Online segmentation of acoustic emission data streams for detection of damages in composites structures in unconstrained environments

V. Placet[1] & E. Ramasso[2] & L. Boubakar[1] & N. Zerhouni[2]
FEMTO-ST Institute, UMR CNRS 6174 - UFC / ENSMM / UTBM,
[1]*Department of Applied Mechanics,*
[2]*Department of Automatic Control and Micro-Mechatronic Systems,*
24 rue Alain Savary, F-25000 Besançon, France.

ABSTRACT: An approach for unsupervised damage detection in ring-shaped Organic Matrix Composites (OMC) under loading based on acoustic emissions (AE) is proposed. It relies on a specific clustering algorithm called Gustafson-Kessel (GK) that manages fuzzy memberships to clusters and complex cluster's shape. A methodology is proposed to 1) make the algorithm robust to initialisation in order to obtain reproducible results and reliable statistical models representing OMC damages, 2) detect and assess AE activity (AEA) over time for AE data mining to emphasize the more relevant AE data in a huge amount of AE hits, 3) adapt the statistical models based on statistical process control using imprecise updating rate automatically tuned.

1 INTRODUCTION

Composite materials are getting increasing importance in structural applications and require reliable design rules. Most of the current theories and tools have insufficient accuracy to predict the deformation, strength and damage of composite materials exposed to a combination of in-service multiaxial stresses and environmental loadings (Kaddour and Hinton 2012). Typically the observed failure consists of inter-fibre matrix cracking, fibre breakage and a variety of interfacial failure (like fibre-matrix debonding, splitting or inter-ply delamination). These damages are almost always accompanied by releases of heat and stress-wave propagation due to microstructural changes.

Acoustic emissions (AE) are relevant for damage detection and the monitoring of their evolution (Huguet et al. 2002, Momon et al. 2012). Analysis of AE signals is performed by pattern recognition techniques (PRT) which are considered as suitable tools to identify distinct "groups" in AE signals based on a large number of relevant features. The groups are called "clusters" when unsupervised PRT are applied and, in that case, PRT are called clustering algorithm, whereas groups are called classes when using supervised or partially supervised PRT (Momon et al. 2012, Ramasso et al. 2012). Clusters can then be associated to damages using knowledge from the field (Fig. 1).

The detection of clusters highly depends on several conditions, such as the experimental configuration, the material, the geometry of the specimen and, as well, on the existence of AE sources not correlated to specimen failure. AE signals are also influenced by attenuation, dispersion and the position of the source. In this context, supervised PRT would build particular damage models dependent on these conditions, and thus, unsupervised PRT appear more suitable for AE analysis in environments with limited constraints.

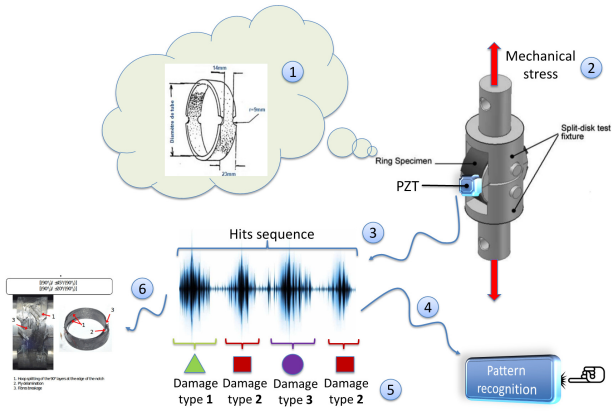


Figure 1: Bridging the gap from AE data to damages illustrated: The ring-shaped specimen (1) considered is submitted to quasi-static loading (2) that generates AE data (3) which are analysed using PRT (4) in order to estimate data-driven damage models (5). The test is performed until failure (6).

The main objective of this work is unsupervised damage detection using clustering algorithms, where each cluster is supposed to represent a damage family. Detecting damages accurately is a difficult problem since one may face with several challenges:

Challenge 1 The *choice of features*. According to the

algorithm used for damage detection, different subsets of features may lead to different results.

Challenge 2 *The number of damage families* is not always well defined and well known.

Challenge 3 *Robustness of algorithms to initialisation* of clustering algorithms has to be ensured for practical real-life applications in order to retrieve results easily.

Challenge 4 The *revision* of models obtained by clustering without re-training (using all past data but only the current ones).

In this paper, we consider tubular composite structures (described in Section 2) submitted to quasi-static loading up to failure. The detection and monitoring of damages is based on AE data for which some specific unsupervised PRT are proposed to tackle challenges 1 and 4 cited before:

Batch estimation of damage data-driven models (Fig. 2). This first tool is proposed to analyse the data in batch mode, i.e. using all data from an experiment after failure (Section 3).

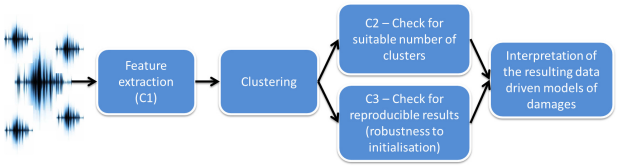


Figure 2: Bridging the gap from AE data to damages by building statistical modelling in batch mode.

Online updating of damage data-driven models (Fig. 3) to take into account possible changes in the structure over time. The updating is performed online without re-training, using a suitable criterion for data *cleansing* that drastically decrease time-consumption (Section 4).

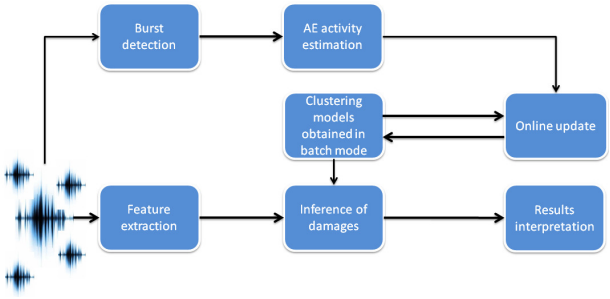


Figure 3: Models updating: online processing.

2 EXPERIMENTAL PLATFORM

This work deals with the health assessment of tubular composite structures. Such structures are used in many application fields, such as speed rotors, flywheels, pressure vessels, transportation systems and so on. Their stress state is most of the time complex (multiaxial and heterogeneous) due to the combination of loads which makes particularly difficult the prediction of damage occurrence. Health was assessed on composite split disks submitted to quasi-static loading up to failure. The tests were performed according to ASTM D2290 "Apparent hoop tensile strength of plastic or reinforced plastic pipe by split disk method". Rings were produced by machining filament-wound carbon fibre reinforced epoxy tubular structures intended for the manufacturing of flywheel rotors with a $[(90^o)_6]$ lay-up configuration.

The transient elastic waves were recorded at the material surface using a multi-channels data acquisition system from Euro Physical Acoustics Corp. (MISTRAS Group). The system is made up of miniature piezoelectric sensors with a range of resonance of $250 - 325$kHz, preamplifiers with a gain of 40dB and a $20 - 1000$kHz filter, a PCI card with a sampling rate of 2MHz and the AEWin software. Two sensors were coupled on the specimen faces using a silicon grease. The calibration of the system was performed after installation of the transducers on the specimen and before each test using a pencil lead break procedure. A part of the ambient noise was filtered using a threshold of 45dB. The acquisition parameters: PDT (Peak Definition Time) = 60μsec; HDT (Hit Definition Time) = 120μsec and HLT (Hit Lock Time) = 300μsec were identified using preliminary measurements. A damage scenario for each specimen was presented (Placet et al. 2012) using infrared thermography, optical observation and analysis of the mechanical behaviour. Fig. 4 is a partial scenario for the considered specimen.

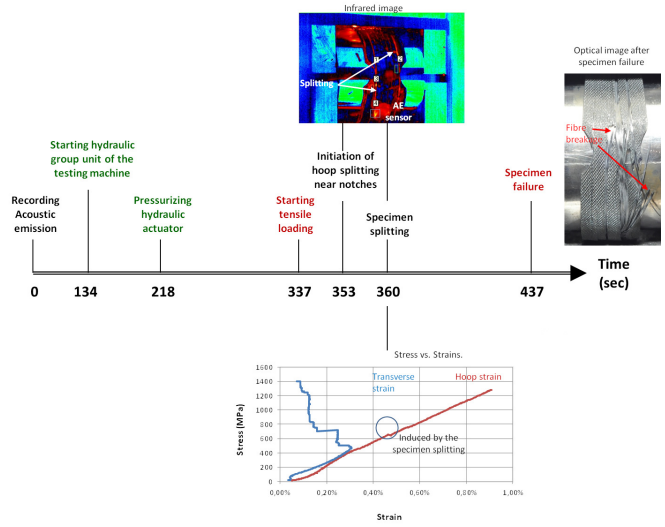
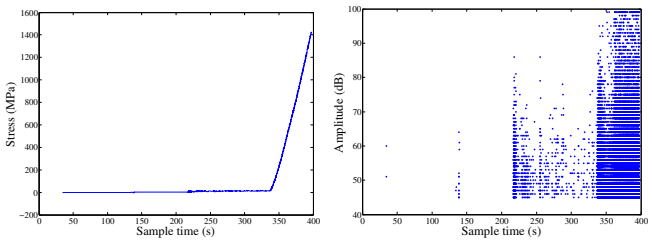


Figure 4: The scenario.

The stress vs. time is depicted in Figure 5(a) as well as the amplitude of acoustic emissions in Figure 5(b). The latter shows many AE hits all along the experiment (48000 data points) with a huge proportion characterised by low values (< 75dB). As expected, most of the AE activity is present after 337sec. but a large

number of AE hits are detected when the hydraulic actuator is being pressurized (> 218sec.). The goal of the clustering algorithm is to find out which type of damage appear at each time despite this huge amount of data. In particular, the size of the clusters are not expected to be the same, i.e. the number of AE hits per damage can differ from a damage to another. Detection of damage based on AE signals in a noisy environment is particularly crucial for industrial application. An increase in the amplitude and frequency thresholds could filter these noise signals but also important events related to the matrix damage.



(a) Stress (b) Amplitude

Figure 5: Stress and features vs time.

The next section is dedicated to the presentation of the Gustafson-Kessel algorithm investigated in this paper to establish a scenario by processing AE.

3 BATCH SEGMENTATION

In the following K denotes the number of clusters.

3.1 *Why the Gustafson-Kessel algorithm? Which features? Which number of clusters?*

Many approaches for AE data processing presented in the literature rely on the Principal Component Analysis (PCA) to pre-process the data. The PCA is used to "automatically" select a subset of features but it also modifies the feature space by combining the original features. This algorithm assumes that 1) the linear combination of features improves the relevancy of the principal components and 2) a large variance implies meaningfulness.

Other approaches rely on a specific subset of features, for example frequency-based ones (Sause et al. 2012) or other subsets selected by prior knowledge such as energy, rise time, duration, amplitude, and so on (Gutkin et al. 2011). The subsets, together with the number of clusters, can also be selected by a greedy approach consisting in applying the clustering algorithm on combinations of features and selecting the one which maximises a given criterion (Halkidi et al. 2001, Sause et al. 2012). The goal of the criterion is generally to evaluate the quality of the partition provided by the clustering. Most of criterion are based on the Euclidean distance to assess the membership of a AE hit to a given cluster. However, the applicability of this approach is limited to clustering algorithms which are based on the Euclidean distance.

The PCA is generally used jointly with the K-means. The main reason to account for the performance of this couple is actually due to the link between both tools (Ding and He 2004). Compared to usual approaches based on K-means or FCM (Momon et al. 2012) that use the Euclidean distance, the GK algorithm investigated here takes the distribution of the data points into account with a modified *Mahalanobis distance* for each cluster which is iteratively adapted to fit ellipse-shaped clusters. By looking at some results of AE segmentation in the literature, the use of ellipses seems more adapted than circles to represent AE data. In the GK algorithm, the covariance between each pair of features is estimated so that possible redundancy or complementarity between features can be taken into account.

In this paper, GK is investigated using a specific subset of features obtained from AEWin software: ***E**nergy*, ***A**bsolute energy*, ***A**mplitude*, ***R**everberation frequency* and ***A**verage frequency*. As proposed in (Barat et al. 2010), a median filter (size 9) was applied, to remove some spikes in particular AE hits due to electric and electromagnetic noise. To keep important information contained in energy-based features, only the two others were filtered. Compared to average (such as performed in PCA), the median does not create new artificial values from original features. The features were normalised using a simple standardisation because each feature does not contribute equally to the multidimensional analysis due to the scale. It is a feature-wise and reversible process that enables PRT to converge more easily.

3.2 *Robustness to initialisation using the ARI*

A proposed methodology to cope with the problem of initialisation is summarized in Figure 6.

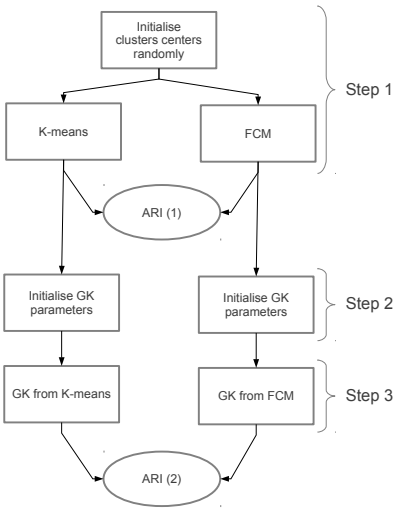


Figure 6: Method to initialisation GK.

To illustrate the problem and the solution, let consider the set of the aforementioned features. An initial position of $K = 4$ and 6 clusters was selected randomly, and then K-means and FCM were run. Both generated partitions were compared using the Adjusted Rand Index (ARI) (Nguyen et al. 2009) which is a value in $[-1, 1]$ that tends to 1 when the two partitions are close. To emphasize the impact of initialisations, we consider 100 initial clusters' centers and run K-means and FCM on features. The initial positions were drawn randomly: First, K values were obtained by dividing the scale of each feature space into K equally-spaced groups, 50 samples were drawn with report to a uniform distribution and 50 other samples drawn according to a gaussian distribution with unit variance both centered on the K groups' center (Fig. 7). The partition from K-means and FCM were compared the 100 runs using the ARI.

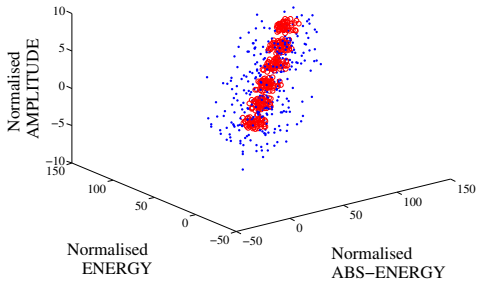


Figure 7: Sampling initial data points in feature space ($K = 6$).

At least, when using the same initial position of clusters, both algorithms are expected to provide an ARI close to 1. On the opposite, K-means and FCM generate different partitions with an ARI spread only $[0, 1]$ with mean 0.522 ± 0.11 (Fig. 8). GK was run 100 times using the clusters' position of K-means and FCM (as proposed in Fig. 6). Since the partitions of both latter algorithms are generally different (with report to the ARI), one can expect different partitioning results using GK. On the opposite, the latter performs well whatever the initialisation is provided either by K-means or FCM (ARI close to 1 for every cases, Fig. 8). Figure 8 also emphasizes that $K = 6$ led to an ARI with limited variance compared to $K = 4$ for K-means, while GK was able to converge in both cases.

Conclusively, according to the ARI, GK seems the most robust clustering algorithm compared to K-means and FCM with report to initialisation conditions when considering this specimen. Our many experiments with different configurations of the specimens led us to the same conclusion. Moreover, the partition generated by GK (with the proposed initialisation) seems closer to the expected behavior of the structure than the other algorithms (see next sections).

3.3 *Validation of the number of clusters*

One possibility to select a well suited value for K, is to try different K and select the one that optimises a given criterion. Among all criterion available in the literature, the so-called Davies and Bouldin index (DB) (Halkidi et al. 2001) is frequently applied to quantify the "goodness" of a partition of AE data. The DB index depends on the average distances of all patterns in clusters i and j to their respective cluster's centers. We propose to use the same distance as in the clustering algorithm, that is here the modified Mahalanobis distance given by $d_{ik}^2 = \|x_k - v_i\|_{A_i}^2 = (x_k - v_i)^T A_i (x_k - v_i)$, where $A_i = [\rho_i \det(F_i)]^{1/q} F_i^{-1}, i = 1 \dots K$, ρ_i is the cluster volume of the i-th cluster (generally assumed equal to 1), F_i is the fuzzy covariance matrix given by $F_i = \sum_{k=1}^{N} (u_{ik})^\beta (x_k - v_i)(x_k - v_i)^T / \sum_{k=1}^{N} (u_{ik})^\beta$ where β sets the fuzzyness of the partition (generally assumed equal to 2), N the total number of points, u_{ik} is degree of membership of data point x_k to the i-th cluster and v_i is the i-th cluster's center (see (Gustafson and Kessel 1978) for details).

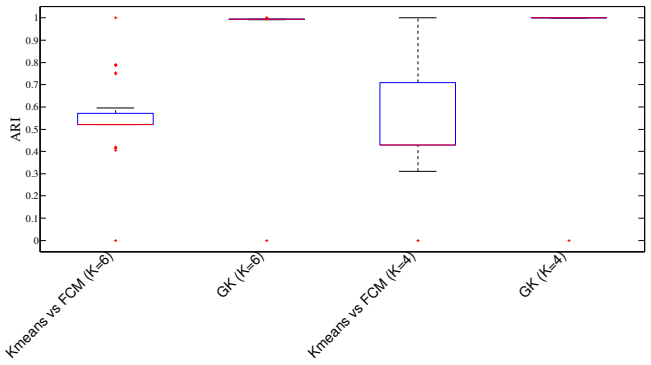


Figure 8: ARI for K-means vs FCM, and for GK with $K = 4$ and 6. Boxplot represents the median value, first and third quartile and extreme points.

To estimate the number of clusters, we considered two main configurations: *Conf1* – GK with selected features, *Conf2* – K-means+PCA as usually applied for AE data analysis. The PCA was applied on a subset of features (after standardisation) made of rise time, counts, energy, duration, amplitude, average frequency, RMS, ASL, reverberation and initiation frequency, strength and absolute energy (available in AEWin software). For the PCA and K-means algorithms, the built-in MATLAB functions were used with "replicate" and "singleton" options for improved convergence and empty clusters management. For these two cases, the DB index was estimated (on average over 10 runs):

Conf1 For GK with selected features (and median filtering): 0.472 ± 0.027 ($K = 4$), 0.513 ± 0.03 ($K = 5$) and 0.582 ± 0.02 ($K = 6$). ARI equal to 1 in all cases in one iteration showing that the data were easily partitioned using GK for the three cases. $K = 4$ appeared DB-optimal.

Conf2 For K-means+PCA (without filtering): 0.9205 ± 0.02, 0.9602 ± 0.01, 1.0123 ± 0.07 for $K = 4, 5, 6$. $K = 6$ appeared DB-optimal.

With the PCA, GK does not generally led to an ARI equal to 1 meaning that the data may not be well sep-

arated. In the sequel, $K = 6$ was selected for comparison purpose between K-means+PCA and GK.

3.4 *Sequence of damages*

To compare both K-means and GK approaches, several tests were performed. Figures 9 and 10 present the sequence obtained with GK and K-means respectively. The vertical axis corresponds to the log CSCA that is the logarithm of the cumulative occurrence of AE hits in a given cluster (see (Ramasso et al. 2012) for details). The number of AE hits per cluster for all configurations is given in Table 1. The cluters are also represented in the Duration/Amplitude space (Fig. 11) for better understanding.

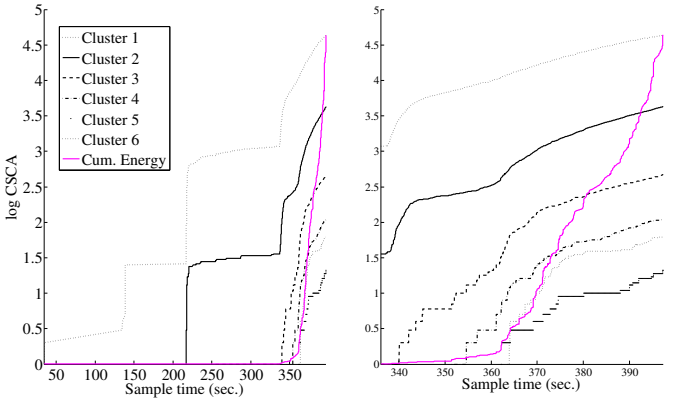


Figure 9: GK with selected features: damage sequence. On the right, a zoom of the graph on the left between 340 and 400 sec.

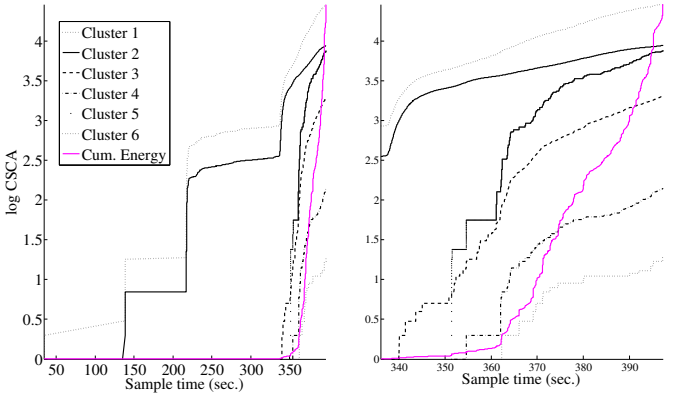


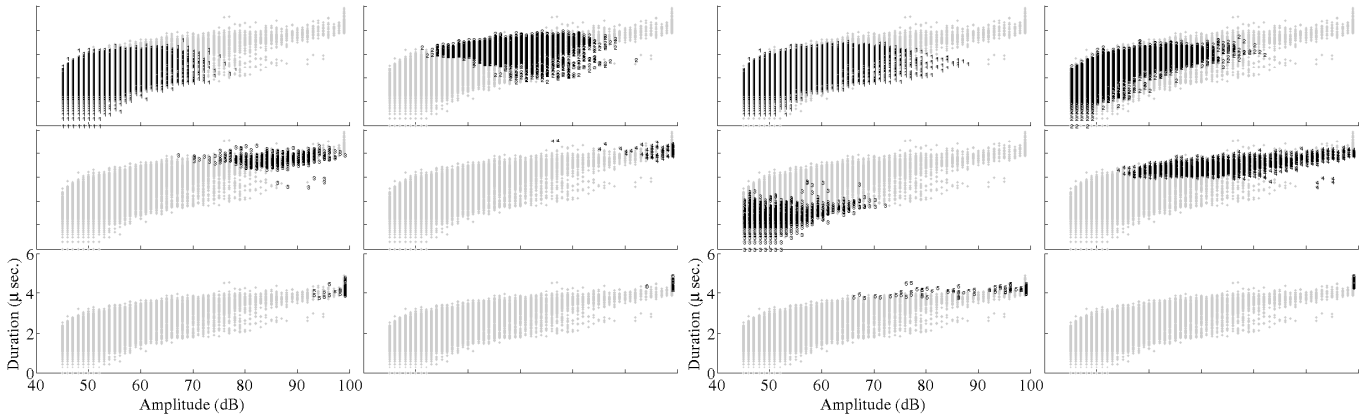
Figure 10: K-means+PCA: damage sequence.On the right, a zoom of the graph on the left between 340 and 400 sec.

Conf1 / Fig. 9 – GK with selected features provided a sequence of damages that began at 34sec. by the cluster corresponding probably to electric and electromagnetic noise. In fact, at this time the machine was turned off and the recorded events probably originated from the experimental apparatus surrounding the tensile machine. The number of hits increased significantly when the hydraulic power unit of the machine was turned on, at 134sec. This cluster was made of 89% of AE hits (Tab. 1) and was characterised by low durations and low amplitudes (Fig. 11(a)), which are consistent with the characteristics of the electric noise and electromagnetic interferences. The second cluster appeared around 218sec., when the hydraulic actuator of the tensile machine was pressurized. This latest was attributed to the hydraulic noise generated by the servo valve of the testing machine. This was consistent with the increase in activity in this cluster observed after 337sec., when the load was applied to the specimen. Cluster 2 was made of 9% of the AE hits and was located at mid amplitude and durations. Several seconds after starting the tensile test (around 342sec.), cluster 3 appeared. This one is obviously describing the rubbing of the specimen on the clamps (half-cylinder) since until 350sec. the tensile stress did not significantly increase when the actuators moved widely. This self-alignment of the clamps can generate much friction between the connecting arms, the pins, the half-cylinders and the specimen. Amplitudes and durations were higher in this cluster than in the two previous ones. Cluster 4 started around 353sec., when the stress started forthrightly to increase (i.e. after the self-alignment of the clamps). The number of hits increased widely in this cluster up to mid-load. This cluster was attributed to the matrix micro-cracking. The evolution of the log CSCA of this cluster was characterised by "stairs", during the first part of the tensile test, corresponding to a high number of AE hits at a given instant, probably when micro-cracking coalesced. Cluster 5 was likely to correspond to the hoop splitting. This one started around 360sec. in agreement with the optical and infrared observations (Fig. 4) and also in agreement with the change in sign of the transverse strain also represented in Fig. 4. For this notched ring, numerical simulations clearly showed a positive transverse strain in the gauge part of the specimen, while a negative strain was calculated for unnotched rings. This change of sign experimentally observed traduces the complete hoop splitting of the specimen at the edge of the notches. The hits in this cluster were characterised by high amplitudes and durations which were typically attributed in the literature to interfacial failure. Almost simultaneously cluster 6 started. The log CSCA was characterised by a few "stairs" corresponding to rare events (with respect to the total number of hits). This cluster was composed of only 21 hits with the highest amplitudes and energies. One can also observe that these stairs corresponded to huge increases in the cumulated energy (Fig. 4). This cluster undoubtly represents the fibres breakage.

Conf2 / Fig. 10 – The sequence provided by K-means was quite different. Only the two last clusters were similar, may be due to their specific characteristics (in particular highly energetic). The four first clusters had a much higher proportion than with Conf1. Clusters 1 and 2 appeared frequently and were made of many AE hits (67% and 16%, Tab. 1). Amplitudes and durations spread from 45dB to 88dB and covered $2/3$ of the duration range. The occurrence of cluster 2 around 135sec. corresponded to the hydraulic group unit starting that may not generate dam-

Conf1	42973 (89.73%)	4242 (8.86%)	467 (0.98%)	111 (0.23%)	62 (0.13%)	21 (0.04%)
Conf2	32297 (67.46%)	7762 (16.21%)	5470 (11.43%)	2187 (4.57%)	141 (0.29%)	19 (0.04%)

Table 1: Number of AE hits in each cluster for the four configurations (sorted by size): GK + selected features (Conf1), Kmeans+PCA (Conf2).



(a) Conf1: GK + selected features (col. 1-2, lines 1-3)

(b) Conf2: K-means + PCA

Figure 11: Duration (y, log., $\in [0, 4.9]$) vs Amplitude (x, dB, $\in [45, 99]$) with clusters for Conf1 and Conf2 (3 lines and 2 column for each). Clusters numbers appear in each subfigure.

ages. In GK, this event was classified in the first cluster. Cluster 3 was made of low amplitudes and durations and started early (338sec.) corresponding to the tensile test starting. Thus, this cluster may be related to rubbing and friction noise. Cluster 4 started when the scissions were observed (Fig. 4), with "stairs" during splitting. Compared to GK, the number of AE hits was quite important meaning that more than 5400 AE hits were related to scissions while this value is divided by more than 10 with GK.

4 ONLINE UPDATING

The online GK algorithm was proposed in (Georgieva and Filev 2009) for online data streams partitioning, i.e. clusters' parameters are updated as new data points arrive. The possibility to add/remove clusters (Serir et al. 2012) is not considered here. The main contribution is a method to adapt the updating rate automatically.

4.1 *Acoustic Emission Activity (AEA) estimation*

When looking at the time-index of recorded waves (Fig. 12(a)), more or less AE hits appear. The acquisition time is thus irregular and continuous. Making it discrete (here with sample time $\tau = 0.25$sec.), one obtains the result in Figure 12(b) that characterises the AE activity (AEA). We observe such a pattern in several configurations with 5 main phases: 1) an increasing phase (accomodation), 2) a decreasing phase, 3) a steady phase, 4) a gradual increasing phase (possibly damages) and 5) some peaks with decreasing trend (fibre breakage). The peaks (Fig. 12(b)) correspond to an activity level during which some relevant AE hits may appear. The level can thus be used for the segmentation of AE data (called Online Segmentation Algorithm, OSA).

Formally, let $t \in \Re$ the acquisition time, $T = \{t_1, t_2 \dots t_j \dots t_N\}$ the set of time instants of AE hits, and $T(j)$ the j-th element. A bin of time-instants is $B_i = \left[B_i^-; B_i^+\right), i = \{1 \dots R\}$ with $B_i^- = T(1) + (i-1)\tau$ and $B_i^+ = T(1) + i\tau$, i.e. a bin is defined as an interval with width τ, and R is the total number of bins. The number of AE hits in the i-th bin is

$$H(i) = \sum_{j=1}^{N} \mathbb{1}\left(B_i^- \leq T(j) < B_i^+\right), i = 1 \dots R. \quad (1)$$

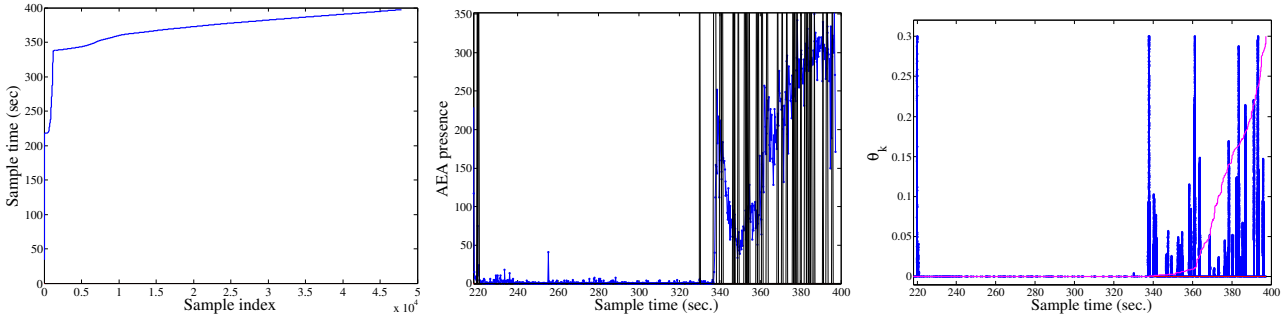
The level of activity is quantified by merging consecutive bins B_i and B_{i-1} in increasing trend until a peak. Let $\mathcal{I} = \{I_1, I_2 \dots I_j \dots\}$ the set of bins after merging at least two consecutive bins:

$$I_l = \{B_i \in \mathcal{B} \mid H(i) \geq H(i-1), B_{i-1}^+ = B_i^-\} \quad (2)$$

If a bin is composed only of data points belonging to the "largest" cluster – i.e. the more frequent cluster, corresponding possibly to noise or to phenomenon which have a limited impact on the structure, for example, the cluster 1 in Fig. 11(a) – the AEA is not taken into account. This simple rule allows one to filter out some AEA, for example, in Fig. 12(b), no AEA was kept between 218sec. and 338sec. despite peaks.

4.2 *Evolving GK (EGK) algorithm*

Let q the dimension of the feature space (here $q = 4$), $x_k \in \Re^q$ is the feature vector at instant k (for example amplitude, energy, absolute energy, reverberation



(a) Time index

(b) AEA presence (black intervals) and discrete time (continuous line)

(c) Level of AE activity

Figure 12: Making time discrete, detection of AEA, and AEA's level for the first specimen.

frequency and average frequency). Given x_k, let p the closest cluster given by minimizing the Mahalanobis-like distance to each cluster, i.e. $p = \text{argmin}_{c=1}^{K} d_{ck}^2$. The p-th cluster's updating is performed by modifying its parameters $v_{p,k}$ (center) and $F_{p,k}$ (covariance) for the current time k. As in Kohonen's rule, the closest cluster's center is moved towards x_k:

$$v_{p,k+1} = v_{p,k} + \theta\Delta \ , \tag{3}$$

where $\Delta = x_k - v_{p,k}$ and $\theta \in]0,1[$ is the updating rate. The inverse of the fuzzy covariance matrix as well as its determinant, which are *used to estimate the distance to each cluster*, are recursively adapted:

$$F_{p,k+1}^{-1} = (I - G\Delta)\, F_{p,k}^{-1}\, (1-\theta)^{-1}\,, \tag{4}$$

where I is the identity matrix and $G = (1-\theta)^{-1} F_{p,k}^{-1} + \Delta^{\mathrm{T}} \left(\theta^{-1} + \Delta\,(1-\theta)^{-1} F_{p,k}^{-1} \Delta^{\mathrm{T}}\right)^{-1}$, and let $\xi = 1 - \theta + \theta\Delta F_{p,k}^{-1}\Delta^{\mathrm{T}}$, the determinant is

$$\det\left(F_{p,k+1}\right) = (1-\theta)^{n-1} \det\left(F_{p,k}\right) \xi\ . \tag{5}$$

4.3 *Automatic tuning of the updating rate θ*

Let θ_k the updating rate at time-instant t_k:

$$\theta_k \in [\theta_{\min}; \theta_{\max}], \theta_{\min} > 0, \theta_{\max} < 1 \tag{6}$$

According to authors (Georgieva and Filev 2009), $\theta \in [0.05, 0.3]$ and is constant. We propose to modify its value according to the AEA. The updating rate is considered imprecise by considering a possible range $[\theta_{\min}; \theta_{\max}]$. When high AEA is present and when it is not due only to noise, the updating rate should be high ($\theta_{\max}$), while it should be low ($\theta_{\min}$) when no AEA is detected.

To estimate the value of θ_k with report to AEA, we make it dependent on the number of hits in a bin. This value has to be transformed into a normalised value in $[0,1]$ that will be scaled onto $[\theta_{\min}; \theta_{\max}]$. In online mode, the maximum number of AE hits is not known in advance so it is proposed to normalise the number of hits by using a time-window with length W sec. over previous peaks. Let λ_k the resulting signal:

$$\lambda_k = \frac{\#\text{ hits in the peak}}{\text{maximum }\#\text{ hits in the time window}} \tag{7}$$

Using statistical process control (Angelov et al. 2010) (Chap. 12), the Mahalanobis-like distance can be upper-bounded by $\chi^2_{q,\beta}$, i.e. the maximal distance for a given probability of false alarm $1-\beta$, where $\chi^2_{q,\beta}$ is the chi-square distribution with q degrees of freedom. Beyond that distance, the null hypothesis "the process is in control" is rejected and thus the current data point is not considered. As proposed in (Angelov et al. 2010) (Chap. 12), β was set to 0.0455. Given the Mahalanobis-like distance d_{pk} between point x_k and the closest cluster (p), we thus assert whether data point x_k is located out of the cluster:

$$\mathcal{M}_{p,k} = \min\left(1; \frac{d_{pk}^2}{\left(\det\left(F_{p,k}\right)\right)^{1/q} \chi^2_{q,\beta}}\right) \tag{8}$$

such that if point x_k is a false alarm then the distance $\mathcal{M}$ is set to 1, otherwise the assigned value is the normalised Mahalanobis-like distance. The value of θ_k is finally estimated by:

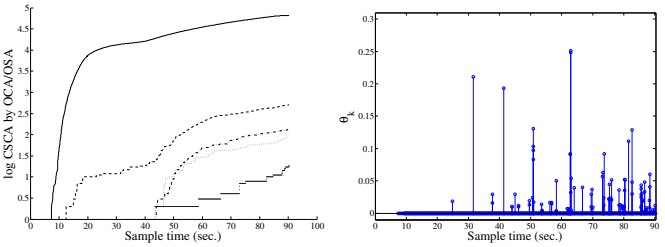
$$\theta_k = \frac{\lambda_k\left(1 - \mathcal{M}_{p,k}\right)}{1 + \left(\dfrac{t - t_0}{\gamma}\right)^2} \tag{9}$$

where t_0 and γ represent the location and the scale of the Cauchy function used to adjust the imprecision of θ_k within a bin. If the current instant k belongs to the bin I_l, the most important part of the bin is assumed to be the median value of time-instants in that bin ($\text{median}_{j \in I_l}\, I_l(j)$). The scale γ is traditionally given by the difference between the third and first quartile of the time-instants in the bin. The computation of these parameters requires to wait for a complete burst

which may imply a small time delay (a few milliseconds in the worst case). An illustration is depicted in Figure 12(c) for the considered specimen.

To illustrate the online clustering (OCA) with online segmentation (OSA), we estimated the cluster/damage at each time-step given AE data from a *new* specimen with a similar lay-up configuration. The value of θ_k over time for the new specimen was found using the AEA-based approach proposed above, and this value was used to update clusters parameters obtained previously. Figure 13(a) shows the evolution of the sequence over time estimated gradually (point after point) with continuous updating. Cluster 1 did not occur (while being the most frequent in Fig. 9) but cluster 2 was the most frequent here. This cluster corresponded also to noise in Figure 9 (or to a cluster not linked to damages).

The updating rate is depicted in Figure 13(b) where only some data points were allowed to update the models, due to the false alarm detection procedure that only kept data points close to clusters. Imprecision about the parameter is well-managed since the value of θ_k varies from 0 to 0.3 according to the proximity to clusters.



(a) Online clustering algorithm (OCA) results

(b) θ_k adapted over time

Figure 13: Second specimen, OSA / OCA results.

5 CONCLUSION AND FUTURE WORK

From this study, several observations can be listed:

The choice of the clustering validity index to assess the number of clusters K depends on the algorithm and on features. However, both K and features have to be linked with expected damages.

Several algorithms starting with the same initial points are not able to converge to the same solution. The PCA seems to increase this difference and, thus, does not seem to be adapted to the specificities of AE data.

The proposed methodology using GK and some selected features emphasized the accomodation phase, micro-cracking, scissions/splitting and fibre breakages. GK manages uncertainty using ellipse-shaped clusters and always converged to same solution given different initialisations.

The Online Segmentation Algorithm (OSA) allows AE data mining by detecting some useful AE hits based on the acoustic activity (AEA).

The Online Clustering Algorithm (OCA) allows to detect the clusters as new data arrive based on an updating rate automatically set within a specified range to manage imprecision about its value.

ACKNOWLEDGMENT

This work has been supported by the Labex ACTION project (contract “ANR-11-LABX-01-01”).

REFERENCES

Angelov, P., D. Filev, & N. Kasabov (2010). *Evolving Intelligent Systems: Methodology and Applications*. IEEE Press Series on Computational Intelligence.

Barat, V., Y. Borodin, & A. Kuzmin (2010). Intelligent AE signal filtering methods. *Journal of Acoustic Emission 28*, 109.

Ding, C. & X. He (2004). K-means clustering via principal component analysis. In *Int. Conf. on Machine learning*.

Georgieva, O. & D. Filev (2009). Gustafson-kessel algorithm for evolving data stream clustering. In *Int. Conf. on Computer Systems and Technologies - CompSysTech 09*.

Gustafson, E. & W. Kessel (1978). Fuzzy clustering with a fuzzy covariance matrix. In *IEEE Conf. on Decision and Control*.

Gutkin, R., C. Green, S. Vangrattanachai, S. Pinho, P. Robinson, & P. Curtis (2011). On acoustic emission for failure investigation in CFRP: Pattern recognition and peak frequency analyses. *Mech. Syst. and Signal Processing 25*, 1393–1407.

Halkidi, M., Y. Batistakis, & M. Vazirgiannis (2001). On clustering validation techniques. *Journal of Intelligent Information Systems 17*, 107–145.

Huguet, S., N. Godin, R. Gaertner, L. Salmon, & D. Villard (2002). Use of acoustic emission to identify damage modes in glass fibre reinforced polyester. *Composite Science Technology 62*, 1433–1444.

Kaddour, A. & M. Hinton (2012). Benchmarking of triaxial failure criteria for composite laminates: Comparison between models of ’Part (A)’ of ’WWFE-II’. *J. Compos. Mater. 46*, 2595–2634.

Momon, S., N. Godin, P. Reynaud, M. RMili, & G. Fantozzi (2012). Unsupervised and supervised classification of AE data collected during fatigue test on CMC at high temperature. *Composites Part A: Applied Science and Manufacturing 43*, 254–260.

Nguyen, X., J. Epps, & J. Bailey (2009). Information theoretic measures for clustering comparison: Is a correction for chance necessary? In *Int. Conf. on Machine Learning*.

Placet, V., F. Trivaudey, & M. Boubakar (2012). On the relevance of using a notched specimen to determine the hoop strength of composites with the split disk method. *Composite Structures Journal*. Under review.

Ramasso, E., V. Placet, R. Gouriveau, L. Boubakar, & N. Zerhouni (2012). Health assessment of composite structures in unconstrained environments using partially supervised pattern recognition tools. In *Annual Conference of the Prognostics and Health Management Society*.

Sause, M. G. R., A. Gribov, A. R. Unwin, & S. Horn (2012). Pattern recognition approach to identify natural clusters of acoustic emission signals. *Pattern Recognition Letters 33*, 17–23.

Serir, L., E. Ramasso, & N. Zerhouni (2012). Evidential evolving Gustafson-Kessel algorithm for online data streams partitioning using belief function theory. *International journal of approximate reasoning 53*, 747–768.