrode i at time j, $r_i(j)$ represents the filtered signal and N is the total number of channels. A moving-average filter was then applied with a window of 13 samples to reduce the noise [6]. As an additional preprocessing step, decimation was done on the recorded signals. Given that the sampling rate of the Emotiv neuroheadset is 128Hz, we used a decimation factor of 6 which is half the decimation factor used for the 240Hz dataset reported in [7] .

C. Feature Extraction

In our analysis, we used data recorded from all 14 channels available in the Emotiv neuroheadset. In our proposed feature extraction method, PCA is done on each channel samples separately using the training dataset to obtain the principal components [6]. In the testing phase, data channels are projected onto their corresponding principal components. The projected data from all channels are then grouped on principal component significance bases as shown in Fig. 2. Each feature vector is formed by grouping similar principal components from all channels. The final feature vectors are then fed to an ensemble classifier. The final score is obtained as the weighted sum of each individual classifier score as illustrated in Fig. 2.

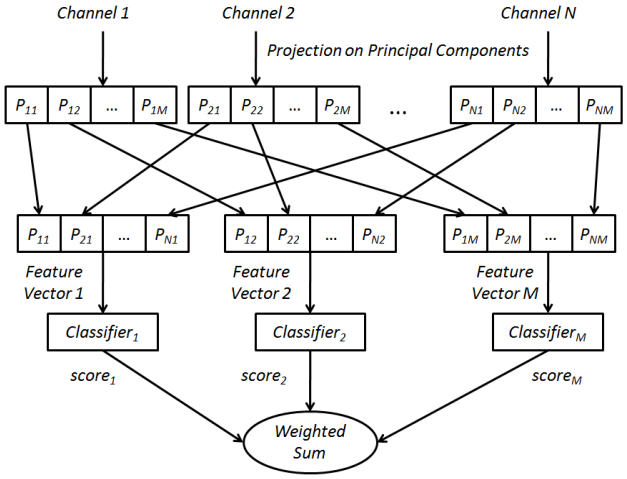


Figure 2. Proposed feature extraction and ensemble classifier. Each channel is projected using its principal components. The corresponding projections are concatenated to constitute a feature vector. A classifier is then trained for each principal component. The final score is a weighted sum of individual classifiers scores.

We compare the performance of the ensemble PCA classifier to that achieved using a concatenated feature vector [6]. In the concatenated feature vector, training data from all channels are concatenated into one vector, then PCA is performed on the concatenated vectors. The obtained principal components are then used in the testing phase to project the testing data to get the final feature vector [6].

D. Classification Methods

Classifying a signal as P300 or not is a binary classification problem. In our analysis, we used linear classifiers whose decision boundary takes the form

$$w^T x + b = 0 \quad (2)$$

where w is the weight vector, b is the bias term, and x is the feature vector. The linear classifiers we examined in this study are Linear Discriminant Analysis (LDA) and Fisher Linear Discriminant (FLD), where the two methods are not the same when the bias term is dropped and this is the case with P300 classification [13].

For each character, a total of 12 feature vectors that correspond to the intensification of 6 rows and 6 columns are classified. Since the goal is to determine one target row and one target column, we determine the target row and column as those that maximize $w^T x + b$. The bias term b is dropped since it is constant among all rows/columns, thus, the classifier score takes the form

$$score = w^T x \quad (3)$$

The target character predicted row r is then determined by

$$r = \arg\max_{row}\left(w^T x_{row}\right) \quad (4)$$

and the target character predicted column c is determined by

$$c = \arg\max_{col}\left(w^T x_{col}\right) \quad (5)$$

The final score is the weighted sum of principal component-based individual classifiers

$$final\_score = \sum_{i=1}^{M} (weight_i * score_i) \quad (6)$$

where M is the number of classifiers, and the score of each classifier is

$$score_i = w_i^T x_i \quad (7)$$

and the weights are based on the eigenvalues of the principal components as we proposed in [6].

Similarly, the predicted row for the target character is the row with maximum final score across all rows

$$r = \arg\max_{row} (final\_score_i) \quad (8)$$

and the predicted column for the target character is the column with maximum final score across all columns

$$c = \arg\max_{col} (final\_score_i) \quad (9)$$

III. Results

A. Cross-validation

We examined the performance of LDA and FLD classifiers on the training datasets of each subject using four feature extraction approaches: PCA ensemble classifier with and without decimation compared to the concatenated feature vector-based classifier with and without decimation. We used a decimation factor of 6. First, we performed 6-fold cross-validation with non-overlapped datasets. This was done by dividing the 42 characters training dataset into 6 sets of 7 characters, where 5 sets were used for training (i.e. 35 characters) and 1 set for testing (i.e. 7 characters). We examined the performance for different post-stimulus time windows (i.e. the length of the feature vector of each individual channel) in the range of 625 ms to 1s. We also examined the performance for different thresholds to select the number of principal components from PCA analysis with values 0.99999, 0.9999, 0.999, 0.995, and 0.99. The window with maximum accuracy averaged across the thresholds was selected. We then selected the best threshold for the selected window as the threshold that maximizes the average accuracy across thresholds of the selected window using a moving average window of length 3.

For the accuracy of statistical significance tests to compare the methods, we redid the cross-validation using only the selected window and threshold parameters. We formed 9 overlapped datasets with 32 characters used to train the classifiers and the other remaining 10 characters were used for validation with an overlap of 6 characters. Each of the four feature extraction methods (EnPCA: PCA ensemble, EnDecPCA: PCA ensemble with decimation, ConPCA: concatenated PCA and ConDecPCA: concatenated PCA with decimation) was tested on the data recorded from all 14 channels and using ISI interval of 300 ms. Fig. 3 illustrates the classification accuracy obtained for the 3 subjects averaged across the 9 overlapped datasets (mean ± SD) for ISI of 300 ms for each feature extraction and classification method. In this figure, maximum classification accuracy obtained for cross-validation datasets across different post-stimulus time windows and different PCA thresholds is reported. As can be seen, all classifiers and feature extraction methods performed equally well with an across-subjects classification accuracy of 86.5±5.9% for the PCA ensemble classifier using LDA and 86.1±4.9% for the PCA ensemble classifier using FLD. The average classification accuracy across all ensemble classifier methods was 86.29±5.4%. This indicates the utility of the PCA ensemble classifier when applied to data recorded using the Emotiv neuroheadset.

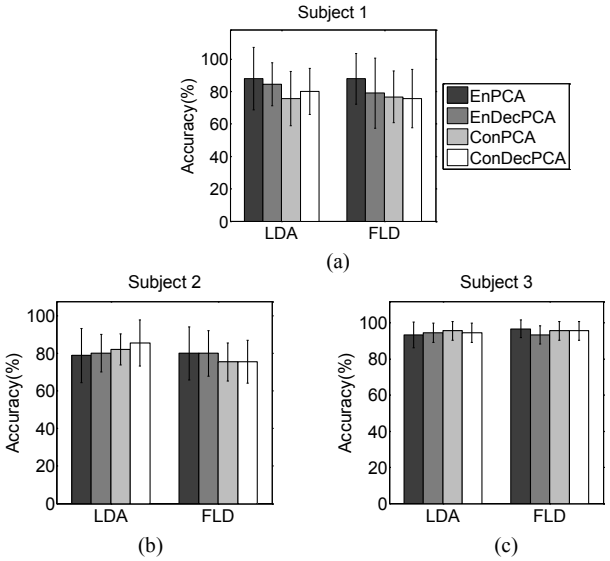


Figure 3. (a) Subject 1, (b) Subject 2 and (c) Subject 3 classification accuracy for different approaches using an ISI of 300 ms.

B. Performance Analysis

We investigated the impact of varying different parameters on the performance of the PCA ensemble classifier. In this analysis, we used our PCA ensemble classifier with FLD without decimation which was the best method on average across all subjects and ISIs. First, we investigated the influence of changing the ISI on the classification accuracy where three different ISIs of 75, 225, and 300 ms were examined. As can be seen in Fig. 4, the average classification accuracy at ISIs of 225 and 300 ms is significantly higher than that of 75 ms ($P < 0.01$, Wilcoxon rank-sum test). This is expected as the overlap between P300 signals corresponding to successive intensified target rows/columns is reduced as the ISI increases [9, 10].

Second, we examined the effect of the post-stimulus time window size on the performance with ISI fixed at 300 ms. Fig. 5a illustrates the accuracy averaged across all subjects showing no significant effect on the accuracy ($P < 0.01$, Wilcoxon rank-sum test). As a result, our approach is independent of the choice of the post-stimulus time window size which simplifies its tuning for different subjects.

Finally, we examined the effect of the PCA significance threshold on the performance with the ISI fixed at 300 ms. As can be seen in Fig. 5b, a significant increase in the accuracy occurs as the threshold is set to a value above 0.999 ($P < 0.01$, Wilcoxon rank-sum test). This indicates that a fixed high threshold could be used for all subjects without the need for significant tuning.

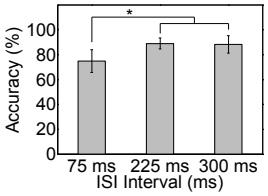


Figure 4. Average classification accuracy across subjects for the PCA ensemble classifier (FLD) for each ISI. *P < 0.01, Wilcoxon rank-sum test.

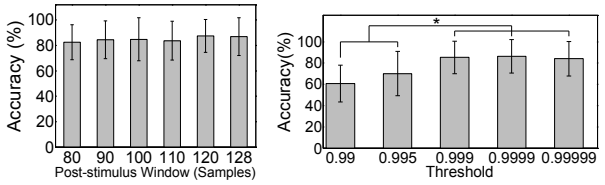


Figure 5. Average classification accuracy across subjects for the PCA ensemble classifier (FLD) for (a) different post-stimulus time windows for the best threshold and (b) different thresholds for the best post-stimulus window. *P < 0.01, Wilcoxon rank-sum test.

C. Offline Testing

Finally, the 40 character testing data were tested for an ISI of 300 ms using all methods for the three subjects. The results are summarized in Table I. Each classification method was examined using the same feature extraction methods investigated in the cross-validation above. We used the number of correct characters as a measure of classification accuracy as opposed to the number of correct classifications of the feature vectors.

Although the cross-validation results reported in Fig. 3 do not show a significant improvement for the ensemble PCA classifier compared to the other approach, the results shown in Table I indicate that using FLD for the ensemble PCA classifier without decimation achieves best performance compared to other approaches. In addition, the results demonstrate less across-subjects variability compared to the concatenated feature vector-based methods which is consistent with our previous results obtained using the benchmark BCI competition III dataset [6].

IV. Conclusion

We examined the performance of the PCA ensemble classifier on data recorded using the Emotiv neuroheadset. We compared the performance of the method to that obtained using a concatenated feature vector-based classifier. Our results indicated that Emotiv neuroheadset can have acceptable results for P300 speller applications using the PCA ensemble classifier. Our approach has the advantage of having lower computational complexity compared to the concatenated feature vector method where the size of the training covariance matrix in the concatenated feature vector method is $N.M \times N.M$ where N is number of channels and M is the number of samples/channel, while for our ensemble classifier, the training covariance matrix size for each principal component is $N \times N$. In addition, the concatenated feature vector size is $N.M$ while for our ensemble classifier is N. This makes the PCA ensemble classifier more suitable for low power devices such as tablets and smartphones. Our analysis revealed that increasing the inter-intensification interval of rows/columns results in an increase in the overall accuracy on the expense of the data transfer rate. Our analysis also indicated that the performance is independent of the post-stimulus window size and the PCA significance threshold. Therefore, these two parameters can be determined without any subject-dependent tuning which simplifies the use of the presented approach.

TABLE I. Offline Test using ISI of 300 ms

Classifier	Feature Extraction	S1	S2	S3	Accuracy
LDA	EnPCA	92.5	40	57.5	63.3±26.7%
	EnDecPCA	92.5	42.5	65	66.7±25%
	ConPCA	90	30	67.5	62.5±30.3%
	ConDecPCA	90	22.5	72.5	61.7±35%
FLD	EnPCA	92.5	47.5	80	73.3±23.2%
	EnDecPCA	90	47.5	80	72.5±22.2%
	ConPCA	92.5	37.5	72.5	67.5±27.8%
	ConDecPCA	82.5	17.5	72.5	57.5±35%

References

[1] B. Graimann, B. Allison, and G. Pfurtscheller, "Brain–computer interfaces: A gentle introduction," in *Brain-Computer Interfaces*, ed: Springer, 2010, pp. 1-27.

[2] C. Kranczioch, C. Zich, I. Schierholz, and A. Sterr, "Mobile EEG and its potential to promote the theory and application of imagery-based motor rehabilitation," *International Journal of Psychophysiology,* vol. 91, pp. 10-15, 2014.

[3] Y. Liu, O. Sourina, and M. K. Nguyen, "Real-time EEG-based human emotion recognition and visualization," in *Cyberworlds (CW), 2010 International Conference on*, 2010, pp. 262-269.

[4] A. Campbell*, et al.*, "NeuroPhone: brain-mobile phone interface using a wireless EEG headset," in *Proceedings of the second ACM SIGCOMM workshop on Networking, systems, and applications on mobile handhelds*, 2010, pp. 3-8.

[5] A. Stopczynski, C. Stahlhut, J. E. Larsen, M. K. Petersen, and L. K. Hansen, "The Smartphone Brain Scanner: A Portable Real-Time Neuroimaging System," *PLoS ONE,* vol. 9, p. e86733, 2014.

[6] A. S. Elsawy, S. Eldawlatly, M. Taher, and G. M. Aly, "A principal component analysis ensemble classifier for P300 speller applications," in *Image and Signal Processing and Analysis (ISPA), 2013 8th International Symposium on*, 2013, pp. 444-449.

[7] D. J. Krusienski, E. W. Sellers, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "Toward enhanced P300 speller performance," *Journal of neuroscience methods,* vol. 167, p. 15, 2008.

[8] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain–computer interfaces," *Journal of neural engineering,* vol. 4, pp. R1-R13, 2007.

[9] E. W. Sellers, D. J. Krusienski, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "A P300 event-related potential brain–computer interface (BCI): the effects of matrix size and inter stimulus interval on performance," *Biological psychology,* vol. 73, pp. 242-252, 2006.

[10] D. J. McFarland, W. A. Sarnacki, G. Townsend, T. Vaughan, and J. R. Wolpaw, "The P300-based brain–computer interface (BCI): effects of stimulus rate," *Clinical Neurophysiology,* vol. 122, pp. 731-737, 2011.

[11] B. Blankertz*, et al.*, "The BCI competition III: Validating alternative approaches to actual BCI problems," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on,* vol. 14, pp. 153-159, 2006.

[12] D. J. McFarland, L. M. McCane, S. V. David, and J. R. Wolpaw, "Spatial filter selection for EEG-based communication," *Electroencephalography and Clinical Neurophysiology,* vol. 103, pp. 386-394, 1997.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*: Wiley-interscience, 2001.