Detecting Rolling Elements Bearings Faults

Chapter · February 2020

1 author:

Jaafar Alsalaet
University of Basrah

Detecting Rolling Elements Bearings Faults

Dr. Jaafar Alsalaet

College of Engineering-University of Basrah

Abstract

Rolling elements bearings are the most important machine elements. Failure of these elements may cause catastrophic breakdown and result in costly downtime. Bearing condition monitoring thus plays an important role in the current machinery maintenance strategies. Due to the complex structure of rolling elements bearings, the observed vibration signals are normally corrupted by noise and random patterns especially in the early stages of bearing defects. It is important to early discover the incipient bearing anomaly before problem development which may lead to catastrophic failure. A number of signal analysis techniques, both in time and frequency domains, have been introduced to extract useful information from the noisy vibration signal. In this workshop, some traditional and advanced bearing vibration analysis techniques will be explored with some practical examples.

Contents

1. Introduction

All rotating machines have bearings to support the rotating parts and isolate them from the stationary parts. Most of these bearings are rolling-element bearings. A rolling element bearing comprises of inner and outer races, a cage and rolling elements. There are five types of rolling elements that are used in rolling-element bearings: balls, cylindrical rollers, spherical rollers, tapered rollers, and needle rollers.

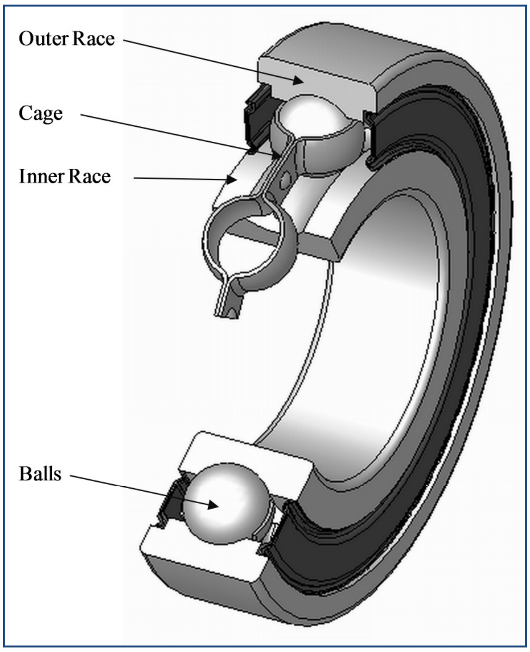

Figure 1.1 Deep Groove Ball Bearing

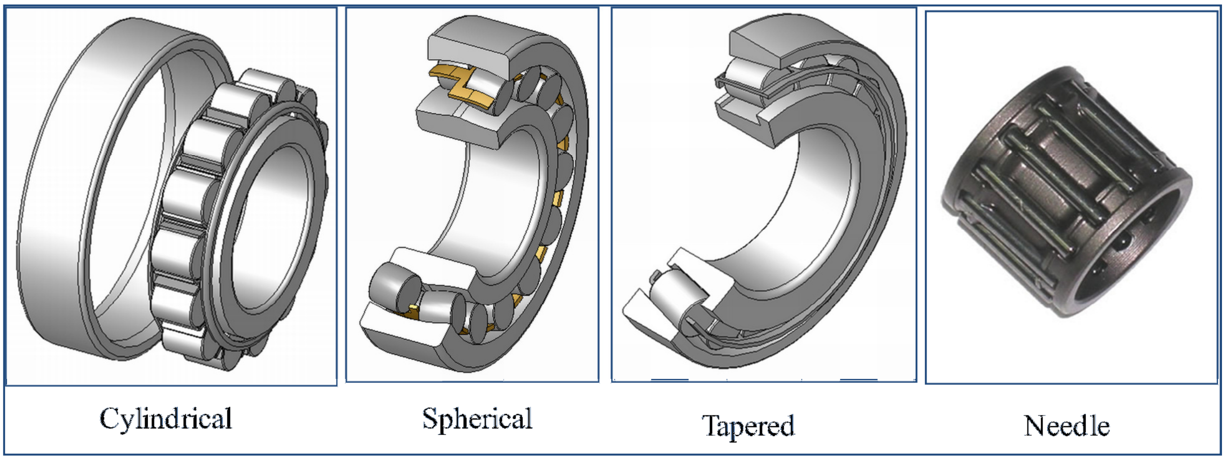

Figure 1.2 Types of Rolling Bearings

Bearing Vibration Patterns

Defects can occur in any of the parts of the bearing and will cause high-frequency vibrations. In fact, the severity of the wear keeps changing the vibration pattern. In most cases, it is possible to identify the component of the bearing that is defective due to the specific vibration frequencies that are excited. Raceways and rolling element defects are easily detected. However, the same cannot be said for the defects that crop up in bearing cages. Though there are many techniques available to detect where defects are occurring, there are no established techniques to predict when the bearing defect will turn into a functional failure.

In section 3 and next section, we will introduce modulation process and how vibration is excited by bearing faults. We will see how bearing defects generate both the bearing defect frequency and the ringing random vibrations that are the resonant frequencies of the bearing components. Bearing defect frequencies are not integrally harmonics to running speed. However, the following formulas are used to determine bearing defect frequencies. There is also bearings database available in the form of commercial software that readily provides the values upon entering the requisite bearing number.

$$
\begin{aligned}
BPFI &= \frac{N}{2}\left(1+\frac{D_b}{D_p}\cos\theta\right)\times RPM \\
BPFO &= \frac{N}{2}\left(1-\frac{D_b}{D_p}\cos\theta\right)\times RPM \\
FTF &= \frac{1}{2}\left(1-\frac{D_b}{D_p}\cos\theta\right)\times RPM \\
BSF &= \frac{D_p}{2D_b}\left(1-\left(\frac{D_b}{D_p}\cos\theta\right)^2\right)\times RPM
\end{aligned}
\tag{1.1}
$$

Where D_b : ball or roller diameter

D_p : pitch circle diameter of the bearing

N : number of balls

θ : contact angle

BPFI : Ball Pass Frequency (Inner Race)

BPFO : Ball Pass Frequency (Outer Race)

FTF : Fundamental Train Frequency (Cage)

BSF : Ball Spin Frequency (Rolling Element)

Rules

- The sum of BPFI and BPFO equals to the number of rolling elements
- *BPFO = No. of rolling elements x FTF*
- Simpler equation to estimate the frequencies (for N between 8 and 12 element):

$$BPFO \approx 0.4 \times N$$

$$BPFI \approx 0.6 \times N$$

Sources of Vibration in a Roller Bearing

1. Friction: when two surface are sliding one on another, there will be a friction regardless whether there is a lubricant or not and the specifications of the lubricant. Vibration due to friction is random and not periodic. It generates broadband noise, vibration at wide range of frequencies.
2. Stress waves: if we have metal-to-metal contact, we will get very short duration pulses which excite high frequency. The pulses may be periodic or not.
3. Resonance: when surfaces impact, wide range of frequencies is generated. This will excite the natural frequencies of the machine components. The natural frequencies may be very high for roller bearings (above 20 kHz). These very high frequencies can be detected using special techniques and will be helpful in determining bearing condition as will be discussed later.

Bearing Failure Stages

Detecting rolling element bearing faults is the highest priority for most vibration analysts. Detecting the fault at the earliest stage is a priority for maintenance engineer. Before we get into the specifics of the four stages of bearing failure, I would like to describe how the vibration changes in general during operation.

If a bearing is poorly lubricated, we can detect an increase in the level of “noise” at very high frequencies. It is not a specific, single frequency; instead it will depend on a number of factors to do with the machine’s construction. Suffice to say that you cannot hear it; it is well above your hearing range.

As the state of lubrication worsens, the level of the noise will increase, but the frequency of the noise will slowly reduce – it will move from very high frequencies to high frequencies. That is not to say that you can't detect the condition at lower frequencies; it is stronger at the higher frequencies.

As the film of lubricant between the bearing surfaces is reduced further, we will have more and more metal-to-metal contact, causing "stress waves" to be generated. Stress waves (also referred to as "shock pulses") are like ripples in a pond; the moment the metal surfaces make contact, a wave of energy races away from the point of contact at the speed of sound. It all happens very quickly.

The subsurface defects will slowly develop due to the extreme forces experienced within the bearing. The difference is that these defects are likely to be localized; at the bottom of the outer race for example. The "noise" from the bearing due to poor lubrication is relatively constant (it is random, therefore not periodic), whereas when a fault condition develops (e.g. a crack or spall), a new source of periodic vibration will be introduced; that are bearing defect frequencies in eq. (1.1). If the damage was on the outer race, each time a rolling element passes that location there will be a spike in the vibration. When the point of damage is between rolling elements, there is no vibration (well, less vibration). The good news is that we can calculate the frequency of this vibration (we can determine how often the rolling element will pass that point). The bad news is that the vibration is very, very low in amplitude.

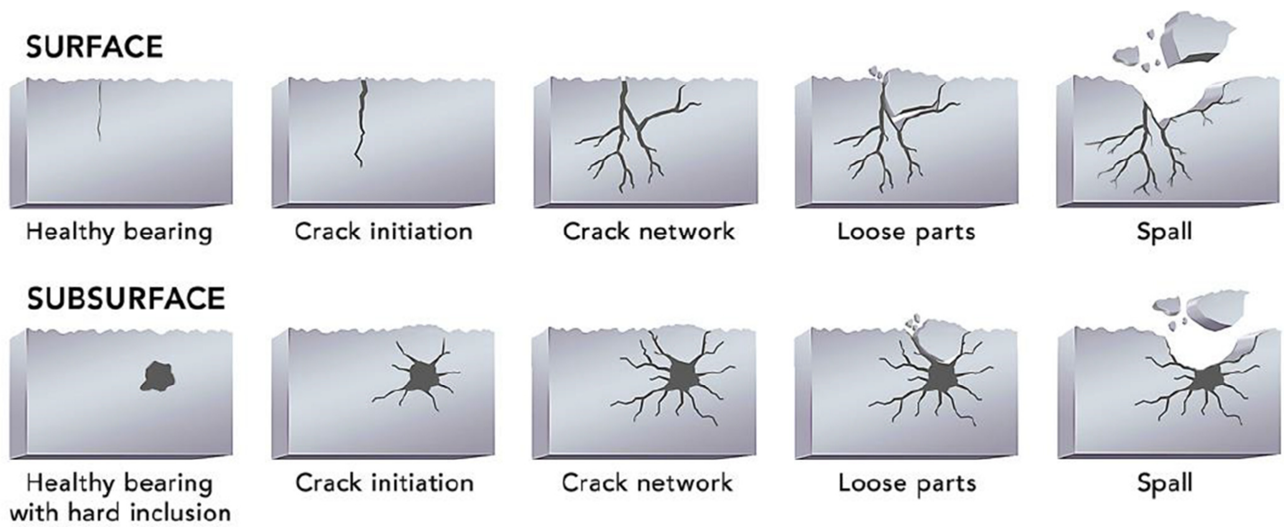

Figure 1.3 Crack and Spall Development [2]

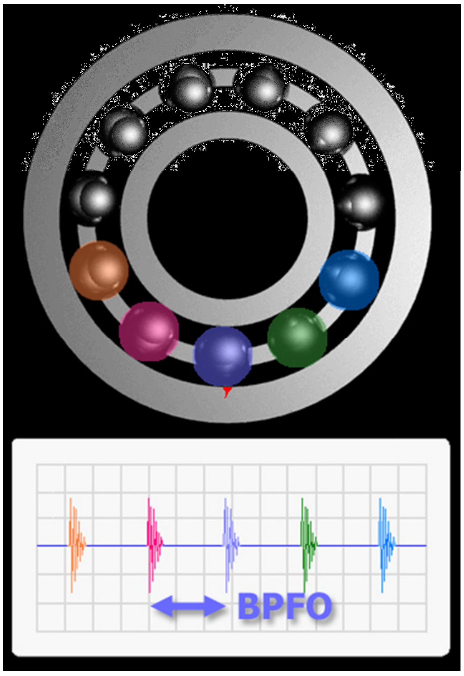

Figure 1.4 Defect in the Bottom of Outer Race [1]

Bearing failure stages are divided classically into four stages.

Stage One

In this stage the defect is minor. If the bearing is removed, there will be no signs of damage as the damage is predominantly sub-surface. The metal-to-metal contact causes random vibration of high frequencies (5 to 40 kHz). The high-frequency detection techniques such as Shock Pulse, Spike Energy, ultrasound, and PeakVue and ultrasound measurement (acoustic emission) are effective in this stage.

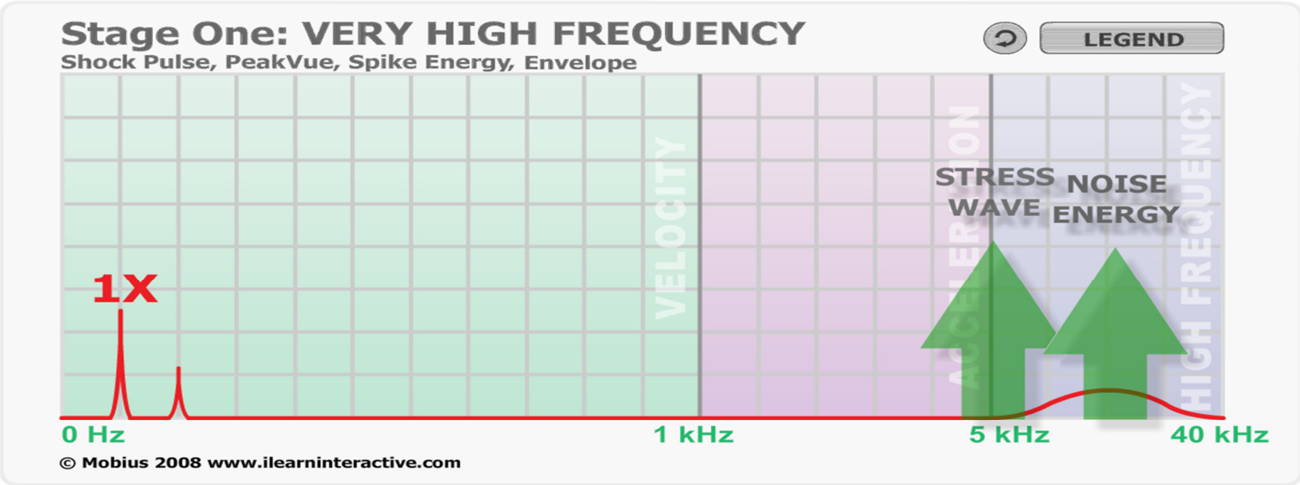

Figure 1.5 Stage One of Bearing Failure [1]

Stage Two

As the fault develops, the sub-surface defects will grow, eventually breaking through to the surface, causing spalls, cracks, flakes, etc. The forces of the impacting will be greater, and there will definitely be periodicity to the vibration. Bearing resonance frequencies may be excited that will show peaks in the frequency range 1000 to 5000 Hz. The high frequency techniques will continue to be effective. Enveloping (demodulation) will also be effective, with peaks visible at the bearing forcing frequencies (BPFO, BPFI, BSF and FT – depending on the nature of the fault) along with harmonics. Harmonics of the bearing forcing frequencies may also be visible in the acceleration spectrum.

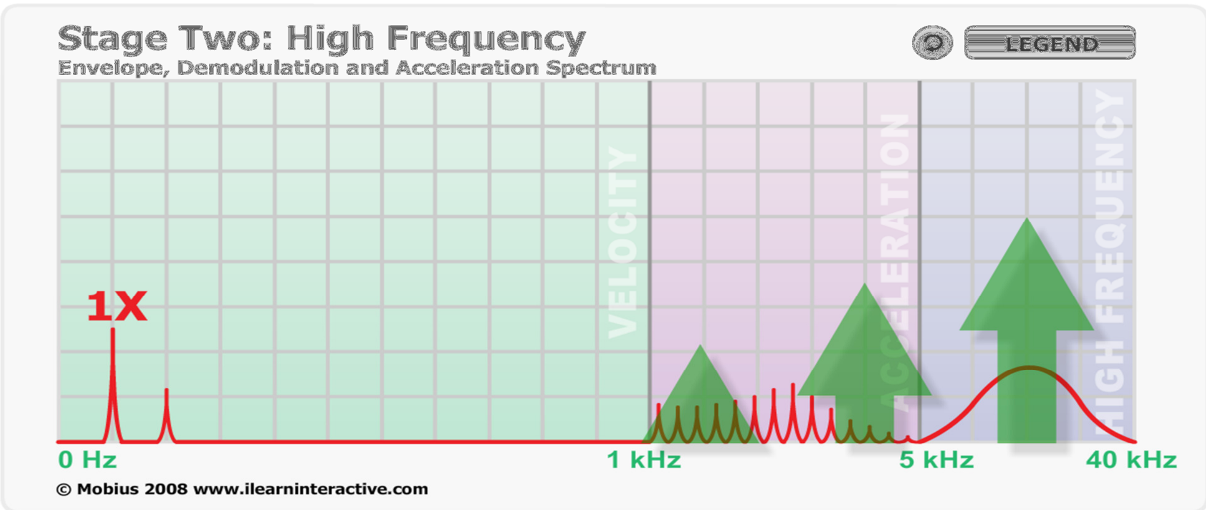

Figure 1.6 Stage Two of Bearing Failure [1]

Stage Three

In this stage, wear is usually now visible on the bearing and may expand through to the edge of the bearing raceway. The high frequency techniques will still indicate the presence of a fault. The peaks in the envelope spectrum will continue to grow in amplitude. There will be peaks in the velocity spectrum that correspond to the bearing forcing frequencies (BPFO, BPFI, BSF and FT) and their harmonics depending on the fault condition. In some cases, sidebands around the bearing defect frequency may appear due to amplitude modulation of fault.

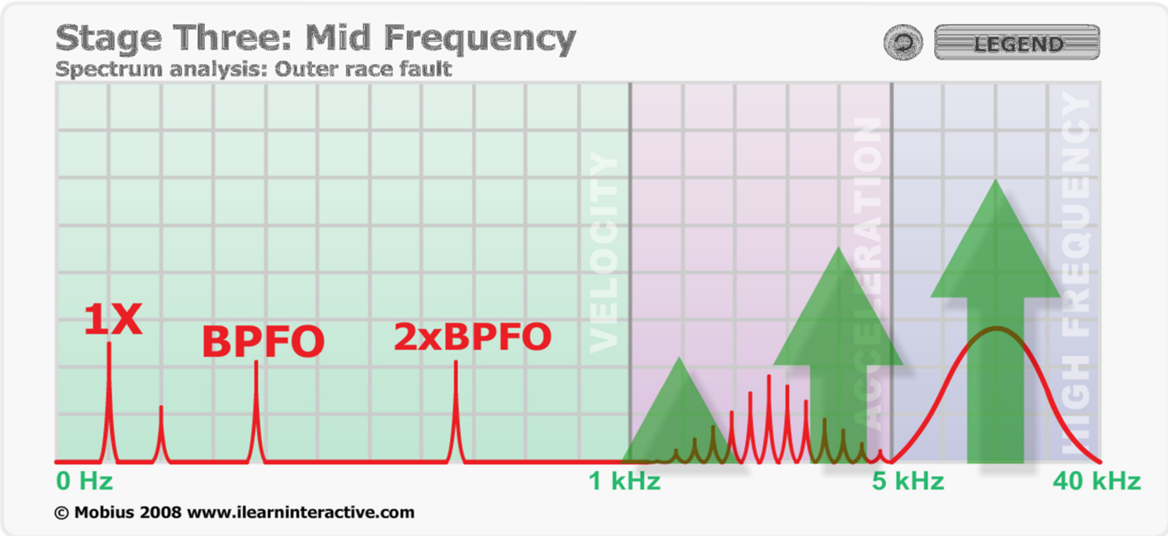

Figure 1.7 Stage Three of Bearing Failure [1]

- Outer race defect: When the outer race become defective, such as a spall in the outer race, there will be BPFO component and when the fault is developed the BPFO harmonics may become significant in the spectrum. If the outer race in rotating, there will be sidebands at outer race rotating frequency due to load variation on the defective portion.
- Inner race defect: If the inner race is defective, there will be BPFI and its harmonics when the fault is developed. Since the inner race is rotating, the load area is also rotating and, hence, vibration amplitude is varying according to the position of the defect or spall. Sidebands will exist around the BPFI and its harmonics, spaced at the rotation frequency.
- Ball defect: If the ball or rolling element is defective, then BSF component is generated with its harmonics. Vibration at 2xBSF may be high due to contact with the inner and outer race. There will be also sidebands at FTF (cage) because the balls are rotating at FTF around the outer race. If the outer race is fixed, the sideband spacing is below 0.5x, but if the outer race is rotating spacing will be higher than 0.5x.

Stage Four

The bearing has considerable damage in this stage. The high frequency detection techniques may still be used but the amplitude of the high-frequency vibration is now reduced due to the smoothing of the sharp edges. There is now so much damage that the vibration loses its periodicity. The peaks at the BDF will drop down, and seemingly random peaks will appear and the noise floor will rise up. This is true for the velocity spectrum and the envelope spectrum.

As more metal is removed, the clearance between bearing parts is increased and the 1X component and its harmonics with noticeable noise floor may be presented due to looseness. The overall vibration velocity RMS is increased.

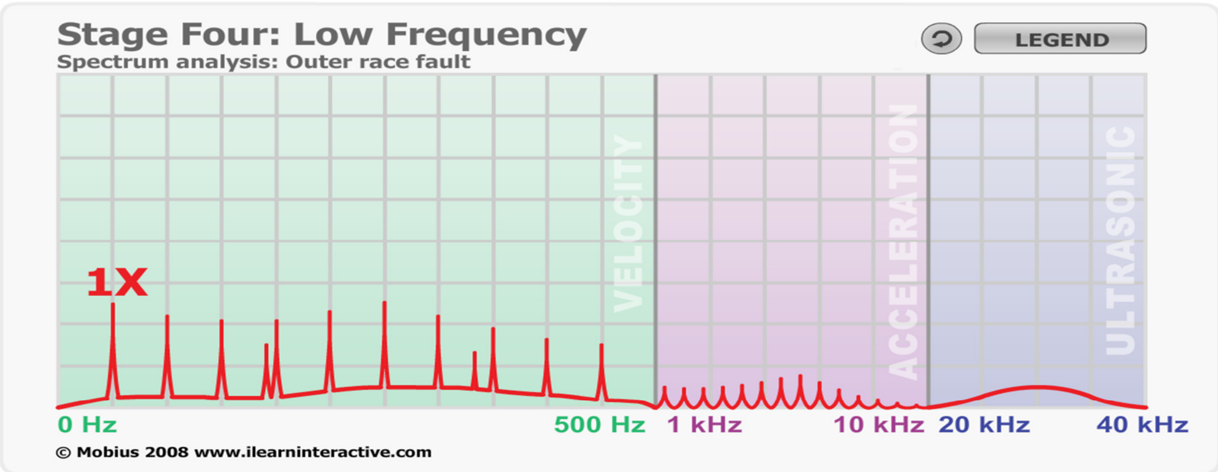

Figure 1.8 Stage Four of Bearing Failure [1]

2. Condition Based Maintenance and Precision Maintenance

If we do a survey of the maintenance philosophies employed by different process plants, we would notice quite a bit of similarity despite the vast variations in the nature of their operations. These maintenance philosophies can usually be divided into four different categories:

- Breakdown or run to failure maintenance
- Preventive or time-based maintenance
- Predictive or condition-based maintenance
- Proactive or prevention maintenance.

Run to failure maintenance

The idea behind run to break maintenance, as the name imply, is to run the machine until failure or breakdown occurs and then replacing the parts that are damaged when the machine is completely stopped. This maintenance procedure is inefficient due to a number of factors:

- Maintenance department perpetually operates in an unplanned 'crisis management' mode.
- It may also lead to sub-sequent failures and, hence, other parts may be damaged resulting in higher maintenance cost.
- When unexpected production interruptions occur, the maintenance activities require a large inventory of spare parts to react immediately.

Preventive maintenance

Also named "planned maintenance", "calendar-based maintenance" and "scheduled maintenance". The philosophy behind preventive maintenance is to schedule maintenance activities at predetermined time intervals, based on calendar days or runtime hours of machines so that they will not fail. This philosophy has less impact on the criticality of the plant used for equipment that does not run continuously, and where the personnel have enough skill, knowledge and time to perform the preventive maintenance work. The main disadvantages are:

- Scheduled maintenance can result in performing maintenance tasks too early or too late.
- It is possible that, without any evidence of functional failure, components are replaced when there is still some residual life left in them.
- It is possible that reduced production could occur due to unnecessary maintenance.
- In many cases, there is also a possibility of diminished performance due to incorrect repair methods.

Condition based maintenance

Mechanical and operational conditions are periodically monitored, and when unhealthy trends are detected, the troublesome parts in the machine are identified and scheduled for maintenance. The machine would then be shut down at a time when it is most convenient, and the damaged components would be replaced.

- ✓ When the plant runs in predictive mode, the condition of the running machines is known.
- ✓ The unexpected failure can be prohibited or at least reduced so that it is not frequently occur.
- ✓ We can plan maintenance schedule and we can plan production schedule around the knowledge of the plant condition. That means "reactive" action is no longer a problem since most work is planned.
- ✓ There is less need of staff overtime and stress is reduced

Proactive or prevention maintenance

This philosophy is based on some strategies such as "Precision Maintenance", "Reliability Based Maintenance" and "Reliability Centered Maintenance". This philosophy emphasis on tracing all failures to their root cause. Each failure is analyzed and proactive measures are taken to ensure that they are not repeated. Proactive maintenance is the strategy used to anticipate the failure before it occurs. It includes a wide variety of practice and strategies and utilizes various technologies. It utilizes all of the predictive/preventive maintenance techniques discussed above in conjunction with root cause failure analysis (RCFA).

RCFA detects and pinpoints the problems that cause defects. It ensures that appropriate installation and repair techniques are adopted and implemented. It may also highlight the need for redesign or modification of equipment to avoid recurrence of such problems. As in the predictive-based program, it is possible to schedule maintenance repairs on equipment in an orderly fashion, but additional efforts are required to provide improvements to reduce or eliminate potential problems from occurring repeatedly. Again, the orderly scheduling of maintenance allows lead-time to purchase parts for the necessary repairs. This reduces the need for a large spare parts inventory, because maintenance work is only performed when it is required.

RCFA requires additional tools and skilled staff in order to be implemented. Signal analyzers such as FFT, signal enveloping, high frequency detection, phase analysis and other techniques may be required to accomplish the tasks.

Regarding bearings problems, precision maintenance implies that in addition to the bearing condition monitoring, it is also important to consider:

- Early detection of bearing faults using HFD techniques for example.

- Fault localization, is it in the outer race, inner race, balls or cage. Is it in the loading area or other area?
- Causes of the fault. Is it due to over-loading of the bearing or other reason such as contamination, lack of lubrication or sparks. Additional tools such as ultrasound measurement to detect sparks and oil analysis to assess oil condition are required to accomplish the task.
- Root cause elimination by proper design/re-design, spark elimination and proper lubrication have to be done before replacing the bearing.

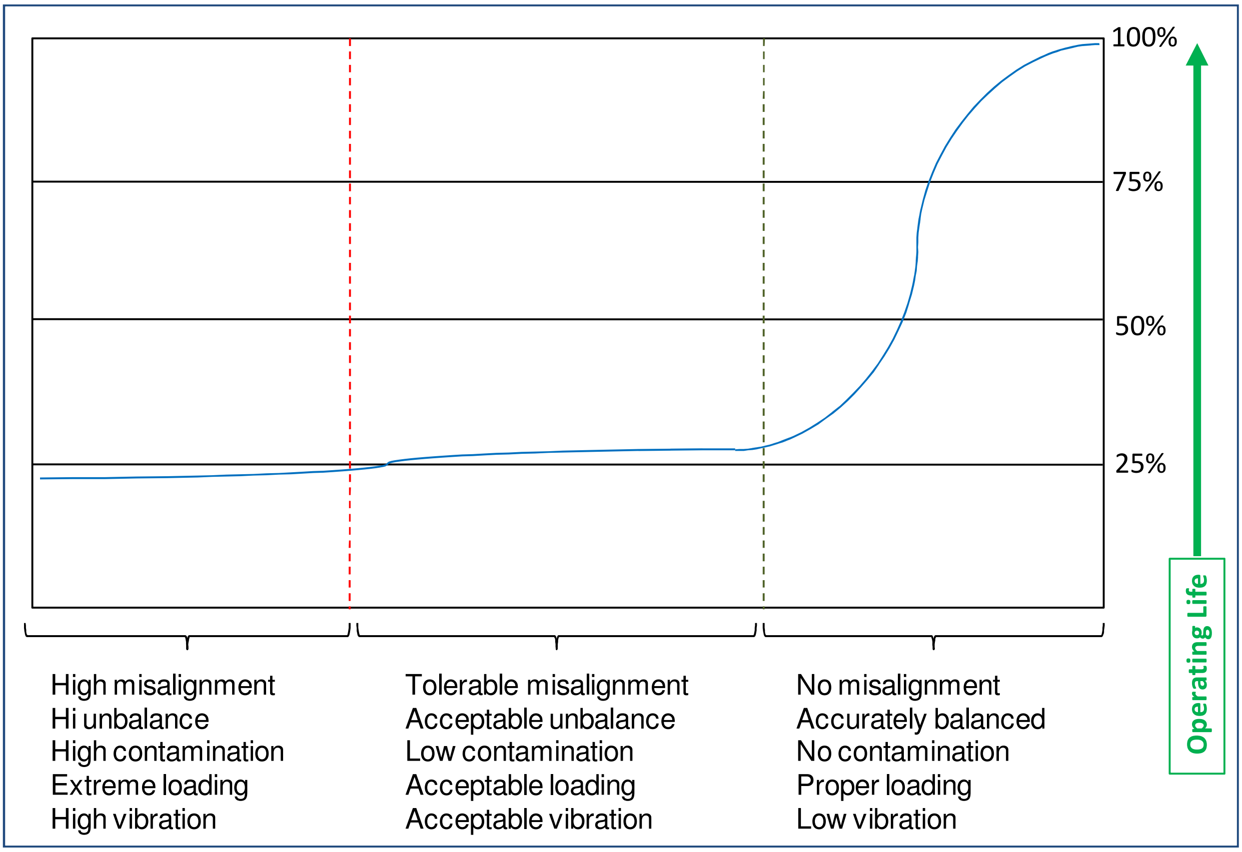

Figure 2.1 Expected machine life vs. operating conditions

3. Time Domain Techniques

RMS, Crest and Kurtosis Factors

Root Mean Square

RMS value is an indication of the power contained in a signal and it is widely used in vibration severity measurement.

$$RMS = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} x_i^{\,2}} \tag{3.1}$$

The overall RMS value of vibration in velocity units is not suitable to detect bearings problems when common 10 to 1000 Hz filter is applied due to the effect of common vibration problems. The RMS of the filtered vibration signal in acceleration can be used to assess bearing condition. For example, the high frequency band of 1 to 20 kHz is used for acceleration severity measurement. The frequency band of 20 to 50 kHz is used in ultrasonic measurement in dB [3]. Spectral emitted energy in the band of 250 to 350 kHz is used by some instruments to measure acoustic emission in dB [4].

Crest Factor

Crest factor is the ratio of the peak value (positive or negative) in a signal to its RMS value. It ranges from 1 for square waves to 1.414 for pure sinusoidal signals and higher values for signals with pulses or short duration activities. Trending of this factor for roller bearings shows an increase of Crest factor (rise of short duration pulses with respect to RMS value) then eventual drop as the fault develops and gets broader. However, there is no vibration analysis software package currently trends Crest factor.

Kurtosis Factor

The kurtosis factor or forth statistical moment is given by the following equation:

$$K = \frac{\frac{1}{N}\sum_{i=0}^{N-1}(x_i - \overline{x})^4}{\left(\frac{1}{N}\sum_{i=0}^{N-1}(x_i - \overline{x})^2\right)^2} \tag{3.2}$$

High kurtosis factor indicates the presence of repeated impulses. It lends itself to the detection of whether the spectrum contains small peaks distributed along a broad frequency range or several peaks positioned at certain locations.

Time Waveform

The time waveform is a plot of the time signal as it is measured. It is rather a complex topic in vibration analysis and used to confirm the results of FFT in some situations such as in bearings and gears fault diagnosis.

Any defect in one (or more) of the bearing parts will excite short duration pulses when the rollers pass over the defect. These impulses will, in turn, excite the bearing natural frequencies as a consequence of impacting. The repetition rate (frequency) of these impulses depends on the rotating speed, bearing geometry and the location of the defect. The duration of these impulses is very short as compared with the interval between them (period); therefore, the energy is distributed at a very low level over the period. As a result, the impulses can easily be obscured by noise or other frequency components.

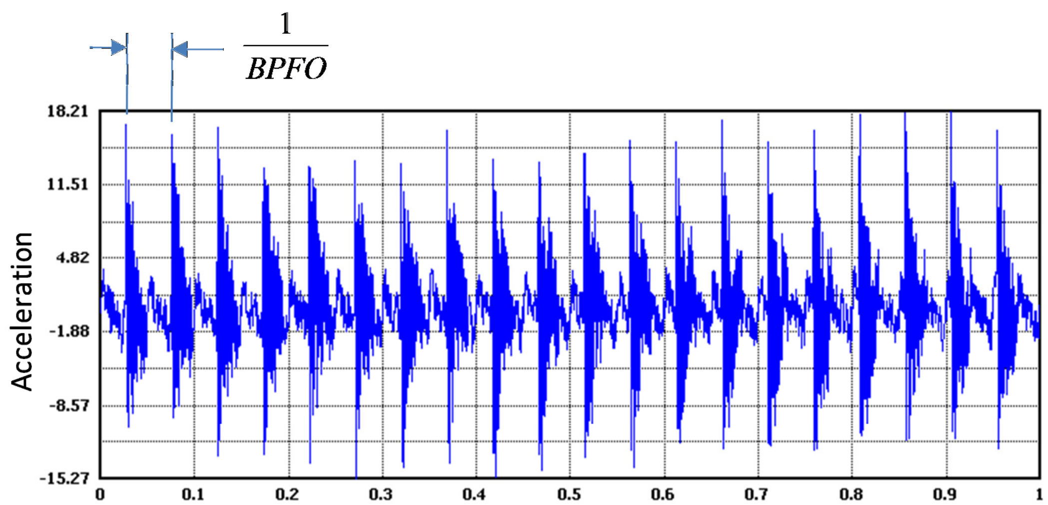

Figure 3.1 Typical time waveform due to defective outer race

Figure 3-1 shows a typical time waveform for a bearing with outer race defect. The repeated pulses are corresponding to the impacts generated when balls passes over the defect (spall or pitfall). The duration between each two successive pulses equals to the reciprocal of BPFO frequency in Hz. However, in many cases bearing vibration pattern is obscured by other patterns and noise and the resulting signal will be random and noisy such that it is difficult to detect bearing patterns unless using other techniques such as enveloping or Wavelet transform. More about time waveform will be coming in the signal envelope detection section 9.

4. Frequency Domain Techniques

FFT Spectrum

FFT spectrum is useful to separate the different frequency signals or components. The spectrum is the plot of the signal components amplitudes vs. frequency. It is obtained from the time waveform by a process called Fourier Transform. Each component in the FFT spectrum has its own frequency and amplitude. The order of the component is simply its frequency divided by the shaft rotational frequency.

FFT spectrum is used to explore the bearing fault stages earlier in sec. 1. It is powerful tool to analyze machinery vibration and diagnose most of problems including bearing problems. However, one must remember the following facts about using FFT spectrum in bearing diagnosis:

1. In the earlier stages of bearing fault, the very high frequency range of 5 to 50 kHz may not be noticeable unless using filtration due to obscuring of the signal by other frequencies. Also, the acceleration units must be used.
2. As the damage develops, bearing defect frequencies begin to appear in the spectrum according to the damage location.
3. Since the signal of vibration due to impacting is not sinusoidal, there will be harmonics of the fault frequency in the spectrum in most cases.
4. If the inner race is rotating, the BPFI component will be amplitude modulated with 1X frequency according to the position of the defect with regard to loading area. This will be reflected as sidebands on both sides of BPFI in the spectrum.
5. If the outer race is rotating, the BPFO component will be amplitude modulated.
6. If one or more ball is defected, the BSF will be amplitude modulated with 1X cage frequency since the balls pass through loading area at a rate of cage frequency.

Figure 4-1 shows the FFT of a bearing with defected rotating inner race. The harmonics of BPFI are more visible when using acceleration units.

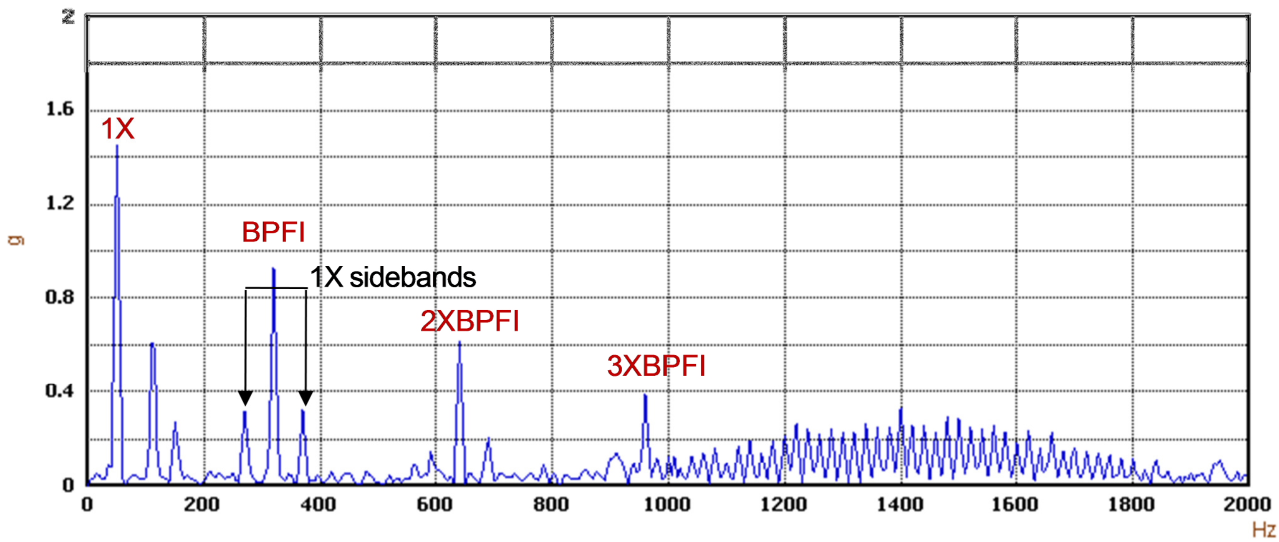

Figure 4.1 FFT of bearing vibration with defected inner race

FFT Waterfall

FFT waterfall is a cascaded view of the successive FFT spectra taken at different intervals. In other words, it is frequency-time representation of a signal. The x-axis represents frequency, y-axis is the amplitude and z-axis is time.

Trending waterfall is the plot that can be used to show the progress of vibration along extended periods of time by comparing FFT spectra along the entire monitoring period, for example along one year or more.

Another waterfall plot is the plot used to represent the signal in time-frequency such that the signal is decomposed to a number of blocks with a constant time interval. The FFTs of these blocks are plotted in cascade fashion. This representation is useful to track changes of the signal along a short period such as during machine run-up or coast-down. We can track the frequency and amplitude of each component and obtain useful information such as resonance conditions. Figure 4-2 shows the FFT waterfall of a faulty bearing where vibration is collected along 8 seconds during run-up and the FFT of signal blocks is obtained each 0.1 sec interval successively and plotted. The variable frequency and amplitude modulation are clearly seen on the plot.

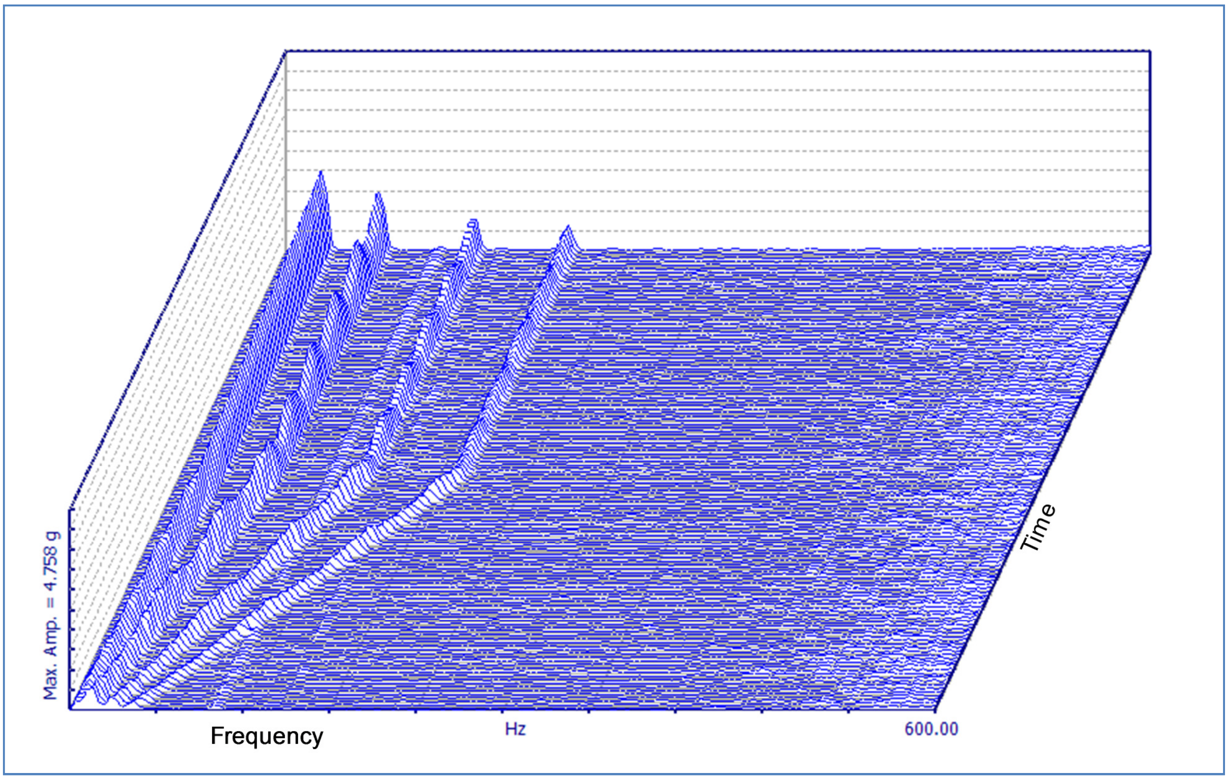

Figure 4.2 FFT waterfall of a faulty bearing during run-up

5. Power Cepstrum

Cepstrum is simply the inverse Fourier transform of the logarithmic power spectrum. The words cepstrum, quefrency, rahmonics and gamnitude in cepstrum analysis correspond to spectrum, frequency, harmonics and magnitude respectively in FFT analysis [5]. The quefrency is measured in seconds and it is actually a "delay time" or periodic time rather than absolute time. According to the above definition, cepstrum can be expressed by:

$$C_{AA}(\tau) = \mathcal{F}^{-1}\left\{\log X^{2}(f)\right\} \tag{5.1}$$

Where $C_{AA}(\tau)$ are the cepstral components while $X(f)$ are the spectral components. The importance of cepstrum analysis comes from the fact that it can detect repeated patterns (family of harmonics) present in the vibration signal. This feature is very important in the analysis of roller bearing vibration. Harmonics in the frequency domain are represented by a single quefrency component in the cepstrum domain. Hence, early detection of harmonics is possible even when the fundamental frequency component is not detectable. Furthermore, the harmonic spacing can easily be indicated from the reciprocal of the quefrency value. Also, the value of the main cepstral peak is good trend parameter because it represents the average over a large number of individual harmonics.

Some researchers showed that detection of incipient faults using Cepstrum is possible even when the fundamental defect frequency is absent in the spectrum. The fundamental frequency may be absent due to two reasons; an average and shift effect which causes a slow migration of the fundamental impact frequency from its computed value; and an inter-modulation effect which translates defect related information to frequency locations unrelated to the fundamental impact frequency [6].

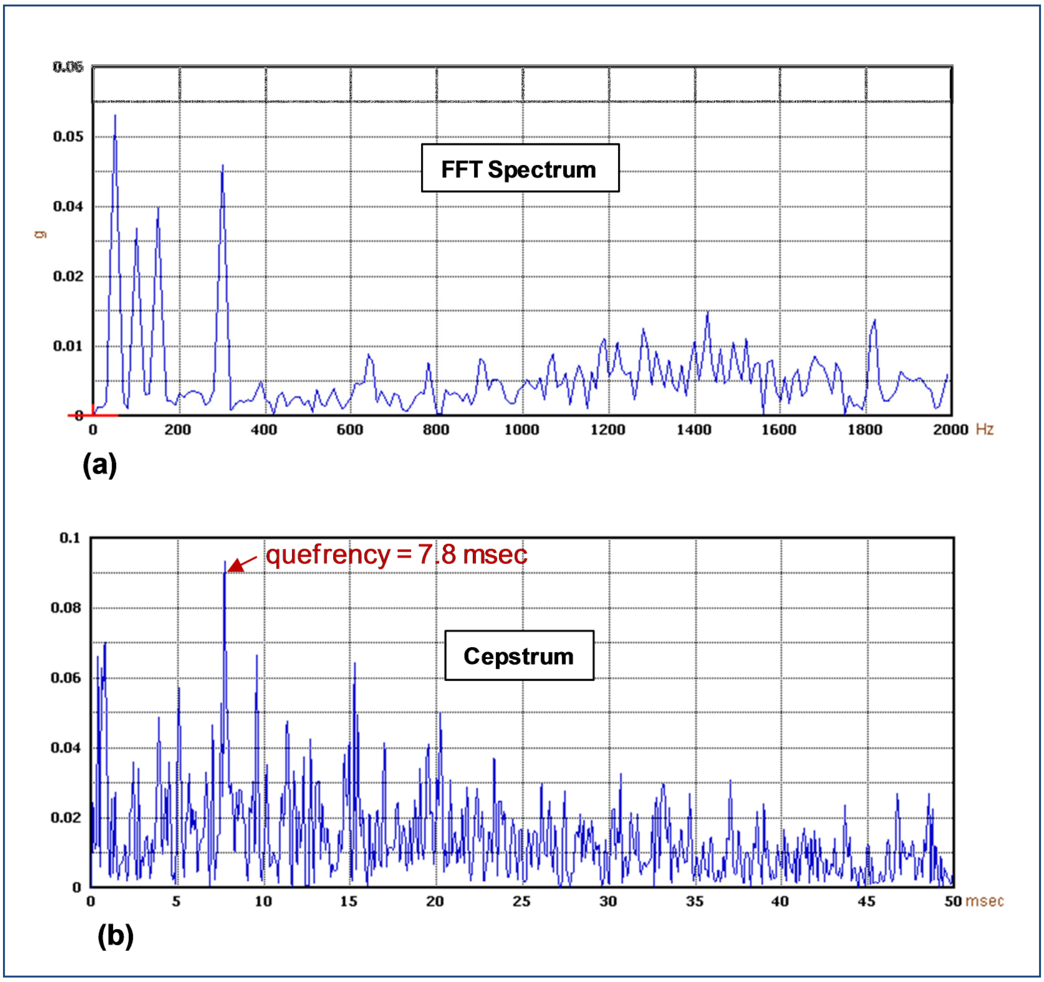

Figure 4.3 (a) FFT Spectrum, (b) Cepstrum plot

Figure 4.3 shows the FFT and cepstrum analysis for a ball bearing with defected balls in the earlier stage. Despite that BSF, corresponding to 2.6X = 129 Hz, and its harmonics are not detectable on the spectrum, the repeated patterns of BSF are successfully detected on the cepstrum plot by the peak quefrency component of 7.8 msec = 1/129.

6. Shock Pulse Method

The original "Shock Pulse Method", patented by Eivind Sohoel (SPM Instruments) in 1969. In 1970 SPM Instruments began to exploit products that are based on this method. The SPM approach measures the mechanical shock speed, measuring the compression wave produced when rolling-element and race interact and, with damage or failing lubrication, eventually, collide. When contact occurs, a surface reaction manifests as a compression wave and travels at the indicated speed of sound within the given material.

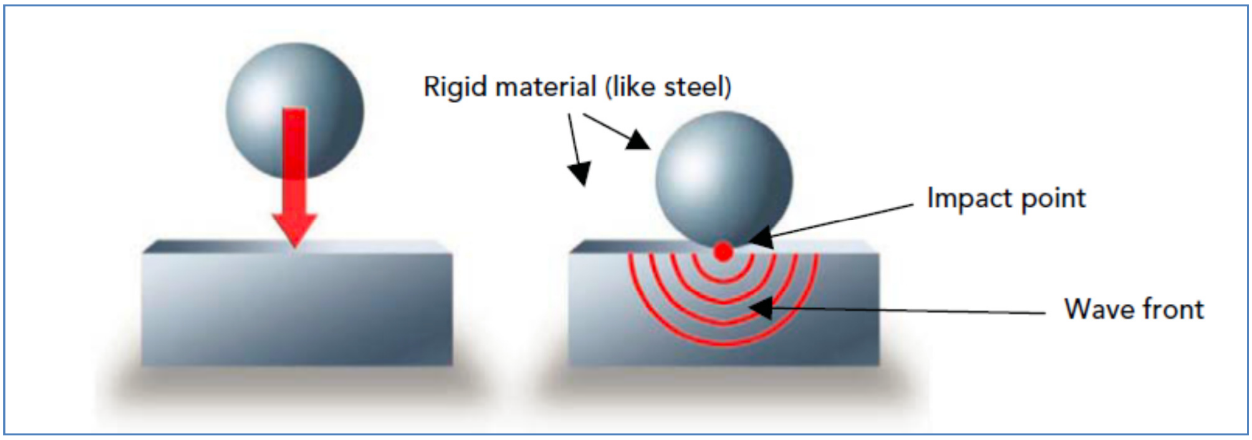

Figure 6.1 Elastic wave front as a result of metal to metal collision [7]

The SPM transducer is an accelerometer having 32kHz resonance frequency used to capture the stress waves. A fault-free bearing, that is well designed and properly selected for this application, is used to generate the "Carpet Value" or baseline relative to RPM and shaft diameter. Empiric testing established a table of such normal noise level expectations, or a quantifiable "carpet value" measured in dB. A brand new bearing entering service already has an expected Shock Pulse value measured in dB. Also, it would generate a regular and unvarying noise to our ear.

Even at that point in time, minor noises from impulses will rise above the average carpet noise. The maximum observed excursion reflects the number of shocks from all sources. A phenomenon such as the slight "catching" of an out-of-round imperfection would generate a rise of the maximum (peak value) rather than in the overall noise. Both the carpet level and maximum value develop over time. This yields not two, but rather, three values: the carpet, the peak, and the difference between them.

The amplitude of a shock pulse is proportional to the speed of the colliding. In the case of a rolling element bearing, the speed of the rolling elements is proportional to the diameter and the rotational speed of the shaft. The dependence of the shock pulse amplitude on the relative speed of the objects at the moment of impact creates problems when establishing alarm acceptance levels. To solve this problem, a normalization factor (dBi or HDi) is introduced that is effectively normalizing the shock pulse amplitude regardless of the rolling element speed. The result is that the shock pulse reading is presented on a normalized scale. By defining the diameter of the bearing and by measuring RPM (or by manually entering RPM) the normalization factor called dBi for different diameters and RPMs (or HDi in the SPM HD method) can be calculated.

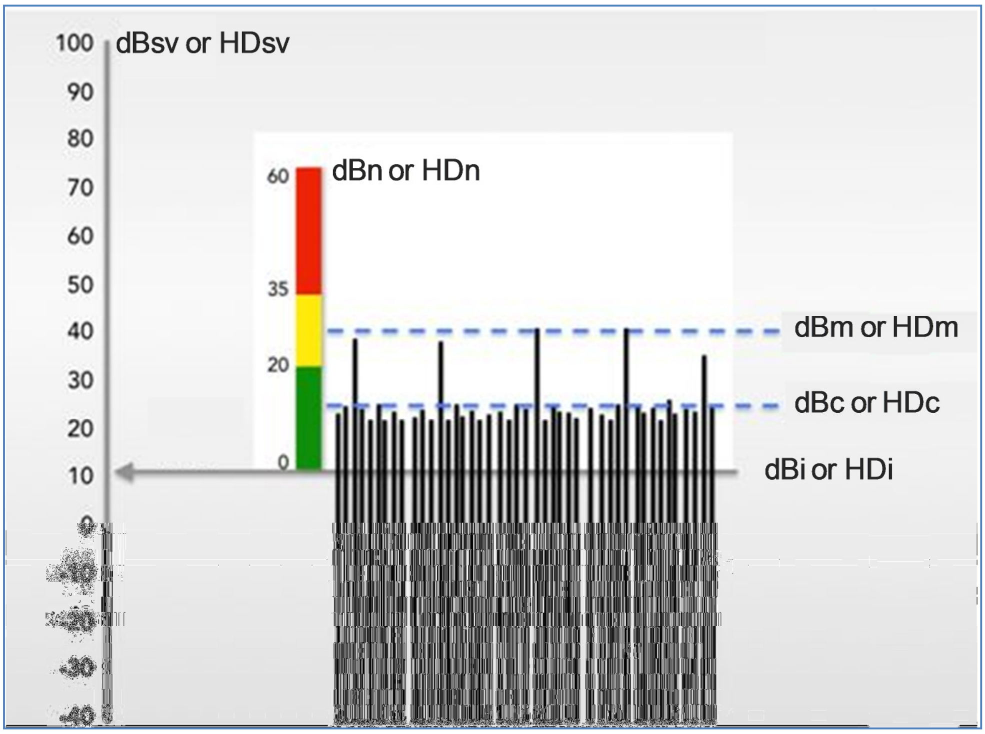

Figure 6.2 Description of SPM measuring parameters [7]

Based on the above, the SPM measuring parameters are (see Figure 6.2):

$dBsv$ *or* $HDsv$: this is the unit of non-normalized (raw) shock pulses.

dBi *or* HDi: normalization factor, depends on RPM and bearing dimensions and it can be calculated from empirical relations.

dBn *or* HDn: this is the unit after normalization.

dBm or HDm: the is the highest shock pulse found during measurement time in the normalized scale (referenced to *dBi*).

dBc or HDc: this is the threshold level where there exits 200 shock per second expressed in the normalized scale (referenced to *dBi*).

In Figure 6.2, the peak shock pulse before normalization is 40 dB, while *dBi* is calculated to be 10 dB. Hence the normalized *dBm* = 40 – 10 = 30 dB indicating some problem according to the severity level. Likewise, the non-normalized carpet level is 22 dB, hence the normalized *dBc* = 22 – 10 = 12 dB.

Recent SPM meters involve not only level detection but also shock pulse spectrum for bearing defect frequency confirmation.

7. Spike Energy

Spike energy (SE) measurement was proposed by IRD Mechanalysis (currently Entek IRD). This method employs the overall vibration in the frequency range 5 – 50kHz to assess the condition of rolling elements bearings. The IRD accelerometer type 970, which has a natural frequency of 27kHz, is utilized to collect vibration signal.

Currently, the SE measurement is based on a bandpass filtered signal with lower cut-off (highpass filter) set at 100Hz, 200Hz, 500Hz, 1kHz, 2kHz or 5kHz. While the upper frequency (lowpass filter) is permanently set at 65kHz [7]. The highpass filter selection depends on the machine speed and it tends to remove the low-frequency components resulting from common problem such as unbalance and misalignment which may obscure the pulses resulting from a bearing defect. Usually acceleration peak-peak value in g's (1 g = 10 m/s^2) is employed as monitoring parameter and, hence, denoted as gSE. The following table shows the recommended filter setting for gSE measurement as advised by Entek IRD:

HP Filter setting	Speed range (RPM)
500 Hz	0 to 100
1000 Hz	100 to 1000
2000 Hz	1000 to 1500
5000 Hz	1500 and above

Figure 7.1 shows the flow chart of Spike Energy signal processing. Spike Energy is a demodulation process that tends to detect the pulses in the filtered signal by a peak-peak detector followed by a decay circuit to hold the peak and slowly decay it. The decay time constant is manually selected by the user or automatically selected by the processing software according to measurement F_{max}.

The peak-to-peak detector in SE circuit is very sensitive to the defect frequency as compared to other envelope detection or demodulation methods. Unlike other demodulation techniques which may cause distortion to the detected peaks due to the use of low-pass filter in envelope detection, Spike Energy detection circuit preserves the severity of defects by holding the peak-to-peak amplitude of the impulses and enhances the fundamental defect frequency and its harmonics by applying a proper decay time constant.

Many users have observed the fluctuation of Spike Energy amplitude, when observed over time (whether using analog or modern digital meters). Unlike the quickly reacting classical parameters, gSE levels show a laborious rise time and slow decay. This feature is inherent to gSE processing. What it means: cumulative or multiple event impulse energy is required to push the gSE level up. A period of quiescence is needed to settle back down. The delay period is required to prevent the amplitude modulation in the signal from imposing fluctuations in the measured gSE value. Recently, the decay time becomes proportional to the frequency manifest within the signal. Despite the fact that gSE is a good indicator for bearing fault, it can easily be compromised. Also, measurement must be trended to identify the acceptable limits.

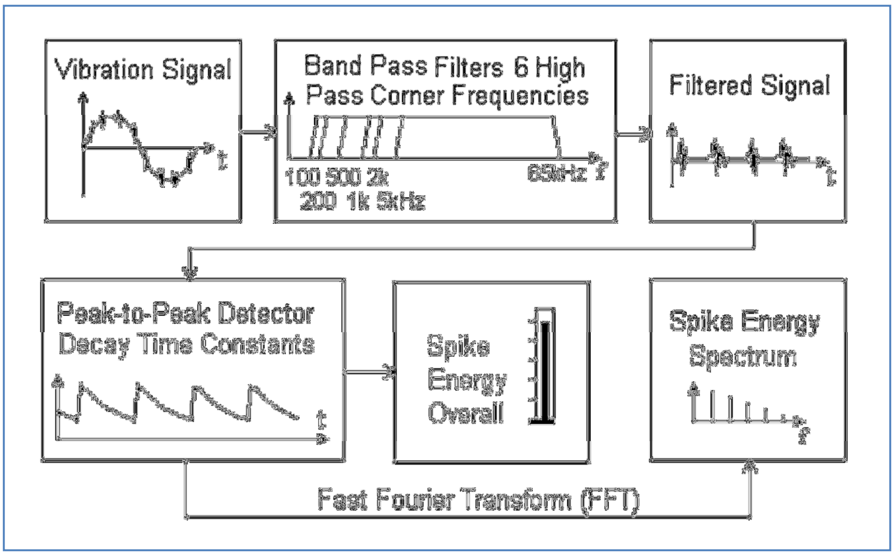

Figure 7.1 Spike Energy processing [7]

8. High Frequency Detection

This technique was introduced by SKF as a competitive to Spike Energy and SPM. It is based on measuring the peak of the signal in the frequency range 5 to 60kHz. Unlike gSE, there is no further processing to the signal other than bandpass filtering. This technique provides reasonable performance as a condition detection system.

9. Envelope Detection and Demodulation

Envelope may be defined as the outer shape of the signal. When a high-frequency signal is amplitude modulated with a low-frequency signal, the envelope would be waveform of the latter signal. Envelope detection is the process of demodulation in order to extract the low-frequency content. For bearing vibration, envelope detection is useful in identifying the intensity of the pulses and also in finding the repetition rate of these pulses. Repetition rate is related to the bearing characteristics frequencies (BPFI, BPFO, BSP and FTF) and can be found by spectrum analysis of the demodulated signal.

Lowpass Filter Enveloping

Enveloping has special importance in identifying bearing faults in the earlier stages. The process of extracting the envelope is shown in Fig. 9.1. The signal is bandpass filtered to remove the low-frequency content (related to common machinery problems) and very high frequency noise. Then, the signal is rectified and low-pass filtered or integrated to obtain the envelope [9]. It is important to select the appropriate filter setting for envelope spectrum processing such that no details are discarded.

As compared to Spike Energy method which detects peak-peak value, envelope processing tends to average the rectified peaks during low-pass filtration resulting in smoothed out peaks with altered peak values. Hence, the envelope processing is more suitable to obtain the demodulated time waveform or demodulated spectrum. Figure 9.2 shows a typical envelope spectrum for a faulty ball bearing. The bearing characteristic frequency will be shown as a peak in the spectrum (50 Hz in this example). There will be multiples of this frequency due to the impulses in the envelope time waveform.

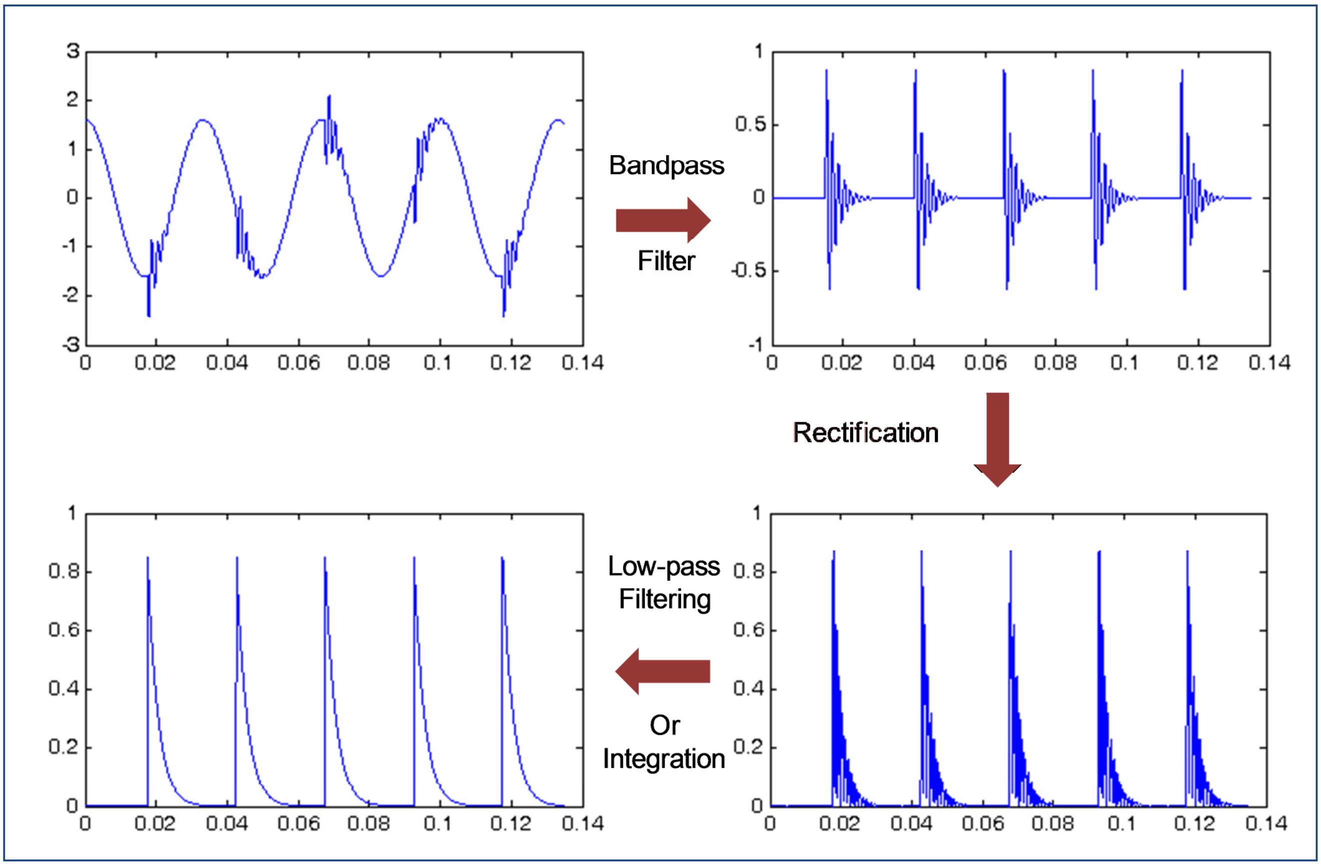

Figure 9.1 Envelope processing by low-pass filtering

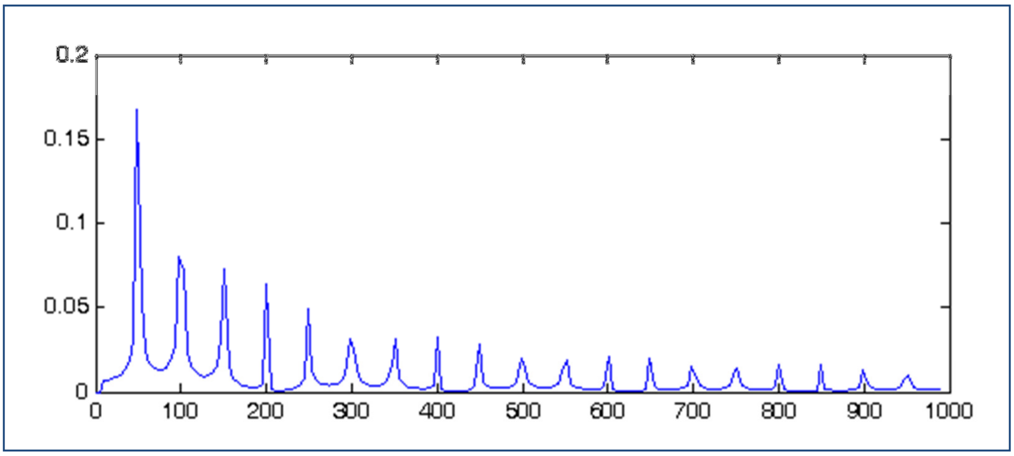

Figure 9.2 Envelope spectrum for a faulty roller bearing

The maximum frequency of the envelope spectrum (Fmax) is normally set at 1:5 of the highpass filter setting or sometimes at different ratio (1:2 for example). This is because that the bearing characteristic frequencies are far less than resonance frequencies which are smoothed out during envelope detection. It is important to select the appropriate filter setting for envelope spectrum processing such that no details are discarded. The following settings, which are provided by *SKF*, are used in some vibration analyzers and were adopted in HiDAC-8 portable machinery fault diagnostic platform as predefined settings.

Table 9.1 Enveloping Setting suggested by SKF

Enveloping Settings Microlog

Filters	Frequency Band	Speed Range	Analyzing Range
1	5 – 100 Hz	0 – 50 RPM	0 – 10 Hz
2	50 – 1,000 Hz	25 – 500 RPM	0 – 100 Hz
3	500 – 10,000 Hz	250 – 5,000 RPM	0 – 1,000 Hz
4	5,000 – 40,000 Hz	2,500 – ... RPM	0 – 10,000 Hz

Hilbert Transform Enveloping

As an alternative to the rectification and low-pass filtration used in demodulation process, the Hilbert transform can be used to obtain envelope which will preserve the actual pulses amplitudes. The steps to obtain the demodulated signals are as follows:

1. The signal is bandpass filtered as usual

2. The Hilbert transform of the filtered signal is obtained as follows;

- Take the FFT of the signal
- Cancel out the negative spectrum (the frequency components beyond 0.5 x Number of points) and double the positive spectrum.
- Take the inverse FFT to obtain the Hilbert transform.

3. Calculate the envelope $A(t)=\sqrt{x_R^2(t)+x_I^2(t)}$, where $x_R(t)$ is the real part of the Hilbert transform and $x_I(t)$ is the imaginary part of it. (Not that $x_R(t)$ is exactly the same as the input real time signal).

PeakVue Technique Enveloping

Another envelope detection technique developed by Emerson is the so called PeakVue technique [15. The technique is based on digitally sampled vibration signals at high sampling rate to capture the stress waves. The steps include bandpass filtering as in previous methods, rectification and then application of PeakVue envelope detection on the sampled data. Instead of using low-pass filtration of signal which could severely alter the amplitudes of peaks, the rectified signal is divided into time slots of Δt equal the reciprocal of 2.56 times the maximum demodulated frequency.

$$\Delta t = \frac{1}{2.56 F_{de\,\mathrm{mod}}} \tag{9.1}$$

The peak at each time slot represent the sample at the given time as shown in Fig. 9.3.

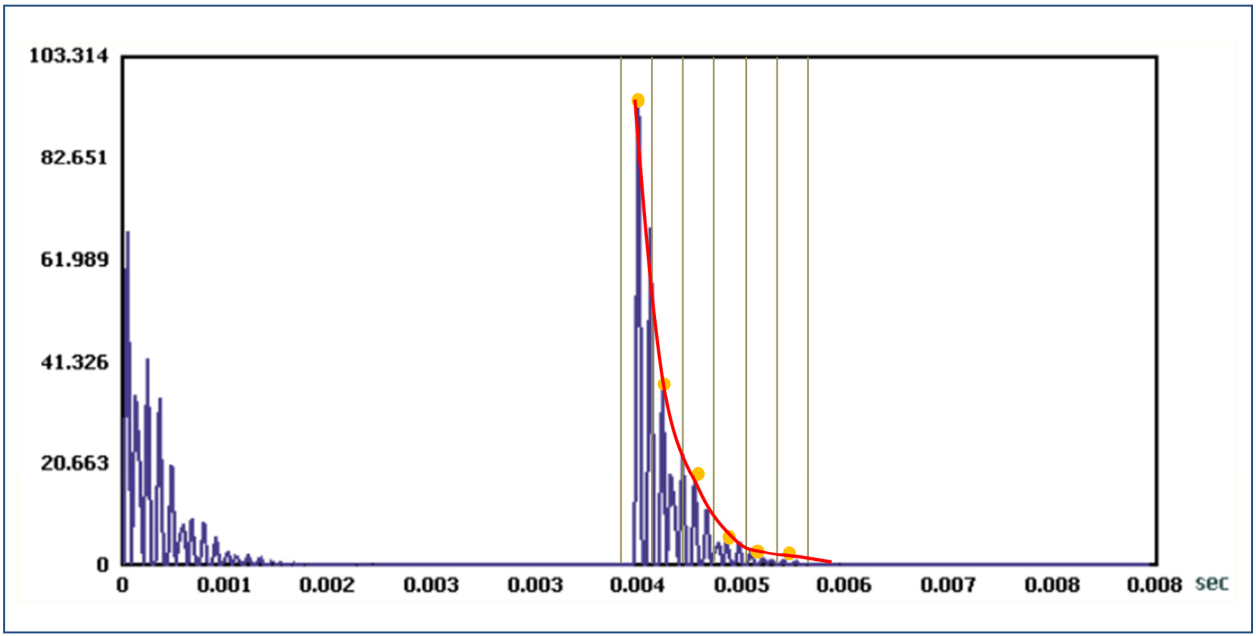

Figure 9.3 PeakVue Enveloping

The above technique is simple, yet powerful and proved to be able to preserve the actual peaks even when using low values for the demodulated frequency as shiwn in Fig. 9.4 [15]. Moreover, autocorrelation may be used to enhance the SNR and clearly show the repetitive patterns in time waveform. The autocorrelation coefficients C_j are ranging from -1 to +1 and are calculated from equation (9.2) for sampled data signal.

$$C_j = \frac{R_j}{R_0}$$
$$Where\ R_j = \sum_{i=0}^{N/2-1} x_i x_{i+j} \tag{9.2}$$

In the autocorrelation process, the random noise will be averaged out due to the fact that negative and positive values are subtracted while repetitive patterns are summed up. To reduce the calculations of autocorrelation function, FFT-based autocorrelation technique can be used.

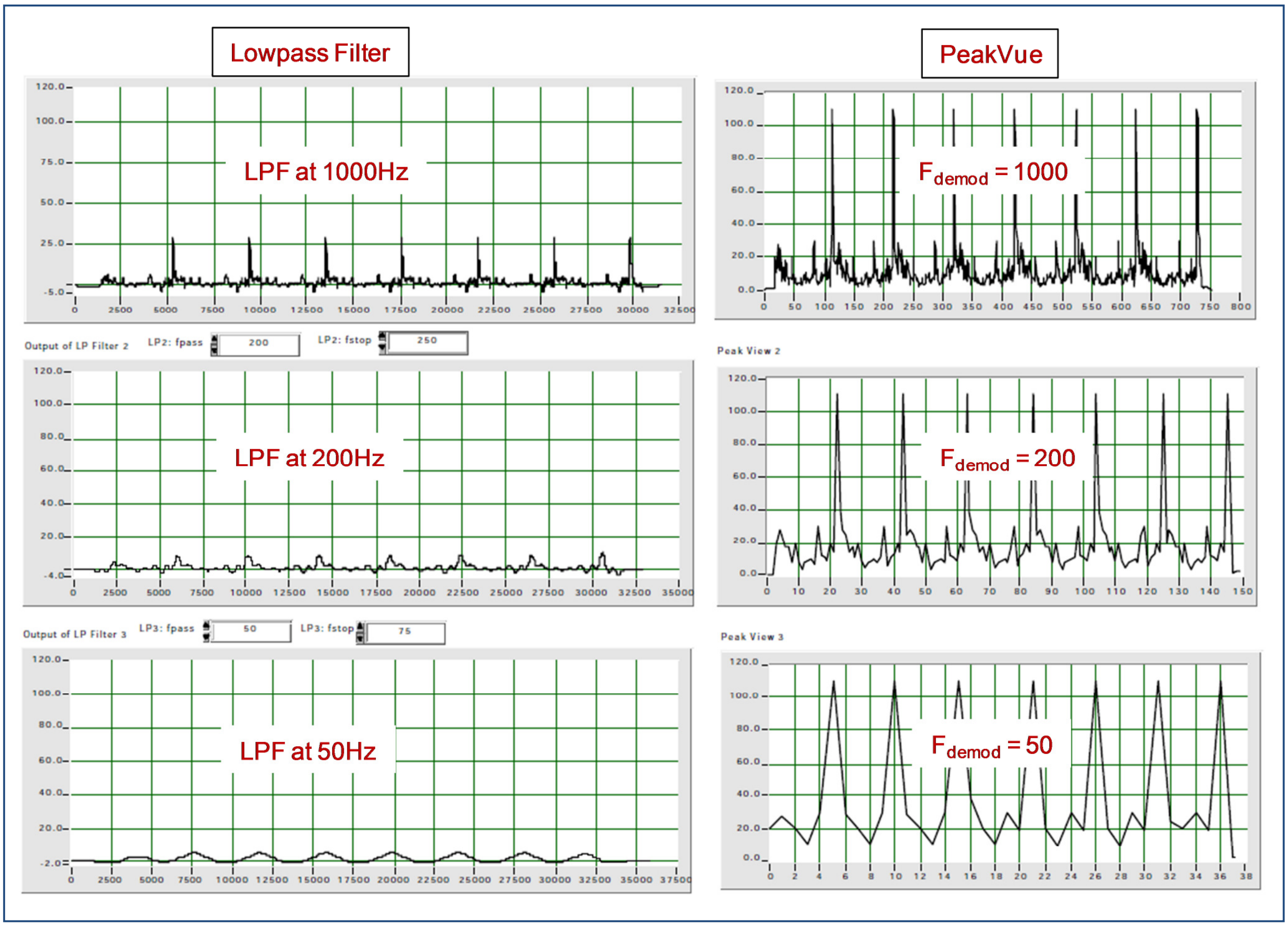

Figure 9.4 PeakVue vs. Lowpass Filtering Techniques [15]

Case Study: Multistage RO Water Pump

This is high pressure multistage pump belongs to Reverse Osmosis unit in SCPI company in Basra. Motor power is 22 kW, running at 1470 RPM and directly coupled to a multistage pump. The pump is installed in early of 2016. After few months of operation, the spike energy readings of motor drive end bearing begin to increase. In March 2017, the spike energy reading increased to 0.53 gSE as shown in Figure 9.5. Few days later, it reaches 0.94 gSE with noticeable high frequency sound. The envelope time waveform and FFT are collected to provide more information about the case.

The four characteristic frequencies of the installed bearing (type FAG-6310) are as follows; Ball-Pass Frequency Inner race (BPFI) = 4.95x, Ball-Pass Frequency Outer race (BPFO) = 3.048x, Fundamental Train Frequency (FTF) = 0.381x, and Ball Spin Frequency (BSF) = 1.98 x.

Figure 9.6 shows the envelope time waveform and FFT spectrum of bearing vibration using PeakVue technique. The time waveform of the demodulated signal clearly shows high amplitude peaks. The FFT spectrum of demodulated signal shows large peak at exactly BPFI 4.95X and its multiples with multiples of 1X sidebands.

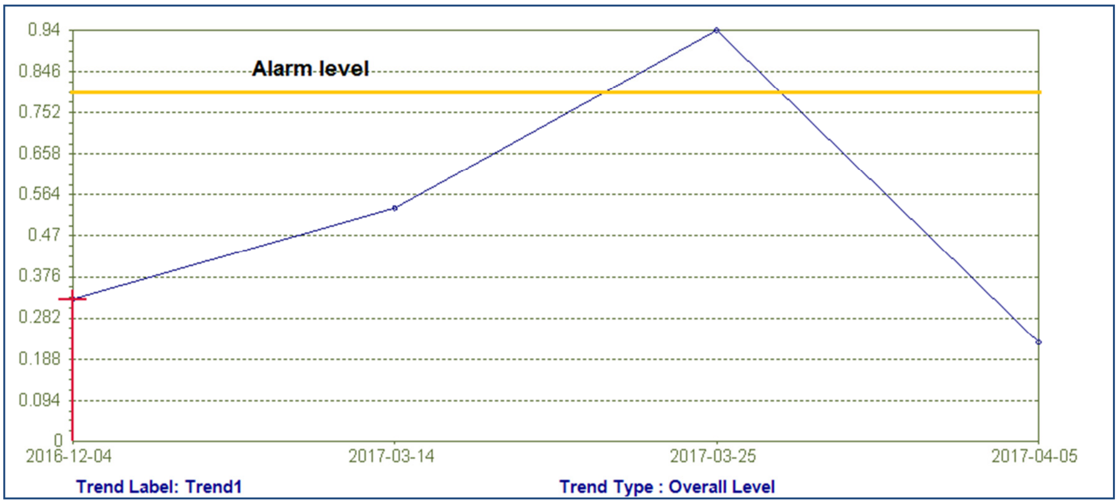

Figure 9.5 Trend of motor bearing spike energy vibration for RO pump

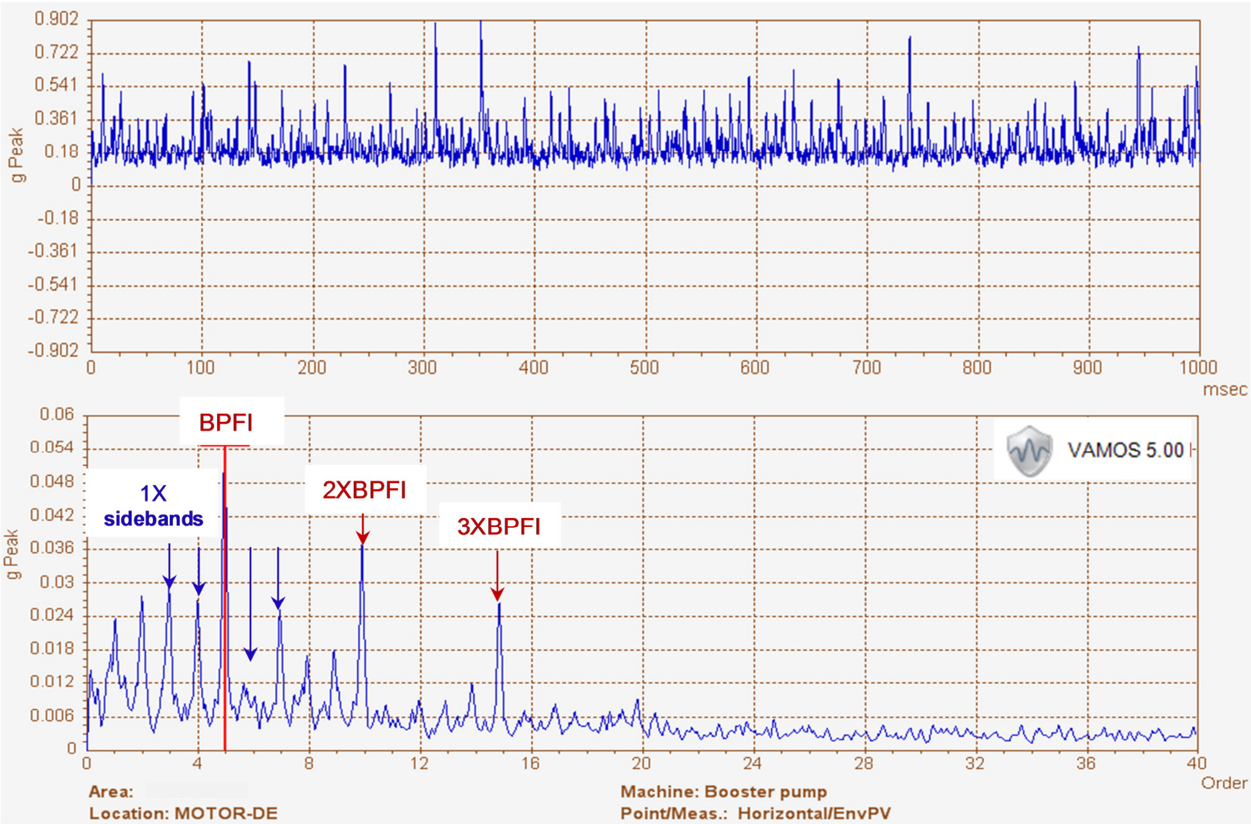

Figure 9.6 Envelope time waveform and FFT spectrum for RO pump bearing vibration

When the bearing is replaced and a complete maintenance procedure is performed on the pump, vibration level decreased to 0.23 gSE.

Case Study: Catacarb Pump P202-B

This is a heavy duty pump used to pump hot catacarb solution in the Southern State Company of Fertilizers (SCF) in Khor Elzubayr in Basrah. The pump consists of 570kW induction motor running at 2930 RPM and a three-stage centrifugal pump directly coupled to the motor. The pump is shown in Figure 9.7. The pump main rotor is periodically replaced each several months due to severe vibration problems that cause its failure. There was high vibration in the pump in both drive-end and free-end with vibration velocity RMS of 14.2 mm/s in the horizontal direction of the pump non-drive-end.

Vibration signatures from the horizontal direction on the non-drive-end are collected using velocity, acceleration and PeakVue envelope methods. Inspecting the acceleration spectrum, there are harmonics of 1X component up to 7X with high peak at 5X as shown in Figure 9.8. The 5X component coincides with Blade Passing Frequency (BPF = No. of blades x shaft speed) of the impeller. This frequency is a characteristic of a pump and it is the result of pumping pulsation when the blades inject fluid through the outlet. However, high amplitude BPF indicates some pump related problems. The suspected problem is that the impeller is axially eccentric inside the pump casing.

Figure 9.7 P202-B Catacarb pump

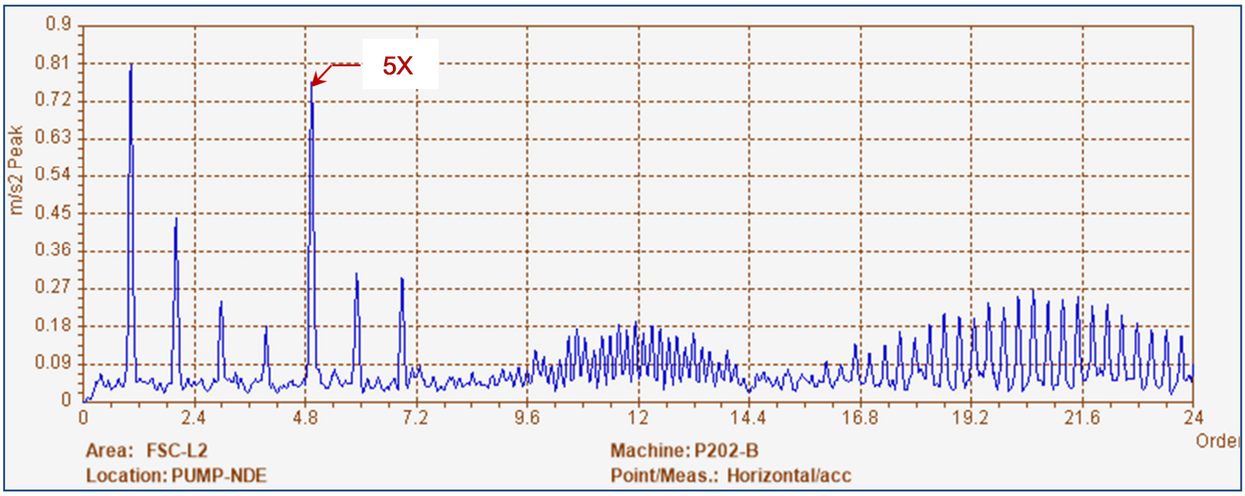

Figure 9.8 Acceleration spectrum of non-drive-end of P202-B

Another problem is detected in the non-drive-end of the pump. The PeakVue peak-peak value is 9.38g which is above alarm limit and indicate bearing problem. Double angular contact bearings type FAG 7311B are used in the non-drive end. The bearing has BPFI = 7.086X, BPFO = 4.915X, BSP = 2.046X and FTF = 0.409. When envelope demodulated spectrum is used to study the problem, the bearing tone of BPFO 4.915X and its multiples are clearly shown in the demodulated spectrum as shown in Figure 9.9. This tone was completely obscured by the nearby 5X component of the blade pass frequency in the normal acceleration spectrum.

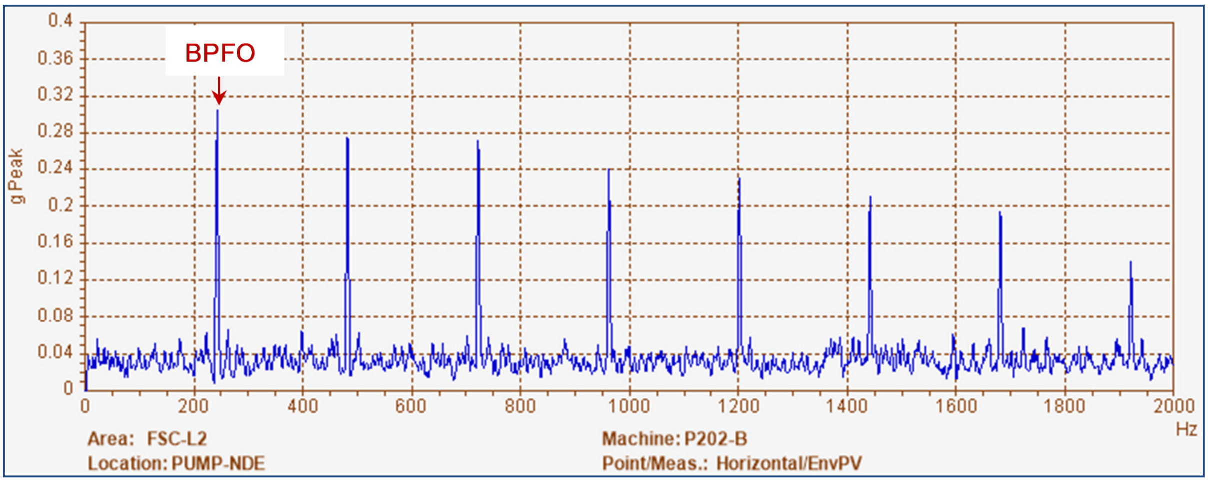

Figure 9.9 PeakVue demodulated spectrum of non-drive-end in P202-B

The lack of lubrication due to insufficient oil level in the oil sump was the problem. Bearings are lubricated by oil ring as shown in Figure 9.10 where the ring sprays the oil on the bearings. The maintenance staff indicated that the oiler bulb was replaced with incorrect type which was the reason of oil level being insufficient in the sump.

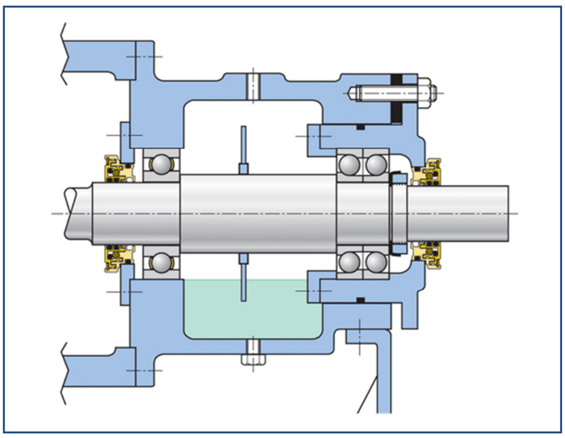

Figure 9.10 Lubrication system of non-drive end bearings in P202-B

10. Signal Enhancement Techniques

When bearing vibration signal is contaminated with noise or other vibration sources, fault detection process needs more advanced techniques. When the fault develop, it becomes easier to detect the fault by common techniques such as FFT. When the SNR (signal to noise ration) is poor, normal time waveform and FFT become ineffective in detecting faults. To improve the performance of signal analysis techniques, the SNR has to be enhanced. Some signal enhancement techniques will be explored here.

Signal Decomposition

The time domain signal acquired from a machine consists of periodic signal and additive noise. Some techniques are used to cancel out noise and extract the useful data. Among these techniques are:

Leakage-free sampling

In this case, sampling period is locked to the rotational speed such that the analysis length is integer multiples of shaft rotations. Encoders are used to achieve this. However, encoders cannot be installed to every machine and need special hardware provision that make this method impractical for everyday vibration data collection.

Time domain averaging

Time domain averaging is the technique based on doing averaging process in time domain such that periodic signals are summed up while noise is cancelled out. Suppose that vibration signal $x(t)$ is composed of periodic signals $f(t)$ and additive nose $s(t)$. When summing several periods of the signal, the repetitive signals add coherently while the noise incoherently [10].

$$\hat{x}(t_i) = Nf(t_i) + \sqrt{N}s(t_i) \tag{10.1}$$

It is clear that SNR will be enhanced by a factor of $\sqrt{N}$. The averaged samples are synchronized with trigger signal obtained from a tacho sensor such as Hall effect or photo sensor. Hence, this technique is called Time Synchronous Averaging (TSA). When using digital sampling and to cope with speed variation during rotation, the data may be re-sampled in order to obtain fixed number of points for each revolution. The sample may span one trigger period or more. Figure 10.1 shows TSA where averaging period is one revolution (one trigger period).

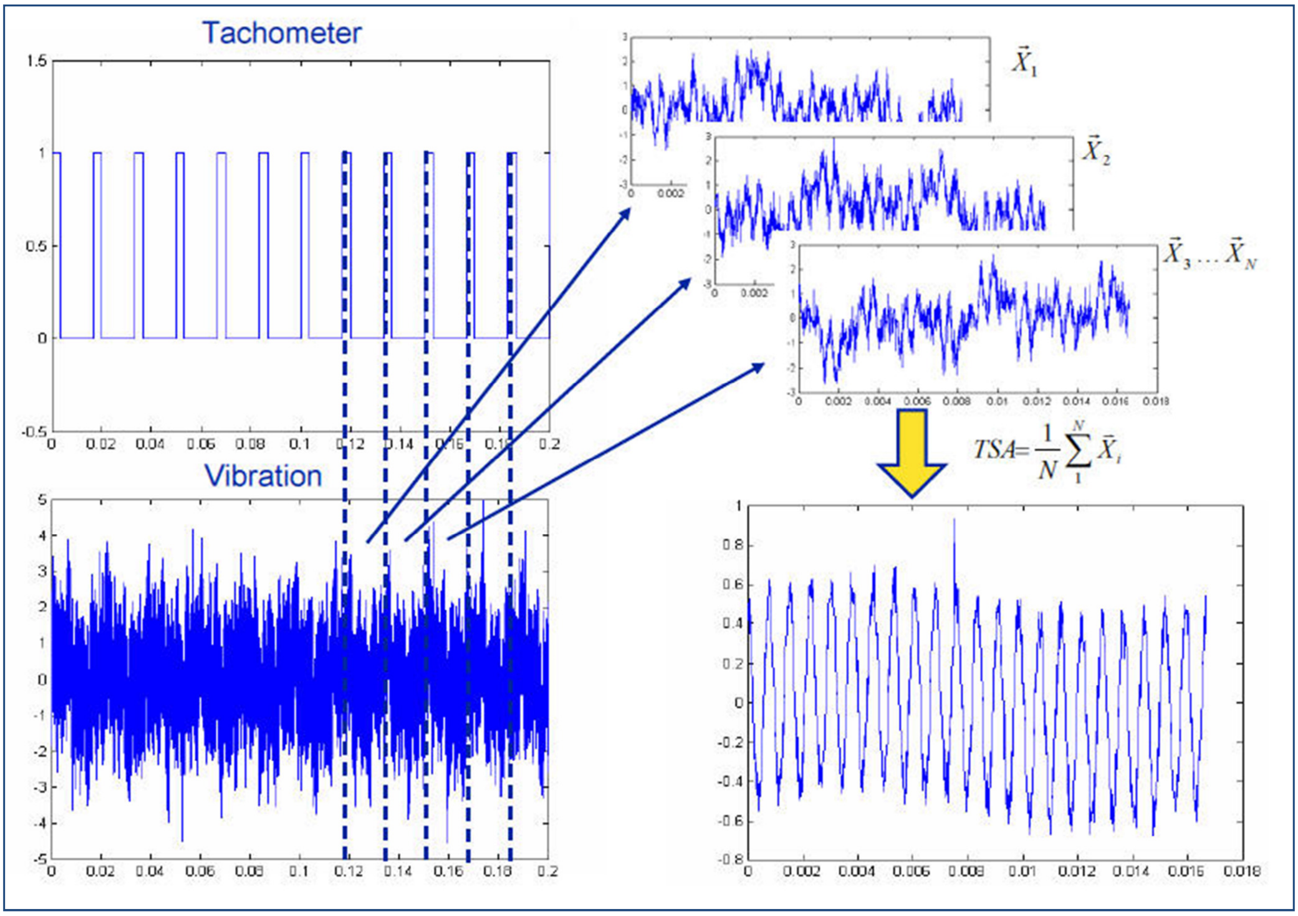

Figure 10.1 Illustration of time synchronous averaging

Adaptive Noise Cancellation

The general concept of Adaptive Noise Cancellation (ANC) is depicted in Figure 10.2. The monitored vibration signal v_0 is contaminated with noise n_0 picked by the primary vibration sensor. A reference sensor is used to pick a reference noise n_1 correlated to n_0 in some way. The reference noise n_1 is filtered by adaptive filter in order to match n_0 as close as possible. Adaptive filter coefficients are estimated by Least Mean Square estimator which minimizes the output power of the noise canceller. The output of the adaptive filter is subtracted from the combination $v + n_0$ in order to obtain best estimate of the required signal. ANC has been successfully applied to monitor bearing vibration [10].

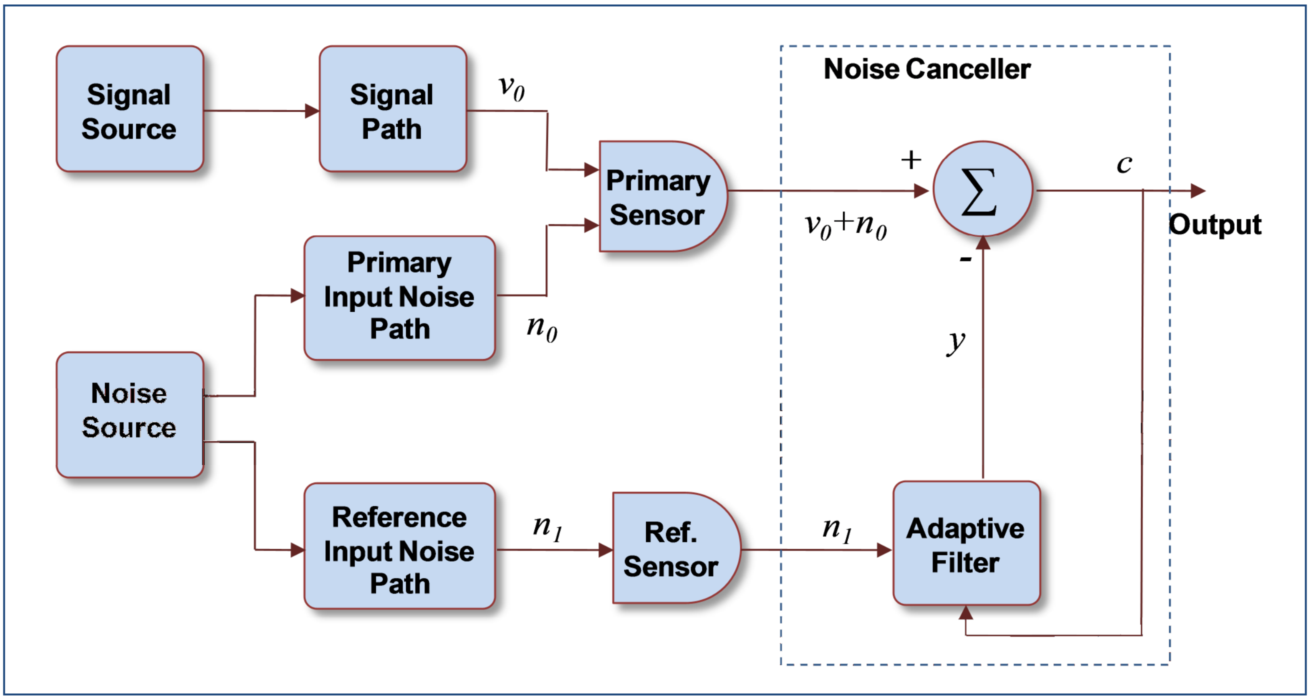

Figure 10.2 Adaptive Noise Cancellation

Adaptive Linear Enhancer

The problem with previous scheme (ANC) is that one requires a reference signal which is strongly correlated with the noise signal. The performance of an ANC scheme is usually limited by the availability and quality of the reference signal.

An alternative mode of operation for an adaptive filter, based on ANC, is provided by an adaptive line enhancer (ALE), as shown in Figure 10.3. This is a simple variation on the ANC requiring only a single input signal. In this case the reference signal is obtained by delaying the input signal by a fixed number of samples, Δ. The adaptive filter then endeavors to predict the signal x_k from the delayed samples. The result is that any input components which are predictable over the delay appear at the filter output, y_k , whilst the error signal contains those components which are unpredictable over the delay [11].

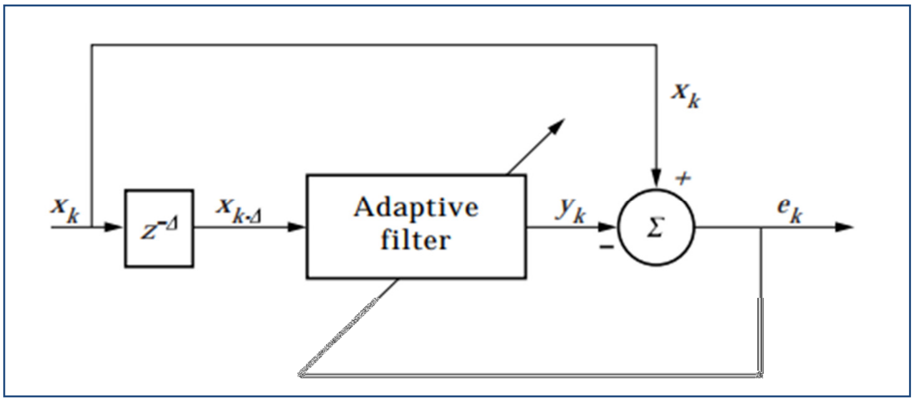

Figure 10.3 Adaptive line enhancer [11]

11. SPM High Definition and Symptom Enhancing

SPM High Definition (HD) is based on the same principles of traditional SPM method. Elastic wave propagation in bulk material is detected by a tuned shock pulse transducer. Same type of transducer is used in SPM HD method. However, the underlying interfacing circuitry and processing algorithm is completely different from traditional SPM method.

The block diagram (black box) of SPM HD algorithm is shown in Figure 11.1. As SPM Instruments developed commercial products that use this technology, minimal information is provided about the underlying processing techniques.

Figure 11.1 SPM HD block diagram [7]

The inputs to the SPM HD device are the shock pulse vibration, RPM and bearing data. The RPM is needed to calculate the normalization factor HDi as well as to synchronize data sampling as this method uses sophisticated digital techniques. Bearing data is used to calculate normalization factor and also to be used to extract SPM HD Time and HD Spectrum [7].

If the RPM is fixed, manual entry for its value is sufficient. But when the speed is varying or fluctuating, it is essential to use accurate speed sensor to read the RPM. Single pulse per revolution or multiple pulses can be used. When multiple pulses are used, such as when using coupling holes or bolts as triggering means, the angular spacing must be accurate in order to avoid smearing/distortion in the processed data.

HDm and HDc

In section 6, SPM is presented and HDm and HDc are described (corresponding to dBm and dBc). These scalar values are described in dB scale and specify the operating conditions of the bearing. SPM HD uses high efficiency A/D converter with digital filters to improve the SNR. Sufficient measuring time must be used to ensure proper results. SPM Instruments recommend using measuring time that spans 50 revolutions (10 rev. minimum) in order to obtain stable HDm measurement.

$$\text{Measuring time} = 50 \times (60/\text{RPM})$$

HDm and HDc are not affected by RPM measurement accuracy since RPM is not used in the calculation of these scalar values directly except when calculating HDi. To measure HDc, the total measuring time is divided into 5 ms time slots. In each time slot, the strongest shock pulse is identified and temporarily stored. When the measuring time is up, the weakest among these stored values is selected as the HDc value (there are 200 5-ms slots in one second). This gives an approximation of the level where more than 200 shocks per second are found.

SPM HD Time Signal

The time signal is a good tool to pinpoint bearing defects and it could be more useful than FFT spectrum in many cases [7]. The HD technology aims to produce clear and crisp time signal. Advanced digital processing techniques are used in this technology. Linear scale is used for time signal.

SPM HD is analogous to time synchronous averaging as the SNR is enhanced by a factor equal to square root of *Symptom Enhancing Factor* SEF. However, the algorithm is different as TSA cannot be used with bearing faults because they are normally not synchronous with RPM (fractional multiples of RPM). Also, the FFT averaging is used to obtain averaged spectrum and cannot be used to average time signals. SPM HD looks for and enhances repetitive pulses even when they are buried into noise. The algorithm involves digital sampling using efficient A/D converter, rectification and envelope detection by low-pass filtering. The symptom enhancement algorithm is applied to emphasis the repetitive pulses in the time signal. The SEF has considerable effect on the clarity of the results and SNR but the measuring time will be, consequently, larger. Figure 11.2 shown the effect of using Symptom Enhancement on the time signal [2]. SPM provided no information about the maths of Symptom Enhancement technique, but one can suspect that autocorrelation is used just like the PeakVue technique of Emerson.

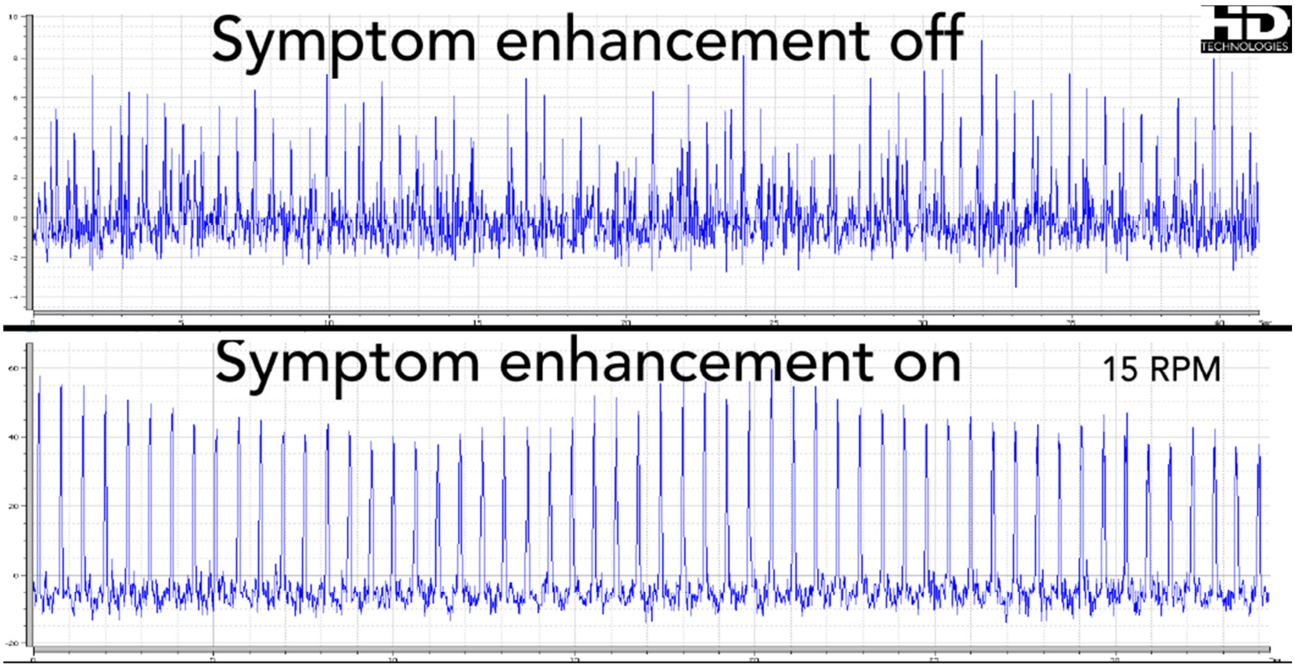

Figure 11.2 Effect of Symptom Enhancement on the SPM HD time signal [2]

SPM HD Spectrum

If a Fast Fourier Transform (FFT) is applied to the SPM Time Signal HD, an SPM Spectrum HD is created. This is useful to interpret and locate bearing faults. The SPM Spectrum HD is also presented on a linear scale.

Figure 11.3 shows the HD Spectrum for a bearing with cracked inner race. The severe amplitude modulation of BPFI with 1XRPM is clearly shown on the HD Spectrum since there are multiple sidebands (spaced at 1XRPM) around BPFI and its harmonics. It is also important to mention that since HD Spectrum is obtained from the SPM Time HD, the rectification of the signal during envelope detection resulted in a non-zero DC level. This DC level is reflected into a spectral component at 0 Hz (or 0 Order) with its sidebands [7].

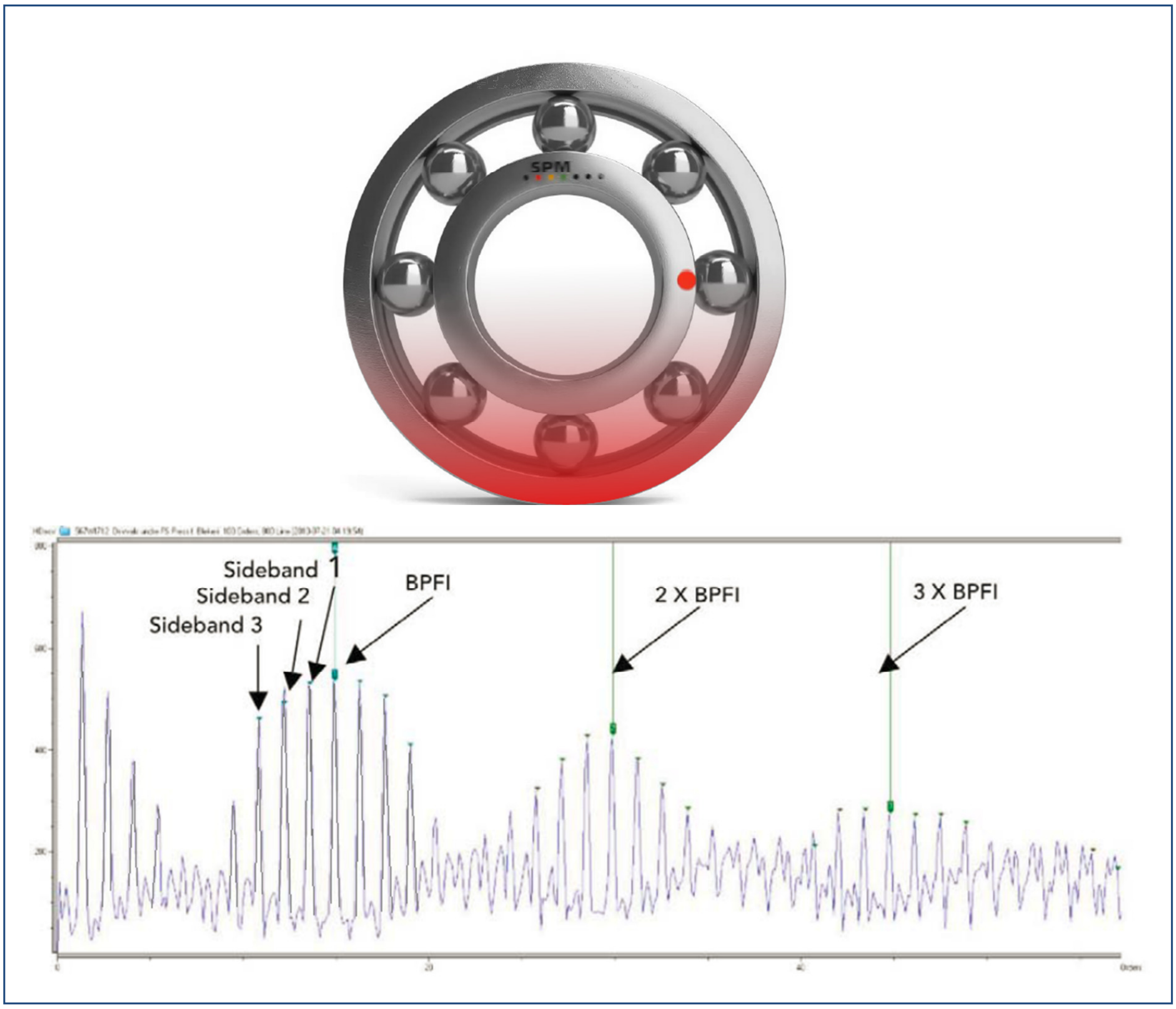

Figure 11.3 SPM HD Spectrum for a bearing with cracked inner race [7]

Hallsta Paper Mill [2]

This case describes using SPM HD method to study low-speed bearings problems in Hallsta paper mill in Sweden. Application of this technique in condition monitoring and analysis of vibration signals had led to important cost saving due to early detection of faulty bearings and changing the philosophy of maintenance program. The paper mill is shown in Figure 11.4.

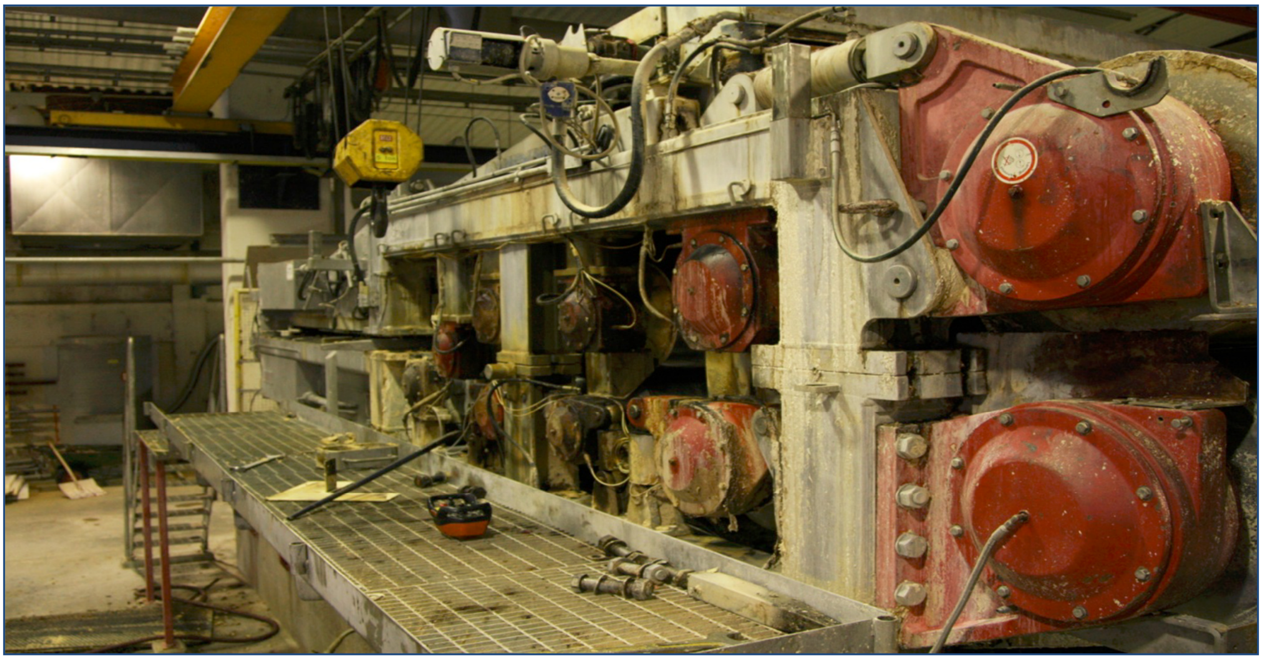

Figure 11.4 Hallsta Paper Mill during maintenance [2]

For the lower side S-roller bearing, type SKF 22320 spherical roller bearing, vibration monitoring showed high HDm readings (> 60 dBm) as shown in Figure 11.5.

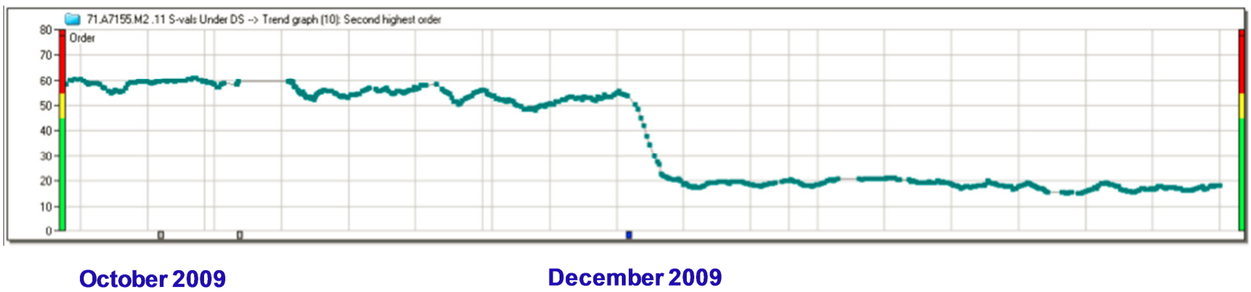

Figure 11.5 S-roller bearing HDm values from Oct. 2009 onward [2]

The roller rotates at 15 RPM. For that low speed bearing, the SPM HD method is applied to obtain time waveform of the demodulated signal with Signal Enhancement Factor of 10. The measuring time for the setup extends from 5 to 8 hours due to large SEF and low frequency. The HD demodulated time waveform before bearing replacement in December 2009 is shown in Figure 11.6. The time waveform clearly shows inner race pattern modulated with 1xRPM.

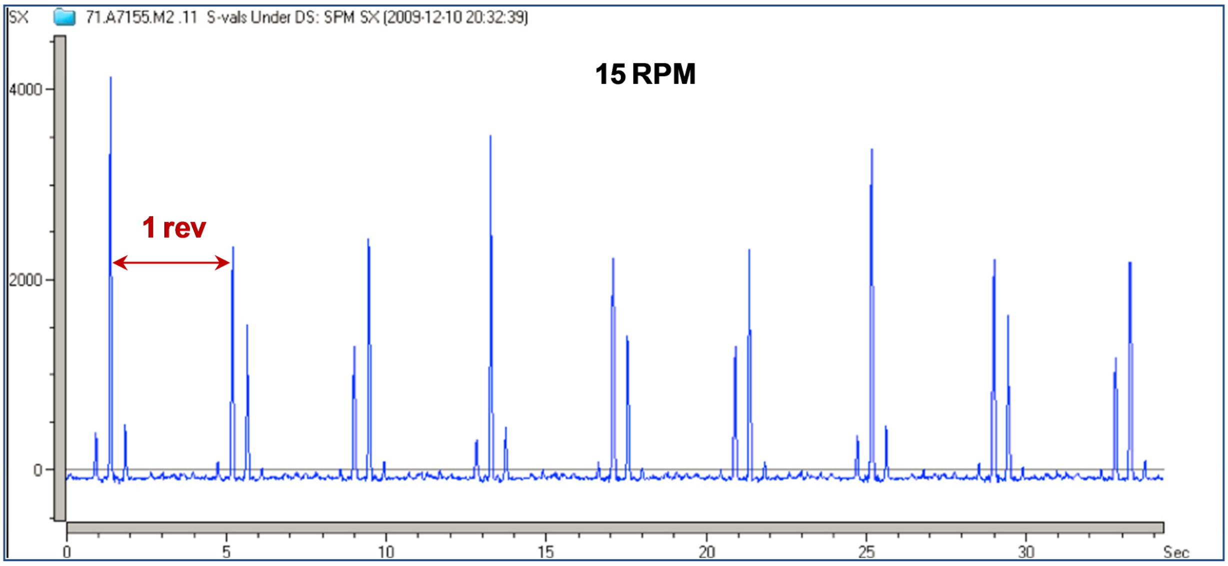

Figure 11.6 Time waveform of S-roller bearing before replacement [2]

When the bearing is replaced, the HDm readings decreased to noticeable level (20 dBm) as shown in Figure 11.5 above. The defective bearing is shown in Figure 11.7.

Figure 11.7 SKF 22320 bearing photo show cracked inner race [2]

12. Wavelet Transform

Although Fourier transform (FT) is very powerful tool in signal analysis, there is a major disadvantage that limits its use in analyzing transient or non-stationary signals. Fourier expansion has only frequency resolution, i.e. there is no time information in this type of analysis. Despite the fact that FT can identify the frequencies contained in a signal, the time at which these frequencies occur is unknown. In fact, FT detects the averages of the individual components over the measurement time. On the contrary, Wavelet transform (WT) is capable to examine the signal in both time and frequency simultaneously which permits its use to analyze signals that are described to be non-stationary, aperiodic, noisy and transient [12].

Wavelet transform is a method of converting a function (signal) from one form into another for the purpose of making the features of the signal more amenable to study. It has the ability of representing the signals in both time and frequency. Given a function $x(t)$ which is square integrable, the Continuous Wavelet Transform (CWT) coefficient $W(a,b)$ is the inner product of $x(t)$ and a normalized, dilated and translated wavelet function $\psi_{a,b}(t)$ [12];

$$W(a,b) = \left\langle x(t), \psi_{a,b}(t) \right\rangle = \int_{-\infty}^{\infty} x(t)\psi_{a,b}^{*}(t)dt$$

Where the shifted and dilated wavelet function is given by: $\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)$

Where $\psi(t)$ is called mother wavelet function. The selection of the wavelet function depends on the type of the signal to be analyzed as well as the physical phenomenon or process being interrogated. Examples of mother wavelets are the Haar, Morlet and Mexican hat.

In the CWT, the wavelet is dilated and shifted either continuously or at very fine resolution. This will result in highly redundant coefficients. This redundancy can be removed by proper discretization of the scaling and shifting factors given that the scaled wavelet functions satisfy the condition of orthogonality. Dyadic grid scaling is normally used in Discrete Wavelet Transform (DWT) with dyadic wavelet functions (father wavelet).

Application of wavelet transform in bearing signature analysis offers many advantages over the traditional analysis techniques. The non-stationary and impulsive signals can be better analyzed by using the time-frequency distribution of the wavelet transform. Moreover, the filtering and denoising capabilities of the DWT can be found effective in improving the signal to noise ratio. Figure 12.1 shows the 3D wavelet transform representation of vibration signal (scalogram) where both time and frequency (scale) representations of the signal can be viewed.

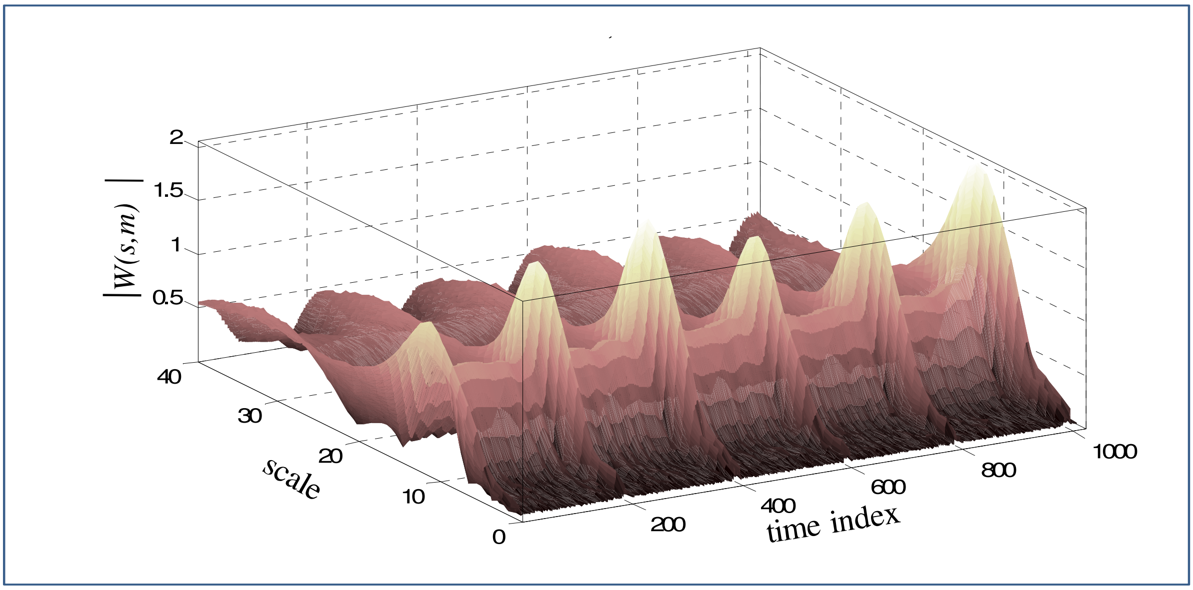

Figure 12.1 Wavelet transform representation (scalogram)

Case Study: P304 Screw Oil Pump

This is a screw-type oil pump, shown in Figure 12.2. It belongs to the Southern State Company of Fertilizers in Khor Elzubayr. It is used to pump the oil at high pressure for the sealing system in a CO2 compressor. It has a 160kW electrical motor running at 2970 RPM and directly coupled to the pump. This pump was installed in 2010 as a replacement for an old pump. After two years of operation, the staff of SCF reported increased vibration acceleration level on the drive-end support of the motor. The four characteristic frequencies of the bearing (type FAG-6219) are as follows; Ball-Pass Frequency Inner race (BPFI) = 5.899x, Ball-Pass Frequency Outer race (BPFO) = 4.1x, Fundamental Train Frequency (FTF) = 0.41x, and Ball Spin Frequency (BSF) = 2.692x.

Inspecting the FFT spectrum in Figure 12.3, it is clear that multiples of 2XBSF are presented with large amplitude at exactly 6XBSF. The presence of BSF is indication of defected balls as mentioned above. However, 2XBSF is usually reported due to the fact that the defected ball rolls over the inner and outer race each one revolution. Also, it has been observed that BSF is presented due to a number of conditions such as ball inaccuracies, lack of lubrication and high preloading. These harmonics of BSF are commonly accompanied by sidebands at FTF.

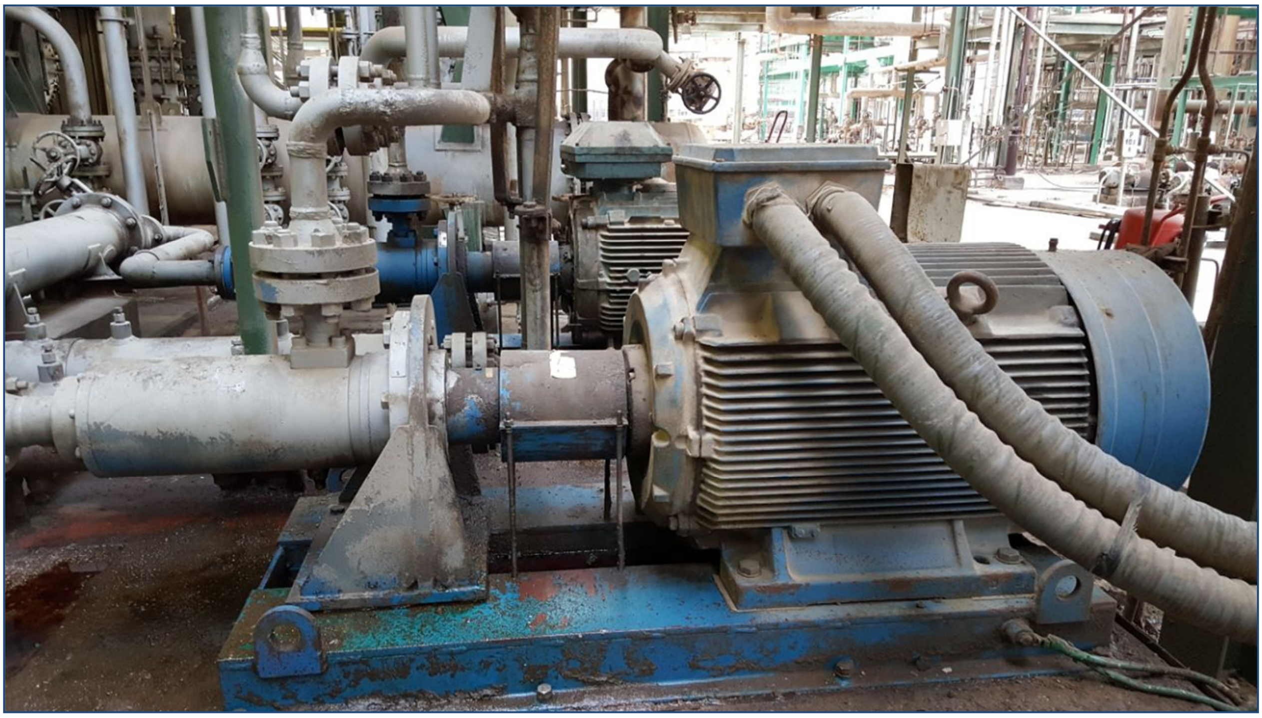

Figure 12.2 P-304 screw oil pump

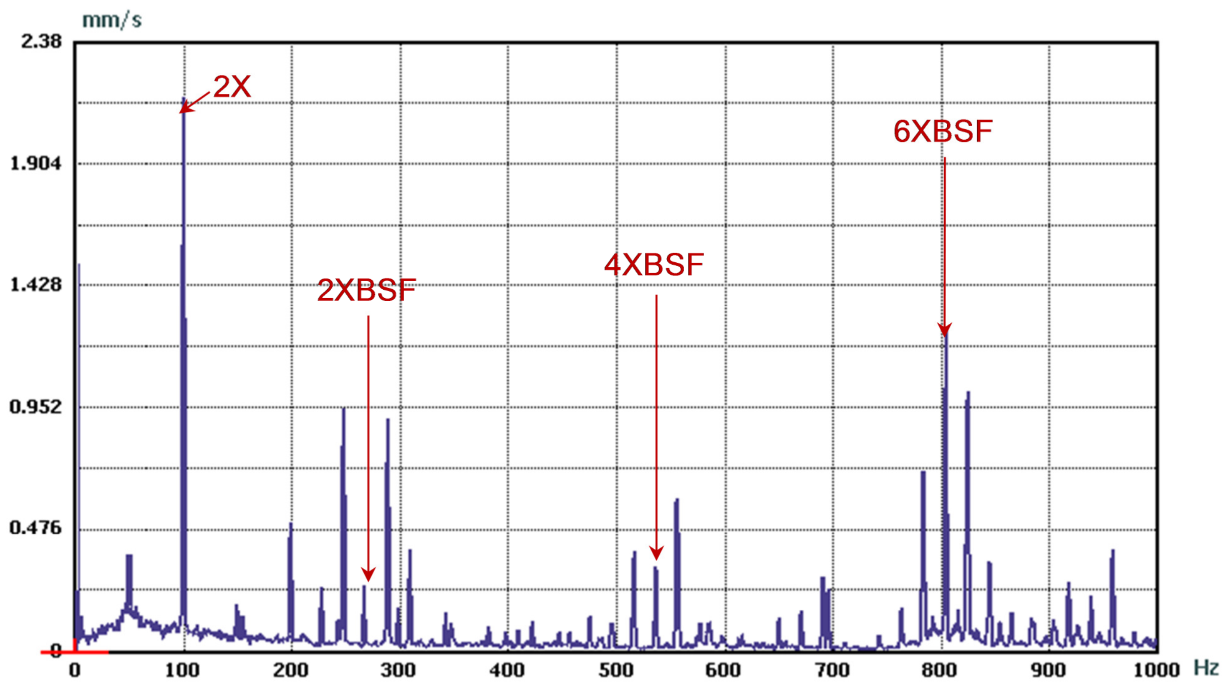

Figure 12.3 P304 Motor drive end bearing vibration spectrum

For further investigation, vibration time signal is viewed. Both direct and filtered signals are displayed. Filtration is done by using CWT to obtain the filtered signal at frequency of 803 Hz, corresponding to the largest harmonic of BSF. The original time signal and the cross-section plot of CWT are shown in Figure 12.4. The original time signal is very noisy such that the amplitude modulation can hardly be identified. On the contrary, the amplitude modulation of 803 Hz component with the FTF is clearly visible on the cross-section plot. Since there are five peaks within the time of 0.25 second, the modulation frequency is simply the reciprocal of the peaks period or numerically 5/0.25 = 20 Hz.

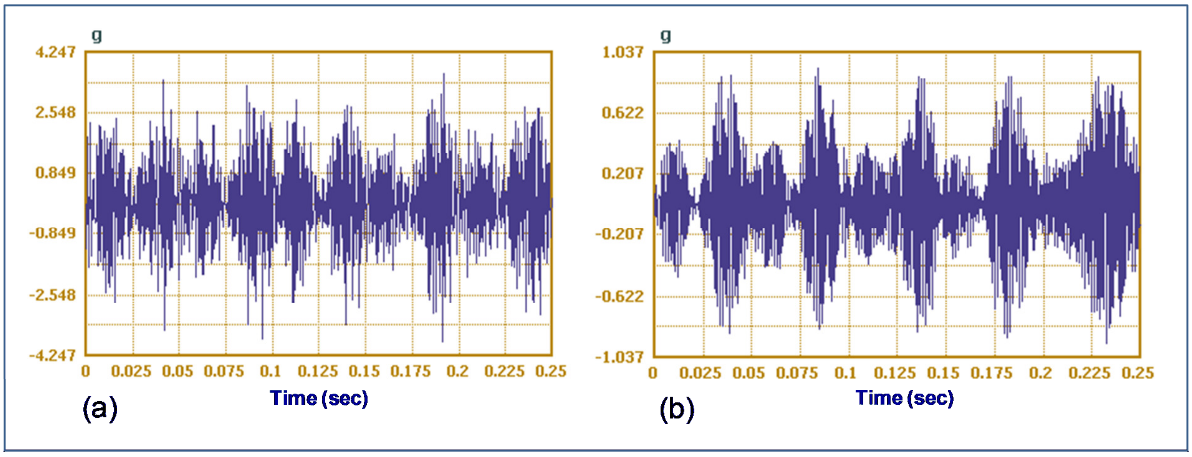

Figure 12.4 Time waveform of bearing vibration for P304 motor; (a) direct, (b) filtered at 803Hz

Figure 12.5 shows the bearing that was taken out from P-304 motor. The metal oxidization and rust clearly shown on the cage due to escape of water and other corrosive materials into the motor. The pump undergoes breakdown periods from time to time which could be the reason for contamination.

Figure 12.5 P-304 motor bearing

13. Automated Diagnosis and Advanced Techniques

Artificial Neural Networks

All the techniques presented in the above sections for fault diagnosis need human interpretation. Some techniques are easier to interpret while other techniques need deep knowledge and experience. Regardless of the processing technique used, automatic interpretation can improve the efficiency and reliability of fault diagnosis. Artificial neural networks (ANN) can identify and classify real data such as vibration signals. ANN consists of a number of richly interconnected artificial processing neurons called **nodes**, collected together in layers forming a network.

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process.

Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of ANNs as well. Figure 13.1 shows a multi-layered ANN.

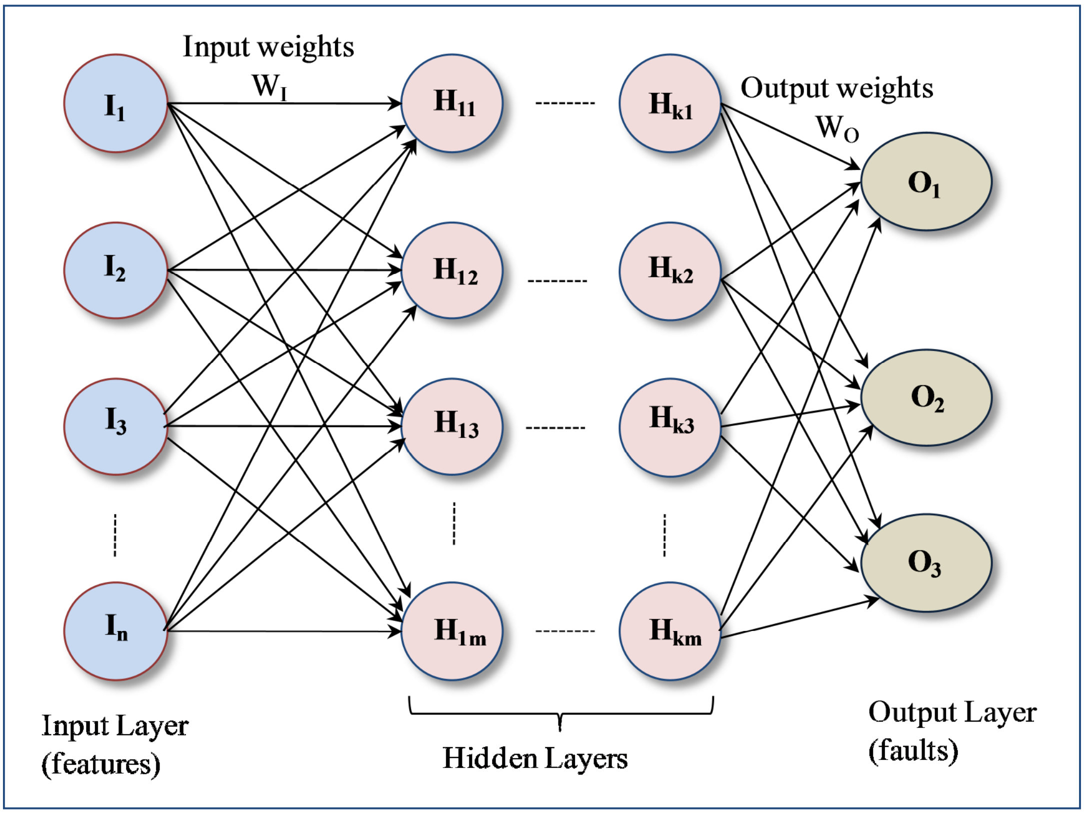

Figure 13.1 Multi-layered ANN

In artificial neurons, inputs are multiplied by weights and then calculated by an activation function. Another function estimates the output of the artificial neurons. The number of nodes within the input and output layers depends on the number of independent variables required to define the problem. The number of hidden layers and their nodes are selected by trial and error methods.

The general procedure for building an ANN is:

1. Define the problem input variables. These are the features or patterns that are measured, observed or estimated. For vibration diagnosis problems, these inputs can be the features of vibration signal such as RMS value, peak, crest, amplitudes of certain orders or any other processed value of the techniques discussed above or other techniques.
2. Define the problem output. These are the faults or conditions need to be detected.
3. Set the number of hidden layers and nodes.
4. Train the ANN by using training data of known inputs-outputs to find the appropriate connection weights. Back propagation is the most applied technique to find the weights.

5. Check the output of the ANN with training data and with testing data of known output to confirm the accuracy of the network. If the performance is not satisfactory, then the number of hidden layers and nodes can be adjusted and the whole process is repeated.

Drawbacks of ANN:

1. Need of precise training data that is not always available.
2. Human intervention still needed in the selection of training data and network topology.
3. Valid for already known and programmed faults. New faults or conditions cannot be interpreted without using additional techniques such as expert systems.
4. Computational and programming complexity especially for large networks with multiple inputs and outputs.

Principal Component Analysis

Principal Component Analysis (PCA) is a tool used to remove redundancy in possibly correlated data set. PCA is one of the dimensionality reduction tools and it is the backbone of many methods in features extraction and data retrieval. One of the most important perceptions which are provided by PCA is to recognize which variables in the system are more important, which are redundant, and which are noises. This is basically accomplished by projecting original correlated variables into a new space of uncorrelated variables, which are called, Principal Components (PCs).

The few first Principal Components are selected because they contain most important data while the other PCs are neglected because they contain an un-useful and a noisy data. To understand the math of PCA, suppose we have a matrix of data X with *M* vector, where each vector represents a measured (observed) variable with *N* time samples. The matrix is *MxN* sized.

$$\underline{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & & & \\ \vdots & & & \\ x_{M1} & & & x_{MN} \end{bmatrix} \tag{13.1}$$

The first step in PCA is to normalize the data. The zero mean and unity variance are examples of scaling techniques used in this regard. The next step is multiplying the normalized data by its transpose to produce a covariance matrix C:

$$C = X\ X^T \tag{13.2}$$

Matrix C, which is *MxM* sized, captures the correlations between all possible pairs of measurements for which the diagonal and off -diagonal terms are called the variance and covariance, respectively. More importantly, the correlation values reflect the noise and redundancy in the measured data. Large values in the diagonal terms correspond to interesting proper data while small diagonal terms indicate low SNR. On the other hand, large off-diagonal values correspond to high redundancy (highly correlated data) while small off-diagonal terms indicate low correlation.

The matrix C is de-factorized by using Singular Value Decomposition (SVD) to find the Eigen values of the problem. Some of the highest Eigen values are considered and used to construct a loading matrix P with the *Q* top Eigen values. So P is *MxQ* matrix. The aim of P matrix is to minimize the dimensionality of the original data by projection of the data on the P matrix to construct scores matrix T:

$$T = X^T P \tag{13.3}$$

The scores matrix is *NxQ* sized. The noise-free data can be constructed from the scores matrix as follows:

$$\hat{X} = (T\,P^T)^T \tag{13.4}$$

The squared prediction error (SPE) between the original and estimated data, given in the below equation, is good measure for fault development.

$$SPE(i) = \sum_{j=1}^{N} (x_{ij} - \hat{x}_{ij})^2 \tag{13.5}$$

When there is no fault, the SPE is normally small (below certain threshold). When fault is developed, the squared error is increased due to presence of strong features or principal components indicating some problem [13]. Also, the principal components can be used as input to an ANN driven automated diagnosis system. Figure 13.2 shows the Eigen values and SPE for normal and faulty bearing with different kinds of faults. High Eigen value, and consequently SPE, in the faulty bearings are easily detectable. All faulty bearings are characterized with SPE beyond the threshold limit set for a normal bearing [14].

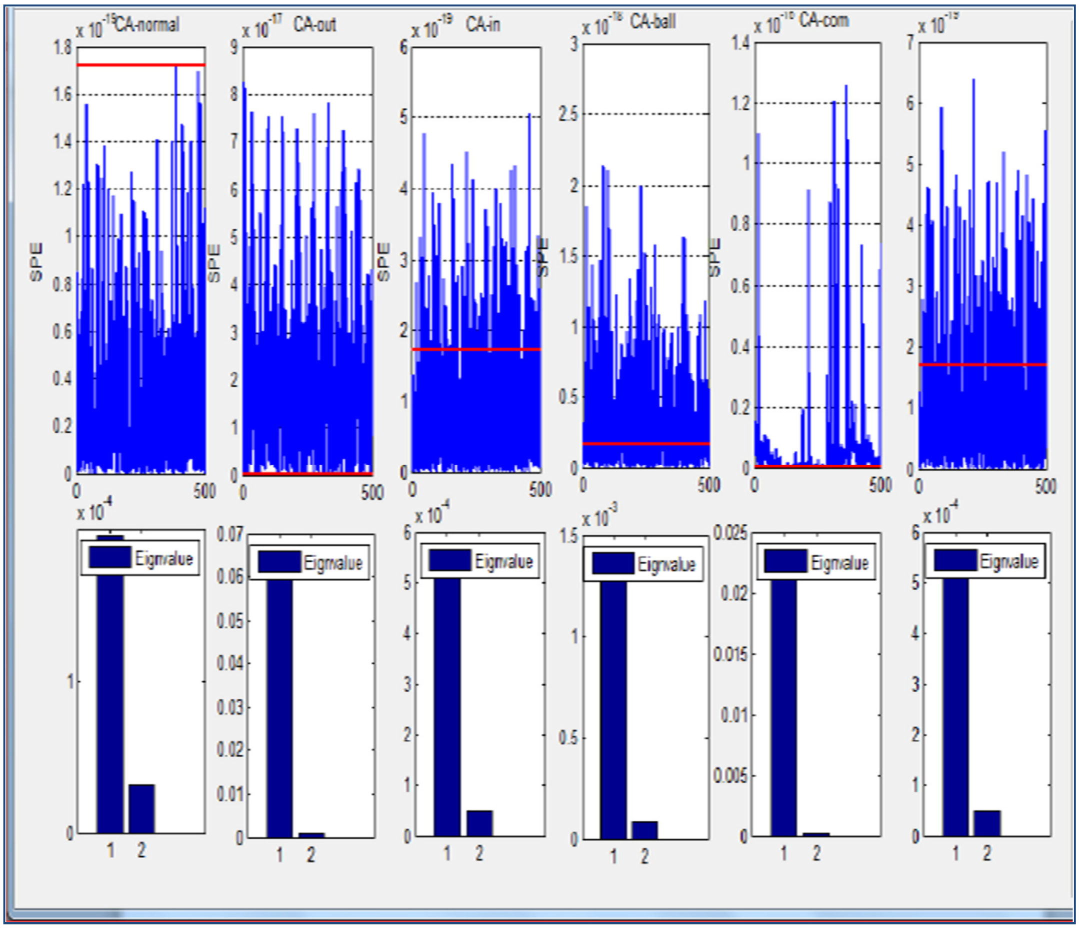

Figure 13.2 Eigen values and squared prediction error for normal and faulty bearings [14]

Support Vector Machine

SVM is computational learning method that based on the statistical foundations of learning theory. It is one of the classification methods to classify linearly separable (initially) data. Suppose some given data points each belong to one of two classes as shown in Figure 13.3, and the goal is to decide which class a new data point will be in. The classification in this case is based on two properties (features), so that each input data is 2-dimensional vector. The output is either Class1 (filled dots) or Class2 (empty dots). There are many hyperplanes that can classify the data, but in case of SVM, the optimum hyperplane is chosen such that the distance from it to the nearest data point on each side is maximized. More generally, a support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection.

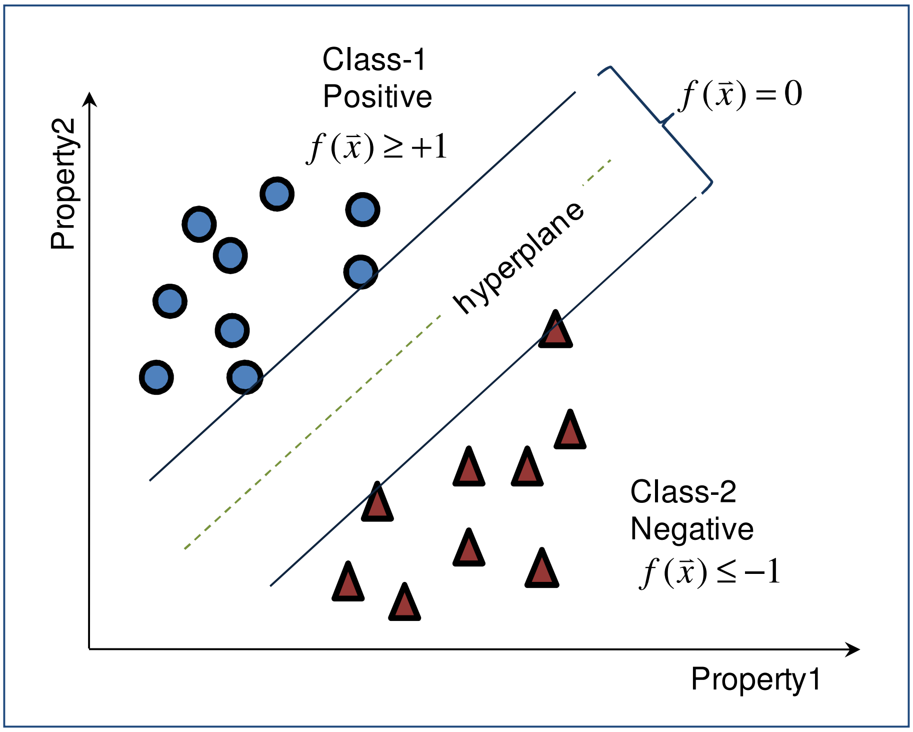

Figure 13.3 SVM classification of two classes

Suppose a set of input vectors $\vec{x}_i$ that can be classified into two classes; Class1 and Class2. Let y_i be the class selector such that y = +1 for Class1 and y = −1 for Class2. The SVM can be trained to find the function of hyperplane. The hyperplane is the plane whose data satisfy the following condition:

$$f(x_i) = \vec{\varpi} \cdot \vec{x}_i + b = 0 \ \ \forall \vec{x}_i \ \ \textit{on the boundaries of hyperplane} \tag{13.6}$$

Where The function *f()* must satisfies the following (classification) conditions:

$$f(\vec{x}_i) = \vec{\varpi} \cdot \vec{x}_i + b = \begin{cases} \geq +1 & \textit{for} \ \ y_i = 1 \\ \leq -1 & \textit{for} \ \ y_i = -1 \end{cases} \tag{13.7}$$

Where $\vec{\varpi}$ is the support vector whose dimension is the same as input vectors and b is a scalar. The values of support vector and the scalar are estimated by exploiting optimization techniques so that the classification error is minimum. The objective of optimization process is to obtain the maximum separation or the widest hyperplane.

SVMs can also be used in non-linear classification tasks by exploiting kernel tricks. The data to be classified is mapped onto a higher-dimensional feature space, where the linear classification is possible. The *n*-dimensional input vectors are mapped into *k*-dimensional feature space.

Multi-class classification is also possible with SVM. The above discussion deals with binary classification where the class labels can take only two values: 1 and −1. In the real-world problem, however, we find more than two classes for examples: in fault diagnosis of rotating machineries there are several fault classes such as bearing faults. The earliest used implementation for SVM multi-class classification is **one-against-all** methods. It constructs q SVM models where q is the number of classes. The ith SVM is trained with all of examples in the ith class with positive labels, and all the other examples with negative labels. Another major method is called **one-against-one** method. This method constructs $q(q-1)/2$ classifiers where each one is trained on data from two classes.

Figure 13.4 shows the flowchart of fault diagnosis using continuous Wavelet transform, PCA and SVM techniques [14]. The CWT is used as filtration technique to obtain filtered time data at specific scales that are corresponding to the bearing faults frequencies BPFI, BPFO, BSF and FTF. PCA is utilized to reduce dimensionality of the data and extract features. The output of PCA is used as input to the SVM classifier. The classifier must be firstly trained using sufficient known data. Then the algorithm can be used to classify the fault from an unknown case.

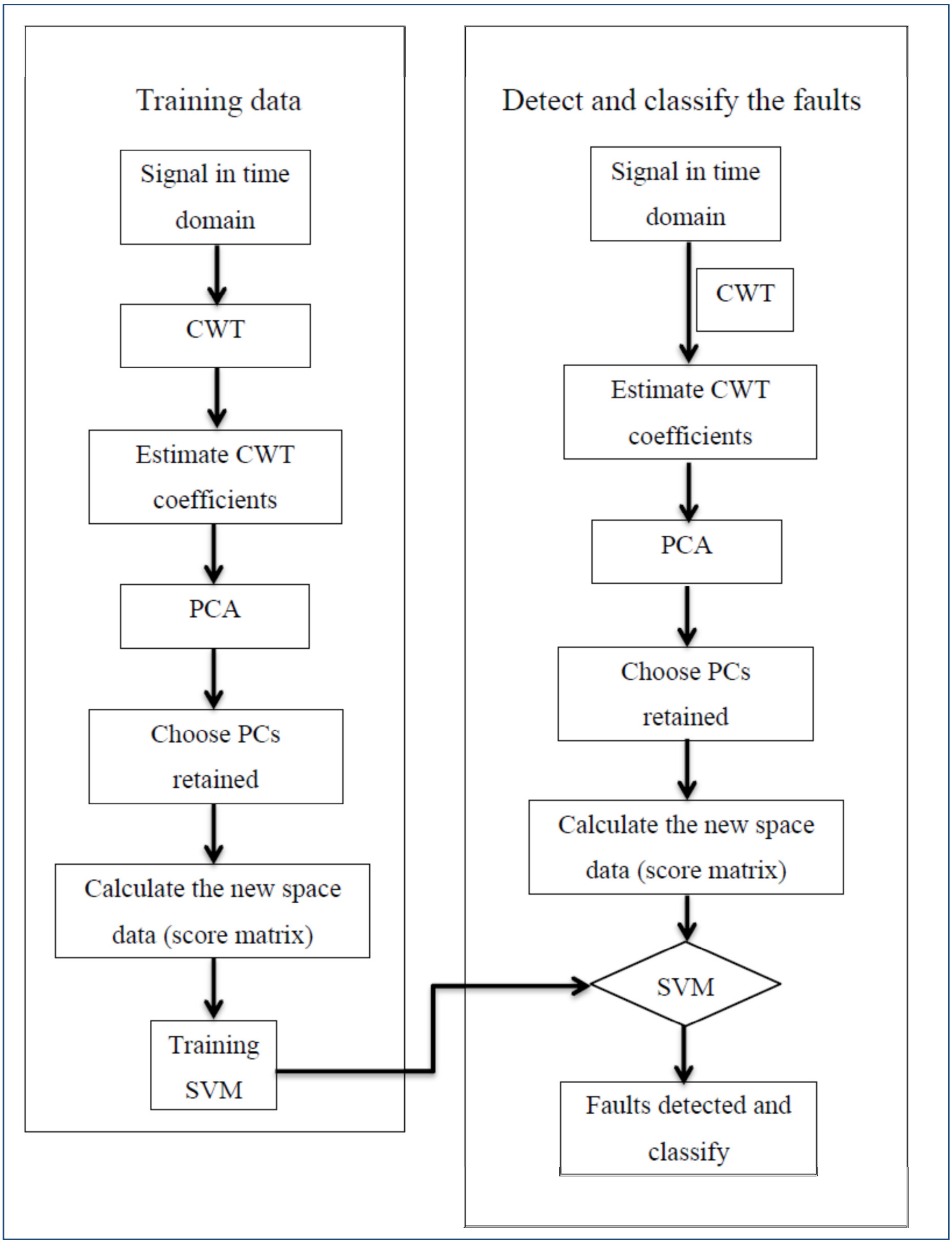

Figure 13.4 Flowchart of fault diagnosis based on SVM

Closing Remarks-Gap between academic research and professional practice

Gap between academic research and ongoing practice is seen more or less in many engineering and technology fields. While academic researches and scientific papers focus on developing advanced techniques or new methods, business organization, in general, rely on well-established, mature and traditional approaches. As related to the author profession, vibration analysis, and from both the academic and professional practice, as university professor and consultant in many industrial organizations, some of the reasons for that gap are:

1. The published researches are irrelevant to the practitioners.
2. Research papers are published in journals that are not accessible to practitioners (pure scientific or academic journals) and reviewed by academic reviewers.
3. Business organizations and system developers have their own developed techniques and do not want to invest in other (new) techniques. SPM and PeakVue are examples.
4. Some techniques are based on special and sophisticated software that are already covered by license or cannot be incorporated in the existing systems. Examples are Matlab and ANSYS based techniques.
5. Every new technique need to be well tested and take its way to the practical application.
6. Lack of practice. Some researches deal with specific cases in the laboratory and do not cover other practical situations where conditions differ and cases overlap.

While bridging the gap is out of scope of this work, a suggestion is to make the researches more practice-focused by applying Action Research, Collaborative Research and Decision Support Systems.

References

[1] Vibration Analysis Reference, Mobius Institute, http://www.mobiusinstitute.com

[2] Accurate Condition Monitoring of Ultralow and Low Speed Machines using SPMHD and HDENV Technologies with Case Studies, Tim Sundstorm, SPM Instrument, IMVAC2017, Dubai, 2017.

[3] Rolling Element Bearing Analysis, Brian Graney and Ken Starry, *Material Education*, Vol. 70, No. 1, pp 78-85.

[4] Enveloping and Spectral Emitted Energy, SKF Application Note CM3014.

[5] Frequency Analysis, Randall, R. B., Bruel & Kjaer, Denmark, 1987

[6] Effect of Inter-Modulation and Quasi-Periodic Instability in the Diagnosis of Rolling Element Incipient Defect, C. C. Osuagwu and D. W. Thomas, ASME J. Mech. Des 104(2), 296-302 (Apr 01, 1982).

[7] An Introduction to the SPM HD Method, Tim Sundstorm, SPM Instruments AB, Oct. 2010.

[8] Spike Energy Measurement and Case Histories, Ming Xu, ENTEK IRD International Corporation.

[9] Nathan Weller, Acceleration Enveloping-Higher Sensitivity Earlier Detection, GE Energy, http://www.ge-mcs.com/download/orbit-archives/2001-2005/2nd_quarter_2004/ 2q04accelenvel.pdf

[10] Rolling element bearing fault diagnostics using the blind deconvolution technique, Mahdi Karimi, PhD thesis, Queensland University of Technology, Sept. 2006.

[11] The enhancement of impulsive noise and vibration signals for fault detection in rotating and reciprocating machinery, S. K. Lee and P. R. White , J. Sound and Vibration, 1998, 217(3), pp. 485-505.

[12] Design and Construction of a Multi Channel Wireless-Based Vibration Monitoring and Analysis System, Jaafar K. Alsalaet, PhD thesis, University of Basrah, 2012.

[13] Multi Fault Diagnosis of the Centrifugal Pump Using the Wavelet Transform and Principal Component Analysis, K. Berli, K. McKee, R. Entwistle, I. Mazhar, and I. Howard, Proceedings of the 9th IFToMM International Conference on Rotor Dynamics, pp. 555-566. Springer International Publishing, 2015.

[14] Advanced Vibration Analysis in Bearing and Gearbox Faults Detection, Sajjad N. Abdulkareem, MSc. Thesis, College of Engineering, University of Basrah, 2015.

[15] Description of PeakVue and Illustration of its Wide Array of Applications in Fault Detection and Problem Severity Assessment, James C. Robinson and James E. Berry, Emerson Process Management Reliability Conference 2001.