ISSN 2007-9737

Benchmarking of Averaging Methods Using Realistic Simulation of Evoked Potentials

Idileisy Torres-Rodríguez*,[1], Roberto Díaz-Amador[2], Beatriz Peón-Pérez[3],
Alberto Hurtado-Armas[1] , Alberto Taboada-Crispi[1]

[1] Universidad Central "Marta Abreu" de Las Villas,
Informatics Research Center, Santa Clara,
Cuba

[2] Universidad Católica del Maule,
Departamento de Medicina Traslacional,
Facultad de Medicina, UCM, Talca,
Chile

[3] Hospital Manuel Piti Fajardo,
Departamento de Electromedicina, Santa Clara,
Cuba

{Itrodriguez, ataboada]@uclv.edu.cu, rodiaz@ucm.cl

Abstract. The objective of this research is to conduct a comparative evaluation of various averaging methods for estimating evoked potentials using realistic simulations. Simulated signals are commonly employed to assess pattern recognition algorithms for event-related potential estimation since obtaining gold standard records is challenging. The simulations used are considered realistic because they allow for variations in potential latency, component width, and amplitudes. Background noise is simulated using an 8th order Burg autoregressive model derived from the analysis of a real dataset of auditory evoked potentials. The simulations incorporate actual instrumentation and acquisition channel effects, as well as power line interference. Three averaging methods for estimating the evoked potential waveform are compared: classical consistent average, weighted average, and reported average. The comparisons are conducted in two scenarios: one without artifacts and the other with 20% contamination by artifacts. The results of the comparative evaluation indicate that the trimmed average offers the best trade-off between the estimated signal-to-noise ratio (SNR) value and bias.

Keywords. Evoked Potentials, averaging methods, realistic simulation, benchmarking, SNR, bias.

1 Introduction

The utilization of simulated signals allows for training, evaluating, or comparing different digital signal processing techniques or pattern recognition algorithms, providing researchers with an unlimited number of test signals for experimentation [1]. While monitoring brain activity through electroencephalographic recordings is widely practiced, assessing the methods for analyzing these signals poses a challenge due to the absence of a reliable gold standard for comparison.

However, to assess various algorithms proposed for signal analysis and pattern detection, researchers often resort to using simulated signals instead of real signals, which typically conform to oversimplified models that do not accurately represent reality. In [2], a system is introduced for generating simulations of evoked potential recordings that exhibit a high level of realism.

This simulation takes into consideration potential variations in latency, width, and amplitude, which are common in real-world scenarios. Event-related potentials in actual

doi: 10.13053/CyS-28-1-4894

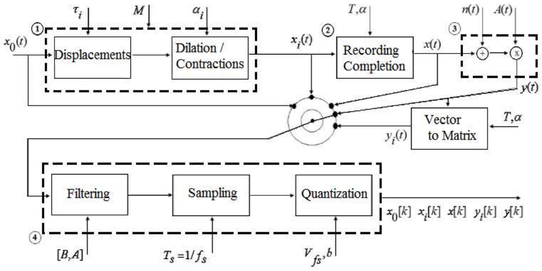


Fig. 1. General scheme for event-related potentials in wide sense simulation

Table 1. Characteristics of the designed Butterworth high-pass filter

Filter's design characteristics	Value
f_c in the passband	30 Hz
f_c in the stopband	15 Hz
Attenuation in the passband	1 dB
Attenuation in the stopband	-6 dB
Order	2

contexts can be contaminated by additive and multiplicative noise, as well as affected by recording instrument effects such as analog filtering, sampling, and quantization. Unfortunately, these aspects are often overlooked in most current evaluations.

All these parameters were estimated from real signals described in [3]. Using realistic simulations of evoked potentials and their associated noise and interference, different methods of robust estimation of the evoked response waveform will be evaluated.

2 Methods

2.1 Selection of the Parameters for Realistic Simulation

The simulation scheme (Fig. 1) used was previously proposed in [3] to simulate event-related potentials in a wide sense. The selection of the parameters for the realistic simulations of evoked potentials was carried out in [2].

In block 1 of Fig.1, $x_o(t)$ represents the initial fundamental epoch, which is defined for all $t \in [0, a)$, and serves as a reference for generating $(M - 1)$ additional epochs. The goal is for the initial waveform to closely resemble the waveform of the potential under study.

In this specific case, the clean recordings of Auditory Evoked Potentials from healthy individuals, obtained from the database published in [4] and described in [3], are chosen as the basic waveforms for simulation. These signals specifically correspond to auditory brainstem responses (ABR) and are characterized by their short-latency potentials.

The parameter τ_i , which accounts for the variations in relative displacements due to latency in $x_o(t)$, is simulated using a normal distribution with a mean of zero and a standard deviation on the order of 0.2 ms. This value is derived from the average standard deviations of the component V latencies, as indicated in the study conducted in [5].

Similarly, the parameter α_i, representing the variations in the width of $x_o(t)$, is simulated using a normal distribution with a mean of zero and a standard deviation on the order of 0.07 ms, based on the values reported in [5]. The simulation allows for selecting different values of M, depending on the desired size of the resulting matrix set.

The selection of the event period T is determined based on the stimulation period, which in this instance comprised 2002 samples (equivalent to 41.7 ms), representing the time interval between each applied stimulus. In this particular case, the width of the analysis window, denoted as a, was set to 884 samples (equivalent to 18.4 ms).

These parameter values can be adjusted to accommodate the specific choice of the initial basic epoch and the potential being simulated. Regarding the additive noise component, denoted as $n(t)$, it is generated as a sum of various sources.

In this case, it consists of the estimated background noise, the 60 Hz interference, and its harmonics, as well as the alpha rhythm commonly present in many signals from the selected database. To process this noise, a low-pass filter is applied using the coefficients estimated by an 8th-order Burg model. Subsequently, a high-pass filter is employed, following the specifications detailed in Table 1, as outlined in the approach presented in [2].

To simulate the alpha rhythm, which appears in certain signals and must be considered when analyzing signal non-homogeneities that can impact the estimation of the average signal, a white noise signal is employed. This white noise signal is subjected to bandpass filtering using a second-order Butterworth approximation with cutoff frequencies ranging from 9 to 11 Hz, as detailed in [6].

The amplitude of the alpha rhythm is randomly distributed between 30 μV and 50 μV, following a normal distribution. This distribution aligns with the analysis carried out on the dataset, ensuring consistency with the characteristics of the actual signals. The parameters associated with filtering, sampling, and quantization are derived from the description provided in the database acquisition documentation.

2.2 Average Methods

The coherent average also referred to as the arithmetic mean (M_Mean as denoted in this study), can be computed using the ensemble matrix formed by the simulated M evoked responses [7].

In this context, the response p_i to the i-th stimulus is considered to be the sum of the deterministic component of the evoked signal or response s, along with an asynchronous random noise r. The model for each of the M simulated responses can be expressed using formula 1:

$$p_i = s + r_i\,, \qquad 1 \le i \le M. \tag{1}$$

The estimation of the deterministic component of the signal, denoted as $\hat{s}$, can be obtained using formula 2, with N representing the number of samples comprising each response [1]:

$$\hat{s}(n) = \tfrac{1}{M}\textstyle\sum_{i=1}^{M} p_i(n), \quad 1 \le n \le N. \tag{2}$$

The application of signal averaging assumes that the underlying noise is stationary and follows a normal distribution with a mean of zero. Additionally, the noise variance should be consistent and equal across all potentials.

However, this condition is not always met, which can impact the effectiveness of the coherent average. To address this limitation, various methods have been proposed in the literature, including weighted averaging and robust averaging [1]. In the case of estimating the deterministic component of the signal, a weighted average approach is employed [8], as described by formula 3:

$$\hat{s}_w(n) = \sum_{i=1}^{M} w_i p_i(n), \quad 1 \le n \le N. \tag{3}$$

In the weighted average (Weighted_Mean) approach, each evoked response is assigned a weight based on specific criteria. One commonly used criterion is to assign weights based on the variance of the estimated noise in each response.

In this method, a smaller weight is assigned to potentials with higher levels of noise [5, 9, 11, 12]. Equation (4) represents the formulation corresponding to these weight assignment criteria:

$$w_i = \frac{1}{\sigma_i^2}\left(\sum_{j=1}^{M}\frac{1}{\sigma_j^2}\right)^{-1}, \quad i = 1, \cdots, M. \tag{4}$$

In formula 4, σ_i^2 represents the noise variance in the i-th potential. If the noise variance were constant across all records, the optimal value of w_i would be $1/M$, corresponding to the traditional average. Both the coherent average and the weighted average are linear techniques and perform well when the noise follows a Gaussian distribution [1, 8].

However, these techniques have limitations when occasional artifacts with large amplitude values (outliers) are present. In the literature, a family of estimators known as trimmed mean has been proposed to mitigate contamination by outliers [1, 5, 8]. The trimmed estimators are based on the median, which serves as the Maximum Likelihood (ML) estimator of s when the noise is assumed to follow a Laplacian distribution [1].

To compute the median, the samples in the ensemble matrix are ordered based on their amplitudes independently for each time point relative to the stimulus, regardless of other time points. This independence allows the median averaging to be unaffected by the nonstationarity of the noise within an epoch.

In the trimmed mean methods, the coherent average is combined with the median to obtain the

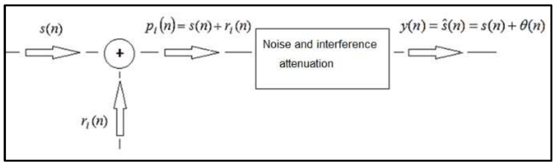

$s(n)$
$p_i(n) = s(n) + r_i(n)$
Noise and interference attenuation
$y(n) = \hat{s}(n) = s(n) + \theta(n)$
$r_i(n)$
+

Fig. 2. Signal and noise modelling

final response. The ensemble matrix is ordered, and a portion of extreme values is discarded or modified, while all other values are used for averaging similarly to conventional mean averaging. It is important to note that the rejection of extreme values differs from the concept of artifact rejection.

The *α*-trimmed mean (Trimmed_Mean) is one of the most popular trimmed estimators [9]. Equation (5) represents the estimation of the deterministic signal using the α-trimmed mean:

$$\hat{s}_{rec}(n) = \frac{1}{M - 2 \cdot K} \sum_{i=K+1}^{M-K} p_i\,(n). \tag{5}$$

If we compare formula 5 with formula 3, it becomes apparent that the weights can be assigned using the following equation:

$$w_i = \begin{cases} \dfrac{1}{M - 2 \cdot K}, & K + 1 \leq i \leq M - K, \\ 0, & in\ other\ case, \end{cases} \tag{6}$$

where α represents the percentage of trimming, M denotes the number of responses, and $K = \alpha M$ corresponds to the number of observations that are eliminated from each extreme of the ordered matrix.

2.3 Quality Measures

Given that the acquired signal (*p*) can be modelled as the desired signal (*s*) plus additive noise *(r*), as shown in formula 1, by applying different techniques to estimate the desired signal, attenuating the different existing interferences (Fig. 2). The output (*y*), after applying these techniques to estimate the desired signal ($\hat{s}$), can be seen as the combination of the desired signal (*s*) plus a remaining noise (*θ*).

The quality of the estimation of a signal segment can be expressed in terms of several parameters. Some of them are the signal-to-noise ratio (SNR) and the bias, which are expressed in equations 7 and 8, respectively [10]:

$$SNR = \frac{\sum_{j=1}^{N} s^2[j]}{\sum_{j=1}^{N} \theta^2[j]}, \tag{7}$$

$$b_\theta = \frac{1}{N} \sum_{j=1}^{N} |\theta[j]|. \tag{8}$$

In each of the above equations, *N* represents the total number of samples of the segment to be evaluated, *θ* is the remaining noise in the signal (signal obtained after attenuating the noise minus the ideal signal). The subscript *j* refers to the *j*th sample of the affected parameter and *s* is pure ideal signal. It is important to reach a compromise between bias and SNR (Equations 7 and 8).

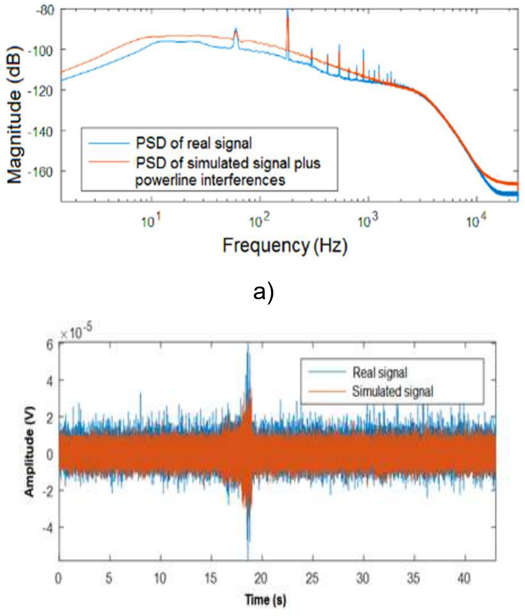
Since bias is the factor that indicates the distortion that is introduced by using a given noise and interference attenuation method, it is necessary to achieve high SNR values but low bias values. Unfortunately, in real situations, the pure ideal signal is not available a priori, so it is impossible to use these measures. But in a controlled environment, such as when using simulated signals, these measures can be used.

2.4 Experiment Description

To evaluate the averaging methods described using simulated signals, 100 data sets of 2,000 epochs each were obtained, without adding artifacts, and then the same 100 data sets of 2,000 epochs, where the 10%, 20%, and 30 % of the total samples of the array were randomly contaminated with outliers.

It was decided to add this level of artifacts based on other experiments found in the literature [6]. From each data set, 512 epochs were randomly selected, 100 times, thus forming a Monte Carlo experiment of 100 runs.

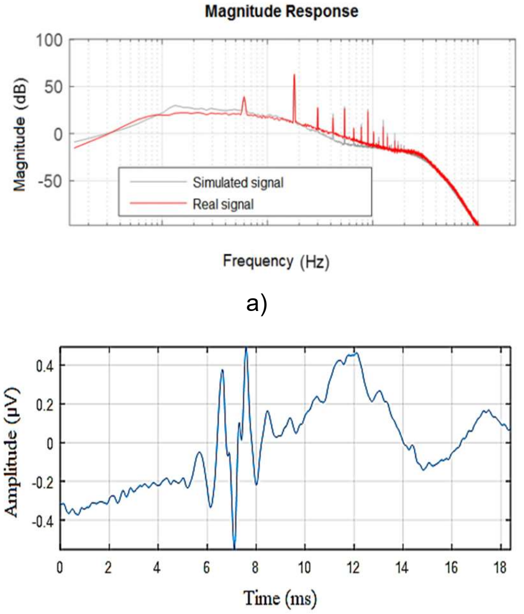
Evoked responses were then estimated using the classic Ensemble Average (M Mean), the Trimmed Average (Trimmed Mean), with a 20% trim factor, and the Weighted Average (Weighted Mean). The signal-to-noise ratio and bias values were calculated on the estimated signals to compare the estimation methods.



a)

b)

Fig. 3. a) The spectra of the simulated background noise and the reference signal. b) Example of the simulated background noise and the reference signal in the time domain



a)

b)

Fig. 4. a) Spectral comparison between simulated and real records. b) initial epoch

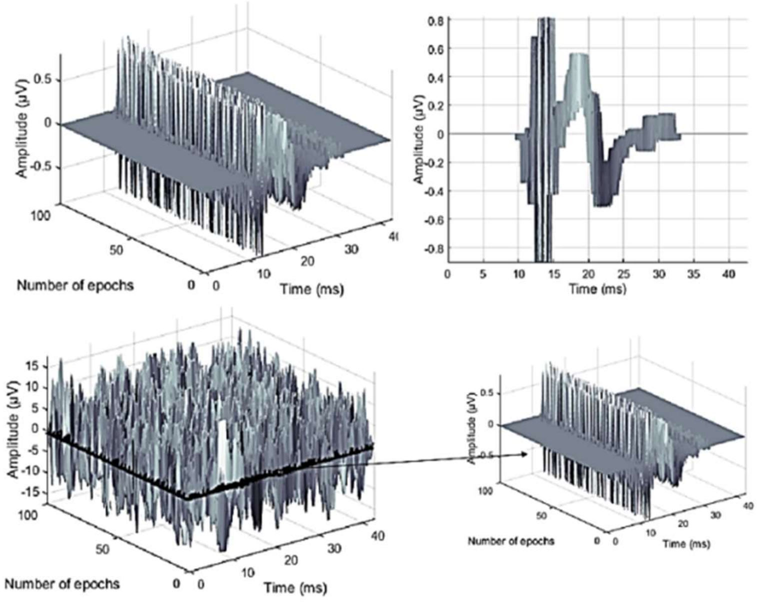


Fig. 5. Effects of relative displacements that may be present in a real scenario and of low SNR that may be present in a real environment

3 Results and Discussion

3.1 Simulated Noised

To generate the simulated noise signals, we first generated Gaussian white noise, which was subsequently subjected to low-pass filtering using coefficients estimated by the model. Next, the noise was high-pass filtered using a filter with the specifications outlined in Table 1.

Additionally, 60 Hz powerline interference was added to the noise signal. In Fig. 3a, we can observe a comparison of the spectra between one of the background noise signals simulated using this approach and the actual reference signal.

Figure 3b illustrates the comparison of the signals in the time domain. It is worth noting that the analysis was conducted on one-second segments of the signal to ensure stationary conditions. The simulated signal demonstrates the variations that have occurred in the signal's variance over time.

3.2 Simulated EP (Evoked Potential) Records

Figure 4a presents a comparison in the frequency domain between a simulated signal generated according to the specifications described earlier. In this case, the initial epoch, denoted as $x_o(t)$, corresponds to subject 6 and was selected randomly. To perform the spectral comparison, the "dirty" record that served as the source for the initial epoch (Fig. 4b) was chosen.

The tests yielded a consistent NRMSE adjustment of 92.5%. Randomly selected initial epochs were used, and simulated noise, interferences, and artifacts were added, resulting in a signal-to-noise ratio of -26.71 dB. The top part of Figure 5 visually demonstrates the effects of relative displacements that can occur in real scenarios.

This simulation example includes noise, interference, and artifacts. The lower part of Figure 5 shows the ensemble matrix, which combines evoked responses with noise and interference. Due to the high level of noise and interference, with an initial signal-to-noise ratio of -26.71 dB, it is not possible to discern any waveform associated with the desired signal.

Table 2. SNRs in dB were obtained with a Monte Carlo experiment of 100 runs on Simulated Data Sets

Average Methods	0% artifacts	10% artifacts	20% artifacts	30% artifacts
Initial SNR (dB)	-26.04 ± 1.17	-30.02 ± 1.06	-32.54 ± 0.93	-34.15 ± 1.01
Mean	-0.20 ± 1.09	-5.90 ± 0.35	-5.68 ± 0.78	-7.92 ± 0.23
Weighted Mean	1.96 ± 0.29	0.82 ± 0.04	-0.25 ± 0.20	-1.30 ± 0.04
Trimmed Mean	-0.66 ± 0.77	-3.88 ± 0.53	-0.81 ± 0.35	-1.83 ± 0.33

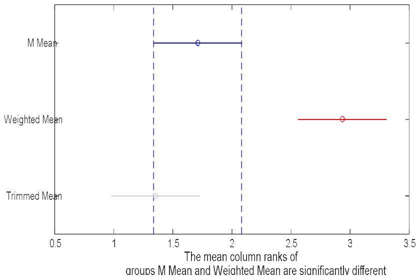


Fig. 6. Differences between the mean SNR values obtained for the averaging methods in the data set without artifacts

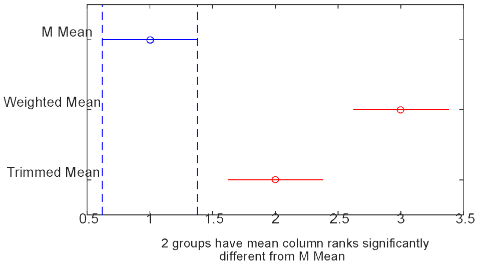


Fig. 7. Differences between the mean SNR values obtained for the averaging methods in the data set with 10% of the samples with artifacts

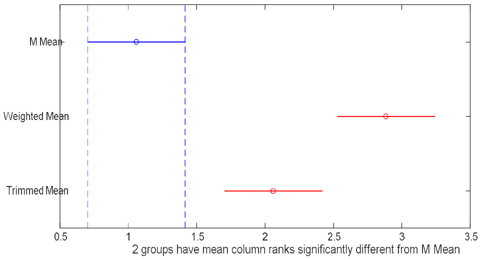


Fig. 8. Differences between the mean SNR values obtained for the averaging methods in the data set with 20% of the samples with artifacts

ISSN 2007-9737

It should be noted that the achieved level of realism in this simulation is significantly higher than in previous simulations, making direct comparisons with previous simulations inappropriate.

The codes used in these simulations are available in the GitHub repository for benchmarkingpurposes[1].

3.3 Estimation of Event-Related Potentials Using Realistic Simulation

Table 2 displays the average Signal-to-Noise Ratio (SNR) values and their corresponding standard deviations obtained from the experiment described in section 2.4.

The results indicate that the Weighted Mean method consistently yielded the highest SNR values across all cases. This suggests that the Weighted Mean method performed better in terms of minimizing the impact of noise and maximizing the clarity of the desired signal compared to the other methods evaluated in the experiment.

Based on the results obtained, a Friedman test was performed to analyze whether there were significant differences in at least one of the averaging methods used for estimation. A value of $p < 0.05$ was obtained, so at least two combinations have significant differences concerning their means.

A post-hoc was performed using the Bonferroni test with an alpha of 0.05 to determine which combinations have differences. Figure 6 shows the multicomparison between the three average methods used, it can be seen how there are differences between the Weighted Mean method with respect to the other two. The differences between the M Mean and Trimmed Mean are not significant.

A similar analysis was performed when 10%, 20%, and 30% of the samples were contaminated (Fig.7 - Fig.9). In this case, the results of the SNR values have significant differences between the three methods. No critical distance overlaps. In all cases, with 0% artifacts, 10%, 20% and 30%, the Weighted_Mean method offered the best results in terms of the value of the signal-to-noise ratio.

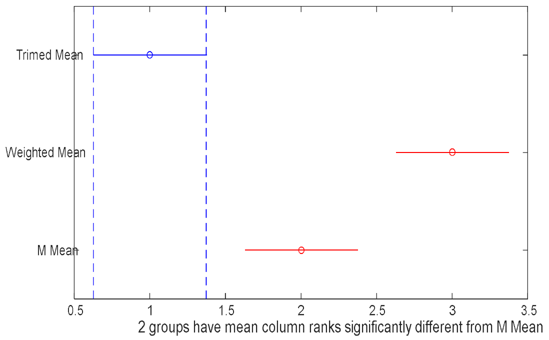


Fig. 9. Differences between the mean SNR values obtained for the averaging methods in the data set with 30% of the samples with artifacts

Table 2. Modified bias in μV obtained with a Monte Carlo experiment of 100 runs on a Simulated Database

Average Methods	0% artifacts	10% artifacts	20% artifacts	30% artifacts
Mean	0.16 ± 0.07	0.24 ± 0.03	0.28 ± 0.02	0.31 ± 0.02
Weighted Mean	0.09 ± 0.01	0.10 ± 0.01	0.23 ± 0.01	0.19- ± 0.01
Trimmed Mean	0.17 ± 0.01	0.21 ± 0.01	0.20 ± 0.01	0.15 ± 0.01

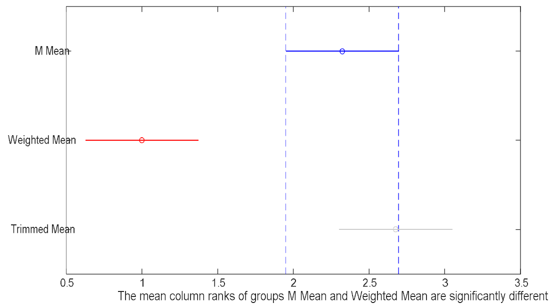


Fig. 10. Differences between the mean values of bias obtained for each averaging method in the data set without artifacts

With the bias, an analysis similar to that performed with the SNR was performed, and the lowest distortion values of the resulting signal were obtained for the weighted average when there

[1] https://github.com/itrodriguez/SimuldorEP/tree/main

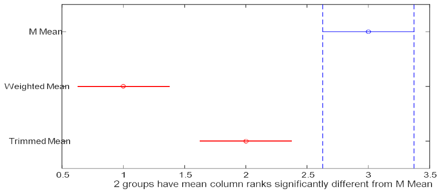


Fig.11. Differences between the mean values of bias obtained for each averaging method in the data set with 10% of the samples with artifacts

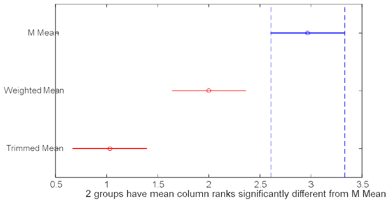


Fig. 12. Differences between the mean values of bias obtained for each averaging method in the data set with 20% of the samples with artifacts

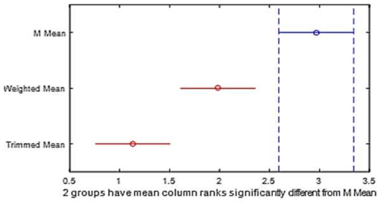


Fig.13. Differences between the mean values of bias obtained for each averaging method in the data set with 30% of the samples with artifacts

were 0% artifacts and for the trimmed average when there were 30% artifacts.

Figures 10, 11, 12, and 13 show the Bonferroni-Holm post hoc analysis with alpha equal to 0.05 to assess the differences between the mean values of bias obtained for the data set without artifacts and with outliers, after finding through a Friedman test that at least one of the methods had significant differences.

When the data set has 0% artifacts, the lowest degree of distortion is presented by the weighted average, with significant differences concerning the average and trimmed average, however, there are no significant differences between these last two methods.

When the samples are contaminated with artifacts, the best results are offered by the trimmed mean for 20% and 30%, significantly different from the other two methods.

In the case of the SNR calculation, it is not the one that offers the highest value, but let us remember that the objective of these two measures is to provide a compromise ratio. So, when there are artifacts, the best compromise is offered by the trimmed mean.

4 Conclusions

Simulated raw recordings of evoked potentials provide a controlled dataset for benchmarking new methods in detecting evoked responses.

Burg's method, utilizing an 8th-order autoregressive (AR) model, offers a reliable estimate of the background noise. Simulations can also incorporate interferences commonly found in real signals, such as 60 Hz powerline interference, alpha rhythm, and instrumentation channel noises. Furthermore, the simulation scheme allows for the introduction of out-of-range values and impulsive noise.

In the benchmarking study of Averaging Methods using Realistic Simulation of Evoked Potentials, it was observed that the weighted average method yields superior results when the data is free from artifacts.

However, in cases where artifacts are present, the trimmed mean method provides the best compromise in terms of performance.

Acknowledgments

This work has been carried out thanks to the doctoral program of Informatics and Computational Sciences of the Central University "Marta Abreu" of Las Villas.

References

1. **Krol, L. R., Pawlitzki, J., Lotte, F., Gramann, K., Zander, T. O. (2018).** SEREEGA: Simulating event-related EEG activity. Neurosci Methods, Vol. 309, pp. 13–24. DOI: 10.1016/j.jneumeth.2018.08.001.

2. **Torres-Rodríguez, I., Díaz-Amador, R., Peón-Pérez, B., Hurtado Armas, A., Taboada-Crispi, A. (2023).** Realistic simulation of event-related potentials and their usual noise and interferences for pattern recognition. In: Rodríguez-González, A.Y., Pérez-Espinosa, H., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Olvera-López, J.A. (eds) Pattern Recognition, Proceedings of 15 th Mexican Conference Pattern Recognition, 2023. Lecture Notes in Computer Science, pp. 201–210. DOI: 10.1007/978-3-031-33783-3_19.
3. **Silva, I., Epstein, M. (2010).** Estimating loudness growth from tone-burst evoked responses. The Journal of the Acoustical Society of America, Vol. 127, No. 6, pp. 3629–3642. DOI: 10.1121/1.3397457.
4. **Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Stanley, H. E. (2000).** Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation, Vol. 101, No. 23, pp. e215–e220. DOI: 10.1161/01.CIR. 101.23.E215.
5. **Valderrama, J. T., De la Torre, A., Alvarez, I., Segura, J. C., Thornton, A. R. D., Sainz, M., Vargas, J. L. (2014).** Automatic quality assessment and peak identification of auditory brainstem responses with fitted parametric peaks. Computer methods and programs in biomedicine, Vol. 114, No. 3, pp. 262–275. DOI: 10.1016/j.cmpb.2014.02.015.
6. **Leonowicz, Z., Karvanen, J., Shishkin, S. L. (2005).** Trimmed estimators for robust averaging of event-related potentials. Journal of Neuroscience Methods, Vol. 142, No. 1, pp. 17–26. DOI: 10.1016/j.jneumeth.2004. 07.008.
7. **Torres-Rodríguez, I., Ferrer-Riesgo, C. A., P. de Morales-Artiles, M. M., Taboada-Crispi, A. (2020).** Performance evaluation of average methods in the time domain using quality measures for automatic detection of evoked potentials. VIII Latin American Conference on Biomedical Engineering, CLAIB 2019, IFMBE prodeedings, Vol. 75, pp. 12–20, DOI: 10.1007/978-3-030-30648-9_2.
8. **Pander, T. (2015).** A new approach to robust, weighted signal averaging. Biocybernetics and Biomedical Engineering, Vol. 35, No. 4, pp. 317–327, DOI: 10.1016/j.bbe.2015. 06.002.
9. **Torres-Rodríguez, I., Ferrer-Riesgo, C. A., Oliva Pérez, J. C., Taboada-Crispi, A. (2019).** Performance of different average methods for the automatic detection of evoked potentials. In: Nyström, I., hernández-Heredia, Y., Milián-Núñez, V. (eds.) Iberoamerican Congress on Pattern Recognition, Springer, Vol. 11896, pp. 629–636, DOI: 10.1007/978-3-030-33904-3_59.
10. **Novis, K. Bell, S. (2019).** Objective comparison of the quality and reliability of auditory brainstem response features elicited by click and speech sounds. Ear Hear, Vol. 40. No. 3, pp. 447–457, DOI: 10.1097/AUD.00 00000000000639.

Article received on 04/07/2023; accepted on 13/10/2023.
**Corresponding author is Idileisy Torres-Rodríguez.*