Integration of Machine Learning on Distributed Acoustic Sensing surveys

Camille Jestin, Clément Hibert, Camille Huynh, Gaëtan Calbris & Vincent Lanticq

EGU General Assembly 2020
Introduction
Introduction

- Scattering effect occurs everywhere along the fiber
- The backscattering light contains the information of strain from where it was generated

**DAS principle**

- Laser pulse propagating through the fiber
- Small part of the pulse back to the equipment due to scattering effect
- Acoustic and Vibration signals
Introduction

DAS principle: Strain acquisition

- Optical phase directly related to strain applied to the fiber core over a gauge length
- Time derivative of the strain -> strain rate
- Strain-rate unit: nm/m/s

\[ \Delta \phi = \frac{4\pi nG \xi}{\lambda \varepsilon} \]
Introduction

DAS principle: Acquisition parameters

- Parameters to adjust:
  - Fiber distance
  - Optical power
  - Pulse width
  - Pulse rate frequency
  - Spatial sampling resolution
  - Gauge length
  - Derivation time
Machine Learning applied to DAS surveys
Context of the study

**Pipeline monitoring for intrusion detection:**
- Third party works detection and location using DAS is commonly applied in different contexts
- Challenge in identifying the origin of the signal:
  - Necessity of pattern recognition for relevant alarm.
  - Source and amplitude analysis for determining the threat at the pipeline neighbourhood.
  - The source identification must be fast, accurate and robust.
  - For its application to DAS data, the used method must be able to handle a big amount of data.

**Solution:** A Machine Learning algorithm enabling Classification of patterns before the alert release
Detection of signal of potential interest

Machine Learning with Random Forest algorithm

Signals Classification

Decision of alert release

IF
Identified class = Risk
AND
Corresponding energy band > threshold:
ALERT

Spectral content between 5Hz and 95Hz
Machine Learning applied to DAS surveys

The training and testing chains

Signal event Database (Sub-Samples)

Signal event Database

Machine Learning with Random Forest algorithm

Signals Classification

Comparison True/Predicted

Testing chain

Training chain

Database

Class 1

Class 2

...
Test of the use of the **supervised classifier named Random Forest algorithm**, an ensemble learning method based on the use of decision trees.

**Discriminating attributes:**
- Duration
- skewness
- $f_{\text{max}}$
- Kurtosis
- Spectral properties
- 50+ features

**Training set:**
Class determination

**Random Forest**

Tree 1

Tree 2

Tree n

Decision = **1** Classification
### Waveform Attributes: 23

1. Duration
2. Max/Mean ratio
3. Max/Median ratio
4. Ascending/Descending time ratio
5. Kurtosis of raw signal
6. Kurtosis of signal envelop
7. Skewness of raw signal
8. Skewness of signal envelop
9. Number of peaks in autocorrelation function
10. Energy in 1<sup>st</sup> third part of autocorrelation function
11. Energy in remaining part of autocorrelation function
12. Ratio of 11 and 10
13-17. Energy of the signal filtered in 5-10Hz, 10-30Hz, 30-50Hz, 50-75Hz and 75-99Hz
18-22. Kurtosis of the signal filtered in 5-10Hz, 10-30Hz, 30-50Hz, 50-75Hz and 75-99Hz
23. RMS between decreasing part of the signal and \( I(t) = Y_{\text{max}} - \frac{Y_{\text{max}}}{t_f - t_{\text{max}}} t \)

### Spectral Attributes: 17

24. Mean of the Discrete Fourier Transform (DFT)
25. Max of the DFT
26. Frequency at the maximum DFT
27. Frequency at the centroid
28. Central frequency of the 1<sup>st</sup> quartile
29. Central frequency of the 3<sup>rd</sup> quartile
30. Median of the normalized DFT
31. Variance of the normalized DFT
32. Number of peaks in normalized DFT
33. Number of peaks (>0.75 DFT<sub>max</sub>)
38. Spectral centroid
39. Gyration radius
40. Spectral Centroid width

### Pseudo-Spectrogram Attributes: 17

41. Kurtosis of max of all DFTs as a function of time
42. Kurtosis of median of all DFTs as function of time
43. Mean ratio between max and mean of all DFTs
44. Mean ratio between max and median of all DFTs
45. Number of peaks in the curve of temporal evolution of DFTs max frequency
46. Number of peaks in the curve of temporal evolution of DFTs mean frequency
47. Number of peaks in the curve of temporal evolution of DFTs median frequency
48. Ratio between 45 and 46
49. Ratio between 45 and 47
50. Mean distance between max and mean of all DFTs as function of time
51. Mean distance between max and median of all DFTs as function of time
52. Number of peaks in the curve of centroid frequency spectrum DFT
53. Number of peaks in the curve of max frequency spectrum DFT
54. Ratio between max frequency and centroid frequency DFTs
55. Mean distance between 1<sup>st</sup> quartile and median of all DFTs as function of time
56. Mean distance between 3<sup>rd</sup> quartile and median of all DFTs as function of time
57. Mean distance between 3<sup>rd</sup> and 1<sup>st</sup> quartiles of all DFTs as function of time

Hibert et. al, 2014, Provost et al., 2017; Hibert et al., 2017
Exemple of pipeline monitoring
Pipeline monitoring: Third party works classification

Manual Compactor

Energy band [5 - 95]Hz

Spectrogram

Energy band [5 - 95]Hz

Spectrogram
Pipeline monitoring: Third party works classification

Excavation

Energy band [5 - 95]Hz

Spectrogram
Pipeline monitoring: Third party works classification

Drilling

Energy band [5 - 95]Hz

Spectrogram
Pipeline monitoring: Third party works classification

Jack hammer

Energy band [5 - 95]Hz

Spectrogram
Pipeline monitoring: Third party works classification

Sheet pile

Energy band [5 - 95]Hz

Spectrogram
Pipeline monitoring: Third party works classification

Circular saw

Energy band [5 - 95]Hz

Spectrogram
Transportations

Energy band [5 - 95] Hz

Spectrogram
Then:

• In this study, we work on 7 classes of event, numbered from 1 to 7.

• Because DAS acquisition can generate traces every few meters along fibres of tens of kilometres, two methods are used for classification using Random Forest algorithm:

  1. The first one is signal based: The algorithm is using each single trace/station for the signal classification.
  2. The second one is event based: A cluster of stations, identified as recording the same event, is used by the algorithm for the source signal classification. The majority of votes will release the final ID of the event.

• Three parameters are used to check the efficiency of the pattern ID using Machine Learning: Precision, Recall and Accuracy
Quality Control parameters

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Negative</td>
</tr>
<tr>
<td>Negative</td>
<td>True negative</td>
<td>False positive</td>
</tr>
<tr>
<td>Positive</td>
<td>False negative</td>
<td>True positive</td>
</tr>
</tbody>
</table>

Accuracy: \[
\frac{\text{True positive} + \text{True negative}}{\text{Positive} + \text{Negative}}
\]

Precision: \[
\frac{\text{True positive}}{\text{True positive} + \text{False positive}}
\]

Recall: \[
\frac{\text{True positive}}{\text{True positive} + \text{False negative}}
\]

F1 Score: \[
2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]
Pipeline monitoring: Results

**First approach:** use of the same number of samples for each class

The first classifier was trained using 50 samples of each class

1 - False alarm rate for each class

1 – Lack of detection rate for each class

Global good classification rate
First approach: use of the same number of samples for each class

Classes:
1. Manual compactor (75)
2. Excavation (551)
3. Drilling (125)
4. Jack hammer (1726)
5. Palplanche (105)
6. Circular saw (321)
7. Transportation (538)

For all studied events: Classification with this algorithm is **88.46% correct** with an accuracy of 91.28%
**Second approach:** Training samples are taken proportional to their natural distribution occurrences

Our classifier was trained using a half of the total dataset

1 - False alarm rate for each class

1 – Lack of detection rate for each class

**Global good classification rate**
**Second approach:** Training samples are taken proportional to their natural distribution occurrences.

For all studied events:
- Classification with our algorithm is **100% correct** with an accuracy of **98.69%**.

**Classes:**
1. Manual compactor (75)
2. Excavation (551)
3. Drilling (125)
4. Jack hammer (1726)
5. Palplanche (105)
6. Circular saw (321)
7. Transportation (538)
Conclusions

- Random Forest algorithm appears to be relevant (fast and robust) for the classification of acoustic events recorded with DAS.

- Tests on other field sites are under process to demonstrate the efficiency of our Machine Learning method on different contexts.

- Different fields of application of this algorithm are possible: intrusion detection along pipelines, in perimeters, seismic event detection and classification (volcanoes, glaciers, etc.).

- Tests on data processing in flux for real-time event detection and classification are under process.