Proceedings of the
7th Graz Brain-Computer Interface Conference 2017 DOI: 10.3217/978-3-85125-533-1-48

NON-STATIONARITY AND INTER-SUBJECT VARIABILITY OF EEG CHARACTERISTICS IN THE CONTEXT OF BCI DEVELOPMENT

T. Krumpe [1], K. Baumgärtner [1], W. Rosenstiel [1], M. Spüler [1]

[1] Department of Computer Engineering, Eberhard Karls University Tübingen, Tübingen, Germany

E-mail: tanja.krumpe@uni-tuebingen.de

ABSTRACT: In a vision of a perfect brain-computer interface (BCI), a user would be able to use the system instantly without the need for a subject-specific calibration, and the performance would remain stable and not deteriorate over time. However, this remains a vision, due to two characteristics of the electroencephalography (EEG) signals: non-stationarity and inter-subject variability. Inter-subject variability describes the fact that the EEG of each person is different and for sufficient BCI communication, the BCI needs to be calibrated separately for each user. Non-stationarity describes a change over time of the EEG signals leading to a decrease in BCI performance with prolonged use. In an approach to better understand these issues, we analyzed the event-related potentials (ERP) and spectral EEG data of 23 subjects in terms of these characteristics. We found that both issues highly affect the data, but we were able to identify a method that nearly eliminates non-stationarity, whereas inter-subject variability remains a major issue that needs to be further addressed.

INTRODUCTION

Using EEG for the signal acquirement in BCI applications is very common and broadly used in research but has some obstacles that prevent a user friendly usage out of the lab. Since scalp EEG recordings represents the summarized activity of a great number of neurons, small changes in the mental state of the user can already make a difference in the overall activity that can be measured. Therefore, training phases before every session are necessary to adapt the system to the current EEG characteristics and mental state of the user in order to guarantee a good performance of the system. Even with long training phases before the start of a session major decrease in performance can occur with increasing time of the session. This will result in a loss of usability and increasing frustration of the subject, which is naturally not desired. Non-stationarity of the signal can be the cause of this, introduced for example by mental changes of the subject (fatigue, disengagement,..) or technical changes (drying electrode gel), leading to differences in the appearance of the trained target signals which in turn leads to a failure of the classifier. Decreasing performance with duration of a session or especially in between offline and online sessions due to a change of the target signal has been observed numerous times. An online, real-time, adaption of the classifier to the upcoming changes in the brain activity is one way to deal with this issue. The classifier is recalibrated by integrating currently recorded data into the already existing data. Supervised [1] as well as unsupervised [2,3] methods have been proposed for the implementation of adaption mechanisms, partly solving this issue. It has also been suggested that the combination of different types of classifiers can reduce the issue of non-stationarity, as they complement each other [4].
Apart from issues within a single session or between online and offline sessions, there are also issues between different subjects that have hardly been solved so far. The issue can be referred to as inter-subject variability which prevents the successful transfer of a previously trained classifier to a new subject, since the differences in EEG signals are usually too big between subjects even though the same task is performed. In some cases normalization methods have been used in order to deal with this variability as for example scaling the data to the mean and standard deviation of a certain number of baseline trials [5]. This can reduce the problem but still cross-subject classification is significantly worse than within subject classification, leaving cross-subject classification an open problem. Other methods like transfer learning, supervised as well as unsupervised, provide the same portion of solution. Information from previously collected trials and subjects can be used to infer knowledge to new and unknown data. Approaches using hierarchical Bayesian models based on Gaussian probability distributions [6,7] or k-Nearest Neighbor approaches [8] for training and optimizing a classifier based on old and new data have been introduced. Almost all approaches are still adaptive since the collection of new and subject specific data is necessary to update the classifier and to integrate subject specific information into the classification approach, therefore persisting the necessity of individual training.

In contrast to this, other approaches were implemented in which researchers used the distinct differences of EEG signals of different subjects during the same task to their advantage. It could be shown that an identification or authentication of a specific subject out of many is possible. Armstrong and colleagues used ERP characteristics [9] and Palaniappan features based on the power spectrum [10] for the authentication of a specific subject which was successfully with almost no errors. This opens up new possibilities for applications using EEG-BCI technology.

The aim of the analysis of this paper is to quantify the extent of non-stationarity and inter subject variability in EEG data to better understand these properties. A large dataset from a standard BCI application was chosen for an anew analysis with respect to the mentioned factors. The data set allowed within and between session as well as between subject comparisons which were evaluated with classical correlation and classification measures. The following sections will describe the properties of the chosen dataset, the methods used for quantification of the two stated issues and several approaches to deal with non-stationarity in EEG data.

MATERIALS AND METHODS

Data: A dataset of Spüler and colleagues [11] consisting of EEG recordings of 23 subjects participating in a P300 speller experiment was used for the analysis. The dataset consists of 2 sessions on two different days for each subject. The participants can be divided into three groups according to their age and health status. Group 1 consists of 9 subjects between age 20 - 28, Group 2 of 8 subjects between age 39 - 52 and the third group was a patient group with severe motor impairments consisting of 6 subjects between the age of 36 - 63. The montage was, in standard 10/20 positions, with electrodes at positions F3, Fz, F4, T7, C3, Cz, C4, T8, CP3, CP4, P3, Pz, P4, PO7, PO8, Oz. Ground and reference electrodes were placed at the left and right mastoid, respectively and the signal was sampled at 256 Hz. The P300 speller consisted of a 6x6 matrix in which the 12 rows and columns of symbols were flashed in random order. Each intensification lasted for 62.5 ms and the matrix remained blank for 125 ms between flashes. In one sequence, each row and column were flashed exactly once. One trial contained 2-10 sequences, depending on the subject. This dataset was chosen as it was comparably big and it included two sessions per subject. Therefore, it provided the opportunity to investigate changes across subjects and changes over time within a session as well as across sessions. To that end, inter subject variability as well as non-stationarity can be evaluated and compared within this dataset.

Data processing: For the analysis, ERPs and power spectra of the data were evaluated. The power spectra were calculated for each trial separately with Burgs maximum entropy method [12] (modelorder 16), from 1 to 30 Hz in 1 Hz bin. In contrast to that, ERPs were averaged over several target intensifications, to improve the signal to noise ratio. One ERP in the analysis was calculated by sequentially averaging over 50 target intensifications using a sliding window approach with a step size of 10. This lead to an average number of 237 trials ($\pm$ 49) and 145 ($\pm$ 68) ERPs for session one and 273 trials ($\pm$ 80) and 169 ($\pm$ 68) ERPs for session two.

Inter subject variability: Evaluating the effect of inter-subject variability was done by a classification approach that aimed to assign ERPs and power spectra to the subject they originated from. Two different approaches were tested, one with the aim to identify a subject correctly on the basis of the ERPs (or power spectra) and the other one with the aim to authenticate a subject on the same signals. Since distinct differences between subjects are assumed to be present, a SVM classification approach should be able to separate the data of different subjects according to their differences. For the identification scenario a one vs one classification was implemented in which it is tested how well a subject can be identified within a set of subjects. Therefore, one classifier was trained on one session for each pair of subjects and tested on the second session of all possible pairs. To determine the accuracy a multiclass-classification was performed.

For the authentication scenario a one vs all classification was implemented. One individual classifier was trained for each subject to distinguish between data of the subject itself and the data of all others. Again one session was used for training, the other session used for testing. In this case sensitivity and specificity were used as performance measures to account for the highly unbalanced classes. If the signals of one subject can be extracted from a variety of signals, it validates that the signal of a single subject does stand out and is distinct. It can be seen as an authentication approach since a yes or no decision is made, answering the question if the signal belongs to the person in question.

In consequence both classification approaches reveal a measure to quantify the inter-subject variability within the given dataset. The higher the performance measures the higher the variability between subjects. For both approaches a C-SVM from the libsvm implementation [13] for Matlab was used in a 5-fold cross-validation. The data of all 16 electrodes was used for the classification.

Non-stationarity: To evaluate the non-stationarity of the signal linear regression models were fit to the ERPs (and the power spectra) and the time of their occurrence in the recording (1...n), again in a 5-fold cross-validation. The occurrence in the recording was labeled consecutively with increasing numbers representing the time of appearance. This was done individually for each subject and session. The regression models are evaluated by calculating R^2 values to estimate how well the time of recording can be predicted from the EEG signals. The R^2 value denotes the proportion of variance in the target variable that is explained by the predicted values. Systematic changes over time that can be described by a linear function should lead to a strong correlation, therefore, quantifying non-stationarity to a certain extent. In addition to quantifying non-stationarity, several methods to decrease the influence of non-stationarity were tested and evaluated. The measure for quantification of non-stationarity after the application of the tested approaches remained the same: A fitted linear regression model and its corresponding R^2 values, representing the correlation between actual and predicted time in recording.

(1) First covariate shift adaption was applied to the data. It reduces trial to trial variability in the distribution of spectral power, by normalizing the power with an averag-

DOI: 10.3217/978-3-85125-533-1-48

ing approach shifting over a certain number of preceding trials. The originally proposed window size of w = 15 was used [14].

$$(t) = P(t) - \frac{1}{w}(\sum_{i}^{w} P(t-i)) \qquad (1)$$

The hypothesis is that an overall reduced variability might erase the change in signal introduced by non-stationarity. (2) As a second approach lateral symmetry was calculated on the data for the electrode pairs (F3-F4, T7-T8, C3-C4, Cp3-Cp4, P3-P4, Po7-Po8). It reveals lateral disparities or asymmetries between the two hemispheres and could possibly minimize systematic changes that are present in the signal. Two different ways of calculating the difference were applied. Once the signal of the two electrodes was subtracted before transferring it into the frequency domain, and the other time the signal was subtracted after transferring it to the frequency domain.
(3) As a last approach event related desynchronization or synchronization (ERD/S) was computed by using the difference of the power spectral density of the trial and a time frame of equal size shortly before (pretrial) [15]. The quotient of the difference and the power of the pretrial reveals the ratio of how much the power has changed due to an event by a (de-)synchronization of firing neurons. Again the hope is that this mathematical approach might erase a possibly present systematical change in the signal.

$$ERD/S = \frac{P_{trial} - P_{pretrial}}{P_{pretrial}} \qquad (2)$$

RESULTS

Tab. 1 shows the results of the analysis concerning non-stationarity in the data. It includes the R^2 values between the actual time in recording of a trial and the predicted time (with regression methods) for session one. Session two revealed the same trend which is why only the results of one of the two is shown. The individual columns represent the different signals that were used for the analysis, standard ERP and power spectrum and the transformed data according to the four suggested methods. When looking at column two and three it can be seen that the prediction quality is much higher for the spectral density distribution than for the ERPs on average for all subjects. When comparing those two columns to the remaining ones in the table it can also be seen that applying lateral symmetry (difference calculated after) to the data reduces the R^2 value notably, whereas ERD/S reduces it to almost 0. The other two methods have smaller or no notable effects on the correlation with the time of recording. Tab. 2 shows the results of the evaluation of inter-subject variability. It is quantified by the classification performance measured in accuracy or sensitivity and specificity depending on the mode of classification. The table is divided into two parts, each representing one mode of classification. Identifying the signals of an individual subject (across sessions), represented in the left part of the table, works better on the basis of ERPs than on the basis of the spectral density distribution (One vs One). The same observation can be made for the authentication approach (One vs All). The One vs All approach reaches very high specificities (TNR) for both signal types (above 97%), whereas the sensitivity (TPR) is much lower in both cases, but significantly better for ERP than spectral data (0.76 vs 0.49 on average respectively). In both approaches it can be seen that the variance of the performance between subjects is rather high, especially for classification on the spectral data. Both tables reveal that there seem to be differences between the three groups of subjects more or less pronounced throughout the various approaches.

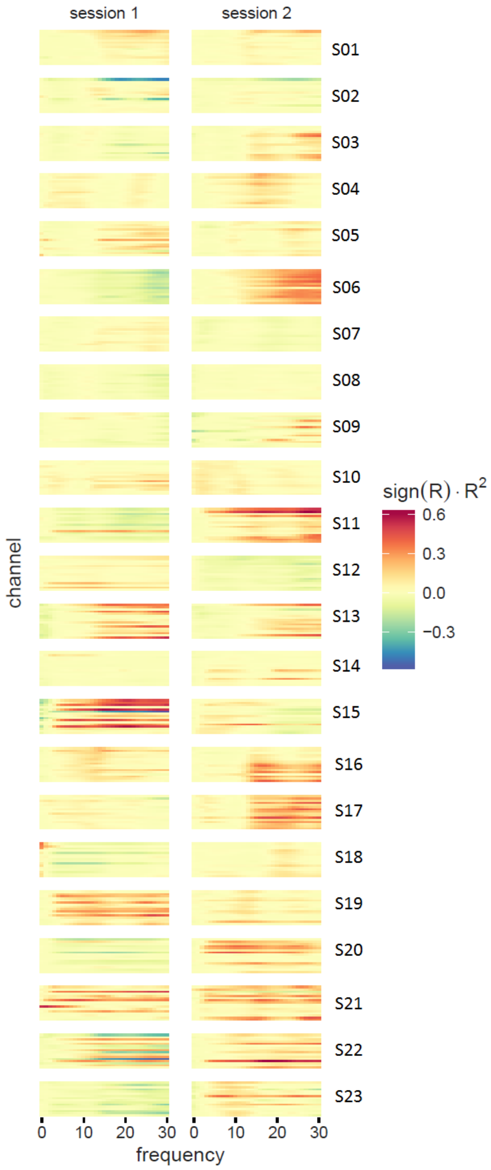

Figure 1: The colors indicate the squared correlation (R^2) between spectral features and the time of recording for each subject and both sessions. Channels from top to bottom: Po8, Oz, Po7, P4, Pz, P3, Cp4, Cp3, T8, C4, Cz, C3, T7, F4, Fz, F3)

Table 1: Correlation (R^2) of the actual time and predicted time of each trial in the recording, regression based, for 23 subjects of 3 groups. a) Covariate Shift adaption, b) Lateral symmetry (before), c) Lateral symmetry (after), d) ERD/S

Sub	ERP	Spec	a)	b)	c)	d)
S01	0.20	0.38	0.40	0.38	0.22	0.01
S02	0.37	0.80	0.62	0.82	0.71	0.11
S03	0.37	0.33	0.63	0.20	0.04	0.03
S04	0.20	0.23	0.57	0.72	0.04	0.01
S05	0.30	0.67	0.46	0.60	0.45	0.12
S06	0.29	0.22	0.43	0.29	0.06	0.06
S07	0.50	0.24	0.51	0.35	0.08	0.07
S08	0.20	0.16	0.49	0.31	0.06	0.01
S09	0.21	0.26	0.36	0.10	0.05	0.02
Mean	**0.29**	**0.37**	**0.50**	**0.42**	**0.19**	**0.05**
S10	0.16	0.41	0.46	0.47	0.19	0.01
S11	0.49	0.73	0.36	0.48	0.35	0.02
S12	0.14	0.26	0.20	0.32	0.15	0.03
S13	0.34	0.81	0.62	0.76	0.71	0.04
S14	0.30	0.22	0.66	0.44	0.09	0.06
S15	0.19	0.83	0.32	0.73	0.69	0.07
S16	0.28	0.54	0.35	0.59	0.14	0.02
S17	0.13	0.67	0.43	0.57	0.32	0.04
Mean	**0.25**	**0.56**	**0.42**	**0.55**	**0.33**	**0.04**
S18	0.12	0.85	0.36	0.89	0.75	0.11
S19	0.08	0.81	0.54	0.80	0.58	0.01
S20	0.18	0.78	0.37	0.80	0.76	0.10
S21	0.02	0.74	0.60	0.66	0.77	0.01
S22	0.14	0.89	0.25	0.56	0.87	0.04
S23	0.08	0.60	0.71	0.75	0.55	0.03
Mean	**0.10**	**0.78**	**0.47**	**0.74**	**0.71**	**0.05**
Mean	**0.23**	**0.54**	**0.46**	**0.55**	**0.37**	**0.04**
std	0.13	0.26	0.14	0.22	0.30	0.04

Especially notable are the high R^2 values for spectral data of the patient group in contrast to the other two groups of subjects. Fig. 1 visualizes the analysis of the non-stationarity by plotting the R^2 values for each channel and the respective spectral features of each subject for both sessions. It is included as a showcase to show the variance between and within subjects to highlight the problem statement.

DISCUSSION

The results presented in Tab. 1 and Tab. 2 revealed that non-stationarity and inter-subject variability can be measured and quantified with the proposed methods. Both are highly present in the dataset underlining the importance of awareness of these issues during BCI development. Non-stationarity was assessed by detecting a systematic change of EEG characteristics over time by linear regression methods. It is strongly present in the power spectra of the signal and a little less prominent in the ERP data. Depending on the application, this change in signal might not be a relevant issue, but to ensure the use of valid features that are related to the cognitive process of interest and not to a systematically introduced artifact, this effect needs to be eliminated.

Table 2: Classification performance quantified in accuracy or TPR (sensitivity) and TNR (specificity) in a single-subject approach - training on session 1 - testing on session 2, with a C-SVM and a linear kernel in a 5-fold cross-validation.

	One vs All				One vs One	
Sub	ERP		Spec		ERP	Spec
	TPR	TNR	TPR	TNR	Acc	Acc
S01	0.80	1.00	0.95	0.98	0.99	0.90
S01	0.95	1.00	0.61	0.99	0.99	0.79
S03	0.98	1.00	0.81	0.95	0.99	0.91
S04	0.82	1.00	0.80	0.97	0.97	0.91
S05	0.88	1.00	0.40	0.99	0.85	0.18
S06	0.59	1.00	0.44	0.93	0.65	0.18
S07	0.90	1.00	0.53	0.94	0.85	0.72
S08	0.78	0.98	0.85	0.99	0.91	0.90
S09	0.86	1.00	0.39	0.97	0.96	0.51
Mean	**0.84**	**0.99**	**0.64**	**0.96**	**0.90**	**0.66**
S10	0.67	1.00	0.96	0.99	0.70	0.95
S11	0.75	1.00	0.00	0.99	0.95	0.00
S12	0.85	0.99	0.48	0.95	0.92	0.45
S13	0.86	1.00	0.38	0.97	0.99	0.53
S14	0.97	1.00	0.85	0.99	0.99	0.72
S15	0.25	0.99	0.00	0.96	0.33	0.01
S16	0.63	0.99	0.00	0.96	0.80	0.00
S17	0.94	1.00	0.18	0.98	0.93	0.38
Mean	**0.74**	**0.99**	**0.36**	**0.97**	**0.82**	**0.38**
S18	0.80	1.00	0.66	0.90	0.80	0.88
S19	0.33	0.99	0.22	0.94	0.56	0.24
S20	0.60	0.99	0.01	1.00	0.75	0.00
S21	0.53	1.00	0.26	1.00	0.65	0.27
S22	0.82	0.98	0.59	0.94	0.88	0.70
S23	0.97	1.00	0.89	0.95	0.97	0.94
Mean	**0.67**	**0.99**	**0.44**	**0.95**	**0.77**	**0.51**
Mean	**0.76**	**0.99**	**0.49**	**0.97**	**0.84**	**0.53**
std	0.20	0.01	0.32	0.02	0.17	0.35

Our analysis was able to show that the systematic change over time can be reduced by using the lateral symmetry (R^2 of 0.30), whereas ERD/S can reduce the effect of time to a minimum (R^2 of 0.04) if not eliminate it completely. The systematic change over time affecting ERPs was rather weak. The R^2 value accounted for a squared correlation of 0.23 on average, leading to the assumption that the effect is merely present or can at least not be predicted well with linear regression methods. Non-stationarity therefore seems to be a bigger issue when dealing with frequency-domain than with time-domain features. Interestingly the patient data (subject 18-23) showed a much higher correlation with time in the power spectra than all other subjects did. Nevertheless ERD/S provided a solution for this equally well to subjects with high and also to subjects with low correlation of the power spectra with time of recording. No further universal trends can be derived concerning the different groups of subjects. At this point it needs to be mentioned that the evaluated non-stationarity is linear only, non-linear non-stationarity, which also exists, has not been accounted for in this paper and needs to be addressed further. It can be observed that the variance is in general very high for all evaluated measures between the groups but also within the groups leading over to the topic of inter-subject variability. A first assumption can be made that the inter-subject variability is supposedly very high since a high variance between the R^2 values can be observed. No pattern across subjects is visible, hence for each subject other features provide discriminability. Inter-subject variability was investigated with a classification approach and the achieved performance measures suggest differences between subjects are reflected more strongly in ERPs than in the power spectra. The assignment of the current trial to the correct subject, in a pairwise or overall comparison, can express the stability of the signal across time and sessions or the great variability between subjects. Both are equally valid assumptions that do not exclude each other. The results showed that an assignment of the correct subject based on the ERPs was possible in both classification approaches with good performance values whereas the assignment on the basis of the power spectra worked less well. Since the train and test set were taken from two different sessions it can be assumed that ERPs contain very specific sections that can be identified across sessions. Due to the high performance values it can also be assumed that differences between the subjects must be very distinct, again at least for the ERPs. The performance values therefore suggest, that a biometric use of P300 ERPs could be feasible for an identification as well as an authentication of the subject in question. Regarding the classification on the power spectra it can be said that the variability between subjects must be severe since a high (close to perfect) specificity can be achieved. The rather low sensitivity leads to the assumption that the signal is likely to be not unique enough or too different between sessions that the identification is not viable. This means that an authentication is possible, though with a high rejection rate on the power spectra in this very scenario. Overall it can be said that an authentication on the basis of brain signals is a possibility for future applications as the specificity is very high despite a rather large rejection rate. Since it is desirable with respect to security aspects to rather need several trials to be successfully authenticated, than to grant access to someone that is not allowed to have access, a real world usage seems to be feasible.

CONCLUSION

Non-stationarity in the power spectra of EEG signals can be modeled with linear regression models and almost be eliminated by using ERD/S instead of the plain power signal. Therefore, using ERD/S could prevent a decline in classification performance with increasing time of the experiment or the need of a recalibration during a session. Inter-subject variability was quantified by classification approaches revealing that differences between subject specific ERPs (power spectra) must be very distinct as an authentication of the correct subject to a corresponding signal was possible reliably. It remains a big issue that needs to be further addressed in terms of BCI development, but it can be turned to an advantageous feature when considering subject authentication and identification as an application. It can be suggested that P300 ERPs work as a biometric measure for identifying subjects, whereas the spectral features of a P300 were less suitable for that cause.

REFERENCES

[1] Shenoy, Pradeep, et al. ”Towards adaptive classification for BCIPart of the 3rd Neuro-IT and Neuroengineering Summer School Tutorial Series.” Journal of neural engineering 3.1 (2006): R13.

[2] Spüler, Martin, Wolfgang Rosenstiel, and Martin Bogdan. ”Adaptive SVM-based classification increases performance of a MEG-based Brain-Computer Interface (BCI).” International Conference on Artificial Neural Networks. Springer Berlin Heidelberg, 2012.

[3] Spüler, Martin, Wolfgang Rosenstiel, and Martin Bogdan. ”Online adaptation of a c-VEP brain-computer interface (BCI) based on error-related potentials and unsupervised learning.” PloS one 7.12 (2012): e51077.

[4] Lotte, Fabien, et al. ”A review of classification algorithms for EEG-based brain–computer interfaces.” Journal of neural engineering 4.2 (2007): R1.

[5] Spüler, Martin, et al. ”EEG-based prediction of cognitive workload induced by arithmetic: a step towards online adaptation in numerical learning.” ZDM 2016: 1-12.

[6] Kindermans, Pieter-Jan, et al. ”Integrating dynamic stopping, transfer learning and language models in an adaptive zero-training ERP speller.” Journal of neural engineering 11(3); 2014: 035005.

[7] Kindermans, Pieter-Jan, et al. ”A P300 BCI for the masses: Prior information enables instant unsupervised spelling.” Advances in Neural Information Processing Systems. 2012.

[8] Wu, Dongrui, Brent J. Lance, and Thomas D. Parsons. ”Collaborative filtering for brain-computer interaction using transfer learning and active class selection.” PloS one 8(2); 2013: e56624.

[9] Armstrong, Blair C., et al. ”Brainprint: Assessing the uniqueness, collectability, and permanence of a novel method for ERP biometrics.” Neurocomputing 166 (2015): 59-67.

[10] Palaniappan, Ramaswamy. ”Two-stage biometric authentication method using thought activity brain waves.” International Journal of Neural Systems 2009, 18(1) : 59-66.

[11] Spüler M, Bensch, M, Kleih S, Rosenstiel W, Bogdan M, Kübler A. Online use of error-related potentials in healthy users and people with severe motor impairment increases performance of a P300-BCI, Clinical Neurophysiology 2012;123(7): 1328-1337.

[12] Burg, John Parker. The relationship between maximum entropy spectra and maximum likelihood spectra. Geophysics 1972; 37(2):375-376.

[13] Chang, Chih-Chung, and Chih-Jen Lin. LIBSVM: a

library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2011; 2(3): 27.

[14] Spüler M, Rosenstiel W, and Bogdan M. Principal component based covariate shift adaption to reduce non-stationarity in a MEG-based brain-computer interface. EURASIP Journal on Advances in Signal Processing (2012); 2012(1): 1-7.

[15] Pfurtscheller G, and Aranibar A. Event-related cortical desynchronization detected by power measurements of scalp EEG. Electroencephalography and clinical neurophysiology 1977; 42(6): 817-826.