21st Australian International Aerospace Congress
ISBN number 978-1-925627-90-9

Normal Paper ☒

Blind peak detection in vibration spectra using Region-based Convolutional Neural Networks

Georgios Protopapadakis [1], Pietro Borghesani [2], Cédric Peeters [1], Nico Herwig [2], Wade Smith [2], Zhongxiao Peng [2], Jan Helsen[1]
[1] *Acoustics and Vibrations Research Group (AVRG), Vrije Universiteit Brussel, Pleinlaan 2, Brussels, 1050, Belgium*
[2]*School of Mechanical and Manufacturing Engineering, University of New South Wales, Sydney, NSW-2052, Australia*

Abstract

Detecting peaks in vibration spectra can provide early indications of potential faults or anomalies, facilitating pre-emptive maintenance to prevent costly breakdowns. It can also be used in combination with other methods. Existing methods for peak detection encounter challenges stemming from noisy data and an unknown number of expected peaks. In response, this paper proposes a Region-based Convolutional Neural Network for blind peak detection. In this instance, the model is trained on complex simulated data, with the potential for future adaptation to incorporate real data. This flexibility allows for a dynamic trade-off between data availability and machine-specificity. The proposed method comprises a classification part and a regression part. Initially, regions are classified as peak or no-peak, followed by the regression part predicting the location and size of the peak for the identified region. Overall, this methodology identifies peaks in spectra without necessitating prior knowledge about the signal or machine kinematics. The performance of the peak detection model is tested on real gearbox data. By eliminating the need for a priori knowledge and ensuring rapid execution, the method emerges as an initial solution for peak detection, contributing to the enhancement of signal analysis methodologies.

Keywords: Blind peak detection, signal processing, monitoring

Introduction

The importance of spectral peak detection has been recognised as a key method for vibration-based monitoring of mechanical equipment, particularly in industries where the components are continuously moving in a predefined manner, such as the rotation of wind turbines and generators [1]. These spectral peaks correspond to periodicities within the rotating components, and their amplitude and frequency can indicate specific faults. Although they can be used directly for monitoring, by following the amplitude or the change in the shape of the peak, information can be extracted for other metrics or methodologies. For example, the estimation of instantaneous angular speed (IAS) in a multi-order probabilistic approach [2].

Many methods exist for the identification of harmonics in a signal. The most known and widely used are MUSIC [3] and ESPRIT [4], adapted from direction of arrival (DOA) applications, to finding harmonics in time series data. The main concept for DOA and peak detection is common, and it is to retrieve harmonics from signals. The drawback of MUSIC is that the number of harmonics must be pre-defined to have accurate results, and noise must be limited. ESPRIT reduced the issue of noise in signals with low signal-to-noise ratios (SNR). Nevertheless, until today, issues still arise caused by noise. Hawarri et al. have proposed a method to further reduce the noise issue by standardising the spectral data [5]. Although this work was a step forward, issues remain when the machine is operating in a transient regime.

Machine learning has been explored to improve Direction of Arrival (DOA) or peak detection capabilities [6]. A comprehensive review of deep learning approaches for DOA problems is available in [7]. While deep learning holds promise for improving predictive

accuracy, a major limitation is its "black box" nature, where results are often not easily interpretable. This lack of transparency can be a deterrent for safety-critical applications. Nevertheless, certain machine learning models offer the potential for greater interpretability, making them more suitable for such applications.

Region Proposal Networks (RPN), introduced by Ren et al. in their Faster RCNN work [8], stood out as both highly effective in object detection and relatively interpretable, offering insights into its decision-making process. In fact, Faster RCNN has seen successful application in peak detection across various other fields [9, 10, 11, 12].

Although RCNNs have been effectively used for peak detection in general, they are yet to be applied to vibration spectra. Given the unique challenges of vibration data, such as low signal-to-noise ratios, non-Gaussian noise, and an unknown number of peaks, traditional methods often struggle with accurate peak detection. This work aims to address these challenges by developing a harmonic detection model tailored to vibration spectra, ignoring for the moment the resonances and the modes of the system.

The following sections detail the model development methodology, with the third section presenting the training results and the accuracy of the model on raw spectral data from offshore wind turbines. The conclusions are provided in the final section.

Methodology – Model Structure

As for the most common case of RCNN-based object recognition within images, labelling of an input sample (in this study a spectrum) is based on the definition of *anchors*. Anchors consist of a set of predefined rectangular boxes with various sizes and aspect ratios. These boxes are slid over the image at regular intervals in both directions, in order to determine if an object is present in that area and its size and location relative to the anchor itself. This labelling usually consists of 5 elements: the presence of an object, x and y location with respect to the centre of the anchor, and width and height as a percentage of the anchor's corresponding dimensions. This labelling is then used to classify the presence and location of objects starting from a CNN-preprocessed version of the image itself. Each pixel of such a pre-processed image, containing many channels, is then fed to a series of dense layers which outputs the 5 elements required for the classification (object or not) and regression (location and size) tasks.

This study borrows the architecture of Faster RCNN and applies it to spectra for the detection, location and sizing of peaks corresponding to periodic or almost-periodic forcing function harmonics in a mechanical signal. In other works using RCNNs for peak detection [9, 10, 11, 12], images of spectra or spectrograms are used as inputs to the model, in order to allow using the standard RCNN algorithm. However, this study is based on the more justifiable and efficient choice of using the actual 1D vectors of the signal's spectral log-amplitude as inputs. For the details of the functioning of traditional 2D RCNNs, the reader is referred to Ref. [8].

The inputs to the network used in this work are one-dimensional log-amplitude spectra. Each log-amplitude spectrum $\mathbf{X}_i$ is composed of a series of real values $X_i[n]$ with $n = 0, \dots, L-1$, each corresponding to the logarithm of the spectral amplitude at the normalised frequency $n = f[n]/\Delta f$ where Δf is the frequency resolution of the spectrum and $L = 2048$ frequency points. The spectra are obtained by fast-Fourier transform (FFT) of time-domain signals $x_i[k]$.

The network structure is shown in Figure 1. The network is composed of a U-Net (in red), and two fully-connected networks (object detection and object location). The U-Net takes as input the spectrum and applies a series of convolutional-max pooling layer pairs, each with $N = 64$ filters, kernel size $K = 13$ (except the first layer which has $K = 17$) and pool size $P = 2$. As shown in [13], each convolutional layer is equivalent to a liftering operation (filtering in the cepstrum domain), which for each filter (or lifter) allows only certain spectral patterns to be retained. The max-pooling just allows each subsequent convolutional layer to operate over a downsampled spectrum (coarser frequency resolution), i.e., enabling the filters (kernels) to span over larger bandwidths. In line with the U-Net concept, the outputs of all convolutional layers

are retained to increase the capacity of the network to focus on spectral patterns of different scale. These are upsampled to a uniform dimension and corresponding frequency resolution (matching the original spectrum) and concatenated along the channel dimensions (i.e., stacked). The output of such U-Net is therefore a series of 256 pre-processed versions of the original spectrum, each with a potentially different combination of liftering operations applied to it.

These are fed to the almost-identical object detection and object location/sizing networks. The first layer of each network is equivalent to a dense layer, even if it's implemented as a convolutional layer with kernel size $K = 1$ and $N = 64$ filters. This means that the same 64-neuron dense layer is applied to each spectral point independently, i.e., for each frequency, all the differently liftered versions of the corresponding spectral amplitude are combined linearly into 64 output values and then saturated using a ReLU activation function. The output of this first dense layer is then fed to a second dense layer, again implemented as a convolutional layer with kernel size $K = 1$. This last layer is different for the two networks: in the object detection network it has $N = 4$ filters/neurons and a sigmoid activation function, while in the object location/sizing network it has $N = 8$ filters/neurons, still with a sigmoid function.

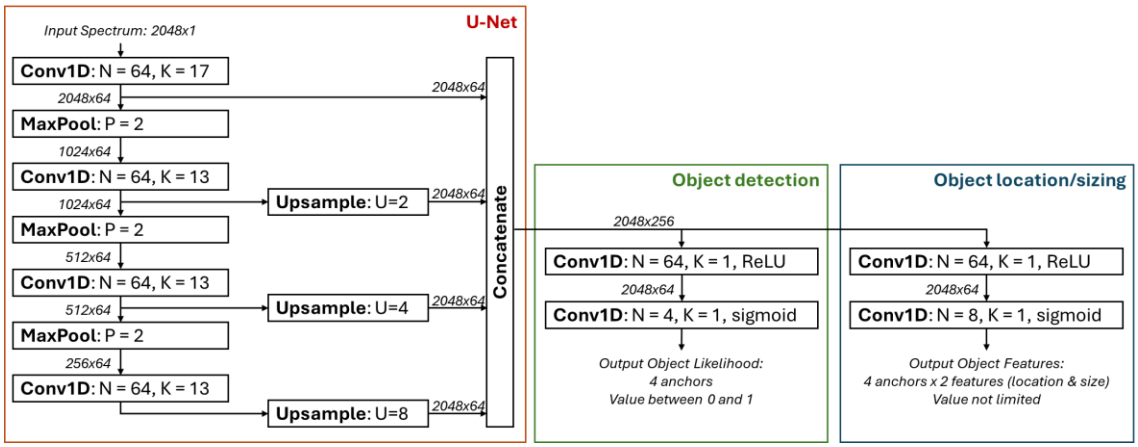

Figure 1. Structure of the network

The dimensionality of the outputs is determined by the input labelling process, following the standard RCNN approach. The labelling involves $A = 4$ *anchors* of varying sizes (5, 11, 22, and 45 frequency points). Each anchor is slid along the frequency axis, centred at each frequency point. For each frequency $f[n]$ and anchor a, if the maximum intersection-over-union (IoU) between the peaks and the anchor exceeds 0.7, the n-th frequency component of the spectrum $X_i[n]$ is labelled as containing a peak for anchor a (i.e., $y_i[n, a] = 1$). When this condition is met, the location and size of the peak relative to the anchor a centred in $f[n]$ are recorded as $\ell_i[n, a]$ and $s_i[n, a]$, respectively.

The network is then trained with input spectra $X_i[n]$ and labels $y_i[n, a]$, $\ell_i[n, a]$ and $s_i[n, a]$ with $n = 0, \dots, L-1$ and $a = 0, \dots, A-1$. The number of anchors is reflected in the dimensionality of the network's outputs: the object-detection network outputs $\hat{y}_i[n, a]$ with 4 channels ($a = 0, \dots 3$), and the object-location/sizing outputs $\hat{\ell}_i[n, a]$ and $\hat{s}_i[n, a]$, i.e., a total of $2 \cdot 4 = 8$ channels. The sigmoid activation at the end of the object-detection network is justified by the need to have a $0 < \hat{y}_i[n, a] < 1$ to represent the likelihood of having a peak in the corresponding anchor (classification). Despite the regression nature of the object location/sizing network, both location and size are bound by the anchor maximum dimension, thus encouraging the inclusion of a sigmoid function. The network is trained using a loss function which combines the three different outputs for each anchor:

$$\mathcal{L} = \sum_{i=0}^{I-1} \sum_{n=0}^{N-1} \sum_{n=0}^{A-1} \mathcal{L}_i^{(y)}[n, a] + b\mathcal{L}_i^{(\ell)}[n, a] + b\mathcal{L}_i^{(s)}[n, a] \,. \tag{1}$$

A binary cross-entropy component is used for the object likelihood output $\hat{y}_i[n, a]$ (classification component). The constant b was tuned for optimal accuracy,

$$\mathcal{L}_i^{(y)}[n,a] = -(y_i[n,a]\log(\hat{y}_i[n,a]) + \hat{y}_i[n,a]\log(y_i[n,a])). \tag{2}$$

A modified mean-squared error is used for the location $\hat{\ell}_i[n,a]$ and sizing $\hat{s}_i[n,a]$ of the peak

$$\begin{aligned}\mathcal{L}_i^{(\ell)}[n,a] &= y_i[n,a]\cdot\left|\hat{\ell}_i[n,a]-\ell_i[n,a]\right|^2,\\ \mathcal{L}_i^{(s)}[n,a] &= y_i[n,a]\cdot|\hat{s}_i[n,a]-s_i[n,a]|^2.\end{aligned} \tag{3}$$

In this case, the term $y_i[n,a]$ ensures that only location and size of actual peaks are considered. As for all RCNNs, the classification/regression is followed by a non-maximum suppression (NMS) of adjacent peaks, which simply eliminates all identified peaks with a reciprocal IoU greater than 20%, except for the peak showing the maximum object-detection score $y_i[n,a]$.

Training and Experimental Results

The training of the algorithm is done purely on simulated data, to test the capability of this approach to work in data-scarce environments and still perform well over experimental signals. A total of 1500 numerical signals $\mathbf{x}_i$ were created in the time domain, and then transformed into log-amplitude spectra $X_i[n]$ (each of length $L = 2048$ frequency points) using an FFT. Each signal $x_i[k]$ is composed as an additive combination of variable-speed harmonic components and noise $v_i[k]$, all convolved with a multi-degree-of-freedom impulse response $h_i[k]$:

$$x_i[k] = h_i[k] * \left(\frac{1}{m_i}\sum_{h=0}^{C_i} A_{i,c}\sin\left(2\pi\theta_{i,c}[k]\right) + v_i[k]\right) \tag{4}$$

The angular quantities $\theta_{i,c}[k]$ are obtained by numerical integration of corresponding angular speeds $\omega_{i,c}[k]$, which are in turn generated independently for each signal i and component c using a smoothed random walk. The real-valued amplitudes $A_{i,c} \in [0.45, 1.4]$ and the integer number of harmonics $5 \le C_i \le 40$ are generated using independent distributions. The combination of harmonics is then normalized with respect to its maximum value m_i and white noise $v_i[k]$ is added with an average signal-to-noise ratio $SNR = 25$, varying for each signal. The convolution with the impulse response $h_i[k]$ is actually obtained in the frequency domain using a corresponding Frequency Response Function (FRF) $H_i[n]$, generated for each signal independently with a number of random poles and zeros between 1 and 4. The frequency location of poles and zeros and the damping ratios are randomly and independently generated for each signal with the former between 10% and 90% of the available frequency range and the latter between 0.25% and 2%. As mentioned earlier, the real-valued logarithm FFT of $x[t]$ is used for training the model.

The data are split into training (90%) and test (10%) data sets. The results for the peak-detection component of the network (object detection) are presented in Table 2.

Table 2: Results of the object detection component of the network

Data set	Precision	Recall	F1-Score
Training	0.916	0.819	0.865
Testing	0.918	0.823	0.868

It can be seen that for both datasets there is a good precision (false peaks are <10%), while the recall is still relatively low (~18% of the peaks are not identified). This could be attributed to the fact that the threshold for detections in the NMS is quite high, but it is also exacerbated by the relatively low SNR experienced in the vicinity of minor peaks. Further work could analyse the effect of NMS and relative peak/background amplitude on these metrics.

The results for the peak location/sizing part are shown in Table 3, which contains the RMSE of the location and size of correctly identified peaks. As can be seen, the location is quite well estimated, with an RMSE error about twice the frequency resolution. The estimated peak widths

are less accurate with an RMSE>4 frequency points. This is very likely related to the local SNR of each peak, which would strongly affect the "perceived" width of the peaks. The importance of this error is also likely dependent on the spectral resolution and the actual application.

Table 3: Results for object detection

RMSE for	**Location** (points)	**Size** (points)
Training set	1.89	4.01
Testing set	2.11	4.27

For a qualitative, but more intuitive understanding of the performance of this approach, especially in terms of detection and sizing of the most relevant peaks in the spectrum, real accelerometer data from an offshore wind farm is utilised. The two test signals are selected randomly from different turbines across the same wind farm, one being in a quasi-stationary case (small IAS variation) and one undergoing a speed transient.

The model is applied directly to the real-valued logarithmic FFT spectra of experimental data after being trained with the numerical signals discussed before. An object score threshold of 45% is employed for NMS, allowing the inclusion of less certain detections, without compromising the precision.

The variability in the frequency width of harmonics across operating conditions and along the spectral axis can lead to anchors being either undersized or oversized for specific peaks, or to the need to use a large number of anchors. To mitigate this, multiple resolutions of the spectrum can be analysed by the model. This approach avoids increasing the number of anchors or extending the peak width variability in the training signals. This would result in unnecessary complexity, and more elaborate training datasets, potentially compromising detection performance. For this study, results are obtained by applying the methodology with the base resolution Δf first, and then again on two halves of the spectrum with a resolution $\Delta f/2$(so that the length was always 2048 points). Duplicates were removed based on IoU.

Figure 2 and Figure 3 present the results obtained from wind turbine data. The peaks detected by the network are marked with a red "X", and the red shaded boxes show the network's

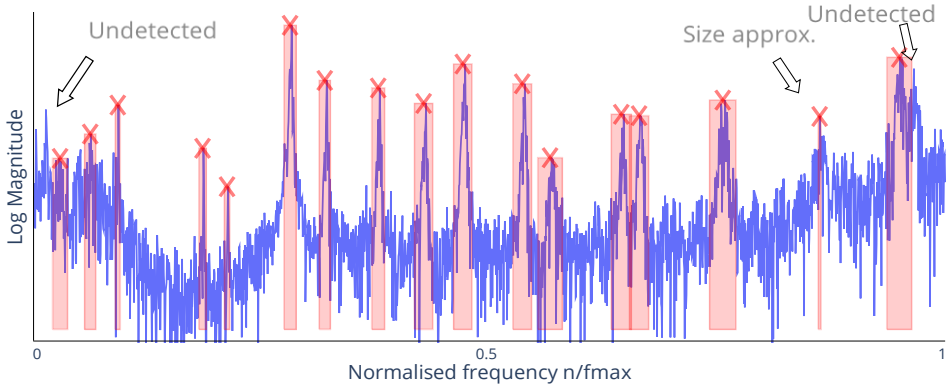

Figure 2: Results for quasi-stationary case (frequency is normalised)

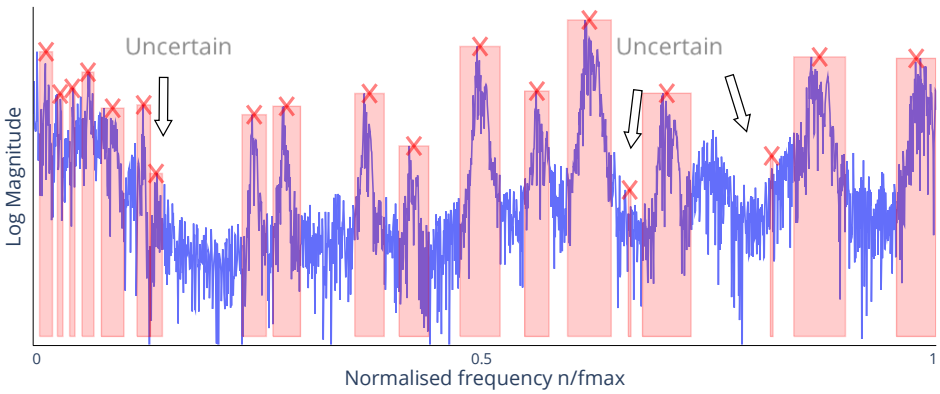

Figure 3: Results for transient case (frequency is normalised)

for the peak width. Due to its lower variation in speed, the first example (Figure 2) shows narrow peaks compared to the second (Figure 3). In both cases, most of the visually distinguishable harmonics are correctly identified by the model. In the stationary case, it is possible to see two peaks that are undetected, one at a very low frequency and one close to the maximum frequency of analysis. The peak locations and widths tend to be accurately captured, especially in the stationary case where they are well separated.

Conclusions

The proposed peak detection method introduces a novel approach for applying RCNNs to vibration spectra analysis, advancing the automation of peak detection without needing prior knowledge of spectral characteristics. This flexible model, trained initially on simulated data, can be used in real-world conditions as shown in a real wind farm case. Despite some margin for improvement, the model's successful detection of the dominant harmonics demonstrates its practical applicability for vibration analysis and condition monitoring tasks.

Acknowledgements

This research was supported by funding from the VLAIO blauwe cluster CSBO Core project. The authors would like to acknowledge FWO (Research Foundation - Flanders) for supporting Georgios Protopapadakis (V469723N).

References

[1] R. B. Randall, A New Method of Modeling Gear Faults, Journal of Mechanical Design 104 (2) (1982) 259–267

[2] G. Protopapadakis, C. Peeters, Q. Leclere, J. Antoni, J. Helsen, Enhancing instantaneous angular speed estimation with an adaptive multi-order probabilistic approach, at SSRN (2024).

[3] R. Schmidt, Multiple emitter location and signal parameter estimation, IEEE Transactions on Antennas and Propagation 34 (3) (1986) 276–280

[4] D. Kundu, Modified music algorithm for estimating doa of signals, Signal Processing 48 (1) (1996) 85–90

[5] Y. Hawwari, J. Antoni, H. Andre, Y. Marnissi, D. Abboud, M. El-Badaoui, Robust spectral peaks detection in vibration and acoustic signals, IEEE Transactions on Instrumentation and Measurement 71 (2022) 1–13

[6] A. M. Elbir, Deepmusic: Multiple signal classification via deep learning, IEEE Sensors Letters 4 (4) (2020) 1–4

[7] Ge, Shengguo, Li, Kuo, Rum, Siti Nurulain Binti Mohd, Deep Learning Approach in DOA Estimation: A Systematic Literature Review, *Mobile Information Systems*, 2021, 6392875, 14 pages, 2021

[8] S. Ren, K. He, R. B. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2015) 1137–1149

[9] H. Kim, S. Sim, Automated peak picking using region-based convolutional neural network for operational modal analysis, Structural Control and Health Monitoring 26 (2019)

[10] J. Zeng, H. Wu, M. He, Image classification combined with faster r–cnn for the peak detection of complex components and their metabolites in untargeted lc-hrms data, Analytica Chimica Acta 1238 (2023) 340189

[11] Y. Ji, S. Zhang, W. Xiao, Electrocardiogram classification based on faster regions with convolutional neural network, Sensors 19 (11) (2019)

[12] S. Jeong, H. Kim, J. Lee, S.-H. Sim, Automated wireless monitoring system for cable tension forces using deep learning, Structural Health Monitoring 20 (4) (2021) 1805–1821

[13] Borghesani, P., Herwig, N., Wang, W., & Antoni, J. (2023, January). Embedding signal processing knowledge in neural networks - an application to gear diagnostics. In *AIAC 2023: 20th Australian International Aerospace Congress* (pp. 669-677). Melbourne, Australia.