Mechanical Systems and Signal Processing (2000) **14**(5), 763–788
doi:10.1006/mssp.2000.1304, available online at http://www.idealibrary.com on IDEAL®

OPTIMISATION OF BEARING DIAGNOSTIC TECHNIQUES USING SIMULATED AND ACTUAL BEARING FAULT SIGNALS

D. Ho and R. B. Randall

School of Mechanical and Manufacturing Engineering, The University of New South Wales, Sydney, NSW 2052, Australia. E-mail: b.randall@unsw.edu.au

(Received 11 January 1999, accepted 7 March 2000)

In this paper, bearing fault vibrations are modelled as a series of impulse responses of a single-degree-of-freedom system. The model incorporates slight random variations in the time between pulses so as to resemble actual vibration signals. Although the bearing fault harmonics in the raw spectrum are caused by the random fluctuations to smear over one another, they remain quite clear in the spectrum of the envelope. However, the envelope spectrum is still prone to masking by discrete and random noise. Therefore, the simulated bearing fault signals were used to investigate the efficient application of self-adaptive noise cancellation (SANC) in conjunction with envelope analysis in order to remove discrete frequency masking signals. Two ways of combining these techniques have been suggested, both of which require the original signal to be band-pass filtered and frequency-shifted in order to reduce the number of samples to be processed by SANC. The subsequent envelope analysis can then be performed by using the Hilbert transform technique or band-pass rectification. Band-pass rectification is simpler but requires extra zero padding above and below the demodulation band, making the length of the signal processed by SANC twice as long as with the former method, but still only a fraction of the length of the original signal. On the other hand, the Hilbert technique requires an extra forward and inverse discrete Fourier transform operation compared with band-pass rectification. These two methods reduce the masking effects in the envelope spectrum by removing pseudo-sum frequencies or placing them outside the frequency range of interest. This is illustrated with examples of simulated and actual vibration signals. The removal of discrete frequency noise using SANC is also demonstrated for actual vibration signals. The threshold for which analysing the squared envelope or its higher powers gives an improvement in the envelope spectrum has also been defined using simulated and actual vibration signals. The treatment in the paper is qualitative and non-mathematical for purposes of clarity, but reference is made to a quantitative treatment of the effects of masking.

1. INTRODUCTION

Vibration analysis has been used extensively in bearing diagnostics of rotating machinery [1, 2]. The measurement of the vibration signals involves the installation of a permanent or temporary transducer as close as possible to the bearing that is being monitored so that the clearest fault signal can be obtained. The fault signal results from the vibration generated when the fault—either on the inner race, outer race or on the rolling elements—interacts with the other rolling surfaces. The kinematic and kinetic interaction between the bearing fault and the rolling surfaces is quite complex and is beyond the scope of this paper. However from a condition monitoring point of view, the vibrations generated by a fault are most important and many methods have been devised to extract the fault frequencies from the measured vibrations.

0888–3270/00/050763 + 26 $35.00/0

1.1. BEARING FAULT SIGNALS

When a fault in one surface of a bearing strikes another surface, a force impulse is generated which excites resonances in the bearing and the machine. The successive impacts produce a series of impulse responses which may be amplitude modulated as a result of the passage of the fault through the load zone or of the varying transmission path between the impact point and the vibration measurement point. The spectrum of such a signal would consist of a harmonic series of frequency components spaced at the bearing defect frequency with the highest amplitude around the resonance frequency. These frequency components are flanked by sidebands if there is an amplitude modulation.

The vibration signal and the corresponding spectrum described above are for a series of equally spaced force impulse excitations. In reality, there is a slight random fluctuation in the spacing between each force impulse because the load angle on each rolling element changes as the rolling elements enter and leave the load zone. This means the rolling diameter of each rolling element is different and some will want to roll faster than others; however, the cage keeps them apart at a certain mean spacing and they all travel around the bearing race at an average of the cage speed. The random fluctuation can be so small that the variation in the spacings may not be detectable to the eye; however the effect on the spectrum is much more pronounced. The slight random fluctuation causes the frequency components to smear laterally, and at higher frequencies they may smear over more than one complete harmonic spacing. In the first part of this paper, a model is devised which incorporates the random fluctuations of a bearing fault vibration.

1.2. ENVELOPE ANALYSIS

A well-known method used to extract bearing defect frequency components from the signal is envelope analysis. Figure 1(a) shows a simulated series of impulse responses without random fluctuation and Fig. 1(b) shows its corresponding spectrum. The higher harmonics of the defect frequency in Fig. 1(b) are quite clear (even though the low harmonics are very small) unlike the spectrum for the series of impulse responses with slight random fluctuation where the defect components are smeared into each other [Fig. 1(e)].

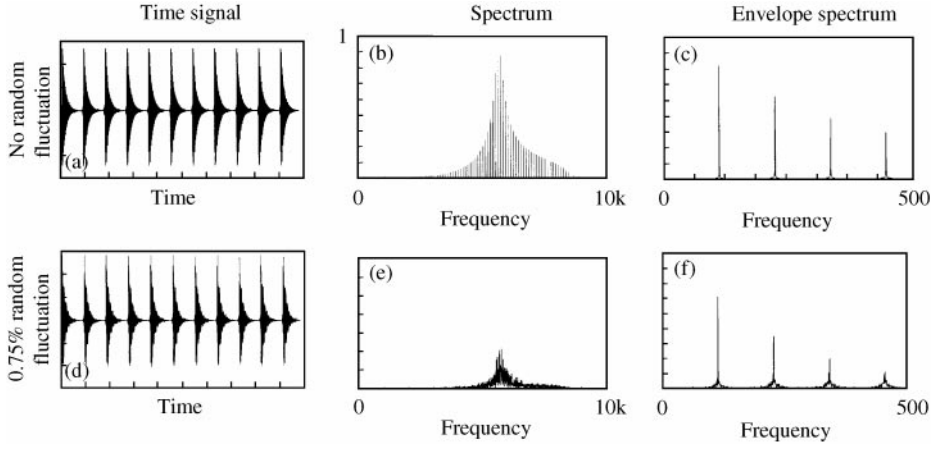

Figure 1. Bearing fault pulses with and without random fluctuations: (a) A series of bearing fault impulses at equal time spacing. (b) The spectrum of (a). (c) Envelope spectrum of (b) demodulated around the resonance frequency. (d) A series of bearing fault impulses with 0.75% random fluctuation. (e) The spectrum of (d). (f) Envelope spectrum of (e) demodulated around the resonance frequency.

The defect frequency components in the envelope spectrum are not as sensitive to random fluctuations and they are still clearly visible [Fig. 1(f)].

The simplest method to perform envelope analysis is to pass the signal through an analogue high-pass filter to remove the low-frequency noise and then by rectifying and frequency analysing the signal, the defect frequency components can be determined in the envelope spectrum. Envelope analysis can be made more efficient by digitising the signal and band-pass filtering it in a frequency region where there is a high signal-to-noise ratio, for example typically around a resonance. The envelope spectrum can then be obtained by using the Hilbert transform technique. This latter method, which manipulates a digital signal, has the advantage of reducing the size of the record that is being processed and gives complete flexibility in choosing cut-off frequencies for the band-pass filter.

In this paper, a study was made of whether analysing the squared envelope rather than just the envelope signal would give a better result in the envelope spectrum. The new model for bearing fault vibrations, which includes the random fluctuations, was used in this study to simulate a mixture of bearing fault signals with random and discrete noise in varying amounts. For clarity, the treatment is largely graphical and non-mathematical, in particular of relationships involving convolution, but reference is made where appropriate to more quantitative treatments of some operations, such as the effects of masking.

1.3. SELF-ADAPTIVE NOISE CANCELLATION

Self-adaptive noise cancellation (SANC) is a further development of adaptive noise cancellation (ANC). ANC uses a reference signal coherent with one of two components in a primary signal to separate the two components. SANC requires only one signal but uses a delayed version of the primary signal as a reference signal, and can be used to separate discrete frequency components from random components, which have a short correlation length. Since bearing fault vibrations have a slight random fluctuation as explained earlier, SANC is able to attenuate any discrete frequency components masking the bearing fault frequency components. This method can be used for the removal of masking components which reside in the same frequency region as the bearing fault components and thus cannot be removed by band-pass filtration. Situations may arise where the bearing fault manifests itself as discrete frequency components and in such cases SANC will not improve the separation.

SANC is usually applied prior to envelope analysis as it removes any discrete masking noise from the signal before it is demodulated to obtain the envelope spectrum. This paper introduces a method for manipulating the signal so that SANC and envelope analysis can be used together efficiently and without the generation of additional masking frequency components.

2. SIMULATION

2.1. MODELS FROM THE LITERATURE

A number of authors have contributed to the modelling of bearing fault signals. However from the diagnostic point of view, the most important aspects of these models are those which directly affect the appearance of the envelope spectrum. Epps and McCallion [3] found that for an outer race notch fault there is a negative ramp in the acceleration signal which represents the entry of the rolling elements into the fault, followed by the response of the bearing system when the rolling element strikes the other edge of the fault. Epps and McCallion also found the separation between the point of entry and the point of impact to be in proportion to the size of the fault; thus it could be a very useful trending parameter.

However, this determination is not usually possible because of the presence of background noise and gear vibration.

The simulation of a bearing fault signal in this paper does not include the above-mentioned negative ramp because it is the higher frequency range that is usually demodulated for diagnostics and thus the spectral information linked to the initial dip of the signal will be filtered out. Only one resonance is considered in this simulation, because often when performing envelope analysis, a single resonance is demodulated in a region where the bearing fault signal has increased most above the noise.

An outer race fault is generally found in the region of the load zone and the defect resembles a depression in the raceway. As a rolling element rolls in and out of the fault, it will destress and restress, respectively. The rolling element drops into the depression and most of the load will be supported by the adjacent rollers or balls. Some bearing signals have been found to be made up of two-impulse responses resulting from the abrupt destressing and restressing of the bearing components [4]. However for cases where detection of initial faults is considered, the fault will be small and modelling with a single-impulse response will be more appropriate. Modelling with two-impulse responses can easily be incorporated if the delay between the two impulses is known.

McFadden and Smith [5, 6] investigated the way in which bearing fault signals (shown as displacement rather than acceleration) are manifested in the envelope spectrum and they developed a model for the bearing fault signal by considering aspects such as the repetition frequency of the impulses, the bearing load distribution and transmission effects due to the variation in the point where the rolling element impacts the fault. By applying these concepts to outer race faults, inner race faults and rolling element faults under axial and radial loading, they were able to predict the spectral pattern corresponding to all the possible cases.

Su and Lin [7] took this a step further by considering the bearing signal under varying loading due to shaft unbalance and roller errors. These two effects cause the loading distribution to have a different effect on the bearing signal than those assumed in the McFadden and Smiths' model.

2.2. A NEW MODEL WITH RANDOM FLUCTUATION

The response of a single-degree-of-freedom (sdof) system is the simplest model of the vibration signal due to an incipient fault. The effects of multiple resonances and the negative ramp at the start of the response mentioned earlier are neglected in the present model, because the frequency components resulting from these effects are not included in a typical envelope analysis process. The response of the bearing to faults can be modelled by a spring–mass–damper system whose spectra for the displacement and acceleration are stated in equations (1) and (2), respectively.

$$\text{Displacement} = \frac{F_0/k}{(1 - r^2) + \mathrm{i}2\zeta r} \tag{1}$$

$$\text{Acceleration} = \frac{-F_0\omega^2/k}{(1 - r^2) + \mathrm{i}2\zeta r} \tag{2}$$

where F_0 is the amplitude of the input force at frequency ω, k the spring constant, r the ratio of frequency over natural frequency and ζ the damping ratio.

For the spectrum of the displacement, the low-frequency region is dominated by the spring line which is horizontal and the high-frequency region is dominated by the mass line which is a decreasing curve proportional to $1/\omega^2$. The situation is the opposite for the spectrum of the acceleration, that is the spring line at the low frequencies is now an ω^2

parabola while the mass line at the high frequencies is a constant. For a periodic series of impulse responses the spectrum as stated in equation (2) will be sampled at the repetition frequency of the impulse responses. However, the frequency components near 0 Hz will have low amplitudes because of the parabolic spring line and often these are below the level of background noise.

If all impulse responses are excited at exactly equal intervals, then the spectrum will be a line spectrum as shown in Fig. 1(b). However, if the impulse responses are not equally spaced and the spacing 'T' varies randomly [8, 9], then the higher harmonics will broaden and begin to merge into one another. This random fluctuation occurs as a result of clearances in the cage and the variation in radial load on the rolling elements as they pass through the radial load zone. As each rolling element experiences a change in load angle, its rolling diameter also changes. Thus, some rolling elements will attempt to roll faster than others, but the cage keeps them at a nominal speed by inducing random slip. This tugging forward and backward motions will result in slight variations in the spacings of the excitation impulses which is responsible for the smearing effect in the spectrum.

If the ball-pass frequency is at 1 Hz and the speed fluctuation is 1%, then there will be complete smearing of harmonics at harmonic 100 ($100\ \text{Hz} * 1\% = 1\ \text{Hz}$). The amplitude of the whole spectrum will decrease because the energy is now spread to the regions between each harmonic [Fig. 1(e)]. With no random fluctuation, the ball-pass frequency spacing can be seen clearly in the spectrum but with random fluctuations, the smearing effect prevents these spacings from being detectable, even around the resonance. However, the ball-pass frequency components can be detected in the envelope spectrum quite clearly even with random fluctuation [Fig. 1(f)].

2.3. COMPUTER-GENERATED BEARING FAULT SIGNALS

The bearing fault signals can be generated as a series of impulse responses of a sdof system. From first principles it can be shown that the complex spectra for the displacement and acceleration responses are given by equations (1) and (2). Figure 2 outlines the procedures followed in producing a series of acceleration impulse responses with finite sampling frequency, random fluctuations between each impulse spacing and load modulations. The length of each impulse response is varied by changing the number of samples used for equation (2) but keeping the sampling frequency and resonance frequency the same. Equation (2) assumes an infinite sampling frequency but since this is not realisable and a real signal is desired, a finite sampling frequency has to be set and a two-sided spectrum is needed. The sampling frequency, resonance frequency, ball-pass frequency and the modulation frequency are arbitrary variables which can be changed in the simulation.

Figure 3(a) is the amplitude spectrum corresponding to equation (2) modified by a low-pass filter as discussed below. The resulting single-impulse response in Fig. 3(b) exponentially decreases toward zero but there is a fluctuation at the end of the pulse which is not part of the response as defined by equation (2). This fluctuation occurs as a result of applying the simulated low-pass filter to the response. The impulse response of the low-pass filter is shown in Fig. 3(c) and this is symmetrical about time zero but the non-causal part is shown at the end of the period. It can be seen that the fluctuation towards the end of the filter response [Fig. 3(c)] is similar to the fluctuation at the end of the response of the sdof system [Fig. 3(b)].

The low-pass filter was simulated by a double-half hanning window in the region of the Nyquist frequency [Fig. 3(a)]. Although the gain of the filter is similar to that of real filters, it has zero-phase shift which means the filter's impulse response will have non-causal components as shown in Fig. 3(c). A linear phase shift can be applied to the whole spectrum

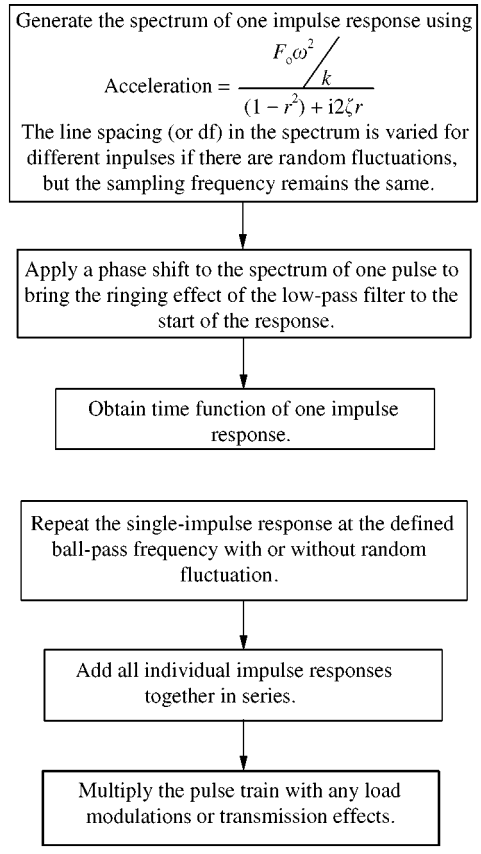

Figure 2. Procedures for generating a series of impulse responses with or without random fluctuations.

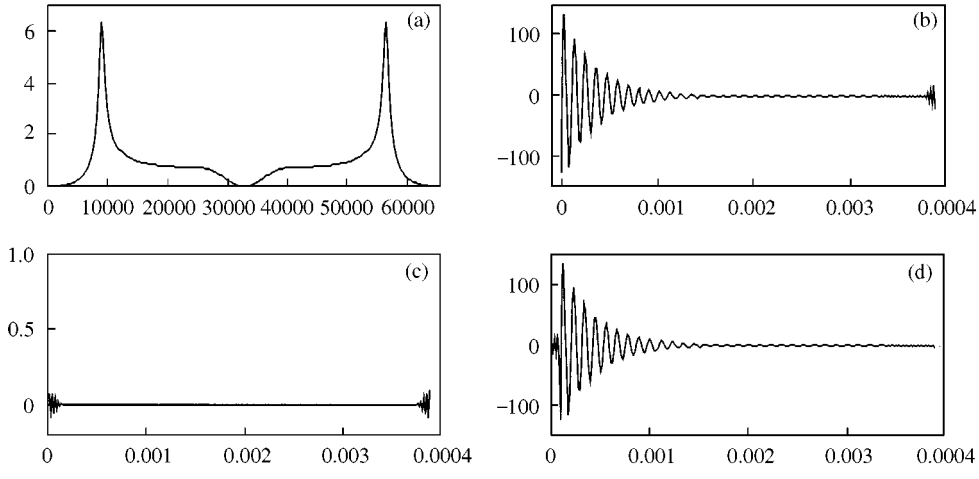

Figure 3. (a) Two-sided amplitude spectrum corresponding to equation (2) with low-pass filtering around Nyquist frequency. (b) A single-impulse response. (c) Impulse response of the low-pass filter used in the simulation. (d) A single-impulse response with the most significant part of the filter response brought to the front of the signal.

so that the most significant non-causal components are now brought to the front of the bearing fault pulse [Fig. 3(d)] as for a physical low-pass filter.

The response as defined by equation (2) has a delta function at time zero because an impulse excitation will impart an infinite acceleration (or a step velocity) to the mass of the sdof system. However, due to the low-pass filtration in the simulation the delta function is reduced to a finite negative value as shown in Fig. 3(b). In reality, when the rolling elements roll into a bearing fault the excitation can be modelled as excitation of the sdof system by a velocity step of the base rather than of the mass. This does not incur an initial impulse but the subsequent vibration response is the same.

3. ENVELOPE ANALYSIS BY HILBERT TRANSFORM TECHNIQUES

As shown earlier, the spectrum of a series of pulses with random spacing fluctuation is smeared and this prevents the ball-pass frequency harmonic spacings from being detected. However, this frequency spacing is quite clear in the envelope spectrum and can be used as a reliable source of information for bearing diagnostics. The envelope spectrum can be obtained by envelope analysis, which extracts the envelope of a signal. This can be done using analogue techniques whereby the signal is band-pass filtered, rectified and frequency analysed to generate the envelope spectrum. In general, the choice of pass band of the analogue filters is limited and the filter characteristics are poor compared with the digital methods to be described. Alternatively, amplitude demodulation can be performed digitally either using full base-band rectification [10] or Hilbert transform techniques [11, 2]. If no band-pass filtration is required, the former method may be preferred as it only involves one forward discrete Fourier transform (DFT). However, the latter method is to be preferred for envelope analysis of a band-passed region because it reduces the total number of samples to be processed and even though it requires two forward DFT and one inverse DFT operations, two of these are much smaller than the first, and only marginally increase the computation time.

The envelope of a signal $x(t)$ can be obtained by taking the amplitude of a complex signal formed from $x(t)$ as the real part and its Hilbert transform as the imaginary part. The Hilbert transform can be obtained by shifting the phase of the original spectrum and then inverse Fourier transforming back to the time domain. Rather than manipulating the phase of the frequency components, the Hilbert transform can alternatively be generated from the one-sided spectrum of a signal. This technique is often performed for a frequency region where the signal-to-noise (SNR) ratio is highest such as a resonance in the high-frequency region. The chosen bandwidth is first filtered digitally by setting the amplitudes of the unwanted frequency components to zero and then shifted so that the lowest frequency component in the chosen band is now at the low-frequency end [Fig. 4(a) and (b)]. To ensure a one-sided spectrum, the frequency band is padded with zeros to double the length in order to set the negative frequency components to zero [Fig. 4(c)]. The inverse Fourier transform of such a spectrum is an analytic signal which is complex and whose imaginary component is the Hilbert transform of the real component [Fig. 4(d)]. This procedure has effectively replaced each pair of positive and negative frequency phasors of the original two-sided spectrum with one phasor of double-amplitude spinning in the positive direction (analytic signal) and whose combined amplitude is the envelope of the signal.

4. SQUARED ENVELOPE AND HIGHER POWERS

The envelope process extracts the amplitude (or the envelope) of the modulation and this can be obtained by using the Hilbert technique shown in Fig. 4. However, it is the square of

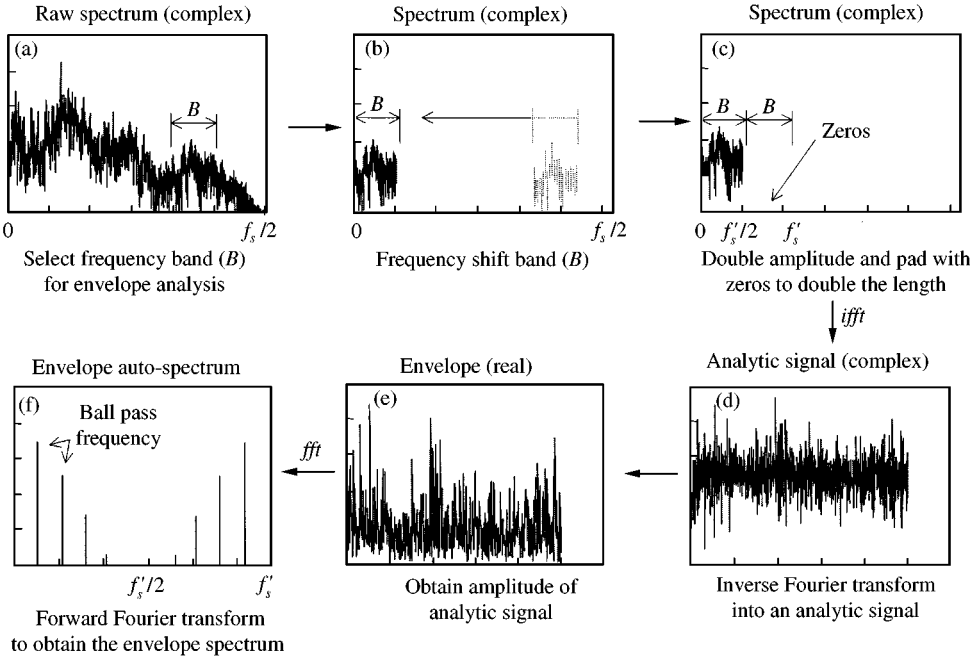

Figure 4. Procedure for envelope analysis using Hilbert transform technique. In practice, only a frequency span up to the Nyquist frequency ($f_s'/2$) is needed in the envelope spectrum for diagnostic purposes. Note, the original sampling frequency (f_s) is much larger than the sampling frequency of the analytic signal (f_s'). This also means the number of samples in the analytic signal is reduced.

the envelope which is more fundamental as the square root of this must be taken to produce the amplitude in Fig. 4(e). It is of interest to know which to use, because it has been found that analysing the squared envelope can improve the SNR ratio in certain situations.

There may be cases where it is very difficult to find a frequency band dominated by bearing components because of interference from discrete frequencies. Such is the case in a helicopter gearbox where there are literally dozens of sets of meshing gears, and where use can be made of ANC [12, 13] and SANC [14–18] to remove the related gear-induced frequency components from the band. If the SNR is greater than unity then there is an advantage in analysing the squared envelope rather than the envelope itself, because the SNR increases as a result of the squaring operation [19]. Conversely, if the SNR is less than unity then squaring the envelope signal will make the situation worse. Thus, it is important to define the SNR ratio so that it can be known when squaring the envelope will produce a more favorable result.

In bearing diagnostics, it is often desirable to be able to detect up to the third harmonic of the bearing defect frequency in the envelope spectrum. However, this is not always possible as the higher harmonics may have decreased in amplitude to such a stage that they are below the background noise. A SNR of 1 is defined here as the situation where squaring the envelope does not improve or degrade the detection of up to the third harmonic of the bearing fault frequency in the envelope spectrum. A SNR of higher than 1 will result in an improvement and the converse will occur if the SNR is lower than 1.

The SNR does not just depend on the mean-squared ratio (MSR) between the bearing and the noise components in the original spectrum, it is also affected by the amount of smearing of the bearing pulses. For a MSR of 1, that is where the power of the bearing signal in the demodulated band is the same as the power of the random noise, depending on the

bandwidth of the smeared bearing components they may be protruding above or hidden in the noise. If the bearing frequency component has an infinitesimally small bandwidth, then its power spectral density (PSD) will be infinite and thus could be distinguished from the noise very easily. As the bandwidth widens, the PSD will decrease until it is level with the noise in which case it is masked by the noise. This is why a SNR of unity in the envelope spectrum does not necessarily correspond to a MSR of unity in the original spectrum. The amount of smearing also has to be taken into account.

For each value of fluctuation of the bearing pulses, there is a corresponding MSR where squaring the envelope signal will not improve or degrade the detection of the third harmonic. Such a case signifies the condition where the SNR is unity. A study was conducted with simulated bearing fault signals to find the relationship between the signal-to-noise MSR and the random fluctuation of the bearing fault pulses for the case where SNR is unity. The results are shown in Fig. 5 and the line represents the threshold where improvements can be gained from using the squared envelope. For MSRs above the line, squaring the envelope will improve the envelope spectrum by at least enhancing the third harmonic with respect to the background noise. However, using the squared envelope for MSRs below the line will increase the level of the noise relative to the bearing harmonics.

Figure 6(b) and (c) shows a typical example of the effects of squaring a simulated signal with a MSR of 0.3 and a percentage fluctuation of 0.5%. Figure 6(a) shows the unmasked envelope spectrum.

Figure 7 is an example of how squaring can give an improvement for practical signals. This signal was recorded from a paper mill and Fig. 7(a) is the envelope power spectrum

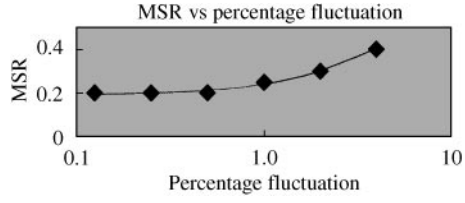

Figure 5. Threshold of mean-squared ratios (MSRs) where values above the line mean that squaring the envelope improves the SNR for different percentage fluctuations.

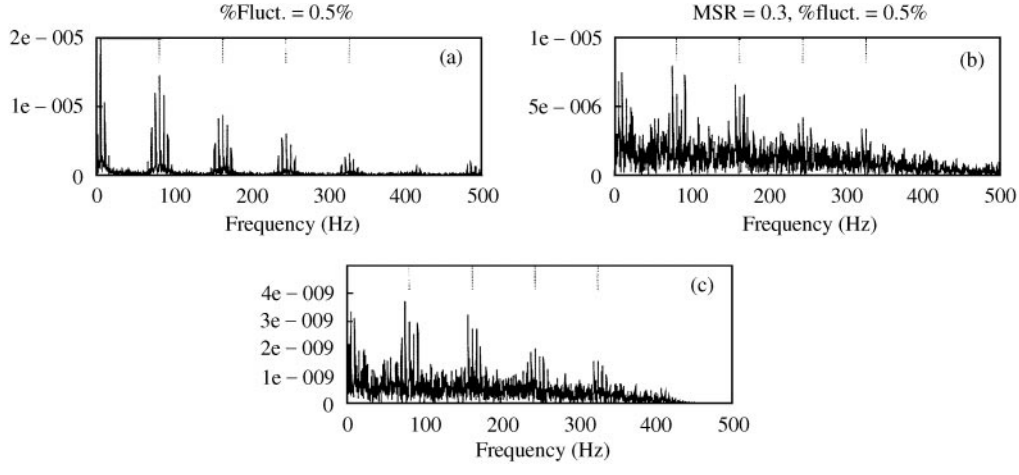

Figure 6. (a) Envelope spectrum of a simulated inner race fault signal. (b) Envelope spectrum of bearing signal and noise with no squaring. (c) Spectrum of the squared envelope.

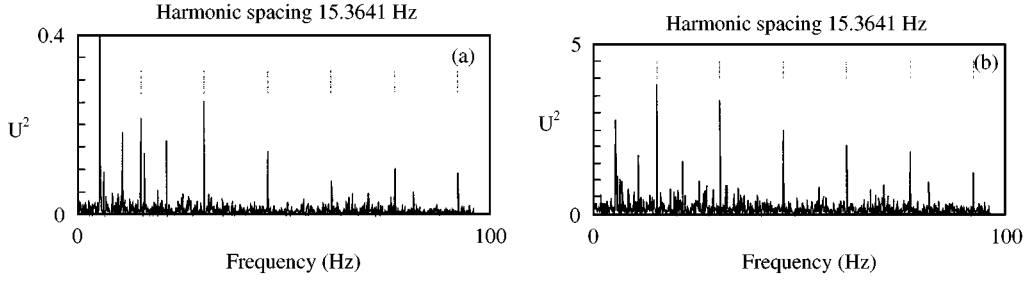

Figure 7. (a) Envelope power spectrum of paper mill signal demodulated from 5.2–5.6 kHz. (b) Power spectrum of the squared envelope from the paper mill signal.

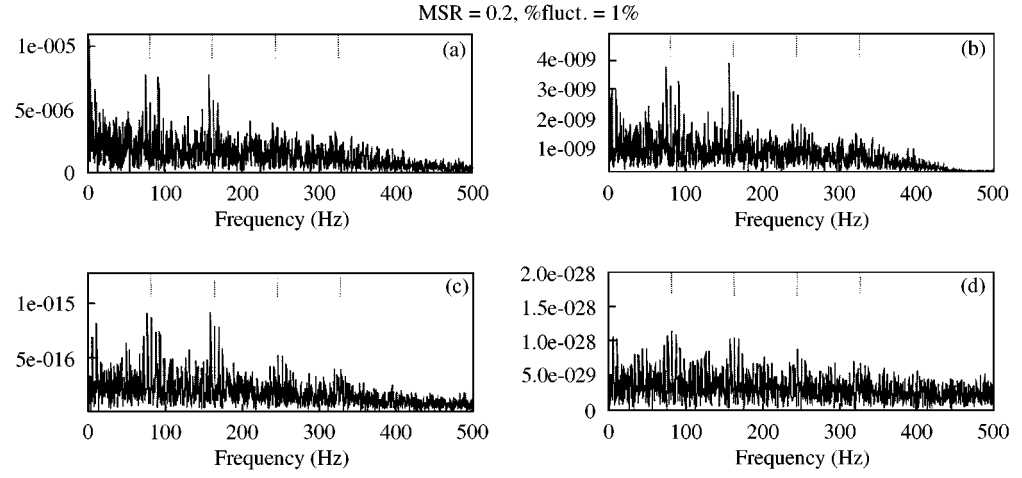

Figure 8. (a) Envelope spectrum of a simulated bearing fault signal and noise with no squaring. (b) Spectrum of the squared envelope. (c) Spectrum of the envelope raised to the fourth power. (d) Spectrum of the envelope raised to the eighth power.

demodulated in the range 5.2–5.6 kHz. The band-pass filtration process has reduced the background noise considerably but by squaring the envelope, the extraneous components coming from a lubricator were further reduced. The spectral comparison for this demodulated band indicated an increase of approximately 12 dB as a result of the bearing fault, which implies that the MSR was well above the values tested in Fig. 5.

If the SNR (as defined in this paper) is greater than unity, then squaring the envelope will enhance the bearing frequency components in the envelope spectrum. If this is the case, then raising the envelope to a higher power should improve the SNR even further. This theory was tested with simulated signals and measured signals. Figure 8(a) shows the envelope spectrum of a simulated inner race fault with noise giving an MSR of 0.2. Figure 8(b)–(d) shows the envelope spectra obtained by raising the envelope of the simulated signal to the power of 2, 4 and 8, respectively. Raising to the power of 4 gives a definite advantage over a power of 2 but a further increase does not seem to give any benefit. Note that even if obtained from a one-sided spectrum, the squared envelope signal is real and thus raising it to even higher powers introduces pseudo-sum frequencies.

Figure 9(a)–(c) shows the envelope spectrum of a signal obtained from a railway bearing with an inner race fault. By comparing the demodulation band for the fault and fault-free

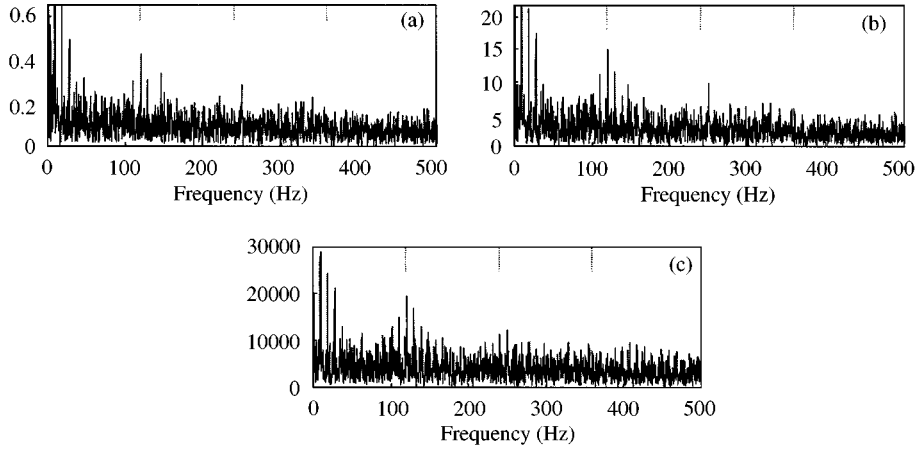

Figure 9. (a) Envelope spectrum of an actual bearing fault with no squaring. (b) Spectrum of the squared envelope. (c) Spectrum of the envelope raised to the fourth power.

cases, the signal-to-noise MSR was found to be 0.28 and the percentage fluctuation of the ball-pass frequency was approximately 0.9%. There is a definite advantage in raising the envelope signal to the fourth power as the second harmonic and its sidebands become clearer by protruding above the background noise.

5. ENVELOPE ANALYSIS AND SANC

Envelope analysis is an amplitude-demodulation process used to obtain the bearing defect harmonics from the spectrum for fault diagnostic purposes. However, this process also demodulates noise components which may be in the same frequency region as the bearing fault and therefore mask the diagnostic information. Digital signal processing techniques such as SANC may be used to remove harmonic families of discrete masking components first before the signal is demodulated using envelope analysis. A schematic diagram of SANC is shown in Fig. 10. The adaptive filters in the SANC are able to differentiate and remove the discrete frequency components from the bearing fault frequency components which have random-like properties as explained earlier.

Figure 11 shows the processes for using SANC in conjunction with envelope analysis. The SANC is not applied to the entire signal because this would require removal of too many discrete components at one time and make the whole adaptive process very inefficient. A better way to use SANC is to band-pass filter and frequency shift the signal so that the noise in other frequency regions can be eliminated and the total number of samples in the signal reduced.

There are two methods suggested for performing envelope analysis as shown in Fig. 11. Depending on which method is used for envelope analysis, the signal used as an input to the SANC will have to be zero padded in different ways. It would be faster—in terms of the number of DFT operations—to implement band-pass rectification rather than the Hilbert technique for envelope analysis because the former only requires the absolute value operation in the time domain while the latter would need a forward DFT, a inverse DFT and zero padding in the spectrum. However, as shown in Section 5, band-pass rectification will require zero padding to double the length (compared with the Hilbert technique) to form a real-valued signal for SANC that is free from spurious sum frequency components.

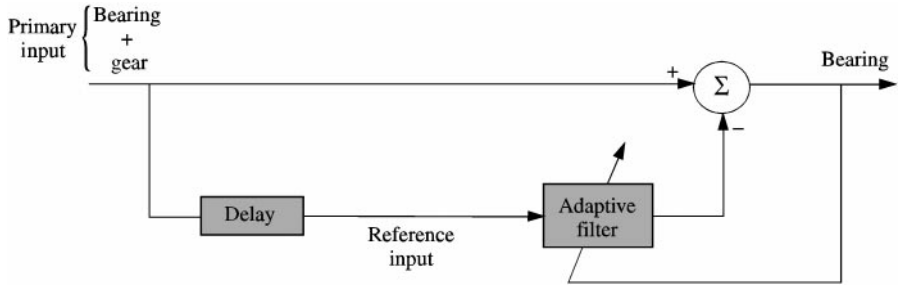

Figure 10. Schematic diagram of SANC.

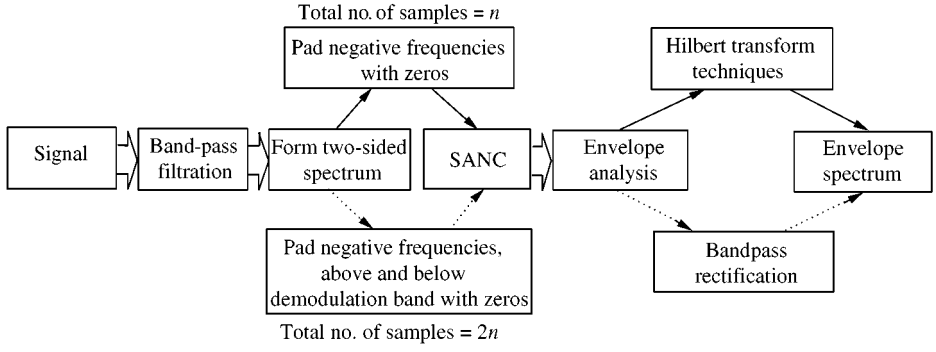

Figure 11. Two ways in which to use SNAC in conjunction with envelope analysis.

Whether to use the Hilbert technique or band-pass rectification depends on the software of the user and the type of signal processing techniques used to reduce the noise. Both methods have their advantages and disadvantages; it is the aim of this paper to provide an understanding of the consequences of both methods. When dealing with band-passed, frequency-shifted signals, care must be taken to ensure that pseudo-frequency components are not generated, that is, extraneous frequency components in the envelope spectrum which did not come from the original signal, but from the way the signal was manipulated. The next section of this paper compares how the two techniques used for envelope analysis, that is band-pass rectification and the Hilbert transform technique, are affected by pseudo-masking frequency components.

6. MASKING EFFECTS

6.1. HILBERT TRANSFORM TECHNIQUE

The envelope spectrum does not just contain the bearing diagnostic information. There may be components deriving from discrete frequency components (such as various sidebands of gearmesh frequencies) and random frequency components. An analysis of masking effects in digital envelope analysis was done in [20] where the bearing fault modulations, masking by discrete frequency components and narrow-band random components were mathematically modelled as signals with a one- or two-sided spectrum. Whether the spectrum is one- or two-sided does affect the way masking components appear

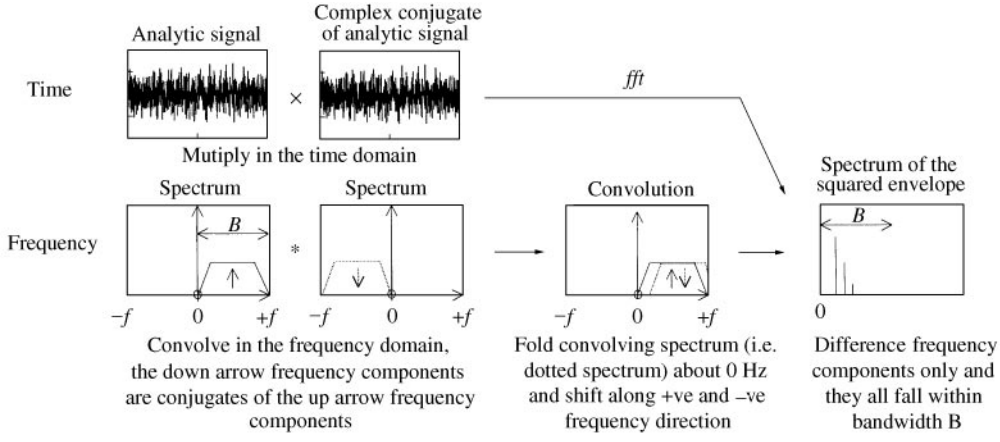

Figure 12. Obtaining the squared envelope of a signal with a one-sided spectrum.

when the amplitude of the raw signal is taken. Figure 12 shows the process of obtaining the squared envelope of a complex analytic time function and the corresponding convolution in the spectrum [21]. This operation constitutes the first step in obtaining the amplitude and is followed by the square root process. It can be seen that as the two spectra are shifted relative to each other, the biggest difference frequency is B; however if the spectrum is two-sided then the biggest difference frequency would be $2B$ because the difference between the positive and negative frequencies would generate sum frequency components. Therefore, envelope analysis by the Hilbert transform technique, i.e. the upper option in Fig. 12, will only have half as many masking components appearing in the envelope spectrum as the band-pass rectification technique.

Randall and Gao [20] determined that the modulation frequencies representing the bearing fault are always present additively in the envelope spectra but they can be masked by discrete or random noise. Full base-band rectification of a signal was considered to give minimal interference from the masking signals because the sum and double-frequency components will not usually appear in the vicinity of the bearing fault components. This assumes that the highest frequency masking component is at 80% of the Nyquist frequency, and its double-frequency component will fold back to 40% of the Nyquist frequency in the spectrum. In general, the bearing fault components are assumed to be in a region much lower than this. The general analytical equation for the rectification process in [10, 20] showed also that higher order terms were generated as a result of this procedure.

Rectified band-pass filtered and frequency-shifted signals are more susceptible to masking effects than rectified base-band signals. It was mentioned earlier that for base-band rectification, the sum and double frequencies may fold back to appear between zero and the Nyquist frequency, but in general they are in a region that is higher than the bearing fault components. However for rectification of band-pass filtered signals, the sum and double-frequency components will be mixed in with the diagnostic information because the frequency range has been restricted so that the entire demodulated band is of interest; that is, the first three or four harmonics of the bearing fault frequency usually span over the entire frequency range of the envelope spectrum. The sum and double-frequency components which appear for this type of rectification are actually pseudo-components because they do not relate to any frequency spacings in the original signal. Rather they relate only to the frequency-shifted band-pass filtered signal, which is dependent on the selection of the demodulation band.

6.2. BAND-PASS RECTIFICATION

The upper option for envelope analysis in Fig. 11 is the preferred method if no extra zero padding is used for the band-pass rectification technique. This is because the upper technique produces fewer masking frequency components. However, the band-pass rectification can yield the same result if appropriate zero padding is used. The following sections explain the benefits of zero padding for band-pass rectification and how it affects a bearing fault signal that is contaminated with discrete and random frequency components.

6.2.1. The squaring process

The band-pass rectification technique for envelope analysis in Fig. 11 can be modelled in a two-step process. First the signal is squared and then the square root is extracted. Thus, the effects of the square root process do not come directly from the original signal, but from the square of the original signal. Although this may appear to be trivial, a deeper understanding can reduce masking due to the squaring operation when performing rectification on a band-pass filtered and frequency-shifted signal. Consider a real-valued time signal which has been band-pass filtered and frequency-shifted for use with the SANC process [Fig. 13(a)] [18]. The SANC output can then be rectified to obtain the envelope signal. The first part of the rectification process is squaring in the time domain, which is

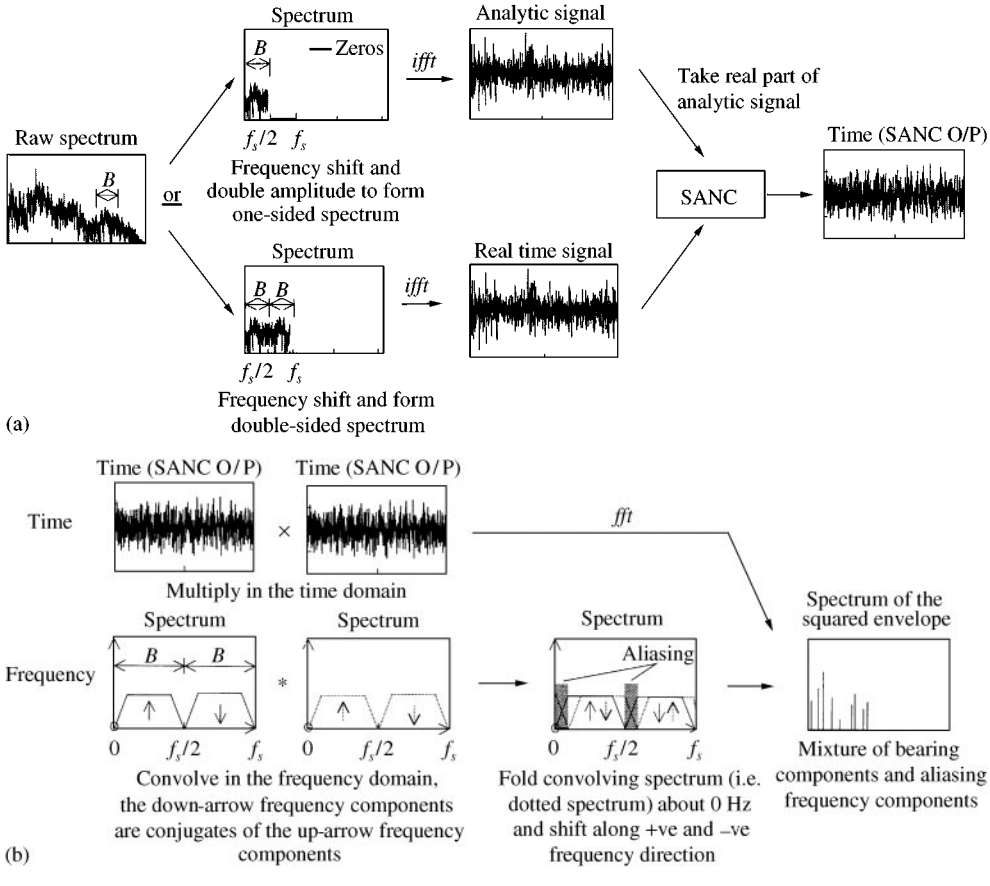

Figure 13. (a) Forming real time signal for input into SANC. (b) Steps for obtaining the spectrum of the squared envelope signal.

equivalent to the spectrum being convolved with itself in the frequency domain. This procedure involves folding the convolving spectrum about zero frequency and then integrating the products of the two spectra for each value of (positive or negative) frequency shift. However, as soon as the convolving spectrum is shifted in the positive or negative direction, the negative components from one spectrum will begin to mix with the positive components from the other spectrum [Fig. 13(b)]. This is due to the circular effects of the DFT, which requires the time and frequency functions to be periodic. The subsequent sum of the products will include this extra contribution from the mixing and hence the envelope spectrum will include both sum and difference frequency components.

In order to avoid the mixing of the negative and positive frequency components, the spectrum should be zero padded to double the size compared with generating the analytic signal. The padding of zeros should be done in the way shown in Fig. 14(a) [18] where there are zeros at frequencies below and above the demodulated bandwidth '*B*' in addition to the zeros which are used to pad the negative frequencies. This now makes the frequency span two bandwidths wide rather than one. By having zeros padded around zero frequency, the negative frequency components will not interfere with the positive frequency components until the frequency shift is greater than one whole bandwidth '*B*' [Fig. 14(b)]. The result of the convolution is a spectrum that contains the modulation information within

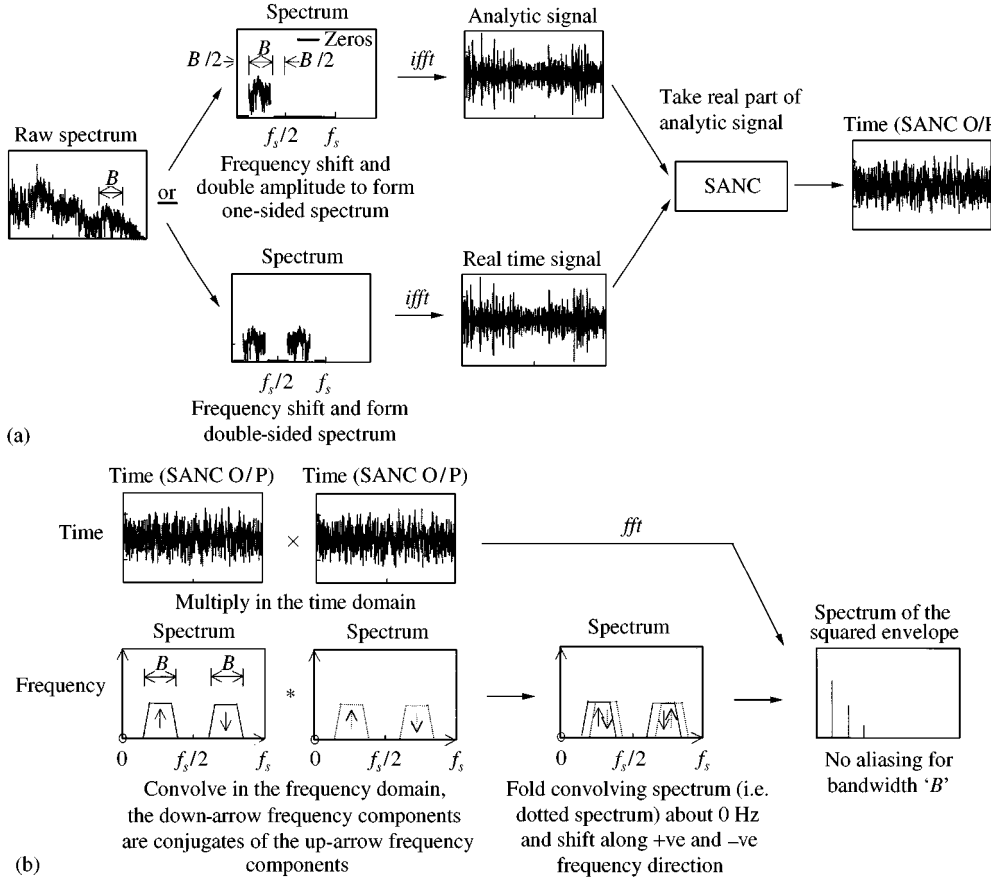

Figure 14. (a) Recommended way of forming real time signal for input into SANC. (b) Steps for obtaining the spectrum of the squared envelope signal.

one bandwidth span ('B') from 0 Hz, above which are the pseudo sum-frequency components.

There is no need for extra padding with zeros if the squared envelope of an analytic signal is required; such is the case for the Hilbert transform technique. An analytic signal can be generated using the procedures shown in Fig. 4. In order to obtain the amplitude, the analytic signal is multiplied by its complex conjugate and then the square root extracted. The squaring process can again be modelled as a convolution process as shown in Fig. 12. Since an analytic signal has no negative frequency components, there will be no pseudo-sum frequency components in the spectrum of the squared envelope. Therefore, the spectrum of the squared envelope will be the same for the frequency range of interest whether it is obtained by the Hilbert technique or the band-pass rectification technique with appropriate zero padding.

6.2.2. Effects of the square root process

The second-half of the rectification process is the square root operation. Convolution in the frequency domain was sufficient to explain all the frequency components in the spectrum when the time function was squared. However, the square root operation is more complex as it produces the higher order components of the function that is being square rooted.

The rectification of a sinusoidal function will result in an infinite number of harmonics of the double frequency with decreasing amplitude. This can be interpreted as due to the cusps on the time axis. Figure 15(a) shows the windowing of a cosine function with period '$2T$' by a rectangular window of width 'T' in the time domain [21]. The resultant spectrum is the convolution of the two spectra. The windowed cosine function is then convolved with a series of delta functions spaced at 'T' in the time domain so that it now becomes the rectified version of the original cosine function. The resultant spectrum is sampled at multiples of '$1/T$', which are the harmonics of the double frequency. Since infinite sampling frequency is not possible in simulation, the spectrum is folded around multiples of the Nyquist frequency and the higher harmonics will now appear as aliasing components. However by choosing a higher sampling frequency, the amplitude of the aliasing components can be made lower.

If the rectified signal in Fig. 15(c) is squared, the time function will no longer have the sharp cusps and the whole signal is smooth as shown in Fig. 15(d). This squared function is simply a cosine signal with double the original frequency and a DC offset.

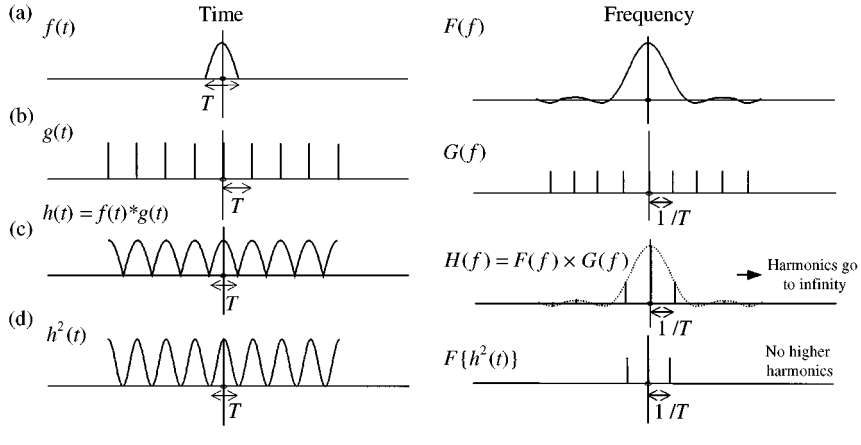

Figure 15. The effects of rectifying a cosine function.

6.3. DISCRETE FREQUENCY COMPONENTS

The benefits of using appropriate zero padding for band-pass rectification are shown in the following sections using simulated and actual vibration signals. It illustrates how envelope analysis by the band-pass rectification technique can produce the same result as the Hilbert transform technique for the frequency range of interest.

Figure 16(a) shows a spectrum containing three discrete frequency components with 1 kHz sampling frequency. These three discrete frequency components may be masking components in the same frequency region as the bearing fault components. Comparisons are made between envelope analysis by different techniques to investigate which produces the worst masking. A band from 10 to 30 Hz was chosen for envelope analysis and Fig. 16(b)–(i) shows the results of the four methods that may be used: (i) Hilbert technique, (ii) Band-pass rectification using zero padding above and below the demodulation band, (iii) band-pass rectification with zero padding above the demodulation band only, and (iv) no zero padding [21]. The left-hand column has the spectra of the squared envelope while the right-hand column has the spectra of the envelope. By using the Hilbert technique, only the difference frequency components are present in the spectrum of the squared envelope [Fig. 16(b)] and the square root process in rectification introduces their higher harmonics [Fig. 16(c)].

Band-pass rectification with zero padding above and below the demodulation band will give the same spectra as the Hilbert technique but with two exceptions: (i) the presence of the pseudo-sum frequencies outside the range of interest [Fig. 16(d)], and (ii) the spectrum of the envelope has higher harmonics of the pseudo-sum frequencies and mixing of the sum and difference frequency components [Fig. 16(e)].

Band-pass rectification with zero padding only above the demodulation band will force the pseudo-sum frequency components to mix with the difference frequency components in the frequency range of interest as shown in Fig. 16(f) and (g). If no zero padding is used, all the frequency components above 20 Hz in Fig. 16(f) and (g) will be folded back into the frequency span of the envelope spectrum. This is shown in Fig. 16(h) and (i). Comparison of the different techniques in Fig. 16 shows that for the three discrete frequency masking components, the Hilbert technique would have the least interference in the envelope spectrum but a comparable result can be obtained through band-pass rectification with appropriate zero padding.

6.4. RANDOM COMPONENTS

For random noise components, the interference in the envelope spectrum is shown in Fig. 17. The noise in the frequency range of interest decays at the same rate for the Hilbert technique and the band-pass rectification with appropriate zero padding. For the other two methods, although the noise starts at the same level at zero frequency, it extends over double the bandwidth.

Band-pass rectification with appropriate zero padding can be modelled as the convolution of two consecutive rectangles because of the zero padding around the zero frequency [Fig. 17(a)]. This produces a triangular spectrum, which slopes to zero after a shift of 'B'. For band-pass rectification with zero padding only above the demodulation band [Fig. 17(b)], the noise decays to zero at the Nyquist frequency ('$2B$'). This means the noise is decaying half as fast as the previous case. If no zero padding is used for band-pass rectification, then all the noise components above the Nyquist frequency in the previous case will be folded back so that they add up with the frequency components in the bandwidth 'B'. This may mask higher harmonics of the bearing fault frequency components if the signal does not have a sufficient SNR ratio. A comparison of the random noise

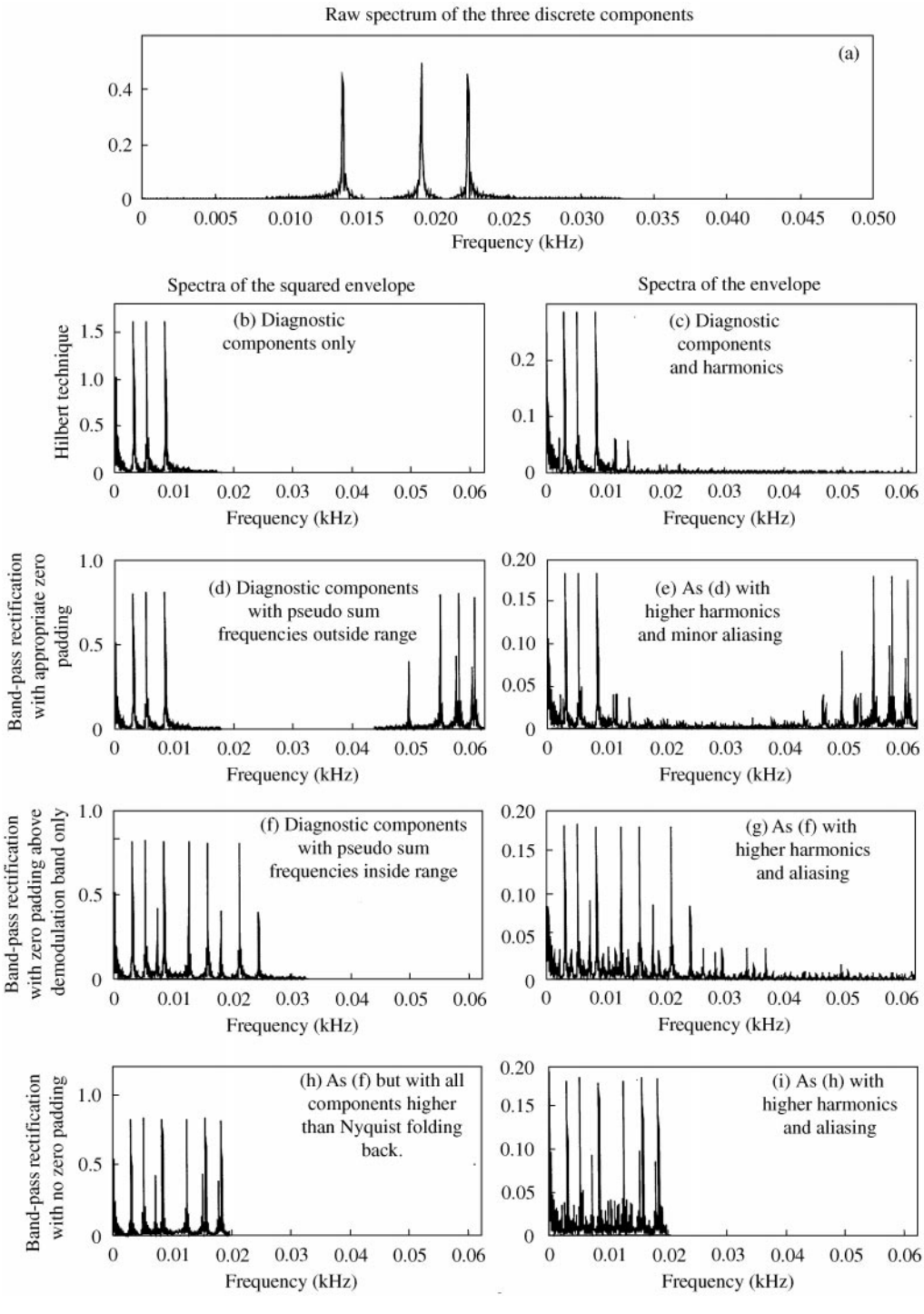

Figure 16. (a) Raw spectrum of three discrete frequency components. (b) & (c) Envelope analysis using the Hilbert technique. (d) & (e) Envelope analysis using band-pass rectification with zero padding above and below demodulation band. (f) & (g) Envelope analysis using band-pass rectification with zero padding above demodulation band only. (h), (i) Envelope analysis using band-pass rectification with non zero padding.

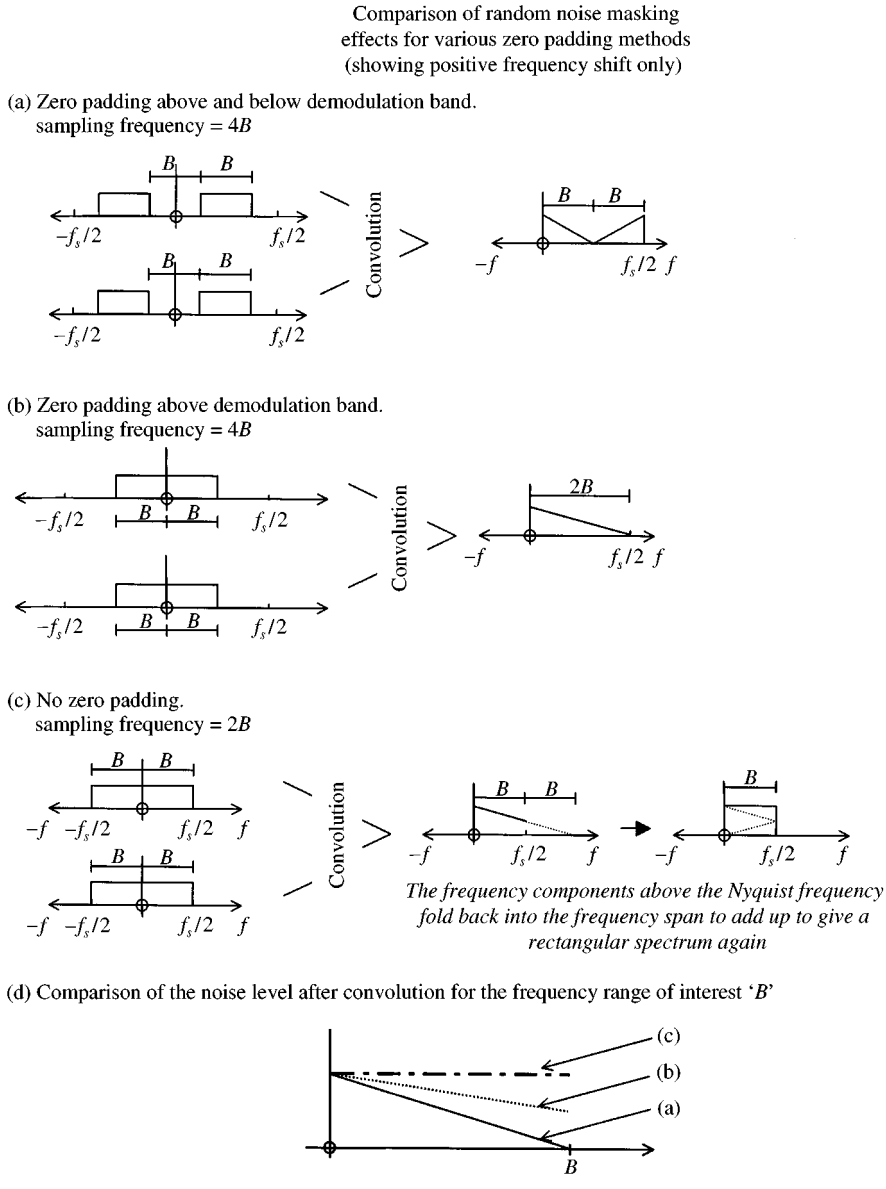

Figure 17. Comparison of random noise making effects for various zero padding methods.

masking in the envelope spectrum as a result of the convolution for the three cases is shown in Fig. 17(d).

6.5. SIMULATION OF COMBINED DISCRETE AND RANDOM MASKING

A simulated bearing fault signal obtained using the model described in Section 2 was mixed with two families of discrete frequency components and random noise for a frequency range around the bearing fault's resonance frequency. Figure 18(a) shows the spectrum of the band to be demodulated. The broad-band, low-amplitude part of the spectrum is

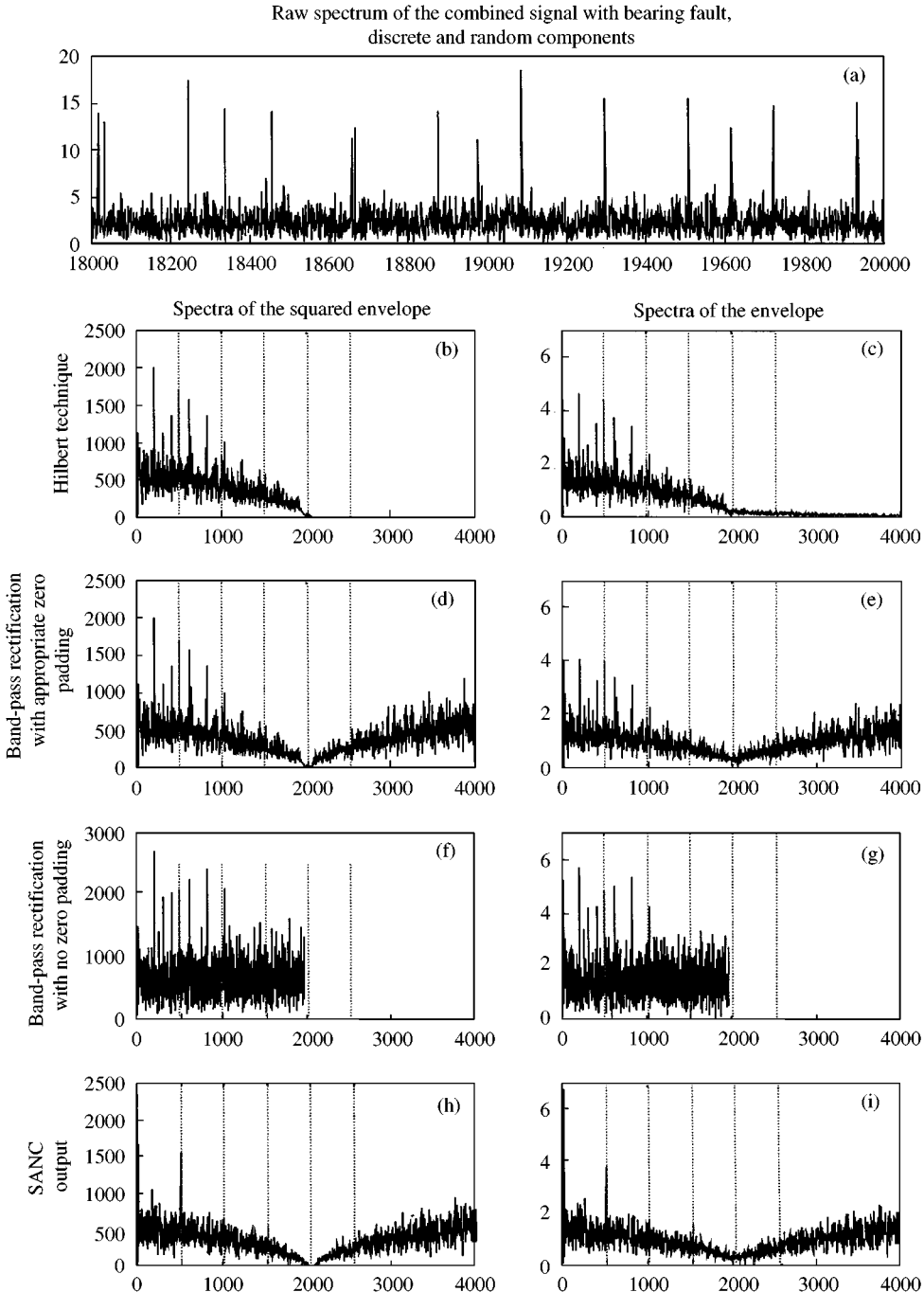

Figure 18. (a) Raw spectrum of bearing components, random an discrete frequency masking noise. (b), (c) Envelope analysis using the Hilbert technique. (d), (e) Envelope analysis using band-pass rectification with zero padding above and below demodulation band. (f), (g) Envelope analysis using band-pass rectification with no zero padding. (h), (i) SANC output.

a combination of the bearing fault signal and the random noise. The former is smeared because the spacings of the impulses have a 2% random fluctuation. The spectra of the squared envelope and the spectra of the envelope are shown in the same format as before

but the band-pass rectification method with zero padding only above the demodulation band has been omitted.

Again similar observations can be made for the various techniques. With appropriate zero padding, band-pass rectification gives the same spectrum as the Hilbert technique for the frequency range of interest. For the case where no zero padding was used, the random noise remained at the same level across the whole frequency span whereas the noise in the first two techniques decreased to zero over the frequency range of interest. In these simulations, the third harmonic of the ball-pass frequency has a low amplitude and is not prominent even using the Hilbert technique. However, a case may occur where the higher harmonics of the ball-pass frequency could be seen more clearly if the background noise were decaying as shown in Fig. 18(b)–(e). Figure 18(h) and (i) shows the envelope spectrum of the output from the SANC process where the discrete frequency masking components have been removed leaving the first three harmonics of the ball-pass frequency. The envelope analysis technique used in this case is band-pass rectification with appropriate zero padding.

6.6. PRACTICAL EXAMPLES

The various envelope techniques were applied to a practical signal recorded from a rig which is used to test faults in actual helicopter gearboxes [22]. There was a roller fault in one of the taper roller bearings deep inside the gearbox. By demodulating a frequency band from 9 to 11 kHz using the Hilbert technique, it was possible to detect the first two harmonics of the ball spin frequency (BSF) and their sidebands spaced at the cage frequency [Fig. 19(a)]. The same envelope spectrum can be obtained with the band-pass rectification method with zero padding above and below the demodulation band [Fig. 19(b)]. However, for the band-pass rectification method without zero padding there is a lot more background noise in the envelope spectrum as shown in Fig. 19(c). This is due to the interaction between the positive and negative frequency components in the convolution process. Although the harmonics of the BSF are still above the background noise, some of the sidebands are masked or are now at the same level as the background noise. This may cause a problem for automated recognition techniques which attempt to identify a certain pattern of frequency components above the background noise.

This situation would not be improved with the application of SANC because the masking frequency components have random properties. Figure 20 shows the benefit that can be gained from using SANC if the masking components are discrete. Signals were recorded from a parallel shaft gearbox rig, which has a seeded outer race fault in one of its bearings. Figure 20(a) shows the raw spectrum of the vibration signal and the subsequent demodulation band. The spectrum of the envelope obtained using the Hilbert technique is shown in Fig. 20(b) where the first two harmonics of the ball-pass frequency outer race (BPFO) are just above the background noise. By zero padding the demodulation band as suggested in Section 5.1, a real signal was formed and this was then entered into the SANC process. The spectrum of the rectified SANC output is shown in Fig. 20(c) where the first 5 harmonics of the BPFO can now be seen clearly. After inspecting the raw spectrum more closely, it was found that the 20th harmonic of the gearmesh was a dominant feature in the demodulation band and this included the shaft modulations which appeared as sidebands. The SANC process was able to remove the discrete shaft frequency modulations which masked the bearing fault frequency components.

6.7. EFFICIENCY OF THE RECOMMENDED WAY OF PADDING WITH ZEROS

To use the efficient fast Fourier transform (FFT), it is desirable to work with a number of samples that is a power of 2 even if this means padding zeros to the next largest power of 2.

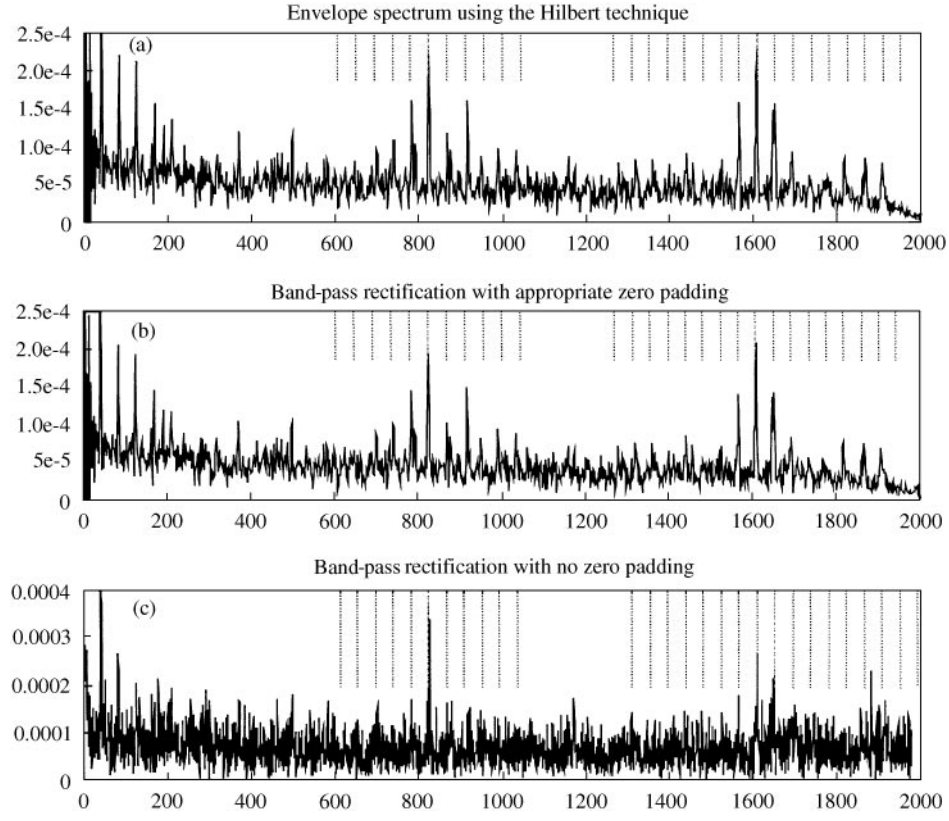

Figure 19. Signals recorded from a helicopter gearbox. (a) Spectrum of the envelope obtained using the Hilbert technique. (b) Spectrum of the envelope obtained using band-pass rectification with recommended zero padding. (c) Spectrum of the envelope obtained using band-pass rectification without zero padding.

Thus, whenever the recommended way of padding with zeros is implemented, the total number of samples in the band-pass filtered time signal should be a power of 2. Figure 21 shows that the total number of samples of a band-pass filtered time signal with appropriate zero padding will be four times the next highest power of 2 from the number of samples in the chosen frequency band [18]. The vertical dashed line indicates the number of samples that would be processed by a digital filter if the total time record had 65 536 samples, it will shift left or right depending on the number of samples in the time record.

It can be clearly seen that for a small frequency band with many fewer samples than the total time record, the zero padded method explained in the previous section uses less samples than the digital filtering method when implementing band-pass filtered and frequency-shifted rectification. The efficiency advantage of using the zero padding method reduces as the size of the chosen frequency band increases. For a chosen frequency band equivalent to a quarter of the sampling frequency, both methods use the same number of samples and if the band is wider still, the number of samples used by the zero padding method exceeds the digital filtering method. However, the former technique ensures that the envelope analysis process is free from sum frequency components, which is not guaranteed by the latter. Also, even though the number of samples used is larger, the FFT processes involved in the zero padding technique are much faster than the convolution process of the digital filter; thus padding with zeros for frequency bands over a quarter of the sampling frequency is still recommended.

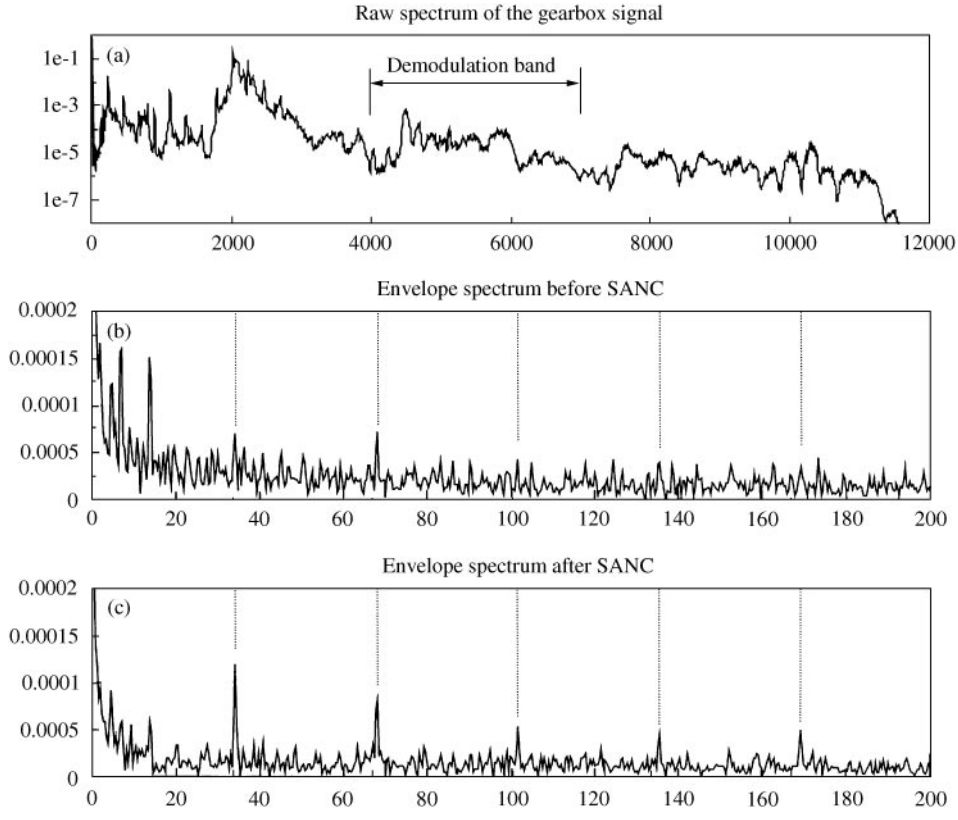

Figure 20. Signal obtained from a parallel shaft gearbox. (a) Original spectrum – demodulation band: 4–7 kHz. (b) Spectrum of the envelope obtained using the Hilbert technique. (c) Spectrum of the envelope after SANC obtained using band-pass rectification with zero padding.

7. CONCLUSION

This paper has investigated how bearing faults can be simulated digitally, including the random fluctuations in spacing of the excitation pulses resulting from clearances in the bearing and the changing load angle experienced by each rolling element. The random fluctuation forces the higher harmonics of the ball-pass frequency to smear over one another in the raw spectrum but the repetition frequency of the excitation impulses is still clear in the envelope spectrum.

Only factors which affect the way bearing faults are manifested in the envelope spectrum were considered in the model; therefore details such as the separation between the point of entry and point of impact, and multiple resonances were not included. The model uses the impulse response of a sdof system to generate a series of pulses to simulate the successive impacts between the rolling elements and the bearing fault. The restriction of the frequency content by a finite sampling frequency meant that a low-pass filter was applied to the signal, but with zero phase shift. The resulting non-causal filter response was compensated by applying a linear phase shift to the spectrum which moved the entire impulse response to the positive time domain.

The simulated bearing fault signals were used to investigate when analysing the squared envelope would improve the envelope spectrum with respect to the background noise. It was found that in general if the random spacing fluctuation of the bearing fault signal is less

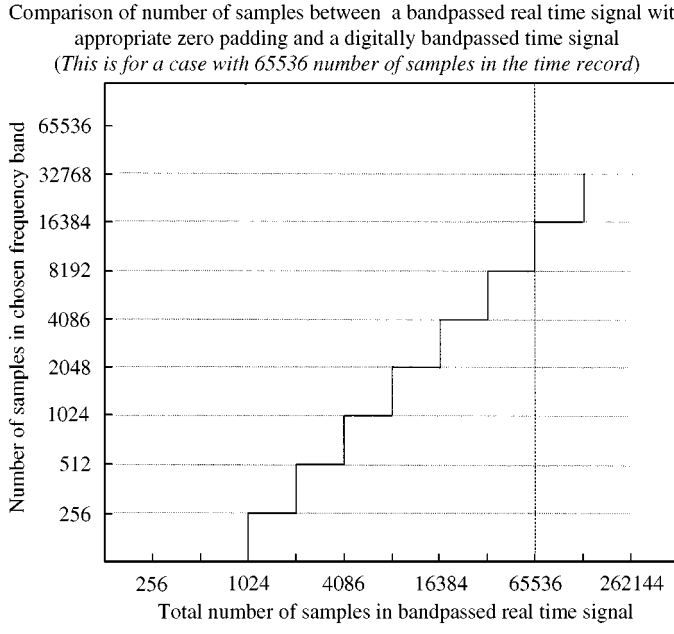

Figure 21. Efficiency of recommended zero padding compared with digital filtering. ——, recommended padding with zeros to avoid aliasing, -------, digital filtering using all samples.

than 1%, then analysing the squared envelope will give an improvement if the MSR of the bearing signal to noise in the demodulated band is greater than 0.2. The study also takes into account the random fluctuation of the bearing signal. Analysing the envelope raised to the fourth power gives a further improvement in the spectrum if the SNR is above unity.

In using the squared envelope technique, it is important to zero pad above and below the demodulated band so that the pseudo-sum frequencies will not fall within the frequency range of interest in the spectrum of the squared envelope.

Techniques such as SANC can be used to remove discrete frequency noise from the signal before performing envelope analysis. Since this requires real-valued signals as inputs, band-pass rectification is a faster way of performing envelope analysis than the Hilbert transform technique. However, band-pass rectification will also need the extra zero padding used for analysing the squared envelope in order to move the pseudo-sum frequencies outside the frequency range of interest and thus time records will have double the number of samples.

Analysing the squared envelope rather than the envelope was found to have the advantage of not including the higher harmonics of the sum frequency components in the envelope spectrum. This is because rectification produces higher harmonics, and in addition some of these are aliased if they are above the Nyquist frequency. This effect cannot be removed from digital signals by low-pass filtering but can only be reduced by using a higher sampling frequency.

The Hilbert transform technique uses the one-sided spectrum of the demodulation band to generate an analytic signal whose amplitude is the modulation signal. By multiplying the analytic signal (which is complex) with its complex conjugate, the convolution effect in the frequency domain only gives the difference frequency components. If the square root operation is then applied, higher harmonics of the difference frequency components will be present.

ACKNOWLEDGEMENTS

This research was conducted with the support of Aeronautical and Maritime Research Laboratory (AMRL), Melbourne, Australia. The helicopter gearbox signals were supplied by the Naval Air Warfare Center Aircraft Division and made under the Helicopter Integrated Diagnostic Systems (HIDS) program. The HIDS data were acquired at their Helicopter Transmission Test Facility, formerly of Trenton, New Jersey, and now relocated to Patuxent River, Maryland.

The Association of American Railways is acknowledged for their supply of rail vehicle bearing measurements from a test rig.

REFERENCES

1. P. D. McFadden and J. D. Smith 1984 *Tribology International* **17**, 3–10. Vibration monitoring of rolling element bearings by the high frequency resonance technique—a review.
2. R. B. Randall 1997 *5th International Congress on Sound and Vibration, Adelaide, Australia.* Developments in digital analysis techniques for diagnosis of bearings and gears.
3. I. K. Epps and H. McCallion *4th Annual Vibration Association of New Zealand Conference, Christchurch, New Zealand*, May 1993.
4. D. Core and G. Edgar 1984 *RCA Engineer* **29**(5), 71–77. Techniques for the early detection of rolling-element bearing failures.
5. P. D. McFadden and J. D. Smith 1984 *Journal of Sound and Vibration* **96**, 69–82. Model for the vibration produced by a single point defect in a rolling element bearing.
6. P. D. McFadden and J. D. Smith 1985 *Journal of Sound and Vibration* **98**, 263–273. The vibration produced by multiple point defects in a rolling element bearing.
7. Y. T. Su and S. J. Lin 1992 *Journal of Sound and Vibration* **155**, 75–84. On initial detection of a tapered roller bearing: frequency domain analysis.
8. R. B. Randall 1987 *Frequency Analysis 3rd edn*, pp. 179–180. Copenhagen: Bruel & Kjaer.
9. A. Barkov and N. Barkova, 1995 *Sound and Vibration* June 10–17. Condition assessment and life prediction of rolling element bearings—Part 1.
10. Y. Gao, R. A. J. Ford and R. B. Randall 1996 *Transactions of the Institution of Engineers, Australia* **ME21**, 137–145. Multi-carrier demodulation of amplitude modulations by rectification—application to detect mechanical faults.
11. R. B. Randall 1986 *IFToMM International Conference on Rotordynamics, Tokyo.* Hilbert transform techniques in machine diagnostics.
12. G. K. Chaturvedi and D. W. Thomas 1982 *Journal of Sound and Vibration* **104**, 280–289. Bearing fault detection using adaptive noise cancelling.
13. C. C. Tan and B. Dawson 1987 *International Tribology Conference, Melbourne.* An adaptive noise cancellation approach for condition monitoring of gearbox bearings.
14. R. B. Randall and Y. Li 1995 *2nd International Conference on Gearbox Vibration and Diagnostics, ImechE, London*, 73–80. Diagnostic of planetary gear bearings in the presence of gear vibrations.
15. D. Ho and R. B. Randall 1997 *5th International Congress on Sound and Vibration, Adelaide, Australia.* Effects of time delay, order of fir filter and convergence factor on self adaptive noise cancellation.
16. B. Widrow and S. Stearns 1985 *Adaptive Signal Processing*, pp. 349–351. Englewood Cliffs NJ: Prentice-Hall.
17. S. Stearns and R. David 1988 *Signal Processing Algorithms*, pp. 252–269. Englewood Cliffs NJ: Prentice-Hall.
18. D. Ho and R. B. Randall 1998 *Comadem '98 Launceston, Australia.* Improving the efficiency of SANC in its application to bearing diagnostics.
19. R. B. Randall, D. Ho, Y. Gao 1998 *ISMA 23: Noise and Vibration Engineering Conference, Leuven, Belgium.* Bearing diagnostics by digital analysis techniques in high background noise situations.
20. R. B. Randall and Y. Gao 1996 *6th International Conference on Vibrations in Rotating Machinery, ImechE, Oxford*, 351–359. Masking effects in digital envelope analysis of faulty bearing signals.

21. D. HO, R. B. RANDALL and Y. GAO 1999 *The 2nd Australian Congress on Applied Mechanics ACAM99 Canberra, Australia*. Comparison of envelope analysis by the Hilbert transform technique vs rectification.
22. R. B. RANDALL and Y. GAO 1998 *Comadem '98 Launceston, Australia*. Using CPB spectrum differences to train neural network to recognise faults in helicopter gearbox bearings.