QSSI: A new Similarity Index for Qualitative Time Series.  
Application to classifying of voltage sags.

Fco. I. Gameroa, Joaquim Meléndeza, Joan Colomerb

cControl Engineering and Intelligent Systems group - eXiT  
Institute of Informatics and Applications  
University of Girona (Spain)  
(gamero, quimmel, colomer)@eia.udg.es

Abstract: This work is focused on defining and implementing a new similarity criterion for sequences of symbolic representations. The proposed algorithm returns a normalized index related to the degree of matching between sequences of qualitative labels. Performance of this method has been tested in the classification of voltage sags (transient reduction of voltage magnitude) gathered at 25kV distribution substations. The objective is to assist monitoring systems in locating the origin of such disturbances in the transmission (HV) or distribution (MV) system. The promising classification accuracy achieved when this method was used with test data suggests that the presented algorithm could be applied satisfactorily and confirms its utility in classification approaches.

Keywords: Fault Location, Power Distribution Systems, Time Series, Similarity, Qualitative Representation

1. Introduction
The motivation of this work is the interpretation of continuous signals involved in process monitoring as sequences of qualitative states in order to develop monitoring strategies working in higher abstraction level. Expert System based monitoring applications have not been fully successful because of interface difficulties between measured process signals and expert knowledge bases (KB); that is, they lack of good human-like interpretation of signals. For this reason, it is necessary to develop new tools to deal with signals coming from process, allowing a more abstracted view of them and
providing criteria and metrics to establish relations and comparisons among them. The goal is to deal with qualitative representations of signal behaviour (tendencies, oscillation degrees, alarms, degree of transient states...) instead of the numeric time series given by sensors and data acquisition systems. Then, it is possible to categorize the time series into a finite number of classes or patterns, enough to represent the system behaviours under study. Thus, the implementation of an intelligent monitoring system will consist in comparing the patterns of process variables with others registered in a case base and previously diagnosed. Unfortunately, time-series inherently contain inaccurate information since measurements are obtained and transferred with imperfect instruments. Some of the problems encountered when comparing time-series are: noise, offset translation, amplitude scaling, and time misalignments among others. The problem of pattern matching in an efficient and precise way in series of data is a non-trivial problem and it is important in a wide variety of applications. Thus, we present a new algorithm, named Qualitative Similarity Sequence Index (QSSI) for the comparison of qualitative sequences. The algorithm returns a normalized similarity value when two qualitative sequences are compared. Based on the QSSI similarity principle, a methodology for classification of voltage sags (short duration voltage reductions) in the area of Power Quality Monitoring will be presented as example of application.

1.1 Related work
The study of qualitative sequences has been often reported by researchers as Qualitative Trend Analysis (QTA) ({Venkatasubramanian, 2003 #332}) or Qualitative Shape Analysis ({Rengaswamy, 2001 #238}). The variation of a process variable with time is called the trend of that variable. The trend of a process variable is composed by simple shapes or symbols, also called primitives by some authors ({Rengaswamy, 1995 #240}). The trend analysis approach is based on monitoring the ordered set of shapes that abstract the trend of each process variable.

The granularity of the shape information also plays a crucial role on variable description. Traditionally, the shape of a trend has been classified into constant, sharp decrease, linear increase, steps, etc. ({Galati, 2006 #388}), and notated by an alphabet. First-order trends are widely used since they are simpler and more robust to noise. {Sundarraman, 2003 #72} describes a process variable as an ordered collection of enhanced atoms. An enhanced atom consists of a first-order shape, the time duration for which that shape is manifested, and the variable magnitudes at the beginning and end of
the shape. This trend including quantitative information is called enhanced trend. Then, they defined a distance measure as the maximum value obtained from matching degree by shape, magnitude and duration for two trends. Synchronization between real and theoretical trends is accomplished by a shape-matching degree table.

Motivated by the concept of episode, defined basically as the set [symbol, duration], some works have presented a formal language for trend representation where the episodes are characterized by seven basic primitives over which first and second derivatives do not change their sign ([Cheung, 1990 #215]; [Janusz, 1991 #230]). Using these primitives, [Dash, 2001 #227] proposed a novel approach to automate the identification of process trends based on an interval-halving procedure. The idea is to fit polynomials to the data and to identify the seven primitives. Later, in [Dash, 2003 #12], the same identification method is employed and two similarity indices are defined based on the aligned qualitative sequences. The first one is just based on the similarity between primitives according to a similarity matrix, and the second one takes into account the time as a weighting factor. Finally, [Maurya, 2007 #418] improves these similarity measures adding shape-based and magnitude-based similarities. In this way they consider the normalized area under primitives and penalize the difference in the magnitude.

The same representation ([Cheung, 1990 #215]) is used in [Wong, 2001 #245] as a basis for classification of tendencies. First, signals are filtered by means of wavelets and next the triangular episodes are obtained. Next, logical fuzzy is used in order to convert the quantitative values of magnitude and duration related to the episodes in another symbolic representation including ’small, medium, large’ symbols representing these quantitative values. As a result the 7 initial symbols become 57. Then classification of data is carried out by means of HMM (Hidden Markov Models).

[Charbonnier, 2007 #432] splits the data into linear segments and classify the latest segments into seven shapes: Steady, Increasing, Decreasing, Positive or Negative Step, Increasing/Decreasing or Decreasing/Increasing Transient. After that, they transform the obtained shapes into 3 types of episodes defined as {steady, increasing, decreasing}, duration, extreme values}. The approach is used to recognize specific situations for Intensive Care Unit patient monitoring by means of simple rules.
In (Meléndez, 2001 #235) an extension of the concept of episode expands the qualitative representation to both the qualitative and quantitative context considering any function of the signal as a basis for representation. This means that it is possible to build episodes, i.e. symbolic representations, according to any feature extracted from the variables (Gamero, 2006 #89). According to this formalism, an episode is a set of numerical and qualitative features. The set includes the time duration that determines the temporal extension of episodes, the qualitative state obtained from the set of the extracted features, and other auxiliary characteristics useful for exploitation and explanation purposes such as classification, regression or reasoning. This definition of episode has been used in this work to obtain a qualitative description of time series (RMS voltage waveforms) as a sequence of episodes.

An important problem encountered when comparing time-series is time misalignment, that is, the unmatching of two sequences due to a distortion (expansion or compression) in the time axis of one or both sequences. A method that tries to solve this inconvenience is Dynamic Time Warping (DTW) (Sakoe, 1978 #61; Silverman, 1990 #64) that uses dynamic programming to align time series with a given template so that total distance measure is minimised. DTW distance between time series is the sum of distances of their corresponding elements. A lot of work has been done to adopt time warping distance, since it is a measure that can measure the similarity between time series that have distortion in the time domain. With regard to qualitative sequences and DTW, an example is presented in (Colomer, 2002 #11). In (Zhou, 2005 #47) another variation of similarity search using time warping is proposed. The technique called segment-wise time warping (STW) combines a natural time scaling transformation through stretched segments and the DTW method. The time complexity of STW is quite high, therefore they need to use an index structure based on a lower bound function. Shou 2005, #363 adopt a multi-step processing technique for similarity queries using DTW. They decompose each sequence into a small number of segments and then they use a version of the Segmented Dynamic Time Warping (SDTW) proposed in (Keogh, 1999 #232). Also the segmented sequences are used to derive lower bounds and finally they develop an index and a multi-step technique that uses the proposed bounds and performs two levels of filtering to efficiently process similarity queries.
Another symbolic representation method called SAX (Symbolic Aggregate approxXimation) is presented in [Lin, 2003 #43]. First, the normalized time series are transformed into Piecewise Aggregate Approximation representation (PAA) (Keogh et al., 2000) and each PAA representation is symbolized into a discrete string. They define a distance measure based on looking up the distances between each pair of symbols and their Euclidean distance.

1.2 Paper organization
This paper is structured as follows. In Section 2, the Qualitative Sequence Similarity Index (QSSI) is presented as a new similarity measure for qualitative sequences. The proposed method will be applied in Section 3 to the classification of sags, originated in the HV and MV power systems. In the illustration example the concept of qualitative representation and episode is used to obtain a useful representation of sags, so the transformation of raw data into qualitative sequences is described too. Finally, main conclusions and some points about future work are given in Section 4.

2. Qualitative Sequence Similarity Index (QSSI)
The methods for computing similarity in the domain of strings over finite alphabets are the basis of the similarity index presented here. A way to represent time sequences is by means of an alphabet, which is often characteristic of the application. This conversion of time series into qualitative implies a discretization in both time (resampling) and magnitude (alphabet). Thus, each pattern is represented by a string of episodes identified by means of a pattern grammar. Then, the sequence comparison problem is reduced to quantifying the degree of similarity or, equivalently, the distance between qualitative sequences. To do this, however, the sequences to be compared must first be aligned.

An algorithm for sequence alignment will, in general, attempt to identify regions of high similarity by maximizing a certain score that quantifies the similarity between the sequences in an optimal (or suboptimal) alignment. Most of the alignment algorithms are based on the technique of dynamic programming (Gusfield, 1997). These algorithms search optimal solutions for a given scoring or cost function. Usually, the scoring
function is based on the definition of a set of operations, basically insertions, deletions and substitutions, with a certain cost associated with each one. The algorithm searches the optimal alignment based on the application of these operations. Next, a criterion is needed to judge whether the two sequences share a sufficient degree of similarity. However, the majority of alignment algorithms lack normalization that would appropriately rate the obtained score with respect to the length of the sequences or will serve as a reference index for quantifying significance of the comparison. The proposed method, named Qualitative Sequence Similarity Index (QSSI), represents a normalized measure of the maximum similarity score between two strings, the source string S and the target string T. The analysis of similarity between two sequences was defined in terms of traces (Sankoff and Kruskal, 1983). A trace from S to T consists of a set of lines used to link the same elements that exist in S and T, as represented in Fig. 1. Each pair of elements connected by a line constitutes a match.

Fig. 1 A possible trace for two sequences.

The QSSI algorithm is based on the Needleman–Wunsch algorithm for global sequence alignment (Needleman and Wunsch, 1970). While only a sub-sequence of each of the sequences is aligned in the local alignment procedure, in a global sequence alignment the entire length of both sequences must be aligned. A global alignment is appropriate for comparing series that are expected to share similarity over the entire length. This is the case of sag analysis. The idea behind the approach is to build up an optimal alignment and then to calculate the similarity using the total score of the aligned elements. The QSSI will be performed in three stages:

- Obtaining matches (pairs of elements) using the Needleman–Wunsch algorithm.
- Minimization of the temporal misalignment, that is, the selection of the optimal trace (with highest score).
- Calculate the normalized similarity index.

The following subsections are devoted to explaining these steps in detail.
2.1 Obtaining matches

Given two sequences S and T, with longitude m and n respectively

\[ S = s_1, s_2, \ldots, s_i, \ldots, s_m \quad T = t_1, t_2, \ldots, t_j, \ldots, t_n \]

a temporal alignment is determined by representing in a two-dimensional array, or alignment matrix, all possible pair combinations that can be constructed from the two string sequences. Thus, every possible comparative analysis of both sequences is represented by a pathway W of length K through the alignment matrix (Fig. 2).

\[ W = [w_1, w_2, \ldots, w_k, \ldots, w_K] \]

where \( w_k = [i_k, j_k] \) and \( i_k \) and \( j_k \) denote the time index of each element. Notice that each marked cell in the alignment matrix constitutes a match. The following constraints have to be considered in the construction of the pathway, W, suitable for the alignment of sequences:

- The pathway must to be monotonic. That is, \( i_{k+1} \geq i_k \) and \( j_{k+1} \geq j_k \). This also implies that the lines from each pair of elements cannot cross each other. Fig. 3 shows an example in which three different analyses could be extracted from the previous sequences, but the third is the pathway for which the sum of the assigned cell values is largest. It is noteworthy that this last alignment is obtained through the path in Fig. 2.

- Multiple connections are allowed. An element can have more than one trace (Fig. 4).

- An “N”-shaped configuration is not allowed. If a term has multiple lines, the terms at the other ends of these lines must not have multiple connections (Fig. 5). This is also a common constraint in speech recognition.

Therefore, not all matches are suitable for comparing strings. The implementation of the Needleman–Wunsch algorithm returns only those candidates to traces that fulfil the constraints enumerated above. In this way any relationship representing permutations is avoided, since this destroys the physical significance of a sequence. Also the alignment of the sequences allows a chain to overlap or enclose the other one, that is, for two sequences S and T with longitude m and n respectively and \( m > n \), the possible traces can
be aligned as shown in Fig. 6. The next step will be to choose the optimal temporal alignment based on an accurate analysis of diagonals.

Fig. 2 Contributors to the optimal alignment represented by a pathway $W$ of length $K=4$. $W = [i_k, j_k]_{k=1...4} = [(2, 3), (4, 4), (5, 5), (6, 6)]$

Fig. 3 A pair with three different alignments.

Fig. 4 Example of multiple connections.
2.2 Minimization of the temporal misalignment

A diagonal $dg$ in the alignment matrix is defined as the cells with a position $(i,j)$ where $j-i=dg$. All the matches over a diagonal $dg$, are perfectly aligned in time, whereas matches out of the diagonal indicate the presence of a time misalignment. The goal in this step is to identify the diagonal which minimizes the temporal misalignment between sequences $S$ and $T$.

The time misalignment, represented by oblique lines in Fig. 7, is produced by two basic effects: the existence of symbols without any matching between other matched symbols and the different lengths of matched symbols. Due to their physical significance, these non-matched elements and their durations cannot be ignored. Nevertheless, the importance of the duration depends on the application.

In this way, the similarity obtained between symbols depends on the size (duration) and position of the intervals assigned to each symbol. For example, Fig. 7 displays from left to right the alignments related to diagonals ‘-1’, ‘0’ and ‘1’ corresponding to the example shown in Fig. 2. The optimal diagonal is the one labelled as ‘0’, since this alignment minimizes the global temporal misalignment.

The optimal diagonal $dg_{opt}$ is obtained by calculating the distance of every match to the diagonal. The diagonal that has the minimum accumulated distance, $f(dg)$, is the

\[
\begin{array}{cccccccc}
  a & a & & a & a & & & \\
  \hline
  b & b & & b & b & & & \\
\end{array}
\]

Fig. 5 The “N”-shaped constraint.

\[
\begin{array}{cccccccc}
  a_1 & a_2 & \ldots & \ldots & a_{m-1} & a_m \\
  b_1 & \ldots & b_n \\
  b_1 & \ldots & b_n \\
  \vdots \\
  b_1 & \ldots & b_n \\
\end{array}
\]

Fig. 6 Possible alignments of two sequences.
optimal one. Only the diagonals containing matches are considered in computing the accumulated distance:
\[
d_{\text{opt} \in Dg} = \arg \min_{dg \in Dg} f(dg)
\]
with
\[
f(dg) = \sum_{k=0}^{K} |j_k - i_k - dg|
\]
and \(Dg = \{dg \in [-1, n-1] \mid dg = j_k - i_k\}\)

Fig. 7 Alignments produced by diagonals -1, 0 and 1.

2.3 Calculating the similarity index

Once the sequences are aligned in time, the algorithm calculates a normalized value representing similarity. The index of similarity for QSSI is defined as:
\[
QSSI(S,T) = \sum_{k=1}^{K} Sim(i_k, j_k) * \frac{p(i_k, j_k)}{\max(m,n,K) * p_0}
\]

where \(Sim(i_k, j_k)\) is the local similarity between every match. The function \(p(i_k, j_k)\) is a penalty factor motivated by the temporal misalignment and \(p_0\) is the maximum value of this function when two characters are perfectly aligned. The quotient \(\max(m,n,K) * p_0\) is a normalization factor, so the final value \(QSSI(S,T)\) is normalized between 0 (completely dissimilar) and 1 (equal). The simplest local similarity assigns a value of 1 for identical characters and 0 if they are different. In this case, since the path W contains only matches, all values for \(Sim(i_k, j_k)\) are 1. Then,
\[
QSSI(S,T) = \frac{1}{\max(m,n,K) * p_0} \sum_{k=1}^{K} p(i_k, j_k)
\]

For the example in Fig. 7, the QSSI of sequences are reported in the last row of Table 1. Obviously, the final matching should be the one with the maximum QSSI.
The function $p(i_k, j_k)$ can be designed to deal with the different durations of qualitative symbols according to current requirements. For example, it can assign lower weight to further matching, which is measured from the optimal diagonal. The extreme case will happen when the time influence is not considered by defining $p(i_k, j_k) = p_0$. The complete flow diagram of the QSSI algorithm is depicted in Fig. 8.

Table 1 Calculation of QSSI values.

\[
p(i_k, j_k) = \frac{1}{2|j-i-d_g|} \quad \text{and} \quad p_0=1
\]

<table>
<thead>
<tr>
<th>$d_g$</th>
<th>$\sum_{k=1}^{K} p(i_k, j_k)$</th>
<th>$\max(m,n,K)\cdot p_0$</th>
<th>$QSSI(S,T)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>$3\times0.5+1 \times 0.25 = 1.75$</td>
<td>$7 \times 1$</td>
<td>$1.75/7 = 0.25$</td>
</tr>
<tr>
<td>0</td>
<td>$3\times1 + 1 \times 0.5 = 3.5$</td>
<td>$3.5/7 = 0.5$</td>
<td>$3.5/7 = 0.5$</td>
</tr>
<tr>
<td>1</td>
<td>$1\times1 + 3 \times 0.5 = 2.5$</td>
<td>$2.5/7 = 0.357$</td>
<td>$2.5/7 = 0.357$</td>
</tr>
</tbody>
</table>
3. Illustration

In this section, we show the application of the above mentioned QSSI algorithm on a classification of voltage sags based on similarity of qualitative sequences.

3.1. Classification of sags

The current dependence of industry, commerce, and services on electricity has led to the regulation of power quality. The most common disturbances affecting power quality are voltage sags. Sags are defined as the time interval between the instant when the root mean square (RMS) voltage decreases down to 90% of its nominal value and the instant when it returns to its normal level (a three-phase unbalanced voltage sag is shown in Fig. 9). The duration depends directly on the reaction time needed by the protective system to restore the network to its normal behaviour or the duration of the transient...
fault. The origin of these alterations may be the electrical utility operation, external agents, or specific operating loads, e.g. motors, capacitors or transformers. Since any alteration of the sinusoidal power wave is usually transmitted through the electrical system according to electrical circuit laws (Bollen, 2000), power quality reports are required to evaluate the quality of the electrical supply. Monitoring reports of disturbances are used in litigation to determine the responsibility of various actors in the network. Responsibility for damages suffered by customers is often assigned to the electricity suppliers, even though the source of the sag might originate in other parts of the electrical system, such as generation, the transmission network (HV), the distribution network (MV), or other customers. Consequently, utility companies have increased the number of power quality monitors installed in distribution substations and are very interested in developing reliable methods to efficiently exploit the information contained in these registers in order to automatically discriminate between sags originating in the transmission (HV) and distribution (MV) networks and to assess and diagnose them. Moreover, determining whether a sag has occurred in the distribution or transmission networks precedes the localization and mitigation stages (Hamzah et al., 2004). This application example is focused on monitoring sags registered in 25kV distribution substations in order to assign their origin to the MV or the HV, i.e. upstream or downstream of the transformer (Error! No s'ha trobat l'origen de la referència.).

Fig. 9 a) Example of a three-phase unbalanced voltage sag. b) RMS voltage.
3.2. Qualitative representation of sags
Power quality monitors, installed in the secondary winding of transformers from each substation, allow events and associated waveforms to be detected and registered. To analyze the benefits of using qualitative representations and normalized similarities associated with these representations, instead of considering other type of disturbances such as harmonics, interruptions or flicker, only sags have been used in this work. Sags are registered as time series of three voltages and three currents involving thousands of samples. Information contained in each sample is individually irrelevant; the significant information is given by the existence of specific primitives and their duration. So, previously to the classification, the voltage and current waveforms are transformed into a sequence of episodes, described as a set of a qualitative symbol and its duration. The pre-processing and discretisation stages to obtain them are described below.

3.2.1. Sag pre-processing
In this work, only sags recorded in the delta configuration of monitors (phase to phase voltage) and gathered over the course of a year from eight 132/25kV substations have been considered. Since voltage sags are characterized by the RMS waveform of the three phase voltages, simple pre-processing has to be applied to the original waveforms to obtain the RMS values. These values were obtained using a 1 cycle (20msec, 128 samples) sliding window and applying the Short Fourier Transform (SFT) to estimate the magnitude of the fundamental frequency (50Hz). This simple pre-processing is commonly used to identify basic characteristics of sags such as magnitude and duration (Bollen, 2000). In this work the pre-processing has been applied before obtaining the qualitative description of waveforms in episodes as described in the next subsection.
Using instantaneous values of waveforms, instead of RMS, to describe sags is only applicable when interest relies on phase shift, harmonic distortion or other transient characteristics (Djokic et al., 2005).

3.2.2. Discretisation of raw data

All the RMS waveforms used in this work are composed of 39 periods, which give 4992 samples (128 samples per cycle). They were converted into sequences of episodes using the magnitude of waveforms and their first derivative. These parameters allow the signals to be represented in the same way a human expert would do it. The premise is that this human-based representation is sufficient to distinguish the different classes. Moreover, instead of considering the three phases in the conversion procedure, only the phase falling lowest during the sag was used. This is possible because the interest does not reside on the analysis of faulted phases but in the location up/downstream of the fault independently of which phase was affected. The results corroborate that this simplification is a valid way to obtain a useful representation for classification purposes.

Then, the qualitative conversion procedure is performed following four basic steps:

- **Piecewise Aggregate Approximation (PAA)** is used to reduce the length of RMS waveforms without loosing information. PAA (Keogh et al., 2000) is a dimensionality reduction technique that approximates a time series by dividing it into M equal-length segments and using the average value of the samples in each segment as the data reduced representation. M must be selected to preserve the shape of the original time series. In this application, since the original waveforms contain 128 samples per cycle, a good choice for the segment length is 64 samples. Disturbances lasting less than half a cycle (64 samples) are not considered sags. Thus, dividing the original 4992 samples by 64 results in a sequence of length M=78, where each sample represents the equivalent of a half cycle while the original shape and the minimal information required are preserved.

- **Low Pass Filtering**: A 3-sample-length mean filter has been applied to the RMS waveform. The purpose of this filter is to smooth the dynamics of the signal and avoid spurious transitions between qualitative states. A similar effect would be obtained by defining a dead-band in the transitions between qualitative states.

- **Computing the first derivative**: The derivative of the filtered sequence is calculated to identify transitions between qualitative states.
• *Qualitative representation.* Finally, a qualitative representation is obtained from the evaluation of the magnitude and the derivative. A qualitative state is assigned to every sample according to the qualitative value of the derivative. According to the previous definition of episode, the consecutive samples with the same qualitative state constitute an episode and their length is defined by the number of them. The type of the episode, useful for classification purposes, is then obtained by adding the qualitative value of the magnitude. This gives information about the depth of the sag. In this way, episodes represent a constant behaviour (derivative) between two time instants preserving the magnitude at these instants.

In the application five categories (Steady State, Fall, Rapid Fall, Rise and Rapid Rise) have been defined for the derivative and six for the magnitude, resulting in a family of 30 possible types of episodes (Table 1). The six levels of magnitude are based on the analysis of distribution of the depth of the sags. The histogram in Fig. 11 show this distribution for the sags registered in 25kV substations. Most of them (61%) arrive until a minimum magnitude between 24kV and 18kV with a majority arriving between 20 kV and 22kV. Thus, qualitative limits have been defined for each 2kV around this majority class. Below 18kV an inferior number of sags is given, resulting in a more disperse distribution. This band has been split into two categories with the threshold at 12kV to have a similar *representativity* of the resulting categories. This categorization results in a minimum of 10% of available sags in each category.

<table>
<thead>
<tr>
<th>Table 2 Types of episodes as ASCII symbols (ASCII code).</th>
</tr>
</thead>
<tbody>
<tr>
<td>6:Level High (≥24KV)</td>
</tr>
<tr>
<td>5: ↓↓↓↓ (≥22KV)</td>
</tr>
<tr>
<td>4: ↓↓↓ (≥20KV)</td>
</tr>
<tr>
<td>3: ↓↓ (≥18KV)</td>
</tr>
<tr>
<td>2: ↓ (≥12KV)</td>
</tr>
<tr>
<td>1: Level Low (&lt;12KV)</td>
</tr>
</tbody>
</table>
The application of this procedure results in a qualitative description of the waveforms of sags. This information can be represented by a string using the ASCII characters in Table 2 followed by the duration expressed as the number of consecutive samples. The whole procedure can be observed in Error! No s'ha trobat l'origen de la referència.: (a) the three voltage phases are evaluated to select the deepest one; (b) the waveform is pre-filtered using PAA representation with ½ period (64 samples) length segments; (c) the first-derivative is calculated from the PAA waveform; and (d) the two waveforms (magnitude and derivative) are qualified (in the figure integer indexes are used for simplicity) before being converted into a discrete string. The evaluation of the first derivative is used to determine the duration of the episodes, since that establishes an interval with the same behaviour. The qualified magnitude at the end of the episode is preserved in the definition of each episode as an auxiliary characteristic. Notice that the level at the beginning of the first episode is always the same, around 25kV. Thus, the resulting string for the previous example is “v3a3b4d1e2d1m33x5w25”. According to Table 2, it is read as follows: during 3 samples (segments) the waveform falls and finishes at the highest level, during 3 more samples the signal presents a rapid fall and finishes at level 4, then it falls, slightly, during the following 4 samples remaining at the same level 4, and so on. This string containing the distinguished qualitative features will be used to compare this sag against others.
Fig. 12 Process for obtaining the qualitative representation of waveforms. (a) Selection of the phase with greater depth. (b) The waveform is pre-filtered using the PAA representation. (c) The first-derivative is calculated from the PAA waveform. (d) The two waveforms are qualified according to breakpoints.

### 3.3. Comparison approach

The whole set of data collected from nine substations was divided into two subsets. The first subset was used for training and both subsets were used to test the performance of the QSSI algorithm. In the example, due to confidential restrictions on the use of real data, the names of the substations have been omitted and substituted with capital letters from A to I. The proposed method includes three main parts: first, preparation of the data as explained in Section 3.2. Next, each test case T is compared with the representative data set in a dictionary constructed previously. Finally, the test case is classified following a simple adaptive k-NN (k nearest neighbours) approach. These two last steps are described below.

#### 3.3.1. Training

In the first stage, a subset of four representative substations (labelled as A, E, F and I) were studied in order to build a dictionary of representative waveforms of sags. A total of 253 voltage sags classified as upstream/HV (transmission, 128 sags), or downstream/MV (distribution, 125 sags) according to the location of their origin in the
power network was available from these substations. After transforming these waveforms into qualitative sequences a total of 94 representative patterns (40 labelled as HV and 54 labelled as HV) were retained to build the dictionary. Only the sags different enough from the others have been retained in the dictionary. Thus, a sag with a similarity greater than 2/3 with respect to another already in the dictionary is discarded. The dictionary size reveals that even though all substations are connected to a similar HV network, their behaviour is not completely equal. Different loads, measuring standards and equipment in substations, or different distances between the registering and the fault occurring point could also be causes of these differences.

3.3.2. Test
The waveforms registered in the second subset of substations (B, C, D, G and H) were added to the first one, obtaining 528 voltage sags for testing. QSSI has been used to determine the similarity between these sags and those ones in the dictionary. The goal is to determine the class of new sags based on the similarity. The complete set of sags has been classified using a simple k-NN approach as it is explained in the next subsection.

3.3.3. Classification method
Given a test case T, an adaptive $k$-nearest neighbour search algorithm {Ougiaroglou, 2007 #441} returns the set of $k$ most similar cases, $C_1, \ldots, C_k$. In this particular example a minimum of $k=3$ nearest neighbours must be used to infer a class for the case T, and we have fixed $k=6$ as the maximum. Thus, the number of nearest neighbour will vary between $3 \leq k \leq 6$. Similarity is used to order the cases as they are retrieved allowing an early-break heuristic to interrupt the retrieval task once enough cases (50% of $k=6$) of the same class have been retrieved. Here, two classes conforms the solution space based on the origin of the sag: upstream (HV) or downstream (MV). So the class corresponding to T could be determined according to the majority. However, it is not taken into account the closeness degree of similarity between the test and the retrieved cases. Even changing the early-break heuristic could be situations where the amount of HV and MV cases is the same.

An alternative approach is to define a decision variable {Bilska-Wolak, 2002 #442}. This definition represents the likelihood of a specific class in a test case. We define an adaptive/exponential decision variable $DV_{ax}^{HV}$ as:
\[ DV_{aX}^{HV} = \frac{\sum_{C_i^{HV}} Sim(C_i^{HV}, T)^r}{\sum_{i=1}^{k} Sim(C_i, T)^r} \]

\( C_i^{HV} \) are the cases in \( C_1, \ldots, C_k \) which belong to class HV, and \( Sim(C_i, T) \) is the similarity between a case \( C_i \) and the test case \( T \), where similarity values are normalized between 0 and 1. Regarding the exponent value \( r \), it could be depended on some heuristics according to the number of NN for each class. Since this paper is focused on QSSI, the heuristics is not introduced here. (PONER EN TRABAJO FUTURO?). So for classification purposes only, because of the reduced and fixed number of retrieved \( k \)-NN used here, \( r \) has adopted the minimum number of \( k \)-NN retrieved, so, \( r \) is set to 3. That is, \( \sum_{C_i^{HV}} Sim(C_i^{HV}, T)^3 \)

\[ DV_{aX}^{HV} = \frac{\sum_{C_i^{HV}} Sim(C_i^{HV}, T)^3}{\sum_{i=1}^{k} Sim(C_i, T)^3} \]

Then, the decision variable is compared against a given threshold \( \tau \). Whenever the current value of the decision variable equals or goes beyond this threshold, the test case \( T \) is classified as belonging to class HV or not. In Table 3 it can be observed some representative situations and values returned by \( DV_{aX}^{HV} \). Note as \( DV_{aX}^{HV} \) classifies the test cases correctly because it captures the closeness of the similarity to the test case involved in the selected class.

The third situation is a special test where \( Sim(C_1^{MV}, T) \) returns a complete similarity while the second case of the same class returns a very low similarity. This last case penalizes the DV value, however \( DV_{aX}^{HV} \) yields a high value as is expected.

<table>
<thead>
<tr>
<th>Ex.</th>
<th>Sim(C_1^{HV},T)</th>
<th>Sim(C_2^{HV},T)</th>
<th>Sim(C_3^{MV},T)</th>
<th>Sim(C_4^{MV},T)</th>
<th>DV_{aX}^{HV} / T class (( \tau =0.5 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.5</td>
<td>0.5</td>
<td>0.80 / HV</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
<td>0.3</td>
<td>0.6</td>
<td>0.6</td>
<td>0.11 / MV</td>
</tr>
<tr>
<td>1</td>
<td>0.4</td>
<td>0.1</td>
<td>0.4</td>
<td>0.4</td>
<td>0.88 / HV</td>
</tr>
</tbody>
</table>

Table 3. Different situations and values returned by \( DV_{aX}^{HV} \).
3.4. Results

Table 4 shows classification results according to the majority rule and using the decision variable ($\tau=0.5$) for the whole set of substations. The last four columns indicate the assigned class, labelled as HV or MV, for each approach. For example in the substation A there are 48 sags originated in HV and the approach based on the majority rule classifies 47 as HV and 1 sag as MV. For those substations used to build the dictionary, an accurate classification would be expected. However, there are a lot of misclassifications because the majority rule is independent of the similarity degree of the retrieved cases (Table 5). Going back to substation A, now using the decision variable the sags are classified correctly.

<table>
<thead>
<tr>
<th></th>
<th>Example 1</th>
<th>Example 2</th>
<th>Example 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{sim}(C_1^A,T)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{sim}(C_2^B,T)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{Sim}(C_3^B,T)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{sim}(C_4^B,T)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D\text{V}_X^{\sim \alpha}/T$ class ($\tau=0.5$)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Majority rule</th>
<th>Decision variable</th>
</tr>
</thead>
<tbody>
<tr>
<td># sags</td>
<td>HV</td>
<td>MV</td>
</tr>
<tr>
<td>Substation A</td>
<td>sags HV 48</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>sags MV 45</td>
<td>1</td>
</tr>
<tr>
<td>Substation B</td>
<td>sags HV 20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>sags MV 16</td>
<td>2</td>
</tr>
<tr>
<td>Substation C</td>
<td>sags HV 58</td>
<td>58</td>
</tr>
</tbody>
</table>
Table 5. Retrieved cases for the sag originated in HV and classified as MV in substation A.

<table>
<thead>
<tr>
<th>Class</th>
<th>HV</th>
<th>MV</th>
<th>HV</th>
<th>MV</th>
<th>MV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similitude</td>
<td>0.907</td>
<td>0.554</td>
<td>0.474</td>
<td>0.457</td>
<td>0.455</td>
</tr>
</tbody>
</table>

Regarding the misclassifications using the decision variable, Table 6 shows the returned cases for each one, T₁ (substation C) and T₂ (substation I). Both tests belong to class MV, but they have been classified as HV. Similitudes returned for T₁ prove there is no room for precision on the decision variable. On the other hand, T₂ is a rare example since although a complete similitude was found to MV, the 50% more voted belong to HV and there is a case with a similitude very close to 1. This suggests a waveform bad classified in the original set. Fig. 13 shows the waveform of this sag and the waveform of the most similar one in HV. Nevertheless, a revision of the information associated with these registers revealed that both corresponds to the same class and that the misclassification was due to a labelling mistake in the utility.
### Table 6. Retrieved cases for misclassifications using the decision variable.

<table>
<thead>
<tr>
<th>Sim / Class</th>
<th>Sim / Class</th>
<th>Sim / Class</th>
<th>Sim / Class</th>
<th>Sim / Class</th>
<th>$DV_{aX}^{HV}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1 Subst. C</td>
<td>0.759 / HV</td>
<td>0.426 / HV</td>
<td>0.407 / HV</td>
<td>0.769 / MV</td>
<td>0.419 / MV</td>
</tr>
<tr>
<td>T2 Subst. I</td>
<td>0.944 / HV</td>
<td>0.666 / HV</td>
<td>0.659 / HV</td>
<td>1 / MV</td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 13** Two waveforms with a similar QSSI. (a) Classified as HV in the dictionary and (b) T2 originally bad classified as MV.

4.2.3 Comparative analysis. NO SÉ CÓMO ARREGLAR ESTO ENTONCES

Performance of QSSI has been contrasted with another strategy for classification of sags based on the principal component analysis (PCA). The method, described in (Khosravi et al., 2007) is applied to voltage and current RMS waveforms obtained after applying the preprocessing described in Section 2.1. Then PCA is used as a dimensionality reduction strategy to convert waveforms to vectors with a reduced number of independent components preserving the information in terms of variability. Training subsets of HV (MV) sags have been used to obtain HV (MV) PCA models with 10
components representing more than the 90% of variance of original data and to obtain the appropriate thresholds for $T^2$ and $Q$ statistics (or square prediction error). $T^2$ is computed with only the loadings of the retained principal components and represents a distance (square distance) to the centre of the model. It can be interpreted as a measure of the systematic variations of the power system where data have been gathered and a violation of the threshold when projecting new data would indicate that this systematic variation is larger than that captured by the model. On the other hand, $Q$ statistic is a squared 2-norm measuring the deviation of the observations to the lower dimensional PCA representation. $Q$ statistic measures the random variations of the monitored system; consequently a violation of the threshold would indicate that the random noise has significantly changed. When these two statistics are used along with their respective thresholds, it produces a cylindrical region around the centre of the model that is used as a decision boundary. New sags projected inside this region are assigned to the same class as the model whereas those examples falling out of this cylindrical region are associated with the other class. Using these criteria, the origin of sags registered in substations different from those used in the construction of the HV (MV) model has been classified.

The average classification rate for these substations has been calculated and summarized in Table 7. $T^2$ and $Q$ thresholds have been obtained according to statistical criteria (see for example Russell, E.L et al. 2000) from the data available for training. The last two rows contain results when the PCA models are obtained with HV and MV waveforms respectively, while the first and second rows incorporates the average classification rate obtained with QSSI for the substations not used in the creation of the dictionary. Observe that the PCA method gives a very good classification rate (98%) for MV sags when the HV model is used. Nevertheless, the same model fails in predicting its own HV class. The same is reproduced with the MV-PCA model resulting in 74.2% success in predicting the HV class. This happens because waveforms of both classes present similar variability and they are not completely separable. Also the reduced number of available examples used to build the PCA models contributes to this low performance of the model.
Table 7 Comparative results: Classification rate (%)

<table>
<thead>
<tr>
<th>Prediction:</th>
<th>HV</th>
<th>MV</th>
</tr>
</thead>
<tbody>
<tr>
<td>QSSI and</td>
<td></td>
<td></td>
</tr>
<tr>
<td>majority rule</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QSSI and decision var.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sags HV</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>sags MV</td>
<td>-</td>
<td>82.2</td>
</tr>
<tr>
<td>Sags HV</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>sags MV</td>
<td>-</td>
<td>99</td>
</tr>
<tr>
<td>Sags HV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sags MV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HV-PCA model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sags HV</td>
<td>42.0</td>
<td>58.0</td>
</tr>
<tr>
<td>sags MV</td>
<td>1.7</td>
<td>98.3</td>
</tr>
<tr>
<td>MV-PCA model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sags HV</td>
<td>74.2</td>
<td>25.8</td>
</tr>
<tr>
<td>sags MV</td>
<td>71.2</td>
<td>28.8</td>
</tr>
</tbody>
</table>

5. Conclusion and future work

This work is focused on defining and implementing a similarity algorithm (QSSI) for the comparison of qualitative time series. Qualitative representations of signal trends can be used as representations of system behaviour while they reduce the dimension of data and could provide a description in the same way a human expert would do it. First of all, a qualitative representation of waveforms is done. As a result, string chains with temporal extensions are obtained from time series (waveforms of sags in the application example). The QSSI algorithm returns a normalized index related to the degree of matching between sequences of qualitative labels. Performance of this method has been tested in the classification of voltage sags gathered at 25kV distribution substations. The objective is to assist monitoring systems in locating their origin to MV or HV, i.e. upstream or downstream of the transformer. QSSI is used to return a similarity index between new and previously registered and labelled sags. Finally, an adaptive nearest neighbour criterion and a new decision variable are used to classify the new sag.

Therefore, the main contribution of the proposed methodology is the definition of the QSSI algorithm. QSSI is based on the Needleman–Wunsch algorithm for global sequence alignment. The idea behind the approach is to build up an optimal alignment between the two sequences and then calculate the similarity using the aligned elements.

Regarding the application example, the main advantage of using the proposed methodology to locate the origin of sags is that it allows sags gathered in different
substations to be compared and similarities to be found, despite the existence of quantitative differences in the waveforms associated with the electrical characteristics (loads, network topology, transformers, etc.) of each substation.

Results show that while the failure rate is practically zero, the mean value of the success rate is about 93.1% using the majority rule. When the \( k \)-NN algorithm is complemented by the decision variable the success rate reaches the 99.2%. This is motivated because this definition takes into account the closeness degree of similarity between the test and the retrieved cases. The proposed methodology shows promising results for the whole set of substations thanks to QSSI, which is effectively used to obtain a similarity measure between two string chains. Those sags with similar index but wrongly classified exhibit a great analogy in their waveforms with sags previously labelled in a different class. A subsequent revision of the information associated with these registers revealed that there are some labelling mistakes in original data.

The inclusion of a Retaining mechanism once the origin of a sag was localised is expected to improve the performance of the system. The promising classification rates for test data show that the QSSI algorithm has satisfactory usability and confirm that the method presented may prove to be a good tool in classification approaches. With additional research, the algorithm should be extended to deal complex sequences defined by multiple attributes instead just a string. With respect to the application, the addition of quantitative characteristics to strings should improve the classification and reduce faulty classification rates.

**Acknowledgements**

This work has been made possible by the interest and participation of ENDESA DISTRIBUCION and its Power Quality Department. It has also been supported by the research projects DPI2005-08922-C02-02 and DPI2006-09370 funded by the Spanish Government.

The eXiT research group is part of Automation Engineering and Distributed Systems (AEDS), awarded with a consolidated distinction (SGR-00296) for the 2005-2008 period in the Consolidated Research Group (SGR) project of the Generalitat de Catalunya.
References


