Recognition of partial discharge signals in impaired datasets using cumulative energy signatures

Castro Heredia, L.C.; Rodrigo Mor, A.; Wu, Jiayang

DOI
10.1016/j.ijepes.2020.106192

Publication date
2020

Document Version
Final published version

Published in
International Journal of Electrical Power & Energy Systems

Electrical Power and Energy Systems 122 (2020) 106192

Contents lists available at ScienceDirect

Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

Recognition of partial discharge signals in impaired datasets using cumulative energy signatures

L.C. Castro Heredia*, A. Rodrigo Mor, Jiayang Wu

Delft University of Technology, Electrical Sustainable Energy Department, Delft, the Netherlands

ARTICLE INFO

Keywords:
Features
Impaired datasets
Energy function
High-voltage testing
Partial discharges
Classification

ABSTRACT

The problem of impaired data sets refers to data sets containing a vast majority of unwanted signals than signals of interest. With increased interest in partial discharge (PD) testing with arbitrary waveforms and transients, these kind of data sets are becoming more and more common. Traditional clustering techniques cannot be applied due to big differences in spatial densities of the existing clusters in the data set. This paper contributes a simple yet efficient technique to recognize PD signals from noise and other disturbances. The signal recognition features are based on two specific areas extracted from the cumulative energy signal (CE) of each recorded waveform. These areas weigh up the extent to which the recorded signals have a pulse-like shape. A third feature, defined as a shape factor, extracts additional metrics from the CE signal that serves the purpose of accounting for the factors affecting the computation of the proposed recognition features and threshold for data size reduction. These three CE-based features are used to create a graph from which a real PD can be spotted in large impaired data sets. The performance of this technique is tested using PD measurements from superimposed impulse tests on a 150 kV cable system.

1. Introduction

The diagnostics and monitoring of high-voltage equipment by partial discharge (PD) measurements demand robust and accurate processing tools, especially when multiple PD sources are active within the test object or when the measurements are conducted in presence of electrical disturbances and interferences. Whenever this is the case, the resulting phase-resolved PD (PRPD) patterns measured by IEC60270 PD measuring devices may be difficult to interpret.

The alternative approach to overcome these limitations has been the extraction of *features* from each PD signal recorded in time domain and the subsequent application of a clustering technique that allows for a clear classification of individual PD sources by their PRPD patterns [1].

This procedure entails several challenges. Firstly, the waveform of the acquired pulse depends strongly on the nature of the PD source, the traveling path of the PD pulse and the sensor/circuit used to measure the signal [2]. The last two factors are related to every particular measuring set-up, which make it difficult to reproduce results elsewhere. Secondly, it comes the challenge of extracting features from the signals.

A *feature* is defined as an attribute of a class and thus may be as *arbitrary* as needed. In a previous work [3], we used the value of apparent charge, energy and peak amplitude as features for PD clustering purposes. However, nothing implicitly limits what parameter may be used as a feature. Examples are, the morphological gradient in time and frequency domain [4,5] that quantifies some sort of maximum increase of the energy signal, the quantities derived from/or the wavelet decomposition coefficients [6] that may populate an endless list of references, and all the features extracted from the PD frequency spectrum. In particular, the spectral power ratio method [7] that quantifies the ratios between the area of specific frequency bands and the full FFT spectrum of the signal, and the TW-map [8] that quantifies what can be understood as a cluster based on the gravity moment of the FFT spectrum and of the equivalent time of the signal, are two techniques extensively researched and applied on field.

Just as nothing prevents a quantity from being a feature, the number of features used to describe a PD signal is only limited by computational resources. When the features data sets become so big that they are no longer easy to represent visually, carry redundant information and demand large computation power, it is common to apply any of the variety of dimensionality reduction techniques, being the principal components analysis [9] and the t-SNE technique [10] very common examples of them.

Dimensionality reduction is required to remove redundant and ineffective information and to decrease the number of features while still capturing a high portion of information [6,11]. The feature extraction

* Corresponding author.

https://doi.org/10.1016/j.ijepes.2020.106192
Received 27 January 2020; Received in revised form 13 April 2020; Accepted 11 May 2020

step is very important in this procedure and the efficiency of a classifier is highly dependent on the wellness of extracted features.

The result of this is a reduced set of features producing the largest distance between natural clusters. When this is the case, a clustering algorithm shall not find difficulties in classifying each data point to its corresponding parent cluster. The k-means [12] and DBSCAN [13] are spatial density based clustering methods (i.e., cluster methods using the average distance between neighbor points as a parameter) very popular for this purpose.

However, the clustering algorithms, and specially the kind based on spatial density, present strong limitations to discover clusters in data sets with markedly varying densities [14].

In this paper's context, varying density refers to data sets containing very little PD data points as compared to disturbance/noise data points. Data sets of this kind are also known as impaired data sets and are becoming more common due to the increasing interest in studying PD activity under impulses, AC-superimposed impulses and in general under arbitrary waveforms. Under these test conditions, many pulse-shaped noise and disturbances are produced during the firing of the impulses or during the testing period, unavoidably producing heavily impaired data sets.

In [15,16], the processing of impaired data sets has been approached by inspecting each individual event within the total data set, (which drops in efficiency when the data sets grow big).

In this paper, the cumulative energy (CE) is used as signature of the signal shape and from it, two main features are extracted, corresponding to two areas, A_{En} and A_{Ep}, delimited in the CE signal. These areas are normalized by comparing to the CE signal of a Dirac delta signal. In this approach, the ability of the features to produce dense clusters and far-apart from each other as in the conventional clustering approach is no longer the target, but their absolute values instead. The closer the value of the feature to unity the higher the extent to which the signal is pulse-shaped. Thus, our technique can recognize pulse-shaped PD signals within a large impaired data set. This can be paired to a "needle-in-a-haystack" problem where the classification of different types of PD sources (type of needles) is out of the scope of the current paper.

The description of this new technique follows the next structure. In Section 2, an overview of the data sets and the software tools used for data processing is given. In Section 3, the definition of the cumulative energy function of a signal is presented. In Section 4, the definition and calculation of the areas A_{En} and A_{Ep} is described. In Section 5, we propose a shape factor k to quantify the factors affecting the features A_{En} and A_{Ep}, and that allows to define threshold limits. Finally, in Section 6, the results of an application case are presented and discussed.

2. Datasets and tools

In this work, three impaired data sets will be used. They correspond to partial discharge measurements carried out in an environment with high electromagnetic interference [16]. The data set 1 comprises a matrix of 4713×2564, data set 2 and 3 are matrices of 5000×2564. The rows in the matrix represent the number of signals and the columns the digital samples of each signal. These data sets are made available to download in the following link [17].

The methods presented in this paper were coded in the software tool *PDflex* [18]. *PDflex* offers the interactive feature of retrieving the waveform of a signal while the user is hovering the pointer over the data in graphs. Taking into account the large amount of data processed in this paper, this tool served the main purpose of confirming visually whether a given signal is a PD or non-PD signal based on its shape.

3. Signal shape signature

From the signal processing theory, given a discrete signal $x(n)$ of N samples, its total energy E is defined as the sum of its square samples, as in (1).

$$E = \sum_{1}^{i=N} |x(i)|^2 \tag{1}$$

The cumulative energy signal E_c results from evaluating E not for the entire domain of E but for an incremental number of samples instead. Thus, E becomes the discrete signal E_c shown in equation (2).

$$E_c(n) = \sum_{i=1}^{i=n} |x(i)|^2 \quad n \leq N \tag{2}$$

The cumulative energy E_c produces monotonously increasing values and its square operand leads to step-like shape raising at the sample where the main peak of the signal arises. These two characteristics prove relevant for the signal recognition based on its shape and, therefore, hereafter the signal E_c will be used as the pulse shape signature from which recognition features will be extracted.

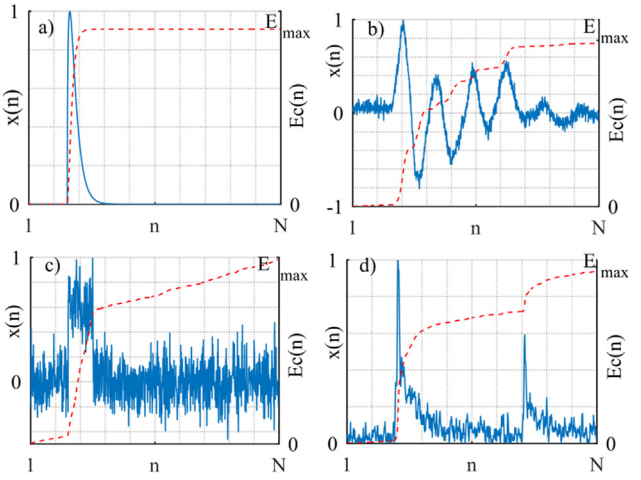

Fig. 1. Examples of the cumulative energy function (dotted line) for different type of signals: (a) fast, unipolar pulse, (b) oscillating pulse, (c) unipolar pule with low SNR, (d) pulse with a second pulse (disturbance) within the same record window.

Fig. 1 shows a collection of examples of the cumulative energy of pulse-shaped signals. As shown in Fig. 1(a), after the main peak of a very fast unipolar pulse, E_c increases step-wise reaching a plateau zone. Since E_c increases monotonously according to (2), any other significant peak or artifact appearing after the main peak will cause an abrupt increase, soaring from the precedent plateau zone as depicted in Fig. 1(b) and (d). In addition, a low signal to noise ratio (SNR) leads to a drift of the baseline of E_c as in the case of signal in Fig. 1(c).

The E_c of the different signals in Fig. 1 serves as a comparative example to show how the shape of E_c diverts from a step-like shape depending on the signal shape and SNR. In the next sections, the procedure for extracting features from E_c will be presented.

4. A_{EN}-A_{EP} graph

The shape of a PD pulse depends on the PD source type, the signal traveling path, the sensor and the measuring circuit used to measure the signal [2]. In this work an assumption is made that a PD signal, when acquired with enough bandwidth, exhibits two main characteristics:

(1) A main peak: although a PD pulse can have an oscillating shape like the example of Fig. 2, the existence of a predominant peak sharpens its pulse-like shape. In other words, the larger the main peak compared to the peaks of the oscillations the more stepped the E_c signal is.
(2) A pre-trigger zone: the main pulse peak appears only after a certain

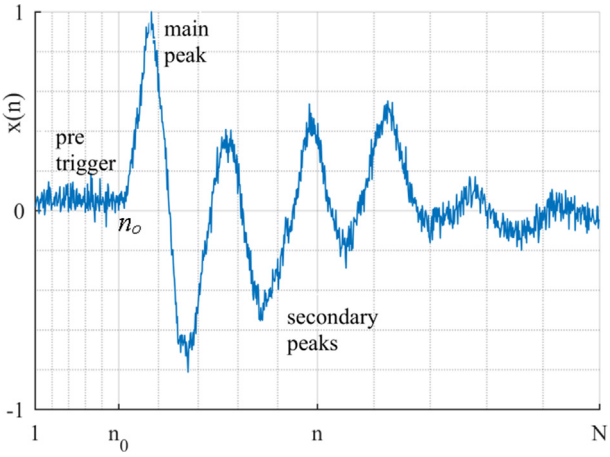

Fig. 2. Definition of the characteristics of a pulse-shaped signal.

number of samples, n_0, from the beginning of the record. In addition, the higher the SNR, the better and the more accurate is the determination of the sample n_0.

Deviations from these characteristics may occur, since noise and other electromagnetic disturbances can trigger the digital acquisition units and distort the signals in such a way that the resulting waveforms are likely to violate these two characteristics.

Based on the cumulative energy, the two areas A_{En} and A_{Ep} shown in Fig. 3(a) are proposed as quantifiers of the extent to which the signals conform to these two characteristics. These two areas result from the intersection of the E_c signal and the baseline connecting the first and last sample N of E_c, see Fig. 3(a). As seen in Fig. 3(b), this baseline represents the E_c of the background noise. The procedure to calculate A_{En} and A_{Ep} is described as follows:

Let $E_c(n)$ be redefined as the cumulative energy normalized to 1 and $g(n)$ be the normalized reference baseline, then A_{En} and A_{Ep} are calculated according to (3) and (4).

$$A_{En} = \sum_{n=n_1}^{n=n_2} (E_c(n) - g(n)) \tag{3}$$

$$A_{Ep} = \sum_{n=n_3}^{n=n_4} (E_c(n) - g(n)) \tag{4}$$

Since $E_c(n)$ and $g(n)$ are discrete functions, a step factor of 1 is considered and the areas A_{En} and A_{Ep} calculated as the sum of the operand elements.

The indexes n_1 and n_2 are the indexes of the samples at the crossing points where the $E_c(n)$ is under the baseline $g(n)$. Likewise, the indexes n_3 and n_4 are the indexes of the samples at the crossing points where the $E_c(n)$ is above the baseline $g(n)$. n_0 is determined as the sample index right at the end of the pre-trigger zone and n_p is defined as the sample index at maximum deviation of E_c with respect to the baseline.

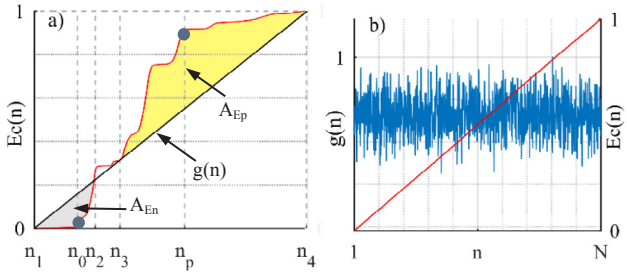

Fig. 3. (a) Definition of the areas A_{En} and A_{Ep} and their limits defined by the crossing of E_c with the baseline, (b) baseline of background noise, $g(n)$.

Therefore, n_p marks the turning point beyond which the accumulation of energy is never as rapid as it was previously.

The indexes n_0 and n_p have to be determined for each signal by means of a routine that translates the coordinate system to one in which the reference baseline is the abscissa axis. After the coordinate translation, finding n_0 and n_p becomes a problem of finding a minimum and maximum respectively. The following MATLAB pseudocode shows the process of finding n_0 and n_p implemented in this paper.

pseudocode: Detection of n_0 and n_p.

```
% translate the coordinate axis to the first sample
Ec2 = Ec - Ec(1);
x = (0:N-1);
% transform to a 2D data set of complex numbers
Ec2 = 1j*Ec2 + x;
% complex reference baseline
bl = Ec2(end);
% translation factor
rot = exp(-1j*angle(bl));
% rotate the coordinate axis by the translation factor
Ec_translated = Ecn2*rot;
% in the new coordinate system, the imaginary part of %Ec_translated becomes the
    magnitude of the sample.
[~, np] = max(imag (Ecn_translated));
[~, n0] = min(imag (Ecn_translated));
```

The output of the max and min operations corresponds to the indexes needed to determine the inputs of equations (3) and (4), while the maximum and minimum values themselves are discarded.

In addition, the values of A_{En} and A_{Ep} calculated as per (3) and (4) are complementary, which means that when one increases the other decreases simultaneously. This behavior is illustrated in Fig. 4 using as example the signal E_c of Fig. 1(a) with two different values of n_0.

If n_0 is swept over the record length, i.e. $n_0 = 1,2,\ldots, N$, then the geometrical space of the possible values of A_{En} and A_{Ep} is the grey and red-shaded regions of Fig. 5.

If the fast signal of Fig. 1(a) is now a Dirac delta signal, then its values of A_{En} and A_{Ep} follow the perimeter of the polynomial curve colored in red in Fig. 5. On the other hand, for non-ideal and discrete signals, the relation between A_{En} and A_{Ep} may be as diverse as listed in Table 1:

The aforementioned possible cases are possible because the shape of E_c depends on the shape of the signal $x(n)$, that for the aims of this paper, has no limitation of any kind. Thus, $x(n)$ can be so dissimilar as those examples in Fig. 6. This variability results in that n_2 not always has the same value of n_3 as can be seen in Fig. 6(a). In other words, E_c is monotonous, but this not imply that the rate of change is steady and smooth. Fig. 6(b) depicts a case where E_c increases steady and smoothly leading to $n_2 = n_3$ although the resulting area A_{En} is bigger than A_{Ep}. More than one area above or under the baseline is also possible. Fig. 6(c) is a case where there are two areas A_{En} under the baseline. The value of A_{En} corresponds to the area with the biggest distance of n_0 with respect to the baseline. Signals missing the pre-trigger zone are also possible as shown in Fig. 6(d). In such a case, the value of A_{En} is zero.

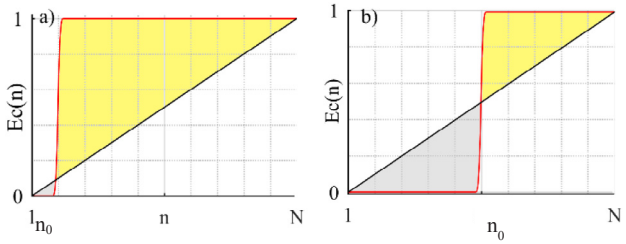

Fig. 4. Variation of A_{En} and A_{Ep} as a function of n_o.

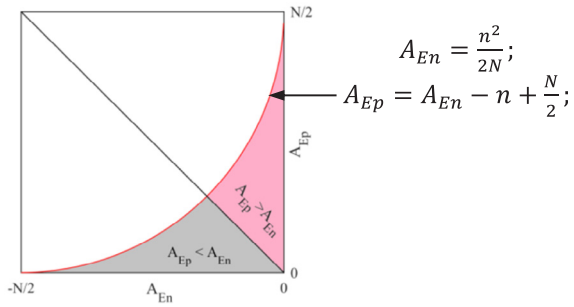

Fig. 5. Geometrical space of A_{En} and A_{Ep}.

Table 1
Example of possible values of A_{En} and A_{Ep}.

Possible result	Example
For a signal conforming strongly to the two characteristics criteria aforementioned: • $n_2 = n_3$ or at least $n_2 \to n_3$, $A_{Ep} \gg A_{En}$	Fig. 6(a)
Otherwise, for a signal poorly conforming to the criteria • A_{En} may be bigger (or smaller) than A_{Ep}. • Several areas under the reference line may exist but A_{En} corresponds to the area with the minimum peak. • A_{En} (or A_{Ep}) may be zero	Fig. 6(b) Fig. 6(c) Fig. 6(d)

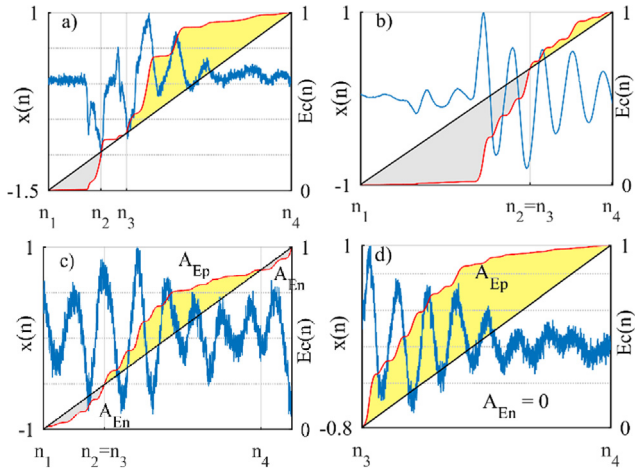

Fig. 6. Examples of signals leading to different relationships between A_{En} and A_{Ep}. (a) $A_{Ep} \gg A_{En}$, (b) $A_{Ep} \ll A_{En}$, (c) multiple areas of A_{En}, (d) $A_{En} = 0$.

4.1. Normalization

Taking into account that the range of A_{En} and A_{Ep} is [0, N/2] and that the geometrical space of Fig. 5, at best, results in a graph with data populating only a fraction of the entire graph area, then a normalization of A_{En} and A_{Ep} was necessary to improve readability of the graph. This normalization is the ratio of the areas from the signal and from a Dirac delta pulse whose E_c signal happens to have the same value of *n2*.

Consider for example the E_c for the signal of Fig. 1(b) and for a Dirac delta pulse that are shown in Fig. 7. The areas corresponding to the Dirac delta pulse are the triangle areas defined by equation (5).

$$A_{En}|_{norm} = \frac{A_{En}}{\frac{(n_2)^2}{2N}} \qquad A_{Ep}|_{norm} = \frac{A_{Ep}}{A_{En} - (n_2) + \frac{N}{2}} \tag{5}$$

In this example, the values of A_{En} for both pulses (gray-shaded areas) are similar, due to the low noise in the pre-trigger zone of the signal and therefore the ratio between them two tend to 1. From the yellow-shaded areas it can be seen that the A_{Ep} of the signal is smaller than that of the Dirac delta pulse due to the oscillations, thus the ratio drifts away from 1.

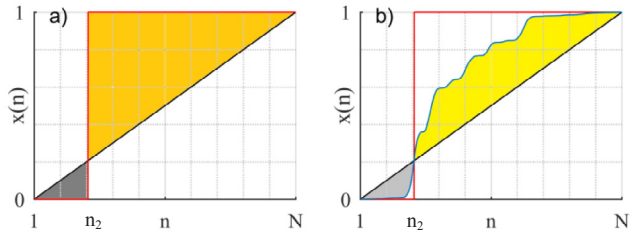

Fig. 7. Comparison of the areas A_{En} and A_{Ep} for a (a) Dirac delta pulse and (b) a signal like the one in Fig. 1(b).

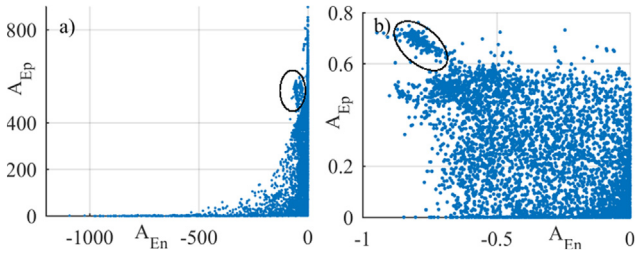

Fig. 8. A_{En} and A_{Ep} graph for data set 1 (a) before normalization, (b) after normalization.

For reference and comparison purposes, the graphs in Fig. 8 depict the results using the data set 1 before and after normalization.

It is worth mentioning that the notation A_{En} and A_{Ep} has remained unchanged after the normalization. Hereafter, A_{En} and A_{Ep} should be understood as normalized values.

Limiting the range to [0, 1] and [−1, 0] makes the interpretation of the graph straightforward: the closer the value of A_{Ep} and A_{En} to 1 and −1 respectively, the better the signal conforms to a pulse-shaped signal. Note for example, that the circled data in Fig. 8(a) is almost merged into the whole data set. After normalization, the circled data is located towards (−1,1) giving clear indication of their pulse-like waveforms and getting apart from other type of waveforms. As confirmation of this, the waveforms of the signals labeled from I to IV in Fig. 9(top) are shown in Fig. 9(I-IV).

The waveform I has a more pulsed shape than II and therefore was found at higher values of A_{Ep}. On the other hand, all waveforms located on the y-axis ($A_{En} \approx 0$) such as IV may be labelled as “low quality” signals because they fail to either have a clear pre-trigger zone or be pulse-shaped. In addition, for waveform III, $A_{Ep} \approx 0$, which results from the offset drift making the signal energy to increase continuously without any stepped increase, see Fig. 12(b).

Worth noticing that the graph of Fig. 8(b) is not meant for clustering purposes, since it is the magnitude of A_{En} and A_{Ep} that bears all importance as it sorts out the non-pulse-shaped signals at the right-bottom and the more pulse-shaped signals at the left-top of the graph.

This kind of shape-based sorting enables to define threshold levels to filter out non-PD signals. They may be hard thresholds or, as it will be described in the next section, thresholds defined from metrics of the E_c signal shape. In this paper, such a definition will be termed shape factor.

5. Shape factor k

As was shown in the section before, data yielding values of A_{Ep} and/or A_{En} close to zero can be directly labeled as low quality signals and therefore they can be removed from the data set. On the other hand, a shape factor can be used as a more sensitive threshold. In this work, we

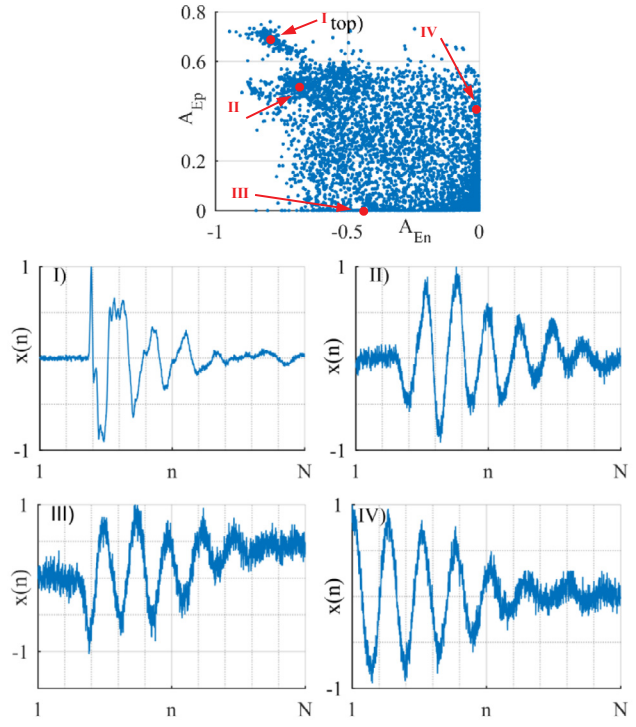

Fig. 9. Waveforms of the signals corresponding to labels in the normalized graph on top.

lump together in the shape factor k metrics extracted from the E_c signal that quantify the effects of the possible instability of n_o and n_p, low SNR, discretization errors and signal offsets on the calculation of A_{En} and A_{Ep}.

The definition of k is given by equation (6) and is illustrated in Fig. 10.

$$\begin{aligned} k &= k_1 \cdot k_2 \cdot k_3 \cdot A_{Ep} \\ k_1 &= \sin(\alpha) = \frac{\Delta E_c}{\left(\frac{n_p - n_0}{N-1}\right)}; \\ k_2 &= \Delta E_c = E_c(n_p) - E_c(n_0); \\ k_3 &= \cos(2\beta) = \cos\left(2 \bullet a\tan\left(\frac{E_C(n_0)}{\frac{n_0}{N-1}}\right)\right); \end{aligned} \tag{6}$$

The first factor termed k_1 is a quantification of how predominant the main peak of the signal is and therefore how steep is the increase of E_c. k_1 tends to 1 for pulses whose E_c increases stepped-wise like the one in Fig. 1(a). In signals with low SNR, like the ones in Fig. 1(c) and Fig. 11(a), although the energy increases stepped-wise ($k_1 \rightarrow 1$) the increase itself may be very low as illustrated in Fig. 11(b). In such a situation, by adding a second factor, $k_2 = \Delta E_c$, signals with low SNR are prevented from scoring a high shape factor.

The third factor is k_3, that quantifies indirectly the background noise level of the signal in the pre-trigger zone. Gaussian noise leads to an E_c that remains close to zero from the beginning of the record until

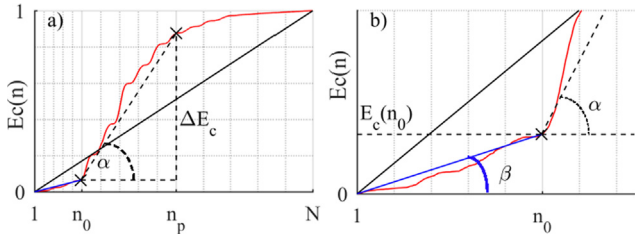

Fig. 10. (a) Metrics in the definition of the shape factor k, (b) zoom-in to illustrate the definition of the angle β.

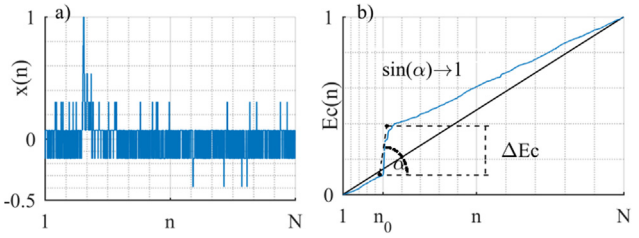

Fig. 11. (a) Signal with low SNR, (b) resulting in a stepped but small increase of energy.

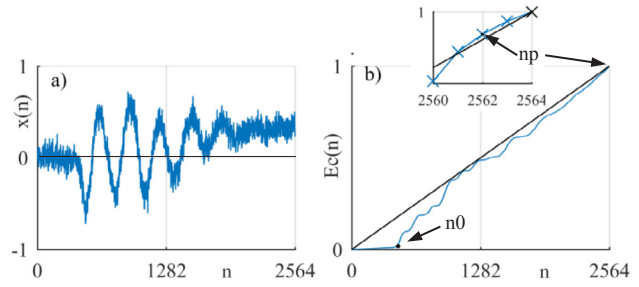

Fig. 12. (a) signal with offset, (b) resulting in a continuous increase of energy.

n_0, which means that cos(2β) tends to 1. Otherwise, E_c drifts closer to the background noise baseline, thus cos(2β) tends to 0 as can be seen in Fig. 10.

Finally, the value of A_{Ep} itself is added to the shape factor in order to balance out high values of k_1, k_2 and k_3 that can result from signals with offsets or due to discretization errors.

Fig. 12 is an example of a signal with a drifted baseline (offset), therefore its E_c increases continuously resulting in an overflowed index of n_p. For this signal, $k_1 = 0.76552$, $k_2 = 0.98144$ and $k_3 = 0.97907$, values that are significantly high despite of the offset of the signal. The zoom-in in Fig. 12(b) shows that the estimation of $A_{Ep} = 0.01$ is very coarse because this area happened to be delimited within the last 4 (out of 2564) samples of the record thus leading to significant discretization error.

Thus, when A_{Ep} is added to k, only the signals with a high energy content above the normalized reference baseline will score a higher shape factor.

By this approach, for the data set 1 only 94 out the total 4713 signals scored a $k > 0.6$, while most of the signals score a shape factor close to zero as represented in the histogram in Fig. 13.

The shape factor is a 1D-dimension vector that in addition can be used as the color map in the A_{Ep}-A_{En} graph of Fig. 14. It can be confirmed that pulse-shaped signals produced simultaneously higher values of both k and A_{Ep}-A_{En}. This result was also confirmed by retrieving and checking the waveforms of all the signals within the circles.

Fig. 14 also serves the purpose of illustrating the enhancement of the shape factor by adding A_{Ep} that was discussed before. Note that in Fig. 14(a), the green shades spread over the entire graph, meaning that

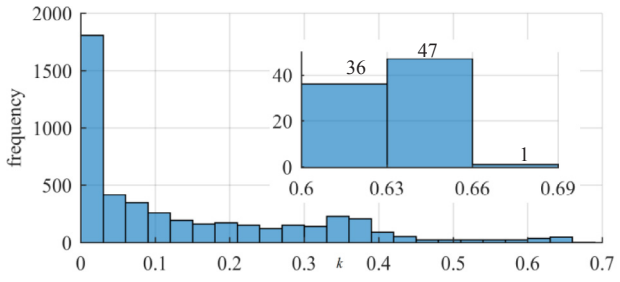

Fig. 13. Histogram of the k values for data set 1.

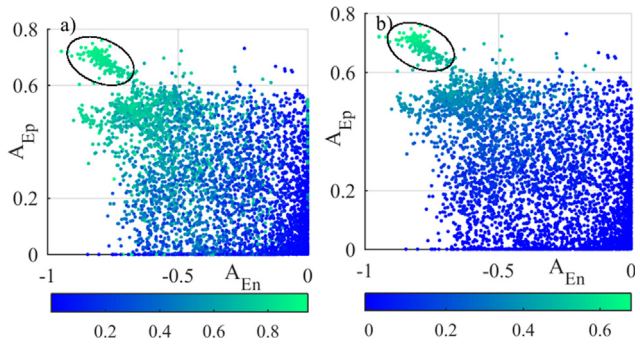

Fig. 14. (a) Color map representing $k_1 \cdot k_2 \cdot k_3$, (b) color map representing $k_1 \cdot k_2 \cdot k_3 \cdot A_{Ep}$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

non pulse-shaped signal can still score relatively high values of $k_1 \cdot k_2 \cdot k_3$. However, as depicted in Fig. 14(b), adding A_{Ep} corrects this situation to a large extent. Thus, the brightest shades of green call attention outright on the more pulse-shaped signals, also acting as a more discriminative threshold than a hard threshold.

6. Application case

In this section the proposed methodology is applied to laboratory measurements on a high-voltage cable as an application case.

The data sets 2 and 3 are measurements collected from superimposed impulse testing on a 16-metre long, 150 kV cross-linked polyethylene (XLPE) extruded cable system, [16].

The test circuit used was that of Fig. 15, which has the purpose of subjecting the high-voltage cable to a testing voltage of 50 Hz AC, and firing impulse transients at around the positive peak of the AC voltage. The collection of the data was done by a Tektronix MSO Series 5 oscilloscope that recorded the signals from two HFCT sensors at both ends of the cable joint. In addition, an artificial defect at the cable joint was created in order to produce surface discharges, which had an inception voltage of 48 kV_{rms}. The description of the test object, the testing circuit and the test program is given below with the sole purpose of presenting the means by which impaired data was collected. Nevertheless, the present contribution focuses only on the application of the A_{Ep}–A_{En} graph on the data sets.

The data set 2 is a case where the AC voltage was set far below the

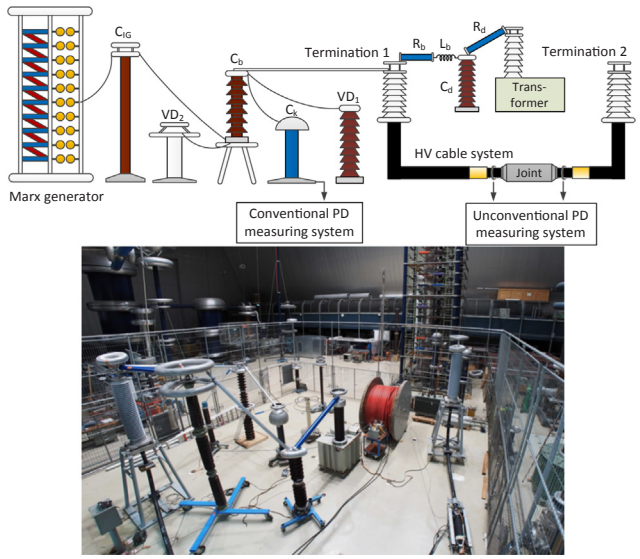

Fig. 15. Superimposed impulse test circuit used for acquiring the data sets 1 and 2. More details can be found in [16].

PD inception voltage, thus leading to a very few PD signals initiated by the impulse voltage. In data set 3, the AC voltage was close to the PD inception voltage, and the testing program stretched to roughly 12 min within which 6 impulses were applied. These resulted in an increase of the PD activity. Some of the test parameters and testing voltages are summarized in Table 2.

The testing program followed the next sequence. The AC voltage is set. At this moment, the FastFrame mode of the oscilloscope is turned on. Next, the Marx impulse generator is switched on. After the charging time of the generator the impulse is fired. The AC voltage is still on and the charging and firing of the impulse generator is repeated as much as needed. The acquisition stops when the number of recorded frames reaches 5000. Each frame length is 2 μs sampled at 1.25 GS/s.

6.1. Data analysis of case 1

The analysis of the case 1 is shown in Fig. 16. Fig. 16(a) shows the peak value of the recorded signals during the 50 s test duration. The signals during the first seconds of the record correspond to the instant at which the Marx generator was switched on and subsequent disturbances during the charging time. After roughly 45 s, the generator is charged to the set voltage and the impulse is fired on top of the AC voltage, leading to the signals after 45 s in Fig. 16(a). Most of the acquired signals are disturbances that triggered the acquisition unit. However, due to the extremely high waveform capture rate of the oscilloscope, actual PD signals also triggered the acquisition.

The imbalance problem of the data captured by this test circuit is shown by the PRPD of Fig. 16(b), where the small PD activity is buried into the disturbance signals. The disturbances due to the impulse firing can be located around 90° (AC voltage peak). The disturbances due to the generator charging and after the impulse firing are phase-independent therefore they appear randomly in the PRPD pattern. After computing the A_{Ep}-A_{En} graph, a bundle of 4 signals out of 5000 were spotted having the highest values towards 1 and the highest values of k as shown in Fig. 16(c). An example of the waveform of these signals is displayed in Fig. 16(d). This set of 4 signals were classified as PD signals based on its distinctive waveforms and opposite polarity. As explained in [16], when a PD signal is originated in the cable joint, the output of the HFCT sensor at each end of the cable joint is similar in amplitude but with opposite polarity. Any signal reaching the sensors from outside the cable joint will result in both sensors output signals having the same polarity. Visual inspection of the waveforms and their polarity of the data points out of the circle in Fig. 16(c) confirmed that those signals had different waveforms as the ones classified as PD signals as well as equal polarity (non-PD signals), thus further validating the results of the A_{Ep}-A_{En} method.

6.2. Data analysis of case 2

In the application case 2, with higher AC voltage level and more impulses being applied, the PD activity increased. In this case, the test duration was 754 s and 6 impulses were applied around the AC voltage peak as shown in Fig. 17(a). The firing of the impulses is spotted around 90° in the PRPD pattern of Fig. 17(b). The data after 180° were also linked to the disturbances created by the impulses based on the non-pulsed shape of their waveforms. The graph A_{Ep}-A_{En} in Fig. 17(c) reveals a larger amount of signals towards 1 and −1 in the graph in correspondence to the higher AC test voltage and larger number of impulses. This is further seen when comparing to the previous study case, the PD signals soared from 4 to roughly 238 and this was enough to draw the PRPD pattern of Fig. 17(d). Note that this pattern and that of the Fig. 18 obtained from the defect when only pure AC voltage was applied are similar. One might expect that with more PD pulses (> 238 signals) the matching of the two patterns would become clearer. Thus, this surface PRPD pattern served as confirmation of the rightfulness of the recognition along with the individual pulse inspection as shown in

Table 2
Testing voltages and acquisition Parameters.

	Number of signals	Sampling rate [GSa/s]	RecordLength (us)	ElapsedTime [s]	AC [kV]	Impulse[kV]
Case 1 (data set 2)	5000	1.25	2	50	35	85
Case 2 (data set 3)	5000	1.25	2	754	44	120

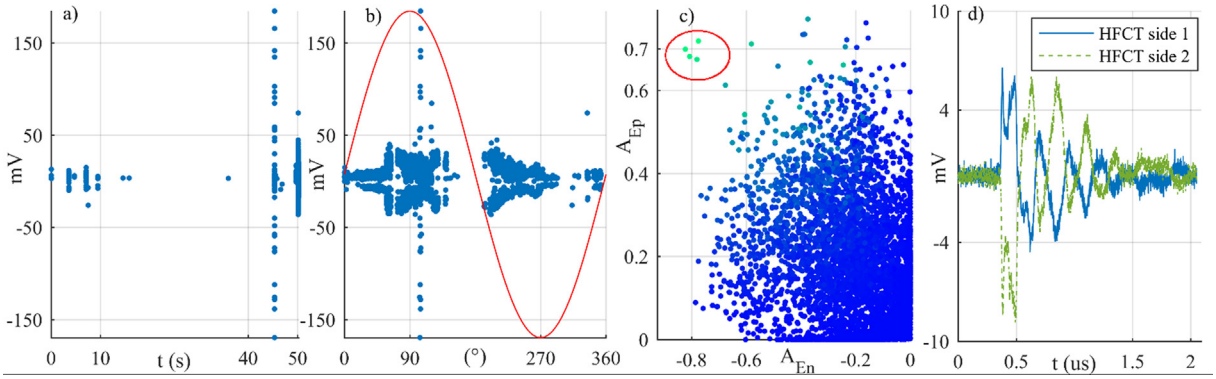

Fig. 16. Application case 1. (a) PD peak voltage in time, (b) PRPD pattern of the whole data set 2, (c) A_{Ep}-A_{En} graph, (d) PD waveforms.

Fig. 16(d), [16].

7. Conclusion

The problem of impaired data sets cannot be approached by traditional clustering techniques. This paper contributed a simple and yet efficient alternative to recognize pulse-shaped signals, namely PD signals, within large data sets containing a vast majority of signals of another shape.

In our approach the resemblance of the energy signal E_c of a recorded waveform to that of a Dirac delta signal was quantified by the features A_{Ep} and A_{En} extracted out of the normalized cumulative energy signal.

Normalization was applied to these features so that their domain would be limited to ± 1, making it easier to read and interpret the graph. Such a characteristic becomes the biggest strength of this method because frees the user from any need for specific or pre-knowledge on the data sets.

In addition, a shape factor k was proposed to quantify the factors affecting the calculation of A_{Ep} and A_{En}. k is a 1D-dimension vector added to the A_{Ep}-A_{En} graph as the color map.

When used in this form, the color map helps in calling the attention outright on the pulse-shaped signals. It also serves the purpose of thresholding the data set. Thus, helping to discard non-PD signals from the data set.

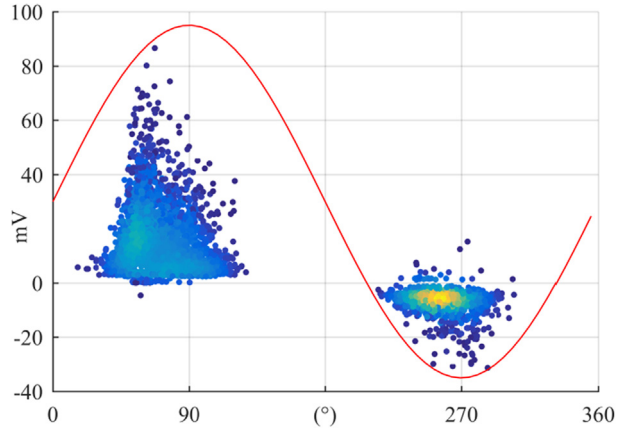

Fig. 18. Example of surface PRPD from the defect under 46kVAC test voltage.

When the amount of non-PD signals is large compared to actual PD signals, extracting high dimensionality features may be expensive from the computational point of view and may turn the analysis very complex. In this regard, this tool proved to be very simple because only a few arithmetic features are extracted from the cumulative energy, which in turn it is a simple concept.

Some of the disadvantages are that the areas A_{Ep} and A_{En} are

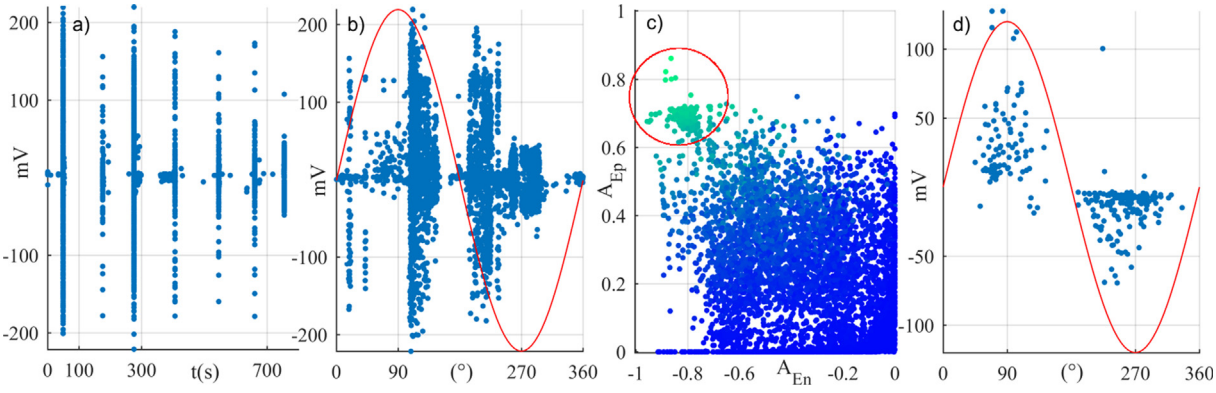

Fig. 17. Application case 2. (a) PD peak voltage in time, (b) PRPD pattern of the whole data set 3, (c) A_{Ep}-A_{En} graph, (d) PRPD pattern of clustered data.

affected by the sampling of the signal. The higher the sampling rate, the better the determination of the crossing points n_x between the E_c and the baseline. Likewise, E_c is affected by the offset of $x(n)$. Therefore, a conditioning stage to remove the offset of signals may be required before the application of this methodology.

CRediT authorship contribution statement

L.C. Castro Heredia: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Writing - review & editing, Validation. **A. Rodrigo Mor:** Writing - review & editing. **Jiayang Wu:** Investigation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Kreuger FH, Gulski E, Krivda A. Classification of partial discharges. IEEE Trans Dielectr Electr Insul 1993;28(6):917–31.
[2] Alvarez F, Ortego J, Garnacho F, Sanchez-Uran MA. A clustering technique for partial discharge and noise sources identification in power cables by means of waveform parameters. IEEE Trans Dielectr Electr Insul 2016;23(1):469–81.
[3] Rodrigo Mor A, Castro Heredia LC, Muñoz FA. New clustering techniques based on current peak value, charge and energy calculations for separation of partial discharge sources. IEEE Trans Dielectr Electr Insul 2017;24(1):340–8.
[4] Zhu M-X, et al. Partial discharge signals separation using cumulative energy function and mathematical morphology gradient. IEEE Trans Dielectr Electr Insul 2016;23(1):482–93.
[5] Zhu M, Liu Q, Xue J, Deng J, Zhang G. Self-adaptive separation of multiple partial discharge sources based on optimized feature extraction of cumulative energy function. IEEE Trans Dielectr Electr Insul 2016;24:246–58.
[6] Ali NHN, Goldsmith W, Hunter JA, Lewin PL, Rapisarda P. Comparison of clustering techniques of multiple partial discharge sources in high voltage transformer windings. 2015 IEEE 11th International Conference on the Properties and Applications of Dielectric Materials (ICPADM). 2015. p. 256–9.
[7] Ardila-Rey J, Martínez-Tarifa J, Robles G, Rojas-Moreno M. Partial discharge and noise separation by means of spectral-power clustering techniques. IEEE Trans Dielectr Electr Insul 2013;20(4):1436–43.
[8] Cavallini A, Contin A, Montanari GC, Puletti F. Advanced PD inference in on-field measurements. I. Noise rejection. IEEE Trans Dielectr Electr Insul 2003;10(2):216–24.
[9] Uriel MM. A new feature space for partial discharge signal separation based on DWT coefficient variance. J Electrical Eng 2018;6:18–27.
[10] Nimmo RD. Methods for Wavelet-based autonomous discrimination of multiple partial discharge sources. IEEE Trans Dielectr Electr Insul 2017;24(2):1131–40.
[11] Weizhong Yan, Goebel Kai F. Feature dimensionality reduction for partial discharge diagnosis of aircraft wiring. In Proceedings of the 59th meeting of the society for machine failure prevention technology, MFPT '05; 2005. p 167–176.
[12] Martínez-Tarifa JM, Ardila-Rey JA, Robles G. Automatic selection of frequency bands for the power ratios separation technique in partial discharge measurements: Part I, fundamentals and noise rejection in simple test objects. IEEE Trans Dielectr Electr Insul 2015;22(4):2284–92.
[13] Zhu M, Wang Y, Chang D, Zhang G, Shao X, Chen J. Discrimination of three or more partial discharge sources by multi-step clustering of cumulative energy features. IET Sci Meas Technol 2019;13(2):149–59.
[14] Levent Ertöz VK, Steinbach Michael. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. Proceedings of the 2003 SIAM international conference on data mining. 2003. p. 47–58.
[15] Tailor N, Edin H, Nikjoo R. Effect of superimposed impulses on AC partial discharge characteristics of oil-impregnated paper. IEEE Trans. Dielectr. Electr. Insul. 2016;23(6):3602–11.
[16] Wu J, Mor AR, Smit JJ. The effects of superimposed impulse transients on partial discharge in XLPE cable joint. Int J Electr Power Energy Syst 2019;110:497–509.
[17] PDflex Examples. Available at: http://pdflex.ewi.tudelft.nl/examples [accessed 13rd of April 2020].
[18] PDflex – Signal Processing Tool. Available at: http://pdflex.ewi.tudelft.nl [accessed 13rd of April 2020].

Luis Carlos Castro was born in Cali, Colombia in 1986. He received the Bachelor and PhD degree in electrical engineering from Universidad del Valle, Cali, in 2009 and 2015 respectively. Currently, he is a post-doc in the Electrical Sustainable Energy Department at Delft University of Technology, in Delft, The Netherlands. His research interests include high-voltage technology, partial discharge testing, accelerated aging of stator insulation and monitoring and diagnostic tests.

Armando Rodrigo Mor is an Industrial Engineer from Universitat Politècnica de València, in Valencia, Spain, with a Ph.D. degree from this university in electrical engineering. In Spain, he joined and later led the High Voltage Laboratory and the Plasma Arc Laboratory of the Instituto de Tecnología Eléctrica in Valencia, Spain. Since 2013 he is an Assistant Professor in the Electrical Sustainable Energy Department at Delft University of Technology, in Delft, Netherlands. His research interests include monitoring and diagnostic, sensors for high voltage applications, high voltage engineering, space charge measurements and HVDC.

Jiayang Wu was born in Nanjing, China in 1988. She received the BSc degree in electrical engineering from the Southeast University, Nanjing, China, in 2010, and the MSc degree in electrical power engineering from the RWTH Aachen University of Technology, Aachen, Germany in 2013. She is currently a Ph.D candidate in the Electrical Sustainable Energy Department at Delft University of Technology, Delft, The Netherlands. Her current research focuses on the effects of transients on the high voltage cable systems.