PROCEEDINGS OF SCIENCE

ONLINE ICRC 2021
THE ASTROPARTICLE PHYSICS CONFERENCE
Berlin | Germany
37th International Cosmic Ray Conference
12–23 July 2021

Denoising cosmic rays radio signal using Wavelets techniques

Watanabe, C.K.O.,[a,b,c,*] **Diniz, P.S.R.,**[b] **Huege, T.**[c,d] **and De Mello Neto, J.R.T.**[a]

[a] *Federal University of Rio de Janeiro (UFRJ), Physics Institute, Rio de Janeiro, Brazil*
[b] *Federal University of Rio de Janeiro (UFRJ), The Alberto Luiz Coimbra Institute for Graduate Studies and Research in Engineering (COPPE), Signal, Multimedia and Telecommunications Laboratory (SMT), Rio de Janeiro, Brazil*
[c] *Institute for Astroparticle Physics (IAP), Karlsruhe Institute of Technology, Karlsruhe, Germany*
[d] *Astrophysical Institute, Vrije Universiteit Brussel, Brussels, Belgium*
E-mail: clara.watanabe@kit.edu, diniz@smt.ufrj.br, tim.huege@kit.edu, jtmn@if.ufrj.br

Estimating a signal embedded in noise requires taking advantage of any prior information about the signal and noise. Until recently, signal processing estimation was primarily Bayesian and linear. Non-linear smoothing algorithms existed in statistics, but these procedures were often ad-hoc and complex. Donoho and Johnstone proved that a simple thresholding algorithm on an appropriate basis could be a nearly optimal non-linear estimator [2].

A radio signal induced from cosmic ray is very well described, for example, by the Daubechies [3] wavelets as a basis, allowing the thresholding to be as safe (or more) as a linear estimation. The best basis search or a pursuit algorithm may also improve the thresholding performance as it adapts the basis to the noisy data.

This work presents the wavelets as a denoising technique for narrowband and gaussian background reduction in radio signals induced from cosmic rays, showing its efficiency in energy reconstruction.

37th International Cosmic Ray Conference (ICRC 2021)
July 12th – 23rd, 2021
Online – Berlin, Germany

*Presenter

PoS(ICRC2021)260

1. Introduction

The radio detection of a cosmic ray is a modern, solid and low-cost technique that uses antennas to detect the electromagnetic component of the air shower [1]. Nowadays, several experiments are running (and planned) worldwide, and in both cases, the community expects to improve the sensitivity of estimating the primary particle energy, composition, and arrival direction.

The main challenge of this type of detection is the background present in the experiment site due to human-made noise, and also the galactic Gaussian noise [4]. The former is dramatically reduced if the experiment site is deployed in a deserted and remote area, while the second is irreducible. When it is not an option to change the site location, it is essential to make an effort on signal processing for narrowband and wideband noise removal.

The wavelets denoising technique is a very well-known technique used nowadays to denoise a signal corrupted with noise, be it Gaussian or impulsive. Although the advantage of the denoising process on a signal is straightforward, for the cosmic ray induced radio signal scenario, it is a central point for the reconstruction of properties of the primary particle, i.e. energy, mass and arrival direction.

This work presents the Stationary Wavelet Transform (SWT) efficiency, for denoising cosmic-ray induced radio signals of a simulated dataset of realistic events. A wavelet mother from the Daubechies [3] family is used. The electric field is generated using the CORSIKA [5] software with the CoREAS [6] radio extension. The simulations represent air showers as measured by the Auger Engineering Radio Array (AERA) [9], and the Offline [10] reconstruction framework is used to apply the antenna and analog chain response and incorporate measured background data from the experiment site. It is expected that both Gaussian and impulsive noise are present in these background samples. The data set consists of 344 simulated air showers.

2. Wavelets

The Fourier transform of a signal $f(t)$ computes the correlation between $f(t)$ and the orthogonal basis of sines. Although this brings relevant information, the limitation is evident when we want to detect singularities in the signal. In general terms, the energy spread of a signal and its Fourier transform cannot both be simultaneously small. As the Fourier transform, the Wavelet transform also computes the correlation between $f(t)$ and the orthogonal basis of a Wavelet mother. In other words, a single wavelet mother is stretched, expanded, and translated in time. It can measure the time-frequency variation of spectral components, but it has a different time-frequency resolution, allowing the characterization of transients with a zooming procedure across scales. The wavelet mother on a discrete resolution is described as

$$\Psi_{n,j} = \frac{1}{\sqrt{2^j}}\Psi\left(\frac{t - n2^j}{2^j}\right). \tag{1}$$

The coefficients are calculated as

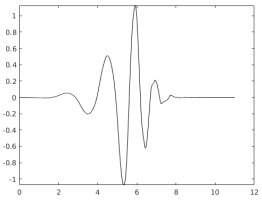
$$\langle f, \Psi_{n,j}\rangle = \int f(t)\Psi_{n,j}(t)\mathrm{d}t, \tag{2}$$

and it is proved [2] that one can similarly recover the function as

$$f(t) = \sum_{j,n}\langle f, \Psi_{n,j}\rangle\Psi_{n,j}. \tag{3}$$

That is the primary mechanism of the *Discrete Wavelet Transform* (DWT) which is very similar to the *Stationary Wavelet Transform* (SWT), where j is a pre-defined maximum number of levels, and the signal length must be a multiple of 2^j.

Concerning the wavelet mother, we have chosen the Daubechies 6 [2] illustrated in figure **??**. The choice of a proper wavelet mother requires previous knowledge of the expected signal because, as shown, the wavelet transform searches for correlations between the wavelet mother and the expected signal. There are no specific



(a) The Daubechies nº6 (db6) wavelet with arbitrary units.

rules for selecting the correct wavelet mother, but we have chosen based on signal shape similarities between cosmic-ray-induced radio signals as predicted by CoREAS plus the AERA detector response and the wavelet mother. It is also worth mentioning that it is possible to create a custom wavelet mother, of course, following the rules of the wavelet function [2]. It is also expected that this work can be improved further by testing different types of wavelets mothers.

2.1 Denoising and Thresholding

We applied the SWT algorithm to denoise radio signals, followed by a thresholding method and the inverse SWT (iSWT) to recover the signal. The corrupted radio signal can be defined as

$$f = X + W, \tag{4}$$

where X is the pure signal and W is an additive noise. Recalling equation (2)

$$\langle f, \Psi_j \rangle \Psi_j = \langle X, \Psi_j \rangle \Psi_j + \langle W, \Psi_j \rangle \Psi_j, \tag{5}$$

where we expand our signal in the wavelet mother basis, and we may name the terms of the equation above respectively as $f_B[j]$, $X_B[j]$ and $W_B[j]$. Rearranging the terms of the equation above we get

$$X_B[j] = f_B[j] - W_B[j]. \tag{6}$$

The denoised signal is then

$$\tilde{F} = DX = \sum_j d_j(x) X_B[j] \Psi_j, \tag{7}$$

where a diagonal operator D is applied that estimates independently each $f_B[j]$ from $X_B[j]$ with a function $d_j(x)$. Among several types of thresholding functions $d_j(x)$ we have chosen what is called the *soft threshold* function [7], that is defined as

$$d_j(x) = \begin{cases} x - T & \text{if } x \geq T \\ x + T & \text{if } x \leq -T \\ 0 & \text{if } |x| \leq T \end{cases} \tag{8}$$

where T is the threshold value. We expect the error between f and $\tilde{F}$ to be as low as possible, and such amount is defined as the *risk*. Donoho and Johnstone proved [8] that if

$$T = \sigma \sqrt{2 \log_e N}, \tag{9}$$

where σ is the f variance and N is f length, the risk is minimal. It is also worth mentioning that there is room for improvements and investigation concerning the choice of T. Furthermore, more refined methods exist [2] that can be applied to our case and even level dependant.

In figure 1 this procedure is performed with an example of simulated radio signal from Coreas plus the AERA antenna response. Figure 2 shows the result of the denoising process of the signal. We may observe an overall reduction of the background without causing any extra distortion on the signal from this procedure.

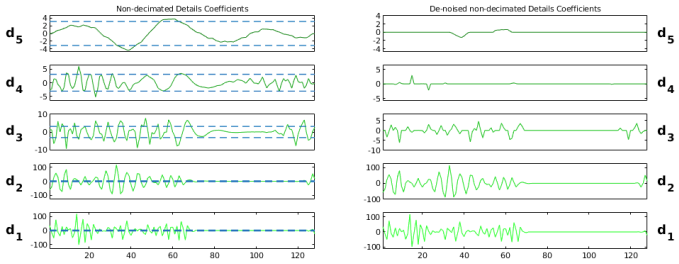
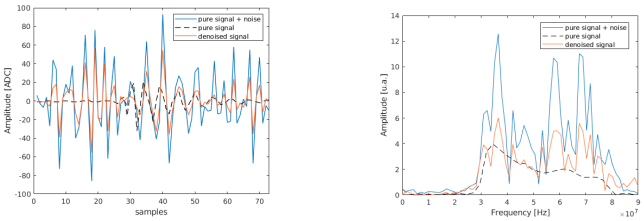


Figure 1: In the left column we show the values of the coefficients of the decomposition of a corrupted radio signal into the wavelet mother basis in 5 levels. In blue the universal threshold calculated from equation (10) is shown. In the right column the coefficients are shown after the soft threshold procedure from equation (8) is applied. We see that this procedure acted with more impact at lower frequencies.



(a) Time domain. **(b)** Frequency domain

Figure 2: This figure shows the denoising method applied to the corrupted signal. In blue, the corrupted signal (CoREAS simulation plus AERA detector simulation) with added noise (measured by AERA) is shown. In orange, the signal after applying the threshold procedure in the coefficients is shown. And in dashed black the pure signal (CoREAS simulation plus AERA detector simulation) is drawn. We may see that the denoising does not cause a relevant source of distortion in frequency spectrum, neither in time domain.

3. The signal energy

Using the procedure above, we could apply this method to the complete dataset and quantify its performance in terms of signal energy reconstructions. However, it is first necessary to describe the simulation steps and specific parameters of the AERA experiment [9].

AERA uses external information from particle and fluorescence detectors to localize the radio pulses from air showers in the signal traces recorded by the antennas. This "signal search window" has a typical width of about $400ns$, i.e $\pm 200ns$ of the timestamp t_{ext} provided by the external detectors. In this time window the most probable peak of the signal t_{Hilb} is estimated. It is given by the envelope of the Hilbert Transform of the signal in this time window [11]. With t_{Hilb} a new smaller window of 200 ns ($t_{\mathrm{Hilb}} \pm 100$ ns) is selected to ensure that most of the signal energy is presented in this subregion, and also that the background present is assumed to be Gaussian [11]. Then, another subregion is selected called the "noise window". It is also a 200 ns window selected after the end of the signal region. This noise window is relevant since it will estimate the background present in the signal region. Regarding equation (4), the energy of the signal X

$$E_X = E_f - E_W, \tag{10}$$

where energy is defined as

$$E_s = \sum_n |x[n]|^2, \tag{11}$$

E_f is the energy of the signal plus noise and E_W is the energy of the noise present in the sample. But if it is assumed that a 200 ns time window is statistically likely to be a Gaussian background [11], then the energy

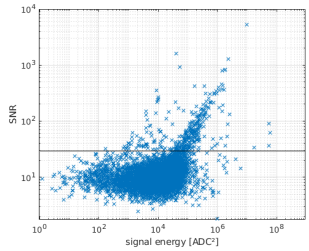
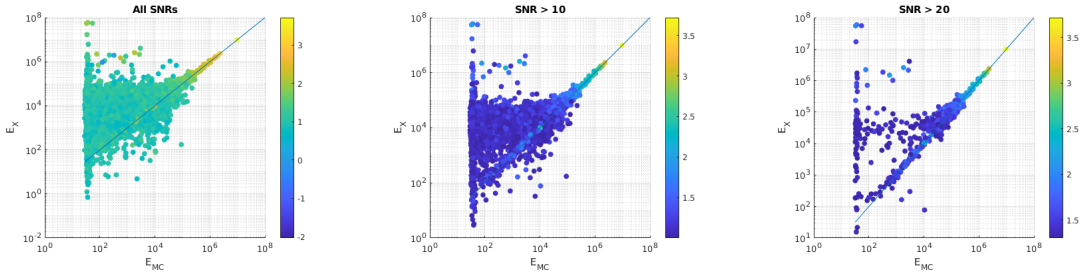


Figure 3: This figure shows the correlation between the SNR (see equation (14)) and the signal energy as defined in equation (13). In black, we show a heuristic line below which linearity is lost, which indicates when the signal energy reconstruction may start to fail (SNR ≈ 20).



(a) The correlation between measured energy and simulated energy for the full SNR range (in log scale)

(b) SNR> 10

(c) SNR> 20

Figure 4: This figure shows the evolution of trace loss due to the SNR criterion. It also shows a better accuracy in the signal energy reconstruction: it lowers the SNR of background traces and increase the SNR of traces with valid signals (given by equation (13)). As a hard cut we expect to only retrieve reliable information for $SNR > 20$.

in the background window E'_W approximates to E_W and

$$E_X \approx E_f - E'_W, \tag{12}$$

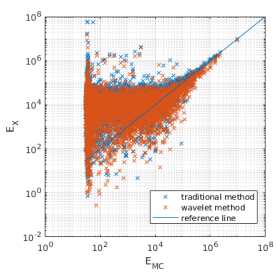
where the calculation of E_f and E'_W are straightforward from data. This method works very well for a wide range of signal-to-noise (SNR) ratio scenarios. SNR for AERA is defined as

$$\text{SNR} = \frac{A^2}{\text{RMS}^2_{W'}}, \tag{13}$$

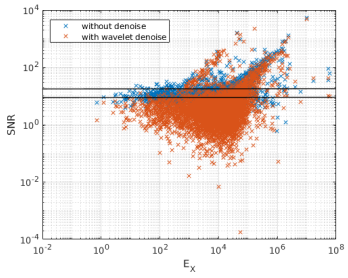
where A is the maximum value of the Hilbert envelope and $\text{RMS}_{W'}$ is the *root-mean-square* of the background window. This SNR definition has a correlation with the signal energy calculated from equation (13), but starts to lose its linearity near the value of SNR ≈ 20 (see figure 3). This means a hard cut on traces, which must satisfy SNR > 20, is applied in this analysis.

4. Analysis

The wavelet analysis effort was motivated as a potential tool for improving the SNR of traces near this noisy region, where the loss of the linearity begins. Ideally this effort will make it feasible to recover more traces that were previously discarded, i.e., to lower the hard cut of the SNR. The signal energy estimation of the wavelet denoised signal is the same as shown before, but now using the denoised version of signal



(a) Signal energy comparison (calculated from equation (8)) between standard method and with a the wavelet denoising process.



(b) The SNR distribution comparison with and without the wavelet filtering process.

Figure 5: We note a systematic decrease of the SNR value on the background region. The black lines represents a heuristic line of the limits of both distributions. Between them it is expected to retrieve data that were previously discarded.

window E'_D and a "denoised" version of the background window E'_W. More precisely,

$$E_D \approx E_f - E'_W \quad \text{where} \tag{14}$$

$$E_f = \tilde{F} \quad \text{(from equation (7))} \quad \text{and} \quad E'_W = \sum_j d_j W_B[j]. \tag{15}$$

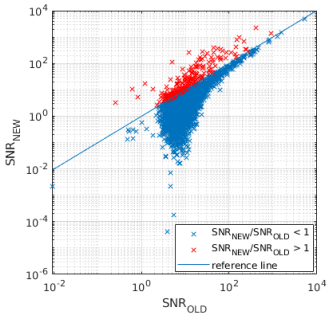
This means that we applied the wavelet filter both in the radio signal region and pure noise window. As we assume that the soft thresholding will not vanish with all noise components, applying the same method in the background window and calculating the energy is an attempt to estimate the remaining background energy present in the denoised signal. In figure 5 (a) is shown how the signal energy estimation is affected by the wavelet technique. It is possible to see that the energy reconstruction is roughly compatible with the previous energy reconstruction.

It is essential to point out that the primary point of this analysis is based on the assumption that the SNR (SNR_{new}) of the wavelet denoised signal will be more accurate. This means that it lowers the SNRs of traces in which no retrievable information is contained, such as antennas far away from the core position by reducing on average its amplitude while increasing the SNR of traces that may have information inside, such as a low amplitude pulse, by lowering part of its background while maintaining its peak amplitude. In other words, the main improvement is in the selection of traces based on their SNR. This analysis looks explicitly for traces with improved SNR. To find signal traces that were improved by the wavelet method, we compare the previous SNR (SNR_{old}) and SNR_{new} with the ratio r defined as

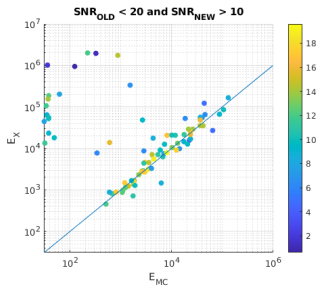
$$r = \frac{\text{SNR}_{\text{new}}}{\text{SNR}_{\text{old}}}. \tag{16}$$

5. Results

If the ratio r is greater than 1, the corresponding traces are selected (see figure 6). The wavelet technique does not improve the energy reconstruction itself (while it does not deteriorate it) since the actual method works very well (even for low SNR values), but selects traces that were previously discarded since they had an SNR that was compatible with background. Of course even in this selection we expect traces that are compatible with background, and thus we need to estimate a new lower cut on SNR_{new}. To do so, we may use the distribution of the SNR_{new} as a function of the signal energy E_X shown in figure 5 (b). An empirical line around $\text{SNR}_{\text{new}} \approx 10$ is sufficient for this particular analysis.



(a) Ratio (r). In red the traces with $r > 1$ are shown.



(b) Selection of traces previously discarded and now recovered by the SNR criterona with $r > 1$, and $\mathrm{SNR_{new}} > 10$.

Figure 6: These events have a compatible signal energy reconstruction with their simulated signal energy. The colorbar displays the $\mathrm{SNR_{old}}$ values, to show that these traces are in the range of $\mathrm{SNR_{old}} < 20$

A value of $\mathrm{SNR_{old}} < 20$, in this dataset, is the upper limit of the traces we wish to analyse, and $\mathrm{SNR_{new}} > 10$ is the lower bound.

Summarizing, to quantify the gain obtained in this analysis, we look for events that should match the criteria of $r > 1$, $\mathrm{SNR_{old}} < 20$ and $\mathrm{SNR_{new}} > 10$. In figure 6 (b) we show the signal energy of traces compared with the Monte-Carlo true simulated signal energy. These showed to be traces with roughly a correct estimation of the signal energy of the radio signal. It is important to mention that we should quantify the energy resolution achieved with this method which is not quantified in this analysis.

This dataset consists in 344 simulated showers, which had 14239 valid traces. Using the $\mathrm{SNR_{old}} > 20$ criteria we select 677 reliable traces, using the wavelet method with $r > 1$ and $\mathrm{SNR_{new}} > 10$ and $\mathrm{SNR_{old}} < 20$ we recover 86 additional traces. This implies an improvement of $\approx 13\%$ of the number of reliable traces.

6. Conclusions

The Wavelet denoising method chosen in this analysis is a standard method for denoising and considered safe and low-risk for the parameters chosen in this case. It was shown to be a potential and easy tool that could benefit the radio community to improve its signal quality.

Of course, as mentioned in this proceeding, there is room for improvement and further investigation concerning the wavelet mother, the threshold method, the energy scale resolution and some specific refinements for each experiment. Also, it would be beneficial to apply this method to other datasets of simulated and experimental data.

7. Acknowledgments

We thank the Pierre Auger Collaboration for allowing us to use the AERA detector simulation and measured AERA background data for this study. The present work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Brazil, process no. 88887.357727/2019-00, the Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), the Deutscher Akademischer Austauschdienst (DAAD) and the Karlsruhe Institute of Technology (KIT).

References

[1] Tim Huege, Radio detection of cosmic ray air showers in the digital era, Physics Reports 620 (2016) 1-52.
[2] Stéphane Mallat, A wavelet tour of signal processing, 2.ed., Academic Press, 1999.
[3] I. Daubechies, Orthonormal bases of compactly supported wavelets, Commun. Pure Appl. Math. 41, 1988.
[4] International Telecommunication Union, Radio noise, Recommendation ITU-R P.372-13, 2016.
[5] Heck, D. and Knapp, J. and Capdevielle, J. N. and Schatz, G. and Thouw, T., CORSIKA: A Monte Carlo code to simulate extensive air showers, FZKA Report 6019, 1998.
[6] T. Huege, M. Ludwig and C.W. James, AIP Conf. Proc. 1535 (2012) 128, arxiv:1301.2132.
[7] D. L. Donoho, De-noising by soft-thresholding, IEEE Transactions on Information Theory, 1995.
[8] Donoho, D. L., and Johnstone, I. M., Ideal Spatial Adaptation by Wavelet Shrinkage, Biometrika, 1994.
[9] Radio detection of cosmic rays with the Auger Engineering Radio Array Tim Huege, on behalf of the Pierre Auger Collaboration EPJ Web Conf., 2019.
[10] S. Argirò, S.L.C. Barroso, J. Gonzalez, L. Nellen, T. Paul, T.A. Porter, L. Prado Jr., M. Roth, R. Ulrich, D. Veberič, The offline software framework of the Pierre Auger Observatory, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 2007.
[11] Energy estimation of cosmic rays with the Engineering Radio Array of the Pierre Auger Observatory A. Aab et al. (Pierre Auger Collaboration), Phys. Rev. D 93, 2016.