Designing spatial filters based on neuroscience theories to improve error-related potential classification

Sandra Rousseau, Christian Jutten, Marco Congedo

2012 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 23–26, 2012, SANTANDER, SPAIN

DESIGNING SPATIAL FILTERS BASED ON NEUROSCIENCE THEORIES TO IMPROVE ERROR-RELATED POTENTIAL CLASSIFICATION

Rousseau Sandra, Jutten Christian and Congedo Marco

Gipsa-lab-DIS, 11 rue des mathématiques Domaine Universitaire 38400 Saint-Martin d'Héres

ABSTRACT

In this paper we present an experiment enabling the occurrence of the error-related potential in high cognitive load conditions. We study the single-trial classification of the error-related potential and show that classification results can be improved using specific spatial filters designed with the aid of neurophysiological theories on the error-related potential.

Index Terms— BCI, single-trial classification, error-related potential, spatial filtering

1. INTRODUCTION

The error-related potential (ErrP) is an event-related potential (ERP) which is generated when a subject commits or observes the commitment of an error. It was first reported in 1991 by Falkenstein et al. [1] and has been since the subject of growing interest. This potential is time-locked to the observation of the error and is mostly characterized by a negative deflection (Ne) [2], followed by a large positivity (Pe) [3]. There exists different kind of ErrPs depending on the agent commiting the error. In BCIs we generally observes the interaction ErrP which occurs when a subject observe an external device commiting an error and the feedback ErrP which is observed when a suject commits an error but becomes aware of it only after an external feedback. Lately several authors have become interested in its integration in BCI systems as a control loop. The integration of the ErrP in BCIs involves two main operations: its single trial detection and the use of this information to on-line modify the system. Since its signal to noise ratio is very low, as any ERP, the ErrP can easily be seen by summing up several trials but is much less detectable on a single-trial basis. The single-trial detection of the ErrP is a crucial point for its integration in BCIs, thus learning its characteristics to design optimal filtering methods is a key point. In this paper we first present an experiment we designed to obtain ErrP data in high cognitive load conditions and study the occurrence of ErrP. Then we propose to apply spatial filtering methods to enhance ErrP signal to noise ratio in order to improve its single-trial detection. We present three different theories on ErrP origin and use them to design different types of filter. These filters are then applied to our data before classification is performed. Filters are then compared in terms of their classification accuracy and the reliability of the corresponding theories is discussed.

2. DATA

2.1. Experiment

The experiment we designed involved a memory game where performance feedback was given after the subject answered questions. No specific reward was given and the experiment was designed so as to induce a high cognitive load. Thus the subject had to focus on other things than just the feedback. The experiment involved two sessions that lasted together approximately half an hour. Each session consisted of six blocks of six trials, for a total of $6\times6\times2 = 72$ trials. Stimuli were presented on a computer screen in front of the subjects and consisted of digits displayed in square boxes. Nine square boxes were arranged in circle on the screen. Each trial started with the display of the score for 3000 ms followed by a fixation cross, which was also displayed for 3000 ms. Then the memorization sequence started, each memorization consisted in a random sequence of two to nine digits appearing sequentially in random positions, with each digit of the sequence randomly assigned to a different box for each sequence. Subjects were instructed to retain positions of all digits. At the end of the sequence the target digit (always contained in the previous sequence) was displayed and subjects had to click on the box where it had appeared. Once the subject had answered, the interface waited for 1500 ms in order to avoid any contamination of ErrP by beta rebound motor phenomena linked to mouse clicking [4], [5]. Then, if the answer was correct, the chosen box background color turned into green ("correct" feedback), otherwise it turned into red ("error" feedback). Subjects were then asked to report if the feedback (error/correct) matched their expectation by a mouse click ("yes"/ "no"). Following the answer a random break of 1000 ms to 1500 ms preceded the beginning of the new trial (see Figure 1).

In order to keep the subjects motivated throughout the experience, the accumulated score was computed at the beginning of each trial. When subjects localized correctly the

978-1-4673-1026-0/12/$31.00

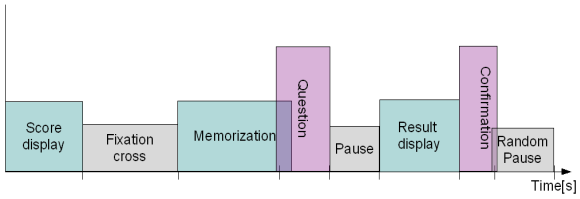


Fig. 1: *Time course of one trial*

target digits their score increased, otherwise, it remained unchanged. The number of digits in the sequence was fixed within blocks (between two and nine) and updated, according to performance, at the beginning of each block. The first block started with four digits for all subjects. The number of digits in the sequences adaptively increased or decreased of one digit at each block, according to subjects performance. The number of digits was adapted with an algorithm tuned to allow about 20% of errors for all subjects, regardless the working memory ability and adapting to fatigue as well as other possible nuisance intervening during the experiment. Between the two sessions the screen was shut down to allow a rest break of 2 - 3 minutes. The use of the adaptative algorithm allowed us to tune the level of the experiment according to each subject in order to have a similar cognitive load for every subject.

2.2. Participants

25 healthy volunteer subjects participated, 14 males and 11 females. Subjects were informed of the procedure before the experiment and filled an information form. All subjects were BCI-naifs at the time of the experiment and none of them reported neurological or psychiatric disorders in the past. Due to the presence of artifacts, four subjects were excluded from analysis. The age of participants ranged from 20 to 30 with a mean (standard deviation) of 24 (2.5). The mean error rate (standard deviation) was equal to 18 (4.6)% of the trials.

2.3. Acquisition and preprocessing

EEG recordings were made from 31 sensors using the extended 10/20 system. Both earlobes, digitally linked, were used as electrical reference. The ground sensor was positioned on the forehead. The impedance of each sensor was kept below $5k\Omega$. The EEG was band-pass filtered in the range 0.1-70 Hz and digitized at 500 Hz using the Mitsar 202 DC EEG acquisition system. Data were bandpass-filtered between 1-40 Hz using an order 4 Butterworth filter with linear phase response. Eye blinks were extracted using ICA (independent component analysis ([6])). One EOG source (or more when necessary) was suppressed for each subject. It was manually selected using both the temporal shape of the source and its topography.

2.4. Observation of the ErrP

In Figure 2 we plot the event-related potential averaged over subjects for correct trials and for error trials for one second post-stimulus (observation of the error). In this figure we see that for error trials the ERP is characterized by a sharp negativity followed by a small positivity which is consistent with previous reports ([7]; [8]). For correct trials, a negativity is also observed but with a much lower intensity.

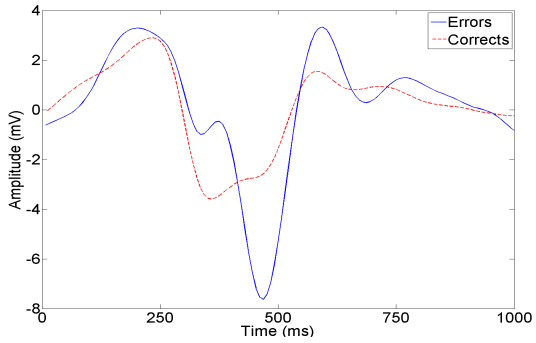


Fig. 2: *Mean potential averaged over 21 subjects at electrode FCz.*
Blue thick line corresponds to error trials and red dashed line corresponds to correct trials

3. METHOD

3.1. Theories

As we said before, after an error trial one can observe the occurrence of the error-related potential which is mostly characterized by a sharp negativity (ERN). However, another event-related potential also appears after correct trials called the correct-related negativity (CRN) [9],[10],[11]. This potential is also characterized by a negativity occurring at the same time as the ERN (but less intense) and localized at the same electrodes. There exists different theories concerning the link between these two potentials and their neurophysiological origins. Here we present three possible theories:

H1: The ERN and the CRN are two completely different phenomena which occur only in one of the two conditions (error or correct). This means that these two potentials are generated by different sources and could be separated by specific spatial filtering. These are two uncorrelated activities.

H2: The ERN and the CRN are generated by the same source which is simply modulated by the value of the outcome. Thus, the differentiation of the ERN and the CRN can only be done by observing the intensity of this source. This theory is supported by many studies

which localized, using blind source separation techniques or fMRI, the source responsible for the ERN and the one for the CRN at the exact same location [12],[11].

H3: The ERN is the sum of two phenomena, one which is linked to the observation of a result or of a conflictual response (and which is common to the CRN), another one which is only linked to the occurrence of an error [13]. Thus we have ERN = CRN + another potential. This means that these two potentials share common neural circuits but that in addition to these, other independent sources are activated only for error trials.

Using these different theories we can build different spatial filters. Here we will develop three different spatial filters corresponding to each of these theories. Theories will then be tested and compared using the classification rate we obtain using each filter.

3.2. Filtering methods

In order to optimize our classification results we have tried to spatially filter our data so as to improve the SNR. Filtering was done using xDAWN algorithm [14], [15]. In this algorithm we consider that the signal is the sum of one target evoked potential plus other possible superimposed evoked potentials (but non-target) and noise. Thus the signal can be written as:

$$X = D_1 A_1 + D_2 A_2 + N \tag{1}$$

where X: represents the signal, D_i: is a Toeplitz matrix whose first column entries are set to zero except for those that correspond to the stimuli of type i. A_i: represents the responses synchronized with the stimuli of type i, N: represents the noise. Evoked responses A_i are estimated as:

$$\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = (D^T D)^{-1} D^T X \tag{2}$$

with $D = [D_1, D_2]$. The aim of xDAWN is then to find the filter U which maximizes the signal to signal plus noise ratio (SSNR):

$$U = argmax \frac{Tr(U^T \sum_1 U)}{Tr(U^T \sum_X U)} \tag{3}$$

with $\sum_1 = (D_1 A_1)^T (D_1 A_1)$ and $\sum_X = X^T X$.

Here D_1 and D_2 will be defined differently according to the theories we use.

H1: $ERN \neq CRN$. We have two types of stimuli: "errors" and "corrects". D1 will be constructed using the stimuli corresponding to erroneous responses and D2 will be constructed using the stimuli of correct responses. A_1 will be an estimation of the mean potential for error trials (ERN) and A_2 will be an estimation of the mean potential for correct trials (CRN).

H2: $ERN = CRN$. We have only one type of stimuli which corresponds to the reaction to an outcome of performance. D1 will be constructed using the stimuli corresponding to correct and erroneous responses. There will be no D2. A_1 will be an estimation of the mean potential for both error and correct trials.

H3: $ERN = CRN + P_1$. We have two types of stimuli "errors" and "reaction to an outcome of performance". D1 will be constructed using the stimuli corresponding to erroneous responses. D2 will be constructed using the stimuli corresponding to correct and erroneous responses. A_2 will be an estimation of the mean potential for both error and correct trials. A_1 will be an estimation of the remaining mean potential after A_2 has been subtracted from error trials.

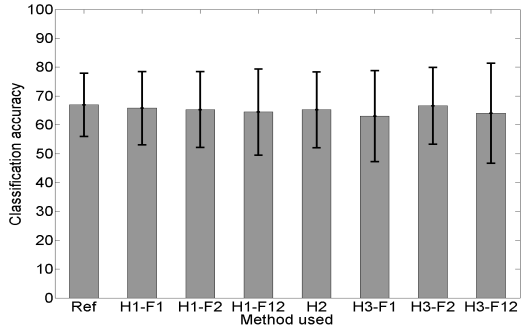
Moreover, here we want to classify both correct and error trials, both might be considered as target trials. Thus for all these theories two types of filters were designed, one to improve the SSNR of $D_1 A_1$ (F_1) and one to improve the SSNR of $D_2 A_2$ (F_2). This means that in equation (3) we will use $\sum_1$ to design F_1 and $\sum_2$ to design F_2.

3.3. Preprocessing and classification

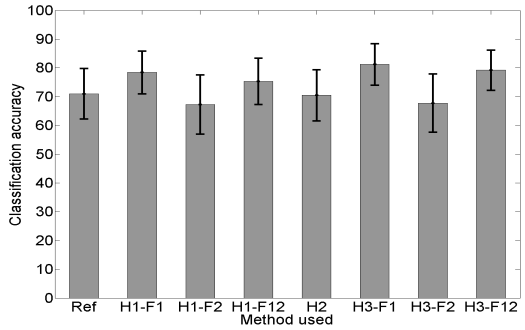
Data were first band-pass filtered between 1-10 Hz and then spatially filtered using the previously described methods. Data were then classified using Bayesian LDA classifier [16]. Since xDAWN algorithm returns filters classified in descending order of performance we used the first two components as features for our classifier. In order to avoid over-learning data were subsampled at 32 Hz. Most studies on ErrP single-trial classification did not perform any spatial filtering but simply selected FCz and Cz signals (based on prior physiological knowledge)[17], [18]. Thus, classification was also performed on raw data (from electrodes FCz and Cz) in order to compare classification results obtained with and without spatial filtering. Classification was performed on 21 subjects. For each subject a leave-one method was used, which means that spatial filters and classifier were learned on the whole data except one and then tested on the remaining data. Classification was performed using F_1 filters, F_2 filters and both.

4. RESULTS

In figure 3 we plot the mean (and standard deviation) results of classification over the 21 subjects for error trials and correct trials using the different methods. First we can see that classification accuracy is higher for correct trials than for error trials, this is also mostly the case in other studies on the ErrP single-trial classification [19],[20],[21]. Moreover we get, for the reference method, an average classification accuracy of 67% for error trials and 70% for correct trials, which



(a) Error trials



(b) Correct trials

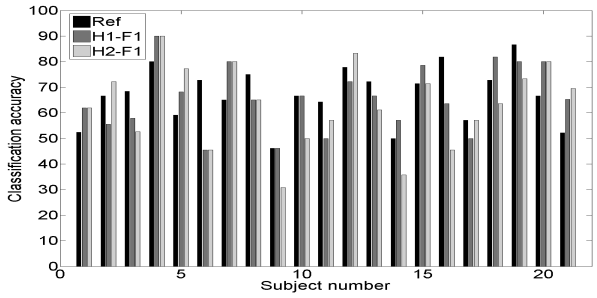
Fig. 3: *Classification results averaged over the 21 subjects. (a) For error trials. (b) For correct trials. Blacks bars correspond to the standard deviation. The height columns correspond to the different classification method used, the first one corresponds to the reference (no spatial filtering), then we have results for the first theory (H1) using filters for A_1 (F1), A_2 (F2) and both (F12), H2 corresponds to results for the second theory (since there is only target signal we have only one type of filter). The three last columns correspond to results for theory 3 (H3) using filters for A_1 (F1), A_2 (F2) and both (F12)*

is a little bit lower than most results obtained in other studies [22],[23],[24], but still very good in regard of the low number of trials available (72 trials with only 20% of errors). We can see that for error trials all filters seem to perform equivalently and provide results similar as those obtained with the reference method, i.e. without spatial filtering. Thus it seems that for detecting errors no spatial filtering is needed or at least that xDAWN filtering method is not optimal, still it does not deteriorate results. For the classification of correct trials it is clear that method H1 and H3 are the most performant ones. Moreover we can also see that the best classfication results are obtained using only the first filter (F1), i.e. when using the ERN as the target potential. H3-F1 slightly outperforms H1-F1 with an average classification accuracy of 81% for H3-F1 against 78% for H1-F1. Both methods clearly outperform the reference method (70% of accuracy). A one-way ANOVA was performed on the classification results to assess the relevance of the observed effect ($F = 8.21$, $p \leq 0.01$). Post-hoc multiple comparison tests showed that both H3-F1 and H1-F1 lead to significantly better results than the reference method. However no significant difference was found between H3-F1 and H1-F1. On the contrary one can see that H2 method gives the same kind of results (69% for H2) as the reference method. In figure 4 we plot the results for each subject for correct end error trials for method H1-F1, H3-F1 and reference. We can see that results for error trials are highly different from one subject to another and that depending on the subject each method can highly outperform the other ones. For correct trials the reference method leads to better results than H1-F1 for only 6 subjects out of 21 ones and than H3-F1 for only 3 subjects out of 21. Thus it seems that H3-F1 is a robust method. Moreover it has to be noted that when reference method outperforms one of the two other methods the difference is always very low (with a maximum difference of 5% only) while when H3-F1 or H1-F1 outperforms the reference method the difference might go up to 24%. When comparing H3-F1 and H1-F1 there does not seem to be any clear difference between both, indeed H1-F1 outperforms H3-F1 for 8 subjects while H3-F1 outperforms H1-F1 for nine subjects, thus there does not seem to be a method that is more reliable than the other.

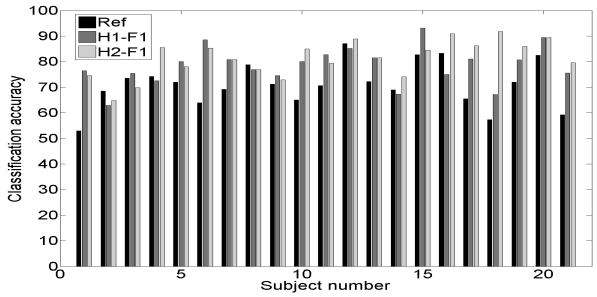
In figure 5 we plot the topographic maps for four different subjects corresponding to the first filter component obtained using the different strategies. We can see that H3 gives more focused and more stable topographic maps. H2 gives highly variable maps which are not always consistent with the litterature on ErrP localization. H1 gives better results than H2 but maps are more spreaded and less stable than those obtained with H3 except for the fourth subject. These observations are consistent with the classification results.

5. DISCUSSION

In this paper we presented a high cognitive load experiment which allowed the generation of ErrP. We obtained satisfying classification results in regard of the very small number of trials available (72 trials) with no particular spatial filtering. In a second time we have proposed to develop spatial filters in order to improve this classification accuracy. The xDAWN method allows us to estimate spatial filters optimized for a given target stimulus on which other undesired stimuli might be superimposed. We developed three different spatial filters based on three different existing theories on the nature of the error-related potential and its link to the correct-related potential. These three theories lead to different estimations of the target stimulus and thus to different spatial filters. First, results showed that classification accuracy could be greatly improved using well defined spatial filters, at least for correct trials. This is very encouraging for improving

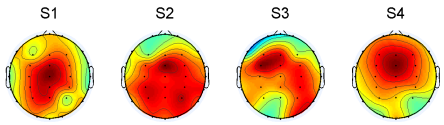


(a) Error trials

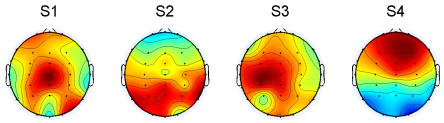


(b) Correct trials

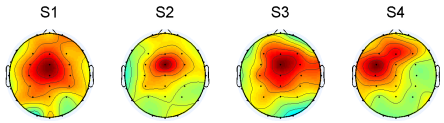
Fig. 4: *Classification results for the 21 subjects. (a) For error trials. (b) For correct trials. Black bars correspond to the reference method and dark gray bars to the H1-F1 method and light gray bars to the H3-F1 method*



(a) H1-F1



(b) H2-F1



(c) H3-F1

Fig. 5: *Topographic maps for four different subjects using the different theories. (a) Filtering is done using H1, we plot the filter obtained for the errors. (b) Filtering is done using H2. (c)Filtering is done using H3, we plot the filter obtained for the errors.*

ErrP single-trial detection since more specific filters could be constructed. Two methods clearly outperformed the reference method (which used no spatial filters but only the selection of relevant electrodes), improving classification results for the ErrP up to 10%. One interesting point is that in neurophysiological studies the most supported theory is the one corresponding to H2 i.e. the one corresponding to the fact that the ERN and the CRN reflect the same phenomenon while in our study it is the one which lead to the poorer results. H3 and H1 lead to similar results with H3 slightly outperforming H1, this can be explained by the fact that, in the case where H3 would be the right theory, H1 might still give us a slightly good estimate of A_1. Indeed the difference between potentials estimated in H1 and H3 is just that in H3 we first subtract the average global potential (i.e. corresponding to both correct and error trials) before estimating A_1. This average global potential can be seen as noise. Thus, H1 will simply lead to a noisy estimation of A_1 but, if this noise is not too high we will get similar results. This paper allowed us to enlight two important points, first that ErrP classification could be improved by spatial filtering and that neurophysiological theories could be used to design these filters and that, in return, these filters and their corresponding classification accuracies could bring information on the relevance of these theories. Further studies should be made in this sense by using other neurophysiological knowledges to develop even more optimized filters.

6. REFERENCES

[1] M. Falkenstein, J Hohnsbein, J Hoormann, and L Blanke, "Effects of crossmodal divided attention on late ERP components.II. Error processing in choice reaction tasks," *Electroencephalogr. Clin. Neurophysiol.*, vol. 78, pp. 447–455, 1991.

[2] A. Gentsch, P. Ullsperger, and M. Ullsperger, "Dissociable medial frontal negativities from a common monitoring system for self-and externally caused failure of goal achievement," *Neuroimage*, vol. 47, no. 4, pp. 2023–2030, 2009.

[3] M. Steinhauser and A. Kiesel, "Performance monitoring and the causal attribution of errors," *Cognitive, Affective, & Behavioral Neuroscience*, pp. 1–12, 2011.

[4] R. Salmelin and R. Hari, "Spatiotemporal characteristics of sensorimotor neuromagnetic rhythms related to thumb movement," *Neuroscience*, vol. 60, no. 2, pp. 537–550, 1994.

[5] G. Pfurtscheller, "Central beta rhythm during sensorimotor activities in man," *Electroencephalography and clinical Neurophysiology*, vol. 51, no. 3, pp. 253–264, 1981.

[6] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*, Academic Press, 2010.

[7] N. Yeung, C.B. Holroyd, and J.D. Cohen, "ERP correlates of feedback and reward processing in the presence and absence of response choice," *Cerebral Cortex*, vol. 15, no. 5, pp. 535, 2005.

[8] G. Hajcak, J.S. Moser, C.B. Holroyd, and R.F. Simons, "It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks," *Psychophysiology*, vol. 44, no. 6, pp. 905–912, 2007.

[9] M. Falkenstein, J. Hoormann, S. Christ, and J. Hohnsbein, "ERP components on reaction errors and their functional significance: A tutorial," *Biological Psychology*, vol. 51, no. 2-3, pp. 87–107, 2000.

[10] P. Luu, T. Flaisch, and D.M. Tucker, "Medial frontal cortex in action monitoring," *Journal of Neuroscience*, vol. 20, no. 1, pp. 464, 2000.

[11] F. Vidal, B. Burle, M. Bonnet, J. Grapperon, and T. Hasbroucq, "Error negativity on correct trials: A reexamination of available data," *Biological Psychology*, vol. 64, no. 3, pp. 265–282, 2003.

[12] C. Roger, C.G. Bénar, F. Vidal, T. Hasbroucq, and B. Burle, "Rostral Cingulate Zone and correct response monitoring: ICA and source localization evidences for the unicity of correct-and error-negativities," *NeuroImage*, vol. 51, no. 1, pp. 391–403, 2010.

[13] M. Ullsperger and D.Y. Von Cramon, "Subprocesses of performance monitoring: A dissociation of error processing and response competition revealed by event-related fMRI and ERPs," *Neuroimage*, vol. 14, no. 6, pp. 1387–1401, 2001.

[14] B. Rivet, A. Souloumiac, G. Gibert, and V. Attina, "P300 speller brain-computer interface: Enhancement of P300 evoked potential by spatial filters," in *Proceedings of the 16th European Signal Processing Conference (EUSIPCO-2008), EURASIP*, Lausanne, Switzerland, August 2008.

[15] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "xDAWN algorithm to enhance evoked potentials: Application to brain computer interface," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 8, pp. 2035–2043, Jan 2009.

[16] D.J.C. MacKay, "Bayesian interpolation," *Neural computation*, vol. 4, no. 3, pp. 415–447, 1992.

[17] X. Artusi, I. K. Niazi, M.-F. Lucas, and D. Farina, "Performance of a Simulated Adaptive BCI Based on Experimental Classification of Movement-Related and Error Potentials," *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, vol. PP, no. 99, pp. 1, 2011.

[18] P.W. Ferrez and J.d.R. Millán, "EEG-based brain-computer interaction: Improved accuracy by automatic single-trial error detection," in *Proc. NIPS 20*, 2007.

[19] P.W. Ferrez and J.R. Millán, "Simultaneous real-time detection of motor imagery and error-related potentials for improved BCI accuracy," in *Proc 4th Intl. Brain-Computer Interface Workshop and Training Course*, Graz, Austria, September 2008.

[20] B. Dal Seno, M. Matteucci, and L. Mainardi, "Online detection of P300 and error potentials in a BCI speller," *Computational intelligence and neuroscience*, vol. 2010, pp. 1–1, 2010.

[21] M. K. Goel, R. Chavarriaga Lozano, and J. del R. Millán, "Cortical Current Density vs. surface EEG for Event-Related Potential-based Brain-Computer Interface," in *5th International IEEE EMBS Conference on Neural Engineering*, 2011.

[22] R. Chavarriaga, P. W. Ferrez, and J. del R. Millán, "To err is human: Learning from error potentials in brain-computer interfaces," in *Int Conf Cognitive Neurodynamics*, R.Wang, Fanji Gu, and Enhua Shen, Eds., Shanghai, China, 2007, pp. 777–782.

[23] P.W. Ferrez and J.R. Millán, "You are wrong!—automatic detection of interaction errors from brain waves," in *Proceedings of the 19th International Joint Conference on Artificial Intelligence, 2005*, 2005.

[24] I. Iturrate, L. Montesano, and J. Minguez, "Single trial recognition of error-related potentials during observation of robot operation," in *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'10)*, 2010.