Article

Journal of Vibration and Control
1–19

DOI: 10.1177/1077546317723227
journals.sagepub.com/home/jvc
SAGE

Automated vibration-based fault size estimation for ball bearings using Savitzky–Golay differentiators

Mohamed AA Ismail[1], Andreas Bierig[2] and Nader Sawalhi[3]

Abstract

Vibration-based fault diagnosis has been utilized as a reliable method for identifying ball bearings health since the 1970s. Recently, there has been an increased research effort to develop methods for fault quantification with the aim of estimating the fault size to allow the service life of a ball bearing to be extended beyond the detection stage. These studies have shown that the vibration signal from a localized spall (e.g. fatigue defect) in a ball bearing exhibits features corresponding to two main events, namely, the entry into and the exit from the spall. The time span between these two events is correlated with the spall size. Studies have shown that the entry into the spall is the more challenging event to identify, which often requires extensive signal processing techniques. This paper introduces an automated vibration-based technique for estimating the size of a spall in a ball bearing under axial loading conditions similar to those of linear electro-mechanical actuators. This technique is based on the extraction of the entry/exit events from the vibrational jerk, which are numerically determined from accelerometer data. The differentiation of the acceleration data to estimate jerk signal is performed using a variant of Savitzky–Golay (SG) differentiators, which provide enhancement for the detection of the entry and exit points. Sensible spall size estimations have been achieved for 24 different scenarios of fault sizes, rotor speeds and loads measured on a test rig provided by DLR (German Aerospace Center).

Keywords

Bearing faults, spall quantification, prognostics, vibration-based condition monitoring, Savitzky–Golay differentiators

1. Introduction

Condition monitoring (CM) techniques have a common objective of detecting system faults as early as possible before they lead to machine failure. CM has two different scopes: fault diagnostics and prognostics (Vachtsevanos et al., 2006). Fault diagnosis involves fault detection and identification, whereas fault prognosis aims to track the growth of a fault from a low- to a high-criticality condition prior to failure. The severity of the fault (e.g. fault size) is predicted to estimate the remaining useful life (RUL) of the faulty part before an urgent intervention should be performed (Ismail et al., 2016).

Vibration-based spall detection for ball bearings by monitoring specific fault characteristic frequencies has been well established since the 1970s (McFadden and Smith, 1984; Randall, 2011a). These frequencies indicate the existence of a spall and they depend on operating speed and bearing geometry, all of which are measurable. However, the spall size cannot be directly extracted through frequency analysis (Randall, 2011b; Jena et al., 2012). The ultimate benefit of estimating the spall size is to allow the service life of a ball bearing to be extended even after detection, as long as the spall size has not yet reached a critical threshold (Sawalhi, 2007; Ismail et al., 2016).

[1]PhD Student at Mechanical Engineering, Technische Universität Braunschweig, Braunschweig, Germany
[2]Institute of Flight Systems, DLR (German Aerospace Center), Braunschweig, Germany
[3]College of Engineering, Prince Mohammad Bin Fahd University, Al-Khober, Saudi Arabia

Received: 22 September 2016; accepted: 6 July 2017

Corresponding author:
Mohamed AA Ismail, Safety Critical Systems and System Engineering, Institute of Flight Systems, DLR (German Aerospace Center), Lilienthalplatz 7, 38108 Braunschweig, Germany.
Email: Mohamed.Ismail@dlr.de

1.1. Related work and challenges in spall quantification

Newly emerging studies (Sawalhi and Randall, 2011; Jena et al., 2012; Kogan et al., 2015; Moustafa et al., 2016) have shown that the fault response of a bearing in fact contains features reflecting two events that are associated with the spall width and how a ball enters and exits the spall. These points are: the entry point into the spall, "A", and the exit point from the spall, "B", as shown in Figure 1(a). When a ball rolls off the main level of the track at the entry point, a short-time excitation is generated due to the gradual decrease of the ball's load (destress). At the exit, the ball strikes the exit edge of the spall, causing a new short-time excitation (an impact), which decays in longer interval due to mechanical damping of the machine, as shown in Figure 1(a).

Two different forms of entry/exit observations have been reported in the literature, namely, step-impulse observations and double-pulse observations. Step-impulse observations have been reported by Sawalhi and Randall (2011) and later by Jena et al. (2012), Kogan et al. (2015), and Moustafa et al. (2016). The entry event "A" was observed to induce a very weak excitation dominated by low-frequency content (a destress), as shown in Figure 1(a). Event "A" was associated with a disturbance of the ball's motion path without any significant striking of the ball against the spall's leading edge, similar to a step response as approximated by Sawalhi and Randall (2011). Subsequently, the ball was found to strongly strike the spall (an impact), giving rise to an impulse response, at the exit "B". The exit event "B" is difficult to identify because of background noise, mechanical damping and oscillating movement of the ball. Double-pulse observations have also been reported by other researchers (Sawalhi, 2007; Moustafa et al., 2016). In this case, the ball rolls off the leading edge and onto the exit edge of the spall with similar excitations, creating two similar pulses, as depicted in Figure 1(b). The first pulse starts after the entry into the spall zone at point "A", and the second pulse occurs before the exit point "B". Each observation, i.e. step-impulse or double pulse, involves specific time and frequency features that are influenced by the spall size and loading conditions (Sawalhi and Randall, 2008).

For the application of step-impulse observations, Sawalhi and Randall (2011) presented two methods of measuring the spall size through intensive signal processing schemes that make use of signal pre-whitening, wavelet operations and the cepstrum to enhance the entry feature under the assumption that the vibration response of the spall will always occur in the step-impulse form. Relying on a similar principle, Kogan et al. (2015) investigated a method for locating the entry/exit points based on knowledge gained from a dynamic model for the ball motion over a spalled zone.

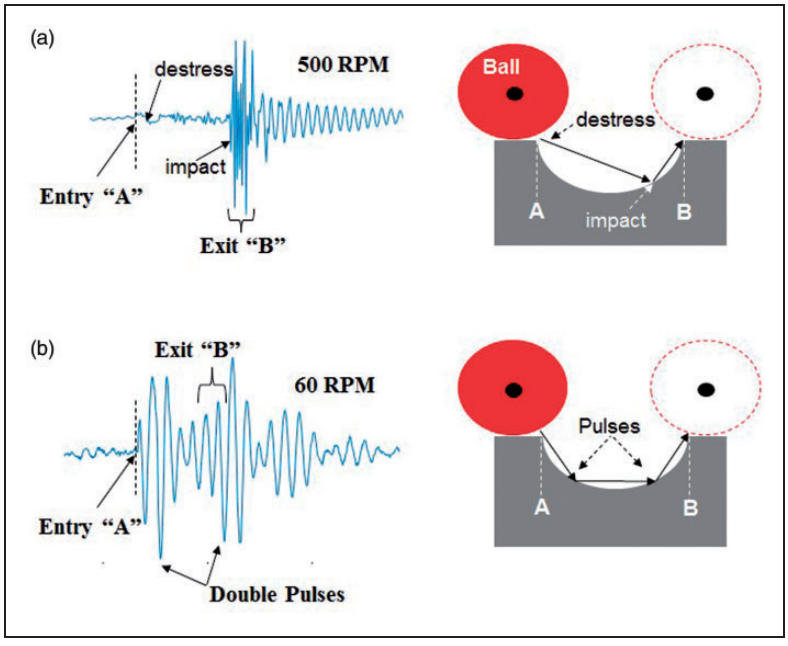

Figure 1. Different response forms of entry/exit of a spall: (a) at 500 r/min, a step-impulse observation is shown, (b) at 60 r/min, a double pulse indicates the spall entry and exit. Both (a) and (b) responses are extracted from the data used in the paper. The entry/exit points are hard to be located directly.

Table 1. Summary of the spall quantification literature based on entry/exit features.

	Study	Speed range (r/min)	Load (kN)	Spall sizes (mm)	Remarks
1	Epps and McCallion (1994)	1500–3000	1.5–7	0.1–3	■ Two events ■ Direct analysis in the time domain
2	Sawalhi (2007)	600–1000	50–100 N m radial torque	0.6–1.2	■ Double pulse observed in data collected from a gearbox ■ Autoregressive models, complex Morlet wavelet and minimum entropy deconvolution
3	Sawalhi and Randall (2011)	800–2400	–	0.6–1.2	■ Step-impulse signals ■ Joint treatment based on wavelet pre-whitening ■ Separate treatment based on cepstrum technique
4	Jena et al. (2012)	1500	–	2.1	■ Step-impulse signals ■ Analytical wavelet transform
5	Moustafa et al. (2016)	10–60	1–10	0.7–6	■ Step-impulse or double-pulse signals depending on the size ■ Instantaneous Angular Speed (IAS)
6	Kogan et al. (2015)	1200	0.14	0.4–2.5	■ Step-impulse signals ■ Band-pass and high-pass filters tuned based on a 3D model
7	This work	60–500	5–8.8	1–4.34	■ Step-impulse or double-pulse signals depending on the speed and spall size ■ Savitzky–Golay differentiators with automatically adjustable parameters

The entry point is assumed to be located at the first maximum in a band-passed signal, whereas the exit point is located based on the decay of the last impulse response in a high pass signal. This method achieves a maximum error of 25% for spall widths in the range of 0.6–2.5 mm. However, no systematic method was provided for determining the best cut-off frequencies for these filters or for selecting a threshold for choosing the exit point from among very similar filtered peaks. The operating speeds and spall sizes addressed in these studies are provided in Table 1. The challenges of spall quantification can be summarized as follows:

- The studies listed in Table 1 were conducted for a limited range of operating conditions and spall sizes. Each study presents an interpretation of a vibration response that matches the investigated operating speeds, i.e. for step-impulse or double pulse.
- The localization of the entry and exit points in different vibration forms for similar conditions, such as those shown in Figure 2(a) and (b), has not been investigated by a generalized criterion.
- The target application considered in this work is a linear electro-mechanical actuator, which is a typical variable-speed application. For this application, the quantification process should be applicable for a wide range of operating speeds.

The objective of this research is to develop a systematic quantification method for spalls in bearings based on the entry/exit features applicable to different operating speeds. The quantification method involves an automated vibration-based technique for monitoring the size of a spall based on the vibrational jerk, which are numerically determined from an accelerometer mounted on the spalled bearing. Once the fault has been detected and confirmed using well-known tools such as envelope analysis (Randall, 2011a), the quantification stage extracts the entry/exit events from the vibration. The fault size is calculated from the measured time between the entry and exit features after appropriate scaling based on the bearing geometry, the sampling frequency and the rotational speed.

This paper is organized into four sections. In Section 2, the underlying idea of the quantification method is introduced, which includes an introduction to Savitzky–Golay (SG) differentiators and the automatic tuning of SG parameters. Section 3 provides descriptions of seeded fault conditions and the quantification results for two detailed

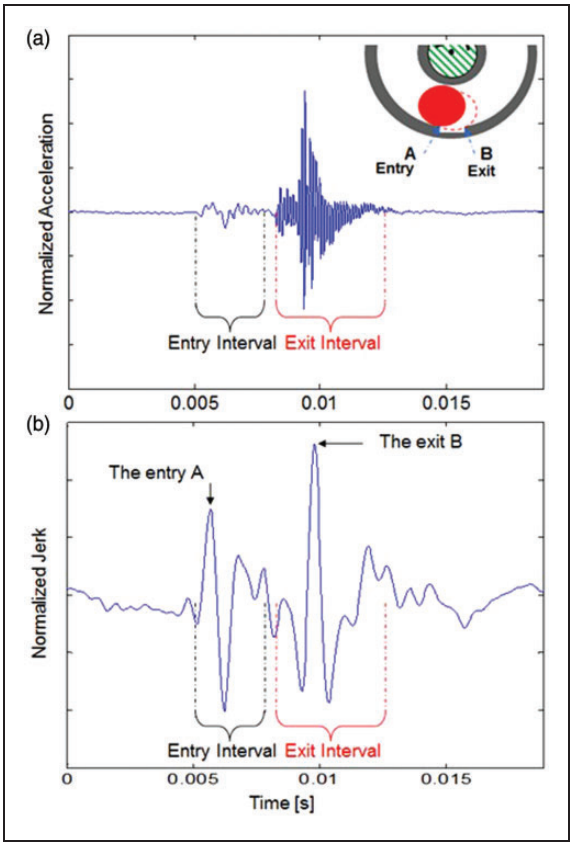

Figure 2. The concept of the entry/exit intervals: In (a), the vibration signal of a spall is shown, where the exact entry/exit points are difficult to locate. In (b), the jerk signal (for a tuned differentiator) exhibits two local maxima, one within the entry interval and one within the exit interval. These maxima may both lie either in the positive or negative sides. It depends on the earlier peak of the entry which is located here in the positive side.

examples and for whole datasets. The last section presents the research contributions.

2. Fault quantification method

The quantification method is based on extracting the average elapsed time between the instant of entry to the spall and the instant of exit. The actual spall size can be calculated as a function of the elapsed time, the operating speed and the bearing geometry (Sawalhi and Randall, 2011) as follows:

$$L = T\frac{0.5\pi f_r\left(D_P^2 - d^2\right)}{D_p} \tag{1}$$

where T is the elapsed time between entry and exit, L is the spall width in mm, f_r is the shaft rotation frequency in Hz, D_p is the pitch diameter in mm, and d is the ball diameter in mm. This formula can be used for both inner- and outer-race faults with a small error depending on the ratio of the ball to the pitch diameters of the bearing.

As an alternative to a direct analysis of the vibration acceleration response for entry/exit points, we here propose the following procedure:

1. A vibration signal consists of two short-time excitations, the entry interval and the exit interval. Both of them are associated with unknown instants (A and B) at which the ball enters and departs the spall as shown in Figure 2(a).
2. The entry and exit instants can be approximately localized by finding the highest rate of change in the vibration signal within the entry interval and the exit interval respectively as shown in Figure 2(b).
3. The rate of change in the acceleration physically corresponds to the jerk, which thus can be utilized for analysis as an alternative to the acceleration signal. The jerk is estimated here by a numerical differentiator applied to accelerometer measurements rather than using a jerk sensor.

A basic representation of an ideal differentiator (ID) is depicted in Figure 3(a); in which, the output signal is proportional to the frequency. In this case, high frequencies of the input signal (involve the background noise) are subjected to large amplifications. Such ideal differentiation is hard to be realized, and a low pass differentiator variant, commonly known as a practical differentiator, are widely used instead (Bakshi and Godse, 2008). The typical frequency response of practical differentiators consists of two parts: an ID and a low pass filter (LPF) part which approximates an integrator as shown in Figure 3(a). For the ID, the output is proportional to the frequency up to a maximum limit. On the other hand, the LPF part provides a smoothing and low pass filtration to higher frequencies. In this work, a practical differentiator is used to provide two signal enhancements for the entry and exit as follows:

1. The tuned differentiator enhances the entry event, which is dominated by a low frequency content, by providing an increasing differentiator gain (i.e. an ID effect for low frequencies) in order to increase the contrast between the entry peak and background noise.
2. The tuned differentiator enhances the exit event, which is dominated by a high frequency content, by providing a decreasing gain (i.e. low pass filtration for high frequencies) in order to maintain the exit peaks close to the amplified entry peaks. An example for entry/exit enhancement for a simulated signal is shown in Figure 3(a–c).

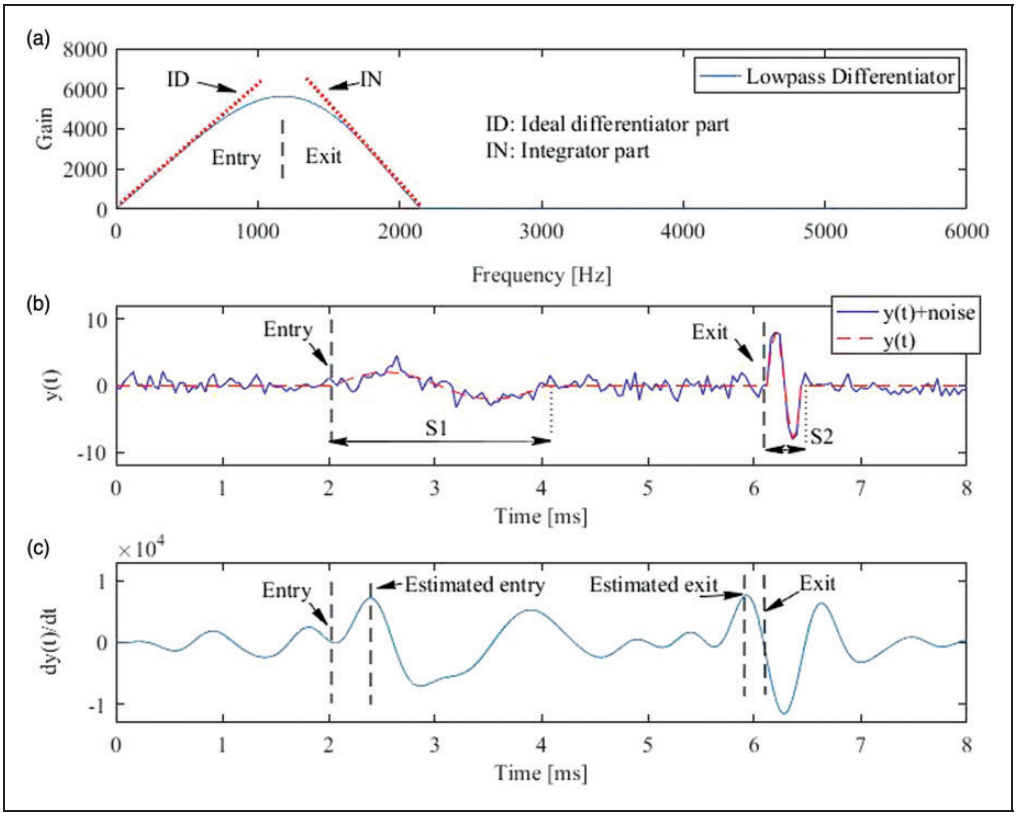

Figure 3. Frequency responses of a practical differentiator in (a) consists of an ideal differentiator and a low pass filter (an integrator), in addition to proposed locations of entry/exit events after tuning. Signal enhancement concept for a simulated signal is shown in (b) and (c). In (b) the entry/exit are modeled by $S1 = 2\sin(2\pi 500t)$ and $S2 = 8\sin(2\pi 2000t)$ respectively, in addition to a noise of 2 SNR (signal to noise Ratio) of variances. The tuned differentiator in (a) is used to estimate the entry/exit peaks in (c) that are better distinguishable than to (b).

Achieving these enhancements requires a specific tuning for differentiator parameters. An automated tuning method for axial-load bearings will be described in detail in the following section.

2.1. Savitzky–Golay filters

The quantification method, introduced in the last section, requires a numerical algorithm to realize such practical differentiators. Savitzky–Golay filters (SGFs) represent a generalized numerical method for smoothing and differentiating noisy data. For example, commonly known moving average filters and finite difference approximation of derivatives can be considered as special SGFs (Schafer, 2011). SGFs were introduced (Savitzky and Golay, 1964) to provide the capability of data smoothing and differentiation while largely maintaining the waveform of the underlying signal (e.g. the width and height). The principle of SGFs is to provide an approximation of a set of noisy data points $x[n]$ of length $2M + 1$ by fitting them to a polynomial $p(n)$ of order N as follows (Schafer, 2011):

$$p(n) = \sum_{k=0}^{N} c_k n^k \quad -M \le n \le M \tag{2}$$

where the c_k denote the coefficients of a polynomial of length $N + 1$. The input data are selected by a window that has a length of $2M + 1$ and is centered at $n = 0$. Two distinct cases of the fitting of these data to $p(n)$ can be identified:

2.1.1. Case 1: Exact solution: if $2M = N$. In this case, the number of unknown polynomial coefficients $(N + 1)$ is equal to the number of data points $(2M + 1)$, and the fit will therefore converge to a unique solution for the coefficients. This case will not cause any smoothing to the data. However, it is useful for deriving finite/central difference formulas to perform direct differentiation for a maximum differentiation order equal to in the number of data points (2).

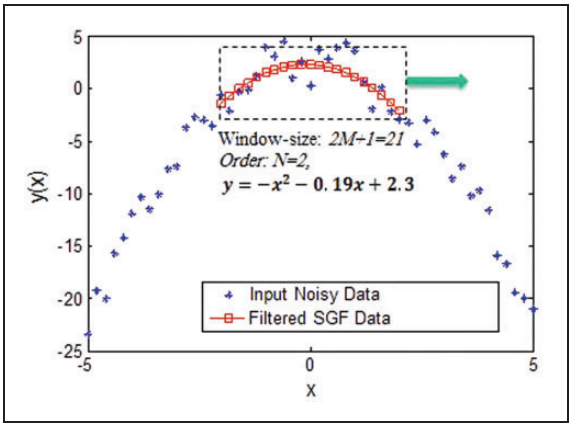

Figure 4. An example of using SGFs for smoothing/differentiating noisy data. A movable data window of 21 points is used to be locally fitted to $y(x) = -x^2 - 0.19x + 2.3$. The filtered smoothed value at the center of the frame ($x = 0$) corresponds to the absolute term $c_0 = 2.3$, additionally two smoothed derivatives are available at the same point, $dy/dx = c_1 = -0.19$ and $d^2y/dx^2 = 2c_2 = -2$. To evaluate a different point "x", the data window is centered at that point and a new polynomial is calculated.

2.1.2. Case 2: Optimal Solution: if 2M > N. This case entails an optimal fitting of the coefficients of $p(n)$. The number of unknown coefficients $(N + 1)$ is selected, according to a certain criterion, to be fewer than the number of data in the frame $(2M + 1)$. An example of this case is depicted in Figure 4.

A computational advantage of SGF is that the signal derivatives are directly associated with the vector of the polynomial coefficients c, as shown in equations (3) and (4):

$$\frac{\partial p(n)}{\partial n} = c_1 + 2c_2 n + \cdots + Nc_N n^{N-1} \tag{3}$$

$$\frac{\partial p(n)}{\partial n}_{n=0} = c_1 \tag{4}$$

The systematic process of SGF-based smoothing/differentiating can be summarized as follows:

– Select a data window of size $F = 2M + 1$ that is centered at the desired point.

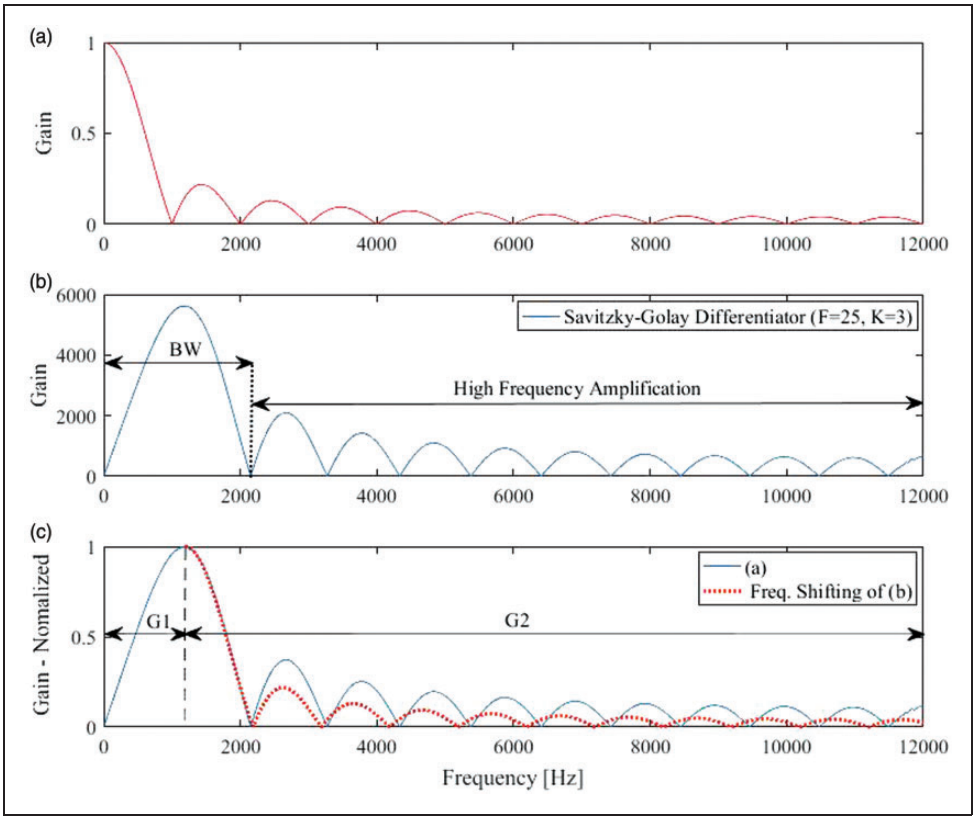

Figure 5. A comparison between a moving average filter (a) and a SGD (b) is shown in based on their frequency responses. The term bandwidth (BW) indicates the largest amplification band which includes an ideal differentiation part (G1) and a low pass smoothing filter (G2). The comparison in (c) shows that G2 part can be approximated by a moving average filter shifted by G1, where for SGDs G1 = 0.5 BW.

– Select a polynomial $p(x)$ of order K: $p(x) = c_0 + c_1x + \cdots + c_kx^K$
– Fit $p(x)$ to the window points by estimating $[c_0.c_1 \ldots . c_k]$.

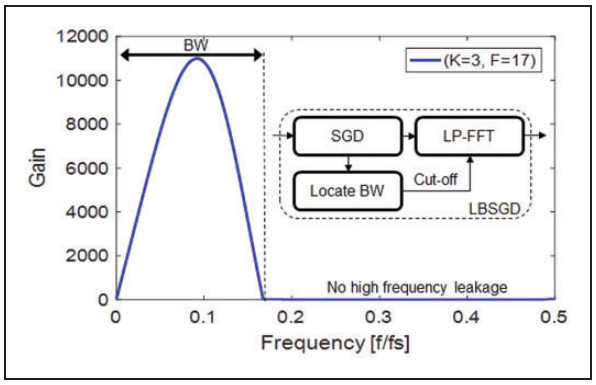

Figure 6. Frequency response of a LBSGD and the corresponding block diagram. The BW is determined using *F*, *K*. High frequencies amplification is completely rejected.

– The smoothed output of the data at the window center $n = 0$ equals to c_0.
– The first derivative of the data, dy/dx, at the window center, is equals to c_1.
– Repeat this process until all data have been scanned.

Here, our scope is limited to Savitzky–Golay differentiators (SGDs) rather than general SGFs.

2.1.3. Spectral characteristics. The frequency response of SGDs is crucial to interpret different enhancement effects to the signal. The first time derivative version of SGDs is defined by: the window size (F) and the polynomial order (K). These parameters are used to define a specific frequency response for the SGD, as shown in Figure 5. The frequency response involves two parts: the first one referred as bandwidth (BW) and a second, unwanted part, the high frequency leakage. The BW part includes the largest differentiator gain, and it matches the frequency response of a practical differentiator shown in Figure 3(a). As shown in

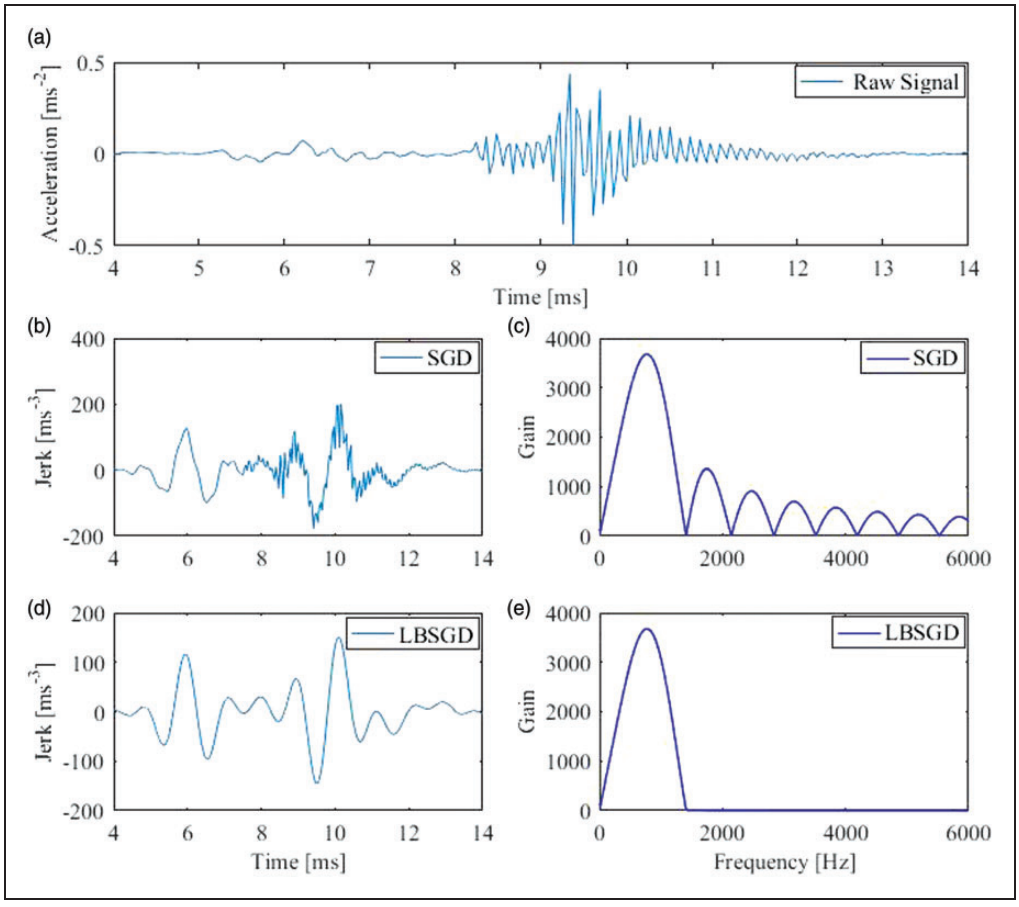

Figure 7. A comparison of differentiating a faulty acceleration signal (a) by a conventional SGD ($K = 3$, $F = 39$) in (b–c) and a LBSGD (BW = 1280 Hz) in (d–e). The SGD enhances the entry/exit peaks but with the presence of the high frequency noise at the exit which is removed by the LBSGD.

Figure 5, a SGD provides two signal enhancements: a rejection of high frequency noise (G2 and higher frequencies) similar to a moving average filter shifted to the middle of BW part, and an ideal differentiation (G1) for lower frequencies.

2.1.4. A low pass SG differentiator. The signal being differentiated is subjected to a large amplification gain (Figure 5(b)) which is linearly affected by the sampling frequency f_s:

$$\frac{dy}{dt} = f_s[y(t) - y(t - dt)] \tag{5}$$

The differentiator gain is focused in the BW part; however, there is also a significant gain at higher frequencies, which disturbs the smoothness of the output signal (Orfanidis, 1995; Schafer, 2011).

To enhance the ability of SGD to reject noisy high-frequency components, we propose to use a modified SGD with a narrow BW, referred as limited BW SGD (LBSGD). A SGD defined by (F, K) is cascaded in series with a digital low-pass filter to eliminate all frequencies beyond the BW limit, as shown in Figure 6. An effective realization of this filter can be achieved using an FFT-based low-pass filter (LP-FFT), which has an approximately ideal low pass filtration (Ben-Ezra, 2009). The required cut-off frequency for the LP-FFT filter is set to be equal to the previous definition of the SGD BW.

The enhanced performance of a LBSGD compared to conventional SGD is depicted in Figure 7. The parameters (F, K) for this comparison were selected based on the automatic tuning of LBSGD which will be discussed in the next section. The selection of (F, K) recognizes a LBSGD of BW = 1280 Hz. As shown in Figure 7(b) to (d), both SGD and LBSGD enhance the detection of the entry/exit points, i.e. highest two peaks; however, the LBSGD provides additional smoothing to the exit point as a result of clearing

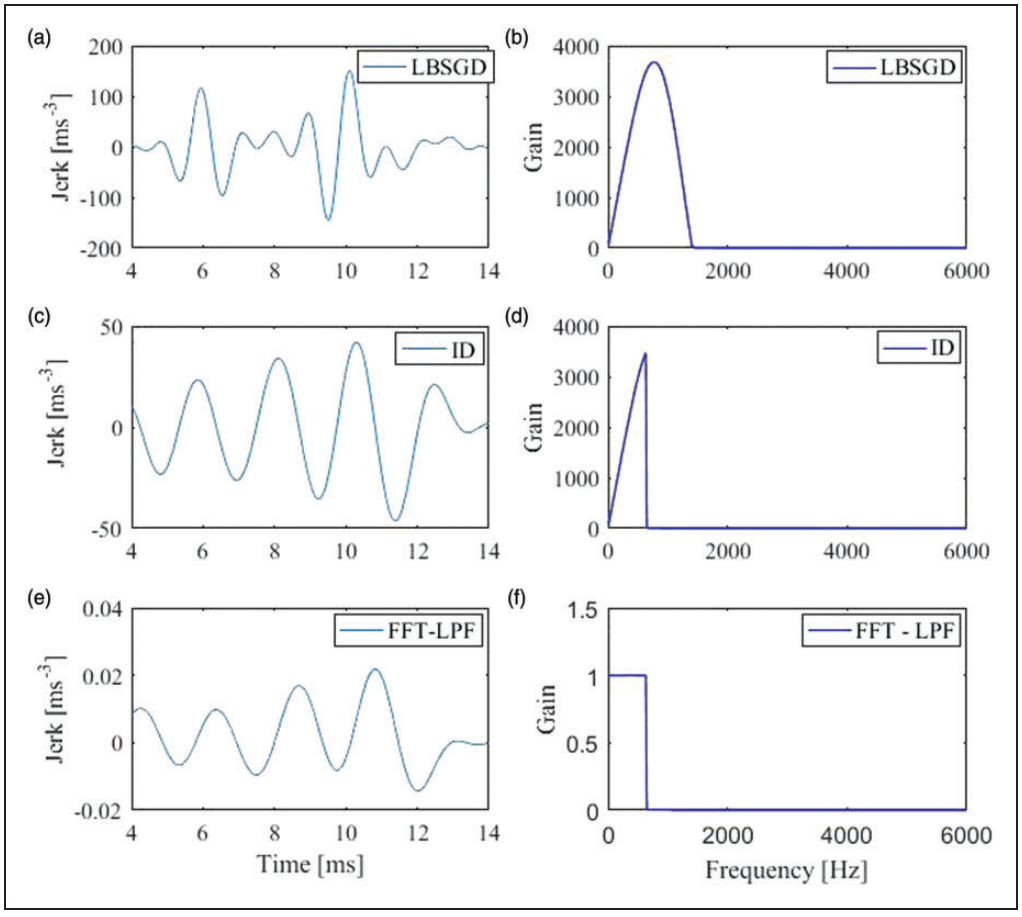

Figure 8. A continued comparison of Figure 7 is shown, including time response of LBSGD in (a) and corresponding frequency response in (b), where BW = 1280. An ideal differentiator filtered to 640 Hz (i.e. to include same differentiator band of LBSGD) is shown in (c–d). A low pass filter with a cut off frequency of 640 Hz is shown in (e–f). Both techniques provide enhancements that merge entry/exit events.

high frequencies. The location of the entry/exit can be performed by peaks detection algorithm.

Another comparison is depicted in Figure 8 by using an ID and a LPF. The ID is estimated by applying a LPF of 640 Hz (i.e. 1280/2) to LBSGD in Figure 7(c) in order to clear the integrator part. The LPF has a sharp cut-off frequency of 640 Hz in order to have common frequency band of LBSGD, ID and LPF. In this case, both of ID and LPF have one type of signal enhancement to whole signal, i.e. the ID gain is proportional to the frequency and the LPF gain is constant independent to signal frequency (i.e. below cut off frequency which equals 640 Hz). For this kind of signals, the short-time feature (not a real periodic signal) of the entry/exit events is merged over the signal, as shown in Figure 8. For example, Figure 8(c) shows the ID signal which has increasing amplitude because of the frequency jump between entry and exit events, i.e. low to high differentiator gains as a result of low to high frequency of the entry and exit respectively. Both ID and LPF provide enhancements to the whole signal rather than local entry/exit events.

2.2. Automated selection of the differentiator parameters

This process is performed offline for a short time acceleration measurement which is corresponding to one or few revolutions of the bearing races. Suppose that five balls roll over an outer-race spall (Figure 9(a)). Five vibration spikes are excited, as observed in the acceleration signal shown in Figure 9(b). We suggest using a number of differentiation operations, denoted by $\emptyset = [\emptyset_1.\emptyset_1.\cdots\emptyset_Q]$. Each of them is a LBSGD described by a pair of (K, F) parameters.

For every spike, we assume that the elapsed time between the two highest peaks (i.e. the highest rates of change) approximates the actual entry-exit separation time, as introduced in Figure 2. As shown in Figure 9(c), as a result of applying an operation $\emptyset_i$, a series of peak pairs (EP1, EP2) is extracted. EP1 denotes the highest peak in the entry interval, whereas EP2 denotes the highest peak in the exit interval. The elapsed time within each pair is then assigned to each spike.

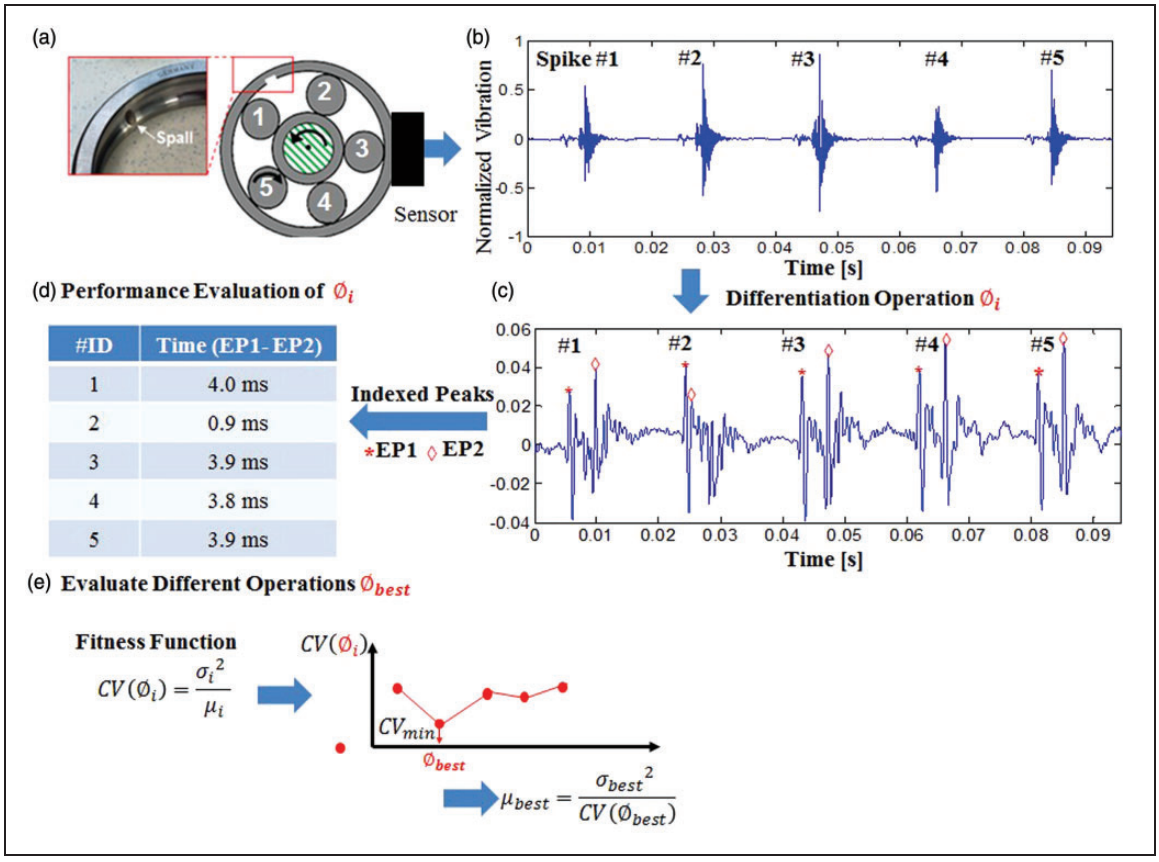

Figure 9. An illustration of LBSGD tuning. An outer-race spall (a) generates five vibration spikes (b) that are subjected to a number of differentiation operations (c). The evaluation of each operation is based on the similarity of the two highest peaks (EP1, EP2) identified within each spike in (c). This similarity is checked by calculating the distribution of the elapsed times between EP1 and EP2 (d). A fitness value is assigned to each differentiation operation using the CV measure, as shown in (e).

The similarity of spike excitation is based on the fact that a vibration signal measured under constant-speed conditions will exhibit repeated similar spall excitations. These excitations may have variable peak magnitudes because of the variable loading profile around the race, but the time separation between them is influenced mainly by the spall width and the speed. Each excitation is induced by the passage of a ball over the same spall. As the speed rate is kept constant for each dataset, the elapsed times between the entry and exit points for all excitations should have similar values. A simple way to visualize this similarity is to calculate a statistical similarity measure of the elapsed times for a group of spall spikes after the application of a differentiation operation, as shown in Figure 9(d). The use of five spikes here is for illustration only; the size of the spike group is related to available data size. For electromechanical actuators, the target application here, the bearing almost turns one or very few revolutions rather than hundreds or thousands revolutions for industrial rotating machinery, e.g. pumps,

We empirically found that the coefficient of variations (CV) can be used to effectively assess the similarity of the excitations. The CV is the variance normalized by the sample mean:

$$CV(\emptyset_i) = \frac{\sigma^2}{\mu} \tag{6}$$

Here, the final evaluation of different differentiation operations is performed based on minimizing the CV, as shown in Figure 9(e).

3. Results and discussion

3.1. Test rig and seeded faults

In this work, seeded faults in the bearings under test (four-point FAG QJ212TVP aerospace bearings, containing 15 balls of 15.87 mm in diameter and with a 85.15 mm pitch diameter) were created through spark erosion. This formed an oval fault geometry defined by the fault depth (fd) and two diameters, namely, the minimum diameter (D1), which represents the spall width, and the maximum diameter (D2), which is perpendicular to the rolling track direction, as shown in Figure 10. A special bearing test rig was used to measure tri-axial vibrations through three accelerometers (model PCB 356A32) and sampling frequency of 25.6 kHz. However, one of the radial vibration measurements is sufficient for the proposed method as the other measurements have very similar signals.

Vibration datasets were acquired for six faults (3 outer-race faults and 3 inner-race faults), described in Table 2, which were chosen to approximate the growth of a spall in three different levels. The bearing under test was fitted vertically in the test rig, and a vertical axial load was applied to the inner race to emulate loading/speed conditions similar to those in linear electro-mechanical actuators. Datasets were collected and processed offline for each fault under four operating conditions as follows: two speeds (500 and 60 r/min) and two axial loads (5 and 8.8 kN).

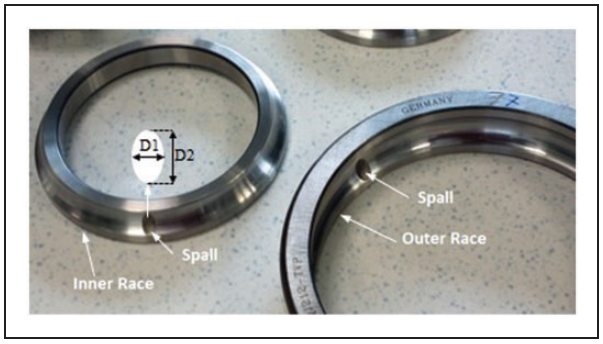

Figure 10. A photograph of a seeded inner-race spall and an outer-race spall.

Table 2. Descriptions of the seeded faults in terms of spall depth (fd), spall width (D1), and spall length (D2).

Fault code	fd (mm)	D1 (mm)	D2 (mm)	Fault type
Inner_1	0.05	1.0	2.6	Inner
Inner_2	0.15	2.1	4.4	Inner
Inner_3	0.40	3.8	6.8	Inner
Outer_1	0.05	1.4	2.6	Outer
Outer_2	0.15	2.4	4.4	Outer
Outer_3	0.40	4.0	6.8	Outer

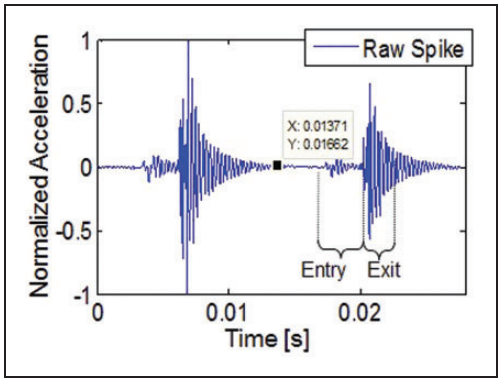

Figure 11. Two spikes from Inner_3 case (Table 2) measured at 500 r/min and 5 kN. The entry and exit intervals both contain several peaks and specific entry/exit points cannot be located.

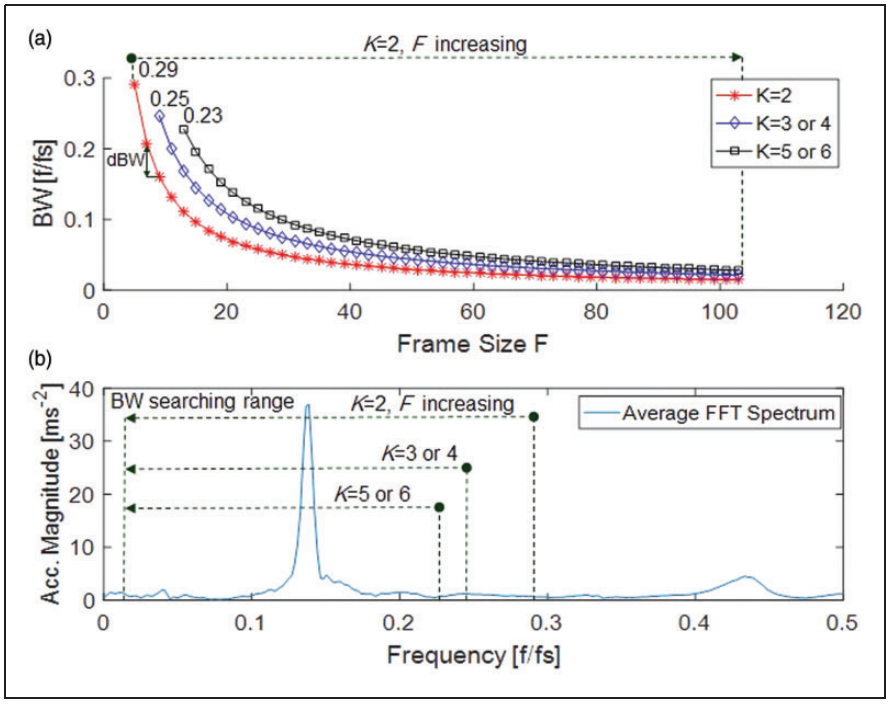

Figure 12. (a) Different BW searching ranges of LBSGD for different (*K*, *F*). (b) The average FFT spectrum of a spike is depicted which involves corresponding scanning ranges for each *K*.

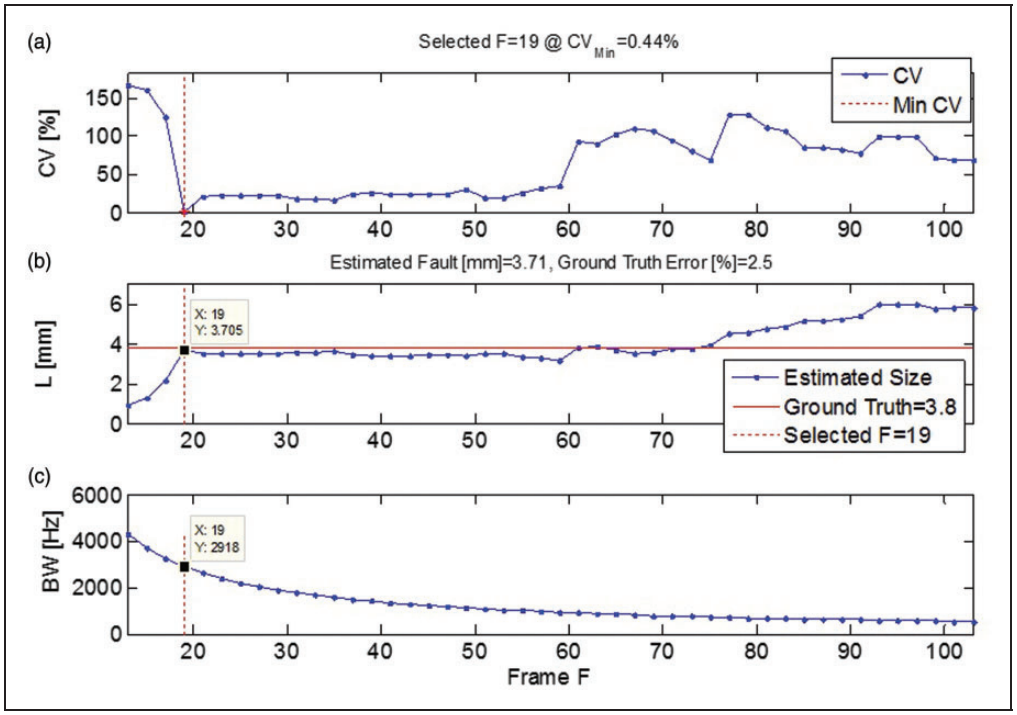

Figure 13. (a) The process of automatically searching for the minimum CV value with $K = 3$ and a varying frame F is shown. The minimum value of CV = 0.44% corresponds to $F = 19$ and an average entry-exit distance of 3.71 mm in (b), which represents an error of 2.5%. (c) The corresponding BW at $F = 19$.

3.2. Detailed quantification procedure

This example is for the dataset of Inner_3 case (Table 2) and measured at 500 r/min and 5 kN. The data was trimmed to 12 spikes (one turn scenario) and normalized with respect to the max peaks of the spikes within the dataset. The magnitude of the signal is not used here because we assume that the load in unknown. Keeping it within a range of ± 1 is useful for comparing different peaks of each spike. This example (Figure 11) represents vibrational spikes similar to the step-impulse case depicted in Figure 1(a).

The quantification procedure follows the steps presented in Figure 9. The vibration spikes are separated and differentiated using a bank of LBSGDs described by their frame size F and polynomial order K.

Practically, there are two possible risks that should be studied before and after tuning the differentiator. The first is the proper selection ranges for F and K, prior to the differentiator tuning, in order to be compatible with resonant frequencies of the signal. The F and K values may produce differentiators with BWs that cannot achieve required enhancements to entry/exit intervals. The second is to perform a numerical test to make sure that the elapsed time between entry/exit peaks does match one of the resonant frequencies of the test rig. This test involves a comparison between final BWs of tuned LBSGD and known resonant frequencies of the test rig. For all datasets used in this paper, the BW does not indicate resonant frequencies.

3.2.1. Selection of differentiator parameter limits. The first step is to determine the K value. As shown in Figure 12(a), at $K=2$ (i.e. the minimum K value which is corresponding to second order polynomial), the corresponding initial value of F, equals $F_{\min} = 2K+1 = 5$. The evolution of the BW over the filter window size in Figure 12(a) is identical for K and $K+1$ starting from $K=3$, a detailed mathematical derivation can be found in (Orfanidis, 1995). Increasing K is subjected to two effects: the first is a decrease in the BW searching range as shown in Figure 12(b); wider searching range is preferable in order to cover the frequency spectrum of the signal. The second effect is related to the BW searching resolution, the dBW in Figure 12(a), which determines the BW incremental step of F. The dBW should be small enough to include narrow changes in the spectrum. We found that setting $K=3$ or 4 accelerates the computation, instead of setting $K=2$, and does not result a significant deviation to the dBW.

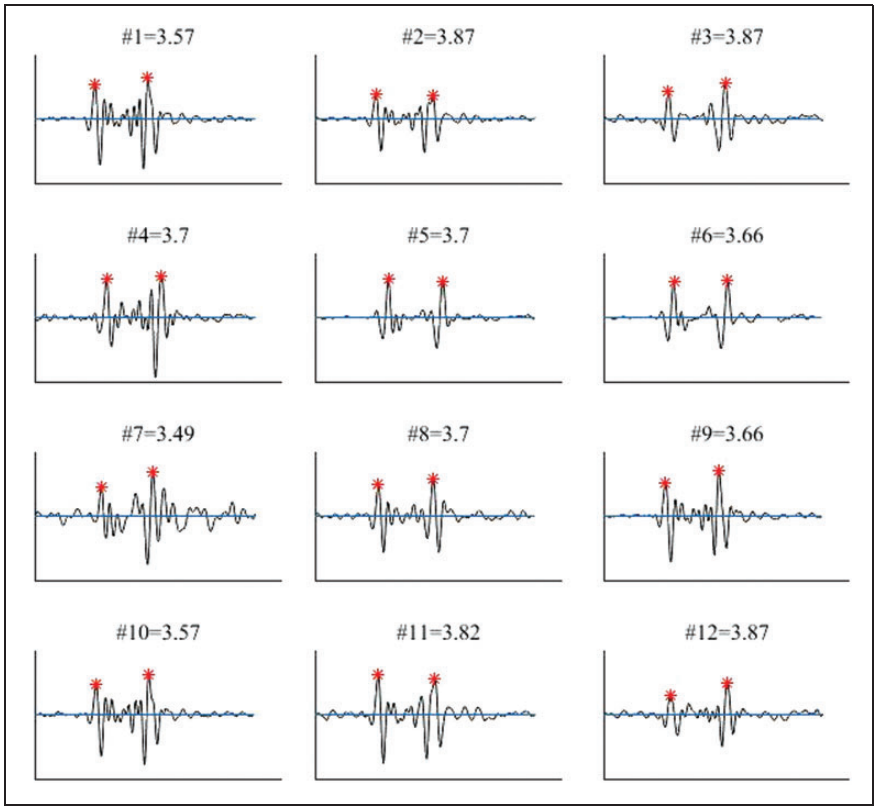

Figure 14. The differentiated signal at $F=19$ is shown. The entry-exit separations (in mm) for the shown spikes are very similar to each other and also close to the actual width of 3.8 mm.

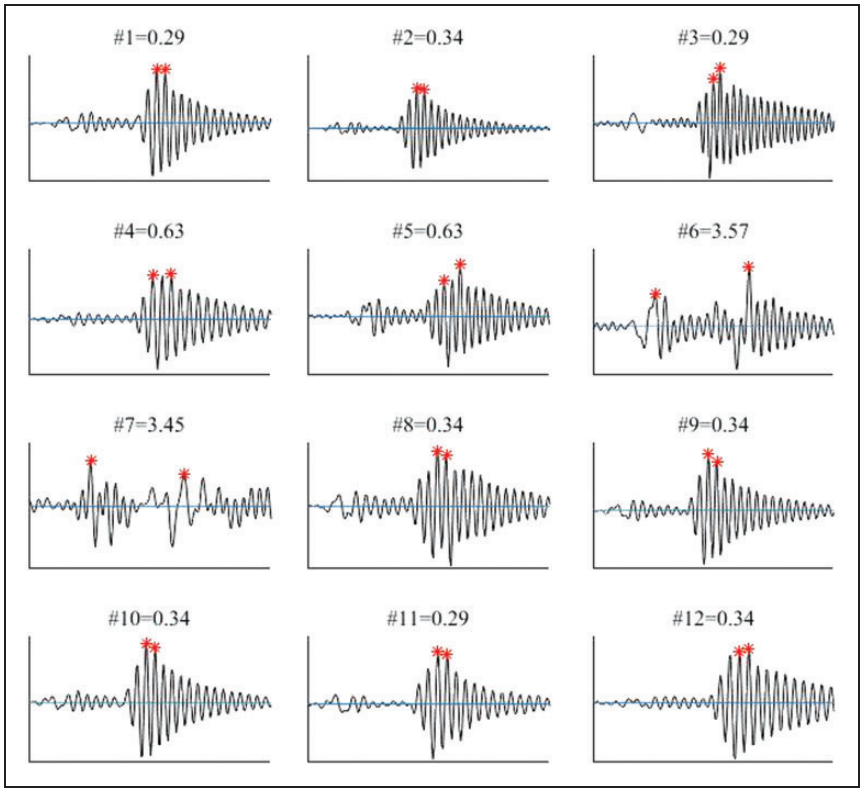

Figure 15. The differentiated signal obtained at an arbitrary $F = 13$ is shown. The entry–exit distances (in mm) for the shown spikes are not similar to each other or to the real value of 3.8 mm.

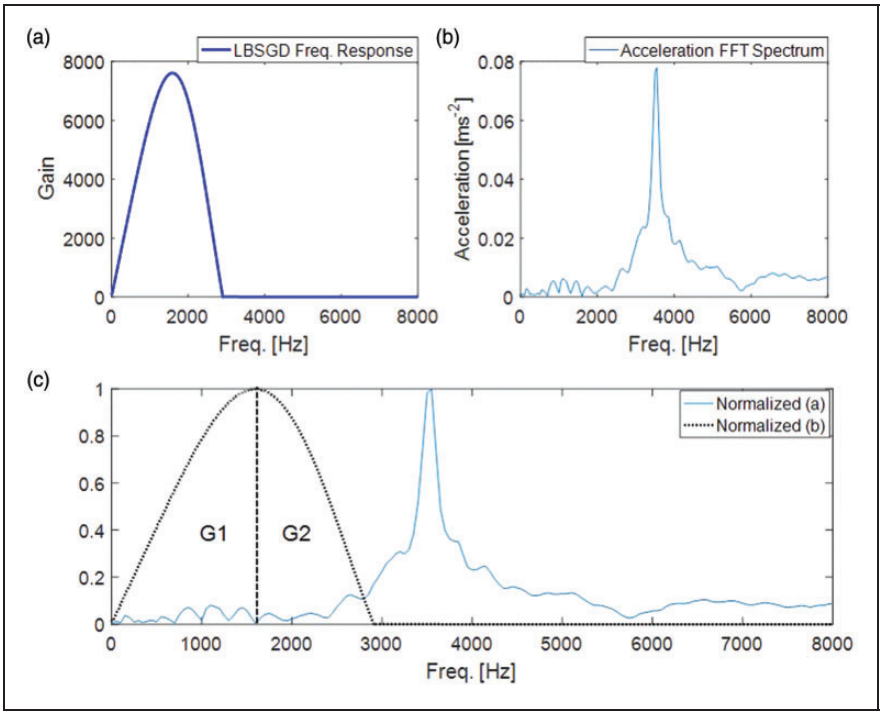

Figure 16. (a) The frequency response of the LBSGD with the optimal parameters $F = 19$ and $K = 3$ is shown and compared to the spectrum of an acceleration spike in (b). The graph (c) involves normalized versions of (a) and (b) in order to locate different differentiator effects G1 and G2.

Figure 12(b) shows the average FFT spectrum of an acceleration spike. The term "average spectrum" is used because the accurate frequency content of the entry and the exit intervals are hard to be separated due to three reasons: (a) their specific intervals are hard locate, as illustrated in Figures 2; (b) the spikes from one turn are not identical in their spectrums since they are subjected to non-identical loading profile; c) the number of samples for entry/exit events are limited by comparing to sampling rate, i.e. about 100 samples for each spike (500 r/min) sampled at 25.6 kHz.

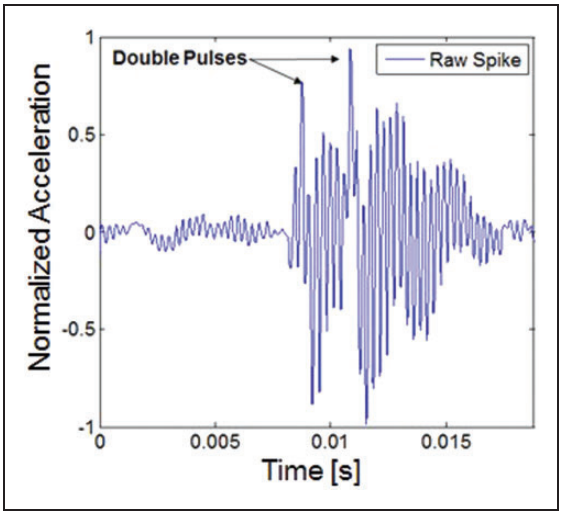

Figure 17. A spike from Outer_2, 500 r/min and 5 kN from example 2 depicted, which indicates the possible double pulse.

Figure 12(b) shows also the searching frequency values for LBSGDs with various combinations of (F, K). With the differentiation order begins by $K=3$ or 4 and the initial frame length set to $F=2K+1=$ 7 or 9, the starting frequency will be 0.25% of sampling frequency.

In accordance with this preliminary investigation, the vibration signal is differentiated using numerous LBSGDs with starting parameters of ($F=9$, $K=3$) and an incremental increase in F, K until either a small CV value (<1%) is achieved or the "no excitation band" is reached. In this dead band, no resonant frequencies are available to be excited by the fault. The excitation dead band can be identified based on observing several average spectrums, e.g. Figure 12(b), for individual spikes to locate the excitation dead band. We experimentally found that this band starts from 2% of the sampling rate or 550 Hz, which corresponds to $F=103$. Consequently, the candidate differentiation parameters are ($F=9$–103, $K=3$–6).

Figure 13 shows the results of applying different LBSGDs with increasing window size. Figure 13(a) indicates different coefficients of variance (CVs) associated with different frames F. According to the CV fitness criterion (discussed in Section 2), the best quantification result corresponds to the most consistent

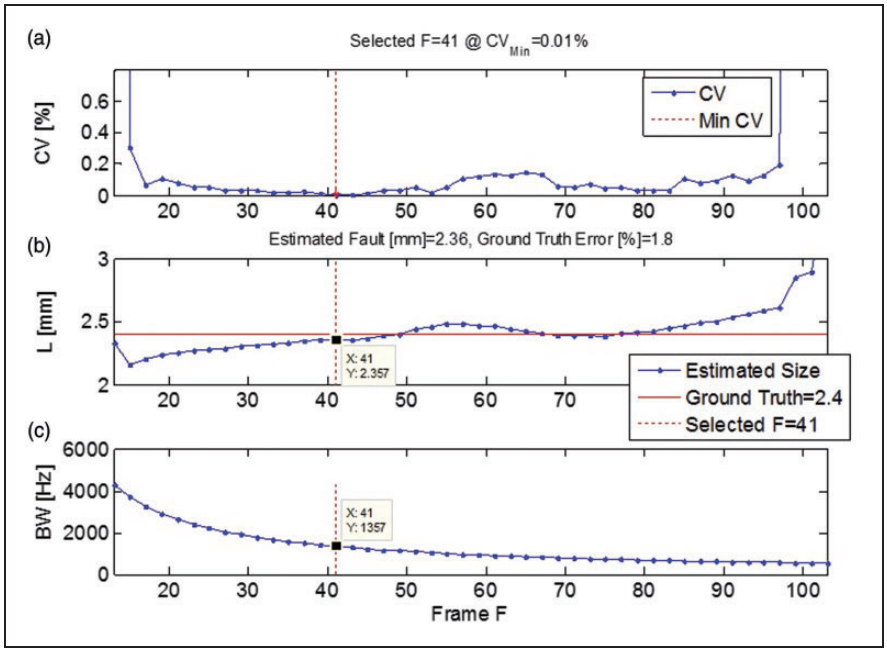

Figure 18. Plot (a) shows the automatic search for the minimum CV value with a varying F. The minimum value of CV = 0.01% corresponds to $F=41$ and an average entry–exit distance of 2.36 mm in (b), which represents an error of 1.8%.(c) shows the corresponding BW at $F=41$.

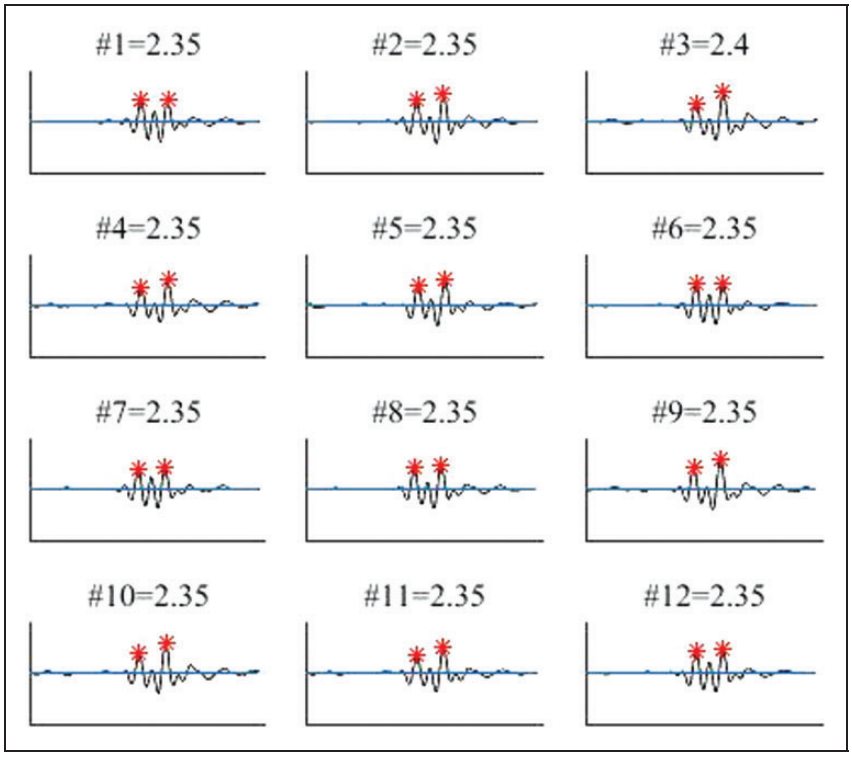

Figure 19. The differentiated signal at $F = 41$ is shown. The entry-exit separations (in mm) for the shown spikes are very similar to each other and also close to the actual width of 2.4 mm.

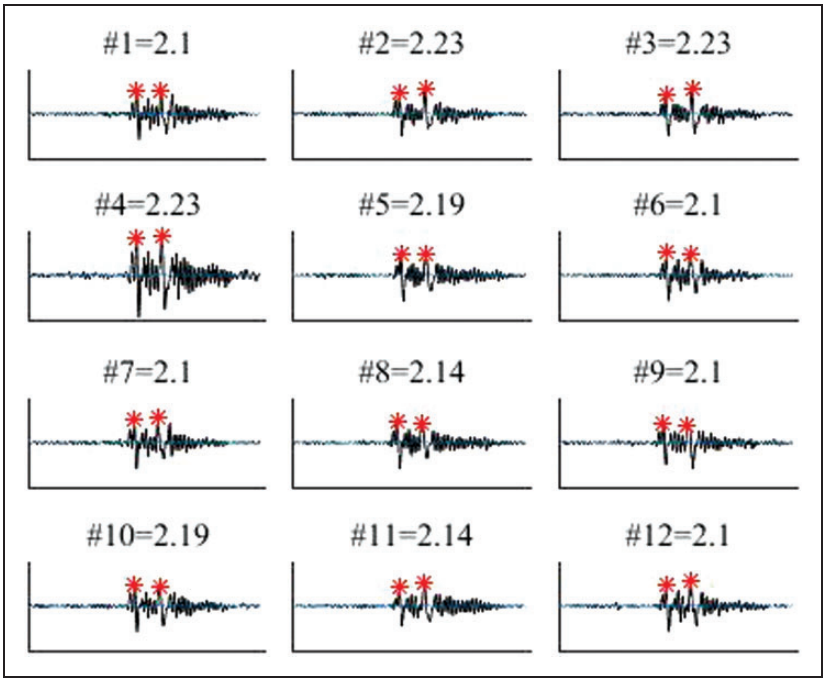

Figure 20. The differentiated signal obtained at an arbitrary $F = 15$ is shown. The entry-exit distances (in mm) for the shown spikes have the average of 2.15 mm; while real size is 2.4 mm.

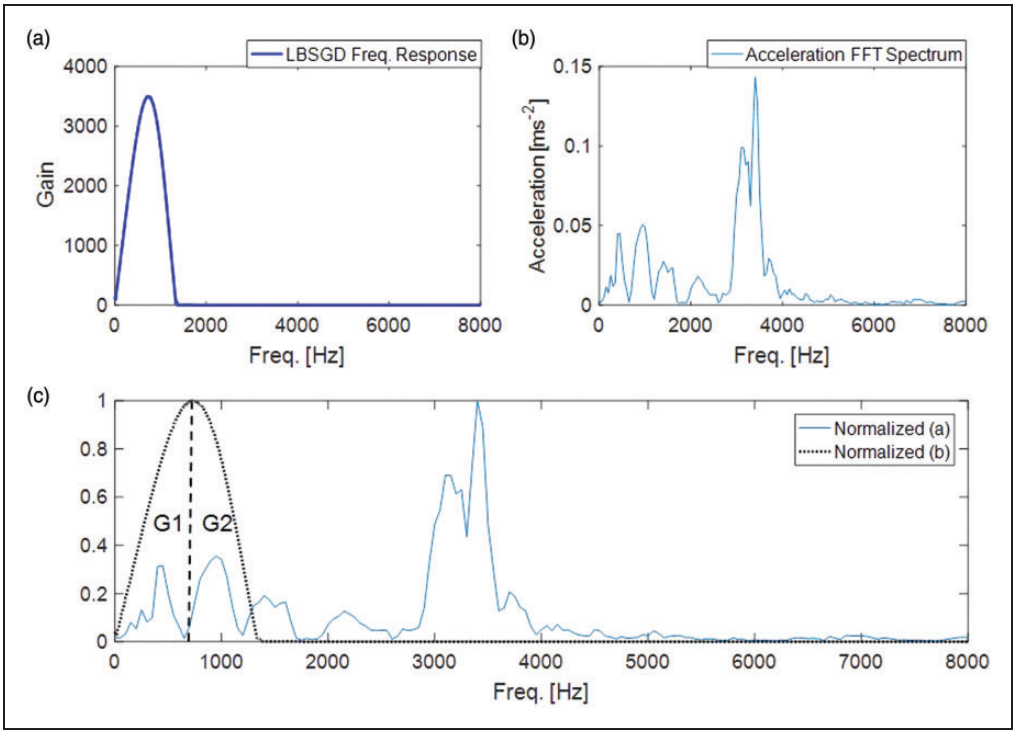

Figure 21. The frequency response of a LBSGD differentiator with the optimal parameters $F = 41$ and $K = 3$ is shown in (a) and compared with the spectrum of a spike from the same dataset in (b). The graph (c) involves normalized versions of (a) and (b) in order to locate different differentiator effects G1 and G2.The peaks in (b) poorly indicate actual resonant frequencies of the test rig because very short length of samples for entry/exit intervals, i.e. about 100 samples (within a spike) measured in 25.6 kHz rate.

Table 3. The fault estimations for the 24 DLR datasets, where N is the speed in r/min, P is the load in kN, W is the spall width, and CV is the minimum coefficient of variance in % achieved for the selection of the SG differentiator.

ID	Speed (r/min)	Load (kN)	Spall width (mm)	Error (mm)	CV_{min}	BW (Hz)	F (Samples)	Observation: double pulse (1) or step impulse (2)
1	500	5	1.0	0.14	0.10	2048	27	1
2	500	5	2.1	0.1	9.30	1792	31	1
3	500	5	3.8	0.09	0.44	2918	19	2
4	500	5	1.4	0.18	0.00	1357	41	1
5	500	5	2.4	0.04	0.00	1357	41	1
6	500	5	4.0	0.12	0.40	627	89	2
7	60	5	1.0	0.2	0.00	550	103	1
8	60	5	2.1	0.14	0.00	589	95	1
9	60	5	3.8	0.05	0.10	550	103	1
10	60	5	1.4	0.1	0.00	1498	37	1
11	60	5	2.4	0.08	0.00	550	103	1
12	60	5	4.0	0.1	0.10	1498	37	1
13	500	8.8	1.0	0.01	2.70	1677	33	1
14	500	8.8	2.1	0.01	11.3	3712	15	1
15	500	8.8	3.8	0.3	9.70	2214	25	2
16	500	8.8	1.4	0.18	0.10	2214	25	2
17	500	8.8	2.4	0.23	0.00	2214	25	1
18	500	8.8	4.0	0.1	4.10	832	67	2

(continued)

Table 3. Continued

ID	Speed (r/min)	Load (kN)	Spall width (mm)	Error (mm)	CV_{min}	BW (Hz)	*F* (Samples)	Observation: double pulse (1) or step impulse (2)
19	60	8.8	1.0	0.15	0.00	550	103	1
20	60	8.8	2.1	0.9	11.2	973	57	1
21	60	8.8	3.8	0.48	17.3	973	57	1
22	60	8.8	1.4	0.21	27.9	550	103	1
23	60	8.8	2.4	1.21	36.0	730	77	1
24	60	8.8	4.0	0.77	0.40	550	103	1

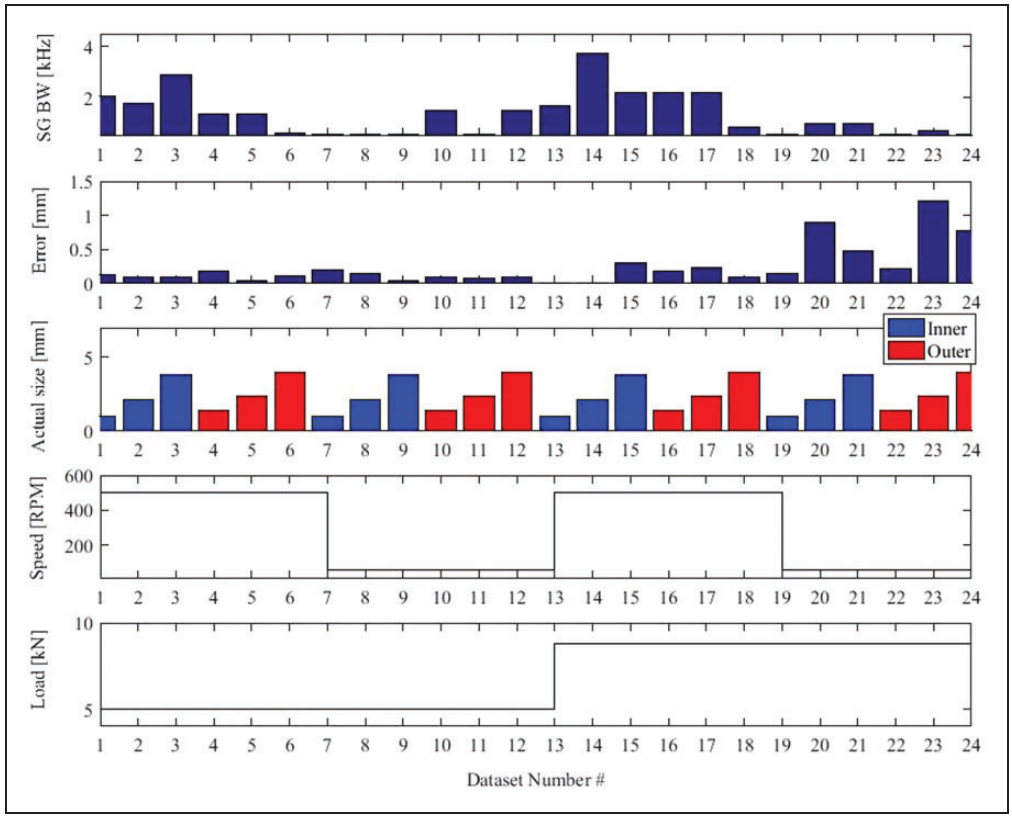

Figure 22. The quantification error versus the actual spall size and operating speed/load for the 24 datasets are depicted. The worst errors are observed for the low-speed and high-load datasets, cases #19–24. The differentiator bandwidth (BW) versus the operating speed and fault size is also shown. There is a strong correlation between the differentiation BW and the speed, whereas the load and spall size contribute minor effects.

distribution of the entry/exit intervals, which can be identified from the minimum in the CV values. In the presented case, the minimum CV is 0.44%, which is achieved for a LBSGD with $F = 19$. Figure 13(b) shows the average spall width L as a function of the frame F. At a CV value of 0.44%, the LBSGD produces a rate-of-change which exhibits peak pairs. The time separation indicates a special width with a mean value of 3.71 mm based on equation (6). This corresponds to 2.5% error with respect to the ground-truth value of the spall. Figure 13(c) shows the corresponding BW of the selected differentiator; at $CV = 0.44\%$, the BW is equal to 2.9 kHz, which includes balanced entry/exit excitations for identifying the fault size. The individual differentiation for signal spikes in the minimum CV of $F = 19$, is depicted in Figure 14, and a worse case is depicted in Figure 15.

The frequency response of tuned LBSGD with the parameters can be seen in Figure 16(a), while Figure 16(b) shows the frequency spectrum of an

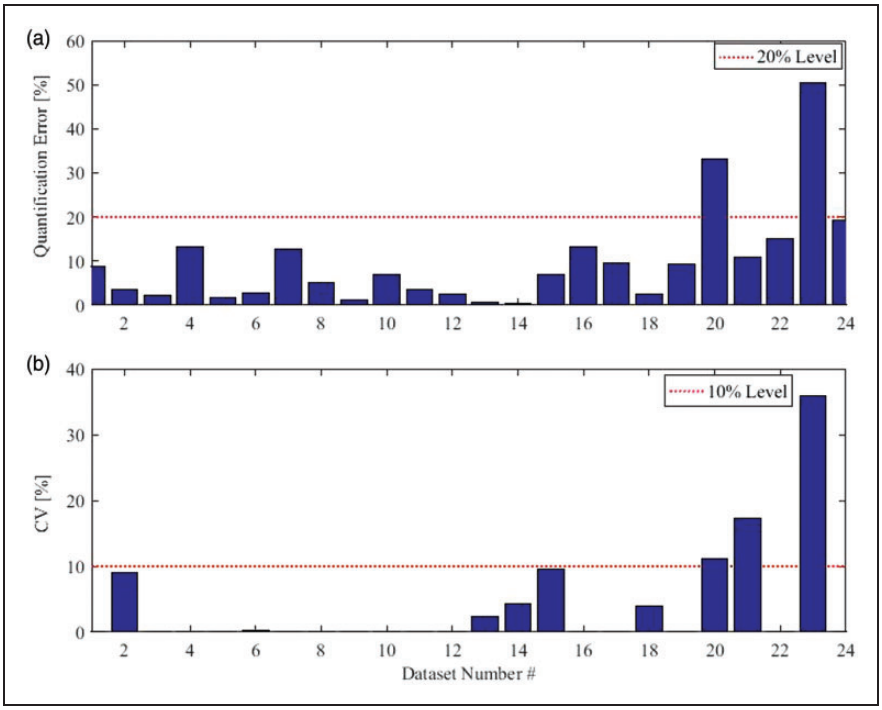

Figure 23. A measure of the level of accuracy achieved in spall quantification is proposed. It is based on comparing the percentage error in the spall-size estimate in (a) with the value of the CV in (b). As the CV increases, the accuracy of the estimated spall width degrades.

acceleration spike. In order to show how the differentiated signal in Figure 14 is calculated, normalized versions of Figure 16(a) and (b) are plotted in Figure 16(c). Then, it is observed that two frequency bands are subjected to a high differentiator gain (ID part), G1 and a low gain G2 (in the LPF part). G1 band is representable to the entry interval and G2 is more representable to the exit interval. The G1 band, which is a low frequency band, has a very low contribution in the acceleration signal; while, tuned LBSGD amplifies it to be balanced with G2 band, as investigated in Figure 16(a).

A second quantification example is depicted in Figure 17, for the dataset corresponding to Outer_2 (Table 2), 500 r/min and 5 kN. This example illustrates vibrational spikes similar to the double-pulse case depicted in Figure 1(a). Figure 18 provides a summary of the search for the optimal filter parameters, which yields a frame $F=41$.

The individual differentiation for signal spikes in the minimum CV of $F=41$, is depicted in Figure 19, and a random case of $F=15$ is depicted in Figure 20.

The frequency response of the differentiator with the optimal parameters can be seen in Figure 21. In this case, i.e. double pulse, both the entry and the exit have similar frequency content, in which tuned LBSGD provides very similar differentiator gains for the entry and the exit. This can be observed on Figure 21(c), both G1 and G2 are very close rather than the step-impulse case in Figure 16(a).

3.3. LBSGD performance for a variety of speeds, loads, and spall sizes

Here, in Table 3, we summarize the quantification results for a variety of spall sizes under two loads (5 and 8.8 kN) and two speed conditions (500 and 60 r/min), as listed in Table 2. Figure 22 provides a summary of the estimation error (E) and the BW under various conditions. The fault estimation error is below 0.25 mm, except for the larger spalls at the higher load as well as lower speed. This error can be related to balls motion which may involve irregular rolling/sliding within large spall zone. The selected BW is significantly influenced by rotation speed, i.e. at the higher speed, the optimal BW is higher than that at the slower speed, with few exceptions. The load and the size of the spall appear to exert little effect on BW values. Figure 23 shows a relation between the quantification error and the corresponding CV value which has a very large value (>10%) for cases of errors larger than 20%.

4. Conclusion

Two vibration responses have been observed for spalled bearings: a double pulse and a step-impulse. At low speed (60 r/min) and for all spall-sizes, the ball rolls over the spall causing a double pulse as a result of destressing (close to the entry) and restressing (close to the exit) events. At higher speed (500 r/min) and

for large spalls, the ball destresses similarly close to the entry but as it has higher kinetic energy it bridges over large spall causing an impact close to the exit edge (impulse). Small and medium spalls measured at 500 r/min almost induce double pulse response because the ball movement within spall zone is restricted. In addition, a fault quantification technique has been presented, which is applicable for both double pulse and step- impulse responses. The quantification begins with the calculation of the jerk of the faulty acceleration data. The jerk is calculated using a LBSGD. The frequency response of the LBSGD consists of an ID part and a LPF part. The LBSGD parameters are tuned to locate both the ID and the LPF within the frequency contents of the entry and exit parts respectively. The tuning is based on the principle of simultaneous enhancement of the entry and exit events by subjecting them to different frequency gains. The LBSGD increases the signal to noise contrast for entry/exit peaks making them more extractable. Among the 24 investigated cases, the technique successfully estimated the fault width (in millimeters) in 22 cases with a maximum error of 20%. Higher errors were reported in the other two remaining cases. Quantification errors can be predicted based on a statistical measure, the coefficient of variation (CV), for the detected entry/exit peaks. The CV has a large value (>10%) for all cases of errors larger than 20%. The performance of the presented technique at higher speed conditions is planned as a future work.

Acknowledgements

We thank the staff of TEKNIKER (IK4, Intelligent Information Systems Unit, Spain) and our colleague Ms. Thu-Hien Pham for their collaboration in preparing the database. We also appreciate valuable comments from Mr. Johann Dauer.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Bakshi UA and Godse AP (2008) *Linear Integrated Circuits & Applications*. Pune, India: Technical Publications Pune.

Ben-Ezra S (2009) FFT filter – clean your signals and display results! Available at: http://se.mathworks.com/matlabcentral/fileexchange/25017-fft-filter-clean-your-signals-and-display-results (accessed 10 October 2015).

Epps I and McCallion H (1994) An investigation into the characteristics of vibration excited by discrete faults in rolling element bearings. In: *Annual conference of the vibration association of New Zealand*, Christchurch, New Zealand, 9–11 May 1994.

Ismail M, Windelberg J, Bierig A, et al. (2016) A potential study of prognostic-based maintenance for primary flight control electro-mechanical actuators. In: *7th conference on recent advances in aerospace actuation systems and components*, Toulouse, France, 16–18 March 2016.

Ismail MA, Sawalhi N and Pham T-H (2015) Quantifying bearing fault severity using time synchronous averaging jerk energy. In: *22nd international congress on sound and vibration*, Florence, Italy, 12–16 July 2015..

Jena D, Singh M and Kumar R (2012) Radial ball bearing inner race defect width measurement using analytical wavelet transform of acoustic and vibration signal. *Measurement Science Review* 12(4): 141–148.

Kogan G, Bortman J and Klein R (2015) Estimation of the spall size in a rolling element bearing. *Insight-Non-Destructive Testing and Condition Monitoring* 57(8): 448–451.

McFadden PD and Smith JD (1984) Vibration monitoring of rolling element bearings by the high-frequency resonance technique: a review. *Tribology International* 17(1): 3–10.

Moustafa W, Cousinard O, Bolaers F, et al. (2016) Low speed bearings fault detection and size estimation using instantaneous angular speed. *Journal of Vibration and Control* 22(15): 3413–3425.

Orfanidis SJ (1995) *Introduction to Signal Processing*. Upper Saddle River, NJ: Prentice Hall.

Randall RB (2011a) *Vibration-Based Condition Monitoring: Industrial, Aerospace and Automotive Applications*. London, UK: Wiley.

Randall RB (2011b) The challenge of prognostics of rolling element bearings. In: *Wind turbine condition monitoring workshop*, Broomfield, CO, 19–21 September 2011.

Savitzky A and Golay MJ (1964) Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* 36(8): 1627–1639.

Sawalhi N (2007) *Diagnostics, prognostics and fault simulation for rolling element bearings*. PhD Thesis, The University of New South Wales, Kensington, Australia.

Sawalhi N and Randall RB (2008) Simulating gear and bearing interactions in the presence of faults. Part I. The combined gear bearing dynamic model and the simulation of localised bearing faults. *Mechanical Systems and Signal Processing* 22(8): 1924–1951.

Sawalhi N and Randall RB (2011) Vibration response of spalled rolling element bearings: observations, simulations and signal processing techniques to track the spall size. *Mechanical Systems and Signal Processing* 25(3): 846–870.

Schafer RW (2011) What is a Savitzky–Golay filter? *IEEE Signal Processing Magazine* 28: 111–117.

Vachtsevanos G, Lewis FL, Roemer M, et al. (2006) *Intelligent Fault Diagnosis and Prognosis for Engineering Systems*. New York: Wiley.