Pattern recognition approach to identify natural clusters of acoustic emission signals

M.G.R. Sause [a,*], A. Gribov [b], A.R. Unwin [b], S. Horn [a]

[a] University of Augsburg, Institute for Physics, Experimental Physics II, D-86135 Augsburg, Germany
[b] University of Augsburg, Institute for Mathematics, Dept. of Computer Oriented Statistics and Data Analysis, D-86135 Augsburg, Germany

ARTICLE INFO

ABSTRACT

A new approach is introduced to identify natural clusters of acoustic emission signals. The presented technique is based on an exhaustive screening taking into account all combinations of signal features extracted from the recorded acoustic emission signals. For each possible combination of signal features an investigation of the classification performance of the k-means algorithm is evaluated ranging from two to ten classes. The numerical degree of cluster separation of each partition is calculated utilizing the Davies–Bouldin and Tou indices, Rousseeuw's silhouette validation method and Hubert's Gamma statistics. The individual rating of each cluster validation technique is cumulated based on a voting scheme and is evaluated for the number of clusters with best performance. This is defined as the best partitioning for the given signal feature combination. As a second step the numerical ranking of all these partitions is evaluated for the globally optimal partition in a second voting scheme using the cluster validation methods results. This methodology can be used as an automated evaluation of the number of natural clusters and their partitions without previous knowledge about the cluster structure of acoustic emission signals. The suitability of the current approach was evaluated using artificial datasets with defined degree of separation. In addition the application of the approach to clustering of acoustic emission signals is demonstrated for signals obtained from failure during loading of carbon fiber reinforced plastic specimens.

1. Introduction

All sorts of rapid microscopic displacements inside solids are associated with the excitation of small elastic waves. This phenomenon is known as acoustic emission and is of particular interest for material science if the microscopic displacement originates from crack formation or crack propagation.

Acoustic emission analysis is a very powerful tool to detect failure events as a function of mechanical loading, especially for monitoring of fiber reinforced composites. Here the typical microscopic failure mechanisms are matrix cracking, fiber–matrix debonding, fiber pull-out, fiber bridging, inter-ply failure and fiber breakage (Schürmann, 2007). Typically the acoustic emission signals recorded during failure of composite materials originate from more than one of these types of failure mechanism. Without additional information about the origin of acoustic emission signals, the interpretation of the recorded signals is often difficult. In addition, the acoustic emission signals are influenced by attenuation, dispersion and the source position within the specimen (Sause and Horn, 2010a). Therefore source identification based on single signal parameters is not feasible and multivariate analysis techniques are required instead.

One suitable approach to distinguish between different acoustic emission signal types is the application of parameter based pattern recognition techniques. To this end, discrimination between noise and non-noise acoustic emission signals is often achieved by unsupervised pattern recognition techniques (Anastassopoulos and Philippidis, 1995; Philippidis et al., 1998). Acoustic emission signals originating from friction and electromagnetic inductions can easily be identified in this respect, due to their inherent characteristic difference to transient acoustic emission signals.

Various approaches focus on detection of characteristic similarities of the recorded signals (Anastassopoulos and Philippidis, 1995; Philippidis et al., 1998; Richardson et al., 1984; Vi-Tong and Gaillard, 1986; Huguet et al., 2002; Ramirez-Jimenez et al., 2004; Marec et al., 2008; Li et al., 2008; Sause et al., 2009; Sause and Horn, 2010a). Utilizing suitable experimental considerations and finite element simulations the respective signal types can be associated with specific failure mechanisms (Richardson et al., 1984; Bohse, 2000; Haselbach and Lauke, 2003; Huguet et al., 2002; Ramirez-Jimenez et al., 2004; Marec et al., 2008; Li et al., 2008; Sause et al., 2009; Sause and Horn, 2010a).

In terms of pattern recognition techniques, the problem faced is the identification of natural clusters of acoustic emission signals. One of the major problems in this concern is an adequate evaluation of the classification results of different clustering approaches. There are many ways to choose features, normalization procedures

* Corresponding author. Tel.: +49 0 821 598 3238; fax: +49 0 821 598 3411.
  E-mail address: markus.sause@physik.uni-augsburg.de (M.G.R. Sause).

and algorithms that will influence the result of the classification process.

In order to find the natural clusters of a dataset without previous assumptions on the number of clusters and previous feature selection an automated screening of feature combinations is a suitable approach. To evaluate the quality of the partition of a dataset, several cluster validity indices are established in literature (Tou, 1979; Davies and Bouldin, 1979; Rousseeuw, 1987; Hubert and Arabie, 1985). Since each cluster validity index shows different performance depending on the occurrence of outliers and the shape of the clusters, a combined evaluation of multiple cluster validation indices is a possibility to increase the reliability of the cluster identification method. Such combined evaluation was already proposed by Günter and Bunke (2003) in the form of a voting scheme to measure the overall performance of the different indices.

The present publication utilizes the voting scheme of Günter and Bunke (2003) in an exhaustive screening of feature combinations to detect natural classes of acoustic emission signals automatically. In the following section, the pattern recognition methodology is introduced. In the subsequent sections the suitability of the approach is demonstrated for artificial datasets and for experimental datasets of acoustic emission signals.

2. Pattern recognition methodology

The goal of the presented unsupervised pattern recognition methodology is to introduce an automated technique to detect promising feature combinations for clustering of acoustic emission signals. In contrast to feature reduction procedures (e.g. based on correlation dendrograms, Anastassopoulos and Philippidis, 1995) an exhaustive search of global optimal combinations of the signal features used is performed.

As already pointed out by Polikar such exhaustive search methods are conceptually simple but computationally devastating since the number of feature combinations to investigate grows combinatorially with the number of available features (Polikar, 2006).

Consequently, before application of the presented exhaustive search method an educated preselection of promising features for description of the respective acoustic emission signals is suggested. However, in principle the method presented has no limitation regarding the number of features, but the exhaustive screening of a large number of feature combinations will lead to increased computation times. A visualization of the complete method is shown in Fig. 1 as a flowchart diagram.

2.1. Determine all feature combinations

The first step consists of selection of all possible feature combinations with a given minimum number of features M to be used for clustering. The total number N of combinations of K previously selected features is thus:

$$N = \sum_{X=M}^{K} \binom{K}{X}. \tag{1}$$

2.2. Preprocessing

As a second step all preprocessing options for each feature combination are applied. In the current investigation the preprocessing consists solely of feature normalization by dividing by the standard deviation. This is based on preceding investigations on classification of acoustic emission signals (Sause et al., 2008, 2009; Sause and Horn, 2010b). For a generalized approach the settings should
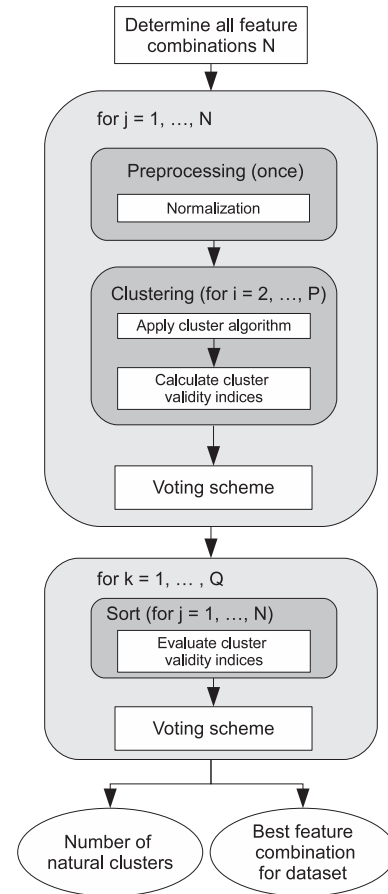

Fig. 1. Flowchart of pattern recognition approach.

be modified to fit the specific needs of the classification problem investigated.

2.3. Clustering

Subsequently the respective partitions for 2, 3, 4, . . . , P clusters are calculated utilizing the k-means algorithm, with P as the maximum number of clusters. To increase the chances of finding the global minimum, MacQueen's algorithm with 10 random cluster center initialisations and a maximum number of 100 iterations to convergence was applied (MacQueen, 1967). As measure of similarity, the Euclidean distance was used. For each partition the Tou-index τ, the Davies–Bouldin-index R, Rousseeuw's silhouette value S, and Hubert's Gamma statistic Γ were calculated. Since these cluster validity indices are comprehensively described in the respective authors' original publications (Tou, 1979; Davies and Bouldin, 1979; Rousseeuw, 1987; Hubert and Arabie, 1985) they are not explicitly repeated. It is worth pointing out, that the τ-index, the S-value and the Γ-statistics indicate good partitions by value maximization, while the R-index indicates good partitions by value minimization.

2.4. Voting Scheme

Based on these $Q = 4$ cluster validity measures the numerical performance of each partition is evaluated in the following scheme adapted from Günter and Bunke (2003):

(1) The number of clusters with the best index performance is given P points.

(2) The number of clusters with second-best performance is given $(P - 1)$ points.

(3) The number of clusters with third-best performance is given $(P - 2)$ points.

(4) The number of clusters with worst performance is given 2 points.

Subsequently the points of all four cluster validity indices are accumulated as a function of the number of clusters and are evaluated for their global maximum in points. The respective number of clusters is defined as numerically optimal and thus selected for the current feature combination.

A visualization of this voting scheme is shown in Fig. 2 for two exemplary feature combinations. For the example in Fig. 2(a), the optimal number of four clusters is detected by each of the cluster validity indices independently. Consequently, the voting points also show a global maximum at four clusters. The second example is shown in Fig. 2(b). Here the optimal number of clusters differs for the cluster indices investigated. In this case the voting scheme yields a combined optimum at four clusters. This number of clusters is found by the R-index, the Γ-statistic and the S-value, while the τ-index would yield three clusters as numerical optimum. This can be caused by outliers, since the τ-index is based on the minimum and maximum distances of data points belonging to the respective clusters.

As discussed by Günter and Bunke (2003) the combined evaluation of multiple indices can improve automated determination of the optimal number of clusters. Since cluster validity indices are often affected differently by outliers and the dataset structure, strengths of different indices are effectively combined by the voting strategy, while weaknesses are reduced.

In this context it is worth pointing out, that the cluster validity indices used were chosen based on their low numerical complexity. There are numerous other cluster separation measures available, but these typically come with increased computational costs. Thus they are less desirable for automated screening of large numbers of feature combinations.

2.5. Detection of globally best feature combination

As a final step, the feature combination with best global performance is determined. By definition, the partition with best global performance is detectable by the extreme values of the associated cluster validity indices. Thus, a second voting strategy is applied, which evaluates the 25 best partitions for each cluster validity index. Subsequently, the individual results of the cluster validity indices are cumulated utilizing a voting scheme:

(1) The partition (feature combination) with best index performance is given 25 points.

(2) The partition with second-best performance is given 24 points.

(3) The partition with third-best performance is given 23 points.

Thus a partition which is favored by all cluster validity indices would get exactly 100 points. Similarly, good partitions would get high numbers of points, while less favored or ambiguous partitions would get fewer points. As shown below, in most cases the partition with best global performance has less than 100 points. Thus the individual cluster validity indices would suggest different partitions. This indicates why the voting strategy is better than simple extreme value search, since it takes into account all four cluster validity indices.

In the following the performance of this pattern recognition method is evaluated using artificial clusters and experimental data.

3. Artificial clusters

Following the approach of Milligan (1985) and the refinement by Qiu and Joe (2006a), artificial datasets with defined degree of separation were generated within the software package R (R Development Core Team, 2010). Subsequently the pattern recognition method described in section 2 was applied to identify the number of natural clusters and to calculate the respective optimal partition. To determine the efficiency of the method, the Rand index was calculated from the known partition of the artificial dataset and the respective partition resulting from the clustering approach (Rand, 1979).

Following the suggestion by Qiu and Joe (2006a) a factorial experiment design was chosen. Three degrees of separation J between clusters were generated utilizing the measure introduced by Qiu and Joe (2006b). The measure is based on the separation of two clusters generated from two univariate normal distributions $N(0, 1)$ and $N(0, A)$. For the values of $A = 8$, $A = 6$ and $A = 4$ the measure of the degree of separation ranges from $J = 0.342$ over $J = 0.213$ to $J = 0.010$ (Qiu and Joe, 2006a,b; 2009). The basic dataset structure consists of ten features including five noisy features. The term noisy refers to a feature with uniform distribution, which is independent of all other noisy and non-noisy features. The number of objects in each cluster was randomly chosen within the range [50, 200] which reflects reasonable variability of the cluster sizes. In addition the clusters were prepared with or without additional outliers. In total 10 outlier data points are
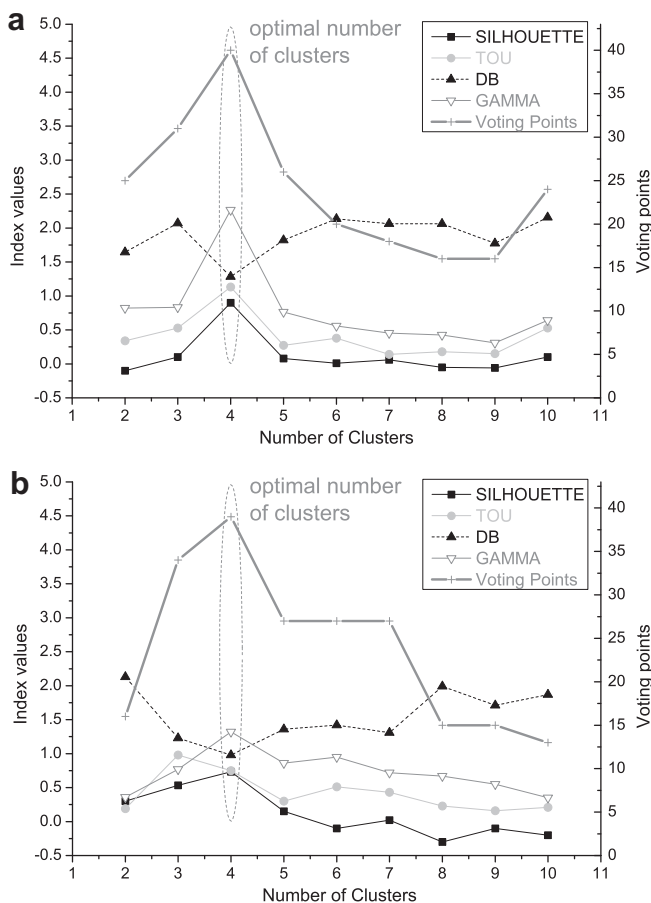

Fig. 2. Evaluation of several cluster index values in voting scheme adapted from Günter and Bunke (2003) for two examples of feature combinations of the same dataset.

generated from a distribution whose marginal distributions are independent uniform distributions. The range of the nth marginal uniform distribution depends on the range $[\mu_n - 4\sigma_n; \mu_n + 4\sigma_m]$ of non-outliers in the nth dimension with mean value μ_n and

Table 1
Factorial experiment design used for investigation.

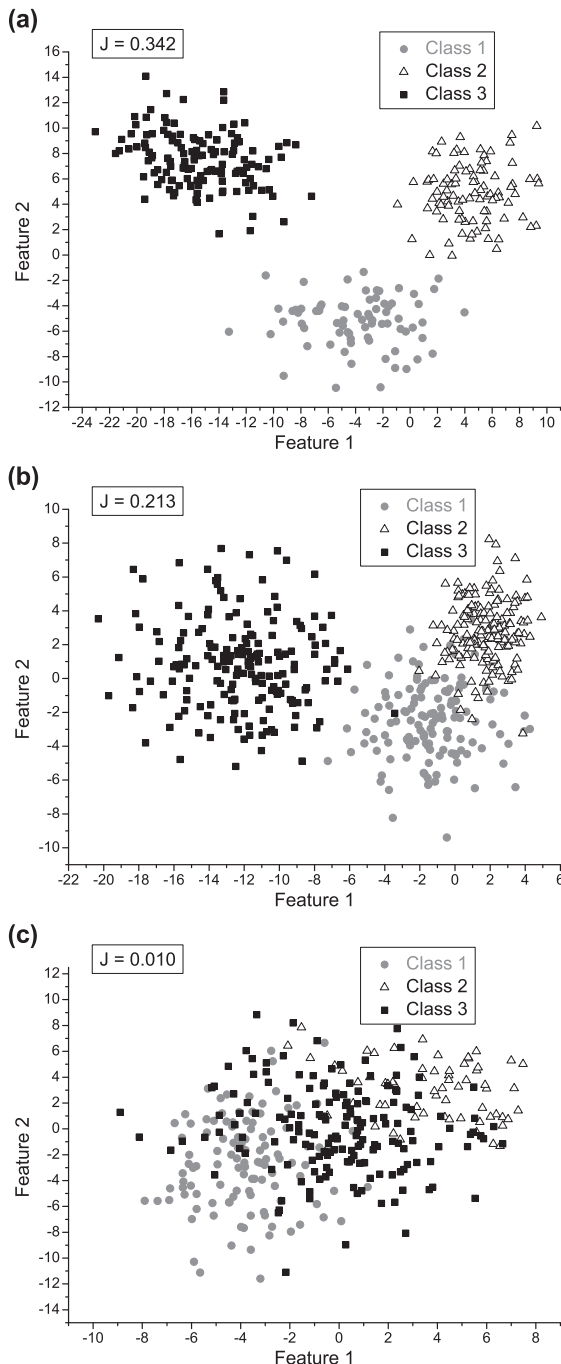| Factors	Levels
Number of clusters	3, 4, 5
Degree of separation	0.342, 0.213, 0.010
Outliers	0, 10
Iterations	3
Total	$3 \times 3 \times 2 \times 3 = 54$ Datasets

Fig. 3. Representative cluster structure in scatter plots of two non-noisy features for degree of separation $J = 0.342$ (a), $J = 0.213$ (b) and $J = 0.010$ (c).

Table 2
Results of investigation for artificial clusters with degrees of separation $J = 0.342$, $J = 0.213$ and $J = 0.010$.

Outliers	Separation	Iteration	Rand-index		
			3 Clusters	4 Clusters	5 Clusters
No	0.01	1	0.7979	0.8509	0.8162
		2	0.7481	0.8494	0.8184
		3	0.8585	0.8742	0.8311
	0.213	1	0.9388	0.9547	0.9831
		2	0.9933	0.9605	0.9675
		3	0.9944	0.9964	0.9812
	0.342	1	1.0000	0.9776	0.9949
		2	1.0000	0.9952	0.9887
		3	1.0000	0.9936	0.9964
Yes	0.01	1	0.7788	0.8502	0.8503
		2	0.7790	0.8338	0.8663
		3	0.7953	0.7673	0.8043
	0.213	1	0.9230	0.9698	0.9694
		2	0.9576	0.9350	0.9377
		3	0.9224	0.9465	0.9455
	0.342	1	0.9563	0.9780	0.9771
		2	0.9678	0.9696	0.9825
		3	0.9611	0.9687	0.9790

standard deviation σ_n of the nth feature (for more details see Qiu and Joe (2006a)). Further, the factorial experiment design includes an evaluation of three, four and five number of clusters and an independent generation of three datasets per configuration as listed in Table 1. This yields 54 artificial datasets in total for the investigation.

An example of the cluster structure of the artificial datasets is shown in Fig. 3 in scatter plots of two representative non-noisy features for each of the three degrees of separation. Clearly, the plots visualize the degree of separation, which is described by Qiu and Joe (2006a) as "well-separated" ($J = 0.342$), "separated" ($J = 0.213$) and "close" ($J = 0.010$).

The pattern recognition approach described in Section 2 was applied to each of the 54 artificial datasets using $K = 10$, $P = 10$ and $M = 4$. The result of this investigation is summarized in Table 2. For all 54 datasets the method is able to detect the correct number of clusters and respective feature combinations automatically. For the partitions found the range of the Rand index is [0.7481, 1]. As a function of decreasing degree of separation, the Rand index values decrease as well. This is attributed to the increasing overlap and the subsequent error of classification using the k-means algorithm. However, independent of the occurrence of outlier data, the method is able to retrieve a suitable partition for each dataset.

4. Clustering of acoustic emission signals

The experimental datasets were obtained from acoustic emission signals recorded during failure of CFRP specimens. The specimens were made of the T800/913 prepreg system with a $[0°/90°/90°/90°/90°]_{sym}$ stacking sequence symmetrical around the medial plane. Here 0° and 90° fiber axis direction refer to the in-plane directions of the specimen, with 0° parallel to the x-direction in Fig. 4. The specimens with dimensions of $100 \times 15 \times 1.4$ mm (length $\times$ width $\times$ height) were loaded in a four-point bending set-up according to DIN-EN-ISO 14125. The acoustic emission signals were acquired using two broadband piezoelectric sensors of type WD in a linear arrangement as shown in Fig. 4. The acquisition settings were 40 dB preamplification with a threshold of 35 dB, 10 MSPs sampling rate and a bandpass ranging from 20 kHz to 1 MHz. Only the acoustic emission signals localized between the two upper supports were used for the pattern recognition process.

The result of parameter based pattern recognition techniques is predominantly influenced by the definition and selection of the
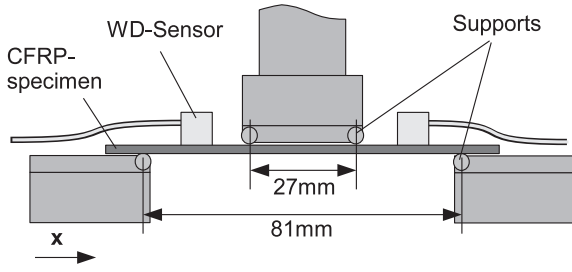
Fig. 4. Experimental setup according to DIN-EN-ISO 14125.

Table 3
Acoustic emission signal features used for pattern recognition.

Feature	Definition
Average frequency (Hz)	$\langle f \rangle = N_{AE}/t_{AE}$
Reverberation frequency (Hz)	$f_{rev} = \frac{N_{AE} - N_{peak}}{t_{AE} - t_{peak}}$
Initiation frequency (Hz)	$f_{init} = \frac{N_{peak}}{t_{peak}}$
Peak frequency (Hz)	f_{peak}
Frequency centroid (Hz)	$f_{centroid} = \frac{\int f \cdot \tilde{U}(f) df}{\int \tilde{U}(f) df}$
Weighted peak-frequency (Hz)	$\langle f_{peak} \rangle = \sqrt{f_{peak} \cdot f_{centroid}}$
Partial power 1–6 (%)	$\int_{f_1}^{f_2} \hat{U}^2(f) df / \int_{0kHz}^{1200kHz} \hat{U}^2(f) df$
	Partial power 1: f_1 = 0 kHz; f_2 = 150 kHz
	Partial power 2: f_1 = 150 kHz; f_2 = 300 kHz
	Partial power 3: f_1 = 300 kHz; f_2 = 450 kHz
	Partial power 4: f_1 = 450 kHz; f_2 = 600 kHz
	Partial power 5: f_1 = 600 kHz; f_2 = 900 kHz
	Partial power 6: f_1 = 900 kHz; f_2 = 1200 kHz

features describing an object of a dataset. In our case, the object is an acoustic emission signal which is a part of a dataset containing all signals for classification. The signal features are certain characteristics of the acoustic emission signals derived from the time or frequency-domain as illustrated in Fig. 5.

As introduced in Section 1, the aim of the application of pattern recognition techniques in the current context is the identification of signals originating from particular failure mechanisms. Based upon the results from previous finite element simulations we solely used frequency features, since there was no indication of a correlation between acoustic emission energy and failure mechanisms in CFRP (Sause and Horn, 2010a). This is not necessarily the case for failure of other composite material types, like failure of multi-layer coatings on CFRP as reported before (Sause et al., 2008, 2009). The approach to distinguish between fiber breakage and matrix cracking based on significant contributions at high frequencies (fiber breakage) or low frequencies (matrix cracking) of the acoustic emission signals was already proposed by various authors (Bohse, 2000; Haselbach and Lauke, 2003; Huguet et al., 2002; Ramirez-Jimenez et al., 2004; Marec et al., 2008; Li et al., 2008). We recently presented a finite element simulation introducing a micromechanical acoustic emission source model able to simulate different types of failure within fiber reinforced composites (Sause and Horn, 2010a,b). It was demonstrated that matrix-cracking, fiber-breakage and interface-failure result in distinguishable frequency spectra in the presented four-point bending setup.

Table 3 holds the investigated frequency features, which are partially derived from the time domain as well. Here the signal duration t_{AE} and the signal rise time t_{peak} as shown in Fig. 5(a) are used to determine the rising and falling parts of the signal. In

addition the overall number of threshold crossings N_{AE} and the number of threshold crossings N_{peak} up to U_{max} are used (see Fig. 5(a)). Other features are directly obtained from the Fourier-Transformation $\tilde{U}(f)$ of the signal $U(t)$ (see Fig. 5(b)). Here f_{peak} defines the frequency position of maximum FFT-Magnitude. It is worth noting, that the features in Table 3 are not meant to provide a comprehensive list for characterization of acoustic emission signals.

For the investigated experimental datasets, the number of objects in the individual datasets is between 525 and 1611. The pattern recognition approach as introduced in Section 2 with $K = 12$, $P = 10$ and $M = 5$ was applied to six experimental datasets. The total number N of feature combinations investigated is thus 3302 for each dataset.

The results of the pattern recognition approach are summarized in Table 4. The table lists the identified number of clusters, the respective identifier (ID) of the feature combination (listed in detail in Table 5) and the associated voting points. In addition, the values of the cluster validation statistics used are given for the respective partition. The last column holds a subjective rating if the found partition and number of clusters is able to detect the expected signal classes within the experimental datasets. In order to find the best visualization of the cluster structure, all pair-wise scatterplots of the identified feature combination were evaluated. The plot of
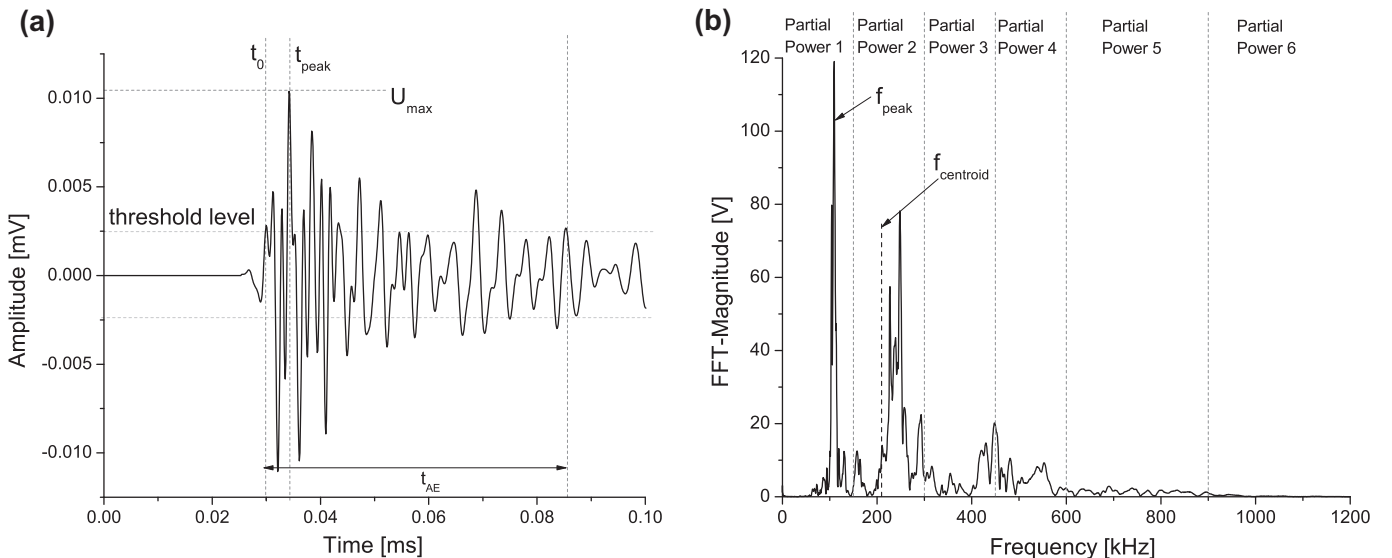

Fig. 5. Extraction of features from acoustic emission signals in time (a) and frequency domain (b).
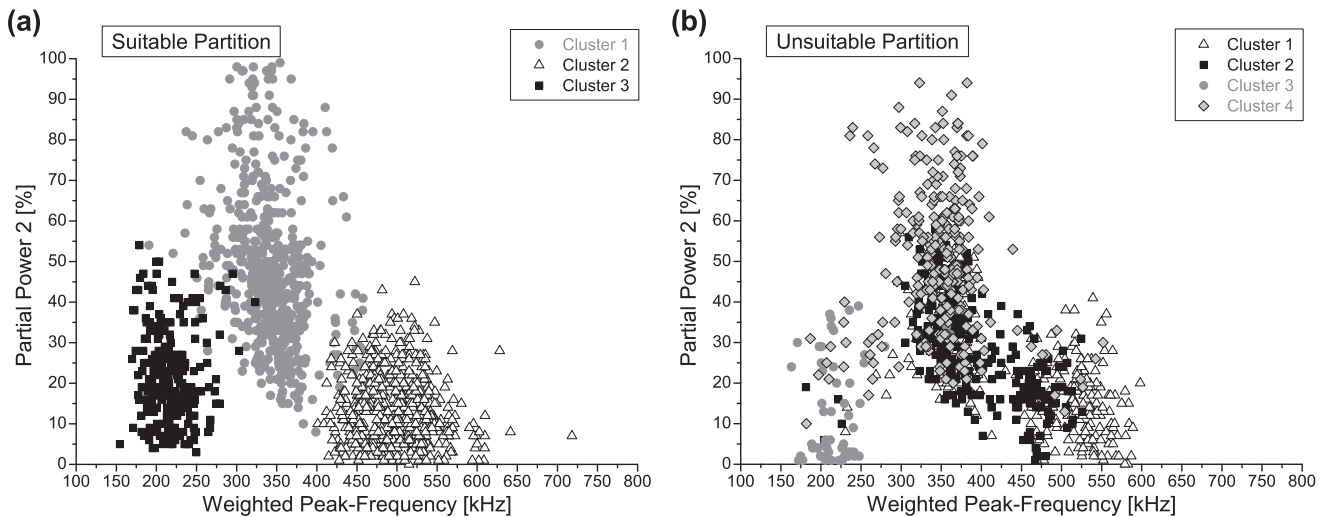
Table 4
Summary of pattern recognition results of acoustic emission signals.

Specimen	Clusters	Points	ID	R	τ	S	Γ	Rating
A	3	90	85	0.8516	1.9792	0.4636	0.6768	Suitable
B	3	90	110	0.8903	2.0652	0.4662	0.6770	Suitable
C	4	53	324	0.6803	1.1241	0.5226	0.6629	Unsuitable
D	3	50	195	0.9813	1.8582	0.4736	0.3657	Unsuitable
E	4	71	119	0.8938	0.8753	0.5027	0.6602	Suitable (noise)
F	3	83	285	0.8466	1.9473	0.4602	0.7007	Suitable

Table 5
List of feature combinations for distinguishing acoustic emission signals.

Feature combination 85	Partial power 1 (%)	Partial power 2 (%)	Reverberation frequency (Hz)	Peak frequency (Hz)	Weighted peak frequency (Hz)
Feature combination 110	Partial power 1 (%)	Partial power 2 (%)	Partial power 4 (%)	Peak frequency (Hz)	Weighted peak frequency (Hz)
Feature combination 119	Partial power 1 (%)	Partial power 2 (%)	Partial power 6 (%)	Peak frequency (Hz)	Weighted peak frequency (Hz)
Feature combination 195	Partial power 1 (%)	Partial power 3 (%)	Partial power 5 (%)	Partial power 6 (%)	Frequency centroid (Hz)
Feature combination 285	Partial power 1 (%)	Partial power 4 (%)	Reverberation frequency (Hz)	Peak frequency (Hz)	Weighted peak frequency (Hz)
Feature combination 324	Partial power 1 (%)	Partial power 4 (%)	Partial power 6 (%)	Peak frequency (Hz)	Weighted peak frequency (Hz)

Fig. 6. Suitable partition of specimen B according to feature combination 110 (a) and unsuitable partition of specimen C according to feature combination 324 (b).

Partial power 2 over weighted peak frequency shown in Fig. 6 was found to be most informative.

For four of the six investigated datasets, the algorithm was able to find suitable partitions with three clusters in terms of this physical correlation. The differences in frequency spectra cause significantly different frequency features and result in the formation of the three distinct clusters as shown in Fig. 6 for specimen B. For specimen E the algorithm finds four clusters. This unexpected additional cluster arises from 13 very similar signals, which could reasonably be identified as noise signals. The partition of specimen C rated as unsuitable is shown in Fig. 6(b). Clearly the assignment is different from that of Fig. 6(a). In particular, the voting points and the respective cluster validity indices values are worse than those of the remaining (suitable) partitions.

Since the experimental setup was identical for all specimens investigated it should be possible to identify a single characteristic feature combination which can be used for pattern recognition of all datasets. Since no unique feature combination was found in the investigation so far, an alternative approach was employed.

In order to find a feature combination suitable for clustering all datasets an alternative strategy was applied. Thus all experimental data sets were combined together resulting in a total number of 5848 objects. Subsequently, the algorithm introduced in Section 2

with $K = 12$, $P = 10$ and $M = 5$ was applied to random subsets of 1000 objects drawn from this new dataset. The result of this investigation is summarized in Table 6.

All subset partitions yield a suitable partition with cluster positions as shown in Fig. 6(a). This is expressed in the comparable values of the respective cluster validity indices and the high values above 60 voting points. The feature combinations 85 and 110 are repeatedly identified as the best feature selection. In particular,

Table 6
Summary of pattern recognition results of acoustic emission signals (random subsets of all signals).

Iteration	Clusters	Points	ID	R	τ	S	Γ
1	3	78	110	0.9122	1.9259	0.4508	0.6503
2	3	60	110	0.9221	1.9391	0.4419	0.6427
3	3	87	110	0.9013	1.9000	0.4435	0.6460
4	3	69	110	0.8881	1.9034	0.4505	0.6515
5	3	71	110	0.9400	1.9193	0.4341	0.6399
6	3	60	85	0.8717	1.8630	0.4576	0.6649
7	3	67	85	0.8773	1.7272	0.4549	0.6607
8	3	62	85	0.8846	1.7555	0.4526	0.6600
9	3	75	110	0.9213	1.9335	0.4325	0.6364
10	3	86	110	0.9109	1.8669	0.4449	0.6408

all subsets favoring feature combination 85 rank feature combination 110 as second-best combination. This suggests choosing feature combination 110 as given in Table 5 as the best selection for pattern recognition of the given datasets with the current preprocessing and cluster algorithm.

5. Conclusion

It was demonstrated that the proposed pattern recognition method is able to identify the correct number of clusters and respective partitions in artificial datasets generated according to the method of Qiu and Joe (2006a). The method is robust against a limited number of outliers and was able to retrieve suitable partitions for degrees of separation of $J = 0.342$, $J = 0.213$ and $J = 0.010$.

In application to acoustic emission signals the method is able to detect suitable partitions, which reflect the physically expected differences in the signal types. In particular, the identified cluster positions reflect the difference in frequency features extracted from the frequency spectra of the signals. This in turn is directly correlated to their physical origin of matrix cracking, interface failure and fiber breakage as suggested by various authors (Bohse, 2000; Haselbach and Lauke, 2003; Huguet et al., 2002; Ramirez-Jimenez et al., 2004; Marec et al., 2008; Li et al., 2008) and demonstrated by finite element simulations (Sause and Horn, 2010a,b).

While the optimal feature combinations identified to find this partition were strictly individual for each single dataset, an alternative strategy applied to random subsets of all datasets yields clearly favored feature combinations and stable partitions. In combination with the results of the single datasets it was observed that high values of voting points are indicative of good partitions. Based on the current investigations, partitions below 60 voting points are not expected to describe the natural clusters hidden within the dataset.

In comparison to other feature selection techniques like dendrograms, breadth-first or hill-climbing algorithms the greatest drawback is the computational efficiency of the proposed method. By selection of statistically representative subsets of the experimental datasets it was demonstrated, that the method is able to retrieve similar results. This drastically improves the computational efficiency of the method. As an alternative to the currently used k-means algorithm, hierarchical cluster algorithms could be used. These are computationally advantageous for calculation of subsequently increasing number of clusters P. Similar increase in performance can be achieved by reasonable limitation of the maximum number of clusters P investigated for each feature combination. For the current investigation this value was chosen as $P = 10$, while higher efficiency could be achieved by dynamic reduction of P based on saturation criteria of the cluster validity indices used. Finally, the minimum number of features M should be chosen adequately high to allow reasonable description of the dataset without losing the discriminative power of single features.

Based on the values of the R-index, the Γ-statistics, the S-value and the τ-index the experimental datasets' degree of separation is found to be slightly worse than those of the artificial datasets with $J = 0.213$. Using this information we may approximately estimate the classification error of the experimental datasets based on a comparison with the respective Rand-Index values of the artificial datasets. This gives a mean error value of 4% for our proposed pattern recognition method.

Moreover, the method can also be used to identify the importance of particular features better than utilizing correlation dendrograms. Since all feature combinations are taken into account, the importance of a single feature can be evaluated by its frequency of occurrence in high ranking partitions. This enables the assessment of the discriminative power of newly defined features of acoustic emission signals.

Acknowledgment

References

Anastassopoulos, A.A., Philippidis, T.P., 1995. Clustering methodology for the evaluation of acoustic emission from composites. J. Acoust. Emission 13, 11–21.

Bohse, J., 2000. Acoustic emission characteristics of micro-failure processes in polymer blends and composites. Compos. Sci. Technol. 60, 1213–1226.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. 1 (2), 224–227.

Günter, S., Bunke, H., 2003. Validation indices for graph clustering. Pattern Recognition Lett. 24, 1107–1113.

Haselbach, W., Lauke, B., 2003. Acoustic emission of debonding between fibre and matrix to evaluate local adhesion. Compos. Sci. Technol. 63 (15), 2155–2162.

Hubert, L.J., Arabie, P., 1985. Comparing partitions. J. Classif. 2, 193–218.

Huguet, S., Godin, N., Gaertner, R., Salmon, L., Villard, D., 2002. Use of acoustic emission to identify damage modes in glass fibre reinforced polyester. Compos. Sci. Technol. 62, 1433–1444.

Li, X., Ramirez, C., Hines, E.L., Leeson, M.S., Purnell, P., Pharaoh, M., 2008. Pattern recognition of fiber-reinforced plastic failure mechanism using computational intelligence techniques. In: IEEE World Congress on Computational Intelligence. Neural Networks. pp. 2340–2345.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Le Cam, L.M., Neyman, J. (Eds.), Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, CA, pp. 281–297.

Marec, A., Thomas, J.-H., Guerjouma, R., 2008. Damage characterization of polymer-based composite materials: multivariable analysis and wavelet transform for clustering acoustic emission data. Mech. Syst. Signal Process. 22, 1441–1464.

Milligan, G.W., 1985. An algorithm for generating artificial test clusters. Psychometrika 50 (1), 123–127.

Philippidis, T., Nikolaidis, V., Anastassopoulos, A., 1998. Damage characterisation of C/C laminates using neural network techniques on AE signals. NDT&E International 31, 329–340.

Polikar, R., 2006. Pattern recognition. In: Wiley Encyclopedia of Biomedical Engineering. John Wiley & Sons Inc., New York.

Qiu, W.-L., Joe, H., 2006a. Generation of random clusters with specified degree of separation. J. Classif. 23 (2), 315–334.

Qiu, W.-L., Joe, H., 2006b. Separation index and partial membership for clustering. Comput. Stat. Data Anal. 50, 585–603.

Qiu, W.-L., Joe, H., 2009. clusterGeneration: random cluster generation (with specified degree of separation). <http://www.R-project.org>.

R Development Core Team, 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0. <http://www.R-project.org>.

Ramirez-Jimenez, C.R., Papadakis, N., Reynolds, N., Gan, T., Purnell, P., Pharaoh, M., 2004. Identification of failure modes in glass/polypropylene composites by means of the primary frequency content of the acoustic emission event. Compos. Sci. Technol. 64, 1819–1827.

Rand, W.M., 1979. Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc., 846–850.

Richardson, J.M., Elsley, R.K., Graham, L.J., 1984. Nonadaptive, semi-adaptive and adaptive approaches to signal processing problems in nondestructive evaluation. Pattern Recognition Lett. 2, 387–394.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Sause, M.G.R., Horn, S., 2010a. Simulation of acoustic emission in planar carbon fiber reinforced plastic specimens. J. Nondestruct. Eval. 29 (2), 123–142.

Sause, M.G.R., Horn, S., 2010b. Influence of specimen geometry on acoustic emission signals in fiber reinforced composites: FEM-simulations and experiments. In: Conf. Proc. 29th European Conf. on Acoustic Emission Testing, Vienna, Austria.

Sause, M.G.R., Schultheiß, D., Horn, S., 2008. Acoustic emission investigation of coating fracture and delamination in hybrid carbon fiber reinforced plastic structures. J. Acoust. Emission 26, 1–13.

Sause, M.G.R., Haider, F., Horn, S., 2009. Quantification of metallic coating failure on carbon fiber reinforced plastics using acoustic emission. Surf. Coat. Technol. 204 (3), 300–308.

Schürmann, H., 2007. Konstruieren mit Faser-Kunststoff-Verbunden, second ed. Springer, Berlin.

Tou, J.T., 1979. DYNOC – a dynamic optimal cluster-seeking technique. Int. J. Comput. Inform. Sci. 8 (6), 541–547.

Vi-Tong, E., Gaillard, P., 1986. An algorithm for non-supervised sequential classification of signals. Pattern Recognition Lett. 5, 307–313.