35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Using frequency-domain features for the generalization of EEG error-related potentials among different tasks

Jason Omedes, Iñaki Iturrate, Luis Montesano, Javier Minguez

***Abstract*— EEG brain-computer interfaces (BCI) require a calibration phase prior to the on-line control of the device, which is a difficulty for the practical development of this technology as it is user-, session- and task-specific. The large body of research in BCIs based on event-related potentials (ERP) use temporal features, which have demonstrated to be stable for each user along time, but do not generalize well among tasks different from the calibration task. This paper explores the use of low frecuency features to improve the generalization capabilities of the BCIs using error-potentials. The results show that there exists a stable pattern in the frequency domain that allows a classifier to generalize among the tasks. Furthermore, the study also shows that it is possible to combine temporal and frequency features to obtain the best of both domains.**

I. Introduction

EEG-based brain-computer interfaces (BCIs) build a communication channel between the user and a device based on brain activity, with a wide range of non-clinical and clinical applications [1]. In all BCIs there is a calibration phase to learn a mapping from EEG activity to the control space that operates the device. This calibration has to be carried out for each subject to deal with the large inter-user EEG variability [1]. In addition, a common procedure is also to recalibrate the BCI for each new task and even for the same task between sessions, to deal with the EEG variability [2], [3]. This is a large shortcoming of current BCI technology as the calibration is a tedious and boring process (that may take between 30 and 45 minutes for error potentials [12]).

Calibration is dependent on the EEG signal used for the BCI. On one hand, for self-generated brain activity (such as the motor imagery of body limbs [4]) there is a body of work to deal with EEG non-stationarities, either to reduce the calibration time [3] or to minimize the impact in the decoding performance [2]. On the other hand, BCIs that rely on external cues such as those using event-related potentials (ERPs) [5], have a better generalization among sessions but do not generalize between different tasks. This is because the amplitude and latency of their components are affected by factors such as spatial attention [6]; stimuli contrast [5]; the probability of appearance of the expected stimulus [5]; the inter-stimulus interval [7]; user-dependent factors such as age and cognitive capabilities [8]; and other cognitive aspects such as the stimulus evaluation time (i.e., the amount of time required to perceive and categorize a stimulus) [5], [9].

Jason Omedes, Iñaki Iturrate, Luis Montesano and Javier Minguez are with the I3A, DIIS, and Univ. Zaragoza, Spain. Javier Minguez is also with Bit&Brain Technologies SL, Spain. eMail: {jomedes, iturrate, montesano, jminguez}@unizar.es. This work has been supported by Spanish projects DPI2011-25892, and DGA-FSE (grupo T04).

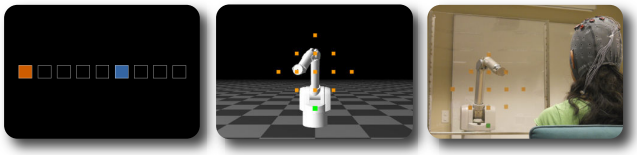

Fig. 1. From left to right, experiments 1 to 3.

The on-line detection of ERPs relies on the fact that these signals are phase-locked to a trigger event [5], [10]. Thus, successful single-trial detection has been carried out mainly in the temporal domain [11], [12], despite of the single-trial temporal variability. A recent study showed that different tasks of these BCIs induce a phase change (i.e. different latencies) in the components of error potentials [13]. As a result, the use of temporal features of the ERP (e.g. amplitudes) significantly degrades ERP detection rate when the tasks in the calibration and execution phase are different. Nonetheless, it is possible to estimate the latency variations between different ERPs and use it to reduce the calibration time of a new task.

Another possible alternative to avoid this degradation of the detection rate could be the use of frequency features, as these features are insensitive to phase shifts (and thus latency) variations. Furthermore, in principle no information from a new task is needed as long as the ERP amplitudes and frequency components remain similar. This paper explores the usage of low frequency components of error potentials as a way of dealing with changes induced by different tasks in the temporal domain. The results show that although the detection accuracy within a single task is better in the temporal domain, there exists a stable pattern in the frequency domain that allows a classifier to generalize among the tasks, and thus BCIs based on these features generalize better. In practice, the study also shows that it is possible to combine temporal and frequency features to obtain the best of both domains.

II. Methods

A. Data Recording

The EEG was recorded using a gTec system with 16 active electrodes (Fz, FC3, FC1, FCz, FC2, FC4, C3, C1, Cz, C2, C4, CP3, CP1, CPz, CP2, and CP4 according to the 10/10 international system), with the reference and the ground placed at the left earlobe and AFz respectively. The EEG was sampled at 256 Hz and power-line notch filtered at 50 Hz.

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

B. Experimental Setup

Six volunteer participants (five males and one female, mean age 27) participated in the study. Participants were instructed to observe movements performed by a device and evaluate them as correct when they were towards a target position and as incorrect otherwise, evoking non-error and error potentials. The participants were asked to restrict eye movements and blinks to specific resting periods.

Three experimental conditions with progressively higher cognitive workload were designed (see Figure 1). In all the experiments, the device performed correct/incorrect movements until reaching a specific goal position. Time between actions was random and within the range $[1.7, 4.0]$ s, with a 20% probability of performing an erroneous movement. The first experiment consisted of a squared cursor that could execute two actions (move one position left or right) in a 1D grid with 9 different equally-distributed positions. The second experiment displayed a simulated robotic arm that could perform four actions (move one position left, right, up or down) in a 2D grid with 13 equally-distributed positions. The third experiment followed the configuration of the second experiment, but using a real robotic arm. For more information about the experiments, see [13]. Each experiment lasted ~ 2.5 hours. They were always executed in the same order as presented above, with a time between sessions of 17.58 ± 10.09 days. For each subject and experiment, approximately 800 trials (around 160 and 640 error and non-error potentials) were acquired.

C. Electrophysiology Analysis

For the time analysis, the time-locked averaged potentials were computed for the error, non-error and difference (error minus non-error averages) conditions at channel FCz. For the frequency analysis, the power spectral density (PSD) of each one-second trial was first computed using the Welch's method with a Hamming window and a window overlap of 50%. Then, the error, non-error and difference average PSDs were computed at channel FCz. The r^2 discriminability test [14] between error and non-error conditions was computed for each channel and time instant (time analysis), and each channel and frequency component (frequency analysis).

D. Feature Extraction

Two different sets of features were extracted.

1) Temporal Features: The EEG was common-average referenced (CAR) and $[1, 10]$ Hz band-pass filtered. Temporal features were the EEG voltages of each trial of eight fronto-central channels (Fz, FC1, FCz, FC2, C1, Cz, C2, and CPz) [12] within a time window of $[200, 800]$ ms (being 0 the stimulus onset) subsampled at 64 Hz, leading to a vector of 312 features. Finally, the features were normalized within the range $[0, 1]$.

2) Frequency Features: The EEG was common-average referenced (CAR). For each of the channels used in the temporal features, the PSD was computed on one second of EEG after the stimulus onset as explained in subsection II-C. The frequency features were the power values of each channel from the theta band ($[4, 8]$ Hz) $\pm$ 1 Hz (as previous studies suggested that the error potentials are generated within this band [10]), which led to a vector of 200 features. Finally, the features were normalized within the range $[0, 1]$.

E. Methods for Single-Trial Classification

Previous studies showed that the usage of temporal features provoke a degradation of performance when training with one experiment and testing with another one (i.e. generalization) [13]. The objective of the present classification study was to analyze whether the frequency features or the combination of both (temporal and frequency) are robust enough to generalize among different tasks (experiments).

Single-trial classification was carried out using a support vector machine (SVM) with a radial basis function (RBF) kernel, as this classifier presents high accuracies when classifying ERPs [15] and error potentials in particular [12]. One important drawback of SVM is its sensitivity to imbalanced datasets. To avoid this drawback, the minority class (i.e. the error class) was oversampled by random replication to match the number of trials of the majority class (i.e. the non-error class) [16].

To study the generalization capabilities of the different feature sets, each task data was divided into a training and a test set composed by 50% of the data each. The classifier was evaluated in two different conditions. First, the baseline accuracy was obtained by using the training and test sets of the same experiment E_j (denoted E_jE_j). Second, the classifier was trained using the train set of an experiment E_i and tested on the test set of another experiment E_j. The train-test combinations considered in the study were E_1E_2, E_1E_3, and E_2E_3, following the combinations studied in [13].

III. Results

A. Results of the Electrophysiology analysis

Fig. 2 (first row) depicts the error, non-error and difference grand averages, for the three experiments. The three difference grand averages of the error potentials have an early negativity and two broader positive and negative components, in agreement with other studies [11], [12]. However, in line with previous works, the latencies of these peaks varied among the three experiments [13] (see figure 2, up-right-most plot). For instance, the latency of the broader negative peak was of 426, 492 and 535 ms for experiments 1 to 3. This variation in latency is also visible with the r^2 metric (Fig. 2 second row). Notice how the r^2 patterns of fronto-central channels present a time shift among experiments.

Regarding the frequency analysis, Fig. 2 (third row) depicts the error, non-error and difference PSD averages for the channel FCz for the three experiments. The difference averages were similar in the theta band for the three experiments (see Figure 2 third row, fourth column). This supports the fact that the main variation of the signals was due to latency differences, but not to amplitude differences (as described in [13]). The r^2 discriminability patterns were in the theta band as suggested in [10]. Notice that the r^2 values were progressively higher among experimental conditions. Despite

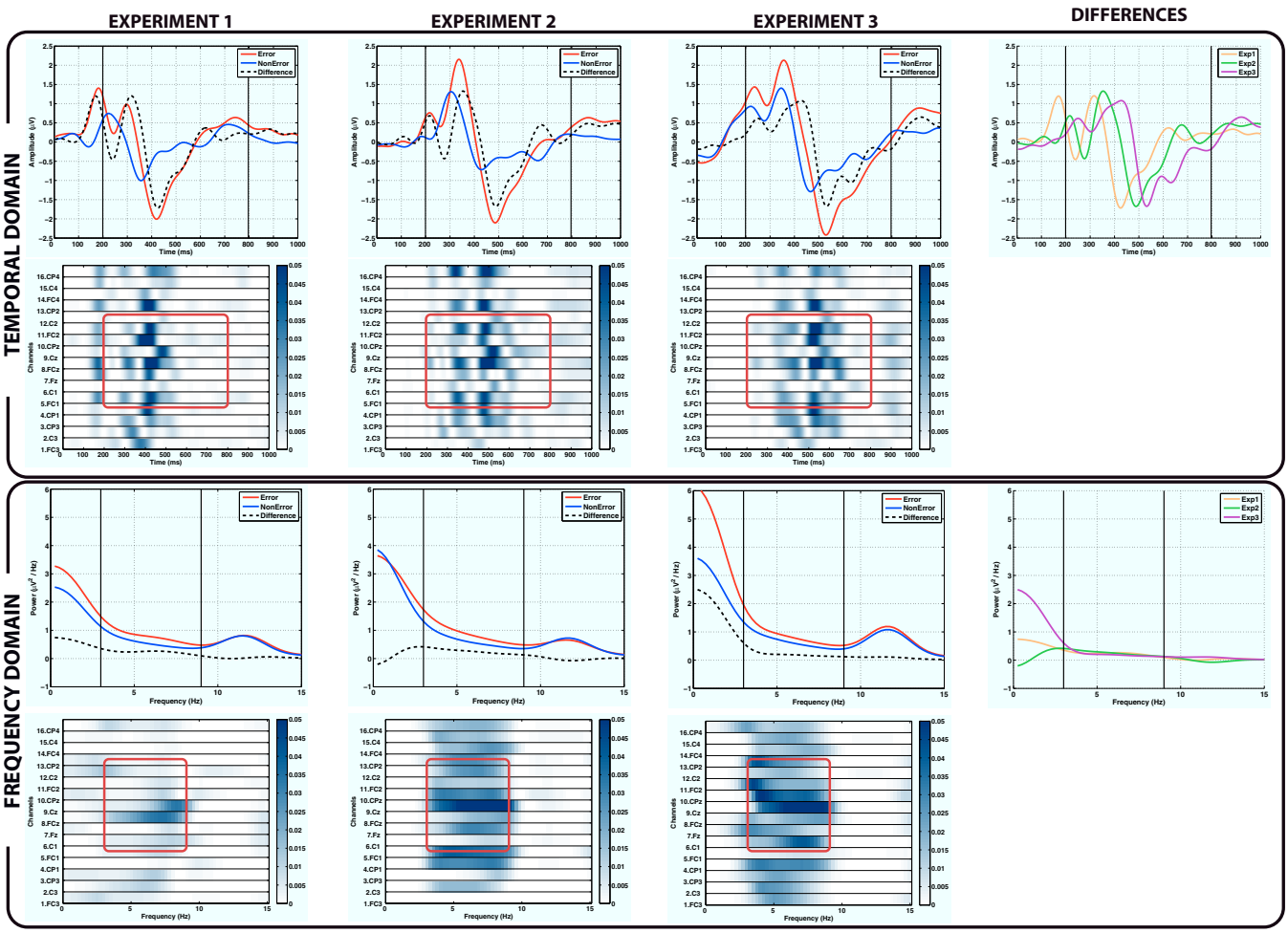

Fig. 2. Electrophysiology results for experiments 1 to 3. First row shows the error, non-error and difference grand averages for channel FCz, and the last column the difference average compared for the three experiments. Second row shows the r^2 test of the temporal signals (x-axis: time, and y-axis: channels). Third and fourth row show the PSD averages (for channel FCz) and the r^2 test of the frequency signals. For each r^2 plot, the squared zone represents the window used for the extracted features.

there is not a clear reason of this increase in the r^2, it could be due to: a user habituation to the protocols (since the three experiments were always executed in the same order from 1 to 3); or a higher cognitive workload that generated stronger error components with greater r^2 values. This increase in separability could hinder the generalization from a more complex experiment to a simpler one (E_iE_j with $i > j$), but not during the opposite generalization (E_iE_j with $i < j$).

B. Classification results

Figure 3 depicts the baseline accuracies of E_jE_j, and the generalization accuracies of E_iE_j for the temporal and frequency feature sets, and for the concatenation of both sets (c.f. subsection II-D), averaged for all subjects.

Regarding the temporal features, the baseline of each experiment had high accuracies, being on average 78.78%, 77.54% and 79.04% for experiment 1 to 3. **However, when generalizing the classifier to another experiment, the use of these results in an accuracy degradation,** mainly due to the latency variations observed in the electrophysiology analysis. In fact, the mean accuracy dropped a 21.09%, 24.36% and 10.21% for the E_1E_2, E_1E_3 and E_2E_3 cases. On the other hand, the use of frequency features resulted on lower baseline accuracies than the temporal ones: 67.29%, 71.33% and 69.67% for experiments 1 to 3. However, the accuracy drop was substantially lower when generalizing the classifier: 3.91%, 4.52%, and 3.44% for E_1E_2, E_1E_3 and E_2E_3. For the baseline classifiers that use the temporal and frequency features, the accuracies presented very similar results to those obtained using the temporal features: 76.17%, 79.31%, and 77.64% for experiments 1 to 3. More interestingly, the generalization classifiers had accuracy drops of 13.11%, 16.23% and 6.35%; but the absolute accuracies were very similar to those obtained with the frequency features: 66.20%, 61.41%, and 71.32%, for E_1E_2, E_1E_3 and E_2E_3. Thus, the use of both set of features at the same time allowed to have the best of time and frequency domains.

These results confirmed that the temporal features had poor task-generalization capabilities due to the latency variations. However, the frequency features generalize better comparing the baseline and the generalization accuracies, suggesting that these features remained similar among ex-

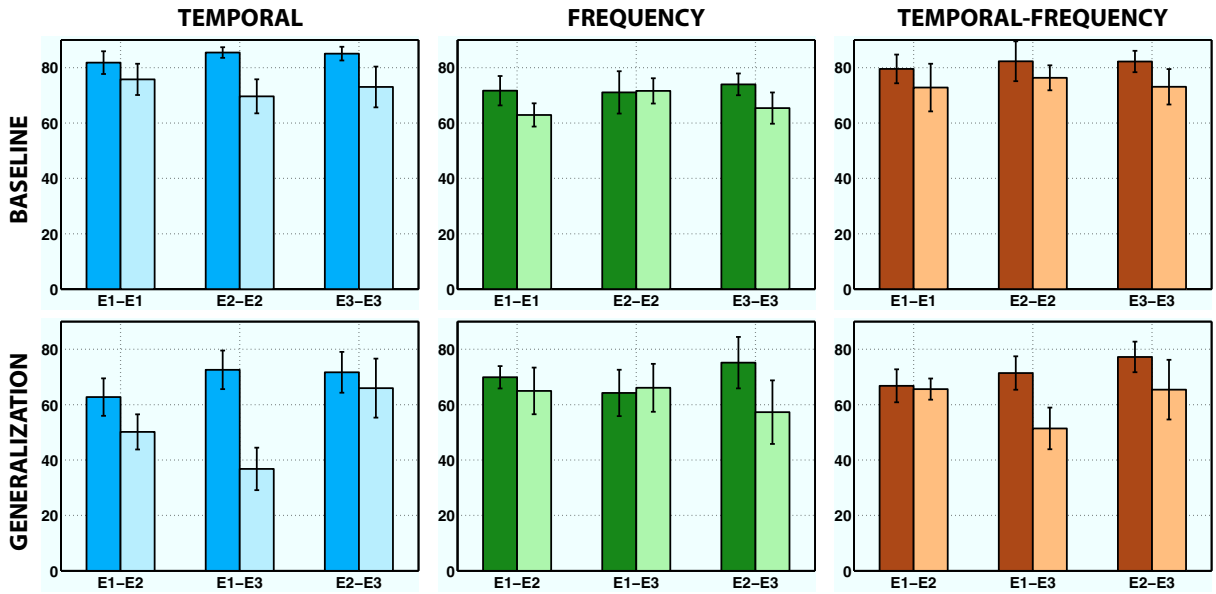

Fig. 3. (Top) Baseline accuracies ± SEM (%) when training and testing with experiment j (denoted E_jE_j) (Bottom) Generalization accuracies ± SEM (%) when training with experiment i and testing with experiment j (denoted E_iE_j). Dark and light colors represent the non-error and error accuracies. Left, middle and right plots show the results when using the temporal, frequency, and combined set of features respectively. Notice that the baseline E_jE_j should be compared to the generalization E_iE_j.

periments.

IV. Conclusions and Future Work

An important issue in current BCI technology is to minimize the calibration time as it is one of the major difficulties especially in the context of patients. For BCIs based on event-related potentials, re-calibration is mainly due to a time shifts present on the potential of interest for each different task. This paper builds on these results showing the presence of these latency changes, and how they affect the temporal features (EEG amplitudes) during the generalization among different tasks (provoking large drops in the accuracies). In addition to this, the paper showed how classifiers based on low-frequency EEG features have better generalization properties among different tasks (completely avoiding the re-calibration process) than those based on temporal features. Furthermore, the combination of features of both domains allows to obtain classifiers with performances similar to the temporal alone on one task, and similar to the frequency alone in generalization (i.e the best properties of both domains). As future work, the authors are studying the use of other frequency features such as wavelets to determine their generalization adequacy.

References

[1] J.d.R. Millán et al., "Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges," *Front Neurosci*, vol. 4, 2010.

[2] C. Vidaurre, M. Kawanabe, P. von Bünau, B. Blankertz, and K.R. Müller, "Toward Unsupervised Adaptation of LDA for Brain-Computer Interfaces," *IEEE Trans Biomed Eng*, vol. 58, no. 3, pp. 587 –597, 2011.

[3] C. Vidaurre, C. Sannelli, K.R. Müller, and B. Blankertz, "Co-adaptive calibration to improve BCI efficiency," *J Neural Eng*, vol. 8, no. 2, pp. 025009, Apr. 2011.

[4] G. Pfurtscheller, D. Flotzinger, and J. Kalcher, "Brain-computer interface: A new communication device for handicapped persons," *J Microcomput Appl*, vol. 16, pp. 293–299, July 1993.

[5] S.J. Luck, *An introduction to the event-related potential technique*, The MIT Press, 2005.

[6] L. Li, D. Yao, and G. Yin, "Spatio-temporal dynamics of visual selective attention identified by a common spatial pattern decomposition method," *Brain Res*, vol. 1282, pp. 84–94, 2009.

[7] E. Sellers, D. Krusienski, D. McFarland, T. Vaughan, and J. Wolpaw, "A P300 event-related potential brain-computer interface (BCI): The effects of matrix size and inter stimulus interval on performance," *Biol Psychol*, vol. 73, no. 3, pp. 242–52, Oct. 2006.

[8] J. Polich, "On the relationship between EEG and P300: Individual differences, aging, and ultradian rhythms," *Int J Psychophysiol*, vol. 26, no. 1-3, pp. 299–317, 1997.

[9] M. Kutas, G. McCarthy, and E. Donchin, "Augmenting mental chronometry: The P300 as a measure of stimulus evaluation time," *Science*, vol. 197, no. 4305, pp. 792, 1977.

[10] J.F. Cavanagh, M.X. Cohen, and J.J.B. Allen, "Prelude to and resolution of an error: EEG phase synchrony reveals cognitive control dynamics during action monitoring," *J Neurosci*, vol. 29, no. 1, pp. 98–105, Jan. 2009.

[11] R. Chavarriaga and J.d.R. Millán, "Learning from EEG error-related potentials in noninvasive brain-computer interfaces," *IEEE Trans Neural Syst Rehabil Eng*, vol. 18, no. 4, pp. 381–388, 2010.

[12] I. Iturrate, L. Montesano, and J. Minguez, "Single trial recognition of error-related potentials during observation of robot operation," in *Proc of the Annual Int Conf of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2010.

[13] I. Iturrate, R. Chavarriaga, L. Montesano, J. Minguez, and J.d.R. Millán, "Latency correction of error potentials between different experiments reduces calibration time for single-trial classification," in *Proc of the Annual Int Conf of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2012.

[14] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, and T.M.Vaughan, "Brain-computer interfaces for communication and control," *Clin Neurophysiol*, vol. 113, no. 6, pp. 767–91, June 2002.

[15] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces," *J Neural Eng*, vol. 4, no. 2, pp. R1–R13, June 2007.

[16] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Proceedings of the 15th European Conference on Machine Learning (ECML)*, 2004, pp. 39–50.