A Stochastic Model for Simulation and Diagnostics of Rolling Element Bearings With Localized Faults

J. Antoni
Lecturer,
Roberval UMR CNRS 6066,
University of Technology of Compiègne,
France

R. B. Randall
Professor,
School of Mechanical and Manufacturing
Engineering,
The University of New South Wales,
Sydney 2052, Australia

This paper addresses the stochastic modeling of the vibration signal produced by localized faults in rolling element bearings and its use for diagnostic purposes. The aim is essentially to provide a better understanding of the recognized "envelope analysis" technique as classically used in the diagnostics of rolling element bearings, and incidentally give theoretical proofs for the specific features of envelope spectra as obtained from experimental data. The proposed model may also prove useful for simulation purposes. First, the excitation force generated by a defect is modeled as a random point process and its spectral signature is derived analytically. Then its transmission through the bearing is investigated in detail in order to find the spectral characteristics of the resulting vibration signal. The analysis finally gives sound justification for "squared" envelope analysis and the type of spectral indicators that should be used with it. [DOI: 10.1115/1.1569940]

1 Introduction

Most frequent faults in rolling element bearings include defects such as cracks, pits and spalls on the inner race, outer race, or rolling elements. Such defects are usually very localized during their early stages which is precisely when they need to be detected. As a result, intensive vibrations are produced by the repetitive impacts of the moving parts of the bearing on incipient defects. Numerous techniques have been proposed over the past three decades to diagnose rolling element bearing in the case of localized from the vibration signals they produce. *Inter alia*, the so-called ''envelope analysis'' or ''high frequency resonance'' technique is probably one of the most valuable and is currently well established in vibration based condition monitoring [1,2,3,4,5]. It is based on the idea that repetitive impacts on a defect excite some resonance—usually in a high frequency range where the signal-to-noise ratio is high—which acts as a modulation carrier. Demodulation of the vibration signal around the carrier then yields the envelope of the signal whose spectral content has been shown to be very relevant in identifying the location of the fault in the bearing, and the shaft on which the bearing is mounted. Since the early and heuristic foundations of the envelope analysis technique, many papers have tried to explain its actual virtues and unbeaten successes when applied to rolling element bearings. These efforts have first focused on proposing a proper model for the vibration signal generated by localized faults. It must be said that the objective of such a model is not to explain the physics of bearing failures but to describe its consequences as observed by the experimenter, i.e. it is phenomenological.

Most likely, the first valuable model for the vibration signal produced by a localized defect is due to McFadden & Smith [2,3]. Therein the repetitive impacts generated by a defect were modeled as a periodic train of Dirac delta functions with period T. Consequently the resonance characteristic in the Fourier domain was sampled at regular intervals $1/T$ (Fourier series). Moreover, McFadden's model had the advantage of explicitly including different sources of amplitude modulations (the radial load distribution, the moving location of impact forces) thus giving a good understanding of the spectral content of the envelope of the resulting vibration signal. This model was later refined by Ho & Randall who pointed out that actual rolling element bearings experience some random slip in their operation so that the train of impacts is slightly random instead of periodic [4]. However small these effects, Ho showed that the resonance characteristic is no longer sampled in the Fourier domain but rather resembles a continuous spectral density where all the harmonics tend to smear over each other. Ho's model resulted in a significantly better description of bearing vibration spectra as observed in the real world, and was next used by Randall, Antoni & Chobsaard to show that bearing signals are quasi-cyclostationary—i.e. their statistics have quasi-periodicity [5]. Incidentally, this observation offered an elegant way for justifying the envelope analysis method from the theory of cyclostationary processes. In a following paper, Antoni & Randall refined their results after specifying that signals from localized faults are not exactly quasi-cyclostationary since the random slips are non-stationary in their nature [6]. However they concluded that the bearing signals could still be treated as pseudo-cyclostationary as a first approximation.

The purpose of this paper is to attempt a complete treatment of the stochastic modeling of bearing vibrations as produced by localized faults, putting together a number of unpublished results and putting the heuristic considerations of [6] on a firmer mathematical foundation. The aim is twofold. Firstly, it is to provide the mechanical community with a model that has proven very satisfactory in describing actual vibration signals and, in particular, their spectra and envelope spectra—including some typical features that have never been explained elsewhere. Secondly, it is to demonstrate how this model permits a proper formalization of the envelope analysis technique as classically used in the diagnostics of rolling element bearings. The paper is organized as follows. A first section addresses the accurate modeling of the nonstationary impacting process as generated by a localized defect on the inner race, the outer race, or on a rolling element. The spectral characteristics of this process are then derived from the theory of regular point processes and important results are deduced concerning the nature of spectral harmonics. In a second section these results are used to investigate the spectral properties of the resulting vibration signal after the impacts have propagated through the system, i.e. as measured on the housing. In particular, the general spectral signature due to a localized defect is found and its mani-

Contributed by the Technical Committee on Vibration and Sound for publication in the JOURNAL OF VIBRATION AND ACOUSTICS. Manuscript received April 2002; revised January 2003. Associate Editor: M. I. Friswell.

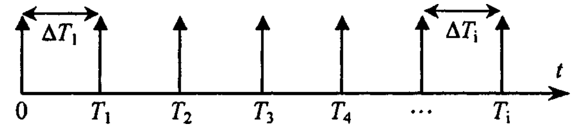

Fig. 1 The impacting process viewed as a point process

festations in a number of spectral indicators (the Fourier transform, the power spectral density, the spectral correlation density, the Fourier transform of the squared signal and the envelope spectrum) are investigated in detail. The relative effectiveness of these spectral indicators in diagnostics is finally discussed in the light of the new results.

2 Modeling the Impacting Process

2.1 Regular Point Process. At the outset, consider the process generated by the repetition of impact forces when a defect in one surface strikes a mating surface. We shall refer to it as the *impacting process* $F(t)$. For a localized defect, each impact may be well described by a Dirac delta function $\delta(t)$ provided the measured signal is sampled at a rate well below the impact spectral bandwidth [2,3,4]. At this stage it is assumed that all impacts have equal magnitudes; magnitudes and signs of impacts will be accounted for later in the text by modulating the impacting process with a suitable time-varying function.

Without loss of generality, the reference time $t=0$ is chosen to coincide with an arbitrary impact which defines the point from which the process is starting to be observed. Hence,

$$F(t)=\sum_{i=0}^{\infty}\delta(t-T_i)\quad\text{where}\quad T_0=0 \tag{1}$$

The stochastic process $\{T_i\}$ governing the arrival of the impacts can be defined in a variety of ways. However it was argued in [6] that an adequate assumption for rolling element bearings is where the inter-arrival times $\Delta T_i=T_i-T_{i-1}$ are independent and identically distributed random variables (see Fig. 1). In turn, this can be shown to define a *stationary Markov process* $\{T_i\}$, that it is to say in which each arrival is only influenced by its immediate predecessor and irrespectively of its index:

$$P\{T_i\leq t_i/T_{i-j}=t_{i-j},j=1,\ldots,i\}=P\{T_i\leq t_i/T_{i-1}=t_{i-1}\}$$
$$=P\{T_1\leq t_1/T_0=0\} \tag{2}$$

It can easily be checked that under these conditions the arrival time process $\{T_i\}$ has a stationary mean $E\{T_i\}=i\cdot T$ but a *nonstationary covariance function* $Cov\{T_i,T_j\}=\sigma_\Delta^2\cdot\min(i,j)$ where $\sigma_\Delta^2=Var\{\Delta T_i\}$. It is specifically this non-stationarity that has not properly been recognized before and actually explains distinctive features of the vibration signal ensuing from a faulty bearing [6].

Now return to Eq. (1) and define $\{N(t)\}$ as the number of impacts that have occurred in the interval $[0,t]$, so that $\{dN(t)\}$ denotes the number of impacts in the infinitesimal interval $[t,t+dt]$. Hence, Eq. (1) simplifies to $F(t)=dN(t)/dt$ which defines an ordinary point process (ordinary means one in which the initial impact occurs at zero time) [7]. For the physical case of interest, it is a sound assumption that the probability of occurrence of an impact in *dt* is proportional to *dt* while the probability of more than one occurrence is negligibly smaller than (*dt*). This property of *regularity* ensures the use of the product density technique to obtain the moments of $F(t)$ [8]. Specifically, we define $f_1(t)$ and $f_2(t,\tau)$ as the *product densities of degree one and two* which may be interpreted as the instantaneous mean rate of impacts respectively at time *t* and at time *t* plus τ. Then,

$$E\{dN(t)^n\}=f_1(t)dt,\quad\forall n>0\ \text{in}\ \mathbb{N} \tag{3}$$

$$E\{dN(t+\tau)dN(t)\}=f_2(t,\tau)dtd\tau,\quad\tau>0 \tag{4}$$

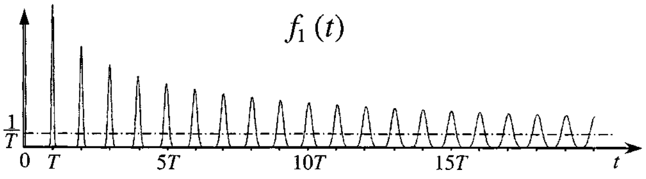

Fig. 2 Product density of degree one for $\sigma_\Delta/T=1/30$

Note the degeneracy occurring at $\tau=0$, where $f_2(t,0)=f_1(t)\delta(\tau)$. By using the terminology of stochastic point processes, explicit solutions will now be found for the first two moments of the impacting process.

2.2 Analytical Forms of the Product Densities. The product density of degree one in Eq. (3) can be expanded into

$$f_1(t)dt=\sum_{i=0}^{\infty}P\{t\leq T_i\leq t+dt/T_0=0\}=\sum_{i=0}^{\infty}\phi_i(t)dt \tag{5}$$

where $\phi_1(t)$ is the probability density function of the ith impact conditioned to the fact that the zeroth occurred at $t=0$ and $\phi_0(t)=\delta(t)$.

Similarly, the product density of degree two is

$$f_2(t,\tau)dtd\tau=\sum_{i=0}^{\infty}\sum_{j>i}^{\infty}P\{t\leq T_i\leq t+dt,t+\tau\leq T_j\leq t+\tau+d\tau/T_0$$
$$=0\} \tag{6}$$

for $\tau>0$. Remembering that the arrival time $\{T_i\}$ is a stationary Markov process,

$$f_2(t,\tau)dtd\tau=\sum_{i=0}^{\infty}\phi_i(t)dt\sum_{k=1}^{\infty}\phi_k(\tau)d\tau,\quad\tau>0 \tag{7}$$

and finally, adding the degeneracy case $f_2(t,0)=f_1(t)\delta(\tau)$ arising when $\tau=0$, one gets the simple expression,

$$f_2(t,\tau)dtd\tau=f_1(t)f_1(\tau)dtd\tau,\quad\tau\geqslant 0 \tag{8}$$

In short Eqs. (5) and (8) give the explicit solutions for first two moments of the impacting process from which those of the vibration signal will later be derived. For the physical process under consideration, it is noteworthy that the product density of degree one f_1 suffices to describe it at least up to the second order (because f_2 factorizes into a product of f_1 terms), thus assigning to the instantaneous mean rate of impacts a major role in this paper.

As a matter of fact, the exact shape of $f_1(t)$ is worthy of further investigation. In view of Eq. (5), the first peak in $f_1(t)$ happens to be the probability density function $\phi_1(t)$ of the first time of occurrence T_1, the second peak the probability density function of T_2 and so on. Therefore the ith peak is the first one convolved with itself i times, i.e.

$$\phi_i(t)=\underbrace{\phi_1(t)*\phi_1(t)*\cdots*\phi_1(t)}_{i\text{-}times} \tag{9}$$

Then, under mild conditions, the bandwidth of the ith peak as measured by its standard deviation is $\sqrt{i}\cdot\sigma_\Delta$ with σ_Δ the standard deviation of $\phi_1(t)$. As the peaks slowly enlarge, their amplitudes decrease accordingly so as to maintain a unit area. This is illustrated in Fig. 2. In the limit, the peaks completely vanish and $f_1(t)$ tends to the constant value $f_1(\infty)=1/T$, that is the mean overall rate of occurrence.[1] However the rate of convergence is extremely small: considering a percentage of random fluctuation of $x/100=\sigma_\Delta/T$, then two peaks completely overlap when their band-

[1] An heuristic proof to this result is that, in the limit, the area under each probability density function is still unity whereas the mean paving is one probability density function per *T* units of time.

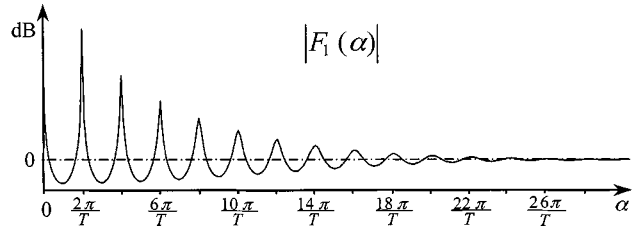

Fig. 3 Fourier transform (modulus) of the product density of degree one for $\sigma_\Delta/T=1/30$

width is greater than their mutual spacing, that is when $\sqrt{i}\cdot\sigma_\Delta > T$ or $i > 10000/x^2$. For random fluctuations of typically a few percent, this means it takes a few thousands peaks for $f_1(t)$ to reach its limit.

Similarly, the product density of degree two can be verified to tend to $f_2(\infty,\tau)=f_1(\tau)/T$. This supports the assertion of reference [6] where it was pointed out that the nonstationary impacting process $F(t)$ ultimately tends to stationarity, yet so slowly that the phenomenon is hardly noticeable in practice over a finite time of observation.

2.3 Spectral Characteristics. Since most processing involved in the diagnosis of rolling element bearings is performed in the Fourier domain, it is now necessary to derive the formulas for the Fourier transforms of the product densities $f_1(t)$ and $f_2(t,\tau)$.

(a) Fourier Transform of the Product Density of Degree One

Combining Eq. (5) and Eq. (9), the Fourier transform of $f_1(t)$ is readily found to yield a continuous density,

$$F_1(\alpha)=\frac{1}{2\pi}\int_{\mathbb{R}} f_1(t)e^{-j\alpha\cdot t}dt=[1-\Phi^*(\alpha)]^{-1} \tag{10}$$

where $\Phi(\alpha)$ is the characteristic function of the first time of arrival T_1—or equivalently of the independent and identically distributed inter-arrival time process $\{\Delta T_i\}$. Equation (10) is known as a "renewal type" equation in the theory of stochastic processes, the study of which requires the exact knowledge of the probability law governing T_1. The Gamma law would probably be a good candidate here as it produces strictly positive inter-arrival times with a peaked probability around the mean value T. However when its variance σ_Δ^2 is small w.r.t. its mean T, the Gamma distribution is well approximated by the Normal distribution with the same mean and variance, thus making the calculations more tractable. Under these assumptions,

$$\Phi(\alpha)\approx\exp\left(-\frac{1}{2}\sigma_\Delta^2\cdot\alpha^2-j\alpha\cdot t\right) \tag{11}$$

from which Eq. (10) is readily found to yield a pole at $\alpha=0$ and a series of finite-energy peaks equi-spaced by $1/T$, with maxima and minima respectively on $\alpha=k/T$ and $\alpha=k/T-1/2$, $k\in\mathbb{Z}$. Figure 3 depicts the behavior of the modulus $|F_1(\alpha)|$ where the percentage of random fluctuation σ_Δ/T has been set rather large for sake of demonstration. In contrast to the time domain, note firstly (*i*) that the magnitude of the peaks falls off more rapidly in the Fourier domain and secondly, (*ii*) that the bandwidth of the successive peaks of $F_1(\alpha)$ remain more or less constant since they are bounded by $1/T$.

(i) Fall off of the peaks. For $\sigma_\Delta^2\ll T$, the relative magnitude of the ith peak w.r.t. the first one decreases almost as fast as $1/i^2$, that is a slope of -40 dB per decade. Since $F_1(\alpha)$ ultimately tends towards a constant amplitude density $F_1(\infty)=1$, this means that there exists a cut-off radian frequency $\alpha_c\approx\sqrt{2}/\sigma_\Delta$ after which all the peaks have faded. Considering the percentage of random fluctuation $x/100=\sigma_\Delta/T$, it is found that $F_1(\alpha)$ becomes almost constant after $i_c=22.5/x$. For example, for a random fluctuation of 2%, this means as few as $i_c=11$ peaks.

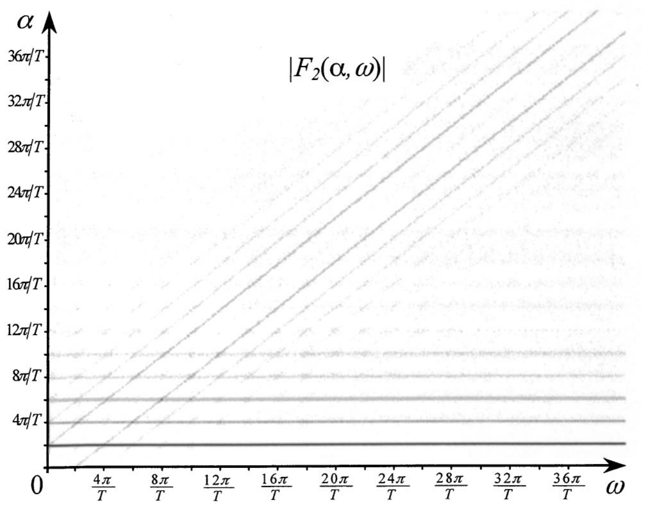

Fig. 4 Double Fourier transform (modulus) of the product density of degree two for $\sigma_\Delta/T=1/30$

(ii) Quality factor. The "apparent increase" of the bandwidth resulting from the fall off of the peaks may be quantified by the quality factor $Q_i\approx T^2/(i\cdot\pi\cdot\sigma_\Delta)^2$ (ratio of the maximum to the minimum of the ith peak). This is naturally found to strongly depend on the percentage of random fluctuation σ_Δ/T.

(b) Double Fourier Transform of the Product Density of Degree Two

The formula for $F_1(\alpha)$ can now be used to compute the double Fourier transform of $f_2(t,\tau)$:

$$F_2(\alpha,\omega)=\frac{1}{4\pi^2}\int\int_{\mathbb{R}^2} f_2(t,\tau)e^{-j\alpha\cdot t}e^{-j\omega\cdot\tau}dt\,d\tau \tag{12}$$

Distinguishing three cases $\tau<0$, $\tau=0$ and $\tau>0$ and after some algebra, one finds

$$F_2(\alpha,\omega)=F_1(\alpha)\cdot[F_1(\alpha)+F_1(\alpha-\omega)-1] \tag{13}$$

This defines a continuous spectral density with marked ridges running along the ω-variable and centered on all $\alpha=k/T$, $k\in\mathbb{Z}$. For large values of ω, $F_2(\alpha,\omega)$ ultimately tends to $F_1(\alpha)$ and thus resembles a pattern of parallel and horizontal ridges along the ω-axis rapidly falling off on each side of $\alpha=0$, as illustrated in Fig. 4. The presence of these parallel ridges distinctively characterizes the (second-order) spectral signature of a random train of impact forces and consequently, that of a faulty rolling element bearing. Moreover, the distance between the ridges indicates the mean rate of occurrence of the fault, thus enabling its identification in the mechanical system.

2.4 Discussion. At this stage, it is instructive to review some former models proposed in the literature in light of the derived results. Clearly, for the deterministic model $f_1(t)$ is a perfectly periodic train of Dirac deltas $\text{III}_T(t)$ and $f_2(t,\tau)$ a two-dimensional version of it, viz $f_2(t,\tau)=\text{III}_T(t)\cdot\text{III}_T(\tau)$. The same applies to their respective Fourier transforms, viz $F_1(\alpha)=1/T\cdot\text{III}_{1/T}(\alpha)$ and $F_2(\alpha,\omega)=1/T^2\cdot\text{III}_{1/T}(\alpha)\cdot\text{III}_{1/T}(\omega)$. The limitation of these formulas arises from the experimental evidence that actual data do not have *line spectra* especially in the vicinity of the high frequency resonance where they are usually demodulated. On the other hand, the simplified stochastic model proposed by Randall & Antoni in [5] leads to $f_1(t)=\text{III}_T(t)*\phi_1(t)$, that is a periodic train of Dirac delta functions low-pass filtered by the probability function $\phi_1(t)$. Therein $f_2(t,\tau)$ turns out to be periodic and low-pass filtered in the t-variable while transient in the τ-variable. These "low-pass filter" and "transient" effects give a better explanation for the *continuity* of experimental spectra in the vicinity of a high frequency resonance [4]. The refined stochastic model proposed herein obviously leads to almost identical properties, yet on the basis of more accurate physical considerations. Of particular concern are the facts that $f_1(t)$ and $f_2(t,\tau)$ are no longer periodic functions neither in the t nor in the τ variables and that the ''low-pass filter'' effect is now replaced by a rapid ''fall-off'' effect. In other words, the Fourier transforms $F_1(\alpha)$ and $F_2(\alpha,\omega)$ are now purely continuous functions in both α and ω (except at $\alpha=0$) where all the former discrete lines (harmonics) have been replaced by *distributed peaks*, gradually broadening with increase in α. Actually, this fact is always observable in envelope spectra, and was one of the main reasons for modifying the stochastic model first mentioned in reference [5].

3 Spectral Statistics of the Vibration Signal

In the preceding section, expressions were derived which describe the spectral signature of a train of pulses as produced by a localized defect. This section now discusses how this spectral signature is transformed after the impacts have propagated through the system, i.e. as it is likely to be measured on the bearing housing by an accelerometer.

3.1 Response of a Rolling Element Bearing to a Random Train of Impacts. Following classical models, the vibration signal produced by a faulty rolling element bearing may be viewed as the response of a linear system driven by the impacting process $F(t)$ [2,7]. For this input-output relationship to be fully comprehensive, we now show that the impulse response of the system should be time-varying and should also accommodate some degree of stochasticity.

At the outset, the impacting process should be modulated by a periodic and positive function $A(t)$ to account for the variations in the impact magnitudes as the defect enters and exits the load zone [2]. Some random modulation might be incorporated in $A(t)$ due to the dependence on the position and the number of the rolling elements in the load zone at time t, but also due to rolling and slip on possibly rough surfaces especially after a defect has appeared and spread to some extent.

Next, let us define $r(t,\tau)$ the structural response at time t of the system subjected to an impulse $\delta(\tau)$ at time τ. In contrast to a static structure, the impulse response $r(t,\tau)$ of a rolling element bearing is time-varying for a variety of physical reasons, the most obvious of which being the variations in the transmission path as the coordinates of the point of impact move w.r.t the location of the sensor, and the variations in the relative angle between the impact forces and the axis of the sensor. For a system operating at constant speed, these variations periodically affect the magnitude, the sign and the phase of the impulse response $r(t,\tau)$ with a period depending on whether the defect lies on the inner race, the outer race or on a rolling element [2]. In addition, $r(t,\tau)$ might have some small random fluctuations to account for unpredictable effects such as contact non-linearities. Figure 5(a) gives a schematic illustration of how the impacting process $A(t)F(t)$ is transformed into a vibration $X(t)$ after passing through the impulse response $r(t,\tau)$.

In practice, the vibration signal $X(t)$ produced by a local fault cannot be observed totally because it is contaminated by other vibrations from a multitude of neighboring sources in the system. Therefore, it is customary to filter it in a frequency band where the signal-to-noise ratio is maximum so that virtually no other sources than that stemming from the faulty bearing are measured by the experimenter. This is usually done by designing a band-pass filter $b(t)$ around a high-frequency resonance of the structure (or the sensor) that is excited by the impacts [1–4]. In order to retain the diagnostic information, the band-pass filter $b(t)$ must have the following properties:

P1: *$b(t)$ is a band-pass filter with central frequency ω_0 much higher than the mean rate of impacts $1/T$,*

P2: *$b(t)$ has an effective duration shorter than the mean inter-arrival time T, or equivalently its spectral bandwidth is larger than the mean rate of impacts $1/T$.*

Therefore, the overall impulse response of the system is obtained from cascading the amplitude modulation function $A(t)$ with the time-varying impulse response $r(t,\tau)$ and finally with the band-pass filter $b(t)$. This is illustrated in Fig. 5(b).

In this procedure, $A(t)$ and $g(t,\tau)=b(t-\lambda)r(\lambda,\tau)$ have some important properties which will make the computation of the input-output relationship tractable. Specifically, because $A(t)$ encompasses all the periodic modulations with possible stochastic effects, it has first and second-order statistics given by:

$$m_A(t)=E\{A(t)\}=m_A\left(t+\frac{2\pi}{\Omega}\right)=\sum_{k\in\mathbb{Z}} a_k e^{j\Omega\cdot t} \tag{14}$$

and

$$R_A(t,\tau)=E\{A(t+\tau)A^*(t)\}=R_A\left(t+\frac{2\pi}{\Omega},\tau\right)=\sum_{k\in\mathbb{Z}} R_A^k(\tau)e^{j\Omega\cdot t} \tag{15}$$

Equations (14) and (15) define a *second-order cyclostationary* process, i.e. a stochastic process with periodic mean and autocorrelation function of intrinsic period $2\pi/\Omega$. Therein Ω is either equal to the speed of the inner race, that of the outer race or that of the cage (relative to the load vector) whether the fault is on the inner race, the outer race, or on a rolling element.

Similarly, $g(t,\tau)$ being a periodic causal Green's function describing the periodically varying transmission path, it expands into:

$$g(t,\tau)=\begin{cases} g\left(t+\dfrac{2\pi}{\Omega},\tau\right)=\sum_{k\in\mathbb{Z}} g_k(t-\tau)e^{j\Omega\cdot t}, & \tau\leqslant t \\ 0 & \tau>t \end{cases} \tag{16}$$

From the above expansion, the mechanism relating the impacting process $F(t)$ to the band-pass vibration signal $Y(t)$ can finally be obtained from the following Stieltjes stochastic integral

$$Y(t)=\int_0^t g(t,\tau)A(\tau)dN(\tau)=\sum_{k\in\mathbb{Z}} e^{j\Omega\cdot t}\int_{\mathbb{R}} g_k(t-\tau)A(\tau)dN(\tau) \tag{17}$$

in which each Fourier coefficient $g_k(t)$ is to be interpreted as a linear, causal and homogeneous impulse response.

3.2 Spectral Characteristics of the Vibration Response. From Eq. (17), the spectral characteristics of the band-pass vibration signal can now be derived and applied to a number of potential indicators for use in diagnostics, namely the Fourier transform of the expected signal, the power spectral density, the spectral correlation, the Fourier transform of the expected squared signal and the power spectral density of the squared signal.

(a) Fourier Transform of the Expected Response

From Eqs. (1), (14) and (17), the expected value (ensemble average) of the vibration signal is

$$m_Y(t)=E\{Y(t)\}=\int_0^t g(t,\tau)m_A(\tau)f_1(\tau)d\tau \tag{18}$$

from which the Fourier transform is found to be:

$$\begin{cases} M_Y(\alpha)=\dfrac{1}{2\pi}\int_{\mathbb{R}} m_Y(t)e^{-j\alpha\cdot t}dt=\sum_{k\in\mathbb{Z}} M_{Y_k}(\alpha-k\Omega) \\ M_{Y_k}(\alpha)=G_k(\alpha)\cdot\tilde{F}_1(\alpha) \\ \tilde{F}_1(\alpha)=M_A(\alpha)*F_1(\alpha)=\sum_{l\in\mathbb{Z}} a_l\cdot F(\alpha-l\Omega) \end{cases} \tag{19}$$

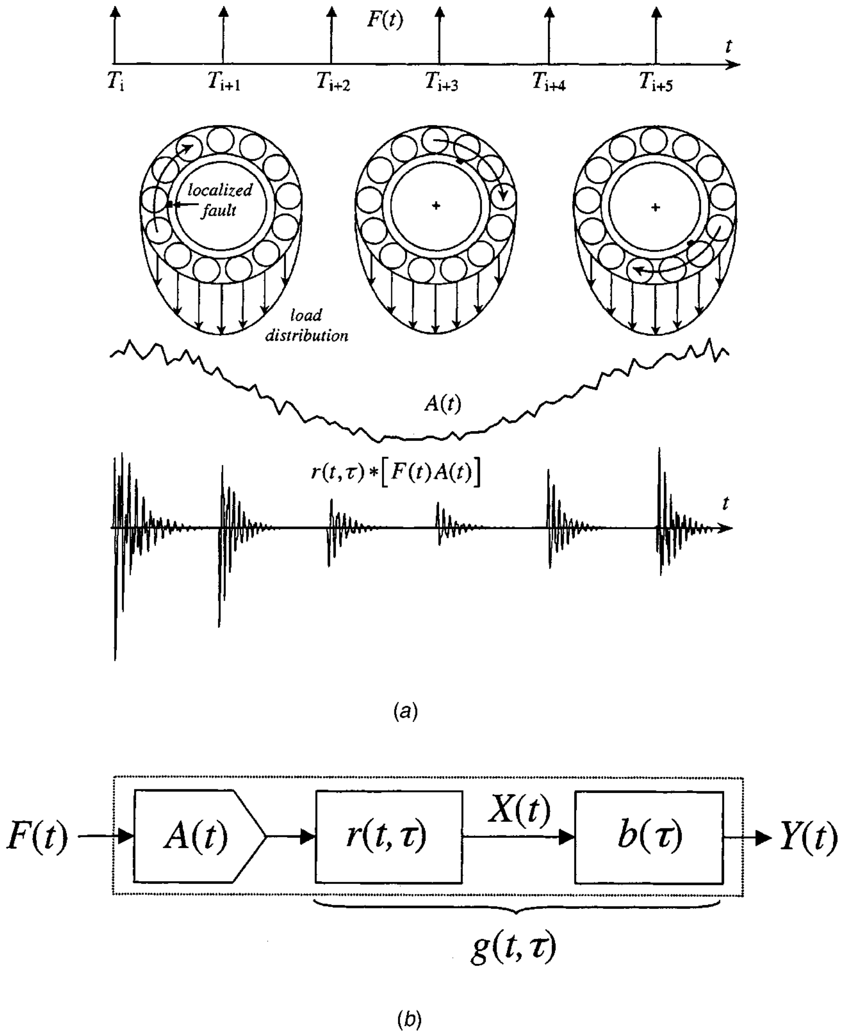

Fig. 5 (*a*) Generation of the resulting vibration signal. *A*(*t*): magnitude of the impacts; *r*(*t*,τ): time-varying (stochastic) structural impulse response at time *t*. (*b*) Scheme of the overall impulse response. *b*(τ) is a band-pass filter that extracts the bearing signal where its signal-to-noise ratio is the highest

where $G_k(\alpha)$ and $M_A(\alpha)$ are respectively the Fourier transforms of $g_k(t)$ and $m_A(t)$. In view of Eq. (19), $M_Y(\alpha)$ is a superposition of shifted functions $M_{Y_k}(\alpha)$; each of them being in turn constructed from shifted and scaled replicas of $F_1(\alpha)$ and then weighted by the frequency response $G_k(\alpha)$. The construction of $M_Y(\alpha)$ results in a *mixed* spectrum with a family of infinite-energy pseudo-harmonics around $\alpha=0$, all equi-spaced by the rotation speed Ω. The number of these pseudo-harmonics directly depends on the number of Fourier coefficients in $g(t,\tau)$ and $m_A(t)$. Note that this specific pattern repeats around all the peaks of $F_1(\alpha)$ at $\alpha=k/T$, $k\in\mathbb{Z}$, but with finite-energy peaks in place of pseudo-harmonics for any $k\neq 0$. Obviously, this makes the very distinctive "*spectral signature*" of a localized defect as it is expected to appear in a faulty rolling element bearing (see Fig. 6). Most importantly its detection forms the main basis of diagnostics since it contains the key characteristic frequencies $1/T$ and Ω that enable the *identification* and the *localization* of faults in complex systems.

However, the problem in the spectral indicator of Eq. (19) is that the frequency support of the band-filters $G_k(\alpha)$ is very likely to be higher than the frequency support of the spectral signature $F_1(\alpha)$ as shown schematically in Fig. 7. In fact, it was already pointed out that $F_1(\alpha)$ falls off by -40 dB per decade down to a cut-off radian frequency α_c of about $\sqrt{2}/\sigma_\Delta$. In order for $G_k(\alpha)$ to overlap with this support, its central radian frequency ω_0 (resonance frequency chosen for demodulation) should be such that $\omega_0\cdot T<100\sqrt{2}/x$ with $x/100=\sigma_\Delta/T$ the percentage of fluctuation. Or equivalently, with i_c the number of peaks in $F_1(\alpha)$ before it dies to 1, ω_0 should be such that $\omega_0\cdot T<2\pi\cdot i_c$. In most instances this condition would not be satisfied if a good signal-to-noise ratio were to be maintained, thus justifying the poor performance anticipated from the Fourier transform of the vibration signal.

(b) Power Spectral Density of the Response

From Eqs. (4), (15) and (17), the autocorrelation function of the vibration response is:

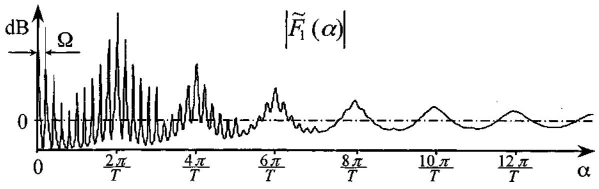

Fig. 6 Typical spectral signature in the vibration signal for $\sigma_\Delta/T=1/30$ and $\Omega=T/10$

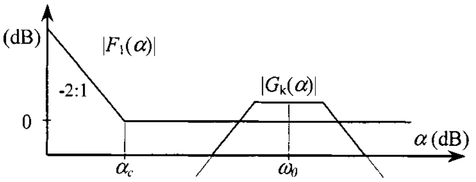

Fig. 7 Illustration of the low-pass filter effect

$$R_Y(t,\tau)=E\{Y(t+\tau)Y^*(t)\}=\int_0^t\int_0^{t+\tau-\nu} h(t+\tau,\nu+\lambda)h^*(t,\nu)R_A(\nu,\lambda)f_2(\nu,\lambda)d\lambda d\nu \tag{20}$$

This is a bivariate function since the vibration signal resulting from the impacting process $F(t)$ is nonstationary. In order to compute the power spectral density, let us first denote by $\overline{R}_y(\tau)$ the "stationarized" autocorrelation function

$$\overline{R}_Y(\tau)=\lim_{W\to\infty}\frac{1}{W}\int_0^W R_y(t,\tau)dt \tag{21}$$

whose Fourier transform then yields the explicit expression for the power spectral density:

$$\begin{cases} S_Y(\omega)=\dfrac{1}{2\pi}\displaystyle\int_{\mathbb{R}}\overline{R}_Y(\tau)e^{-j\omega\cdot\tau}d\tau=\sum_{k\in\mathbb{Z}}S_{y_k}(\omega-k\Omega) \\ S_{Y_k}(\omega)=|G_k(\omega)|^2\cdot\tilde{F}_2(\omega) \\ \tilde{F}_2(\omega)=\dfrac{2}{T}\mathrm{Re}\{F_1(\omega)\}*S_A(\omega) \end{cases} \tag{22}$$

where $S_A(\omega)$ is the Fourier transform of the stationarized version $\overline{R}_A(\tau)$ of $R_A(t,\tau)$ in Eq. (15). The set of Eq. (22) indicate that the principle of construction of $S_Y(\omega)$ is similar to that outlined for the Fourier transform in Eq. (19) because $S_A(\omega)$ contains the same discrete harmonics as $F_A(\alpha)$ and $|G_k(\omega)|^2$ obviously acts in the same frequency band as $G_k(\alpha)$. Therefore, the same conclusion holds in regard to the expected performance of the power spectral density as a diagnostic indicator.

(c) Spectral Correlation Density of the Response

We now demonstrate that the aforementioned shortcomings due to the non-intersection of the low-pass and band-pass filters (see. Fig. 7) can be solved by considering the double Fourier transform of the autocorrelation function $R_Y(t,\tau)$. This yields a quantity called the spectral correlation density[2] [5], very similar to the "generalized spectrum"—within a simple change of variable—used by Lin [9]. The spectral correlation density,

$$S_Y(\alpha,\omega)=\frac{1}{4\pi^2}\int_{\mathbb{R}}R_Y(t,\tau)e^{-j\omega\cdot\tau}e^{-j\alpha\cdot t}d\tau dt \tag{23}$$

is found to have explicit expression

$$\begin{cases} S_Y(\alpha,\omega)=\displaystyle\sum_{k,l\in\mathbb{Z}^2}S_{y_ky_l}(\alpha-k\Omega,\omega-l\Omega) \\ S_{Y_kY_l}(\alpha,\omega)=G_k(\omega)G_l^*(\omega-\alpha)\cdot\tilde{F}_2(\alpha,\omega) \\ \tilde{F}_2(\alpha,\omega)=\displaystyle\sum_{p\in\mathbb{Z}}F_2(\alpha-p\Omega,\omega)*S_A^p(\omega) \end{cases} \tag{24}$$

where $S_A^p(\omega)$ is the Fourier transform of $R_A^p(\tau)$ in Eq. (15). Although involving two frequency variables, the construction of $S_Y(\alpha,\omega)$ is again similar to that outlined in Eqs. (19) and (22). Nevertheless, there is now a domain in the frequency plane (α, ω) where the diagnostic information is totally preserved. Specifically, for small values of α within twice the bandwidth of $g(t,\tau)$ and large values of ω around the resonance frequency ω_0, $S_Y(\alpha,\omega)$ clearly displays the spectral signature of a defect because the band-pass filters $G_k(\omega)$ and $G_l(\omega-\alpha)$ are band-passing in this area. This is a direct consequence of property ***P2***, a schematic illustration of it being shown in Fig. 8. Note that in the domain of concern $S_Y(\alpha,\omega)$ is markedly ridged in the horizontal ω-direction just as $F_2(\alpha,\omega)$ was in Fig. 4.

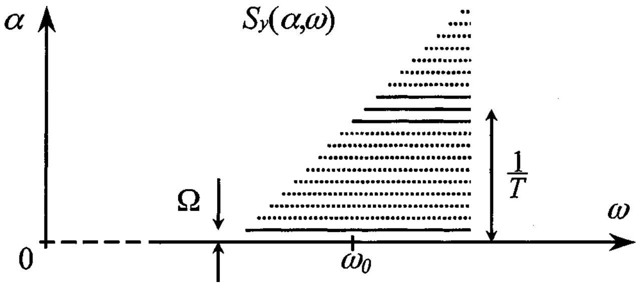

Fig. 8 Scheme of the spectral correlation density

(d) Fourier Transform of the Expected Squared Response

Because it is bivariate, the spectral correlation density may be difficult to compute and therefore it has been suggested to replace it by its integrated version over the ω-variable (while preserving the diagnostic information). In reference [5], this was shown to be equivalent to the Fourier transform of the expected squared signal, i.e.

$$M_{Y^2}(\alpha)=\int_{\mathbb{R}}S_Y(\alpha,\omega)d\omega=\frac{1}{2\pi}\int_{\mathbb{R}}E|Y(t)|^2e^{-j\alpha\cdot t}dt \tag{25}$$

This equation is easily found to be identical in structure to Eq. (19) where the coefficients a_1 are replaced by $R_A^l(0)$ defined in Eq. (15), and where the band-pass filter $G_k(\alpha)$ is replaced by the low-pass filter $P_k(\alpha)=\Sigma_pG_k(\alpha)\cdot G_{k-p}(\alpha)$. The fact that $P_k(\alpha)$ is now necessarily a low-pass filter comes from the convolution of $G_k(\alpha)$ by itself and this is exactly the reason why $M_{Y^2}(\alpha)$ can preserve the diagnostic information whereas $M_Y(\alpha)$ cannot. Indeed, under property ***P2*** the support of $P_k(\alpha)$ necessarily overlaps with that of the spectral signature of the fault, contrary to the scheme of Fig. 7.

A last point to consider is whether to take the square of the raw signal in Eq. (25) or the squared magnitude of its *analytic* version. Strictly speaking, the analytic signal should be used so that the expectation of its squared magnitude truly gives the *squared envelope*. However, minor differences would be found when using the real signal provided it is properly band-pass filtered around a resonance. This point was also addressed in a lot of detail in reference [4].

(e) Power Spectral Density of the Squared Response

In light of the previous demonstration, one can expect the power spectral density $S_{Y^2}(\omega)$ of the squared signal to perform just as well as $M_{Y^2}(\alpha)$. Indeed, $S_{Y^2}(\omega)$ is the exact definition of the "spectrum of the squared envelope" as was proposed in [4]. In order to prove this result, let us invoke property ***P1*** in conjunction with the assumption that the point process $\{dN(t)\}$ is regular (see section 2.1). Then, the following approximation holds

$$|Y(t)|^2\approx\int_0^t p(t,\tau)|A(\tau)|^2dN(\tau) \tag{26}$$

with $p(t,\tau)=|g(t,\tau)|^2$. After taking the Fourier transform of the "stationarized" autocorrelation function of $|Y(t)|^2$, the envelope spectrum is found to be identical to Eq. (22) where $S_A(\omega)$ is replaced by $S_{A^2}(\omega)$—the Fourier transform of the stationarized

[2] There is a simple relation between the spectral correlation density and the power spectral density, viz $S_Y(0,\omega)=S_Y(\omega)\delta(\alpha)$

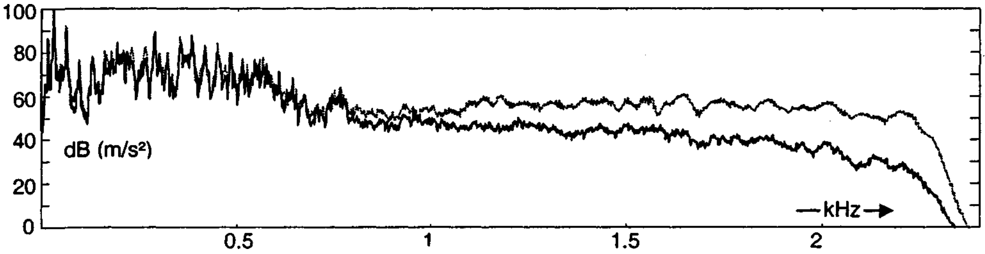

Fig. 9 Power spectral density of a vibration signal in case of no fault (continuous line) and an inner race fault (dotted line)

autocorrelation function $\overline{R}_{A^2}(\tau)$ of the squared process $\{|A(t)|^2\}$—and where $|G_k(\omega)|^2$ is replaced by $|P_k(\omega)|^2$. The fact that $|P_k(\alpha)|^2$ acts as a low-pass filter demonstrates again that $S_{Y^2}(\omega)$ is also a usable diagnostic indicator. Here again, the analytic version of the signal may be preferred in Eq. (26) in order to estimate the power spectral density of the true squared envelope.

3.3 Discussion. It has been proven in some depth why the Fourier transform and the power spectral density generally are poor indicators for diagnosing rolling element bearings in the case of localized faults, a fact that the authors have regularly observed on experimental data. Indeed, even though classical spectral analysis may perform very well in detecting a fault—e.g. through monitoring the relative energy levels in some frequency bands—it rarely helps in recognizing its type nor its location—and this is exactly what diagnostics asks for. For example, Figure 9 compares the power spectral densities of a vibration signal measured on a gearbox before and after one of the rolling element bearings (12 balls, ∅7.12 mm, pitch circle ∅38.5 mm) was purposely damaged by machining a small slot on its inner race. The frequency resolution is 12 Hz. Note that the presence of the fault only shows up at high frequencies. The fact that there is no difference at low frequencies is due to the extremely poor signal-to-noise ratio in that band (observe that most of the sources there relate to harmonics from the gears). Of interest also is the fact that in spite of its increase, the spectral density at higher frequencies is continuous and therefore gives no indication of a fault producing repetitive impacts.

In clear contrast with the Fourier transform and the power spectral density, the same transformations applied on the *squared* signal (or its analytic version) have been shown to solve the problem in a surprisingly simple manner. For example, Fig. 10 displays the power spectrum of the squared magnitude of the analytic signal after band-pass filtering in the frequency band [1.8; 2.2] kHz with a frequency resolution of 2 Hz. Now the specific spectral signature of the inner race fault shows up in good accordance with Fig. 4, with $1/T = 71$ Hz (ball pass frequency on the fault) and $\Omega = 10$ Hz (shaft rotation).

From a theoretical point of view, it is worth mentioning that the Fourier transform of the squared signal preserves the diagnostic information by exploiting the non-stationarity of the signal, while the power spectral density exploits its non-Gaussianity. In fact, the power spectral density of the squared signal is implicitly a fourth-order "stationarized" statistic. Strictly speaking, these two indicators have different theoretical justifications and this is supported by recalling that the former only requires property ***P2***, while the latter requires the more stringent condition ***P2***+***P1***. However, both are inclined to provide envelope analysis—or "squared" envelope analysis—with a strong formal justification.

Comparison of the five spectral indicators, which have been assessed so forth, is summarized in Table 1.

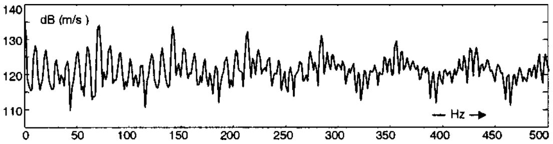

Fig. 10 Power spectral density of the squared envelope

Table 1 Comparison of five spectral indicators in terms of their ability of detecting and identifying localized faults.

Spectral indicators	Diagnostic skills
Fourier transform of expected signal	−
Power spectrum of the signal	−
Spectral correlation density (2-D) of the signal	++
Fourier transform of the expected squared signal	++
Power spectrum of the squared signal	++

4 Conclusion

A comprehensive stochastic model has been proposed for describing and simulating the vibration produced by localized faults in rolling element bearings. Sources of stochasticity were modeled in both the impacting force process—by means of a regular point process—and in the transmission path—by means of a cyclostationary process, thus encompassing a large range of physical situations. These refinements proved very valuable in explaining some of the actual features observed on experimental data. The spectral signature of a localized fault was derived analytically and new results were deduced concerning the nature of spectral harmonics produced by the impacting process. These were shown to be *distributed* and equi-spaced (by the mean rate of impacts) peaks with a *rapid fall-off* that could be quantified as a function of the percentage of stochastic fluctuations. Next, the spectral signature of a defect was shown to duplicate when it propagates through the structure (with shifts equal to the rotation speed of the defect), thus generating additional families of pseudo-harmonics. These results finally helped in investigating the effectiveness of a number of spectral indicators dedicated to the diagnostics of rolling element bearings. From simple considerations on band-pass and low-pass filtering operations, it was demonstrated that both the Fourier transform and the power spectral density of the *squared signal* are the most relevant indicators, thus bringing new supports in favor of "squared" envelope analysis.

References

[1] Darlow, M. S., and Badgley, R. H., 1975, "Applications for Early Detection of Rolling Element Bearing Failures Using the High-Frequency Resonance Technique," ASME Paper 75-DET-46.

[2] McFadden, P. D., and Smith, J. D., 1984, "Model for the Vibration Produced by a Single Point Defect in a Rolling Element Bearing," J. Sound Vib., **91**(1), pp. 69–82.

[3] McFadden, P. D., and Smith, J. D., 1985, "The Vibration Produced by Multiple Point Defects in a Rolling Element Bearing," J. Sound Vib., **98**(2), pp. 69–82.

[4] Ho, D., and Randall, R. B., 2000, ''Optimization of Bearing Diagnostics Techniques Using Simulated and Actual Bearing Fault Signals,'' Mech. Syst. Signal Process., **14**(5), pp. 763–788.

[5] Randall, R. B., Antoni, J., and Chobsaard, S., 2001, ''The Relationship Between Spectral Correlation and Envelope Analysis in the Diagnostics of Bearing Faults and other Cyclostationary Machine Signals,'' Mech. Syst. Signal Process., **15**(5), pp. 945–962.

[6] Antoni, J., and Randall, R. B., 2002, ''Differential Diagnosis of Gear and Bearing Faults,'' ASME J. Vibr. Acoust., **127**, pp. 1–7.

[7] Roberts, J. B., 1966, ''On the Response of a Simple Oscillator to Random Impulses,'' J. Sound Vib., **4**(1), pp. 51–61.

[8] Srinivasan, S. K., et al., 1967, ''Response of Linear Vibratory Systems to Non-Stationary Stochastic Impulses,'' J. Sound Vib., **6**(2), pp. 169–179.

[9] Lin, Y. K., 1965, ''Nonstationary Excitation and Response in Linear Systems Treated as Sequences of Random Pulses,'' J. Acoust. Soc. Am., pp. 453–460.