

# CS229 - Project Final Report:

## Automatic earthquake detection from distributed acoustic sensing (DAS) array data

Ettore Biondi, Fantine Huot, Joseph Jennings

### Abstract

We attempt to automatically detect earthquake events in distributed acoustic sensing (DAS) data via a supervised learning approach. Detecting earthquakes with different magnitudes could potentially provide the ability of predicting major catastrophic events.

## Introduction

Distributed acoustic sensing (DAS) is an emerging technology used to record seismic data that employs fiber optic cables as a probing system. By measuring the backscattered energy of a pulsing laser transmitted down a fiber optic cable, it is possible to measure the strain rate occurring within different sections of the cable [1]. DAS recording systems have been shown to measure data comparable with conventional geophones [2] and have been successfully used in exploration and earthquake seismology settings [3, 4]. Recently, a DAS array has been deployed beneath Stanford campus using existing telecommunication fiber optic cables. The data recorded from this array have the potential for near-surface imaging of the subsurface and early-warning earthquake monitoring.

In order to accurately detect earthquakes for early warning we must extract earthquake signals from the surrounding urban and ambient noise that is constantly recorded by our DAS array. Therefore, in addition to classifying earthquake signals, we also attempted to classify urban and ambient noise present in the array dataset with the end goal of classifying each sample or windows of samples as one of these three types of signal. To perform this classification, we used a supervised learning approach, in which we train a classifier on labeled training data that consists of the processed amplitudes from our DAS array and their corresponding type of signal. Therefore, in our application we aimed to separate the signal into three classes, namely, ambient noise, urban noise, and earthquake signal. We first attempted this classification process using a single sample from our amplitude data (i.e., our feature is a scalar) with Gaussian Naive Bayes, softmax and kernelized support vector machine (SVM) supervised learning models. We found that using this feature space, our estimated models classified nearly all samples as ambient noise (the most dominant class) and not a single earthquake was detected. To improve the performance of our classifier, we then changed the feature space to overlapping windows of amplitude data. We found that using this feature space, we improved the precision and accuracy of the machine learning algorithms.

## Related work

While our investigation does appear to be the first attempt towards using statistical learning techniques for detecting earthquakes recorded on DAS data, work has been done on both statistical learning for earthquake detection and prediction on geophone data [5] and also for surveillance purposes on DAS data [6, 7]. In [8], the authors show that it is possible to estimate the time before a laboratory earthquake occurs. In this work they employ a random forest approach on statistical features constructed using data measured by an accelerometer. Detecting low magnitude earthquakes is fundamental to estimate the time before a major event is about to happen. This observation pushed us to find them in our recorded DAS data.

## Dataset and preprocessing

The fiber optic cable is deployed in Stanford’s telecommunication tunnels in a double loop pattern. Every 8 meters of cable acts as a receiver and records vibration at a sampling rate of 50Hz, creating a data matrix of 300 channels distributed in space, each continuously recording cable strain since September 2016. The array generates contiguous time series that conveniently lend themselves to image processing. However, from a seismological perspective, the data are very poorly coupled with the ground as the fiber is put through concrete blocks and freely hangs between them. Therefore, we preprocessed the data by balancing the amplitudes, bandpass filtering and detrending them. We are focusing our attention on two months of data; specifically, from September 2016 to October 2016. The size of this dataset in terms of memory is approximately 280Gb, which corresponds to 75 billion samples. Because of the class-project time constraint, we decided to work with only 46 hours of data in which different earthquake events with magnitude from 0.3 to 3.5 were recorded by the DAS array. To further decrease the memory requirement, we bandpass the data between 0.23 to 10Hz and halve the sampling rate in both channel and time directions. These steps allowed us to work with an in-core approach rather than an out-of-core one.

## Labeling the data

We labeled the various categories by using complementary data from other sources. For the cars, we proceeded with a methodology similar to [9], where a clustering algorithm (K-means) was applied to the data in the continuous wavelet domain (CWT) and was able to separate different types of seismic signals. By projecting the data over the array’s geometry, a human supervisor can easily hand pick the clusters corresponding to traffic noise. For labeling the earthquake signals, we used the methodology described in [10], in which earthquakes are identified in our dataset by mapping them with the event catalog from nearby USGS recording stations. This procedure allowed us to pick approximately 800 earthquake events. After the labeling procedure we can perform a quality check on our labeled data by overlaying the recorded data with their corresponding labels (Figure 1). For example, Figure 2 displays an example in which an earthquake event was recorded in addition to the typical signal recorded by the array. We decided to classify all the signal within the event duration as an earthquake since it can be interpreted as an anomaly with respect to the urban and ambient noise. Besides using just a single amplitude sample, we also tried to train our machine learning models on overlapping windows of data. Figure 3 shows examples of a window of ambient noise, urban noise and earthquake amplitudes.

## Methods

Using our created feature space of single amplitude samples, we first trained a Naive Bayes classifier in which we assumed an independent Gaussian distribution for each feature. For our implementation we used the Python `sklearn` package [11]. The idea behind this approach is to maximize the joint a probability distribution defined as

$$J(\sigma_j, \mu_j, \phi_j) = \prod_{i=1}^m P(x^{(i)}, y^{(i)}; \sigma_j, \mu_j, \phi_j), \quad (1)$$

where  $x^{(i)}$  represents a feature for the  $i$ -th sample and  $y^{(i)}$  is the corresponding label. Maximizing equation 1 results in the following expressions for the parameters  $\sigma_j$ ,  $\mu_j$  and  $\phi_j$  [12]

$$\phi_j = \frac{\sum_{i=1}^m \mathbb{1}\{y^{(i)} = j\}}{m}, \quad (2)$$

$$\mu_j = \frac{\sum_{i=1}^m \mathbb{1}\{y^{(i)} = j\}x^{(i)}}{\sum_{i=1}^m \mathbb{1}\{y^{(i)} = j\}}, \quad (3)$$

$$\sigma_j^2 = \frac{\sum_{i=1}^m \mathbb{1}\{y^{(i)} = j\}(x^{(i)} - \mu_j)^2}{\sum_{i=1}^m \mathbb{1}\{y^{(i)} = j\}}, \quad (4)$$

where  $j \in \{0, 1, 2\}$ . In addition to Naive Bayes, we applied a multinomial logistic regression in which we maximize the following log-likelihood function

$$\ell(\theta) = \sum_{i=1}^m \sum_{l=0}^2 \mathbb{1}\{y^{(i)} = l\} \log \left( \frac{e^{\theta_l^T \mathbf{x}}}{\sum_{j=0}^2 e^{\theta_j^T \mathbf{x}}} \right), \quad (5)$$

where  $\theta_l$  are the parameters we attempt to estimate for the  $l$ -th class. Lastly, we used a kernelized SVM approach in which we maximize the following loss function

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y^{(i)} y^{(j)} K(x^{(i)}, x^{(j)}) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C, i = 1, \dots, m, \\ & \sum_{i=1}^m \alpha_i y^{(i)} = 0 \end{aligned}$$

where we set  $C = 1$  and we used two the following two kernels

$$K_{rbf}(x^{(i)}, x^{(j)}) = e^{\gamma \|x^{(i)} - x^{(j)}\|^2} \quad (6)$$

and

$$K_{sig}(x^{(i)}, x^{(j)}) = \tanh(\gamma x^{(i)T} x^{(j)} + r). \quad (7)$$

## Results and Discussion

In order to understand how much each class is separable in the single amplitude feature space, in Figure 4a we show the amplitude as a function of class number. We clearly notice that the three classes entirely overlap, thus they are not linearly separable. In addition, Figure 4b displays the normalized histogram of the amplitudes for each class. On the histograms we superimpose their respective Gaussian fitted distribution. It is interesting to see that amplitudes of the ambient noise is the closest to being Gaussian. Table 1 summarizes our results obtained from testing our trained machine learning models on our test dataset which consisted of two hours of recordings. When the single amplitude feature was used (first four rows in the table), most of our machine learning models classified all of the samples as ambient noise, which was the most frequent class present in the training set. Only the kernelized SVM approaches were able to detect some of the urban noise samples. None of the models were able to capture any earthquake sample. The most likely cause of failure of these methods is the high overlap between classes in this feature space. Additionally, the fact that we classified the earthquake in a large window (Figure 2), introduced many mislabeled samples that hampered the classification algorithm. Conversely, the last four rows recap the results obtained when windows of data were employed in the training as feature space. Notice that the number of samples in each class, in this case, is more balanced than the single amplitude feature space. Among the algorithms run on the windows of data, Naive Bayes performed the best. We are not clearly sure of the reason for this behavior. Further investigation of the amplitude distribution within each window could possibly clarify this observation.

| Model \ Class   | Ambient noise |          | Urban noise |          | Earthquakes |          | Total     |          |
|-----------------|---------------|----------|-------------|----------|-------------|----------|-----------|----------|
|                 | Precision     | Accuracy | Precision   | Accuracy | Precision   | Accuracy | Precision | Accuracy |
| Naive Bayes     | 0.88          | 1.00     | 0.00        | 0.00     | 0.00        | 0.00     | 0.77      | 0.88     |
| Softmax         | 0.88          | 1.00     | 0.00        | 0.00     | 0.00        | 0.00     | 0.77      | 0.88     |
| SVM RBF         | 0.88          | 0.96     | 0.19        | 0.09     | 0.00        | 0.00     | 0.80      | 0.85     |
| SVM Sigmoid     | 0.88          | 1.00     | 1.00        | ~0.00    | 0.00        | 0.00     | 0.88      | 0.88     |
| NB Windows      | 0.85          | 0.85     | 0.92        | 0.35     | 0.03        | 0.51     | 0.87      | 0.59     |
| Smax Windows    | 0.53          | 0.68     | 0.56        | 0.37     | 0.02        | 0.06     | 0.54      | 0.51     |
| SVM RBF Windows | 0.36          | 0.60     | 0.22        | 0.00     | 0.01        | 0.11     | 0.28      | 0.29     |
| SVM Sig Windows | 0.22          | 0.03     | 0.48        | 0.85     | 0.03        | 0.06     | 0.35      | 0.45     |

Table 1: Summary of testing precision and accuracy for individual classes and for all classes combined. The first four rows contain the results for the single amplitude feature, while the last four contain the results for the windows of data.

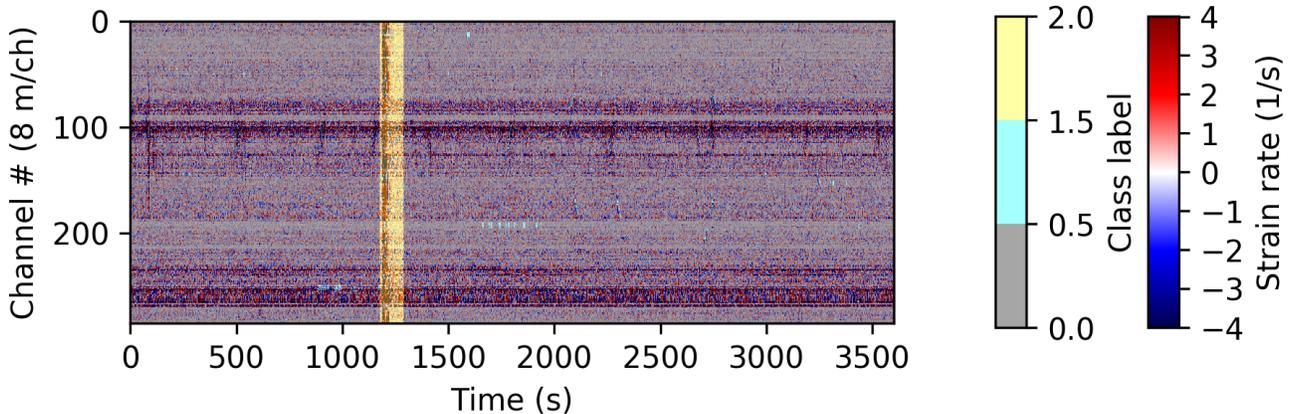


Figure 1: One hour of labeled, recorded data. The gray area corresponds to ambient noise, cyan to urban noise, and the yellow a recorded earthquake. Note that the earthquake signal is present over all channels indicating that all channels were excited simultaneously.

## Conclusion and Future work

We have reported preliminary results on automatic detection of earthquakes from DAS data recorded on Stanford Campus. To this end we have employed different supervised learning techniques, namely, Gaussian Naive Bayes, softmax regression, and kernelized SVM. We tested these algorithms on two different feature spaces. One composed of single amplitude samples, while the other one formed by overlapping windows of data. We found that Naive Bayes on the second feature space performed the best in classifying different noise signatures. In the future, our primary work will be focused on creating accurate synthetics. While we will both train and test on our synthetic examples, the main goal of creating the synthetics will be for us to train our machine learning models on accurately labeled data that approximate the true one. After this step, we subsequently will make predictions on the true test data. We are taking this research direction because we believe that a significant problem that jeopardized our results was the presence of many mislabeled samples.

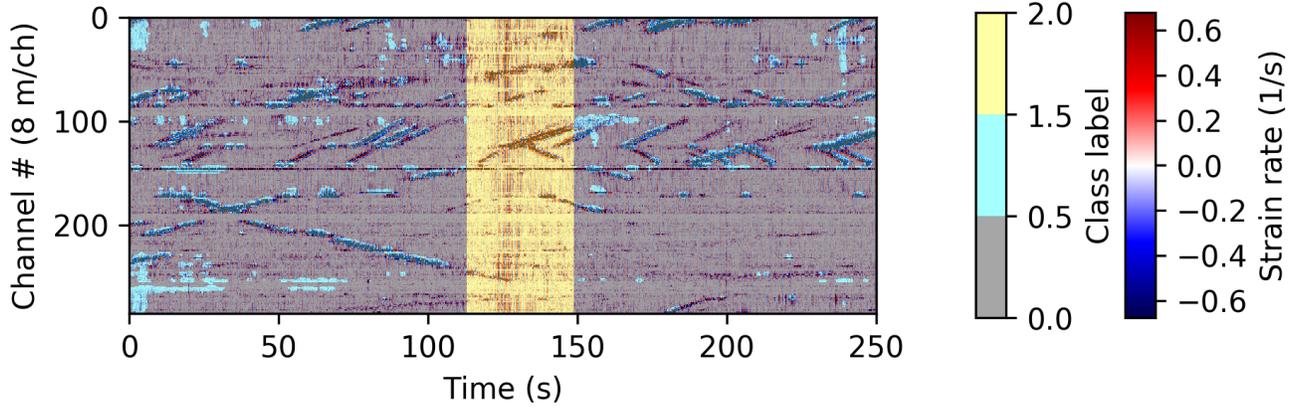


Figure 2: 250 seconds of recorded, labeled data. This section of data contains a recording of a quarry blast which excited all channels simultaneously. The data were bandpassed from 0.23 to 2.0 Hz.

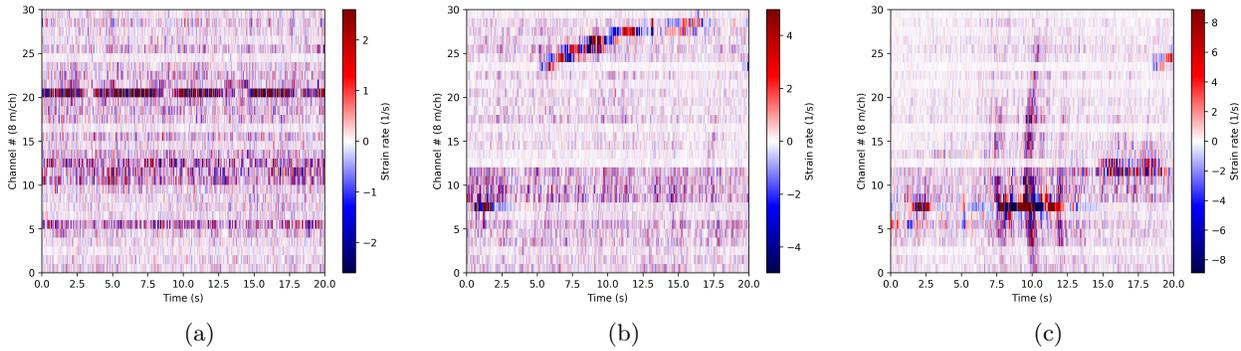


Figure 3: Window containing (a) ambient noise, (b) urban noise, and (c) earthquake amplitudes.

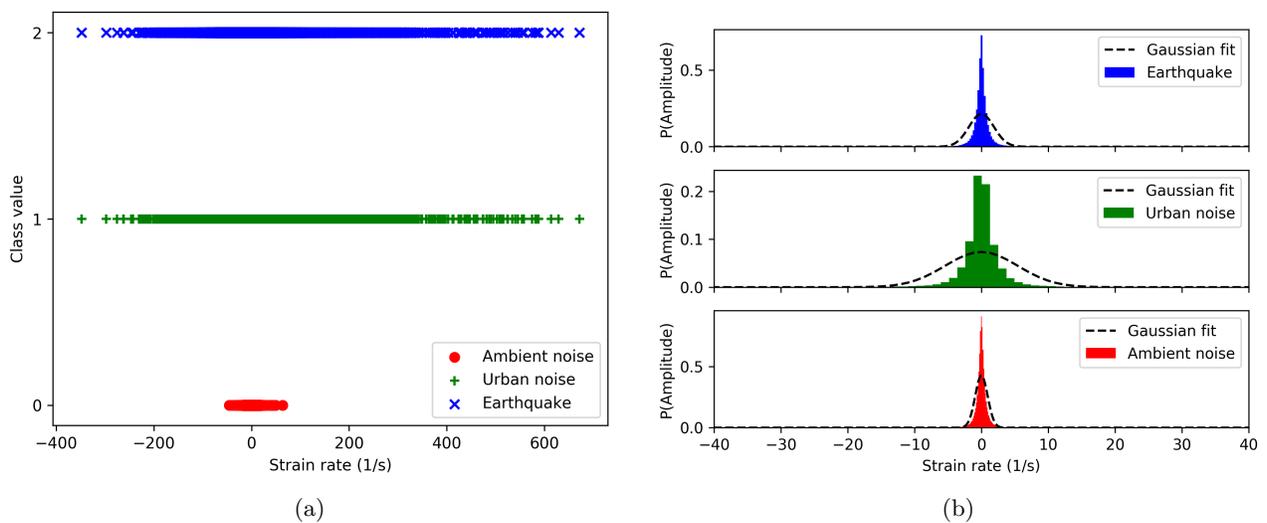


Figure 4: (a) Plot of amplitudes for each class label. Note the amplitudes entirely overlap within the single amplitude feature space. (b) Amplitude histograms for each class where we superimpose their respective Gaussian fit.

## Contributions

Fantine Huot provided the labels for the urban noise. Ettore and Joseph equally contributed to every single step in the project (writing the code, report, poster, etc.)

## References

- [1] S. V. Shatalin, V. N. Treschikov, and A. J. Rogers, “Interferometric optical time-domain reflectometry for distributed optical-fiber sensing,” *Applied optics*, vol. 37, no. 24, pp. 5600–5604, 1998.
- [2] T. M. Daley, B. M. Freifeld, J. Ajo-Franklin, S. Dou, R. Pevzner, V. Shulakova, S. Kashikar, D. E. Miller, J. Goetz, J. Henninges *et al.*, “Field testing of fiber-optic distributed acoustic sensing (DAS) for subsurface seismic monitoring,” *The Leading Edge*, vol. 32, no. 6, pp. 699–706, 2013.
- [3] A. Mateeva, J. Lopez, H. Potters, J. Mestayer, B. Cox, D. Kiyashchenko, P. Wills, S. Grandi, K. Hornman, B. Kuvshinov *et al.*, “Distributed acoustic sensing for reservoir monitoring with vertical seismic profiling,” *Geophysical Prospecting*, vol. 62, no. 4, pp. 679–692, 2014.
- [4] B. Biondi, E. Martin, S. Cole, M. Karrenbach, and N. Lindsey, “Earthquakes analysis using data recorded by the stanford das array,” in *SEG Technical Program Expanded Abstracts 2017*. Society of Exploration Geophysicists, 2017, pp. 2752–2756.
- [5] C. E. Yoon, O. O’Reilly, K. J. Bergen, and G. C. Beroza, “Earthquake detection through computationally efficient similarity search,” *Science advances*, vol. 1, no. 11, p. e1501057, 2015.
- [6] J. Lan, S. Nahavandi, T. Lan, and Y. Yin, “Recognition of moving ground targets by measuring and processing seismic signal,” *Measurement*, vol. 37, no. 2, pp. 189–199, 2005.
- [7] J. Tejedor, J. Macias-Guarasa, H. F. Martins, J. Pastor-Graells, P. Corredera, and S. Martin-Lopez, “Machine learning methods for pipeline surveillance systems based on distributed acoustic sensing: A review,” *Applied Sciences*, vol. 7, no. 8, p. 841, 2017.
- [8] B. Rouet-Leduc, C. Hulbert, N. Lubbers, K. Barros, C. Humphreys, and P. A. Johnson, “Machine learning predicts laboratory earthquakes,” *arXiv preprint arXiv:1702.05774*, 2017.
- [9] F. Huot, Y. Ma, R. Cieplicki, E. Martin, and B. Biondi, “Automatic noise exploration in urban areas,” in *SEG Technical Program Expanded Abstracts 2017*. Society of Exploration Geophysicists, 2017, pp. 5027–5032.
- [10] Y. Siyuan, E. Martin, J. P. Chang, S. Cole, and B. Biondi, “Catalog of northern california earthquakes recorded by das,” *SEP-Report*, vol. 170, pp. 81–96, 2017.
- [11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [12] T. F. Chan, G. H. Golub, and R. J. LeVeque, “Updating formulae and a pairwise algorithm for computing sample variances,” in *COMPSTAT 1982 5th Symposium held at Toulouse 1982*. Springer, 1982, pp. 30–41.