Clustering of EEG Occipital Signals using K-means

Víctor Asanza*, Kerly Ochoa*, Christian Sacarelo*, Carlos Salazar*, Francis Loayza*, Carmen Vaca* and Enrique Peláez*
*Escuela Superior Politécnica del Litoral, ESPOL
Facultad de Ingeniería en Electricidad y Computación
Centro de Tecnología de información
Campus Gustavo Galindo Km 30.5 Vía Perimetral, P.O. Box 09-01-5863
Guayaquil, Ecuador
Emails: {vasanza, kermaoch, csacarel, csalazar, floayza, epelaez}@espol.edu.ec, cvaca@fiec.espol.edu.ec

***Abstract*—Recent studies show that it is feasible to use electrical signals from Electro-encephalography (EEG) to control devices or prostheses, these signals are provided by the body and can be measured on the scalp to determine the intent of the person when it is observing a visual stimulus frequency range detectable by the human eye. This group of signals are very susceptible to noise due to voltage levels that are able to acquire. Therefore, in this work we propose a statistical analysis of the distribution of normal EEG signals in order to determine the need of a pre-processing to remove noise components from electrical grids or other possible sources. This preprocessing includes the design and use of a filter that will eliminate any signal component that is not in the operating frequency range of the EEG occipital area of the brain. Finally, we will proceed to use the k-means algorithm to cluster with signals according to their frequency and temporal characteristics.**

***Index Terms*—Electro-encephalography; Occipital Lobe; Direct Current Artifacts; Butterworth Filter; Signal Preprocessing; Fast Fourier Transform; Clusterization.**

I. Introduction

BIOMEDICAL signals, such as Electroencephalography (EEG) are used to measure electrical brain activity with the help of electrodes that get in contact with the scalp. These signals represent cortical neuronal activities of different lobes of the brain such as frontal, temporal, central, parietal and occipital [1]. These electrodes that are located in the occipital lobe of the brain allow us to detect visual stimulus. The electrical activity of this brain area reflects the same frequency behavior than the visual stimulus [2]. There are many invasive and noninvasive methods to acquire these EEG signals. Invasive measurements require the use of needles or even complicated surgeries; however, the non-invasive method is more accessible to us and easy to perform. The non-invasive method is by far the most common method of measurement because it is superficial and can be performed with minimal risk to the person, in this method the electrodes measure generally cortical electrical activation [3].

The non-invasive method, also known as a surface method, despite being the most common, has interference problems caused by the electrodes used on the scalp; however, it is important that these electrodes are wetted with a conductive gel with components of sodium chloride to reduce the impedance of leather scalp and should avoid relative motion between the electrodes and the head of the volunteer [4]. Other types of perturbations which are susceptible EEG signals is the direct current (DC) and the artifacts that work with alternating current (AC). The amplitude of the EEG signals ranges from microvolts to lower millivolts range - mV (less than 10mV). The amplitude, and the properties of EEG signals in both the time domain and frequency depend on factors such as stimulus intensity, quality electrodes contacts used as a reference, the properties of the scalp skin (e.g., the thickness of the skin, adipose tissue, among others), electrode properties and the amplifier, as well as the conductive gel quality [4], [5].

The development of technologies for rehabilitation of patients with motor impairments, particularly for those who have difficulties to control their movements by diseases such as Parkinson's, or injury to the spine or muscle spasticity, may be possible through measuring bioelectric EEG signals that patients generate and therefore determining the movement intention of the patient. However, given the susceptibility of these signals to noise, some pre-processing methodologies and clustering of the EEG signals have been proposed [6].

In this paper we use a k-means algorithm for clustering pre-processed EEG signals using temporal characteristics and frequency to detect when a person is observing visual stimuli in two frequency ranges.

This paper describes in section II the methodology used for data collection with healthy volunteers. Section III shows the results of the pre-process and subsequent clusterization with the proposed algorithm and in section IV a discussion of these results and some conclusions are presented.

II. Methodology

The data acquisition process begins with capturing EEG signals from 5 healthy skilled volunteers who gave their written consent before performing the experiments. Each volunteer was asked to repeat an experiment for 10 times at different frequencies; each experiment was trigger by a visual

978-1-5090-1629-7/16/$31.00 © 2016 IEEE

stimulus.

The experiment was designed for recording the EEG signals generated by two electrodes: Left Occipital (LO) and Right Occipital (RO). These electrodes were placed on the surface of the occipital area of the scalp. For the experiment the volunteers were sitting in a comfortable chair and were placed at front of a display that generated visible stimuli through white LEDs on the following frequencies: 5, 6, 7, 8, 9, 24, 26, 27, 28, 29 Hz. The frequency of these stimuli was randomly generated by a digital frequency generator. The volunteers during the elicitation process must be completely relaxed in order for the experiment to be successful. Figure 1 shows a volunteer in the acquisition process of surface EEG signals using a noninvasive commercial equipment [7]. Each of these stimuli were 19.5 seconds long, a time that was established through previous laboratory tests as been sufficient for the volunteer to adapt to the visual stimuli.

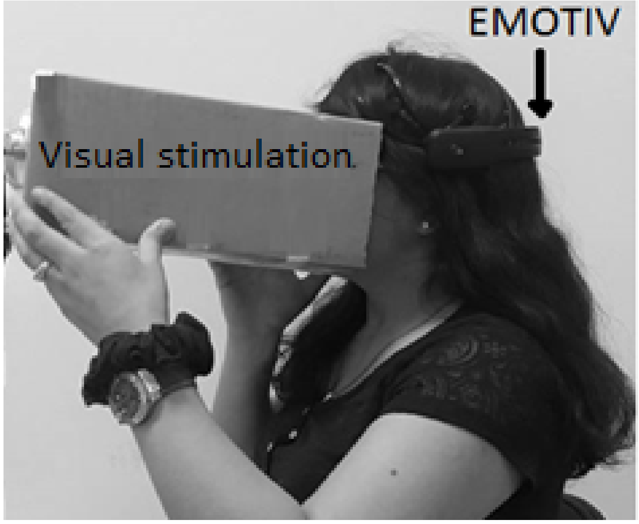

Fig. 1. Visual stimuli generated by a display with LEDs used to acquire the occipital EEG signals.

Each volunteer performed an experiment for each of the 10 visual stimuli frequencies (5, 6, 7, 8, 9, 24, 26, 27, 28, 29). In each experiment the EEG signals generated in the 2 electrodes (LO, RO) of the occipital area was simultaneously recorded. It is important to note that the data acquisition equipment has a sampling rate of 128 samples per second, allowing to acquire 2500 samples, considering that each task has a duration of 19.5 seconds.

The electrodes used to measure the EEG signals were the two occipital areas, as shown in Figure 2. These electrodes capture the visual stimuli generated by the eyeballs.

Figure 3 shows the behavior of the acquired EEG signals containing 2500 samples captured through two electrodes simultaneously around the occipital area. The horizontal lines of the figure represent the minimum and maximum average of the shift (offset) from the acquired signals. This is due to the presence of direct voltage current (DC), also known as DC artifacts [4], [8].

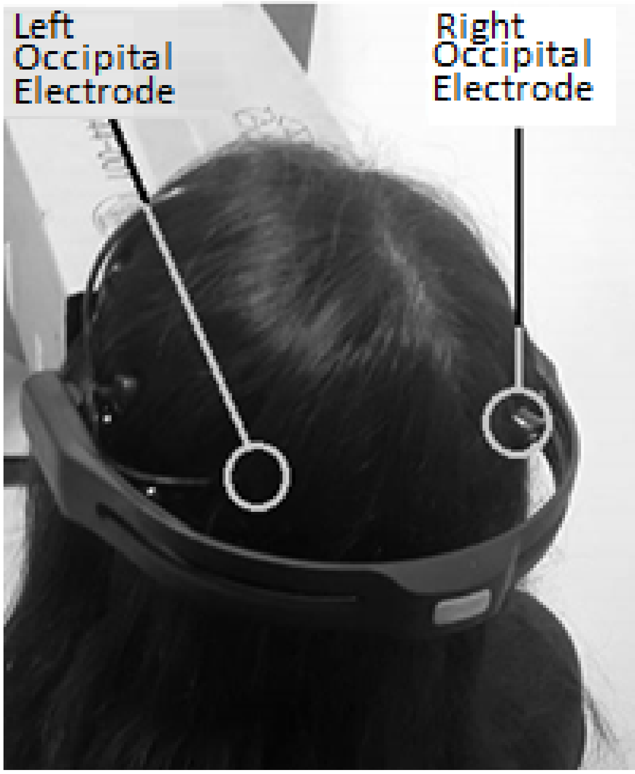

Fig. 2. Distribution of the 2 occipital electrodes Emotiv equipment.

For the simplicity of the analysis, the EEG signals of all experiments were grouped into a single matrix: EEG (2500 rows containing the samples x 20 columns containing the frequencies);

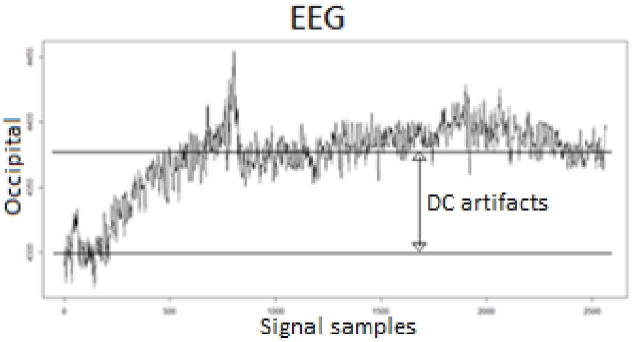

Fig. 3. DC artifacts present in the occipital EEG signals 5Hz visual stimulus.

Once the DC artifacts have been identified, the hypothesis testing normality of the acquired data can be performed. As demonstrated in [9], when data contains noise due to DC artifacts it does not behave as a normal distribution. In the test with a significance level of 5% for a null hypothesis Ho, the EEG signals captured were not normally distributed with zero mean and variance value of 1, whereas for the alternative hypothesis H1, the EEG signals were modelled as a normal distribution with zero mean and variance value of 1. Therefore, there is insufficient evidence to reject the null hypothesis Ho, because the captured signals are very susceptible to interference of noise coming from DC artifacts and the electrical grid. The presence of DC artifacts alters the characteristics of descriptive statistics in the time domain, and the values for mean and variance. Figure 4 shows a) the histogram of the occipital EEG signals with a visual stimuli area of 5Hz, which does not behave as a normal distribution; and, b) the comparison between the distribution of EEG signals vs data from a theoretical normal distribution. Plotting the acquired data, it shows a normal distribution of the samples.

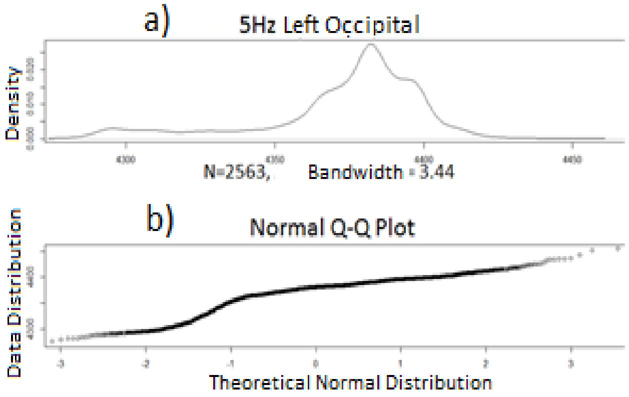

Fig. 4. a) Histogram of the EEG signal without pre-processing to occipital area with 5Hz visual stimulus. b) Comparison between the distribution of the EEG acquired data vs data from a normal distribution.

The results of the histogram for the EEG visual stimuli signals on the frequencies 6, 7, 8, 9, 24, 26, 27, 28, 29 Hz, show similar behavior to what is seen in Figure 4, even in the comparison between the distribution of EEG data vs. a theoretical normal distribution.

The analysis in the frequency domain has also been performed and the presence of noise with its effects was also detected. For this analysis the Fast Fourier Transform (FFT) was applied to the 2500 electrode samples per electrode per visual stimulus per volunteer. Figure 5 shows the signal as having a strong activity near to 0 Hz frequency in the presence of DC artifacts.

Having identified the presence and type of noise in the EEG signals, a Butterworth filter of third order was designed to obtain a frequency response plot as flat as possible and avoid distorting the original signal in the frequency domain [9]. The filter was designed for the frequency range between 5 and 30 Hz, which is the range in which EEG stimuli was generated.

III. RESULTS

After identifying the noise characteristics, the preprocessing of EEG signals was conducted by applying the designed filter. Figure 6 shows the 2500 electrodes samples. There can be seen the complete elimination of noise caused by the presence of DC artifacts. It can also be observed a signal without offsets or tendencies for performing analysis and extraction of temporal and frequency characteristics without major distortions.

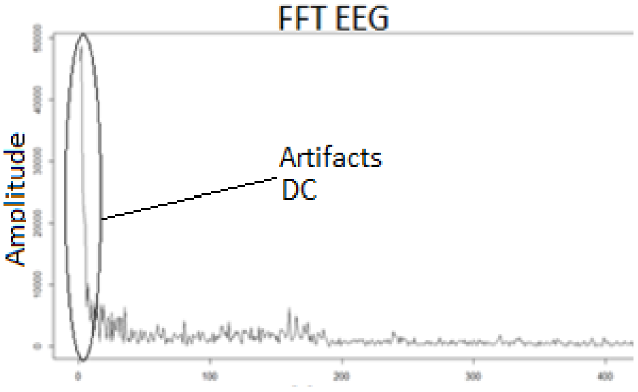

Fig. 5. An analysis to the FFT frequency of the EEG signal.

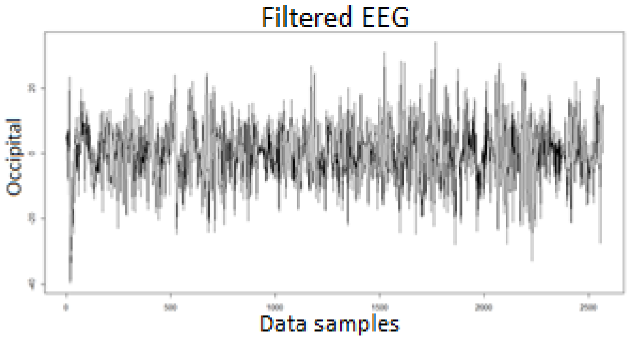

Fig. 6. EEG signal whithout DC artifacts in the 2 electrodes of the occipital area.

Figure 7 shows the signal without DC artifacts. This signal allows to assess cortical activity in the occipital area of volunteers and subsequently to extract descriptive statistics characteristics.

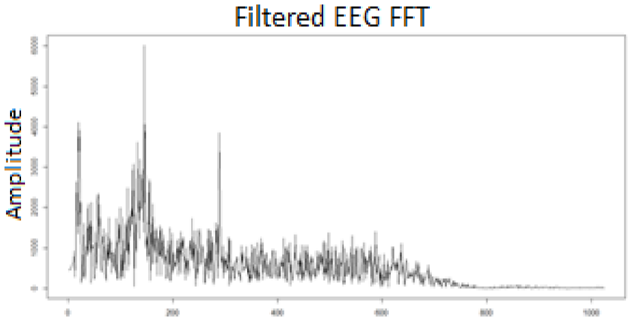

Fig. 7. Frequency analysis filtered with the FFT of the EEG signals.

After pre-processing the EEG signals the testing normality hypothesis was performed to a level of significance of 5%, being the Ho the null hypothesis. The EEG data are not normally distributed with mean value of 0 and a variance value of 1 versus the alternative hypothesis H1: The EEG data are normally distributed with mean and variance values of 0 and 1 respectively. The test indicates that there is sufficient evidence to reject the null hypothesis Ho, therefore we accept the alternative hypothesis H1. Figure 8 shows: a) the histogram of the EEG signals filtering the occipital area with the same 5Hz visual stimulus showing a normal distribution; b) shows the comparison between the distribution of EEG signals versus data of a theoretical normal distribution. The graph shows a normal distributed behavior and because of the perfect slope of graph, it is understandable to infer a normal theoretical distribution.

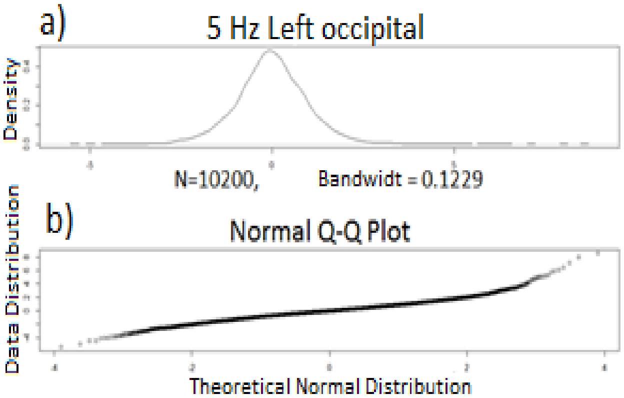

Fig. 8. a) Histogram of the EEG signal with pre-processing to occipital area with 5Hz visual stimulus. b) Comparison between the distribution of the EEG acquired data vs data from a normal distribution.

The results of the histogram of the EEG signals filtered with visual stimulus frequencies: 6, 7, 8, 9, 24, 26, 27, 28, 29 Hz, is similar to what is plotted in Figure 8, even for the distributed EEG data acquired vs. the theoretical normal distribution.

Once the acquired signals were preprocessed from the Left Occipital (LO) and Right Occipital (RO) electrods, the temporal statistical characteristics such as minimum, maximum, median, arithmetic mean, variance (LO, RO) covariance (LO, RO), Correlation (LO, RO) and the maximum frequency rate value of the FFT signal were extracted. Additionally, some characteristics of the signals in the frequency domain were extracted such as WhichMax (LO, RO), Variance (LO, RO) Covariance (LO, RO) and Correlation (LO, RO).

To facilitate the analysis a matrix was developed in which the rows represent the frequency of the visual stimuli and the columns represent the time characteristics and frequency of both electrodes from the left and right occipital areas (LO, RO). The algorithm used for determining the appropriate number of clusters is described below [10]:

1. Select K centroids (K rows chosen at random).
2. Assign each data point with the closest centroid.
3. Recalculate the centroid as the average of all data points in a cluster (i.e., the centroids are p-length mean vectors where p is the number of variables).
4. Assign data points to their closest centroids.
5. Repeat steps 3 and 4 until the observations are not reassigned or the maximum number of iterations (R use 10 as a default) is reached.

The algorithm, uses the enhanced R. Hartigan and Wong [11] algorithm. This means that in steps 2 and 4 each observation is assigned to the cluster with the smallest value of:

$$SS(k)=\sum_{i=1}^{n}\sum_{j=0}^{u}\left(x_{ij}-\bar{x}_{kj}\right)^{2}$$

Figure 9 shows the graph for the sum squared error (SS) values vs the number of clusters. This suggests that the optimal number of clusters is two (k = 2).

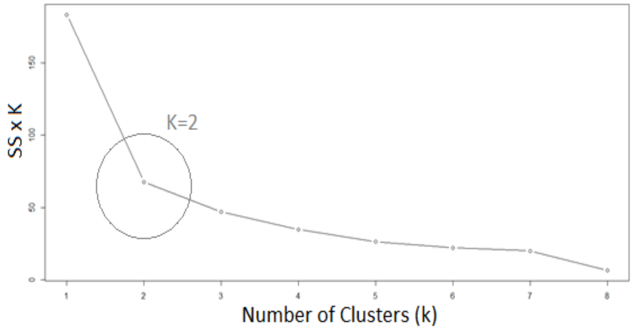

Fig. 9. SS vs k clusters.

Once known the value of k suggested for our dataset, the 2 clusters represent the signals belonging to the group of 5 to 9 Hz and the other group of 24 to 29 Hz. Then we proceed to apply the k-means algorithm to a group data with the following characteristics: a) data with the characteristics of variance (time and frequency), covariance (time and frequency) and correlation (time and frequency). b) The same data, but now with the index maximum frequency. c) The data now without variance (time and frequency), covariance (time and frequency) and correlation (time and frequency). The results that were obtained with the data groups are shown in Table 1. The success was measured by checking the clustering of data, a procedure that was performed with the following stimulus frequency: 6, 7, 8, 9, 24, 26, 27, 28, 29 Hz. obtaining tagged data.

TABLE I
HIT RESULTS USING K-MEANS TO DIFFERENT GROUPS OF FEATURES.

Table Head	Features	%Success
a	With: Var(t,f), Cov(t,f), Corr(t,f)	36%
b	With: WhichMax(f), Var(t,f), Cov(t,f), Corr(t,f)	80%
c	Only With:WhichMax(f)	80%

Figure 10 shows the cluster with the index variables of maximum frequency for both the left occipital electrode WhichMax_f01 (Wmax_f01) and the right occipital electrode WhichMax_f02 (Wmax_f02). This will give us a percentage between between_SS / total_SS = 74.3%.

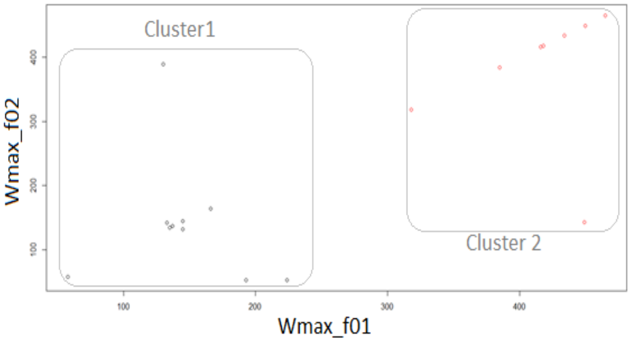

Fig. 10. EEG signals in the frequency range 5-9 Hz (cluster 1) and in the range of 24 to 29 Hz (cluster 2).

IV. Discussion and Conclusions

One of the goals of this research was to use a normal distribution test on EEG signals in both the time and frequency domain without pre-processing in order to detect the presence of noise parameters using descriptive statistics. This detection allowed us to determine the need for filtering the acquired signals. In addition, we had to use the FFT of the EEG signal without pre-processing to appreciate data with strong activity near 0 Hz. This led to the conclusion of being an effect of DC artifacts. It is interesting to note this was also visible in the time domain by detecting trends from these DC artifacts. Therefore, the hypothesis test is a valuable tool that helps us to detect the presence of noise given the temporal effects from low frequency DC artifacts as the average from the variation of the signal during the experiment, as well as its variance.

The design of the display played an important role in reducing the random environment noise, that shows the filter was good enough as preprocessing step.

Another objective of this work was to perform a clusterization with the pre-processed data based on temporal statistical characteristics and frequency. Based on the extracted characteristics two clusters were detected on each of the EEG signals: One in the frequency range of 5 through 9 Hz (cluster 1) and another in the range of 24 to 29 Hz (cluster 2).

As shown before a better cluster is obtaining by measuring the values of the index of maximum frequency. This happens because the occipital EEG signals show the same frequency behavior to visual stimuli. Furthermore, the use of Pearson correlation characteristics between the electrodes of the occipital area does not improve the cluster because the visual stimuli is on both eyes in every experiment for all frequencies.

The results with visual stimuli in the following frequencies 6.0, 7.0, 8.0 26.0, 27.0, 28.0 and 29.0 Hz was approximately 90%, because the 5 Hz frequency is very low for visual stimulation experiments; and, frequencies of 24 and 9 Hz conform the closest group signal that were used in the experiment.

References

[1] J. R. Garcell, "Aportes del electroencefalograma convencional y el análisis de frecuencias para el estudio del trastorno por déficit de atención. primera parte," *Salud Mental*, vol. 27, no. 1, p. 23, 2004.

[2] L. A. Riggs and P. Whittle, "Human occipital and retinal potentials evoked by subjectively faded visual stimuli," *Vision Research*, vol. 7, no. 5-6, pp. 441–451, 1967.

[3] G. Schalk and E. C. Leuthardt, "Brain-computer interfaces using electrocorticographic signals," *IEEE reviews in biomedical engineering*, vol. 4, pp. 140–154, 2011.

[4] I. Iturrate, C. Escolano, J. Antelis, and J. Minguez, "Dispositivos robóticos de rehabilitación basados en interfaces cerebro-ordenador: silla de ruedas y robot para teleoperación," in *III International Congress on Domotics, Robotics and Remote-Assistance for All, Barcelona, Spain*, 2009, pp. 124–134.

[5] G. R. Bermúdez, P. J. G. Laencina, D. Brizion, and J. R. Dorda, "Adquisición, procesamiento y clasificación de señales eeg para el diseño de sistemas bci basados en imaginación de movimiento," *Jornadas de introducción a la investigación de la UPCT*, no. 6, pp. 10–12, 2013.

[6] J. V. Pinzón, R. P. Mayorga, and G. C. Hurtado, "Brazo robótico controlado por electromiografía," *Scientia Et Technica*, vol. 1, no. 52, pp. 165–173, 2012.

[7] K. Jerbi, J. Vidal, J. Mattout, E. Maby, F. Lecaignard, T. Ossandon, C. Hamamé, S. Dalal, R. Bouet, J.-P. Lachaux *et al.*, "Inferring hand movement kinematics from meg, eeg and intracranial eeg: From brain-machine interfaces to motor rehabilitation," *IRBM*, vol. 32, no. 1, pp. 8–18, 2011.

[8] N. A. Badcock, P. Mousikou, Y. Mahajan, P. de Lissa, J. Thie, and G. McArthur, "Validation of the emotiv epoc® eeg gaming system for measuring research quality auditory erps," *PeerJ*, vol. 1, p. e38, 2013.

[9] I. Mesa, A. Rubio, I. Tubia, J. De No, and J. Diaz, "Channel and feature selection for a surface electromyographic pattern recognition task," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5190–5200, 2014.

[10] R. Kabacoff, *R in action: data analysis and graphics with R*. Manning Publications Co., 2015.

[11] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.