State of the Art in Monitoring Rotating Machinery – Part 1

Robert B. Randall, The University of New South Wales, Sydney, Australia

In the last thirty years there have been many developments in the use of vibration measurement and analysis for monitoring the condition of rotating machinery while in operation. These have been in all three areas of interest, namely fault detection, diagnosis and prognosis. Of these areas, diagnosis and prognosis still require an expert to determine what analyses to perform and to interpret the results. Currently much effort is being put into automating fault diagnosis and prognosis. Major economic benefits come from being able to predict with reasonable certainty how much longer a machine can safely operate (often a matter of several months from when incipient faults are first detected). This article discusses the different requirements for detecting and diagnosing faults, outlining a robust procedure for the former, and then goes on to discuss a large number of signal processing techniques that have been proposed for diagnosing both the type and severity of the faults once detected. Change in the severity can of course be used for prognostic purposes. Most procedures are illustrated using actual signals from case histories. Part 2 of this article will appear in the May 2004 issue of S&V.

The vibrations measured externally on operating machines contain much information about their condition, as machines in normal condition have a characteristic "vibration signature," while most faults change this signature in a well-defined way. Thus, vibration analysis is a way of getting information from the inside of operating machines without having to shut them down. Another way of getting information from operating machines is by analysis of the lubricant, and "oil analysis" is useful in machine condition monitoring. This article is concerned only with vibration analysis techniques.

Machine vibrations are measured in two fundamentally different ways – relative displacement of a shaft in its bearings using so-called "proximity probes," and absolute motion of the casing (usually at the bearings) using absolute motion transducers. Proximity probes must be designed into the machines and are typically used on high speed turbomachines with fluid film bearings. They are used for permanent monitoring of relatively simple parameters such as peak-to-peak relative displacement and shaft orbits (in the bearing) and are primarily used to protect valuable and critical machines by shutting them down in the event of excessive vibration. Only in a limited number of situations can long term predictions be made. This is because incipient faults often show up first at high frequencies, to which the relative displacement measurements are not sensitive. Proximity probes have a frequency range up to 10 kHz, but because of the natural reduction of displacement amplitudes with frequency and the dynamic range limitation of proximity probes to 30-40 dB, the limitation is really of harmonic order (to about 10-12 harmonics). The dynamic range limitation is determined mainly by electrical and mechanical runout, i.e. the signal measured in the absence of vibration. The higher dynamic range limit corresponds to the use of "runout subtraction," where the runout measured at low speed can be subtracted from other measurements at high speed. This technique is somewhat dubious over long periods of time where the originally measured runout may have changed.

Since all vibrations represent an alternation between potential energy (in the form of strain energy) and kinetic energy, vibration velocity is the parameter most closely related to stress, and is the parameter used to evaluate severity in most vibration criteria. For the same reason, a velocity spectrum is usually 'flattest' over a wide frequency range, requiring the minimum dynamic range to represent all important components. By comparison, vibration displacement tends to over-emphasize low frequencies (as for relative displacement) while vibration acceleration tends to over-emphasize high frequencies. The latter can sometimes be useful for faults, such as in rolling element bearings, which show up first at high frequencies, but may disguise changes at low frequencies. However, the best and most common transducer for measuring absolute casing vibration is the piezoelectric accelerometer which produces a signal proportional to acceleration. Its dynamic range is so large (160 dB) that it can be combined with electronic integration to give a velocity signal with more than 60 dB dynamic range over three decades in frequency. This cannot be achieved by typical velocity transducers with an upper frequency limit of 1-2 kHz.

For these reasons, the rest of this article mostly assumes measurements made with accelerometers, sometimes integrated to velocity. The meaning of "rotating machines" has been taken to include reciprocating machines such as diesel engines. Because of their importance and ubiquity, the measurement of torsional vibration of the crankshaft is included as a supplementary technique.

Fault Detection

As mentioned above, the use of accelerometers, possibly with integration to velocity, allows the measurement of signals with a frequency range of more than three decades, e.g., 5 Hz-5 kHz or 20 Hz-20 kHz, with very good dynamic range. Such a range can be necessary to detect the full range of possible faults. With fluid film bearings these can extend down to 40% of shaft speed (e.g. oil whirl) up to at least the 400th harmonic of shaft speed (e.g. harmonics of gearmesh and bladepass frequencies). Rolling element bearings often have fault indications at frequencies on the order of 1000 or more times the shaft speed. Criteria exist for vibration severity, such as the ISO Standard 2372 (developed from the German recommendation VDI-2056), and the so-called "General Machinery Criterion Chart,"[1] widely used in the USA, developed from the earlier Rathbone and Yates charts. As mentioned above, these all represent equal velocity criteria, for a wide range of machine sizes and speeds, and can be expressed in terms of RMS levels covering the frequency range 10-1000 Hz. The reason for the upper frequency limitation is not for any good technical reason, other than the fact that much of the data on which it was based were obtained using velocity probes with that frequency range.

The ISO 2372 standard has different criteria depending on the size of the machine, and whether they are flexibly or rigidly mounted. Thus, there must be differences from the "General Machinery Criterion Chart," which does not differentiate. Both criteria are in agreement that equal changes in severity are represented by equal changes on a log amplitude scale, and that a change of 20 dB (vibration velocity ratio of 10:1 or 1:10) is serious. The number of grades between 'good' and 'faulty' differ slightly, but it can be inferred that a significant change is represented by a change of 6-8 dB (vibration velocity ratio of 2-2.5). There is no doubt that typical vibration levels will tend to vary with the size and type of machine, but in one

Based on a paper presented at ISMA 2002, the International Conference on Noise and Vibration Engineering, Leuven, Belgium, September, 2002.

study[2] it was found that even for machines of the same class (ethylene compressors in a petrochemical plant), the mechanical impedance of the bearings varied over a very wide range. That means the same measured vibration level would represent very different internal forces, in particular at different frequencies. Thus, rather than using absolute criteria, a strong argument can be made for detecting faults based on the change from the normal levels at each measurement point, with 6 dB and 20 dB representing significant and serious changes, respectively.

The use of velocity means that there is a better chance that changes at any frequency will affect the overall RMS levels. But it is still evident that monitoring of frequency spectra, rather than overall levels, will have a better chance of detecting changes at whatever frequency they should occur. There are very good reasons why the spectra used for comparison should be of the constant percentage bandwidth (CPB) type, rather than FFT constant bandwidth spectra:

- A 1/18-octave (4% bandwidth, log frequency) spectrum can cover three decades in frequency with 180 spectrum values, whereas a single (linear frequency) FFT spectrum only covers 1 to 1.5 decades, meaning that several have to be used to guard against all possibilities.
- Even minor speed changes, such as those given by slight load variations with an induction motor, make it very difficult to compare FFT spectra, whereas on a log frequency scale, a small speed change (on the order of the bandwidth) can be compensated by a lateral shift of the spectrum. Smaller changes will be included within the bandwidth.
- To aid the comparison of digitized CPB spectra, a mask can easily be made by smearing a reference spectrum to account for the large changes in sample values along the flanks of discrete frequency components due to small speed changes less than the bandwidth.

Figure 1 shows the application of this technique to signals from an auxiliary gearbox on a gas turbine-driven oil pump.[3] It shows the comparison of a spectrum with a mask formed from the original reference spectrum, and the resulting spectrum of exceedances. This comparison is for the situation four months after the first detection of the fault. Two remarks are worthy of mention at this point:

- The maximum change of 20 dB is quite serious, but stabilized at this level, the machine was allowed to run for a further five months before being repaired at a convenient time.
- Despite the significant change of a number of frequency components, the overall RMS value of the signal would not have changed, because of the masking effect of strong adjacent components. The fault was in a bearing, but the spectrum was dominated locally by strong gear-related components.

This fault detection procedure has proven itself to be very robust on a wide range of different cases over more than twenty years.

Fault Diagnosis

Once a significant change indicating a potential fault has been detected, it is usually necessary to perform other signal processing techniques to make a diagnosis of the fault(s), which depend greatly on the type of fault expected. Not much diagnostic information can be extracted from the CPB spectrum, in particular because it is on a logarithmic frequency axis, and this disguises things such as harmonic patterns that are very valuable diagnostically. However, the frequency range where the change occurred is valuable information and guides the selection of the appropriate linear frequency range for FFT spectra to be used diagnostically.

The type of analysis to be applied depends on the type of fault, and so it is interesting to investigate how various faults manifest themselves in the vibration signal.

Shaft Speed Faults. A number of faults manifest themselves at a frequency corresponding to the speed of the shaft. Among these are unbalance, misalignment and cracked shaft, which are difficult to distinguish from each other. This is one area where proximity probes can be useful, as their ability to determine the mean position of the shaft in the bearing as well as the shape of the orbit can help differentiate between unbalance and misalignment. *In general*, misalignment tends to produce a stronger second harmonic of shaft speed and more axial motion. Even though the unbalance force is purely at shaft speed, the nonlinearity of components in the system (such as fluid film bearings) can distort the response motion, leading to higher harmonics, while moment unbalance gives rocking motions with axial components. For constant unbalance distribution and alignment, cracks in shafts give changes at the first and second harmonics of shaft speed, though these may be more evident as a change in phase angle (relative to the phase of a once-per-rev tacho signal) rather than in amplitude. For ‘breathing’ cracks, opening and closing each revolution, the third and other odd harmonics are also excited. Sophisticated rotor dynamic models are now being made of the most critical machines and these provide the best possibilities for distinguishing between unbalance, misalignment and shaft cracks, at least with the machine running at speed and load. Cracks can be detected during rundowns, not so much by a change in critical speed, as by an increased response when harmonics of shaft speed pass through the critical speed(s).

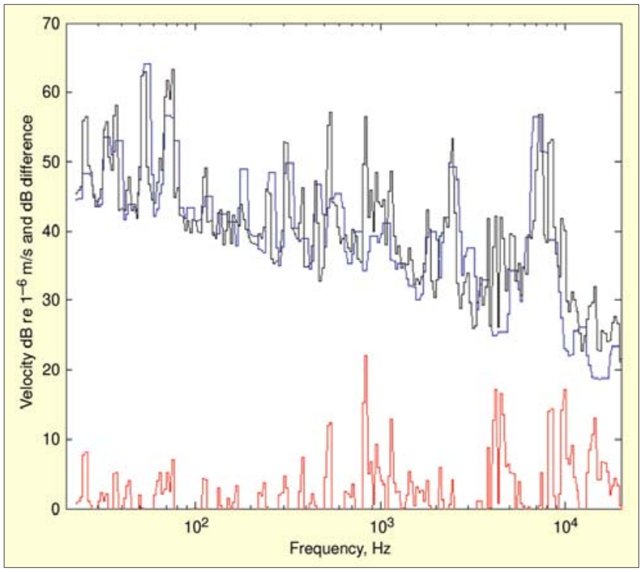

Figure 1. (Upper) Comparison of new spectrum with mask, velocity dB re 10^{-6} m/s. (Lower) dB difference spectrum.

Other faults show up at sub-synchronous frequencies, such as “oil whirl,” usually at 40-48% of shaft speed, caused by a resonant wave in the fluid film bearing. This can sometimes be confused with “hysteresis whirl,” due to hysteretic friction between components on the rotor. The friction forces are such that a self-excited whirl is generated when passing through the shaft critical speed and remains at this frequency as the shaft speed increases. Since many machines run at about twice their first critical speed, this can sometimes be confused (and perhaps even merged) with oil whirl. When oil whirl combines with resonant shaft response it is sometimes called “oil whip.” Exact subharmonics (e.g. 1/2, 1/3) can result from “parametric excitation,” variations in stiffness due to looseness, ‘rubs’ etc., and can be distinguished by their exact subharmonic nature.

Electrical Machine Faults. Electrical machines such as AC motors and generators produce vibrations due to electrical as well as mechanical forces.[4] Stator faults tend to give increases at twice mains frequency, as this is the rate at which the poles of the rotating magnetic field are passing a fixed point (the anomaly) on the stator. For two-pole synchronous machines, this is the same as twice shaft speed, making it difficult to distinguish between a stator fault and misalignment. However, the electrical forces are strongly dependent on the load and varying the load may allow the two effects to be separated. Switching off the power and tracking the second harmonic as it runs

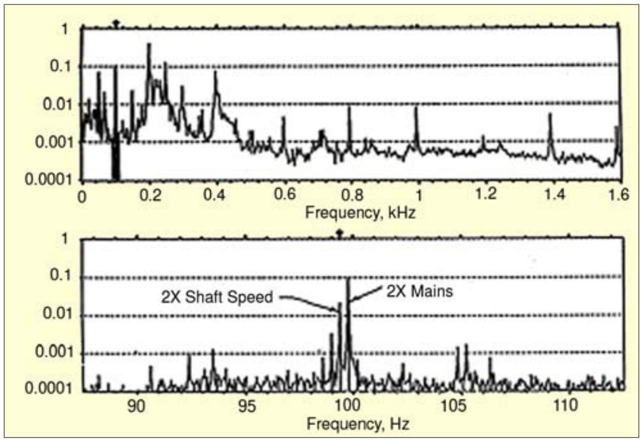

Figure 2. Upper – baseband spectrum. Lower – use of FFT zoom spectrum in marked area of baseband spectrum to separate the harmonics of shaft speed from those of mains frequency.

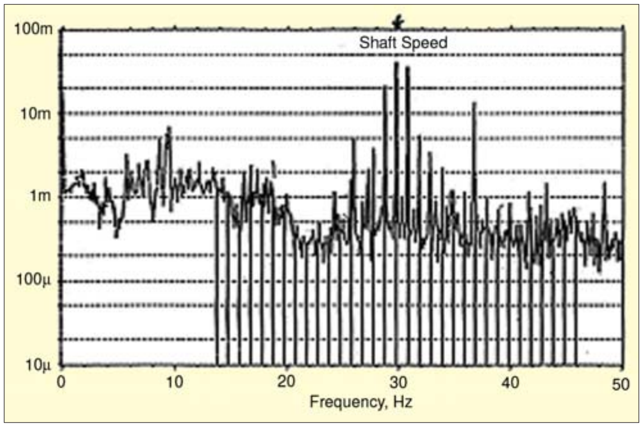

Figure 3. Example of a rotor fault on an induction motor, showing modulation sidebands around the shaft speed component. The main cursor (shaft) frequency is 29.75 Hz and the sideband cursor spacing is 0.25 Hz.

down in speed allows complete separation.

For induction motors[4] the situation is easier in that the shaft speed is less than synchronous and FFT zoom analysis allows separation of the harmonics of shaft speed from those of mains frequency. Figure 2 shows an example, where from the upper baseband spectrum, it appears that the second harmonic of shaft speed is elevated. However, the lower zoom analysis centered on this frequency shows that it is the second harmonic of mains frequency that dominates and the second harmonic of shaft speed is five times lower.

Figure 3 shows an example of a fault on the rotor of a four-pole induction motor (in the USA where the mains frequency is 60 Hz). The main effect is at the shaft speed (corresponding to the rate of rotation of the fault), but it can be distinguished from mechanical unbalance by virtue of the strong modulation sidebands. The sideband cursor shows that these are spaced at 1.0 Hz, which is the number of poles times the slip frequency of 0.25 Hz (synchronous speed 30 Hz minus shaft speed 29.75 Hz). This is the frequency at which the poles of the rotating field pass a given point (the anomaly) on the rotor.[4]

Gear Faults. Gears represent a typical component where the wide frequency range of accelerometers is needed. The basic vibration generating mechanism in gears is the "transmission error" (TE), which can be understood as the relative torsional vibration of the two gears, corrected for the gear ratio. The TE can be expressed as a linear relative displacement along the line of action, which is the same for both gears but represents an angular displacement inversely proportional to the number of teeth on each gear.

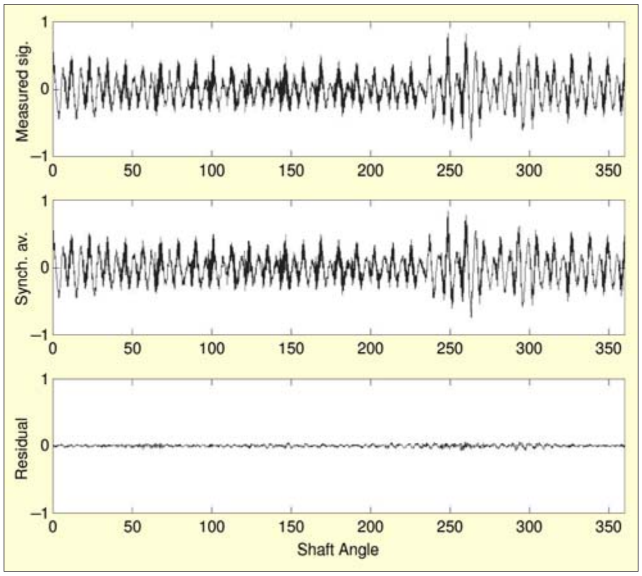

Figure 4. Comparison of an original gear signal (upper) with a synchronous average (middle) and their difference (lower) for a simulated tooth root crack at roughly 250°.

The TE results from a combination of geometric errors of the tooth profiles and deflections due to tooth loading. Thus, even a gear with perfect involute profiles will have some TE under load. It is thus important to make comparisons of gear vibration spectra under the same load to obtain information about changes in condition.

Gear vibration signals are dominated by two main types of phenomena:[5]

- Effects that are the same for each meshing tooth pair, such as the tooth deflection under load and the uniformly distributed part of initial machining errors and/or wear. These manifest themselves at the toothmeshing frequency and its harmonics. Since there is a pure rolling action at the pitch circle and sliding on either side, tooth wear tends to occur in two patches on each tooth. Wear is thus often first seen as an increase in the second harmonic of the toothmeshing frequency.
- Variations between the teeth, which can be localized or distributed more uniformly around the gears. These manifest themselves at other harmonics of the gear rotational speeds, for the gear on which they are located. Localized faults such as cracks and spalls tend to give a wide range of harmonics and sidebands throughout the spectrum, whereas more slowly changing faults such as those due to eccentricity and distortion during heat treatment, tend to give stronger harmonics grouped around zero frequency and as sidebands around the harmonics of toothmesh frequency.

Since even with faults the same geometric shapes always mesh in the same way, the signals produced by gears are basically deterministic, at least as long as the teeth remain in contact.

This is illustrated in Figure 4[6] for a gear with a simulated tooth root crack meshing with a normal gear. The signal for one rotation is seen to be very similar to the synchronous average over several rotations, leaving a residual signal that is close to zero. In this case both gears had the same number of teeth, making the fundamental meshing period the same as one rotation of each gear, but in general this period (after which all gears again have the same orientation) would be much longer. The signal corresponding to each gear can however be extracted by averaging synchronously with the rotation of the gear concerned.

For light load or very large geometric errors the teeth can lose contact and introduce some randomness or chaotic nature to the signals. For condition monitoring it is better for the loading to be sufficient to maintain tooth contact, to ensure that changes in the vibration signals are due to changes in condition.

Bearing Faults. This discussion is limited largely to rolling element bearings, since with fluid film bearings in principle

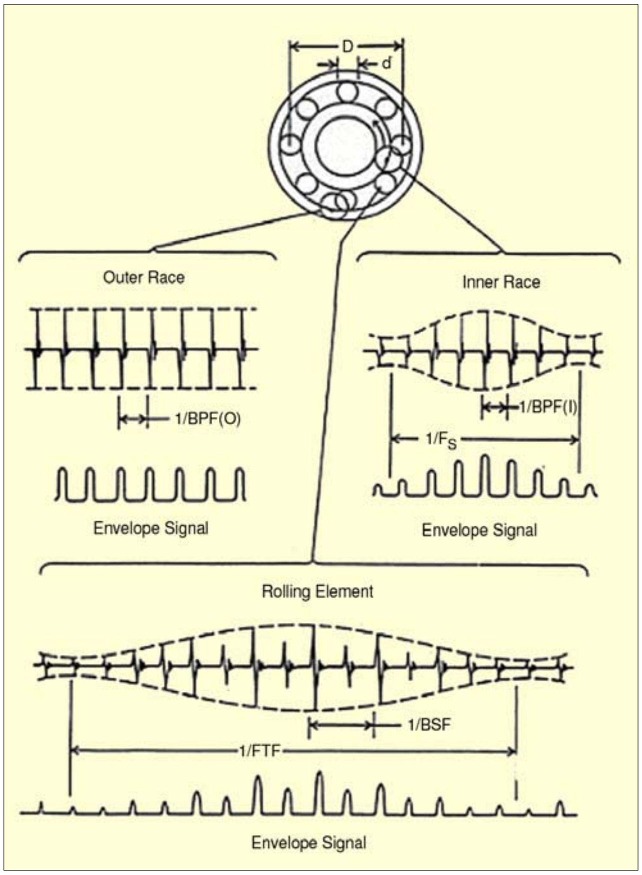

Figure 5. Typical signals and envelope signals from local faults in rolling element bearings. BPFO = ballpass frequency, outer race; BPFI = ballpass frequency, inner race; BSF = ball spin frequency; FTF = fundamental train frequency (cage frequency).

there should not be any metal-to-metal contact and consequent wear. There have been very few studies of detecting the wear of fluid film bearings from their vibration signals, but the operational faults that could give rise to such wear can be monitored by the techniques described in "Shaft Speed Faults." This is also one area where the use of oil analysis can aid vibration analysis, as bearing metals are quite distinctive in their chemical composition.

Rolling element bearings do eventually wear out, and it is very valuable to detect their deterioration at an early stage. Figure 5 shows typical acceleration signals produced by localized faults in the various components of a rolling element bearing, along with the corresponding envelope signals produced by amplitude demodulation. It will be shown that analysis of the envelope signals gives more diagnostic information than analysis of the raw signals.

The diagram illustrates that as the rolling elements strike a local fault on the outer or inner race, a shock is introduced that excites high frequency resonances of the whole structure between the bearing and the response transducer. The same happens when a fault on a rolling element strikes either the inner or outer race. The series of broadband bursts excited by the shocks is further modulated in amplitude by two factors:[7]

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

For the common case of uni-directional load (e.g., completely dominating over unbalance load) outer race faults will tend to occur in the load zone and the bursts will not be modulated as illustrated in Figure 5. On the other hand, inner race faults pass through the load zone at shaft frequency and rolling elements pass through the load zone at the fundamental train frequency or FTF (i.e., cage frequency). Note that the ballspin frequency (BSF) is the frequency with which the fault strikes the same race (inner or outer), so that in general there are two shocks per basic period. If these shocks (or at least their envelopes) were identical, the odd harmonics would vanish and the fundamental frequency would be twice BSF.

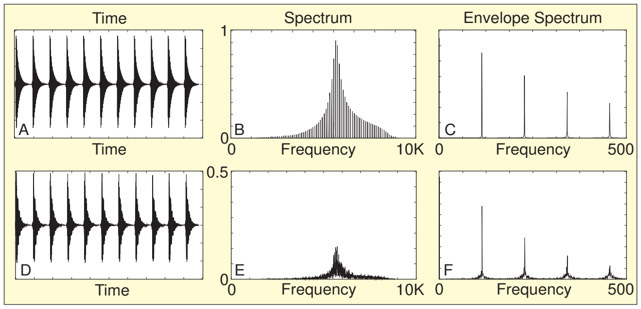

Figure 6. Bearing fault pulses with and without frequency fluctuation; (A, B, C) no frequency fluctuation; (D, E, F) 0.75% random frequency fluctuation.

The formulae for the various frequencies shown in Figure 5 are as follows:

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

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

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

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

where f_r is the shaft speed, n is the number of rolling elements and ϕ is the angle of the load from the radial plane.

These are, however, the kinematic frequencies assuming no slip. In reality there must virtually always be some slip for the following reason. The angle ϕ varies with the position of each rolling element in the bearing as the ratio of local radial to axial load changes. Thus each rolling element has a different effective rolling diameter and is trying to roll at a different speed. The cage ensures that the mean speed of all rolling elements is the same by causing some random slip. This is typically on the order of 1-2%.

This random slip, while small, does give a fundamental change in the character of the signal and is the reason why envelope analysis extracts diagnostic information not available from frequency analyses of the raw signal. It also allows bearing signals to be separated from gear signals[6] with which they are often mixed, as discussed below.

Figure 6 illustrates the effect of the small random frequency fluctuations on the spectrum and envelope spectrum,[8] as typified by an outer race fault. Figure 6a shows a series of high frequency bursts at a rate corresponding to the ballpass frequency with no random fluctuation. It is assumed that just one resonance frequency (e.g., the lowest) is excited, and so the harmonics of the repetition frequency in Figure 6b represent samples of the spectrum of one of the pulses (expressed in terms of acceleration). The values of the low harmonics are obviously very small, and only become significant in the vicinity of the resonance frequency, where their spacing indicates the repetition frequency. The envelope spectrum in Figure 6c, the frequency analysis of the envelope signal obtained by amplitude demodulation of the signal in Figure 6a, has strong low harmonics, as it corresponds to a series of pulses as in Figure 5. The small random fluctuation in the spacing of the bursts in Figure 6d can hardly be seen by the eye. Still, it gives a smearing of the higher harmonics in Figure 6e, so that no diagnostic information can be extracted from the raw spectrum. However, the envelope spectrum of Figure 6f clearly indicates

$$\begin{Bmatrix} G_0 \\ G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \\ G_6 \\ G_7 \end{Bmatrix} = \frac{1}{8} \begin{bmatrix} \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ \uparrow & \nearrow & \rightarrow & \searrow & \downarrow & \swarrow & \leftarrow & \nwarrow \\ \uparrow & \rightarrow & \downarrow & \leftarrow & \uparrow & \rightarrow & \downarrow & \leftarrow \\ \uparrow & \searrow & \leftarrow & \nearrow & \downarrow & \nwarrow & \rightarrow & \swarrow \\ \uparrow & \downarrow & \uparrow & \downarrow & \uparrow & \downarrow & \uparrow & \downarrow \\ \uparrow & \swarrow & \rightarrow & \nwarrow & \downarrow & \nearrow & \leftarrow & \searrow \\ \uparrow & \leftarrow & \downarrow & \rightarrow & \uparrow & \leftarrow & \downarrow & \rightarrow \\ \uparrow & \nwarrow & \leftarrow & \swarrow & \downarrow & \searrow & \rightarrow & \nearrow \end{bmatrix} \begin{Bmatrix} g_0 \\ g_1 \\ g_2 \\ g_3 \\ g_4 \\ g_5 \\ g_6 \\ g_7 \end{Bmatrix}$$

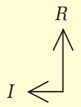

Figure 7. Matrix representation of the DFT.

the average burst frequency, even if the higher harmonics are a little smeared.

Signals such as the one shown in Figure 6d-f, which are not periodic but have a hidden periodicity that can be extracted by demodulation, are known as *cyclostationary*.[9] Because they are often generated in rotating machines along with deterministic discrete frequency signals, they are treated below in a special section.

Reciprocating Machine Faults. Reciprocating machines, such as diesel engines and reciprocating compressors, also produce vibrations with both periodic and cyclostationary components. The latter can be associated with combustion, which occurs every basic cycle, but not identically each time. Signals from reciprocating machines have a different character from those of simple rotating machines, as they consist of a series of impulsive events (combustion, piston slap, valves opening and closing, etc.), and so the most effective analysis techniques must take into account variations in both frequency and time (or crank angle). This is merely an attempt to match what experienced mechanics do with their ears in distinguishing between bearing knock, combustion knock, piston slap, etc.

Combustion-related faults could be recognized by viewing the cylinder pressure signal throughout the cycle, but this requires having a pressure transducer in every cylinder, something that is not very practical even though test engines are sometimes instrumented this way in the laboratory. Efforts are currently being made to reconstruct cylinder pressure signals from external measurements that react directly to the cylinder pressure, such as accelerations of the block or head, or torsional vibrations of the crankshaft.

A simple indication of (complete or partial) misfire is given by viewing the crankshaft torsional vibration directly, as uniform firing on all cylinders gives uniform torque pulses, and corresponding uniform angular velocity fluctuations for the combustion on each cylinder. This is illustrated below.

Signal Processing Techniques

Once faults have been detected, it is necessary to apply a range of signal processing techniques to the vibration signals to try to determine the reasons for the spectral change. In the following, a number of classical and newer techniques are reviewed.

FFT Analysis. As illustrated in Figures 2 and 3, an FFT (fast Fourier transform) spectrum is a powerful diagnostic method, in particular when combined with zoom analysis as in Figure 2b, and a harmonic/sideband cursor as in Figure 3. The FFT is a fast algorithm for calculating the DFT (discrete Fourier transform), of a block of $g(n)$ N samples of data, giving a spectrum $G(k)$ of N frequency lines, using the formula:

$$G(k) = 1/N \sum_{n=0}^{N-1} g(n)\exp(-j2\pi kn/N) \qquad (5)$$

Equation 5 actually assumes that $g(n)$ is one period of a periodic signal, so that the spectrum is that of the corresponding Fourier series. The sample index number n represents time $n\Delta t$, where Δt is the sample spacing, the reciprocal of the sampling frequency f_s. Similarly, the frequency index k represents frequency $k\Delta f$, where Δf is the line spacing, the reciprocal of the record length T (= $N\Delta t$). Because the time signal is sampled, the spectrum $G(k)$ is also periodic, with a period equal to the sampling frequency f_s. In the normal situation where the signal $g(n)$ is real, the negative frequency components are the complex conjugates of the positive frequency components, and there are thus only $N/2$ independent (but complex) spectrum values. Because of the periodicity of the spectrum, the second half (from $f_s/2$ to f_s) actually represents the negative frequency components (from $-f_s/2$ to zero). This also explains why all frequencies in the original signal outside this range ($-f_s/2$ to $f_s/2$) must be removed by a lowpass filter before digitization, as they would otherwise mix with the true components within this range (causing 'aliasing').

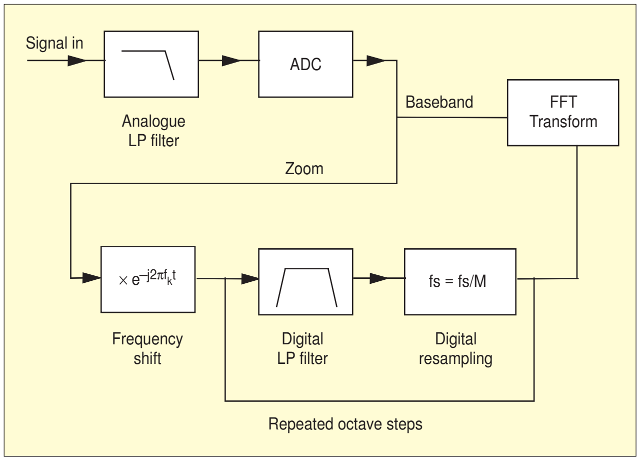

Figure 8. Schematic diagram of FFT zoom process.

Figure 7 is a matrix representation of Equation 5 for $N = 8$. The square matrix shows the orientation of the complex exponential components (unit vectors) for the various values of k (frequency index) and n (time index). Note the orientation of the real and imaginary axes (R and I). The first row represents zero frequency, and the first column zero time.

All the vectors in the first row equal unity, so the zero frequency spectrum value is simply the sum of the time samples divided by N (= 8), giving the average value as expected. The vectors in the second row rotate $-1/N^{th}$ of a revolution per time sample, corresponding to a single rotation over the record length, and thus give the first harmonic of the periodic signal. The next row rotates twice as fast, giving the second harmonic and so on. The fifth row represents half the sampling frequency, and all rows after this are more easily understood as rotations in the opposite direction, giving the negative frequency components. The last row, for example, gives minus the fundamental frequency. The periodicity of the spectrum can be understood by realizing that the first row equally well represents the sampling frequency (one revolution per time sample) as zero frequency.

Zoom FFT. The normal frequency range of an FFT spectrum is from zero to half the sampling frequency, but as shown in Figure 2, it can be an advantage to "zoom in" on a narrower frequency range. This was once done by two techniques known as "non-destructive zoom" and "real-time zoom,[10] where the first was basically a way of obtaining the FFT transform of a long record by combining the results of smaller transforms. This was useful when FFT analyzers typically had a fixed transform size. Now with less restriction, the same result can be achieved by performing a large transform directly. The zoom factor (compared with an original transform) is simply the ratio of transform sizes, and as with non-destructive zoom the longer record must be accommodated in the memory.

With real-time zoom, virtually any zoom factor can be achieved without storing the original record, as long as it can be processed in real-time by a zoom processor. This normally requires special hardware in a dedicated analyzer. The principle is shown in Figure 8, where the input signal is frequency shifted, lowpass filtered and decimated to a lower sampling frequency in real-time. It is only the decimated signal that is stored and FFT transformed, vastly reducing the storage requirement. To zoom in another frequency band would require the original signal to be stored separately (e.g. on a DAT recorder).

Multiplication of the original signal by exp($-j2\pi f_k t$) subtracts frequency f_k from every frequency in it and thus shifts the zoom center frequency f_k to zero. The lowpass filtering permits resampling to a lower frequency without aliasing.

The lowpass filtering and resampling are usually done in octave (2:1) steps, as this can be repeated as many times as desired in real-time to obtain zoom by a factor equal to any power of two. This procedure is explained in a following section on digital filtering.

A zoom processor such as that in Figure 8 can also be used for demodulation as explained in a following section.

Practical FFT Analysis. The DFT actually produces the Fourier series spectrum of a periodic repetition of the record transformed. This must be taken into account when it is used on another type of signal. When the record does not correspond to an integer number of periods of all frequencies in the original signal, the periodic repetition will give a distortion because of the sudden step where the two ends are joined into a loop. The effects of this can be mitigated by applying a "time window" to the signal before transformation, to force the value and slope to zero at the joint and avoid a discontinuity. Since what is analyzed is then an amplitude modulated version of the original signal, spectral peaks are surrounded by sidebands, but these are usually less disturbing than the effects of no special window. The most commonly used window is the Hanning window, one period of a sine-squared function, and if it is scaled so as to read the same value at the center of a discrete frequency peak, the sidebands give extra power by a factor of 1.5. This must be compensated when integrating over a frequency band or calculating the PSD (power spectral density) of a broadband signal. The sidebands introduced by a window function give rise to so-called 'leakage' (of power away from the central frequency), and this is minimized by windows such as Hanning. Where a frequency component falls between two analysis lines (the $G(k)$ of Equation 5), it will be divided between them, and neither will show the true peak value. This is known as the "picket fence effect" and is for example a maximum of 1.4 dB for a Hanning window, and as much as 3.9 dB for a rectangular window (which results when no special weighting is used). The so-called "flat-top window" has been designed to eliminate this picket fence effect, and is thus most useful for signals dominated by discrete frequency components and in particular calibration signals. On the other hand its bandwidth factor is 3.7 (in comparison with the 1.5 mentioned above for the Hanning window) so that the discrete frequency components do not protrude as much from any noise in the spectrum.

As with Fourier series, the results of the DFT are scaled in the same units as the original signal (as follows from Equation 5), but this is only relevant for the discrete frequency components. Note that the positive frequency components must be scaled up by a factor of $\sqrt{2}$ to give RMS values (including the negative frequency part) or 2 to give sinusoidal amplitudes. If the original signal is other than a (quasi-)periodic signal, the output of the DFT must be modified to give correctly scaled results.

If the signal were stationary random, for example, its spectrum should be scaled as a PSD in U^2/Hz to give consistent results, where U represents the units of the original signal. The 'power' in a discrete frequency line from the DFT equals the square of its magnitude (to be multiplied by 2 to get the total mean square value at that frequency including the negative frequency component). This must be divided by the bandwidth in Hz to get an estimate of the PSD of one record of a random signal. Since the line spacing Δf always equals $1/T$ for the DFT, where T is the record length transformed (in seconds), the conversion to PSD can be achieved by a multiplication by T (as well as being divided by the bandwidth factor for the particular window as mentioned above).

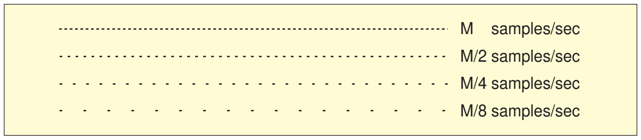

Figure 9. Effect of repeatedly halving the sampling frequency.

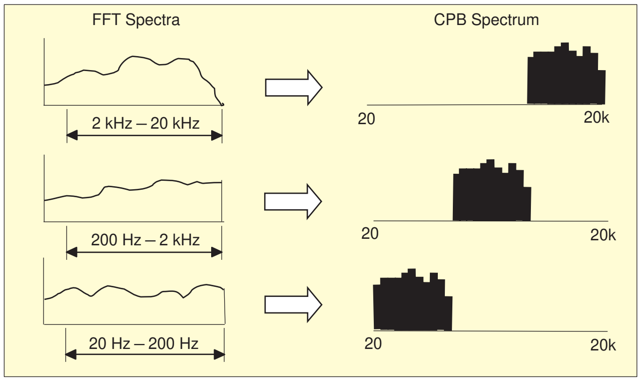

Figure 10. Conversion from FFT spectra to a CPB spectrum.

If the signal transformed is a transient (usually with no weighting), its spectrum should be scaled as an ESD (energy spectral density) in U^2s/Hz. Not only must the amplitude squared value be divided by Δf to give a spectral density, but also the power of the periodically repeated signal must be multiplied by T to give the energy in one period (i.e. the original signal). Altogether, this means a multiplication by T^2.

For stationary random signals, a single estimate of the spectrum is not sufficient, and an average over several estimates is necessary. The SD (standard deviation) of the error in a spectral estimate of a stationary random signal is given by the formula:

$$\varepsilon = \frac{1}{2\sqrt{BT}} \tag{6}$$

Since with the DFT, the bandwidth B is equal to the reciprocal of the time record length T, the product BT is unity for each transform, and Equation 6 can be replaced by:

$$\varepsilon = \frac{1}{2\sqrt{n}} \tag{7}$$

where n is the number of independent averages. For rectangular weighting, independent records mean nonoverlapping time records, but for windows such as Hanning, the windows can be overlapped by up to 50% and still give almost independent estimates.

Digital Filters. As has just been seen, the FFT provides a very efficient way of obtaining frequency spectra on a linear frequency scale with constant bandwidth, and this is most often advantageous for diagnostic purposes. However, for generating spectra with constant percentage bandwidth (i.e., $1/N^{\text{th}}$-octave) on a logarithmic frequency scale, as in Figure 1, digital filters give considerable advantage, in particular recursive IIR (infinite impulse response) filters.

Digital filters are similar to analog filters in that the output signal is convolved with the impulse response of the filter, and they operate directly in the time domain on continuous (though sampled) signals (as opposed to the blockwise treatment of the FFT process). The coefficients that define the filter properties give a characteristic that is defined in relation to the sampling frequency. Thus, 18 sets of filter coefficients will define the $1/18^{\text{th}}$ octave filters in one octave, but halving the sampling frequency will produce the equivalent filters one octave lower. Before halving the sampling frequency, the signal must be lowpass filtered by a filter that removes the upper octave of frequency information, but this can also be done by a digital filter with the same coefficients for every octave.

Figure 9 illustrates that when the sampling frequency is repeatedly halved for each octave, the total number of samples to be treated per unit time = M (1 + 1/2 + 1/4 + 1/8 + . . .) = 2 M samples/sec so that if the digital filter processor is capable of operating twice as fast as necessary for the highest octave, any number of lower octaves can be processed in real-time. This feature was mentioned in conjunction with the zoom processor of Figure 8.

If the digital filtering cannot be done in real-time, a very large data sample will have to be stored in advance. As an example, to produce the 1/18th octave filters of Figure 1 over three decades in frequency (frequency range 1000:1), each estimate of a spectrum value would have to encompass at least the impulse response time of the filter, approximately 30 periods of the center frequency for a 1/18th octave filter. For the lowest filter in the lowest octave there would have to be six samples per period, and since the sampling frequency would have to be decimated by a factor of 500 from the highest to the lowest decade, this corresponds to almost 100,000 samples in the original record. To achieve a result with only 10 averages would thus require on the order of 10^6 samples in the original record.

CPB spectra can also be obtained by conversion from FFT spectra, as illustrated Figure 10, where each decade is converted separately. The bandwidth of the individual lines in the original FFT spectra (including the effect of any window) must be less than the percentage bandwidth being converted at the lowest frequency in the FFT band. The conversion is achieved by calculating the lower and upper cutoff frequencies of each constant percentage band, and then integrating up the power in the FFT lines (and parts of lines) between the limits. The method indicated in Figure 10 gives a large change in filter characteristic at the junction between decades, but this is not likely to be such a problem with machine vibration analyses as with acoustic spectra.

To reduce the latter problem, some FFT analyzers do the conversion on an octave rather than a decade basis.

Parametric Spectrum Analysis. With Fourier analysis, the spectral resolution is of the order of 1/T, and thus the better the time localization the poorer the frequency localization, and vice versa. This is one expression of Heisenberg's uncertainty principle, and is because no assumption is made about the behavior of the time function outside the window (effectively it is set to zero, which is extremely improbable).

With parametric spectral analysis,[11] better spectral resolution can be obtained for short records, basically because it assumes that the behavior of the function outside the window is most similar to its behavior inside the window. This is valid for sinusoidal or near sinusoidal signals. With parametric analysis, the signal is modelled as the output of a physical system described by a limited number of parameters when excited by a unit white noise input. Thus the frequency response of the system represents the signal spectrum. Generally, the improvement in spectral resolution is accompanied by a deterioration in amplitude accuracy.

MA Models. Perhaps the easiest case to understand is where the system is modelled as an FIR (finite impulse response) filter, in which case the output is the (digital) convolution of the input signal with the finite length impulse response of the filter, as expressed by the equation:

$$y_i = \sum_{k=0}^{M} b_k x_{i-k} \tag{8}$$

where x_i represents the input signal, y_i represents the output signal, and b_k represents the convolution weights or samples of the impulse response. Equation 8 is a convolution equation or "moving average," giving rise to the term MA model. Applying a Z-transform to Equation 8, which is the equivalent of a Laplace transform for discrete time signals, the convolution becomes the product:

$$Y(z) = \sum_{k=0}^{M} b_k z^{-k} X(z) = B(z)X(z) \tag{9}$$

from which comes the transfer function:

$$B(z) = \sum_{k=0}^{M} b_k z^{-k} = b_0 \prod_{k=1}^{M} \left(1 - z^{-1} z_k\right) \tag{10}$$

which has no poles and is thus an "all-zero" model.

This type of model is obviously most efficient when the effective length of the impulse response is short, meaning that it is highly damped and thus without sharp spectral peaks.

To Be Continued. This concludes Part 1 of this article. Part 2 will appear in the May 2004 issue.

References

1. John S. Mitchell, *Machinery Analysis and Monitoring*, Penn Well, 1981.
2 E. Downham and R. Woods, ASME paper, Toronto, September 8-10, 1971.
3. P. Bradshaw and R. B. Randall, "Early Fault Detection and Diagnosis on the Trans Alaska Pipeline," MSA Session, ASME Conf., Dearborn, pp 7-17, 1983.
4. J. Howard Maxwell, "Induction Motor Magnetic Vibration," Proc. Vibration Institute, Meeting, Houston, TX, Apr.19-21, 1983.
5. R. B. Randall, "A New Method of Modeling Gear Faults," *ASME J. Mech. Design*, 104, pp 259-267, 1982.
6. J. Antoni and R. B. Randall, "Differential Diagnosis of Gear and Bearing Faults," *ASME J. Vib. & Acoustics*, 124, Apr. 2002, pp 165-171.
7. P. D. McFadden and J. D. Smith, "Model for the Vibration Produced by a Single Point Defect in a Rolling Element Bearing," *J. Sound Vib.*, 96 (1), pp 69-82, 1984.
8. D. Ho and R. B. Randall, "Optimisation of Bearing Diagnostic Techniques Using Simulated and Actual Bearing Fault Signals," *Mechanical Systems and Signal Processing*, 14 (5), September 2000, pp 763-788.
9. W. A. Gardner, *Introduction to Random Processes with Applications to Signals and Systems*, Macmillan, 1986.
10. R. B. Randall, "Frequency Analysis," Brüel & Kjær, Copenhagen, 1987.
11. S. Braun, and J. K. Hammond, "Parametric Methods," *Mechanical Signature Analysis*, (Editor S. Braun), Academic Press, London, 1986.

The author can be contacted at b.randall@unsw.edu.au.