Acoustic multiplets detection based on DBSCAN and cross-correlation

Theotime de la Selle, Jérôme Weiss, Stéphanie Deschanel

Mechanical Systems and Signal Processing 211 (2024) 111149

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

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

Acoustic multiplets detection based on DBSCAN and cross-correlation

Théotime de la Selle [a,*], Jérome Weiss [b], Stéphanie Deschanel [a]

[a] Univ Lyon, INSA Lyon, Université Claude Bernard Lyon 1, CNRS, MATEIS, UMR5510, Villeurbanne, 69621, France
[b] ISTerre, CNRS and Université Grenoble-Alpes, CS 40700, Grenoble, 38053, France

ARTICLE INFO

Communicated by M. Rebillat

Keywords:
Fatigue crack growth
Acoustic emission
Multiplets
DBSCAN
Cross-correlation

ABSTRACT

Non-destructive detection of fatigue crack propagation in industrial parts remains nowadays a key challenge in various engineering fields. Acoustic emission (AE) signals specific to incremental fatigue crack growth can be detected, cycle after cycle, as precursors to final fatigue rupture. These so-called acoustic multiplets are characterized by strongly similar waveforms, triggered at almost the same load during the fatigue cycle, and arising from the same source. Detecting such multiplets provides information about the crack growth process and for industrial parts in service, allows an early warning of potential failure. We developed a method based on a density-based data clustering algorithm (DBSCAN) working with a dissimilarity metric derived from the cross-correlation of AE waveforms to automatically classify acoustic multiplets in fatigue and other fields. Automatized processes described here allow to use the algorithm both on laboratory and industrial fatigue cases, and are designed to work in-operando. Our methodology is tested on AE signals recorded during different laboratory fatigue tests. This demonstrates the robustness of the algorithm to detect different multiplets for different materials, test conditions, specimen geometries, or acoustic sensors.

1. Introduction

1.1. Context: fatigue of materials

Fatigue of materials still remains nowadays a considerable scientific, technological and industrial problem. It came out since the beginning of the industrial revolution in the naval and railway industries, and remains relevant in present times, as illustrated by more recent dramatic accidents [1]. In these situations, parts and structures subjected to cyclic mechanical or thermo-mechanical stresses can fail under stress levels much lower than the stresses at failure observed under creep or monotonic loading. This fatigue failure is often insidious, consisting in the outcome of a crack propagation process that can extend over several hundreds or thousands of loading cycles without any detectable change in the physical or mechanical properties of the material at the scale of the part or structure in question. This problem was then extended to other fields such as aeronautics [2] and, more recently, to microelectronic and nanotechnology industries [3]. Therefore, the non-destructive, early detection of fatigue crack propagation is a prime objective.

In case of metals, the mechanisms of fatigue crack initiation and propagation are now well identified [4,5], particularly for low-cycle fatigue (LCF), which means that the plasticity threshold is repeatedly reached during cyclic loading. Under these loading conditions, dislocation movements generate localized plastic deformation jointly forming persistent slip bands (PSB) [6]. Stress concentrations around PSBs can lead to the initiation of micro-cracks. After an initial stage (stage I) of propagation limited to a

* Corresponding author.
E-mail address: Theotime.de-la-selle@insa-lyon.fr (T. de la Selle).

https://doi.org/10.1016/j.ymssp.2024.111149
Received 24 February 2023; Received in revised form 12 January 2024; Accepted 13 January 2024
Available online 14 February 2024

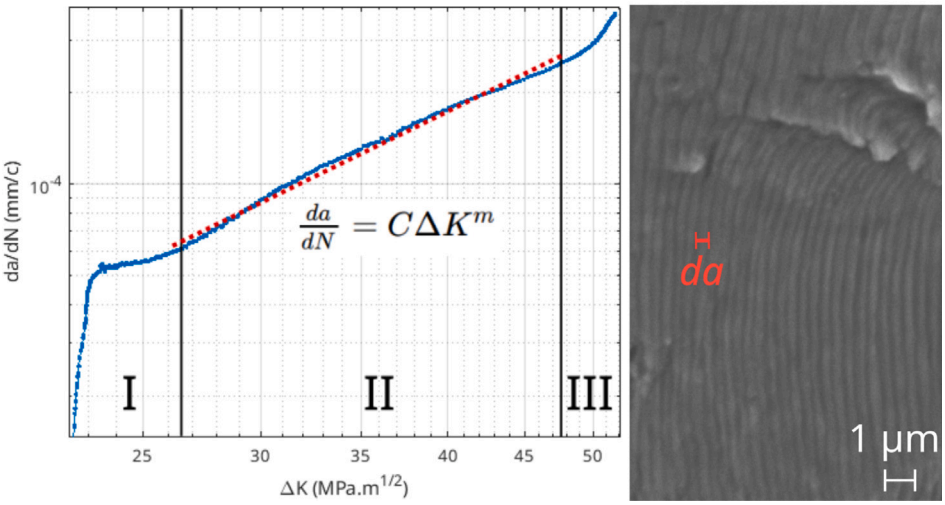

Fig. 1. Left : typical plot of crack growth rate with respect to stress intensity range ΔK for a load imposed $R = 0.1$ fatigue test performed on a 5083 aluminium alloy, showing three fatigue regimes including the Paris' law (regime II). Right : Scanning Electron Microscope (SEM) post-mortem image of fatigue striations on the fracture surface in a 5083 aluminium after the same fatigue test.

few grain sizes (i.e. a few tens or hundreds of micrometers in most classical metallic materials), the Paris' regime (stage II) [7] begins, during which propagation takes place along a plane perpendicular to the largest principal tensile stress. This second stage of propagation can last several thousands of cycles, i.e. often represents a significant fraction of the total lifetime. "Stable crack growth" is the term used as the crack front advances slowly (see stage II in Fig. 1), by successive increments of the order of a few micrometers at most at each cycle, leaving striations that are clearly visible by post-mortem analysis of the fracture surfaces (an example for a 5083 aluminum alloy is shown on Fig. 1 on the right). During the Paris' regime (stage II), crack propagation can be considered as 'slow' in comparison to propagation during stage III (unstable crack growth) or to critical crack growth not due to fatigue. Indeed, crack propagation finally strongly accelerates (stage III), rapidly leading to macroscopic rupture, which is sought to be predicted and avoided. A typical plot of crack growth rate with respect to the stress intensity range ΔK, from a fatigue test performed on a 5083 aluminium alloy, is represented in Fig. 1, illustrating the three stages. Such incremental, repetitive crack propagation is also observed in polymers or metallic glasses, although the mechanisms of crack nucleation are different in these materials.

NDT for fatigue crack growth (FCG) From this perspective, a non-destructive detection of fatigue cracks, while remaining within stage II, well before the onset of stage III preceding macro-failure, appears as a key challenge. As this stage II of cracking occurs slowly inside the material, over a very large number of cycles, its detection and the estimation of the evolution of the crack is often impossible to carry out by visual control, whereas X-ray tomography [8] can be hardly applied to structural components in service. Several methods based on the non-destructive monitoring of mechanical or physical properties have been proposed to estimate the remaining lifetime. Among these, the detection of a modification of physical properties, such as electrical resistivity [9], ultrasonic energy [10], or electrochemical properties [11], have been proposed. However, they all share a common limitation as the change in the physical property is averaged at the macro-scale, so is hardly sensitive to the effect of a single (or few) crack in the beginning of its propagation, and faced with classical signal-to-noise issues, leading to (too) late alarms. Therefore, the early detection and monitoring of fatigue crack growth from NDT remains a major engineering issue.

Acoustic emission (AE), a non-destructive technique based on passive recording of dynamic surface motion caused by elastic stress or pressure waves, has been used as a monitoring tool of fatigue for a long time [12]. Such waves are generated by the spontaneous, sudden release of elastically stored energy during sudden dislocation motions [13,14], brutal phase transformation [15,16], or microcracking [17], resulting in burst-type or continuous transient AE signals. Classically, in AE studies, the burst-type waveforms are recorded, generally on a timescale of µs to ms, when the signal crosses a threshold fixed by the user. Fig. 2 provides an example of AE waveform with some standard descriptors. The horizontal dashed lines represent the detection threshold.

We focus here on burst-type AE during fatigue. Most of the works in this field [18,19] recorded the global AE activity, considering the number of counts, or of detected bursts per cycle. Correlations between these rates and the crack growth rate da/dN, become only significant close to final failure [18]. Therefore, these current methods based on AE suffer from the same limitations than the other non-destructive methods listed above : tracking the crack growth during the Paris' regime from a global measure is extremely difficult and strongly sensitive to the signal-to-noise ratio (SNR). In addition, the non-specific nature of the AE measurements requires extra efforts of classification. As an example, besides crack growth, several mechanisms can produce AE during fatigue, such as dislocation avalanches [20], damage, phase transformations (e.g. martensitic) [16] or twinning [16], and environmental noise. Sophisticated methods were proposed to discriminate the AE signatures of different source mechanisms : e.g. the k-means algorithm for clustering AE signals (clusters representing mechanisms) from classical features of the AE signals (maximum amplitude, energy,

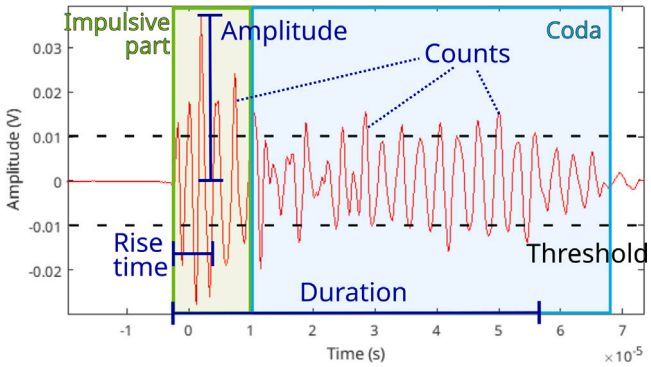

Fig. 2. A typical AE burst-type waveform and some classical description features extracted from it.

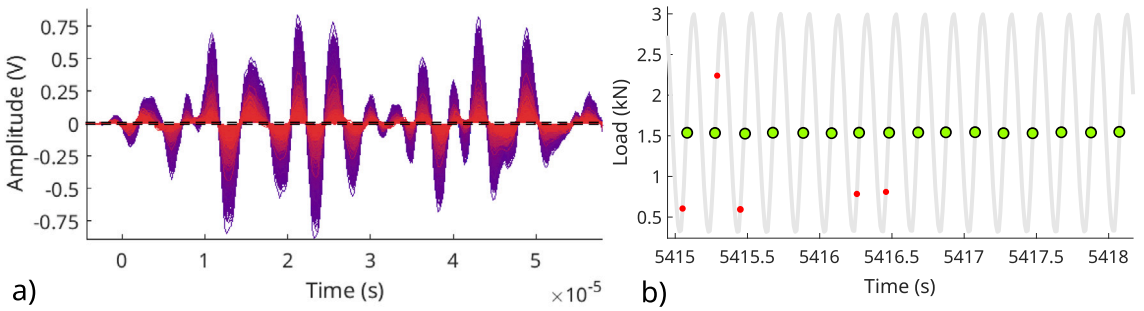

Fig. 3. (a) 3020 superposed waveforms of the multiplet recorded by a PAC nano30 sensor during a fatigue test presented in Fig. 3(b). The colors vary from red to blue depending on the arrival time. Black dashed lines represent the acquisition threshold level. (b) Example of multiplet occurrence during a load imposed ($R = 0.1$) fatigue test on a compact tension 5083 aluminium alloy specimen. Fatigue cycles with AE activity: green dots correspond to AE signals in a multiplet and red ones to other recorded signals.

duration, peak frequency, ...) [16]. Several other approaches in AE pattern recognition and signal source mechanism identification were developed [21–23] but are hardly feasible in-service.

1.2. Acoustic multiplets : a new waveform-based study

A recent work [24,25], conducted by two co-authors of this paper, highlighted, for the first time, very specific acoustic signals, characterized by almost identical waveforms (WF) (Fig. 3(a)), triggered at each fatigue cycle at almost the same stress level (Fig. 3(b)), sometimes over a very large number of successive cycles (several hundreds). By performing AE-monitored fatigue tests on several metallic materials, it has been demonstrated that these specific signals appear during fatigue crack propagation and their sources originate from the crack (1D acoustic localization procedure, see Section 4.2 for further details). It was thus concluded that these so-called acoustic multiplets, or repeating bursts, are a specific signature of the fatigue crack activity (propagation, rubbing or clapping) during stage II. They are reminiscent of repeating earthquakes, or multiplets, first identified in the 80's [26]. In seismology, these multiplets are interpreted as repeated stress releases at a same asperity along the fault [27,28]. In the seismic case, the repeatability is not related to cyclic loading but to a stick–slip mechanism under a slow far-field driving. In our context of fatigue of materials, AE multiplets grant us two major advantages : they allow an early detection of a fatigue crack growth during the Paris' regime in service, while mitigating signal-to-noise ratio problems, and offer a new way to study crack growth mechanisms.

The originality of the method comes from a waveform-based analysis, i.e. it does not rely on AE descriptors used in classical analyses. This new way of investigating AE signals requires to distinguish signals belonging to different multiplets from other sources. A first methodology proposed for this identification was based on the combination of several criteria: the similarity of the waveforms, the repetitive character of the AE from one cycle to another and over a sufficiently large number of successive cycles, and an occurrence at almost the same stress level [25]. However, this method suffers from several limitations: it relies on empirical choices for the definition of various criteria (a cross-correlation threshold to define "similar" waveforms, a minimum load variation, periodicity of signals emission). In addition, in case of poor SNR or for physical reasons that remain to be investigated, waveforms can be missed within a sequence, sometimes during tens to hundreds of cycles, leading to *"silent zones"* within a multiplet and a possible misinterpretation of the data. Finally, in the case of industrial applications, it is often impossible to have an accurate load history and the loading frequency may vary. The goal of this paper is to present a new automatic algorithm allowing the identification and classification of AE multiplets. This approach relies on a minimal number of user-defined parameters, is sufficiently fast to be

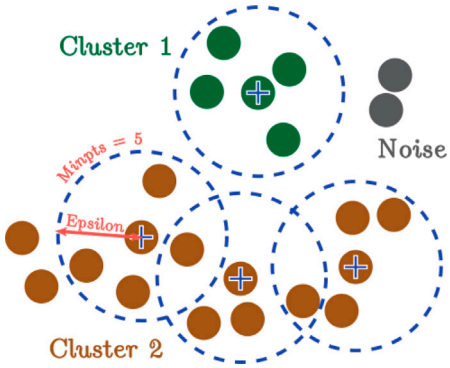

Fig. 4. Example of a 2D application of DBSCAN. Here an euclidean distance is used.

implemented in service and does not require a knowledge of the load history, hardly accessible in industrial applications, to detect properly all the multiplets.

2. Algorithmic method based on cross-correlation and DBSCAN

2.1. Data clustering algorithm DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm proposed by Ester, Kriegel, Sander and Xu in 1996 [29]. It is a non-parametric density-based clustering method: given a set of points in a N-dimensional space, this method clusters points that are very close (points with many nearby neighbors), while labeling as outliers the points that are alone in low-density regions (whose nearest neighbors are too far away). Before developing the application of DBSCAN to AE signals in fatigue, the general principle of this algorithm is briefly presented in 2D, but it works exactly the same way for a larger number of dimensions.

The main idea is that for each point of a cluster the neighborhood of a pre-defined radius has to contain at least a certain number of points, i.e. the density in the neighborhood has to go beyond a threshold [29]. DBCSAN requires fixing two parameters: *Epsilon*, the radius of the neighborhood and *Minpts*, the minimum number of points in this neighborhood. *Epsilon* defines the minimal distance for two points to be considered *linked by density* while *Minpts* defines the minimum number of points linked by density to a single point as a condition for these points to take part into a cluster. Fig. 4 represents a 2D example with two clusters and noise. Circles of radius *Epsilon* are shown around some points, representing the neighborhood. As for any classical clustering algorithm (k-means, hierarchical clustering), the choice of the metric, i.e. the "distance" between two points, is crucial. It determines the shape of a neighborhood and, according to the type of data in each dimension, some metrics are appropriated or not. The euclidean metric is used for this presentation of DBSCAN on a 2D example while for our application the choice of the metric is discussed in Section 2.2.

DBSCAN steps DBSCAN visits each point of the entire data-set from which the clusters will be extracted. All the distances between this visited point and the others are computed. Then, in this set of distances, if at least *Minpts -1* distances are less than *Epsilon*, the corresponding points *linked by density* form a cluster (including the one that DBSCAN is currently visiting). Finally, DBSCAN agglomerates together the clusters having at least one point in common.

At the end of the process, after visiting all the points of the data-set, they are separated into three categories:

- **Core points** belonging to a cluster and *linked by density* to at least *Minpts -1* points
- **Boundary points** belonging to a cluster and linked by density to less than *Minpts -1* points
- **Noise points** not belonging to any cluster.

In conclusion, we stress that DBSCAN has numerous advantages for our study. Main advantages are: non-supervised algorithm i.e. none of the data is labeled before clustering, the number of clusters is not fixed a priori, the algorithm can deal with noise and it is able to detect clusters with complex shapes. Note that in our DBSCAN implementation, each cluster constitutes a multiplet as we are going to cluster waveforms according to their dissimilarities (see multiplet definition in 1.2) and verify manually the detection results corresponds to known multiplets (see Section 4.2). Thus, *noise* points are associated to all other signals coming from other types of sources (overall plasticity, phase transformations, etc.) or from environmental noise. In what follows, therefore, we use the term *multiplets* only.

2.2. DBSCAN combined with cross-correlation of waveforms for multiplets detection

Time series clustering with DBSCAN Even if DBSCAN was initially designed for spatial clustering (the metric being a spatial distance, see Fig. 4), it has also been used to classify time series. However, as explained by [30,31], most of the works on time series clustering with DBSCAN are based on time series *descriptors*, or "*features*" (see Fig. 2). This is the case for the works based on AE, with applications proposed in AE source location [32], in defects identification in roll contacts of machine elements [33], or localization of fractures in rock [34]. Here we intend to cluster AE waveforms with DBSCAN directly from their *shapes*, which has never been proposed to classify AE signals, to our knowledge.

Shape similarity metric selection As presented in 1.2, the detection of multiplets was initially based on waveforms similarity, times of arrival (TOA) and the stress level at which the AE signals are triggered. Here, we only rely on the waveforms without post-recording filtering and their respective TOAs. In particular, we do not rely anymore on the measurement of stress, which might be difficult in industrial cases. A key point in this approach is the choice of an appropriate and robust measure of waveform similarity to implement in DBSCAN as a metric of their closeness. We choose to develop a time series shape similarity metric based on the *cross-correlation function*, which is often chosen for this kind of problematic in other fields [35–37].

Some computational works have tried time series shapes clustering by DBSCAN [30,38,39] with different metrics (euclidean, Dynamic Time Warping, Minimum Jump Cost). To our knowledge, apart from one computational study which introduce, compare and demonstrate the efficiency of a time series shape similarity metric based on the cross-correlation function [40], it is the first time that an application of DBSCAN combined with the cross-correlation function is proposed.

The cross-correlation function The cross-correlation consists in the displaced dot product between two time series. It is often used to quantify the degree of similarity between two signals [35]. In the case of discretized records, the cross-correlation between two signals u and v with the same N samples length is expressed by:

$$corr[u,v](t) = \sum_{n=n_0}^{N} u(n) \cdot v(t+n) \tag{1}$$

When the discrete time series u and v match, the value of $corr[u,v]$ is maximized and the maximization corresponds to the time delay between the two time series. In the case of AE waveforms, the two parameters n_0 and N define a cross-correlation window. Because acoustic waveforms, triggered when the signal crosses the threshold (see 1.1), are transient waves and contain an impulsive part and a coda (see Fig. 2), the choice of this window has to be set precisely for an adequate determination of the maximization point.

Dissimilarity metric Based on a normalization of the cross-correlation function, we define a *dissimilarity metric* δ_{uv} measuring the remoteness between the shapes of two waveforms u and v. δ_{uv} takes values in $[0,1]$ (0 meaning *identical waveforms* and 1 *totally dissimilar*) and is adapted to the DBSCAN algorithm.

$$\delta_{uv} = 1 - \left| \max_t \left(\frac{corr[u,v](t)}{\sqrt{corr[u,u](0) \times corr[v,v](0)}} \right) \right| \tag{2}$$

Partial dissimilarity matrix and Δt_{max} parameter In order to cluster all the waveforms with DBSCAN, one could compute a dissimilarity matrix M which evaluates the dissimilarity between each signals and all others. However, because of the computational cost of the cross-correlation maximum, we propose for real-time analyses to calculate only some chosen coefficients of the matrix and to build what we call a *partial dissimilarity matrix*. For this purpose, a time-based parameter Δt_{max} is introduced to set which coefficients should be computed. For each waveform u, are computed only the coefficients M_{uv} satisfying the condition: $t_v < t_u + \Delta t_{max}$ on TOAs t_u and t_v.

$$M_{uv} = \begin{cases} \delta_{uv}, & \text{if } t_v < t_u + \Delta t_{max} \\ 1, & \text{otherwise} \end{cases}$$

This parameter besides reducing drastically the computing time by reducing the number of coefficients to compute, is also necessary for our application in AE multiplet detection. Indeed, to correctly cluster the signals belonging to multiplets emitted cycle to cycle, it its necessary to limit the detection of waveforms *linked by density* in time (see Section 2.1). Δt_{max} allows this limitation in time because two signals u and v separated by more than Δt_{max} are considered *totally dissimilar* by fixing the coefficients at 1. An example of the resulting partial matrix of dissimilarities is shown on Fig. 5 (the whole matrix is displayed but only half of it has actually been computed because of symmetry).

2.3. Definition of the three parameters in the general case

Three parameter have been defined in the previous section, two corresponding to DBSCAN, *Epsilon* and *Minpts*, and one time-based parameter which defines the computation of the partial dissimilarity matrix, Δt_{max}. These parameters are related to the characteristics of the multiplets, and therefore control the multiplets detection according to the test conditions or the industrial application.

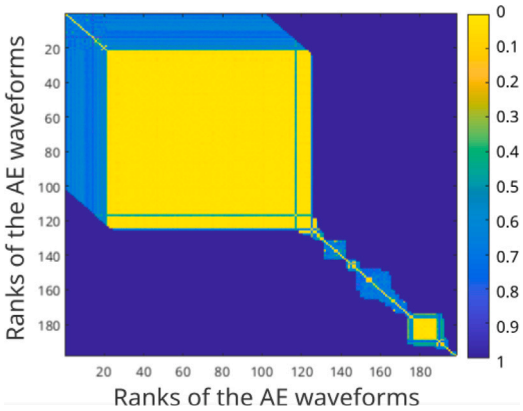

Fig. 5. Example of a partial dissimilarity matrix obtained from a load imposed ($R = 0.1$) fatigue test on a double edge notch tension specimen of 42CD4 steel. The color scale corresponds to the level of dissimilarity. The emerging yellow squares correspond to identified multiplets.

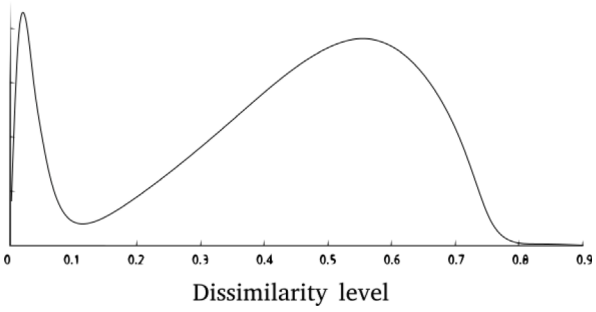

Fig. 6. Example of a distribution of dissimilarities, showing a mode of low δ_{uv} corresponding to large intra-multiplet cross-correlations.

- **Epsilon :** cross-correlation threshold, i.e. the maximum of dissimilarity between waveforms belonging to the same multiplet. $Epsilon$ can be automatically determined (see after).
- **Minpts :** minimal size of a multiplet, i.e. the minimum number of signals within each multiplet. This parameter is user-defined.
- Δt_{max} : sliding time window of detection. This parameter drives the filling of the partial matrix (see above). Δt_{max} is user-defined and depends strongly on the applications of this method.

Automatic determination of Epsilon While Δt_{max} and $Minpts$ are user-defined parameters, $Epsilon$ can be determined automatically from the distribution of computed dissimilarities. In a fatigue test characterized by acoustic multiplets, two modes are expected in this distribution (see Fig. 6): one narrow mode close to 0 corresponding to small dissimilarities between waveforms belonging to multiplets, and a wider one corresponding to uncorrelated source signals. The remaining correlation of these individual signals results from the resonant nature of the AE sensors. Indeed, due to the acquisition chain, and particularly to the response of the sensor, the cross-correlation between two signals with a totally different source is larger than zero. As a matter of fact, a distribution of cross-correlations computed on a dataset taken from an experiment without multiplets would not exhibit the large cross-correlations mode.

In case of a large dataset, this computation could take too much time for a real time application. To mitigate this effect, the distribution is calculated only on several tens of the matrix diagonals.

From such distribution, one can easily select automatically the threshold by finding the first local minimum between the two modes, e.g. around 0.1 for the schematic figure above.

General method summary Fig. 7 summarizes the general multiplet detection method presented above. The inputs are the time-series representing all the waveforms recorded and their corresponding TOAs. The two user-defined parameters allow to adapt the detection according to the applications. As an output, a multiplet (or noise) id is associated to each waveform.

2.4. Adaptation to detect multiplets during fatigue

The method and the three parameters presented previously are not specific to fatigue multiplets and could be applied to find multiplets in other fields. Nevertheless, we will only focus on fatigue multiplets for the rest of the article. In this section, the method is completed and we further explain how to detect the multiplets emitted by fatigue mechanisms.

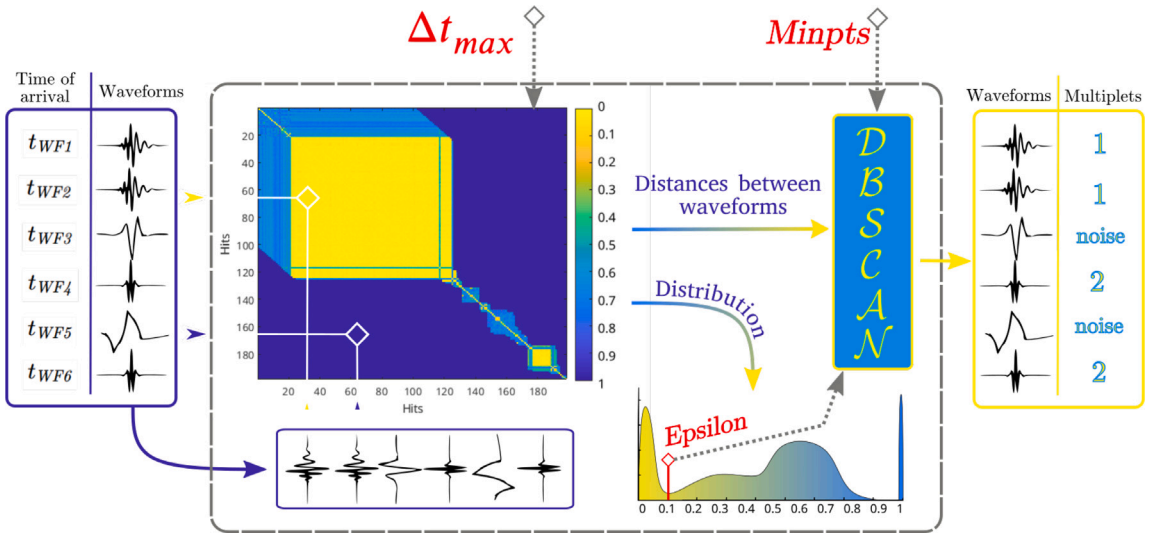

Fig. 7. Architecture of the multiplet detection algorithm.

Setting of Δt_{max} to detect fatigue multiplets Acoustic multiplets resulting from the fatigue of materials are specific because of their periodicity, since the crack emits a signal at each solicitation cycle. Thus, in this case, the sliding time window of detection, Δt_{max}, has to be proportional to the cycling period T_c and the minimum number of signal in each multiplets $Minpts$. Since in practice perfectly periodic multiplets are not recorded and some silent zones are observed in the multiplets, a widening factor d (user parameter) is added:

$$\Delta t_{max} = \frac{T_c \times Minpts}{2 \times d} \tag{3}$$

$d \in]0, 1]$ corresponds to the minimum multiplets density that is imposed in detection.

In the end, the user has only two parameters to choose: $(Minpts, d)$, each having a physical meaning related to the multiplets that can be detected.

Industrial case: approximation of T_c The cycling period T_c is set by the incremental fatigue crack propagation, itself resulting from the solicitation applied on the system. There are two cases: laboratory experiments and industrial applications. Obviously, in laboratory experiments, T_c is known and no extra effort is needed. On the other hand, in the case of industrial applications, T_c may not be known *a priori*, and may vary during the fatigue process. A solution based on the detection of multiplets is proposed below to automatically estimate the period of solicitation T_c that induces fatigue crack propagation.

As we try to approximate T_c, Δt_{max} is not user-defined, and the cross-correlation matrix is fully calculated. Assuming that fatigue multiplets are quasi-periodic with a period T_c, it is possible to approximate T_c by T_m, the mean of the *time between hits (TBH)* of the multiplets detected by DBSCAN based on the full cross-correlation matrix (see Eq. (4)). For this purpose, $Epsilon$ is determined automatically (see above), and we set $Minpts = 10$ to allow the detection of small parts of multiplets for which the periodicity is well defined (not perturbed by silent zones).

$$T_m = \frac{\sum_{k=1}^{N_m} TBH_m^k}{N_m} \text{ and } TBH_m^i = TOA_m^i - TOA_m^{i-1}, \tag{4}$$

where m is the multiplet id, i the index of the signal within the multiplet, and N_m the size of the multiplet m.

If several multiplets are detected, T_m is derived from the one that minimizes the *TBH standard deviation*, σ_m.

To reduce the full matrix and finally obtain a partial matrix corresponding to a well estimated Δt_{max}, the following procedure is iterated:

- The approximation of $T_m^{(n)}$ at iteration n allows to derive a $\Delta t_{max}^{(n)}$, hence to reduce the matrix $M^{(n-1)}$ to $M^{(n)}$
- The detection of multiplets is performed on $M^{(n)}$
- For each multiplet, $\sigma_m^{(n)}$ is calculated
- $T_m^{(n+1)}$ is obtained from the value, among the (possible) different multiplets, associated to the smallest $\sigma_m^{(n)}$

The matrix is reduced at each step n and, as a result, T^n as well. $T_m^{(n)}$ converges to the real period of multiplets T_c and the convergence is checked from a criterion on the residual $r^{(n+1)} = T_m^{(n)} - T_m^{(n+1)}$. The procedure, summarized in Fig. 8, is stopped when $r^{(n)} < r_{th}$, with $r_{th} = 10^{-3}$ s for the experimental examples detailed in Section 4. Since in industrial applications the characteristic loading frequency of the system may vary, this iterative method is relaunched each time a multiplet is detected.

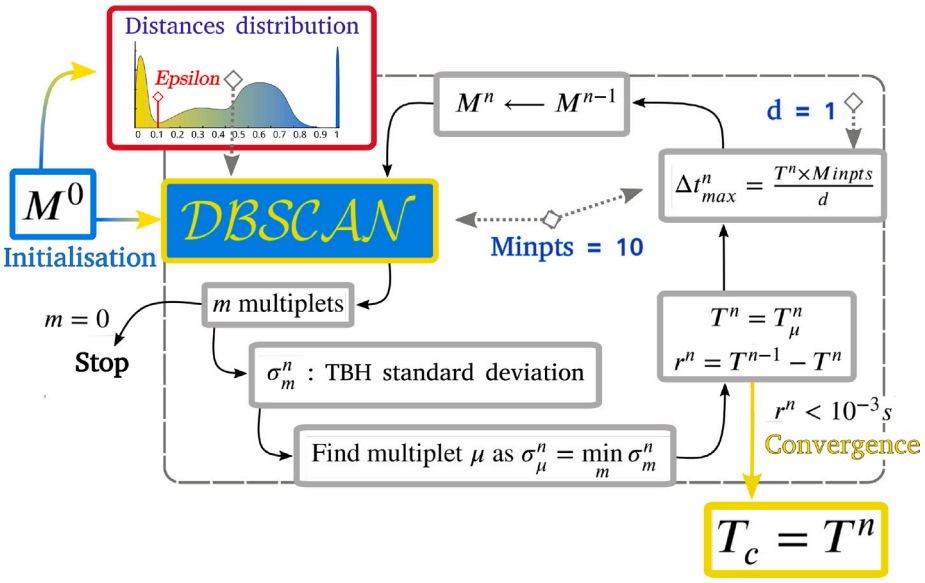

Fig. 8. Iterative procedure to approximate the fatigue period T_c.

To sum up, if a structure subjected to fatigue loading of variable periodicity emits multiplets, our algorithm can be improved by a stand alone procedure to be able to detect and classify them. The procedure that has been presented here estimates, from the recorded signals, the periodicity of the considered multiplet that fits best the data. In a real life fatigue case with a (slowly) evolving fatigue frequency, this procedure is relaunched periodically to re-estimate the evolving emission periodicity and then allows a robust classification. However, in the extreme case of a fully random solicitation and/or if the standard deviation of the cycle duration is large compared to its mean, we cannot ensure that the algorithm will be able to classify properly.

3. Experimental dataset

Our algorithm has been tested on a large dataset of AE recording from diverse fatigue tests in the laboratory (see Table 1 for an overall view). Two Physical Acoustic Corporation sensors, not necessarily the same, are coupled on the specimen surface for each test. Their positions are depicted on Fig. 9. For these tests, waveforms are sampled at 5 MHz during at most 400 µs. Noise reduction was only necessary in the case of compact tension (CT) experiments, as the other specimen geometries were solicited using silent fatigue machines. In case of CT specimens, the high amplitude acoustic noise coming from the grips were filtered by teflon rings inserted in the specimens pin holes. Teflon is commonly used for noise filtering in AE applications. Acquisition thresholds (see Section 1.1) are set by a preliminary cyclic test (few minutes) in the elastic domain under low load. Given that no source of AE emits signals in this condition, the threshold is selected just above the ambient noise recorded by the sensors. The resulting thresholds depend on the fatigue machines and the assembly mounting: in our case, the thresholds are about 32 dB for a specific AE machine (noise reduction) and around 38 dB for a classical hydraulic fatigue machine.

To study the robustness of the method, many test conditions have been used :

- **Materials :** AISI 4140 (42CD4) steel, 5083 and 7075 aluminium alloys
- **Specimen geometries :** Compact tension (CT), double edge notch tension (DENT) and dogbone (DB) (see Fig. 9)
- **Fatigue machines :** Homemade silent (HS) for DENT specimens, classical hydraulic Instron for CT specimens and hydraulic MTS for DB ones
- **Fatigue test loading conditions :** Strain imposed with a strain ratio $R_\epsilon = \frac{\epsilon_{min}}{\epsilon_{max}} = 0.1$, and load imposed with applied force ratios $R_F = \frac{F_{min}}{F_{max}}$ equal to 0.1, 0.5 or 0.7. The loading frequency ranges from 0.2 to 10 Hz, depending on the test (see Table 1).
- **AE sensors (frequency ranges):** μ200 ([500, 4500] kHz), μ80 ([200, 900] kHz), nano30 ([150, 750] kHz)

Due to the different microstructures and compositions, the three materials have different fatigue behaviors, which could influence the emission of multiplets. The geometries impact the propagation and reflection of waves, while the different machines generate more or less experimental noise. Increasing the load ratio allows to reduce the fretting of the crack faces, and thus to study the multiplets coming from this mechanism. Finally, various sensors with various frequency responses were used to show the adaptability of the method. Even if two sensors record the AE activity during each test, the multiplet detection is applied separately on the two sensors.

We did not try to investigate phenomenological differences between low cycle fatigue (LCF) and high cycle fatigue (HCF) conditions. Our goal was to ensure that our detection algorithm worked in different loading situations (strain or stress controlled).

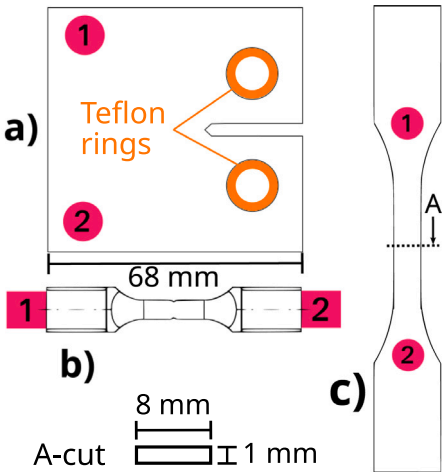

Fig. 9. Specimen geometries and sensors positions: **(a)** CT **(b)** DENT **(c)** Dogbone. Teflon rings are inserted into the CT pin holes to filter noise coming from the grips.

Table 1
7 tests selected covering all different cases.

n°	Material	Geometry	Machine	Loading	Sensors
1	5083	DENT	HS	$R_F = 0.1$	μ200+μ80
2	5083	CT	Instron	$R_F = 0.1$	nano30 (2)
3	5083	CT	Instron	$R_F = 0.7$	nano30 (2)
4	4140	DENT	HS	$R_\varepsilon = 0.1$	nano30 (2)
5	4140	CT	Instron	$R_F = 0.1$	μ200+μ80
6	4140	CT	Instron	$R_F = 0.5$	μ80 (2)
7	7075	DB	MTS	$R_F = 0.1$	μ80 (2)

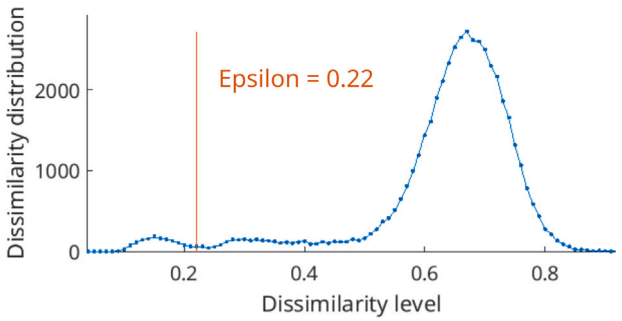

Fig. 10. Example of the automatic determination of $Epsilon$ on AE data recorded by a μ200 sensor during a fatigue test (test n° 1). The orange vertical line represents this threshold $Epsilon$.

However, if we consider that in LCF the yield stress is reached at each cycle while this is not the case in HCF, we can consider that most of our tests belong to LCF.

4. Validation of the method on experimental dataset

4.1. Parameters selection according to the dataset

Epsilon automatic selection The dissimilarity threshold $Epsilon$ is automatically determined from the procedure presented in , while Δt_{max} and $Minpts$ are defined by the user. An example of the automatic selection of $Epsilon$ is shown on Fig. 10. One could notice that the level of dissimilarity within multiplets is relatively large here (around 0.15). *Epsilon* is then automatically selected at 0.21. This can be explained by the large frequency range of the sensor μ200 used in this test, while other sensors are more resonant and show a multiplet dissimilarity mode around 0.05 (see). This demonstrates that our method is adaptable to different sensors with various frequency ranges.

Table 2
Results: approximation of T_c.

n°	Iterations	T_c^{approx}	T_c	Relative error
1	3	0.1999 Hz	0.2 Hz	0.0031%
2	2	5.0005 Hz	5 Hz	0.0096%
3	3	5.0002 Hz	5 Hz	0.0032%
4	3	0.9998 Hz	1 Hz	0.0173%
5	3	1.0003 Hz	1 Hz	0.0272%
6	5	9.9997 Hz	10 Hz	0.0033%
7	4	0.9946 Hz	1 Hz	0.5432%

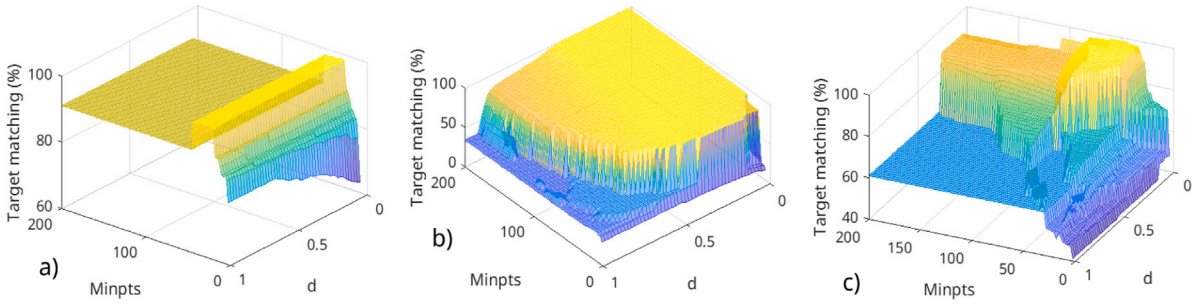

Fig. 11. 3 Sensitivity studies of the parameters $Minpts$ and d. An optimal classification called *target* is defined manually and a comparison is made for each classification performed with all couples $(Minpts, d)$. (a) Short multiplet alone (b) Long multiplet (c) combination of a short and long multiplets with silent zones from a noisy environment (AE acquisition threshold at 48 dB, an extreme case).

Δt_{max} determination As explained in Section 2.4, Δt_{max} is related to the parameters d and $Minpts$ (Eq. (3)) in fatigue applications. These user-defined parameters act as lower thresholds: d defines a lower bound for the density while $Minpts$ is a lower limit for the multiplet size. However, the use of this Δt_{max} formulation requires to know the cycling period T_c. In laboratory tests, T_c is fixed by the operator during the test but in industrial fatigue cases, T_c may not be known, or could change during the lifetime of the component. Consequently, we used the method detailed in to approximate T_c for each test of our dataset (only on one sensor) and to compare with the imposed fatigue periods. Results are summarized in Table 2. Since the errors remain always much less than 1%, we make the approximation $T_c = T_c^{approx}$ and apply the multiplet detection in our laboratory cases as one would do for industrial cases.

Minpts and d setting These two user-defined parameters are directly linked to the minimum size and the minimum density of the multiplets. In practice, ambient noise impacts the recording of AE signals and affects the size and the density of potentially detected multiplets. This means that the choice of these parameters depends on a compromise between sensitivity and specificity: increasing $Minpts$ or d avoids the clustering of non-multiplet signals, but at the risk to miss some multiplets, especially if the associated SNR is relatively low. On the contrary, decreasing these parameters enlarges the detection and tends to misclassify signals in multiplets.

Sensitivity study Fig. 11 shows a sensitivity study over 3 representative examples: a short multiplet, a long multiplet and a combination of a short and long multiplets with silent zones from a noisy environment (AE acquisition threshold at 48 dB). For each of them, the best detection is found by setting manually $Minpts$ and d; this optimal classification result is called *target*. This *target* is compared with all detections performed for $Minpts \in [1, 200]$ and $d \in [0, 1]$. A landscape of similarity to *target* is obtained for each test at all couples $(Minpts, d)$. Yellow regions corresponds to target matching close to 100% and delimit the range of the coordinates $Minpts$ and d.

According to the sensitivity studies performed on several examples, the lower limit of density d has only a significant impact on detection when one or more multiplets exhibit *silent zones* (Fig. 11*c*), which often occurs in noisy environment or sometimes in very long multiplets. The effect of the lower limit of multiplet size $Minpts$ is more or less the opposite: major variations in detection are noticed on small multiplets (in Fig. 11*a* between 10 and 30 signals). Decreasing too much $Minpts$ increase the risk of misclassification because of too high sensibility. In case of long multiplets (Fig. 11*b*) the couple $(Minpts, d)$ can be chosen in a large range ($d \in]0, 0.6]$ and $Minpts\ in [20, 200]$) and thus, for high values of $Minpts$ and d, the risk of misclassification is very low. However, highly noisy environment (Fig. 11*c*) restricts the range of both $Minpts$ and d, because it creates *silent zones* and short multiplets. In this case, d should be taken below 0.3 and the choice of $Minpts$ would be a compromise between sensitivity (avoid missing multiplets) and specificity (avoid false classifications in multiplets).

Computing time on a conventional processor To ensure that the algorithm has a sufficiently low level of time complexity to detect multiplets in real-time in case of industrial NDT, we simulate the *in-operando* computation by running it on a replay of each test of the dataset. On a conventional processor (Intel(R) Core(TM) i5-10310U CPU @ 1.70 GHz), the algorithm is sufficiently fast to detect multiplets at any moment of all the 7 fatigue tests.

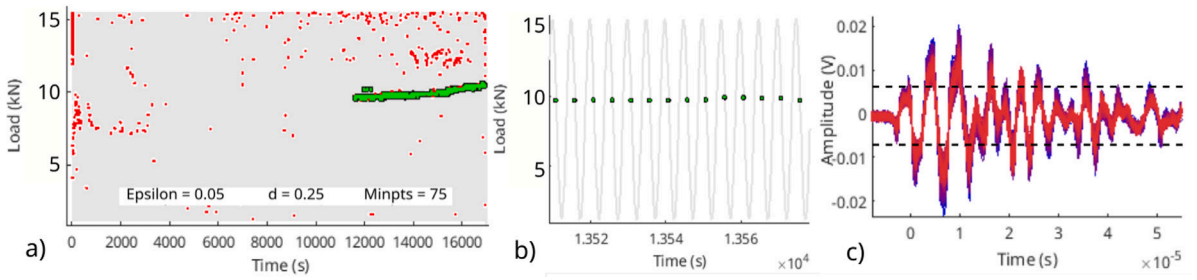

Fig. 12. Example 1 (a) Multiplet detection during a FCG ($R_F = 0.1$) test on a DENT specimen of 5083 aluminium alloy recorded by a μ80 sensor (test n° 1). Green dots are multiplet signals. (b) zoom on one multiplet and (c) the corresponding waveforms superposed.

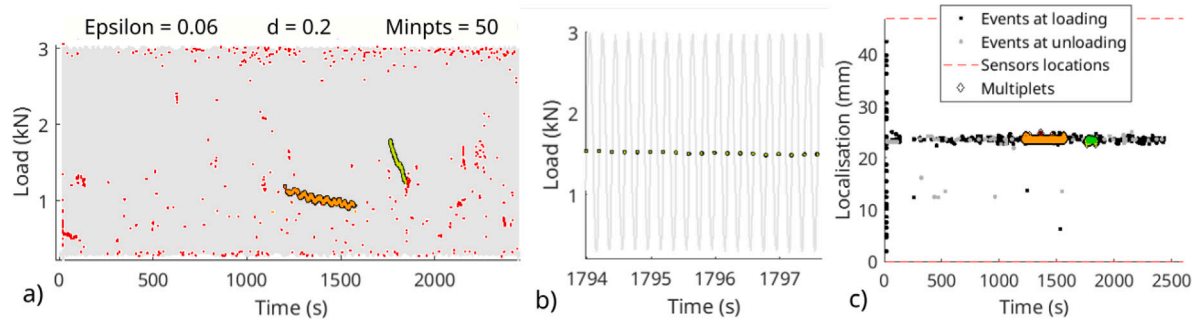

Fig. 13. Example 2 (a) Multiplets detection during a FCG ($R_F = 0.1$) test on a CT specimen of 5083 aluminium alloy recorded by a nano30 sensors (test n° 2). Orange and yellow dots are multiplets signals. (b) zoom on the only one multiplet and (c) the 1D localization of AE events, including multiplets.

4.2. Detailed results for two fatigue tests

Two examples of fatigue multiplets detected by the method presented in this paper are shown in Figs. 12 and 13. For each of these two tests, we show the occurrence of AE hits, belonging or not to multiplet(s), during the fatigue life: AE hits (red dots) are superposed on the loading cycles (gray line) and detected multiplets are emphasized by other colors (blue, yellow and orange). Figs. 12(b) and 13(b) of these two figures are zooming on some cycles to reveal the periodicity of signals belonging to multiplets, and to show their occurrence within a fatigue cycle (at loading or unloading). Fig. 12(c) shows a 1D spatial localization of the signals belonging to the detected multiplet. This 1D localization is performed classically in AE by triangulation based on the arrival times gap $\Delta t = t_2 - t_1$ determined from first threshold crossings respectively on sensors 2 and 1 (see [24] for more details). Here, the location confirms that the associated AE source is located at the crack position. Fig. 12(c) shows the similarity of the waveforms for the detected multiplet.

These examples demonstrate that all clustered signals satisfy the conditions recalled in Section 1.2 to define a multiplet: almost identical waveforms, triggered at each fatigue cycle at almost the same stress level, and arising from the same source associated to the fatigue crack (note that some signals cannot be localized as only one sensor received them). This was verified for all clusters identified by our method, for all fatigue tests in our dataset.

4.3. Specific cases

To further demonstrate the robustness of the algorithm, this section presents the detection of multiplets in some complicated cases.

Multiplets including silent zones As mentioned in Section 1.2, multiplets may include silent zones (see Fig. 14) which could lead to misclassifications: multiplets fragmented into smaller ones, or parts of multiplets missed. As the parameter d, controlling the minimum density, is linked to the maximum acceptable size of a silent zone, our method can properly detect multiplets cut by silent zones. Fig. 14 illustrates that the multiplet (yellow) has been detected and identified by the algorithm even if it is fragmented into different episodes.

Quasi-simultaneous multiplets In some experiments, different fatigue multiplets can be emitted within the same cycles, sometimes from similar sources resulting in rather similar waveforms (see Fig. 15). Thanks to the automatic selection of $Epsilon$ based on the distribution of dissimilarities, the dissimilarity threshold is adapted to separate these multiplets. In such case, the distribution plotted in Fig. 16 exhibits the classical *multiplets mode* at a very low dissimilarity level of about 0.02 and a second mode, next to the

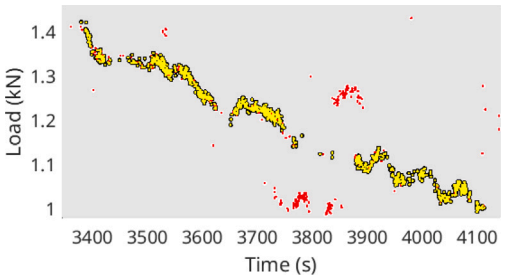

Fig. 14. Multiplet separated by silent zones detected during a FCG ($R_F = 0.1$) test on a CT specimen of 5083 aluminium alloy (test n° 2).

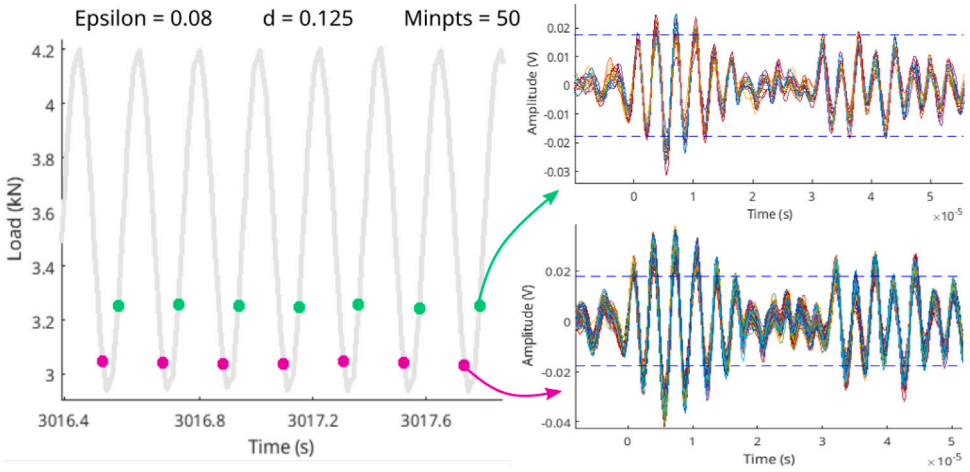

Fig. 15. Multiplets emitted during the same loading cycles, at loading and unloading, detected during a FCG ($R_F = 0.7$) test on a CT specimen of 5083 aluminium alloy (test n° 3).

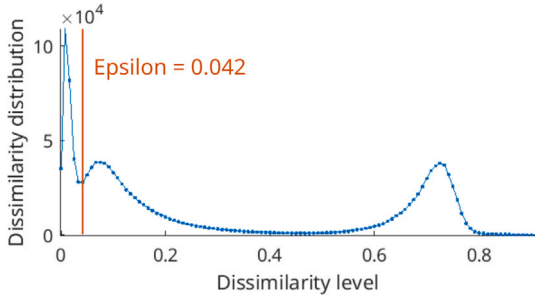

Fig. 16. Distribution of dissimilarities δ_{uv} and the automatic setting of *Epsilon* in the case shown on Fig. 15 (test n° 3).

first, at a higher dissimilarity level of about 0.09. The first mode is composed of the distances δ_{uv} between waveforms within each of the two similar multiplets (distances *intra*-multiplets about $\delta_{uv} = 0.02$) while the second mode corresponds to the distances δ_{uv} computed between waveforms belonging to those two multiplets (distances *inter*-multiplets about $\delta_{uv} = 0.09$). To correctly separate these multiplets in the detection, the parameter *Epsilon* is automatically set at the minimum between the two modes, as explained in Section 2.3. The algorithm returns two multiplets well separated within the loading cycle, with their respective waveforms shown on Fig. 15.

5. Conclusion

This paper present a method, based on the combination of a clustering algorithm DBSCAN and the cross-correlation between acoustic emission (AE) waveforms, to automatically detect and classify AE multiplets. These multiplets consist of groups of AE signals with very similar (highly correlated) waveforms, signature of a unique source. The classification is performed using a dissimilarity metric derived from the cross-correlation function and is designed to any kind of AE applications emitting multiplets.

AE multiplets have been recently identified as a signature of incremental fatigue crack growth, therefore as a potential non-destructive, early warning of fatigue failure. Here we detail the adaptation of our method to fatigue applications, for which multiplets comprise highly similar AE waveforms repeating at each loading cycle at almost the same stress level. As a NDT tool, our algorithm is designed to run in a in-operando mode on AE data, for laboratory tests as well as industrial cases. In industrial situations, the periodicity of the fatigue multiplets, which is closely linked to the (not always well known) loading frequency, is automatically estimated from a dedicated procedure.

Our method is tested and validated on an experimental dataset composed of various fatigue tests with different metallic materials, AE sensors, specimen geometries, types of solicitation, and testing machines. The robustness of the algorithm is demonstrated, in particular, by the analysis of some specific cases: separation of quasi-simultaneous and analogous multiplets, or the detection of multiplets including silent zones as a result of a low SNR.

CRediT authorship contribution statement

Théotime de la Selle: Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jérome Weiss:** Supervision, Validation, Writing – review & editing. **Stéphanie Deschanel:** Project administration, Supervision, Validation, 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.

Data availability

Data will be made available on request.

Acknowledgments

The authors gratefully acknowledge the support of the Agence Nationale de la Recherche (ANR) through grant ANR-19-CE42-0012.

References

[1] W. Schütz, A history of fatigue, Eng. Fract. Mech. (54) (1996) 263–300.
[2] S.K. Bhaumik, M. Sujata, M.A. Venkataswamy, Fatigue failure of aircraft components, Eng. Fail. Anal. 15 (6) (2008) 675–694.
[3] J. Suhr, P. Victor, L. Ci, S. Sreekala, X.R. Zhang, O. Nalamasu, P.M. Ajayan, Fatigue resistance of aligned carbon nanotube arrays under cyclic compression, Nature Nanotechnol. 2 7 (2007) 417–421.
[4] S. Suresh, Fatigue of Materials, Cambridge university Press, 2001.
[5] J. Schijve, Fatigue of Structures and Materials, Springer, Dordrecht, 2010.
[6] W.A. Wood, Formation of fatigue cracks, Philos. Mag.: J. Theor. Exp. Appl. Phys. 3 (31) (1958) 692–699.
[7] P. Paris, F. Erdogan, A Critical Analysis of Crack Propagation Laws, J. Basic Eng. 85 (4) (1963) 528–533.
[8] S. Dezecot, J.-Y. Buffiere, A. Koster, V. Maurel, F. Szmytka, E. Charkaluk, N. Dahdah, A. El Bartali, N. Limodin, J.-F. Witz, In situ 3D characterization of high temperature fatigue damage mechanisms in a cast aluminum alloy using synchrotron X-ray tomography, Scr. Mater. 113 (2016) 254–258.
[9] R. Horstman, K. Lieb, B. Power, R. Meltzer, M. Vieth, G. Aronson, R. Ritchie, Optimization of the electrical potential technique for crack growth monitoring in compact test pieces using finite element analysis, J. Test. Eval. 7 (4) (1979) 208.
[10] J. Papazian, J. Nardiello, R. Silberstein, G. Welsh, D. Grundy, C. Craven, L. Evans, N. Goldfine, J. Michaels, T. Michaels, Sensors for monitoring early stage fatigue cracking, Int. J. Fatigue 29 (9–11) (2007) 1668–1680.
[11] A. Girones, Evaluation of fatigue damage for duplex stainless steels in aggressive environments by means of an Electrochemical Fatigue Sensor (EFS), Int. J. Fatigue 25 (9–11) (2003) 1189–1194.
[12] D.O. Harris, H.L. Dunegan, Continuous monitoring of fatigue-crack growth by acoustic-emission techniques: Purpose of this investigation was to further explore the relationship between crack-growth characteristics and acoustic-emission variables such as instrumentation gain and sensor frequency. The threshold conditions for crack detection were also investigated, Exp. Mech. 14 (2) (1974) 71–81.
[13] G. L'Hôte, S. Cazottes, J. Lachambre, M. Montagnat, P. Courtois, J. Weiss, S. Deschanel, Dislocation dynamics during cyclic loading in copper single crystal, Materialia 8 (2019) 100501.
[14] J. Weiss, W. Ben Rhouma, T. Richeton, S. Dechanel, F. Louchet, L. Truskinovsky, From mild to wild fluctuations in crystal plasticity, Phys. Rev. Lett. 114 (10) (2015) 105504.
[15] S.M.C. van Bohemen, M.J.M. Hermans, G. den Ouden, I.M. Richardson, A study of acoustic emission energy generated during bainite and martensite formation, J. Phys. D: Appl. Phys. 35 (15) (2002) 1889–1894.
[16] M. Shaira, N. Godin, P. Guy, L. Vanel, J. Courbon, Evaluation of the strain-induced martensitic transformation by acoustic emission monitoring in 304L austenitic stainless steel: Identification of the AE signature of the martensitic transformation and power-law statistics, Mater. Sci. Eng. A (2008) 8.
[17] S.L. McBride, J.W. MacLachlan, B.P. Paradis, Acoustic emission and inclusion fracture in 7075 aluminum alloys, J. Nondestruct. Eval. 2 (1) (1981) 35–41.
[18] T.M. Roberts, M. Talebzadeh, Acoustic emission monitoring of fatigue crack propagation, 2003, p. 18.
[19] A. Berkovits, D. Fang, Study of fatigue crack characteristics by acoustic emission, Eng. Fract. Mech. 51 (3) (1995) 401–416.
[20] J. Weiss, W. Ben Rhouma, T. Richeton, S. Dechanel, F. Louchet, L. Truskinovsky, From mild to wild fluctuations in crystal plasticity, Phys. Rev. Lett. 114 (2015) 105504.
[21] N. Godin, S. Huguet, R. Gaertner, L. Salmon, Clustering of acoustic emission signals collected during tensile tests on unidirectional glass/polyester composite using supervised and unsupervised classifiers, NDT & E Int. 37 (4) (2004) 253–264.

[22] M. Moevus, N. Godin, M. R'Mili, P. Reynaud, D. Rouby, G. Fantozzi, G. Farizy, Analysis of damage mechanisms and associated acoustic emission in two SiC/[Si-B-C] composites exhibiting different tensile behaviours. Part I: Damage patterns and acoustic emission activity, Compos. Sci. Technol. (2008) 1250–1257.
[23] A. Sibil, N. Godin, M. R'Mili, E. Maillet, G. Fantozzi, Optimization of acoustic emission data clustering by a genetic algorithm method, J. Nondestruct. Eval. 31 (2) (2012) 169–180.
[24] S. Deschanel, W. Ben Rhouma, J. Weiss, Acoustic emission multiplets as early warnings of fatigue failure in metallic materials, Sci. Rep. 7 (1) (2017) 13680.
[25] S. Deschanel, J. Weiss, Contrôle de la fatigue des matériaux par émission acoustique, 2018, p. 20.
[26] R.J. Geller, C.S. Mueller, Four similar earthquakes in central California, Geophys. Res. Lett. 7 (10) (1980) 821–824.
[27] Hirokazu Moriy, Hiroaki Niitsuma, Roy Baria, Multiplet-clustering analysis reveals structural details within seismic cloud at the soultz geothermal field, France, p. 31.
[28] G. Poupinet, W.L. Ellsworth, J. Frechet, Monitoring velocity variations in the crust using earthquake doublets: An application to the Calaveras Fault, California, J. Geophys. Res.: Solid Earth 89 (B7) (1984) 5719–5731.
[29] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD '96, AAAI Press, 1996, pp. 226–231.
[30] N. Begum, L. Ulanova, H.A. Dau, J. Wang, E. Keogh, A general framework for density based time series clustering exploiting a novel admissible pruning strategy, 2016, p. 19.
[31] P. Trunfio, Service-Oriented Distributed Knowledge Discovery, in: Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, vol. 20121229, Chapman and Hall/CRC, 2012.
[32] Y. Rui, Z. Zhou, X. Cai, L. Dong, A novel robust method for acoustic emission source location using DBSCAN principle, Measurement 191 (2022) 110812.
[33] J. Hillenbrand, J. Detroy, J. Fleischer, Investigation of defects in roll contacts of machine elements with acoustic emission and unsupervised machine learning, IOP Conf. Ser.: Mater. Sci. Eng. 1193 (1) (2021) 012085.
[34] Qingfeng Xue, Yibo Wang, Hongyu Zhai, Xu Chang, Automatic identification of fractures using a density-based clustering algorithm with time-spatial constraints, Energies 11 (3) (2018) 563.
[35] J.G. Proakis, D.G. Manolakis, Digital Signal Processing, third ed., Prentice Hall, 1996.
[36] J.H. Kurz, F. Finck, C.U. Grosse, H.-W. Reinhardt, Similarity matrices as a new feature for acoustic emission analysis of concrete, in: EWGAE 2004 Proceedings, 2004, p. 7.
[37] I. Grabec, Application of correlation techniques for localization of acoustic emission sources, Ultrasonics 16 (3) (1978) 111–115.
[38] Lyudmyla Kirichenko, Tamara Radivilova, Anastasiia Tkachenko, Comparative analysis of noisy time series clustering, p. 13.
[39] Sangeeta Rani, Geeta Sikka, Recent techniques of clustering of time series data: A survey, Int. J. Comput. Appl. 52 (15) (2012) 1–9.
[40] J. Paparrizos, L. Gravano, Fast and accurate time-series clustering, ACM Trans. Database Syst. 42 (2) (2017) 1–49.