Mechanical Systems and Signal Processing 25 (2011) 485–520

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/jnlabr/ymssp

Review

Rolling element bearing diagnostics—A tutorial ☆

Robert B. Randall [a,*], Jérôme Antoni [b]

[a] School of Mechanical and Manufacturing Engineering, University of New South Wales, Sydney, NSW 2052, Australia
[b] Laboratory Roberval of Mechanics, University of Technology of Compiégne, 60205 Compiégne, Cedex, France

ARTICLE INFO

Article history:
Received 23 July 2010
Accepted 29 July 2010

Keywords:
Rolling element bearings
Diagnostics
Cyclostationarity
Spectral kurtosis
Minimum entropy deconvolution
Envelope analysis

ABSTRACT

This tutorial is intended to guide the reader in the diagnostic analysis of acceleration signals from rolling element bearings, in particular in the presence of strong masking signals from other machine components such as gears. Rather than being a review of all the current literature on bearing diagnostics, its purpose is to explain the background for a very powerful procedure which is successful in the majority of cases. The latter contention is illustrated by the application to a number of very different case histories, from very low speed to very high speed machines. The specific characteristics of rolling element bearing signals are explained in great detail, in particular the fact that they are not periodic, but stochastic, a fact which allows them to be separated from deterministic signals such as from gears. They can be modelled as cyclostationary for some purposes, but are in fact not strictly cyclostationary (at least for localised defects) so the term pseudo-cyclostationary has been coined. An appendix on cyclostationarity is included. A number of techniques are described for the separation, of which the discrete/random separation (DRS) method is usually most efficient. This sometimes requires the effects of small speed fluctuations to be removed in advance, which can be achieved by order tracking, and so this topic is also amplified in an appendix. Signals from localised faults in bearings are impulsive, at least at the source, so techniques are described to identify the frequency bands in which this impulsivity is most marked, using spectral kurtosis. For very high speed bearings, the impulse responses elicited by the sharp impacts in the bearings may have a comparable length to their separation, and the minimum entropy deconvolution technique may be found useful to remove the smearing effects of the (unknown) transmission path. The final diagnosis is based on "envelope analysis" of the optimally filtered signal, but despite the fact that this technique has been used for 40 years in analogue form, the advantages of more recent digital implementations are explained.

Contents

☆ Some of the material in this tutorial is adapted from related sections in the book *Vibration-based Condition Monitoring: Industrial, Automotive and Aerospace Applications*, by R.B. Randall, to be published by John Wiley and Sons.

* Corresponding author. Tel.: +61 2 9958 3591; fax: +61 2 9663 1222.
E-mail addresses: b.randall@unsw.edu.au (R.B. Randall), jerome.antoni@utc.fr (J. Antoni).

0888-3270/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ymssp.2010.07.017

1. Introduction

Rolling element bearings are one of the most widely used elements in machines and their failure one of the most frequent reasons for machine breakdown. However, the vibration signals generated by faults in them have been widely studied, and very powerful diagnostic techniques are now available as discussed below.

Fig. 1 shows typical acceleration signals produced by localised faults in the various components of a rolling element bearing, and the corresponding envelope signals produced by amplitude demodulation. It will be shown that analysis of the envelope signals gives more diagnostic information than analysis of the raw signals. The diagram illustrates that as the rolling elements strike a local fault on the outer or inner race a shock is introduced that excites high frequency resonances of the whole structure between the bearing and the response transducer. The same happens when a fault on a rolling element strikes either the inner or outer race. As explained in [1], the series of broadband bursts excited by the shocks is further modulated in amplitude by two factors:

- The strength of the bursts depends on the load borne by the rolling element(s), and this is normally modulated by the rate at which the fault is passing through the load zone.
- Where the fault is moving, the transfer function of the transmission path varies with respect to the fixed positions of response transducers.

Fig. 1 illustrates typical modulation patterns for unidirectional (vertical) load on the bearing, at shaft speed for inner race faults, and cage speed for rolling element faults. The formulae for the various frequencies shown in Fig. 1 are as follows:

Ballpass frequency, outer race:

$$\mathrm{BPFO} = \frac{nf_r}{2}\left\{1-\frac{d}{D}\cos\phi\right\} \tag{1}$$

Ballpass frequency, inner race:

$$BPFI = \frac{nf_r}{2}\left\{1+\frac{d}{D}\cos\phi\right\} \tag{2}$$

Fundamental train frequency (cage speed):

$$FTF = \frac{f_r}{2}\left\{1-\frac{d}{D}\cos\phi\right\} \tag{3}$$

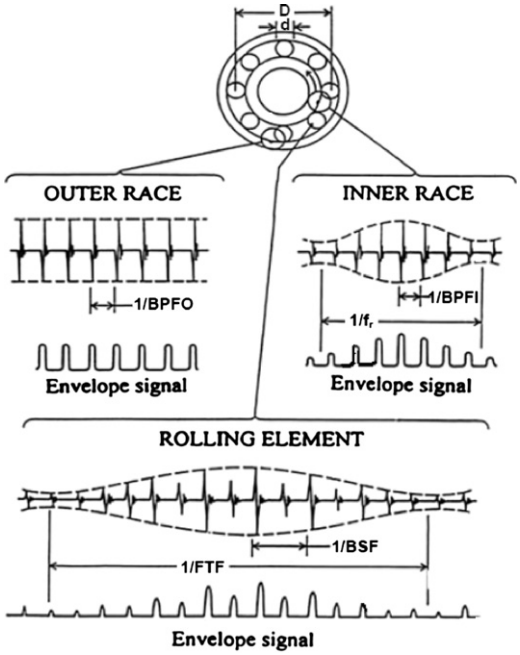

Fig. 1. Typical signals and envelope signals from local faults in rolling element bearings.

Ball (roller) spin frequency:

$$BSF(RSF) = \frac{D}{2d}\left\{1-\left(\frac{d}{D}\cos\phi\right)^2\right\} \tag{4}$$

where f_r is the shaft speed, n is the number of rolling elements, and ϕ is the angle of the load from the radial plane. Note that the ballspin frequency (BSF) is the frequency with which the fault strikes the same race (inner or outer), so that in general there are two shocks per basic period. Thus the even harmonics of BSF are often dominant, in particular in envelope spectra.

These are however the kinematic frequencies assuming no slip, and in actual fact there must virtually always be some slip because the angle ϕ varies with the position of each rolling element in the bearing, as the ratio of local radial to axial load changes. Thus, each rolling element has a different effective rolling diameter and is trying to roll at a different speed, but the cage limits the deviation of the rolling elements from their mean position, thus causing some random slip. The resulting change in bearing frequencies is typically of the order of 1–2%, both as a deviation from the calculated value and also as a random variation around the mean frequency. This random slip, while small, does give a fundamental change in the character of the signal, and is the reason why envelope analysis often extracts diagnostic information not available from frequency analyses of the raw signal. It means that bearing signals can be considered as cyclostationary (see Appendix A). This also allows bearing signals to be separated from gear signals with which they are often mixed, as discussed below.

It should be noted that the argument about variation of rolling diameter with load angle applies equally to taper roller and spherical roller bearings, since by virtue of their kinematics, the ratio of roller diameter to race diameter varies with the axial position, and so there is only one position where there is no slip. The slip on either side of this position is in opposite directions, and generates opposing friction forces which balance, but the location of the no-slip diameter is strongly influenced by the point of maximum pressure between the rollers and races, and is thus dependent on the ratio of axial to radial load, which varies with the rotational position of the roller in the bearing. The same argument cannot be made for parallel roller bearings, which are unable to sustain an axial load, but on the other hand, they would rarely have negative clearance, and the rollers are only compelled to roll in the load zone. Thus, when they enter the load zone, they will tend to have a random position in the clearance of the cage, and the repetition frequency would have a stochastic variation as for other bearing types, even if the deviation of the mean value from the kinematic frequency is less.

Fig. 2 shows the basic reason why there is often no diagnostic information in the raw spectrum. This shows acceleration signals from a simulated outer race fault, with and without random slip. Spectra are shown for both the raw signal and the envelope. The individual bursts are simulated as the impulse response (IR) of a single degree of freedom (SDOF) system with just one resonance, but this could be the lowest of a series. As is quite common, the assumed resonance frequency is

two orders of magnitude higher than the repetition frequency of the impacts. The Fourier series for the periodically repeated IRs are samples of the frequency response function (FRF) of one IR. Because the FRF is measured in terms of acceleration, the spring line at low frequencies is a ω^2 parabola, with zero value and zero slope at zero frequency. Thus, the low harmonics of the repetition frequency have very low magnitude and are easily masked by other components in the spectrum. If the signal were perfectly periodic, the repetition frequency could be measured as the spacing of the harmonic series in the vicinity of the resonance frequency, but as illustrated in Fig. 2(e), the higher harmonics smear over one another with even a small amount of slip (here 0.75%). However, the envelope spectra (Fig. 2(c), (f)) show the repetition frequency even with the small amount of slip, even though the higher harmonics in the latter case are slightly smeared.

As mentioned, the lowest resonance frequencies significantly excited are often, but not universally, very high with respect to the bearing characteristic frequencies. It would for example not be the case for gas turbine engines, where the fault frequencies are often in the kHz range. Even so, the low harmonics of the bearing characteristic frequencies are almost invariably strongly masked by other vibration components, and it is generally easier to find wide frequency ranges dominated by the bearing signal in a higher frequency range. The advantage of finding an uncontaminated frequency band encompassing several harmonics of the characteristic frequency is that bearing fault signals are generally impulsive, but cannot be recognised as such unless the frequency range includes at least ten or so harmonics. If a pulse train is lowpass filtered between the first and second harmonics of the repetition frequency, the result is a sinewave, with no impulsivity at all. The most powerful bearing diagnostic techniques depend on detecting and enhancing the impulsiveness of the signals, and so the fact that low harmonics of the bearing characteristic frequencies can sometimes be found in raw spectra is basically ignored in the rest of this paper. This is because the authors believe that the purpose of a tutorial is to give details of the most widely applicable method to solve the problem at hand, rather than a catalogue of all publications on the subject, which is more the function of a review. As a counter example, a paper by one of the authors [2] was the first to use the cepstrum to diagnose bearing faults, this relying on being able to find separated harmonics of the bearing frequency over a reasonably wide frequency range. It was a high speed machine (an auxiliary gearbox running at 3000 rpm), and a reasonable number of the first 20 or 30 harmonics were separated and gave a component in the cepstrum. On the other hand, the primary method recommended in this paper, envelope analysis, performed equally well if not better in that case, and does not require the harmonics to be separated, as illustrated in Fig. 2, so the cepstrum method has little application.

Even though this tutorial concentrates primarily on the method of envelope analysis (after first having separated the bearing signal from strong background signals which generally mask it), a brief history will first be given here on the development of bearing diagnostics, and a justification for the choice of the proposed method.

1.1. Short history of bearing diagnostics

One of the earliest papers on bearing diagnostics was by Balderston [3] of Boeing in 1969. He recognised that the signals generated by bearing faults were primarily to be found in the high frequency region of resonances excited by the internal impacts, and investigated the natural frequencies of bearing rings and rolling elements, which were often to be found in

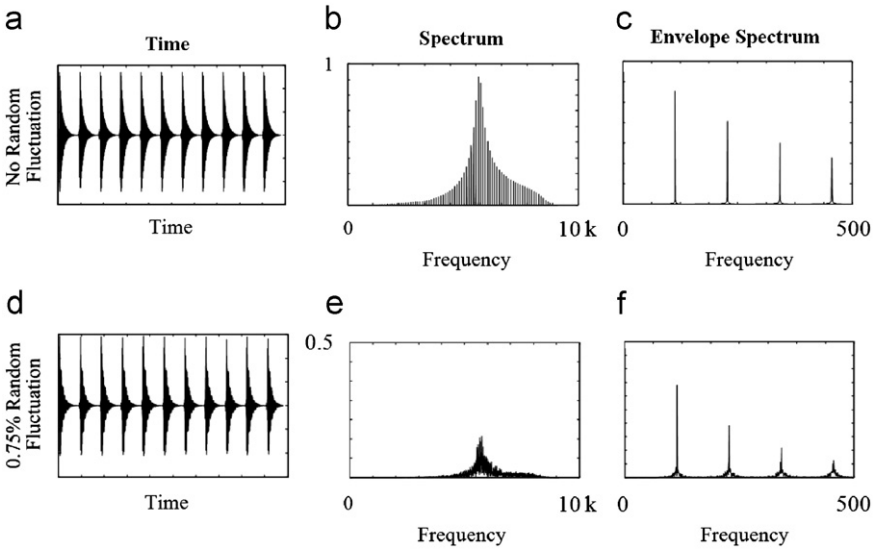

Fig. 2. Bearing fault pulses with and without random fluctuations: (a), (d) time signals; (b), (e) raw spectra; and (c), (f) envelope spectra.

the response vibrations. He pointed out that at such high frequencies, in the tens of kHz, measurable acceleration levels corresponded to extremely small displacements, which could be accommodated in the clearance space between surface asperities of a bearing ring in its housing, even after fitting, and thus natural frequencies were not greatly modified by the mounting. Shortly after, in 1970, Weichbrodt and Smith [4] used synchronous averaging to expose local faults in both bearings and gears. In the former case they sometimes performed averaging on the (rectified) envelope signals. Braun [5] made a fundamental analysis of synchronous averaging in 1975, and the basic technique was also applied to bearing signals [6]. This appears to be one of the first references to the fact that bearing signals are not completely periodic, with a random variation in period. Braun made an analysis of the effects of jitter (of the synchronising signal) and likened this to the random spacing of bearing response impulses. This model, which is effectively Model 1 in the next section (Fig. 4), was much later shown to be incorrect, and so this approach has not been expanded on in this tutorial, even though it can give satisfactory results in some situations.

At around that time, the "high frequency resonance technique" (HFRT), later called "envelope analysis", was developed (see [7] and the first 15 references of [1]). Even though this is described in the previous section as solving the problem of smearing of high harmonics (Fig. 2), this was not the main reason for its development, since it probably was not recognised at the time because of the limited resolution of FFT analysis. The main reason for its development was to shift the frequency analysis from the very high range of resonant carrier frequencies, to the much lower range of the fault frequencies, so that they could be analysed with good resolution. The frequency shifting was done using analogue rectifiers. Even in McFadden and Smith's classic paper on the modelling of bearing fault signals [1], the fault pulses are treated as periodic.

This concept of demodulating high frequency resonant responses led to the development of a number of bearing diagnostic methods, where the demodulated frequency was the resonance of the transducer itself. This includes the "Shock Pulse Meter" (SPM), marketed for some time by the SKF bearing company, and the "Spike Energy" method marketed by IRD. These used the resonance of a conventional accelerometer as the main carrier, in the former case with bandpass filtering around a well-defined frequency of about 32 kHz, and in the latter case a highpass filtering at about 15 kHz, with more tolerance for the transducer resonance. Systems including acoustic emission (AE) transducers, with frequency ranges from 50 kHz to 1 MHz, were also introduced at that time. While often being effective in improving the signal/noise ratio of bearing signal to background noise, this was not universally the case. Ref. [8] describes situations where the transducer resonance happened to coincide with other excitations, such as turbulence and cavitation in pumps, and therefore gave false readings. The authors recommended choosing the appropriate resonance frequency for demodulation in each case.

There has long been a discussion on how to choose the optimum bandwidth for the demodulation associated with envelope analysis. Some recommended searching for a peak at high frequency in response spectra, on the assumption that it would be excited by bearing faults, while others suggested that a hammer tap test would be more likely to identify bearing resonances. In the authors' opinion, prior to the development of the spectral kurtosis (SK) based methods in the current tutorial, the best approach was to demodulate the band with the biggest dB change from the original condition, although this does require having reference signals with the bearings in good condition.

Such methods are not discussed further in this tutorial because the authors believe that the methods proposed herein solve the problems in the vast majority of cases.

Another approach to bearing diagnostics that can be found in the literature, but is not discussed here, are statistical methods based on pattern recognition. These rely on training a pattern recognition system with typical signals representing the different classes to be distinguished. There are two main reasons why such methods are not treated here. One is that they require large amounts of data for the training, and it is very rare that sufficient data can be acquired by experiencing actual faults in practice, including all permutations and combinations of fault type, location, size, machine load and speed, etc., in particular for expensive critical machines. Most published results are not non-dimensionalised and would only apply to a particular bearing on a particular machine for which the system was trained. It is likely that some of these problems will be overcome by fault simulation in the future. The other reason is that the authors believe that the quasi-deterministic approach proposed in this tutorial covers the vast majority of situations, as exemplified by the wide range of different cases treated in Section 6, without requiring excessive amounts of data from failures. Even so, the reader is referred to the Tutorial on "Natural Computing" [9] for a detailed discussion of methods based on pattern recognition.

2. Bearing fault models and cyclostationarity

Bearing faults usually start as small pits or spalls, and give sharp impulses in the early stages covering a very wide frequency range (even in the ultrasonic frequency range to 100 kHz). However, for some faults such as brinelling, where a race is indented by the rolling elements giving a permanent plastic deformation, the entry and exit events are not so sharp, and the range of frequencies excited not so wide. They would still generally be detected by envelope analysis, however.

Cases have been encountered where faults have not been detected while small and the spalls have become extended and smoothed by wear. Although not necessarily generating sharp impacts any more, this type of fault can often be

detected by the way in which it modulates other machine signals, such as the gearmesh signal generated by gears supported by the bearings. Fig. 3 illustrates the case of an extended inner race spall, where the gearmesh signal is modulated by the type of signal shown, a mixture of a deterministic (local mean) part, and an amplitude modulated noise as the rough section of the race comes into the load zone. It should be kept in mind that the rolling elements are on a different part of this rough surface for every revolution of the inner race.

The optimum way to analyse a faulty bearing signal depends on the type of fault present. The main difference is between initial small localised faults, as illustrated in Figs. 1 and 2, and extended spalls, as illustrated in Fig. 3, in particular if the spalls become smoothed.

Both fault types give rise to signals that can be treated as cyclostationary. For a signal to be cyclostationary of order n, its nth order statistics must be periodic. Thus, a first-order cyclostationary signal (CS1) has a periodic mean value (e.g. a periodic signal plus noise) while a second-order cyclostationary signal (CS2) has periodic variance (e.g. an amplitude modulated white noise). The statistics are obtained by ensemble averaging over an ensemble of realisations. The consequences of this are summarised in Appendix A, and much more detail can be found in [10,11].

2.1. Localised faults

For localised faults, the question arises as to the correct way to model the random spacing of the impacts. Perhaps the first publication to model bearing fault signals as cyclostationary was [12], but the results were not very convincing, possibly because the main resonances excited by the faults may have been outside the measured range up to about 6 kHz. Results are shown below (Fig. 18) where localised faults on a very similar sized bearing only manifested themselves at frequencies above 8 kHz. Good results were obtained in [13], by modelling the vibration signals from localised bearing faults as CS2 cyclostationary.

However, the way of modelling the random variation in pulse spacing in Ref. [13] (model 1) was later found to be incorrect, and in [14] a more correct model (model 2) was proposed. As illustrated in Fig. 4, the variation in model 1 was

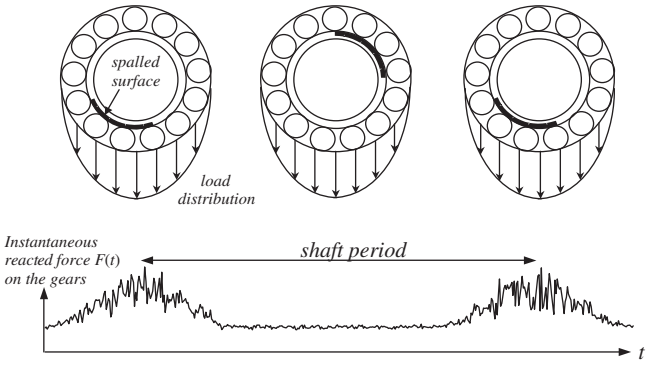

Fig. 3. Typical modulating signal from the effect of an extended inner race fault on a gear signal.

Model 1
Random variable is the jitter δT_i around each period

$$T_i = iT + \delta T_i$$

This gives a truly cyclostationary signal with the uncertainty of occurrence independent of the number of periods into the future

Model 2
Random variable is ΔT_i defined by

$$\Delta T_i = T_{i+1} - T_i$$

This gives a pseudo-cyclostationary signal with the uncertainty of occurrence increasing with the number of periods into the future

Fig. 4. Two models for the variation in period of pulses from a localised bearing fault.

modelled as a random "jitter" around a known mean period, whereas in the correct model it is actually the spacing itself that is the random variable.

In particular, this has implications for the uncertainty of prediction of the location of a future pulse. For model 1, this is constant, and determined by the jitter, whereas in the actual situation, the variation is caused by slip for which the system has no memory, and thus the uncertainty increases with time of prediction into the future (model 2). As pointed out in [14], and put on a firmer mathematical basis in [15], this means that the signals from a localised fault in a bearing are not truly cyclostationary, but are better termed "pseudo-cyclostationary". Fig. 5 (from [14]) shows the practical consequences of this for a signal with a small amount of random variation. It is seen that in terms of interpreting spectra, in particular envelope spectra, usually only at the low harmonics, there is little practical difference in treating the pseudo-cyclostationary signals as cyclostationary.

It was shown in [13] that the spectrum of the squared envelope of a signal (vs cyclic frequency) is equal to the integral of the spectral correlation over all normal frequency. See Appendix A for a discussion of spectral correlation. Fig. 6 shows a typical example.

In this case, the (squared) envelope spectrum contains all the diagnostic information required, namely the harmonics of BPFI, with low harmonics and sidebands spaced at shaft speed, and so there is no real benefit in showing the full spectral correlation.

2.2. Extended spalls

For extended spalls, there will often be an impact as each rolling element exits the spall, and in that case, envelope analysis will often reveal and diagnose the fault and its type. However, there is a tendency for the spalled area to become worn, in which case the impacts might be much smaller than in the early stages. Such extended spalls can still be detected and diagnosed if the bearing is supporting a machine element such as a gear, as discussed in connection with Fig. 3. The typical modulating signal shown in Fig. 3 contains both CS1 (the local mean value) and CS2 (amplitude modulated noise) cyclostationary components. In contrast, for gears supported in healthy bearings, even with a faulty gear, the modulating signal tends to be deterministic (CS1 when mixed with noise), because the same profiles mesh in the same way each time.

Fig. 7 compares the gearmesh modulating signals for a gear fault and an extended inner race bearing fault, and the spectral correlation of the latter. This has discrete characteristics in the cyclic frequency direction, but a mixture of discrete and continuous characteristics in the normal frequency direction. This is because a periodic signal has a periodic autocorrelation function (in the time lag or τ direction) so the Fourier transform in this direction also gives discrete components. Thus the spectral correlation has discrete components in both directions (a "bed of nails"). If the periodic

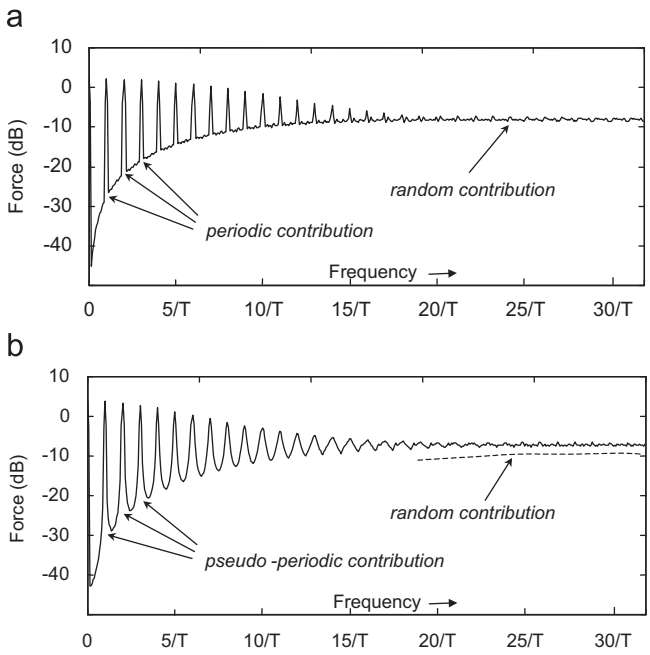

Fig. 5. Frequency spectra for the two models: (a) model 1 and (b) model 2.

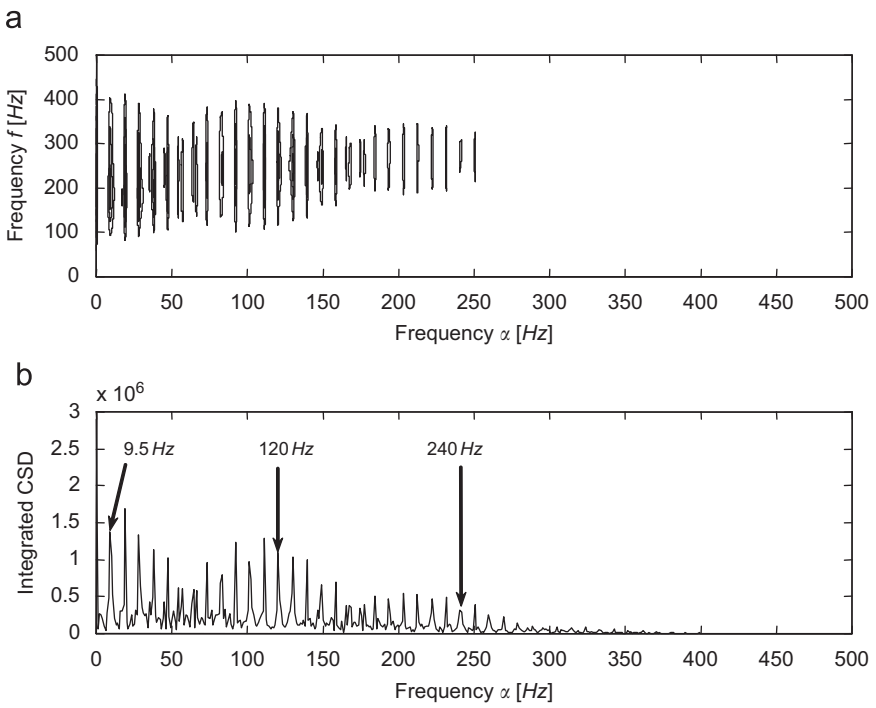

Fig. 6. Spectral correlation and spectrum of the squared envelope for a local inner race bearing fault. BPFI=120 Hz, shaft speed=9.5 Hz.

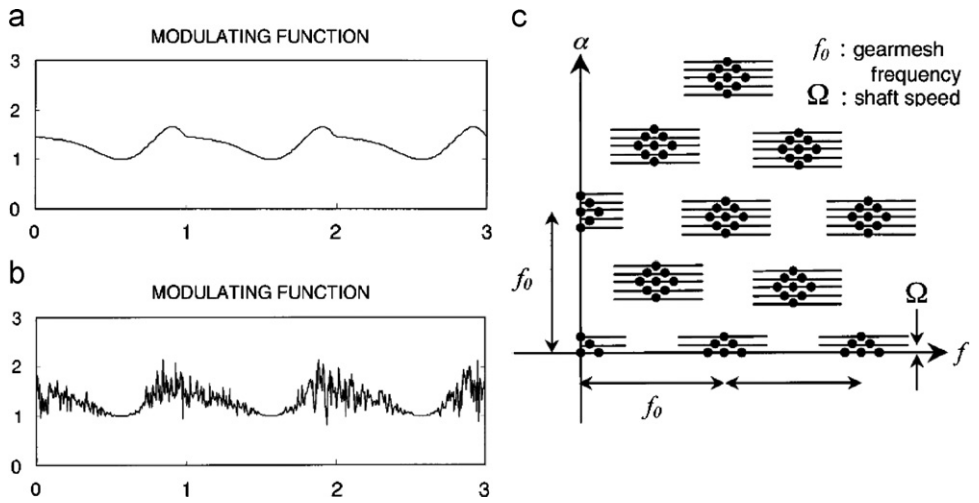

Fig. 7. Spectral correlation for a mixture of first and second-order cyclostationarity, illustrated using modulation by gear and bearing signals: (a) gearmesh modulation by a gear signal; (b) gearmesh modulation by an extended inner race bearing fault; and (c) spectral correlation for case (b). From [14].

components are removed (by one of the methods described below), then only the CS2 components will be left, and they could only come from an extended bearing fault in a case such as this.

Note that the continuous lines in the spectral correlation of Fig. 7 are at the low harmonics of shaft speed (Ω) but also in principle at the harmonics of BPFI and sidebands spaced at shaft speed around them. For an inner race fault the shaft speed is probably the best to use to extract this information, but for an unmodulated outer race fault, components may be found in the spectral correlation at harmonics of BPFO. Note that where the shaft speed is the modulating frequency, the signal is truly second-order cyclostationary (since the cyclic frequency is completely determined) whereas if the modulating frequency is BPFO or BPFI, the signal would be pseudo-cyclostationary.

Fig. 8 (from [14]) shows the spectral correlation, for cyclic frequency equal to shaft speed Ω, for two cases of inner race faults in the same type of bearing. For the localised fault, the difference manifests itself at high frequencies above

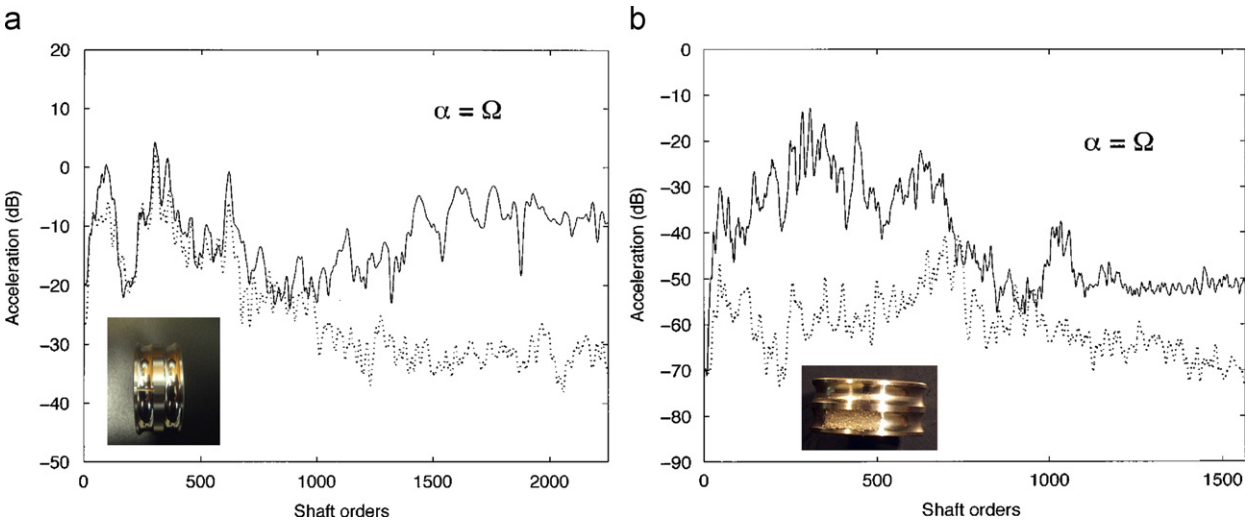

Fig. 8. Spectral correlation evaluated for cyclic frequency equals shaft speed: (a) localised fault and (b) extended fault.

1000 shaft orders, whereas for the extended fault the differences are concentrated at lower frequencies up to 15 times the gearmesh frequency. In the former case the fault was easily detected by envelope analysis, but in the latter case it was much less clear. Fig. 9 (also from [14]) shows an actual case from the input pinion bearing of a helicopter gearbox, where the extended inner race spall was not detected until very late. There was no on-board vibration monitoring, and metal particles were getting trapped in an oil dam, and not reaching the chip detector. By the time these measurements were made on a gearbox test rig, the spall had become smoothed and did not reveal itself by envelope analysis at BPFI, only at the harmonics of shaft speed, and so could have been misinterpreted as a gear fault if this analysis (with removal of deterministic components) had not been done.

3. Separation of bearing signals from discrete frequency noise

One of the major sources of masking of the relatively weak bearing signals is discrete frequency "noise" from gears, since such signals are usually quite strong, even in the absence of gear faults. Even in machines other than gearboxes, there will usually be strong discrete frequency components that may contaminate frequency bands where the bearing signal is otherwise dominant. It is usually advantageous therefore to remove such discrete frequency noise before proceeding with bearing diagnostic analysis.

A number of methods are available, with different pros and cons. The notational convention used throughout the presentation is to denote the measured vibration signal either by $x(t)$ or by $x(n)$ depending on whether continuous-time or discrete-time is most convenient. Filtered versions of x will be systematically denoted by y. For convenience, the spectrum of any signal, say s, will be denoted by $S(f)$ independently of whether s is continuous or discrete-time.

3.1. Linear prediction

Linear prediction is basically a way of obtaining a model of the deterministic (i.e. "predictable") part of a signal, based on a certain number of samples in the immediate past, and then using this model to predict the next value in the series. The residual (unpredictable) part of the signal is then obtained by subtraction from the actual signal value.

The model used for linear prediction is an "autoregressive" or AR model as described by the following equation:

$$y(n) = -\sum_{k=1}^{p} a(k)x(n-k) \tag{5}$$

where the predicted current value $y(n)$ is obtained as a weighted sum of the p previous values.

The actual current value is given by the sum of the predicted value and a noise term:

$$x(n) = y(n) + e(n) \tag{6}$$

As described in [16] the $a(k)$ can be obtained using the Yule–Walker equations, often using the so-called Levinson–Durban recursion (LDR) algorithm.

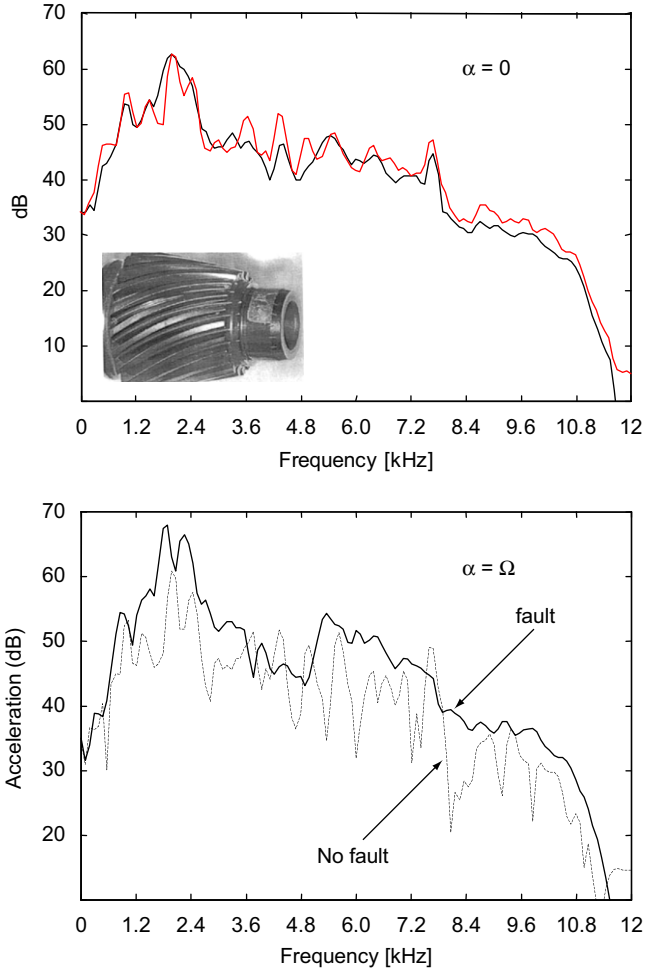

Fig. 9. Comparison of spectral correlation evaluated for cyclic frequency equals zero (normal power spectrum) and shaft speed Ω for the depicted extended inner race spall. The fault is only apparent at $\alpha=\Omega$. Discrete frequency components were removed before the spectral correlation analysis.

Note that Eqs. (5) and (6) can be combined and written as

$$x(n)+\sum_{k=1}^{p} a(k)x(n-k)=e(n) \tag{7}$$

which can be Fourier-transformed to

$$X(f)A(f)=E(f) \tag{8}$$

or

$$X(f)=\frac{E(f)}{A(f)} \tag{9}$$

which can be considered as the output $X(f)$ of a system with transfer function $A^{-1}(f)$ when excited by the forcing function $E(f)$. The transfer function is thus an all-pole filter, and can be interpreted as an autoregressive (AR) system. The forcing function $E(f)$ is white, containing stationary white noise and impulses, and its time domain counterpart $e(n)$ is said to be "prewhitened". Thus, removing the deterministic (discrete frequency) components leaves a prewhitened version of the residual signal, which includes the bearing signal because of the randomness of the latter.

3.2. Adaptive noise cancellation

Adaptive noise cancellation (ANC) is a procedure whereby a (primary) signal containing two uncorrelated components can be separated into those components by making use of a (reference) signal containing only one of them. The reference

signal does not have to be identical to the corresponding part of the primary signal, just related to it by a linear transfer function. The ANC procedure adaptively finds that transfer function, and can thus subtract the modified reference signal from the primary signal, leaving the other component [17].

ANC was proposed many years ago [18,19] as a means of extracting a faulty bearing signal in cases where the primary signal could for example be measured on the faulty bearing of a gearbox, and the reference signal on another remote bearing with no contamination from the faulty bearing signal. However, it does rely on being able to obtain an uncontaminated reference signal, which would not be possible for example on a planetary gearbox, where all signals must be transmitted through the ring gear.

3.3. Self-adaptive noise cancellation

When one of the two components to be separated is deterministic (discrete frequency) and the other random, the reference signal can be made a delayed version of the primary signal, because if the delay is longer than the correlation length of the random signal, the adaptive filter will not recognise the relationship, and will find the transfer function between the deterministic part of the signal and the delayed version of itself. Thus, the separation can be achieved using one signal only, and the procedure is called self-adaptive noise cancellation (SANC). This is illustrated in Fig. 10, applied to the separation of gear and bearing signals, where the gear signal is deterministic.

The adaptive filter in Fig. 10 is a recursive filter, with a number of weights to be determined, but which also updates at each step, which means it can cope with slow changes to the signal or system properties. The recursive algorithm is the so-called least mean squares (LMS) algorithm (Widrow and Stearns [17]) and can be expressed as follows:

$$\mathbf{W}_{k+1} = \mathbf{W}_k - \mu\nabla_k \tag{10}$$

where gradient vector

$$\nabla_k = \frac{\partial E[\varepsilon_k^2]}{\partial \mathbf{W}_k} \tag{11}$$

and μ is a convergence factor, which should be chosen carefully to avoid divergence on the one hand, but not give rise to excessive adaptation time on the other.

Ho [20] shows that a conservative approximation for Eq. (10) is

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \frac{2\mu_n \varepsilon_k \mathbf{X}_k}{(L+1)\hat{\sigma}_k^2} \tag{12}$$

where $\mathbf{W}_k$ is the vector of weight coefficients of the adaptive filter at the kth iteration, μ_n the normalised convergence factor: $0 < \mu_n < 1$, μ the convergence factor: $\mu = (\mu_n/((L+1)\hat{\sigma}_k^2))$, ε_k the output error at the kth iteration, $\mathbf{X}_k$ the vector of input values at the kth iteration, L the order of the adaptive filter, $(L+1)$ the number of filter coefficients and $\hat{\sigma}_k^2$ the exponential-averaged estimate of the input signal power at the kth iteration.

Thus, there are three factors to be chosen for a successful result. Perhaps the most important is L, the order of the adaptive filter, which for mechanical systems, and in particular the separation of gear and bearing signals, is quite large, typically in the hundreds and even thousands.

An empirical study of the optimum choice of filter order was made in Ref. [21], and the basic results (for a single family of equally spaced harmonics or sidebands in the band being treated) are given in Fig. 11. Ref. [21] also contains recommendations for the situation where there is more than one family, and then the minimum frequency spacing plays a role.

For fewer than 12 spectral components some saving can be made, but above this the minimum order corresponds to approximately one period of the frequency spacing. Ho recommends that the order should be at least double the minimum for best results.

In Ref. [22] SANC was likened to prediction theory, and it was pointed out that the asymptotic result in Fig. 11 is in agreement with the analytical result that the bandwidth of the SANC filter is the reciprocal of the period of the frequency being separated, and the filter characteristic is almost identical to that for the equivalent comb filter for synchronous

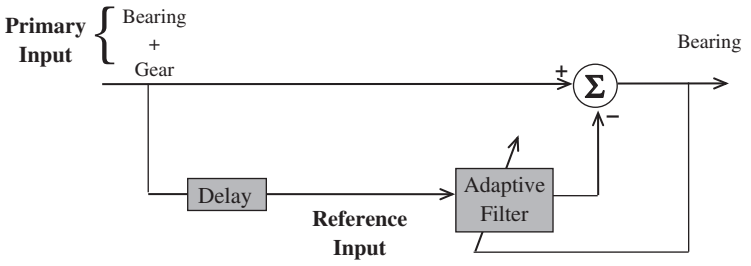

Fig. 10. Schematic diagram of self-adaptive noise cancellation used for removing periodic interference (gear) leaving random signal (bearing).

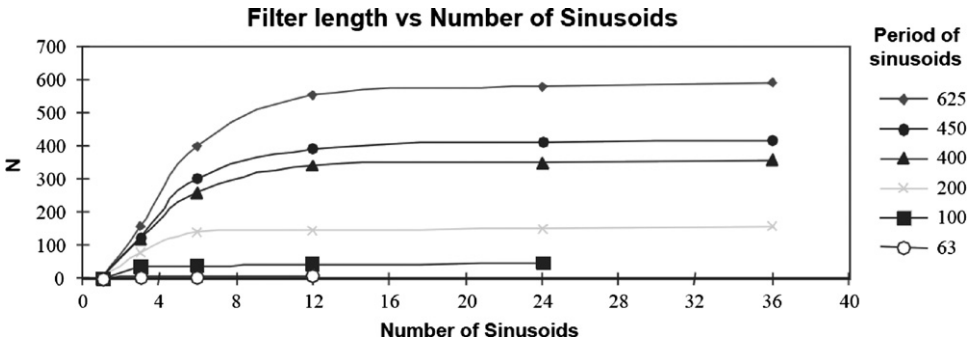

Fig. 11. Minimum filter order vs number of discrete spectrum components.

averaging (e.g. Fig. 13). In [23] it was also pointed out that it can be advantageous to vary the convergence factor exponentially, to start with a larger value and reduce it as the convergence proceeds.

With respect to delay, this should be chosen so as to be longer than the correlation length of the random part of the signal, but shorter than that of the deterministic part. In principle, the latter is infinite, but in practice the correlation does decrease with increasing delay, so the delay should be as short as possible. Some guidance is therefore needed to determine the correlation length of the random part. This will normally be governed by the bandwidth of (resonance) peaks in the band being processed, the narrowest corresponding to the longest correlation length, and this can sometimes be judged by inspection, or by knowledge of typical damping factors of the system. For narrow band noise of bandwidth B, the correlation length is of the order of $1/B$. Thus, for a 3 dB bandwidth of 1% it would correspond to 100 periods of the centre frequency. The actual delay should be made perhaps three times this correlation length.

Even so, in most cases the discrete/random separation technique discussed in the next section would generally give similar results to SANC, but more efficiently and without the possibility of divergence.

3.4. Discrete/random separation (DRS)

The SANC, as being an adaptive filter, is capable of adjusting to small speed variations. The price to pay is a convergence stage which may last for quite long, especially for filters of high orders. In circumstances where the components to be tracked are very stable in time or can be made so by order-tracking the signal in a pre-processing step, adaptation is no longer needed and a much more efficient estimation of the filter is possible in a batch way. Since the goal is to find a filter that predicts a current value of the signal from its delayed version, this can be simply achieved in the frequency domain by computing the frequency response function (FRF) between two blocks of data, $[y(n-N+1),\ y(n-N+2),\ldots,\ y(n)]$ and $[y(n-N+1-\Delta),\ y(n-N+2-\Delta),\ldots,\ y(n-\Delta)]$, distant by some delay Δ. Namely, denoting by $Y(f)$ and $Y_\Delta(f)$ the Fourier transforms of the two blocks at frequency f, the corresponding FRF is

$$H(f)=\frac{\mathbb{E}\{Y(f)Y_\Delta^*(f)\}}{\mathbb{E}|Y(f)|^2} \tag{13}$$

In practice, the expected value in Eq. (13) will be approached by averaging many blocks of data, and these should be windowed before Fourier transforming in order to reduce leakage errors. The use of the FFT in estimating $H(f)$ will definitively speed up the computation by orders of magnitude as compared with the SANC, while virtually achieving the same result. As a matter of fact, Ref. [24] shows that the amplitude of the separation filter is

$$H(f)=\frac{(\rho N/2)|W(f)|^2}{(\rho N/2)|W(f)|^2+1} \tag{14}$$

where ρ=SNR, N is the transform size, and $W(f)$ is the Fourier transform of the window used, scaled to a maximum value of 1 in the frequency domain. This ideally returns a value of unity at the frequencies of the discrete components (where the signals are correlated) and zero at frequencies where there is noise when the product ρN is large. Even for a SNR as low as 10^{-2} (-20 dB), this gives a value of 0.7 for N=512. This is the same as for the equivalent SANC filter. The filter characteristic is somewhat poorer than for the equivalent SANC, in terms of sidelobes for a rectangular window, or noise bandwidth if a window such as Hanning is used, but on the other hand, the DRS is so much more efficient that a longer filter (larger value of N) can be used to give better resolution.

Once the filter $H(f)$ is estimated in the frequency domain, it is applied on the original signal to filter out all periodic components. This is again amenable to a fast implementation by using the FFT on short blocks of data in conjunction with the "overlap/add" method in order to remove artefacts due to the circular convolution which is implicit beyond this approach. Note that if the amplitude spectrum of the filter is used, this gives zero phase shift of the separated signal.

The filter is then non-causal. Note also that the frequency domain method can be used on the one-sided spectra of analytic signals, giving a further advantage when used in combination with demodulation, such as used in bearing diagnostics.

The size of transform N should span 10–20 periods of the minimum frequency to be removed (e.g. the lowest shaft speed) and the amount of delay to be chosen is the same as that recommended for SANC, i.e. about 300 periods of the centre frequency of the demodulated band for envelope analysis.

Fig. 12 shows a typical result [24] of using DRS to separate a gearbox vibration signal into its deterministic components (dominated by the gears) and random components (in this case dominated by a bearing with an outer race fault).

3.5. Time synchronous averaging (TSA)

On rare occasions, there can be a need to remove discrete frequency components with minimum disruption of the residual signal, and this can best be done using time synchronous averaging (TSA), although it can be tedious, since it requires a separate operation for every different set of harmonics in the signal, including a separate resampling in each case so as to give a specified integer number of samples per period, and an integer number of periods to be averaged.

In practice it is done by averaging together a series of signal segments each corresponding to one period of a synchronising signal. Thus:

$$y(t) = \frac{1}{N}\sum_{n=0}^{N-1} x(t+nT) \tag{15}$$

This can be modelled as the convolution of $x(t)$ with a train of N delta functions displaced by integer multiples of the periodic time T, which corresponds in the frequency domain to a multiplication by the Fourier transform of this signal, which can be shown to be given by the expression [5]:

$$H(f) = \frac{1}{N}\frac{\sin(N\pi Tf)}{\sin(\pi Tf)} \tag{16}$$

The filter characteristic corresponding to this expression is shown in Fig. 13 ([25]) for the case where $N=8$, and is seen to be a comb filter selecting the harmonics of the periodic frequency. The greater the value of N the more selective the filter and greater the rejection of non-harmonic components. The noise bandwidth of the filter is $1/N$, meaning that the improvement in signal/noise ratio is $10\log_{10} N$ dB for additive random noise. For masking by discrete frequency signals, it should be noted that the characteristic has zeros which move with the number of averages, so it is often possible to choose a number of averages which completely eliminates a particular masking frequency. The above characteristic is for an infinitely long time signal $x(t)$, and in Ref. [25] it is shown that for the practical situation of a finite length of signal with

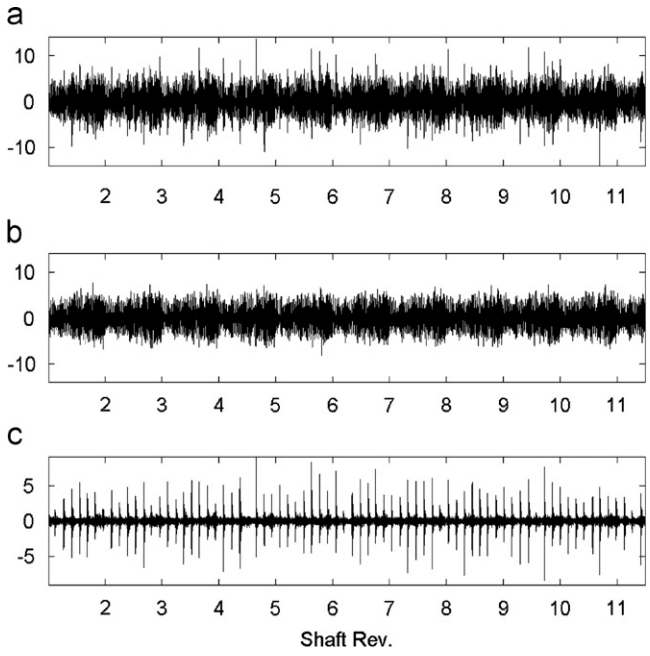

Fig. 12. Example of enhancement of an outer-race fault signal in a gearbox: (a) measured vibration signal, (b) extracted periodic part (gears), and (c) extracted non-deterministic part (bearing).

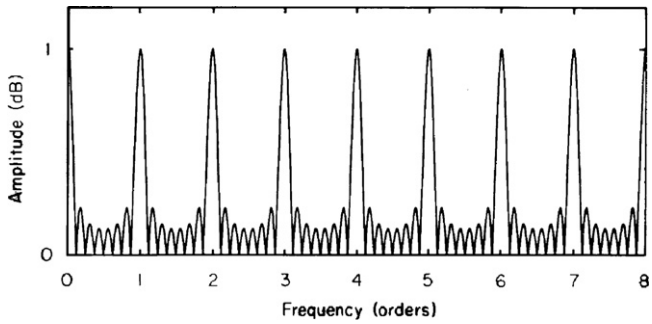

Fig. 13. Filter characteristic corresponding to 8 synchronous averages (from [25]).

finite sampling frequency, the above simplified model is not exactly true for discrete frequency "noise", but that it is often still possible to find an optimum number of averages to completely remove a discrete masking signal.

For good results the synchronising signals should correspond exactly with samples of the signal to be averaged, as one sample spacing corresponds to 360° of phase of the sampling frequency, and thus to 144° of phase at 40% of it, which is a typical maximum signal frequency. Moreover, even a 0.1% speed fluctuation would cause a jitter of the same order of the last sample in a (typical) 1 K record, with respect to the first, and thus an even greater loss of information at the end of the record, after averaging.

Sampling the signal using a sampling frequency derived from the synchronising signal (order tracking), solves both these problems and is always to be recommended. Because of its importance, order tracking is discussed in Appendix B.

To remove a particular harmonic family from a signal, it must be resampled with an integer number of samples per period of that component, and then an integer number of periods N averaged together. The extracted single period is then repeated N times and subtracted from the original signal to give the residual. An alternative way of achieving virtually the same result uses the Fourier transform of the entire record by means of the FFT algorithm. This is most efficient when the transform size is a power of two, so both the number of samples per period and the number of periods should be powers of two. The harmonics of the periodic signal will then be concentrated in single spectral lines (no leakage) and can be removed from the (complex) spectrum. The residual values at the harmonic frequencies should not be set to zero; the best estimate of the noise at these frequencies is the mean of the (complex) noise values in the adjacent lines on each side. The spectrum can then be inverse transformed to the time domain to obtain the whole residual signal.

Fig. 14 from [26] compares the results of synchronous averaging (including the alternative method) with DRS in terms of the residual signal. Fig. 14(a) shows the spectrum of the original signal, from a helicopter gearbox, after order tracking. It is dominated by two sets of harmonics at 567.7 and 1900 Hz, respectively. Since all shafts are geared together, the order tracking would have removed the speed variations of all, though only one could be sampled to have an integer number of samples per period. Fig. 14(b) shows the results of removal of the two sets of harmonics in the frequency domain (the alternative method), requiring a separate resampling of the signal in the second case to obtain an integer number of samples per period, even though only the first was required to remove speed fluctuations. Fig. 14(c) shows the results of removal by synchronous averaging, also requiring separate resampling of the signal in each case. Figs. 14(b) and (c) are virtually identical. Finally, Fig. 14(d) shows the results of applying DRS, where both sets of harmonics were removed in one operation (and in principle any other discrete frequency components as well). The effect of the notch filter can be seen in the residual spectrum, and in particular, as highlighted in the figure, there are two narrow band noise peaks that have been removed along with the discrete frequency harmonics with which they were associated. Occasionally, though very rarely, the coarser notch filter given by DRS might remove relevant information.

Note that the separate resampling for each set of harmonics might have to be done in any case for gas turbine engines with two or three independent shafts, where even after removal of speed fluctuation for one shaft, the residual relative variation of the others might be too great to permit separation by DRS.

Note that order tracking actually changes the "time" axis to rotation angle, but the term TSA is used here as long as the machine has nominally constant speed.

An updated treatment of generalised synchronous averaging will be found in the current series of tutorials in [27].

4. Enhancement of the bearing signals

Even after removal of discrete frequency "noise", the bearing signal will often be masked in many frequency bands by other noise, and may also be rendered less impulsive than at the source if the individual fault pulses are modified by passage through a transmission path with a long impulse response (IR). This is most likely to be the case with high speed bearings, where the bearing fault frequencies are so high, and corresponding spacings so short, that the IR is of the same length as the intervals between them. A method known as minimum entropy deconvolution (MED), which removes the

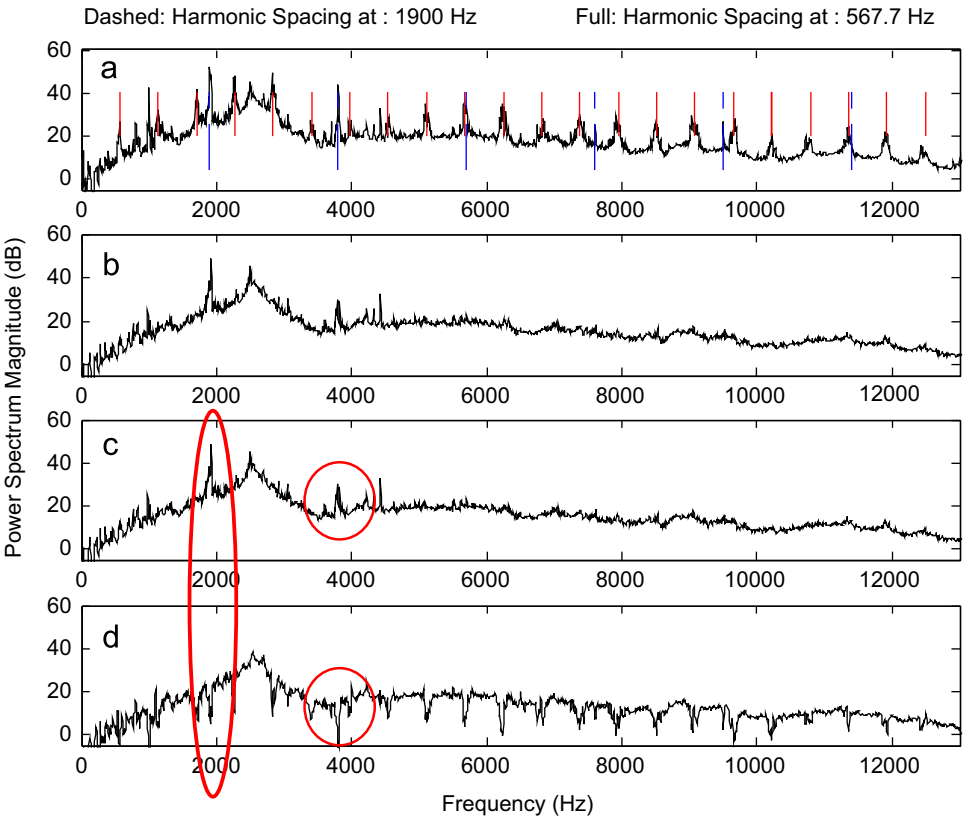

Fig. 14. Power spectrum of (a) order tracked signal; (b) residual signal obtained by setting the rotor related harmonics to the mean of adjacent (noise) lines; (c) residual signal obtained by subtracting the synchronous average and (d) DRS residual (removing discrete components using DRS).

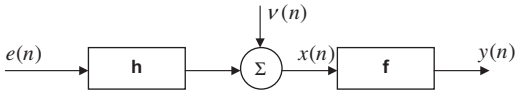

Fig. 15. Inverse filtering (deconvolution) process for MED.

effect of the transmission path, is thus first discussed, and then a number of methods presented which enhance the bearing signal with respect to residual background noise.

4.1. Minimum entropy deconvolution

The "minimum entropy deconvolution" (MED) method is designed to reduce the spread of IRFs, to obtain signals closer to the original impulses that gave rise to them. It was first proposed by Wiggins [28] to sharpen the reflections from different subterranean layers in seismic analysis. The basic idea is to find an inverse filter that counteracts the effect of the transmission path, by assuming that the original excitation was impulsive, and thus having high kurtosis. The name derives from the fact that increasing entropy corresponds to increasing disorder, whereas impulsive signals are very structured, requiring all significant frequency components to have zero phase simultaneously at the time of each impulse. Thus, minimising the entropy maximises the structure of the signal, and this corresponds to maximising the kurtosis of the inverse filter output (corresponding to the original input to the system). The method might just as well be called "maximum kurtosis deconvolution" because the criterion used to optimise the coefficients of the inverse filter is maximisation of the kurtosis (impulsiveness) of the inverse filter output. The MED method was applied to gear diagnostics in [29], and to bearing diagnostics in [30].

Fig. 15 illustrates the basic idea. The forcing signal $e(n)$ passes through the structural filter $\mathbf{h}$ whose output is mixed with noise $v(n)$ to give the measured output $x(n)$. The inverse (MED) filter $\mathbf{f}$ produces output $y(n)$, which has to be as close as possible to the original input $e(n)$. Of course the input $e(n)$ is unknown, but is assumed to be as impulsive as possible.

The filter **f** is modelled as an FIR filter with L coefficients such that

$$y(n) = \sum_{l=1}^{L} f(l)v(n-l) \tag{17}$$

where f has to invert the system IRF h such that

$$f^{*}h(n) = \delta(n-l_m) \tag{18}$$

The delay l_m is such that the inverse filter can be causal. It will displace the whole signal by l_m but will not change pulse spacings.

The method adopted in [29,30] is the objective function method (OFM) given in [31], where the objective function to be maximised is the kurtosis of the output signal $y(n)$, by varying the coefficients of the filter f. This kurtosis is taken as the normalised fourth order moment given by

$$O_k(f) = \frac{\sum_{n=0}^{N-1} y^4(n)}{\left[\sum_{n=0}^{N-1} y^2(n)\right]^2} \tag{19}$$

and the maximum is found by finding the values of f for which the derivative of the objective function is zero, i.e.:

$$\frac{\partial O_k(f)}{\partial f} = 0 \tag{20}$$

Ref. [31] describes how this can be achieved iteratively, when the filter coefficients of f converge within a specified tolerance.

An example of the application to bearing diagnostics is given in Fig. 16 from [30]. As reported in [30], the bearing under test is a high speed bearing similar to those used in gas turbine engines, but mounted in a test rig, where a real spall was induced. In this case linear prediction filtering was used to remove the discrete frequency components related to the harmonics of the shaft speeds of the test machine, as well as prewhitening the signal.

The linear prediction (AR) filtering has made the pulses visible, but because of the high speed (12,000 rpm), the impulse responses excited have a length comparable to their spacing, and tend to overlap. The kurtosis has only improved from −0.40 to 1.25 in the AR operation, but increased to 38.6 after applying MED. This fault is at a fairly advanced stage, but using MED meant that it could be detected much earlier. Note that the kurtosis used here is based on cumulants, and is adjusted so as to give a value of zero for Gaussian signals. This means subtracting 3 from the normalised fourth order (centred) moment.

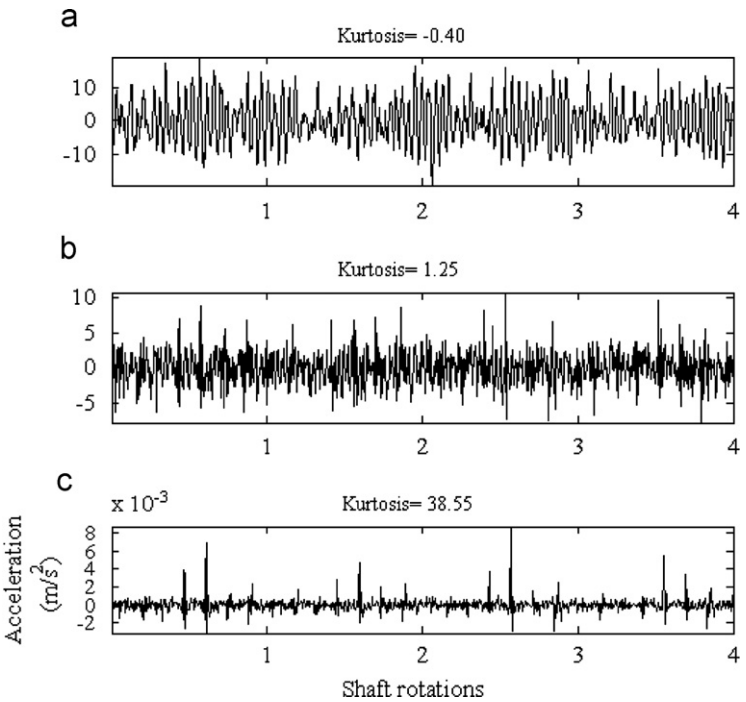

Fig. 16. Example of applying both AR and MED filtering to bearing signals with an inner race fault in a high speed bearing: (a) original time signal; (b) after application of AR filtering; and (c) after additional MED filtering.

4.2. Spectral kurtosis and the kurtogram

From the earliest days of envelope analysis there has been a debate on how to choose the most suitable band for demodulation, with many claiming that it is difficult, and some recommending the use of hammer tap testing to find bearing housing resonances. This problem has now largely been solved by the use of spectral kurtosis (SK) and the kurtogram to find the most impulsive band (after removal of discrete frequency masking).

Spectral kurtosis (SK) provides a means of determining which frequency bands contain a signal of maximum impulsivity. It was first used in the 1980s for detecting impulsive events in sonar signals [32]. It was based on the short time Fourier transform (STFT) and gave a measure of the impulsiveness of a signal as a function of frequency. Kurtosis had long been used as a measure of the severity of machine faults, since its proposal by Stewart et al. in the 1970s, (e.g. [33]) but there was only a vague suggestion that clearer results might be achieved by using filtering in frequency bands, typically octaves, and the concept of spectral kurtosis was not really developed.

The application of SK to bearing faults was first outlined in [34,35], where a very thorough study was made of the definition and calculation of the SK for this purpose.

4.2.1. Spectral kurtosis—definition and calculation

The spectral kurtosis extends the concept of the kurtosis, which is a global value, to that of a function of frequency that indicates how the impulsiveness of a signal, if any, is distributed in the frequency domain. The principle is analogous in all respects to the PSD which decomposes the power of a signal vs frequency, except that fourth-order statistics are used instead of second order. This makes the spectral kurtosis a powerful tool for detecting the presence of transients in a signal, even when they are buried in strong additive noise, by indicating in which frequency bands these take place.

The spectral kurtosis of a signal $x(t)$ may be computed from the STFT $X(t,f)$, that is the local Fourier transform at time t obtained by moving a window along the signal. When seen as a function of frequency, the squared magnitude $|X(t,f)|^2$—i.e. the spectrogram—returns the power spectrum at time t and its further average over time, $\langle |X(t,f)|^2 \rangle$, the PSD as computed by the Welch method. When seen as a function of t, $X(t,f)$ may be interpreted as the complex envelope of signal $x(t)$ bandpass filtered around frequency f and its squared magnitude will then indicate how energy is flowing in that frequency with respect to time. If that frequency band happens to carry pulses, bursts of energy will then appear. This may be simply detected by computing the kurtosis of the complex envelope $X(t,f)$ as follows:

$$K(f) = \frac{\left\langle |X(t,f)|^4 \right\rangle}{\left\langle |X(t,f)|^2 \right\rangle^2} - 2 \tag{21}$$

with $\langle \bullet \rangle$ the time-averaging operator and where the subtraction of 2 is used to enforce $K(f)=0$ in the case $X(t,f)$ is *complex* Gaussian (instead of 3 for *real* signals) [34]. The interpretation of the spectral kurtosis is further illustrated in Fig. 17 in the case of a rolling element bearing signal $x(t)$ modelled as a series of impulse responses $g(t)$ excited by impulses X at times τ_k:

$$x(t) = \sum_k g(t-\tau_k) X(\tau_k). \tag{22}$$

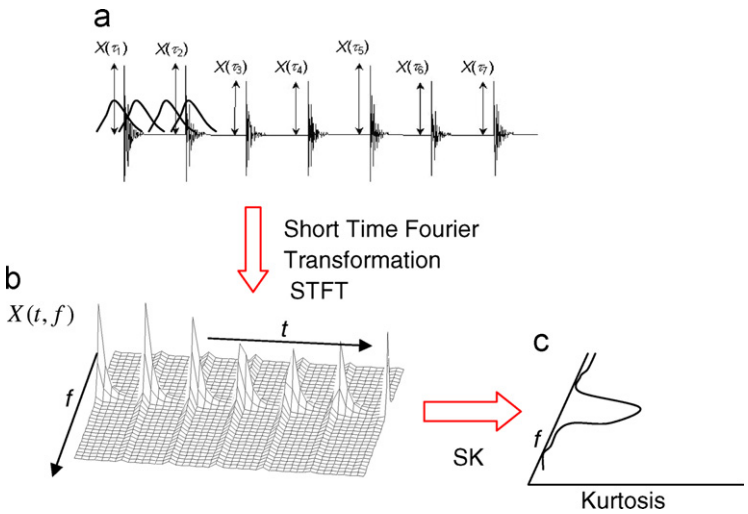

Fig. 17. Calculation of SK from the STFT for a simulated bearing fault signal: (a) simulated time signal, (b) STFT, and (c) SK as a function of frequency.

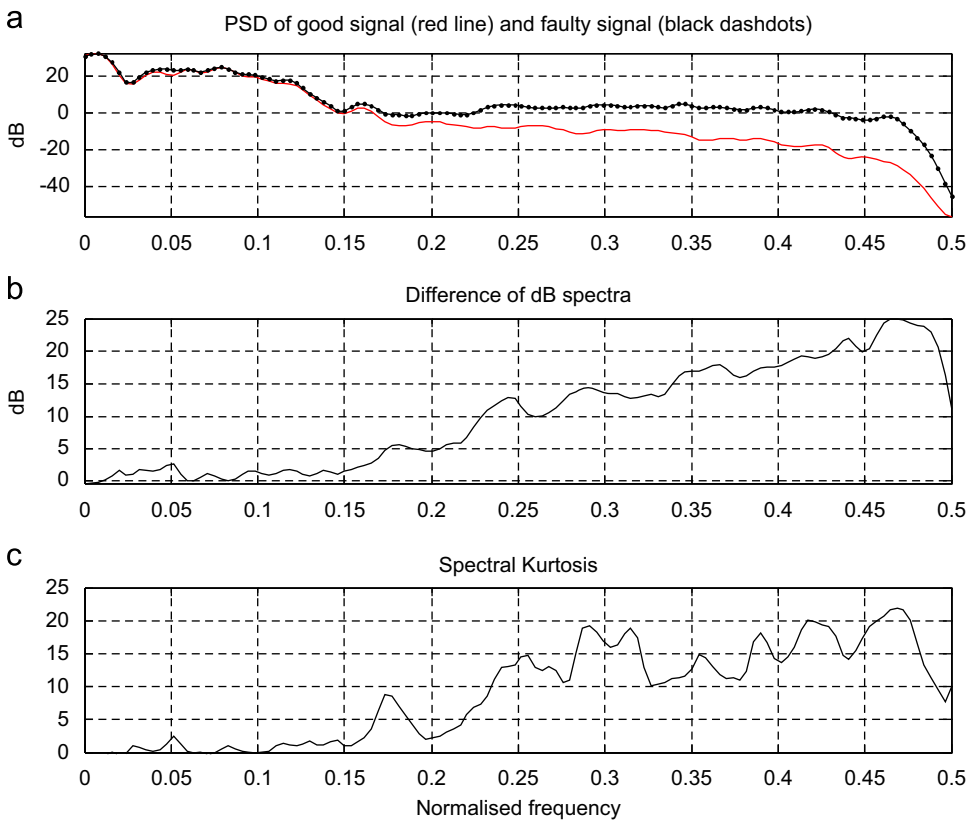

Fig. 18. Comparison of SK with dB spectrum difference for an inner race bearing fault. Frequency scale normalised to sampling frequency 48 kHz: (a) dB spectrum comparison over frequency range 0–24 kHz with and without the fault; (b) dB spectrum difference; and (c) spectral kurtosis.

This example makes it clear that, to obtain a maximum value of the spectral kurtosis, the STFT window must be shorter than the spacing between the pulses, but longer than the individual pulses [35]. Since the spacing between the pulses is most often unknown, a simple strategy is to compute the spectral kurtosis for different putative window lengths and then to select that result with the maximum overall value. A good advice is to perform a similar analysis on the PSD—i.e. as it appears in the denominator of Eq. (21)—so as to control the excessive leakage (smoothing) that too short a window would introduce. This is illustrated in the following example.

Fig. 18 shows a comparison of the spectral kurtosis with the dB spectrum difference caused by an inner race fault in a ball bearing. Note that the shape of the SK curve is very similar to the dB difference, and thus conveys virtually the same information without requiring historical data. It is also interesting that the actual values of SK and dB difference are very similar, though this is somewhat fortuitous because the scaling of the SK is dependent on the choice of window parameters, as stated above.

Fig. 19 gives typical results for how to choose the window length. In this case, window lengths of 32 or 64 samples would be acceptable, 16 giving too much smoothing, and greater than 64 reducing the SK too much (because of bridging between pulses).

In Fig. 17 the peak in the SK is depicted as occurring at the resonance frequency of the impulse responses, but this requires qualification. In [36] the hypothetical case of a repeated SDOF response with a small amount of added noise was investigated. It was found that there was an anomaly in the results; a dip in the SK in the vicinity of the resonance peak, accentuated in the case of a shorter window. This was found to be because the STFT at the resonance peak extends for a longer time than on either side (as can in fact be seen in Fig. 17) and this more than compensates for the higher peak value. The use of prewhitening by AR and/or MED methods prevents this anomaly by making the STFT shorter at the resonance peak.

4.2.2. Use of SK as a filter

Because of the high values it takes at those frequencies where an impulsive bearing fault signal is dominant and its theoretical nullity where there is stationary noise only, it makes sense to use the spectral kurtosis as a filter function to filter out that part of the signal with the highest level of impulsiveness. It has been shown in Ref. [35] that in the case of an impulsive signal $x(t)$ of the type given in Eq. (22) buried in additive stationary noise $n(t)$, the resulting measurement

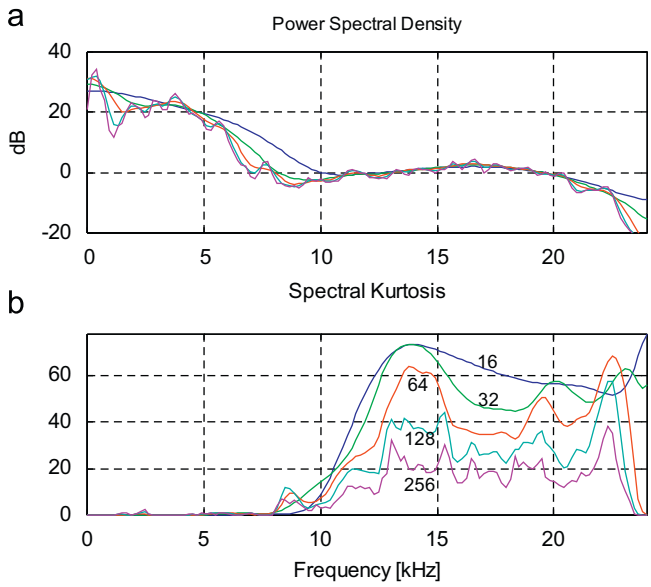

Fig. 19. (a) PSD spectra with the different window lengths and (b) SK calculated for the indicated window length in samples (from [35]).

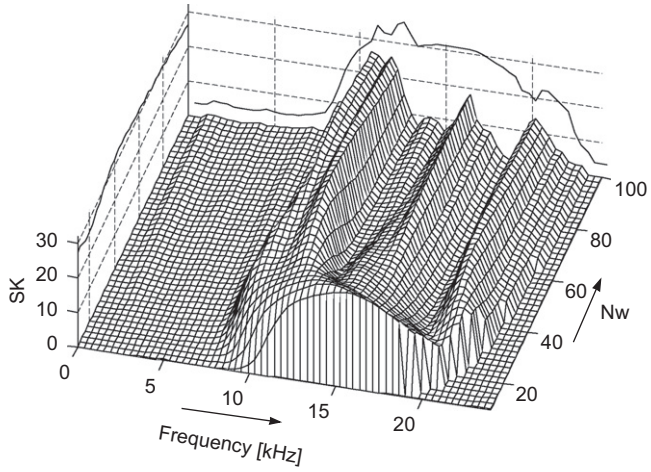

Fig. 20. Kurtogram for a weak bearing fault in a gearbox [35]. N_W is window length defining spectral resolution. SK is spectral kurtosis. Maxima are projected onto each plane.

$y(t)=x(t)+n(t)$ has spectral kurtosis

$$K_y(f)=\frac{K_x(f)}{[1+\rho(f)]^2} \tag{23}$$

where $K_x(f)$ is the spectral kurtosis of $x(t)$ and $\rho(f)=S_n(f)/S_x(f)$ the noise-to-signal ratio. This suggests that the optimal filter that maximises the similarity between the filtered component and the true noise-free signal—i.e. the Wiener filter—is the square root of the spectral kurtosis. Following similar lines, it can also be shown that the optimum filter that maximises the SNR of the filtered signal without regard to its shape—i.e. the matched filter—is a narrow band filter at the maximum value of spectral kurtosis.

4.2.3. *The kurtogram*

As previously pointed out, the spectral kurtosis, and therefore the optimal filter which can be obtained from it, will critically depend on the choice of the STFT window length or, equivalently stated, on the bandwidth of the band-pass filter that outputs the complex envelope $X(t,f)$. One solution is to display the spectral kurtosis also as a function of the latter parameter, thus giving rise to a two-dimensional representation called "kurtogram". This is basically a cascade of spectral kurtoses obtained for different values of the STFT window length as was done in Fig. 19b, but for a much finer grid of values. Fig. 20 shows the kurtogram for a gearbox signal with a weak outer race fault [35]. In this example, an optimal pair

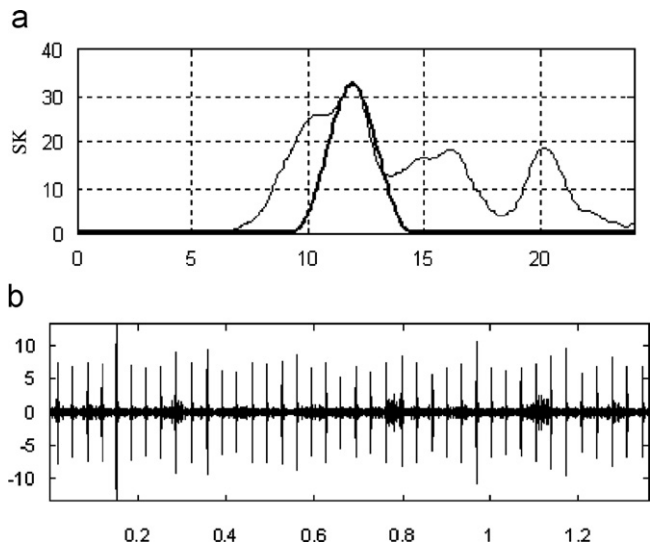

Fig. 21. (a) Optimal bandpass filter compared with SK at N_w=44 and (b) outer race fault signal obtained using the filter of (a) (from [35]).

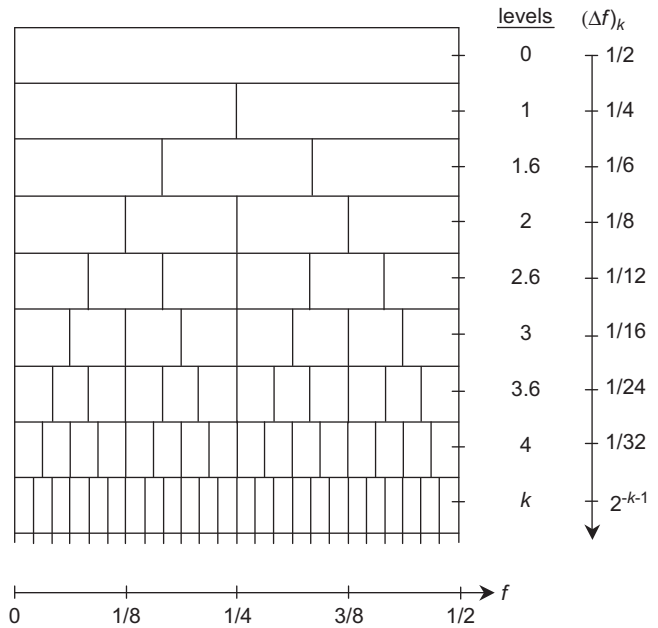

Fig. 22. Combinations of centre frequency and bandwidth for the 1/3-binary tree kurtogram estimator [37].

of values achieves the global maximum for a window length of 44 samples around a frequency of 12.5 kHz. This may be used to design a band-pass filter with similar centre frequency and bandwidth, as illustrated in Fig. 21(a). Finally Fig. 21(b) shows the filtered time signal resulting from this optimum filter. It has a kurtosis of 46.7, compared with 9.9 and 25.8, respectively, for the Wiener and matched filters of Section 4.2.2 [35].

4.2.4. The fast kurtogram

Computation of the kurtogram for all possible combinations of centre frequencies and bandwidths is obviously costly and not convenient for practical purposes. Suboptimal solutions are however conceivable by subdivision of the bandwidths into rational ratios that permit the use of fast multirate processing. The simplest division in this respect is the dyadic one, where bandwidths are iteratively halved, starting from the largest bandwidth that covers the entire signal spectrum—equivalent to a one-sample window length as used in the computation of the conventional kurtosis—and stopping at some narrow bandwidth whose only limit is to leave enough independent time samples in $X(t,f)$ to compute the

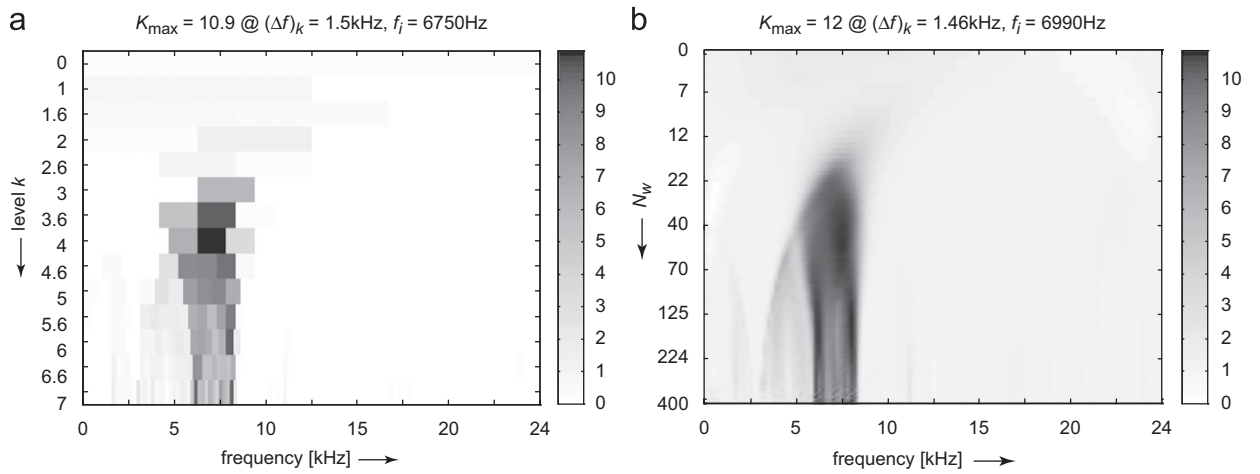

Fig. 23. Comparison of the fast kurtogram with the full kurtogram for an impulsive signal from loose parts monitoring [37].

time averages in Eq. (21). This is rather similar in principle to the FFT algorithm and even more similar to the "discrete wavelet packet transform" (DWPT). In Ref. [37], an even finer decomposition is proposed, based on a "1/3-binary tree", where each halved-band is further split into 3 other bands, thus producing a frequency resolution as illustrated in Fig. 22, in the sequence 1/2, 1/3, 1/4, 1/6, 1/8, 1/12,…,2^{-k-1}, with corresponding "scale levels" k=0, 1, 1.6, 2, 2.6,…, that enforce a similarity in notation with the DWPT transform. It should be noted that, however, in addition to providing a finer resolution than allowed by the DWPT (actually limited to integer values of k only), the proposed solution also has much better filtering characteristics as demonstrated in Ref. [37].

A comparison between the fast kurtogram and the full kurtogram of a signal is illustrated in Fig. 23. Although the former is suboptimal due to its coarser resolution, it happens to return virtually the same result as the latter in terms of transient localisation, but orders of magnitude faster, thus making it ideally suited to industrial applications.

It is pointed out in [37] that the discrete wavelet transform occupies fewer combinations than the fast kurtogram, being limited to constant percentage bandwidth, and that the DWPT gives a poorer frequency characteristic of some filters, as well as being limited to the binary tree.

Nonetheless, since the kurtogram is used to detect series of impulse responses, such as from bearing faults, and these tend to have an approximate constant damping ratio, which manifests itself in the frequency domain as a constant percentage bandwidth structure, it can also be argued that some of the combinations given by the 1/3-binary tree are unlikely, and that a 1/nth-octave wavelet analysis is adequate for seeking filter bandwidth/centre frequency combinations. For this reason Ref. [38] proposes a "wavelet kurtogram", based on non-orthogonal complex Morlet wavelets. These can have any desired bandwidth, but the sequence (in terms of octaves) 1/1, 1/2, 1/3, 1/4, 1/6, 1/8, 1/12… is often used. A number of examples are given below in Section 6.

4.2.5. *Wavelet denoising*

An alternative to spectral kurtosis methods for enhancing bearing fault signals buried in noise is to make use of wavelets. In wavelet analysis, signals are decomposed in terms of a family of "wavelets" which have a fixed shape, but can be shifted and dilated in time. The formula for the wavelet transform is

$$W(a;b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi^*\left(\frac{(t-b)}{a}\right) dt \tag{24}$$

where $\psi(t)$ is the mother wavelet and $\psi^*(t)$ its complex conjugate, translated by b and dilated by factor a. Since this is a convolution, the wavelets can be considered as a set of impulse responses of filters, which because of the dilation factor have constant percentage bandwidth properties. In principle, they are not very different from 1/nth octave filters, but with zero phase shift because the mother wavelet is normally centred on zero time. The dilation factor a is known as scale, but (its inverse) represents log frequency, as for constant percentage bandwidth filters. Wavelets give a better time localisation at high frequencies, and for that reason can be useful for detecting local events in a signal.

Wavelets can be orthogonal or non-orthogonal, and continuous or discrete [39]. Examples of orthogonal wavelets are the Daubechies dilation wavelets [40], which are compact in the time domain, but in principle infinite in the frequency domain. They tend to have irregular shapes in the time domain. Newland [39] describes complex harmonic wavelets, which are compact in the frequency domain, but infinite in the time domain. They have the appearance of windowed sinusoids (harmonic functions) and are typically of one octave bandwidth, although they can be narrower. The advantage of complex wavelets is that the imaginary part of the wavelet is orthogonal to the real part (sine rather than cosine) and thus the overall result is not sensitive to the position (phasing) of the event being transformed (it may be centred on a zero

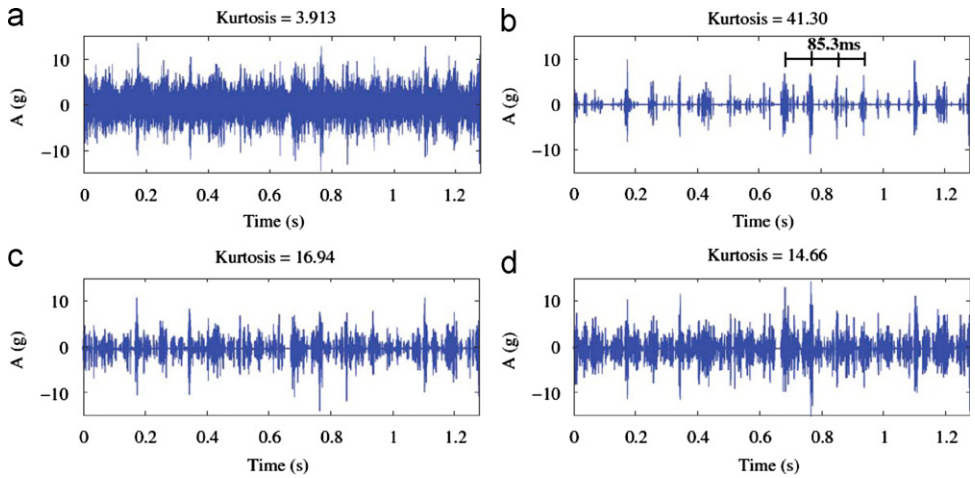

Fig. 24. Example of advanced wavelet denoising: (a) raw vibration signal; denoised signal using NeighCoeff shrink based on (b) DTCWT; (c) DWT (discrete wavelet transform), and (d) SGWT (second-generation wavelet transform). From [43].

crossing of the real part, but this would be a maximum of the imaginary part). The local sum of squares of the real and imaginary parts is a smooth function. Harmonic wavelet transforms can be efficiently produced using FFT methods [39].

Orthogonal wavelets are the most efficient to use when analysis/synthesis is to be performed (e.g. after denoising), or when the significant features of the signal are to be represented with a minimum number of parameters (e.g. as inputs to artificial neural networks). However, it is possible to obtain complete reconstruction of a signal using non-orthogonal wavelets, as long as there is some redundancy or overlap [39]. For analysis purposes, non-orthogonal wavelets such as Morlet wavelets are often more convenient, and in any case it is generally preferable to use redundant "lapped" transforms to aid visual interpretation [39].

Many authors have described the use of wavelets for detecting local faults in gears and bearings (see review in [41]). However, much of the literature on the use of wavelets for machine diagnostics does not take account of the alternative methods described in this tutorial, or makes errors in doing so. The most common error is to show results only in the time domain, on very short records, quite often shorter than the longest modulation period (cage rotation for rolling element faults). Often it is claimed that wavelet analysis is superior to envelope analysis, but it does not appear to have been realised by many authors that the squared modulus of the wavelet coefficients is effectively a squared envelope signal, and much diagnostic information can be gleaned by frequency analysis of such squared envelope signals. This is particularly the case for the complex wavelets, such as the harmonic wavelets just discussed, and the complex Morlet wavelets used in the wavelet kurtogram [38]. As discussed in the next section, frequency analysis of the squared envelope signal often finds fault repetition and modulation patterns which may not be easily seen in time signals, in particular when the modulation is so strong that the pulses are only excited when the faults are in the load zone, and then with greatly varying amplitude.

However, one of the primary applications of wavelets is in denoising of signals in both time and frequency domains simultaneously. Most wavelet denoising is an extension of the work of Donoho and Johnstone [42], who defined two types of thresholding to remove noise, this being defined as any components with amplitude less than a certain threshold value. In so-called "hard thresholding" the retained components are left unchanged, but in "soft thresholding", the noise estimate (the threshold value) is subtracted from them also (symmetrical treatment of positive and negative values). More advanced methods are continually being developed.

Fig. 24 shows the result of denoising acceleration signals from a gearbox with a simulated tooth root crack, using a proposed new "dual-tree complex wavelet transform (DTCWT)" and "NeighCoeff shrinkage" for thresholding, and compares it with other wavelet transforms [43].

On occasion, it is possible that the wavelet denoising might perform better than SK methods.

5. Envelope analysis

As shown in the Introduction (Section 1), the spectrum of the raw signal often contains little diagnostic information about bearing faults, and over many years it has been established that the benchmark method for bearing diagnostics is envelope analysis, where a signal is bandpass filtered in a high frequency band in which the fault impulses are amplified by structural resonances. It is then amplitude demodulated to form the envelope signal, whose spectrum contains the desired diagnostic information in terms of both repetition frequency (ballpass frequency or ballspin frequency) as well as modulation by the appropriate frequency at which the fault is passing through the load zone (or moving with respect to the measurement point).

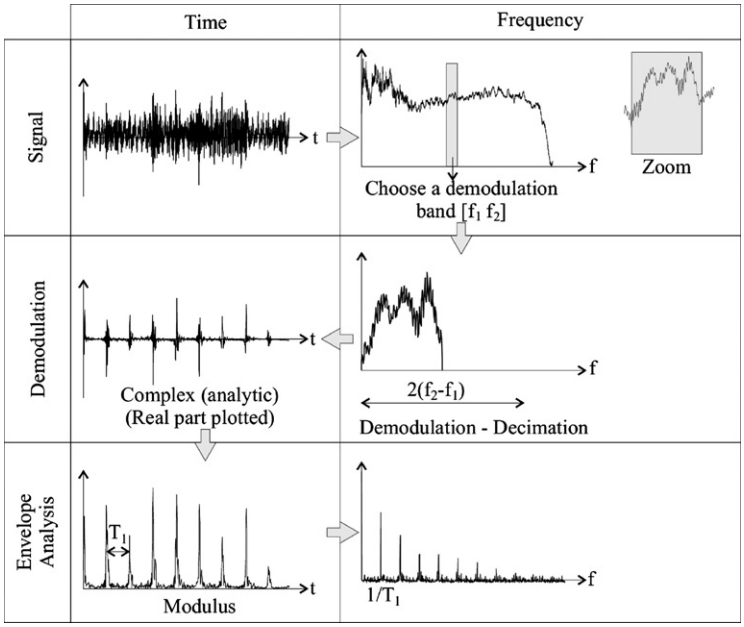

Fig. 25. Procedure for envelope analysis using the "Hilbert transform" method [45].

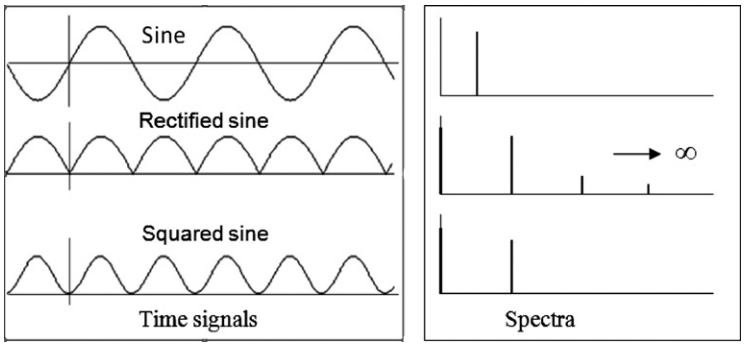

Fig. 26. Potential aliasing given by squaring and rectifying a sinusoidal signal. With just squaring, but not rectification, aliasing can be avoided by doubling the sampling frequency before squaring.

However, the envelope analysis technique was devised more than 30 years ago, e.g. [7], and used analogue techniques with inherent limitations. Considerable improvement can be made by taking advantage of digital processing techniques, rather than slavishly following the analogue method in digital form.

A number of benefits arise from performing the amplitude demodulation using "Hilbert transform" techniques where a one-sided spectrum (positive frequencies only) is inversely transformed to the time domain. This gives a complex time signal (a so-called "analytic signal") whose imaginary part is the Hilbert transform of the real part [44]. An immediate benefit is that the extraction of the section of spectrum to be demodulated is effectively by an ideal filter, which thus can separate it from adjacent components that might be much stronger (e.g. gearmesh frequencies). This was not always possible with analogue filters, and real-time digital filters suffer from the same restrictions on filter characteristic. The application to envelope analysis is shown in Fig. 25 (from [45]).

Fig. 25 depicts the envelope as the modulus of the analytic signal obtained by inverse transformation of the selected one-sided frequency band. In fact, it was shown in [46] that it is preferable to analyse the squared envelope signal rather than the envelope as such. The reason for this is simply explained in Fig. 26, which compares the spectra of a rectified and a squared sinusoid. It should be noted that mathematically the envelope of a signal is the square root of the squared envelope, and likewise a rectified signal is the square root of the squared signal. The square root operation introduces extraneous components that are not in the original squared signal, and which cause masking of the desired information. In Fig. 26 it is seen that the rectified signal has sharp cusps, requiring harmonics extending to infinity to reproduce them. Since the whole operation is done digitally, it is not possible to remove these high harmonics by lowpass filtration (as it

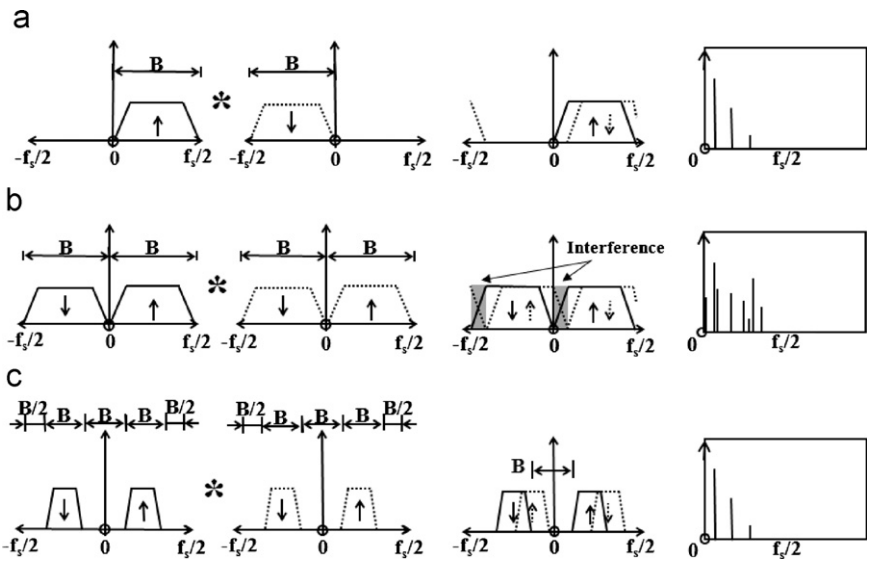

Fig. 27. Generation of spectrum of squared envelope (or signal) for three cases: (a) analytic signal; (b) equivalent real signal; and (c) frequency shifted real signal downward arrow indicates complex conjugate.

was for example with an analogue rectifier), and they alias into the measurement range, causing masking. Note that since the squaring doubles the frequency content of a signal, the sampling frequency should be doubled before the signal is squared or rectified digitally, although as will be seen, this corresponds to the zero padding in Fig. 25 when the analytic signal is processed.

Finally, the benefits of using the one-sided spectrum are illustrated in Fig. 27 (from [46]). If the analytic signal (from the one-sided spectrum) is termed $x_a(t)$, its squared envelope is formed by multiplication with its complex conjugate, and the spectrum of the squared envelope will be the convolution of the respective spectra. Thus:

$$\Im\{x_a(t)x_a^*(t)\} = \Im\{x_a(t)\} * \Im\{x_a^*(t)\} = X_a(f) * X_a^*(-f) \tag{25}$$

When this convolution is carried out, as illustrated in Fig. 27(a), the result only gives difference frequencies, for example sideband spacings, which contain the desired modulation information.

However, for the equivalent real signal $x(t)$, the spectrum of its squared value is simply the convolution of $X(f)$ (the spectrum of $x(t)$) with itself. This is illustrated in Fig. 27(b), and is seen to give the same difference frequency components, but mixed with sum frequencies (the difference of a positive and a negative frequency), which contain no diagnostic information, and only serve to mask the true result.

As illustrated in Fig. 27(c), this interference can be avoided with a real-valued signal, as long as it is frequency shifted so as to introduce zero padding around zero frequency as well as around the Nyquist frequency. This effectively means that the sampling frequency has to be doubled for the same demodulation band, so that transform sizes must be twice as big for the same problem. When Ref. [46] was written, the primary means of separating gear and bearing signals was using SANC (Section 3.3), which required real-valued signals, but the frequency domain based DRS method (Section 3.4) can make use of one-sided spectra, and thus avoid this complication.

Ref. [46] showed that even where the power of the masking noise (random or discrete frequency) was up to three times the power of the bearing signal, in the demodulation band, it was still advantageous to analyse the squared envelope. Using spectral kurtosis, it is usually possible to find a spectrum band where the signal/noise ratio of the bearing signal is much higher.

6. A semi-automated bearing diagnostic procedure

In [47], a method was proposed for diagnosing bearing faults that was successful for a wide range of cases, from high speed gas turbine engine bearings to the main bearing on a radar tower, with a rotational period of 12 s. It can be said to be semi-automated because only a small number of parameters have to be adjusted for each case, these corresponding to, and including, the dimensions and speed of the bearing. As shown in Fig. 28, it combines a number of the techniques described in earlier sections.

It is generally a good idea to start with order tracking (Appendix B), as the separation of discrete frequency and random components will not always be possible unless this is done. Ref. [45] describes a case where it was not possible to use DRS to separate gear and bearing signals until order tracking had been carried out. No tacho or shaft encoder signal was available, but it was found possible to extract the instantaneous speed information by phase demodulation of a number of

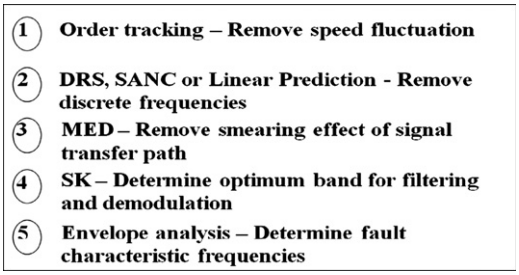

Fig. 28. Semi-automated procedure for bearing diagnostics.

Table 1
Planet bearing frequencies.

Fault	BPFO	BPFI	Cage	Roller
Frequency (Hz)	77.1	117.8	9.8	37.0

gearmesh frequencies, these being phase-locked to shaft speed. The best mapping of shaft angle vs time was given by averaging a small number of estimates with a similar appearance. In this case the random speed variation was only 0.5% peak-to-peak (1203–1209 rpm).

For separation of discrete frequency and random components (e.g. gear and bearing signals) the best choice is generally DRS (Section 3.4) as it poses the minimum problems with regard to choice of parameters. The size of transform N should span 10–20 periods of the minimum frequency to be removed (e.g. the lowest shaft speed) and the delay should be at least three times the correlation length of the bearing signal. Assuming 1% slip, this would correspond to about 300 periods of the centre frequency of the demodulated band. Determining the latter might require one iteration as it is best decided after the SK procedure.

MED need only to be applied for high speed bearings, where the impulse response of the bandpass filtered resonance is of comparable length to the spacing of the bearing fault pulses (BPFI would normally be the highest fault frequency and thus the shortest spacing). This can perhaps best be decided by trial and error, based on whether MED gives an increase in SK.

The optimum band for demodulation should be chosen using a fast kurtogram procedure. Note that the kurtogram is sensitive to large random pulses which may be present in some realisations of a signal. If the final envelope spectrum does not reveal periodic components, even though the SK is high, it should be checked whether such random impulses from an extraneous source are dominant in certain frequency bands.

In the final envelope analysis, it should be recognised that modulating effects are important to the diagnosis. In general, inner race faults would be modulated at shaft speed, and rolling element faults at cage speed. For unidirectional load, an outer race fault would not be modulated, but modulation at shaft speed can occur because of significant unbalance or misalignment forces, and modulation at cage speed can result from variations between the rolling elements. Note that with planetary gear bearings, it is the inner race that is fixed with respect to the load, and so inner race faults tend not to be modulated, whereas the signals from outer race faults are modulated by the frequency at which they pass through the load zone. Since planet gears are analogous to rolling elements in a bearing, the modulation frequency can be calculated by an equation similar to that for BSF.

Ref. [47] illustrates the general procedure by its application to three very different case histories, so a brief summary of those results is given here.

6.1. Case history 1—helicopter gearbox

A test was carried out on a helicopter gearbox test rig at DSTO (Defence Science and Technology Organisation) Melbourne, Australia, where it was run to failure under heavy load. The signals were analysed blind, with no indication of the type of failure. Frequencies corresponding to the planet bearing (which actually failed) are given in Table 1, although all other potential bearing frequencies had to be calculated.

Initial analyses of the signals, even at the end of the test where bearing failure was indicated by the growth of wear debris, showed no indication of the fault in either the time signal or spectrum. The latter was dominated by gear components (harmonics of the main meshing frequencies and their sidebands) over the whole frequency range up to 20 kHz. The kurtosis of the raw signal was −0.6, roughly like noise.

The procedure of Fig. 28 was applied, and in this case the discrete frequency components were removed using linear prediction. Fig. 29 compares the residual of the linear prediction process with the original signal, and it is seen that it has

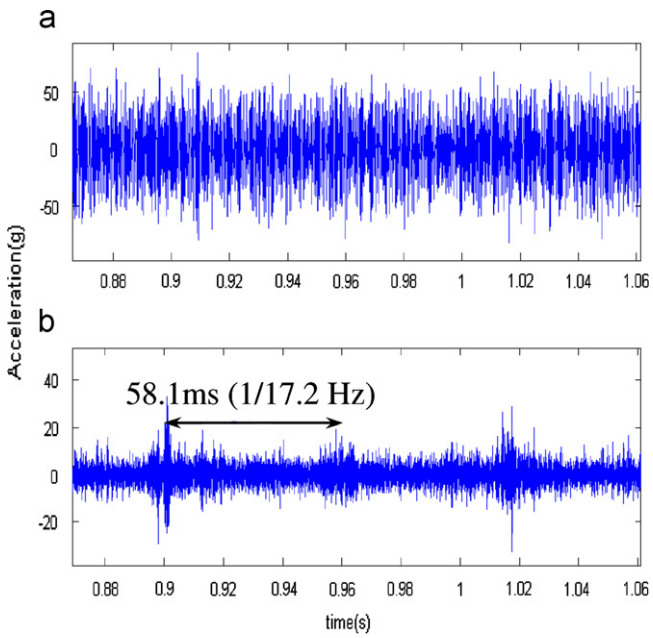

Fig. 29. Time signals (one rotation of the carrier): (a) order tracked signal and (b) residual signal—passage of the three planets is seen.

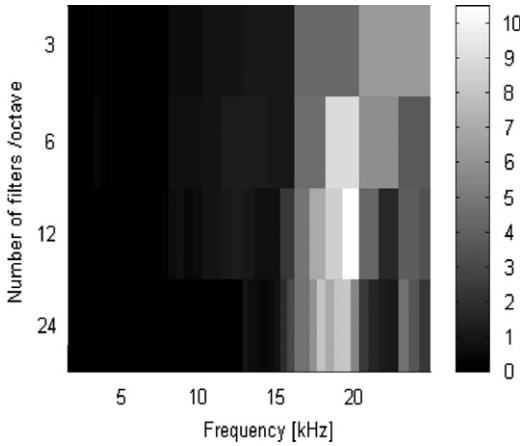

Fig. 30. Wavelet kurtogram for 4 filter banks; namely (3, 6, 12, and 24) filters/octave.

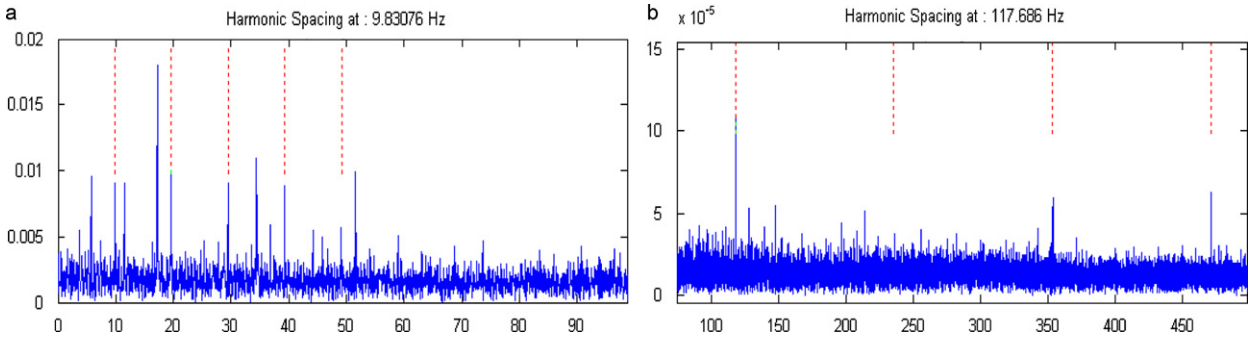

Fig. 31. Squared envelope spectra showing two fault frequencies: (a) cage speed (9.8 Hz) and (b) BPFI (117.7 Hz).

become slightly more modulated, (the kurtosis increased to 2.2), but the three "bursts" are related to the passage of the planets (period 58.1 ms).

The wavelet kurtogram (Section 4.2.4) of the residual signal (of Fig. 29(b)) was next produced using a range of filter banks (3, 6, 12, 24 filters/octave) and the results are shown in Fig. 30. The maximum SK of 12 was obtained using 12 filters/octave (centre frequency of 18,800 Hz and bandwidth 1175 Hz).

Finally, Fig. 31 shows the squared envelope spectra in two frequency ranges, for the last measurement. Fig. 31(a) shows a strong pattern of harmonics spaced at the cage speed of a planet bearing. This basically gives an indication that there is a variation for every rotation of the cage. This can be a cage fault, but is often an indicator of variation between the rolling elements. Fig. 31(b), in a somewhat higher frequency range, shows a strong component corresponding to the ballpass frequency, inner race (BPFI). Because it is a planet bearing, no modulation is expected for an inner race fault, and no modulation sidebands are found in the envelope spectrum. When the gearbox was disassembled, severe spalling was found on the inner race of one planet bearing and three rollers had minor spalls, explaining the modulation at cage speed.

Fig. 32 shows that the kurtosis of the optimally filtered signal follows the same trend as the cumulated oil wear debris, and thus has the potential for making a prognosis of remaining useful life. Note that the wear debris gives no indication of the source, whereas the SK corresponds to one of the planet bearings. In this case, MED was tried but gave no benefit, because the failure was in the low speed part of the gearbox.

6.2. Case history 2—high speed bearing

Measurements were made on a bearing test rig at the FAG bearing company in Germany, on which bearings are tested to failure. At several points through their life, the bearings are dismantled and inspected. The bearing being tested is for a high speed application, and is thus tested at 12,000 rpm, typical of gas turbine bearings. The accelerometer used to capture data was mounted using a magnet, and it is suspected (from inspection of the spectra) that the mounting resonance frequency was of the order of 12 kHz.

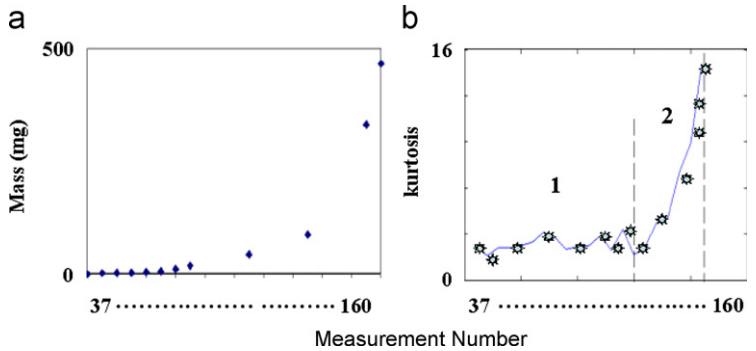

Fig. 32. (a) Accumulated metal wear debris and (b) kurtosis of the filtered signal.

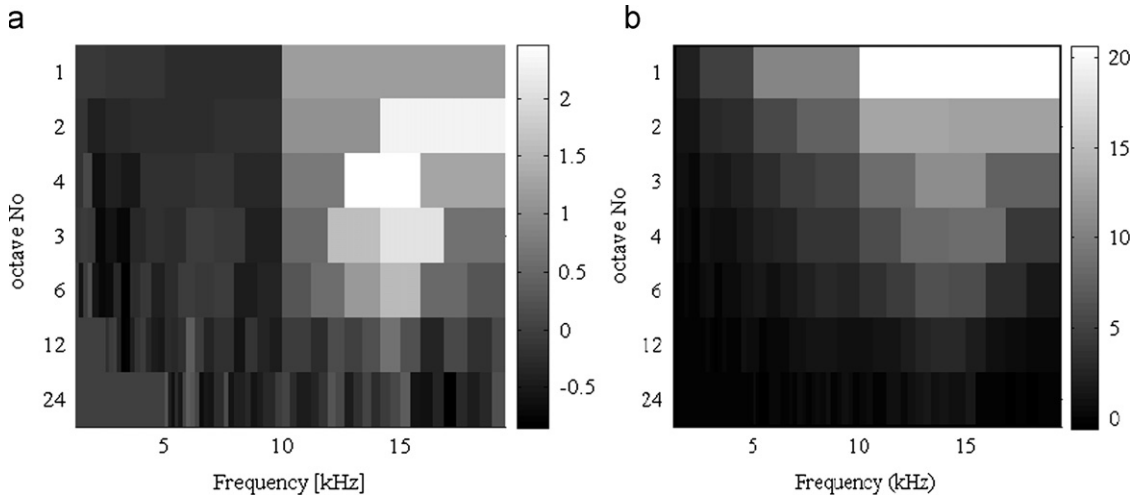

Fig. 33. Wavelet kurtograms: (a) before application of MED and (b) after MED.

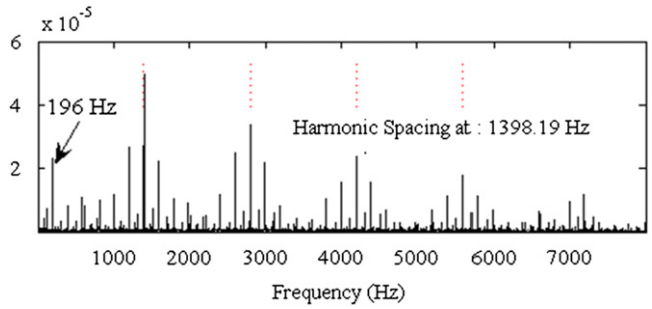

Fig. 34. Envelope spectrum of signal of Fig. 16(c) showing harmonics of BPFI 1398 Hz, and harmonics and sidebands spaced at shaft speed 196 Hz.

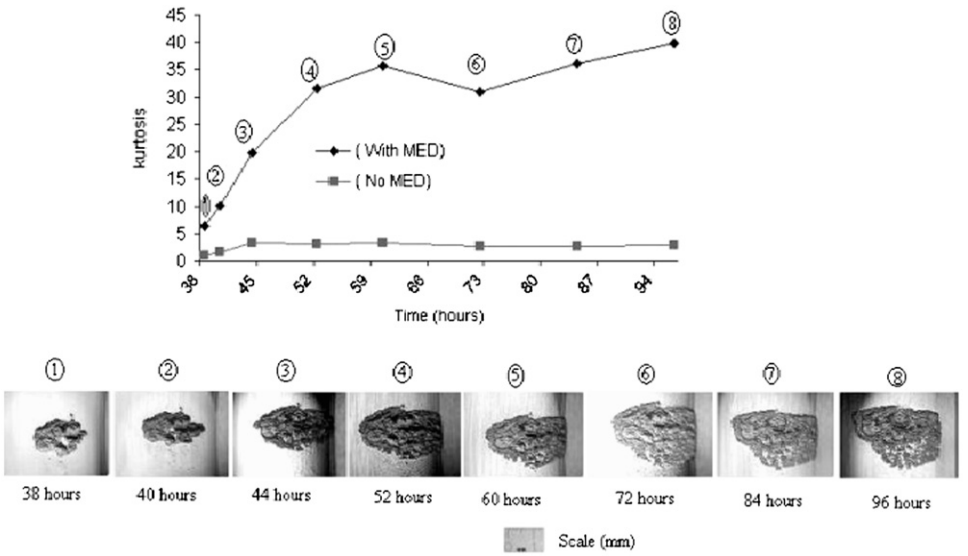

Fig. 35. The development of kurtosis with and without MED, compared against fault size.

This is the case depicted in Fig. 16, where MED gave a considerable improvement in the impulsiveness of the signal because the individual impulse responses were overlapping. The wavelet kurtograms for the signals before and after the application of the MED technique (Fig. 33) show that the SK has been increased from 2.5 to 20, making the impulsiveness apparent.

Fig. 34 shows an envelope spectrum for a late stage of development of the fault, demodulated in the band indicated in Fig. 33(b). It is a typical envelope spectrum for an inner race fault, with a series of harmonics of BPFI (1398 Hz), together with low harmonics of, and sidebands spaced at, shaft speed (196 Hz).

Fig. 35 shows the benefits in using the MED technique to sharpen the impulses. As in the first case history, the SK now correlates well with the fault severity, giving the possibility of using it for prognosis.

6.3. *Case history 3—radar tower bearing*

Measurements were received from before and after a main bearing change on a radar tower. The radar driving system consists of a motor, a gearbox and a spur pinion/ring-gear combination. The motor runs at 1800 rpm (30 Hz) and is connected to a three-stage reduction gearbox. The final tower rotation period is 12 s (0.082 Hz). The ring gear used to drive the tower is an integral part of the bearing and so was changed at the same time, but the drive pinion was left unchanged. The reason for the change was an increase in noise (which may have been partly due to wear of the gears), but the current analysis clearly shows evidence of bearing faults in the old bearing. Compared with the two previous cases, the speed is orders of magnitude less, but even so the same analysis procedure could be used with little operator intervention. Because of the slow speed, the fault impulses were well separated, so there was no need to use MED filtering.

The spectra from before and after replacement gave no indication of the bearing fault, but the raw time signal (Fig. 36(a)) did indicate the fault, as a series of impulses protruding from the background signal. Application of DRS

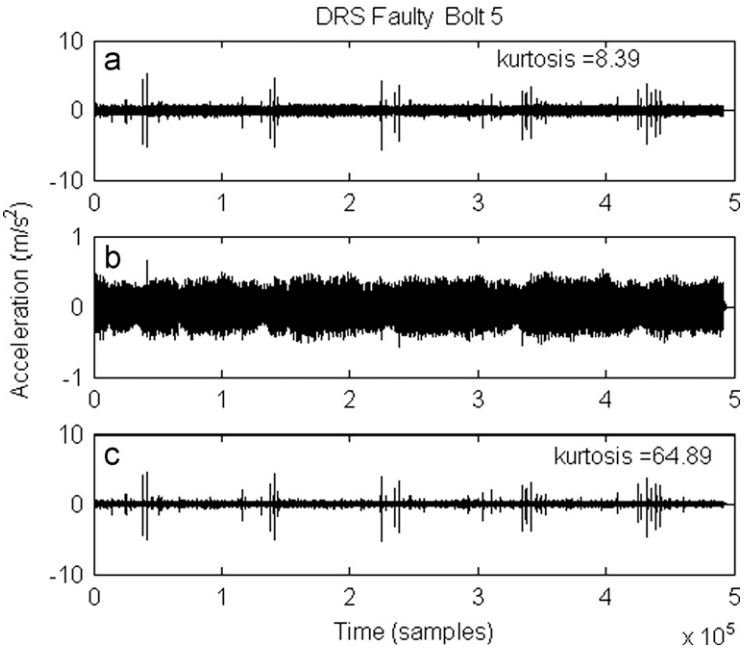

Fig. 36. Results of applying DRS to signal with faulty bearing: (a) original signal; (b) deterministic part; and (c) random part.

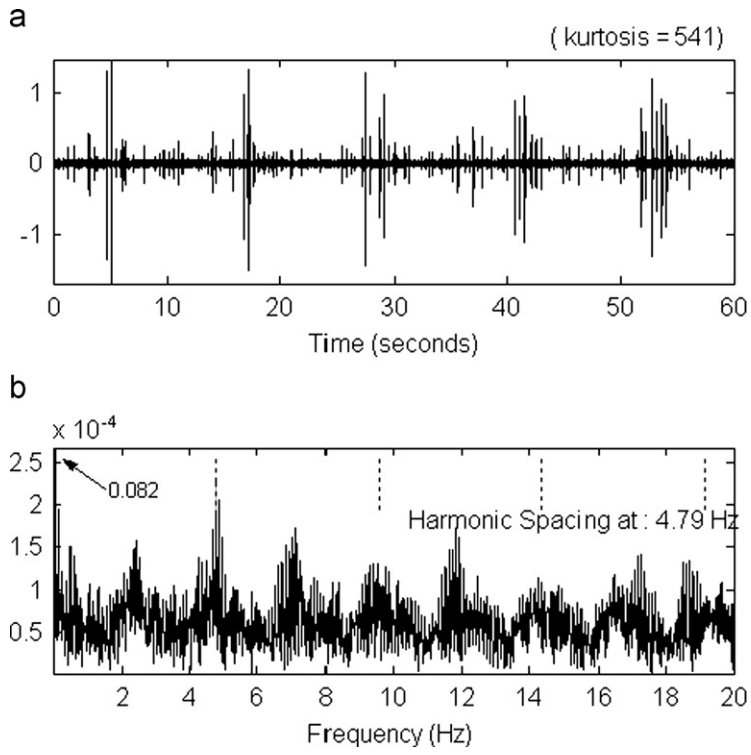

Fig. 37. Effect of optimal SK filtering: (a) time signal and (b) envelope spectrum.

(Fig. 36(b), (c)) showed that the latter was dominated by deterministic components, mainly from the gears (which dominated the spectra). The local impulses visible in the time signal from the old bearing (before and after application of DRS) were not present in the signals from the new bearing. Removal of the gear signal increased the kurtosis from 8.4 to 64.9. Even without using SK, envelope analysis of the signal in Fig. 36(c) revealed the bearing fault frequency (4.79 Hz) with modulation sidebands at rotational speed 0.082 Hz. Interestingly, it is not possible to determine whether the fault(s)

were in the inner or outer race, partly because this is a thrust bearing, and so BPFI and BPFO are the same ($\varphi = 90^\circ$ in Eqs. (1) and (2)).

Moreover, there is a reason why faults in both the fixed and moving races would be modulated at shaft speed, the former because the load was somewhat eccentric, and thus rotating around the bearing, and the latter because of the varying path length to the transducer.

Fig. 37 shows the result of using a wavelet kurtogram to apply the optimum bandpass filter and extract the signal coming from the bearing fault(s). This was found to be a filter of bandwidth 517 Hz centred on 2755 Hz. The kurtosis has increased to the remarkable value of 541. This makes it clear that care must be taken in using the kurtosis directly as an indicator of fault severity; it is affected not only by the size of the individual fault pulses, but also by the spacing between them. If this machine was run at twice the speed, for example, it is likely that the kurtosis would be approximately halved. Thus, in evaluating the kurtosis corresponding to machine faults, account should be taken of the ratio of fault repetition frequency to typical resonance frequencies excited. The damping ratio of the impulse responses also has an influence on the kurtosis (hence the benefits of MED).

The harmonic cursor in Fig. 37 is set on the ballpass frequency of 4.79 Hz, but it will be seen that there are components at multiples of half this frequency also. The reason for the appearance of the half frequency is that these bearings have a special construction with the races in a *V*-shape, and the 118 rollers (with the same length and diameter) being alternately mounted with $\pm 45^\circ$ orientation, so that only each alternate roller contacts one side of the *V*. Thus a fault on one side would give impulses at half the BPF. As mentioned above, there are sidebands spaced at rotational speed (0.082 Hz) around the harmonics of (half) BPF, as well as low harmonics of 0.082 Hz.

It is evident that the semi-automated procedure used to extract the bearing signal of Fig. 37(c) would have detected the bearing fault at a very early stage, long before it protruded above the background gear signal, as in Fig. 36(a).

Appendix A. Cyclostationarity and spectral correlation

Vibration signals are usually classified as deterministic (i.e., whose behaviour can be described exactly by an equation) or random (i.e., whose behaviour cannot be predicted exactly), or a combination of both. Deterministic signals are further categorised as periodic and non-periodic, and random signals as stationary and non-stationary. Cyclostationarity represents a further category including signals which, although not necessarily periodic, are produced by a hidden periodic mechanism. This includes periodic signals as a special case, but also stationary signals (infinite period) and non-stationary signals which exhibit periodicity after passing through a non-linear transform. As such, cyclostationarity happens to comprise most signals generated by rotating and reciprocating machines. As briefly reviewed below, cyclostationary signals endow a rigorous mathematical framework with many properties that which considerably extend the scope of traditional signal processing.

Strictly speaking, an *n*th order cyclostationary signal is one whose *n*th order statistics are periodic or, equivalently stated, one which produces a peak in its Fourier transform after passing through any non-linear transformation involving *n*th power. The simplest example is a first-order cyclostationary signal, say x, which is just a periodic signal p embedded in additive stationary noise n. Its first-order statistics, i.e. its mean value in the ensemble average sense, then reproduces the periodic component p unaltered: $\mathbb{E}\{x\} = \mathbb{E}\{p+n\} = p$. Similarly, a second-order cyclostationary signal is one whose autocorrelation function

$$R_{xx}(t,\tau) = \mathbb{E}\{x(t-\tau/2)x(t+\tau/2)\} \tag{A1}$$

is a periodic function of time, i.e. $R_{xx}(t,\tau) = R_{xx}(t+T,\tau)$ (this is not to be confused with the stationary enforced autocorrelation function computed as the time averaged $\langle R_{xx}(t,\tau)\rangle = R_{xx}(\tau)$). A simple example is provided by a white noise modulated by a periodic amplitude, as illustrated in Fig. A1(a). Note that, in spite of the periodic modulation, this is still a completely random signal. The autocorrelation function, by virtue of the second-order non-linearity it introduces,

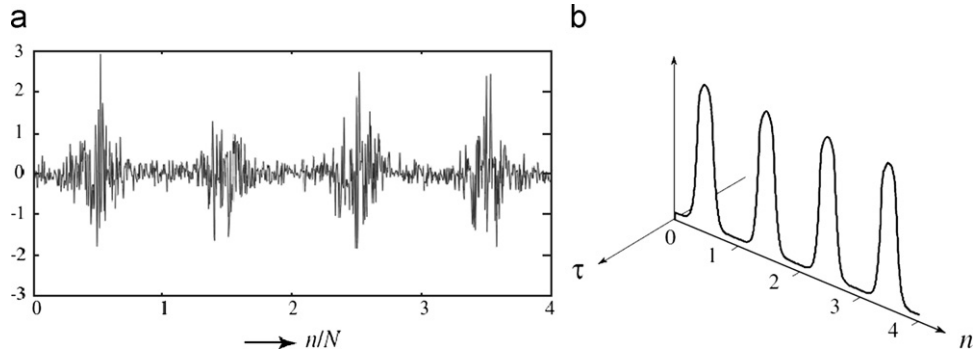

Fig. A1. Example of amplitude modulated white noise (from Antoni [10]): (a) time signal over four periods of cyclic frequency; (b)two-dimensional autocorrelation function vs time (sample) n and time lag τ.

succeeds in revealing the hidden periodicity as shown in Fig. A1(b). For time-lag $\tau=0$, it returns the instantaneous power of the signal which is seen to flow periodically as a function of time. For other values of the time-lag the autocorrelation function is zero because the carrier signal is white noise, but any dependence on τ—compatible with an autocorrelation function—will be possible in general depending on the colour of the signal.

Higher-order types of cyclostationary signals are defined following similar lines as for first and second orders, although not reviewed here. Indeed, restriction to first and second orders of cyclostationarity appears good enough for many practical purposes.

A.1. Spectral correlation

When available for all values of t and τ, the autocorrelation function in Eq. (A1) contains all the information about a second-order cyclostationary signal, yet displaying it in the frequency domain usually provides more insight into the structure of the signal. Since the autocorrelation function is a function of two variables, a two-dimensional Fourier transform is performed, giving rise to the so-called "spectral correlation", whose name will become obvious later:

$$S_{xx}(\alpha,f)=\lim_{W\to\infty}\frac{1}{W}\int_{\mathbb{R}}\int_{-W/2}^{W/2}R_{xx}(t,\tau)e^{-j2\pi(f\tau+\alpha t)}dt\,d\tau. \tag{A2}$$

This involves two frequency variables with very different physical meanings. Frequency f, as being the dual of time-lag τ, indicates the frequency of the carrier signal. Frequency α, as being the dual of time t, indicates the frequency of the modulation and, accordingly, is usually named the "cyclic frequency" or the "modulation frequency". Hence, the spectral correlation may be interpreted as giving the strength of the elementary waves in signal x carried and modulated at all possible combinations (α,f).

Because the autocorrelation function of a second-order cyclostationary signal is transient in time-lag τ and periodic in time t, the corresponding spectral correlation is continuous in f but discrete in α, thus returning a very distinctive signature as illustrated in Fig. A2 for the signal of Fig. A1(a). Since this signal has a white content, the corresponding spectral correlation is flat in the f direction, but since the modulation is periodic, it is non-zero only at those cyclic frequencies which are multiple of the fundamental modulation frequency. It is noteworthy that the spectral correlation at $\alpha=0$ returns the usual power spectral density, thus clearly demonstrating on this example that classical spectral analysis fails to reveal the hidden periodicity of a cyclostationary signal.

The same result can be obtained by correlating the spectrum with itself, hence the name "spectral correlation". This can be explained by considering that for an amplitude modulated noise signal as in Fig. A1, each spectral line has a pair of sidebands spaced at each modulating frequency or harmonic (from convolution of each line with the spectrum of the modulating function). These are not visible because the basic spectrum is white. However, when the spectrum is correlated with itself, this correlation becomes apparent whenever the spectrum shift corresponds to a multiple of the cyclic frequency, suddenly giving a finite value, while being zero for other frequency shifts. Another way of explaining it derives from the definition of Eq. (A2). When the first transform is made with respect to τ, the result at each time is the instantaneous power spectrum (i.e. the product of a complex spectrum with its complex conjugate). When these products are transformed with respect to t, it gives a convolution in the frequency domain, and since complex conjugation of a spectrum corresponds to a reversal of the frequency axis, a convolution of complex conjugates corresponds to a correlation.

Thus, the alternative way to calculate the spectral correlation can be expressed as

$$S_{xx}(\alpha,f)=\lim_{T\to\infty}\mathbb{E}\{X_T(f+\alpha/2)X_T^*(f-\alpha/2)\} \tag{A3}$$

where $X_T(f)$ stands for the Fourier transform of signal $x(t)$ over an interval of duration T. Formula (A3) is probably most convenient when it comes to estimate the spectral correlation from finite-length measurements since it is easily amenable

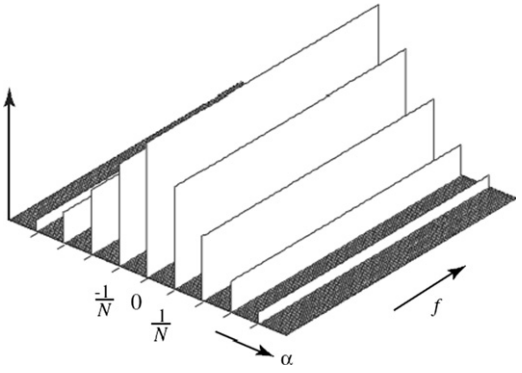

Fig. A2. Spectral correlation for the case of Fig. A1(a) (from [10]).

to a Welch-type estimator where the ensemble average is replaced by an average over blocks of data, with no need for synchronisation on the (unknown) cyclic frequencies α's. Indeed, the spectral correlation is a very good tool to measure precisely the unknown cyclic frequencies in a signal, in particular in the case of localised faults in rolling element bearings, where the fault frequency has some random variation and/or uncertainties related to the value of the load angle ϕ in Eqs. (1)–(4).

Further detailed information on how to estimate and interpret spectral correlations and other functions of interest in cyclostationary analysis is given in [10,11].

One last subtlety arises when the spectral correlation is used to analyse signals which consist of a mixture of first and second-order cyclostationarities. Since the autocorrelation of a periodic signal is both periodic vs time and time-lag, it produces a spectral correlation function discrete in both f and α directions (a "bed of nails"). In some circumstances, the contribution of the first-order part may considerably complicate the visual interpretation of the spectral correlation, and even mask the contribution of the second-order part. In such a situation, it is advisable to analyse the two parts independently after separating them with one the algorithms described in Section 3. In many instances, this will also help in differentiating phenomena of distinct physical origins, such as gear and bearing faults, as illustrated on several occasions in this paper.

A.2. Spectral correlation and envelope spectrum

It can be shown (see inset) that the integral of the spectral correlation over all frequency f is the Fourier transform of the expected value of the squared signal, and so is effectively the spectrum of the squared envelope. An example is shown in Fig. 6 of the application to an inner race bearing fault. Section 2.2 contains a discussion of when the full spectral correlation gives advantages over the envelope spectrum.

Integration of spectral correlation over f:

$$\int S_{xx}(\alpha,f)df = \iint \mathbb{E}\{x(t+\tau/2)x*(t-\tau/2)\}\left(\int e^{-j2\pi f\tau}\,df\right)e^{-j2\pi\alpha t}\,dt\,d\tau = \iint \mathbb{E}\{x(t+\tau/2)x*(t-\tau/2)\}\delta(\tau)\,d\tau e^{-j2\pi\alpha t}\,dt$$
$$= \int \mathbb{E}\{x(t)x*(t)\}e^{-j2\pi\alpha t}\,dt = \underset{t\to\alpha}{\mathfrak{I}}\{E\{|x(t)|^2\}\}$$

the spectrum of the squared envelope

A.3. Wigner–Ville spectrum

Another quantity of interest with cyclostationary signals is the time–frequency representation obtained by Fourier transforming the autocorrelation with respect to the time-lag variable. This defines the so-called Wigner–Ville spectrum

$$W_{xx}(t,f) = \int_{\mathbb{R}} R_{xx}(t,\tau)e^{-j2\pi f\tau}\,d\tau \tag{A4}$$

which returns, at all time t, an instantaneous power spectrum. It is obvious from Eq. (A4) that the Wigner–Ville spectrum of a cyclostationary signal is a periodic function of time. It is also clear from the presence of the ensemble averaging operator in the definition of the autocorrelation function—see Eq. (A1)—that the Wigner–Ville spectrum is equal to the mean of the well-known Wigner–Ville distribution (WVD), i.e. the Fourier transform of the interaction $x(t+\tau/2)x*(t-\tau/2)$. An important consequence of this is that interference terms between incoherent components in the signal average to zero in the Wigner–Ville spectrum while keeping the time–frequency unaltered, a quite unique property in time–frequency analysis.

The Wigner–Ville spectrum for the signal of Fig. A1 is illustrated in Fig. A3. It returns a flat power spectrum at any time instant, which is consistent with the signal being white, with a periodic time modulation proportional to the local value of the instantaneous variance $R_{xx}(t,0)$—the curve at time-lag $\tau=0$ in Fig. A1. Applications of the Wigner–Ville spectrum to vibration signals and some discussion on how to estimate it in practice are discussed in Refs. [10,48].

Appendix B. Order tracking

In analysing rotating machine vibrations it is often desired to have a frequency x-axis based on harmonics or "orders" of shaft speed. This can be to avoid smearing of discrete frequency components due to speed fluctuations, or can be to see how the strength of the various harmonics changes over a greater speed range, for example as they pass through various resonances. If a constant amplitude signal which is synchronous with the rotation of a shaft, for example, is sampled a fixed number of times per revolution, the digital samples are indistinguishable from those of a sinusoid, and thus give a line spectrum, whereas if normal temporal sampling is used the spectrum spreads over a range corresponding to the variation

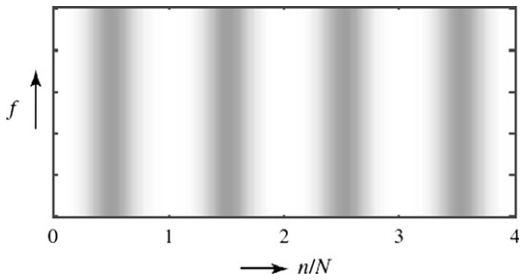

Fig. A3. Wigner–Ville spectrum for the case of Fig. A1(a) (from [10]).

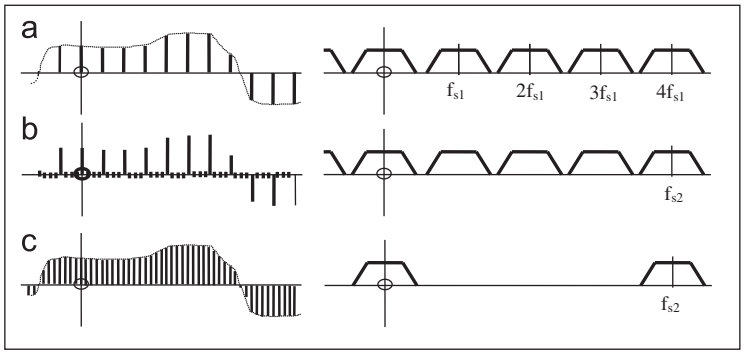

Fig. B1. Digital resampling with four times higher sampling frequency: (a) signal sampled at f_{s1} and its spectrum; (b) addition of zeros which changes sampling frequency to f_{s2}; and (c) lowpass filtration and rescaling.

in shaft speed. Thus, for order analysis it is necessary to generate a sampling signal from a tacho or shaft encoder signal synchronous with shaft speed. It is sometimes possible to use a shaft encoder mounted on the shaft in question to directly provide a sampling signal, but more often the latter has to be generated electronically or numerically.

Formerly, this was done using a phase-locked loop to track the tacho signal and then generate a specified number of sampling pulses per period of the tracked frequency. However, an analogue phase-locked loop has a finite response time and cannot necessarily keep up with random speed fluctuations such as occur with an internal combustion engine from cycle to cycle. The best method is to digitally resample each record based on the corresponding period of the tacho signal, so as to achieve sampling for uniform increments in shaft rotation angle. This is known as "angular resampling". Angular resampling can be done in a number of ways, based on digital interpolation. One way is simply to increase the sample rate by a large factor (say 10), and then select the nearest sample to each theoretical interpolated position. Increasing the sample rate by an integer factor can be achieved in two ways. In the time domain, it can be done by inserting the appropriate number of zeros in between each actual sample, and then applying a digital lowpass filter to limit the frequency range to the original maximum, thus smoothing the curve (it will also require rescaling proportional to the resampling factor). Resampling by a factor of 4 is illustrated in Fig. B1.

The same result can be achieved in the frequency domain by padding the FFT spectrum with zeros in the centre (i.e. around the Nyquist frequency) and then inverse transforming the increased (2-sided) spectrum to the same increased number of time samples. Note that the record length in seconds is the reciprocal of the frequency line spacing in Hz, which is not affected by the zero padding. This latter procedure can also be used to resample a record consisting of an integer number of samples to another (though greater) integer number, and is the basis of the Matlab® function INTERPFT. In general, more accurate interpolation, not limited to a ratio of integer numbers, can be achieved by fitting a curve to a group of samples (e.g. two for a linear curve, three for a quadratic, etc.) and then calculating the value of the polynomial at the interpolated positions. The accuracy of the interpolation can be judged by considering that the interpolation in the time domain corresponds to a multiplication in the frequency domain by a filter characteristic which in general aliases back into the measurement range. For example, choosing the nearest sample value is the same as convolving the original samples with a rectangular function of width equal to the sample spacing, while a linear interpolation corresponds to a convolution with a triangular function of base width twice the sample spacing (the convolution of the rectangular function with itself). In the former case the filter characteristic is a sinc(x) function with zeros at multiples of the sampling frequency, while in the latter case it is the square of the sinc(x) function, which has much smaller sidelobes. However, the sidelobes folding back into the measurement range are not the only source of error. The Fourier transform of the convolving function is always a lowpass filter, which reduces the amplitude of some components in the measurement range. The square of the sinc(x) function, for example, has a more pronounced lowpass filtering effect than the sinc(x) function itself, even though it has smaller sidelobes. These considerations have been discussed in detail by McFadden [49] and Fig. B2 compares the

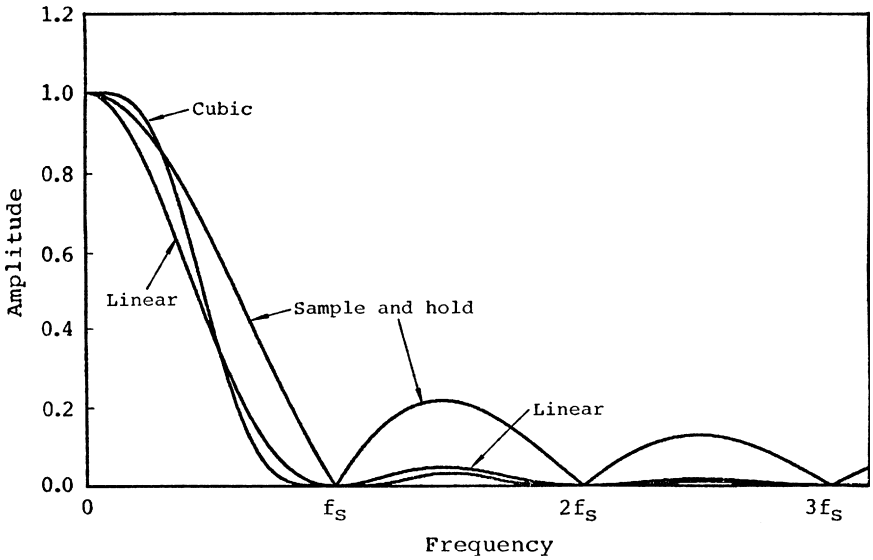

Fig. B2. Comparison of frequency characteristics for interpolation at different orders (from [49]).

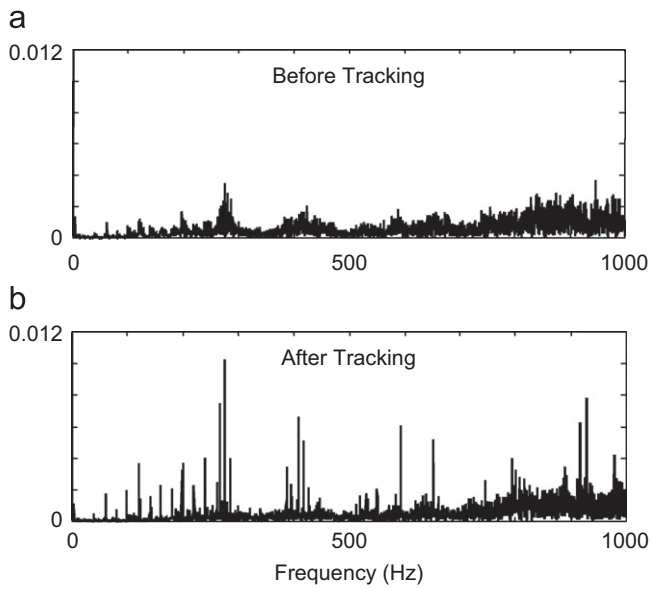

Fig. B3. Use of tracking to avoid smearing of shaft speed related components.

frequency characteristics associated with different orders of polynomial interpolation. From this it can be seen that cubic interpolation not only has the smallest sidelobes, but also the least lowpass filter effect within the measurement range. Note that an even better result could be achieved by first doubling the sampling frequency, so that the measurement range would only extend to about 20% (instead of 40%) of the indicated sampling frequency f_s.

A very efficient means of achieving quadratic resampling of the signal is described in [50].

Yet another approach to angular resampling is to use phase demodulation of some shaft speed related component to obtain a mapping of shaft rotation angle vs time. The mean shaft speed is subtracted during this demodulation operation, but is known precisely, so can be added back to obtain the actual phase vs time relationship. The shaft speed related component can be taken from a tachometer or shaft encoder signal, but can in some cases be extracted from the vibration signal itself, as long as a clear shaft harmonic is detectable in the spectrum. Bonnardot et al. [45] show how this procedure was used to achieve order tracking of a helicopter gearbox signal, which was necessary to obtain separation of gear and bearing signals.

The phase vs time map can be used to determine the times corresponding to uniform phase increments, and thus achieve angular sampling by interpolation between the original time samples. It is generally advisable to resample to an

integer number of samples per shaft rotation, and to start the sampling at a known phase angle, for example a zero crossing of the sinusoid corresponding to the shaft speed vibration. It can be advantageous to make the number of samples per period, and also the number of periods in a record length, equal to a power of 2, so that FFT transforms can be made most efficiently, in particular for very long record lengths.

Note that once deterministic discrete harmonics have been removed, the same phase vs time map can be used to transform back to the time domain, which may be an advantage for analysis of the residual random signal, for example in cases where it is being analysed by operational modal analysis to determine natural frequencies, which are time rather than shaft angle related.

Fig. B3 shows the effect of order tracking on the spectrum of a signal from the gearbox of a mining shovel, with a reasonable variation in speed over the cycle. Without the order tracking, no discrete frequency components are visible in the spectrum.

If order tracking is being performed directly on an analogue signal, it must be ensured that the signal is adequately lowpass filtered to prevent aliasing, in particular when resampling at a lower frequency (for example as a machine speed reduces). Digital filtering can be useful here, as the cutoff frequency varies directly with the sampling frequency, but the initial analogue lowpass filtration must be such that aliasing components do not enter the measurement range. Digital oversampling can solve this problem, as from Fig. B1(c) it can be seen that the sampling frequency can be reduced by a large factor before overlap occurs. In fact, for four times oversampling, the speed reduction factor is 5.9 if a tracking digital filter is used, cutting off at about 40% of the sampling frequency [51].

References

[1] P.D. McFadden, J.D. Smith, Model for the vibration produced by a single point defect in a rolling element bearing, Journal of Sound and Vibration 96 (1) (1984) 69–82.
[2] P. Bradshaw, R.B. Randall, Early detection and diagnosis of machine faults on the Trans Alaska pipeline, in: Proceedings of the MSA Session, ASME Conference, Dearborn MI, September, 1983.
[3] H.L. Balderston, The detection of incipient failure in bearings, Material Evaluation 27 (June) (1969) 121–128.
[4] B. Weichbrodt, K.A. Smith, Signature analysis—non-intrusive techniques for incipient failure identification. General Electric Technical Information Series 70-C-364, 1970.
[5] S. Braun, The extraction of periodic waveforms by time domain averaging, Acoustica 23 (2) (1975) 69–77.
[6] S. Braun, B. Datner, Analysis of roller/ball bearings, ASME Transactions, Journal of Design 101 (1) (1979) 118–128.
[7] M.S. Darlow, R.H. Badgley, G.W. Hogg, Application of high frequency resonance techniques for bearing diagnostics in helicopter gearboxes, Technical Report, US Army Air Mobility Research and Development Laboratory, 1974, pp. 74–77.
[8] H. Engja, M. Rasmussen, J. Lippe, Vibration analysis used for detection of rolling element bearing failures, Norwegian Maritime Research 3 (1977) 23–33.
[9] K. Worden, W.J. Staszewski, J.J. Hensman, Natural computing for mechanical systems research: a tutorial overview, Mechanical Systems and Signal Processing, 2010.
[10] J. Antoni, Cyclic spectral analysis in practice, Mechanical Systems and Signal Processing 21 (2007) 597–630.
[11] J. Antoni, Cyclostationarity by examples, Mechanical Systems and Signal Processing 23 (2009) 987–1036.
[12] A.C. McCormick, A.C. Nandi, Cyclostationarity in rotating machine vibrations, Mechanical Systems and Signal Processing 12 (2) (1998) 225–242.
[13] R.B. Randall, J. Antoni, S. Chobsaard, The relationship between spectral correlation and envelope analysis in the diagnostics of bearing faults and other cyclostationary machine signals, Mechanical Systems and Signal Processing 15 (5) (2001) 945–962.
[14] J. Antoni, R.B. Randall, Differential diagnosis of gear and bearing faults, ASME Journal of Vibration and Acoustics 124 (2002) 165–171.
[15] J. Antoni, R.B. Randall, A stochastic model for simulation and diagnostics of rolling element bearings with localised faults, ASME Journal of Vibration and Acoustics 125 (2003) 282–289.
[16] M.S. Kay, S.L. Marple, Spectrum analysis—a modern perspective, Proceedings of the IEEE 69 (11) (1981) 1380–1419.
[17] B. Widrow, S. Stearns, Adaptive Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1985, pp. 349–351.
[18] G.K. Chaturvedi, D.W. Thomas, Bearing fault detection using adaptive noise cancelling, Journal of Sound and Vibration 104 (1982) 280–289.
[19] C.C. Tan, B. Dawson, An adaptive noise cancellation approach for condition monitoring of gearbox bearings, in: Proceedings of the International Tribology Conference, Melbourne, 1987.
[20] D. Ho, Bearing Diagnostics and self adaptive noise cancellation, Ph.D. Dissertation, UNSW, 2000.
[21] D. Ho, R.B. Randall, Effects of time delay, order of fir filter and convergence factor on self adaptive noise cancellation, ICSV5, Adelaide, 1997.
[22] J. Antoni, R.B. Randall, Unsupervised noise cancellation for vibration signals: part I—evaluation of adaptive algorithms, Mechanical Systems and Signal Processing 18 (2004) 89–101.
[23] J. Antoni, R.B. Randall, Optimisation of SANC for separating gear and bearing signals, in: Proceedings of the Comadem Conference, Manchester, September, 2001, pp. 89–96.
[24] J. Antoni, R.B. Randall, Unsupervised noise cancellation for vibration signals: part II—a novel frequency-domain algorithm, Mechanical Systems and Signal Processing 18 (2004) 103–117.
[25] P.D. McFadden, A revised model for the extraction of periodic waveforms by time domain averaging, Mechanical Systems and Signal Processing 1 (1) (1987) 83–95.
[26] N. Sawalhi, R.B. Randall, Localised fault diagnosis in rolling element bearings in gearboxes, in: Proceedings of the Fifth International Conference on Condition Monitoring and Machinery Failure Prevention Technologies (CM-MFPT), Edinburgh, 15–18 July 2008.
[27] S. Braun, The synchronous (time domain) average revisited, Mechanical Systems and Signal Processing, 2010.
[28] R.A. Wiggins, Minimum Entropy Deconvolution, Geoexploration, vol. 16, Elsevier Scientific Publishing, Amsterdam, 1978, pp. 21–35.
[29] H. Endo, R.B. Randall, Application of a minimum entropy deconvolution filter to enhance autoregressive model based gear tooth fault detection technique, Mechanical Systems and Signal Processing 21 (2) (2007) 906–919.
[30] N. Sawalhi, R.B. Randall, H. Endo, The enhancement of fault detection and diagnosis in rolling element bearings using minimum entropy deconvolution combined with spectral kurtosis, Mechanical Systems and Signal Processing 21 (6) (2007) 2616–2633.
[31] J.Y. Lee, A.K. Nandi, Extraction of impacting signals using blind deconvolution, Journal of Sound and Vibration 232 (5) (1999) 945–962.
[32] R.F. Dwyer, Detection of non-Gaussian signals by frequency domain kurtosis estimation, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Boston, 1983, pp. 607–610.
[33] D. Dyer, R.M. Stewart, Detection of rolling element bearing damage by statistical vibration analysis, ASME Paper, 26–30 September, 1977.

[34] J. Antoni, The spectral kurtosis: a useful tool for characterising nonstationary signals, Mechanical Systems and Signal Processing 20 (2) (2006) 282–307.
[35] J. Antoni, R.B. Randall, The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines, Mechanical Systems and Signal Processing 20 (2) (2006) 308–331.
[36] N. Sawalhi, R.B. Randall, Spectral kurtosis enhancement using autoregressive models, in: Proceedings of the ACAM2005 Conference, Melbourne, February, 2005.
[37] J. Antoni, Fast computation of the kurtogram for the detection of transient faults, Mechanical Systems and Signal Processing 21 (1) (2006) 108–124.
[38] N. Sawalhi, R.B. Randall, spectral kurtosis optimization for rolling element bearings, in: Proceedings of the ISSPA Conference, Sydney, Australia, August, 2005.
[39] D.E. Newland, Wavelet analysis of vibration signals, in: M. Crocker (Ed.), Handbook of Noise and Vibration Control, Wiley, 2007 (Chapter 49).
[40] I. Daubechies, The wavelet transform, time–frequency localization and signal analysis, IEEE Transactions on Information Theory 36 (1990) 961–1005.
[41] Z.K. Peng, F.L. Chu, Application of the wavelet transform in machine condition monitoring and fault diagnostics: a review with bibliography, Mechanical Systems and Signal Processing 18 (2004) 199–221.
[42] D.L. Donoho, I.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, Biometrika 81 (3) (1994) 425–455.
[43] Y. Wang, Z. He, Y. Zi, Enhancement of signal denoising and multiple fault signatures detecting in rotating machinery using dual-tree complex wavelet transform, Mechanical Systems and Signal Processing 24 (2009) 119–137. 10.1016/j.ymssp.2009.06.015.
[44] R.B. Randall, Noise and vibration data analysis, in: M. Crocker (Ed.), Handbook of Noise and Vibration Control, Wiley, 2007 (Chapter 46).
[45] F. Bonnardot, R.B. Randall, J. Antoni, Enhanced unsupervised noise cancellation using angular resampling for planetary bearing fault diagnosis, International Journal of Acoustics and Vibration 9 (2) (2004).
[46] D. Ho, R.B. Randall, Optimisation of bearing diagnostic techniques using simulated and actual bearing fault signals, Mechanical Systems and Signal Processing 14 (5) (2000) 763–788.
[47] N. Sawalhi, R.B. Randall, Semi-automated bearing diagnostics—three case studies, in: Proceedings of the Comadem Conference, Faro, Portugal, June 2007.
[48] J. Antoni, On the benefits of the Wigner–Ville spectrum for analysing certain types of vibration signals, in: Proceedings of the Wespac8 Conference, Melbourne, 2003.
[49] P.D. McFadden, Interpolation techniques for time domain averaging of gear vibration, Mechanical Systems and Signal Processing 3 (1) (1989) 87–97.
[50] R. Potter, M. Gribler, Computed order tracking obsoletes older methods, SAE Paper 891131, 1989.
[51] S. Gade, H. Herlufsen, H. Konstantin-Hansen, N.J. Wismer, Order tracking analysis, Brüel & Kjær Technical Review 2 (1995) 1995.