Mechanical Systems and Signal Processing 133 (2019) 106199

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

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

Uses and mis-uses of energy operators for machine diagnostics

R.B. Randall *, W.A. Smith

School of Mechanical and Manufacturing Engineering, University of New South Wales, Sydney 2052, Australia

ARTICLE INFO

Article history:
Received 10 April 2019
Accepted 13 June 2019

Keywords:
Teager Kaiser Energy Operator
Frequency domain energy operator
Frequency weighted energy operator
Gear diagnostics
Bearing diagnostics
Speed determination
Amplitude demodulation
Frequency demodulation

ABSTRACT

The Teager Kaiser Energy Operator (TKEO) was originally proposed for use in speech analysis as representing the total energy (i.e. kinetic plus potential energy) in a signal. It was shown that for a mono-component carrier, with slowly changing amplitude and frequency, the TKEO is approximately equal to the product of the squares of the instantaneous amplitude and frequency. The TKEO is only strictly defined for mono-components, i.e. signals that can be modelled as a single carrier frequency, modulated in amplitude and frequency in such a way that they can be represented as the real part of an analytic signal, with a one-sided spectrum.

The traditional way of estimating the TKEO was by an efficient time domain operation involving only three adjacent samples, and this can be done in real time, but this implies that all filtering and other processing must use causal processing to retain this advantage. However, causal filters give phase distortion and non-ideal filter characteristics. It is easily shown that the TKEO is approximately equal to the squared envelope of the derivative of the signal, which can alternatively be calculated by efficient non-causal Hilbert transform techniques via the frequency domain, incidentally giving a more accurate result, as well as being virtually as efficient. When combined with other non-causal processing, such as ideal filtering by choice of a specified band in the frequency domain, and ideal differentiation/integration by $j\omega$ operations in the frequency domain, this approach has many advantages in cases where real-time processing is not required, and where the processing can be carried out by post-processing of recorded signals, which can be very long.

Machine diagnostics is one area where real-time processing gives no advantage, and even numerous disadvantages, which accompany causal processing, such as mentioned above. Even in the single situation in machine monitoring where a result might be required rapidly, online monitoring of critical equipment, there is little practical difference in the processing time for causal and non-causal techniques (a maximum of a second or so) as this would rarely be sufficient time to make a decision on whether to shut a machine down, or for its speed to reduce significantly even if it were. The disadvantage of non-causal (batch) processing via Fourier transforms comes from the intrinsic circularity of the latter, where all functions in both time and frequency domains are assumed periodic. However, this has been dealt with since the birth of the FFT algorithm in 1965, and usually means that time records (or spectra) just have to be extended a small amount to allow truncation of wraparound effects.

There are already a considerable number of papers published recommending the use of the TKEO and its variants for machine diagnostics, many claiming that this gives advantages over traditional approaches, for example of amplitude and frequency demodulation based on Hilbert transforms. However, this paper demonstrates that the claimed advantages are invariably false, for the following reasons:

* Corresponding author.
E-mail address: b.randall@unsw.edu.au (R.B. Randall).

https://doi.org/10.1016/j.ymssp.2019.06.017

1) The formulas derived for estimating the instantaneous amplitude and frequency of a mono-component using the TKEO actually give the values for the derivative of the signal, which are not the same. It is true that the time domain TKEO gives better results for a single chirp sweeping over a wide frequency range from zero (because of huge wraparound effects) but this situation does not apply to machine signals because of interference between multiple harmonics of shaft speeds, meaning that the maximum speed range in one record is 2:1.
2) By employing Hilbert transform and non-causal processing techniques, the errors and excessive time of causal time domain processing (for example time domain convolutional filtering, differentiation, etc) are avoided and virtually all other parameters and features of machine faults are estimated more accurately and faster, with considerably more control of frequency bandwidth and waveforms. Multiple differentiations can be achieved with equal accuracy in one operation.
3) Many of the proposed applications of the TKEO do not require the signal to be mono-component, such as the application to bearing diagnostics (since bearing signals do not have continuous phase) and where the only advantage of differentiation (increasing weighting with frequency) can only be realised with a frequency range much greater than the maximum 2:1 limit for a mono-component.

This paper demonstrates all the above claims with a range of typical signals and applications to gear and bearing diagnostics, and rebuts many of the false claims previously made.

1. Introduction

In [1], Kaiser formalised an "energy operator", first proposed by Teager for use in speech analysis, as well as representing it as analogous to the total energy of a spring/mass system, including both kinetic and potential energy, which continuously alternate in an oscillatory system. It has become known as the Teager Kaiser Energy Operator (TKEO). It was shown that in a discretised version it could be very efficiently estimated from three adjacent samples, effectively in real-time, as appropriate to speech analysis. In [2,3] Maragos et al. made extensive further developments, including the use of the TKEO for amplitude and phase demodulation, and presenting error estimates for both the continuous and discretised versions.

A number of authors have proposed using the TKEO for machine diagnostics, e.g. [4–9], some even claiming that it gave better results for amplitude and frequency demodulation than Hilbert transform techniques. The main aim of this paper is to show that where estimations do not have to be made in real-time, the most efficient (and accurate) way to estimate an energy operator equivalent to the TKEO, is by using Hilbert transform techniques via the frequency domain, as it allows non-causal, zero phase shift, signal processing operations to be used. This advantage is most in evidence in cases where an instantaneous frequency is being estimated by frequency demodulation. It has long been known that the TKEO is approximately equal to the product of the instantaneous squared amplitude and squared frequency of a signal, but usually not appreciated that this only corresponds to total energy when the vibration signal is in terms of displacement (so that the square of its derivative, velocity, represents kinetic energy (KE)). It turns out that the TKEO is actually an approximation of the squared envelope of the derivative of a signal, so if the signal is velocity, its squared envelope directly represents total energy, and if the measured signal is acceleration (most often the case) the TKEO represents the squared envelope of the jerk. Independent of the type of signal measured, the squared frequency is approximately the ratio of the squared envelope of its derivative to the squared envelope of the signal, but this cannot be directly derived from the TKEO which only gives the numerator. It can, however, be estimated using the above-mentioned Hilbert transform techniques via the frequency domain, and in fact as shown in this paper, the most accurate estimate (of the frequency itself, rather than its square) can be obtained using a very similar operation, with the same computational effort.

Machine diagnostics is a typical case where real-time operation is not required, as information is usually being sought days, weeks or months in advance of when the condition may become serious, and certainly the few seconds' delay involved in non-real-time processing is immaterial, even for online monitoring of critical machines. Another aim of the paper is to show that where information about instantaneous frequency is not specifically required, for both bearing and gear diagnostics, there can be advantages in relaxing the requirement of a mono-component, to gain some benefit from the differentiation over a wider frequency range than the 2:1 range imposed by the normal situation with machine vibration signals, as explained below. The disadvantage of using non-causal post processing FFT techniques is the wraparound errors associated with the circularity of the FFT algorithm, but transform sizes can be made very large in post-processing, and the small affected sections at the ends can usually be discarded.

2. Formulations and equations

The TKEO is defined in both continuous and discretised forms [1–3], as given in Eqs. (1) and (2), respectively.

$$\Psi_c(x(t)) = [\dot{x}(t)]^2 - x(t)\ddot{x}(t) \tag{1}$$

$$\Psi_d(x[n]) = (x[n])^2 - x[n+1]x[n-1] \tag{2}$$

Returning to Kaiser's mass/spring oscillator, with slowly varying amplitude and frequency (so that $\dot{A}(t)$ and $\dot{\omega}(t)$ can be ignored), if

$$x(t) = A(t)\cos\phi(t), \quad \text{where } \omega(t) = \dot{\phi}(t) \tag{3}$$

then

$$\dot{x}(t) = -\dot{\phi}(t)A(t)\sin\phi(t) + \dot{A}(t)\cos\phi(t) \approx -\omega(t)A(t)\sin\phi(t) \tag{4}$$

$$\ddot{x}(t) \approx -\dot{\omega}(t)A(t)\sin\phi(t) - \dot{A}(t)\omega(t)\sin\phi(t) - \omega(t)A(t)\cos\phi(t)\dot{\phi}(t) \approx -[\omega(t)]^2A(t)\cos\phi(t) \tag{5}$$

and

$$\Psi_c[x(t)] = [\dot{x}(t)]^2 - x(t)\ddot{x}(t) \approx [\omega(t)]^2[A(t)]^2\left(\sin^2\phi(t) + \cos^2\phi(t)\right) = [\omega(t)]^2[A(t)]^2 \tag{6}$$

which is the (approximate) squared envelope of $\dot{x}(t)$, in other words the sum of the squares of $\dot{x}(t)$ and its Hilbert transform $\omega(t)A(t)\cos\phi(t)$, which is the parameter whose square is proportional to the instantaneous PE. The approximations made in the above derivation are the same as those made in the time domain version of the TKEO, as developed in detail by Maragos et al. in [2,3], where a complete error analysis is made, for both the continuous and discretised versions. It includes the cases of mono-components representing pure amplitude modulation, pure frequency modulation, and combined amplitude and frequency modulation.

Note that the TKEO only gives physical energy when $x(t)$ is a displacement signal. In the general case, it will be an approximation to the squared envelope of the derivative of a signal, so in the common case where $x(t)$ is acceleration it will be the squared envelope of the jerk (derivative of the acceleration).

If Envsq$[x(t)]$ is the squared envelope of $x(t)$, i.e.

$$\text{Envsq}[x(t)] = x^2(t) + \tilde{x}^2(t) \tag{7}$$

where $\tilde{x}(t)$ is the Hilbert transform (HT) of $x(t)$, then

$$\Psi_c[x(t)] \approx \text{Envsq}[\dot{x}(t)] \tag{8}$$

and since

$$[A(t)]^2 = \text{Envsq}[x(t)] \tag{9}$$

then from (6)

$$[\omega(t)]^2 \approx \frac{\text{Envsq}[\dot{x}(t)]}{\text{Envsq}[x(t)]} \tag{10}$$

3. Comparison of time domain and frequency domain methods

3.1. Amplitude and frequency demodulation

A generally modulated mono-component carrier can be represented as

$$x(t) = A(t)\cos(2\pi f_c t + \phi_m(t)) \tag{11}$$

where $A(t)$ is the instantaneous amplitude, and $\phi_m(t)$ is the phase variation (rad) around the linearly increasing phase of the carrier frequency f_c. The corresponding frequency modulation signal $f_m(t)$ is the derivative of the phase modulation, converted to Hz, and is thus $\frac{1}{2\pi}\frac{d}{dt}(\phi_m(t))$. $x(t)$ is seen to be the real part of the rotating vector

$$x_a(t) = A(t)\exp j(2\pi f_c t + \phi_m(t)) = A(t)[\cos(2\pi f_c t + \phi_m(t)) + j\sin(2\pi f_c t + \phi_m(t))] \tag{12}$$

In [3] formulas are developed for achieving amplitude and frequency demodulation, using the TKEO, and are given here as Eqs. (13) and (14), respectively.

$$A(t) \approx \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \tag{13}$$

$$\omega(t) \approx \sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \tag{14}$$

Eq. (14), for example, is similar to Eq. (10), since:

$$\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]} = \frac{\text{Envsq}[\ddot{x}(t)]}{\text{Envsq}[\dot{x}(t)]} = \frac{\text{Envsq}[\dot{y}(t)]}{\text{Envsq}[y(t)]} \ , \ \text{where } y(t) = \dot{x}(t) \tag{15}$$

Thus, the frequency obtained using Eq. (14) is the instantaneous frequency (IF) of $y(t)$, the derivative of the original signal $x(t)$, rather than the IF of the signal itself. As will be shown below, in general these are not the same.

3.2. Hilbert transform methods

A signal such as $x_a(t)$, with a 1-sided spectrum (frequency varying in a range around $+f_c$), is complex, and is known as an analytic signal. Its imaginary part can be shown to be the Hilbert transform of its real part [10]. A very efficient way to perform a Hilbert transform is thus to transform to the frequency domain, double the amplitude of positive frequency components, set negative frequency components to zero, and transform back to the analytic signal $(x(t) + j\,\tilde{x}(t))$ in the time domain [10]. The analytic signal of Eq. (12) is a product of two terms so its spectrum will be the convolution of the Fourier transforms of these two terms, with a bandwidth less than the sum of the two bandwidths. The spectrum will then be 1-sided if there is no overlap of modulation sidebands over zero frequency, i.e. the sum of the half bandwidths of the amplitude and phase modulation terms is less than the carrier frequency. The bandwidth of the amplitude modulation term is the same as that of $A(t)$, but that of the phase modulation term is theoretically infinite. However, as discussed in [11], only a certain number of sidebands are significant, and the effective bandwidth of the phase/frequency modulation component can be determined from the maximum frequency deviation and rate of change of frequency.

The HT $\tilde{x}(t)$ of $x(t)$ can alternatively be calculated directly in the time domain by the following convolution operation:

$$\tilde{x}(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau} d\tau = x(t) * \frac{1}{\pi t} \tag{16}$$

although in general this will be more computationally expensive. In practice, because of the singularity at $t = 0$, numerical computation is somewhat complicated, but efficient programs have been developed to produce FIR filter based Hilbert transformer impulse responses with exact phase response but some amplitude distortion.

One of the first papers to recommend use of the TKEO for machine diagnostic problems [4] based this recommendation partly on a comparison of the TKEO methods with Hilbert transform methods for the particular case of determining the instantaneous frequency of a chirp signal with constant amplitude but frequency varying linearly between zero and 300 Hz. Fig. 1 (from [4]) illustrates this.

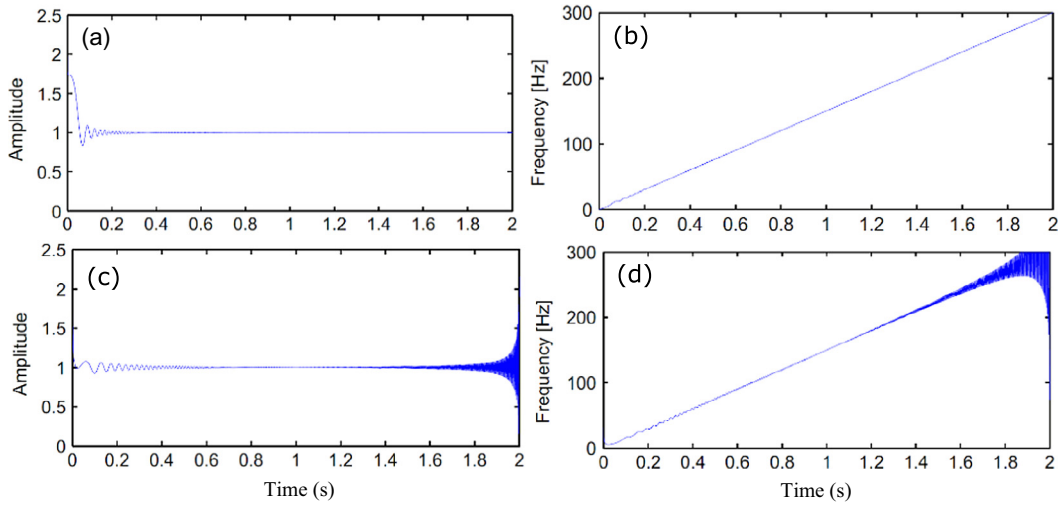

Fig. 1. Comparison (from [4]) of amplitude and instantaneous frequency estimates using TKEO (Eq. (2)), and HT (frequency domain) techniques (a, b) TKEO (c, d) Hilbert (a, c) Amplitude (b, d) Frequency.

However, in [12] this comparison is taken up in detail, and it is shown that the difference is mainly between time domain and frequency domain execution, and in particular only applies for a narrow class of problems which is not of much interest for machine diagnostics.

Fig. 2 (from [12]) shows the result of applying the HT in the time domain. This was done using the Matlab® function FIRPM in mode "Hilbert" (Order M = 140). It is seen that the amplitude characteristic is if anything slightly better than the TKEO (Fig. 1(a)) at low frequency, whereas the phase characteristic has somewhat bigger end effects, though much less than the frequency domain HT version. Note that the problem at zero frequency can be explained by the fact that a zero frequency carrier has no possibility to incorporate left-hand sidebands, these always being necessary for both amplitude and frequency modulation.

On the other hand, it is pointed out in [12] that such a single component chirp, sweeping from zero, virtually never occurs in machine vibration signals, since most important potential carrier frequencies for mono-components, such as shaft speeds, gearmesh frequencies etc., have multiple harmonics. It is shown in [11] that the maximum speed range which can be accommodated in one record and still allow a mono-component to be separated by filtration, is 2:1, before the sidebands around the second harmonic start to overlap with those around the first. Another important point is that speed sweeps from zero are rarely of interest in any case, because machine vibrations are usually negligible below a speed of, say, 5% of full speed, being completely stiffness controlled and extremely low level in terms of velocity or acceleration.

Fig. 3, also from [12], shows that if the frequency sweep range of the chirp is <2:1 (in this case 160–300 Hz), the end effects for the HT frequency domain implementation are much smaller. They are basically due to the wraparound effects of the circular FFT transform used in this case. In a later example, it will be shown that even this can normally be reduced in extent by appropriate smoothing, so the disadvantage of frequency domain implementation lies only in these limited end effects, while considerable benefits arise from the non-causal signal processing, avoiding most phase distortion. This will directly be the case if a mono-component has to be isolated by a bandpass filter, which was the first approach used in speech analysis, for example, to apply the TKEO to multi-component signals [13], to split them into a sum of mono-components. To retain the advantage of real-time operation, which is the primary advantage of the TKEO, such filters have to be causal convolutive time domain filters, which give amplitude and phase distortion in the vicinity of any cut-off frequencies. Later approaches (e.g. [14]) attempt to separate multi-components with overlapping frequency ranges, but still use causal real-time processing.

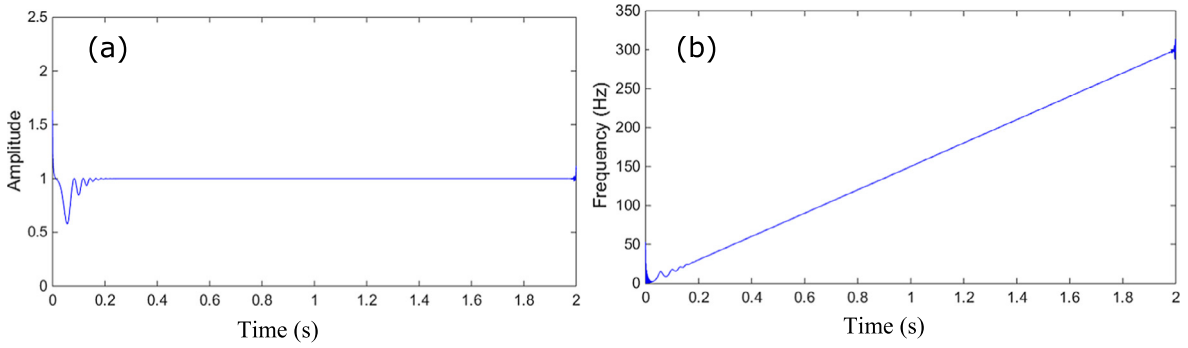

Fig. 2. Estimated amplitude and frequency of the chirp, applying HT in the time domain [12] (a) Amplitude (b) Frequency.

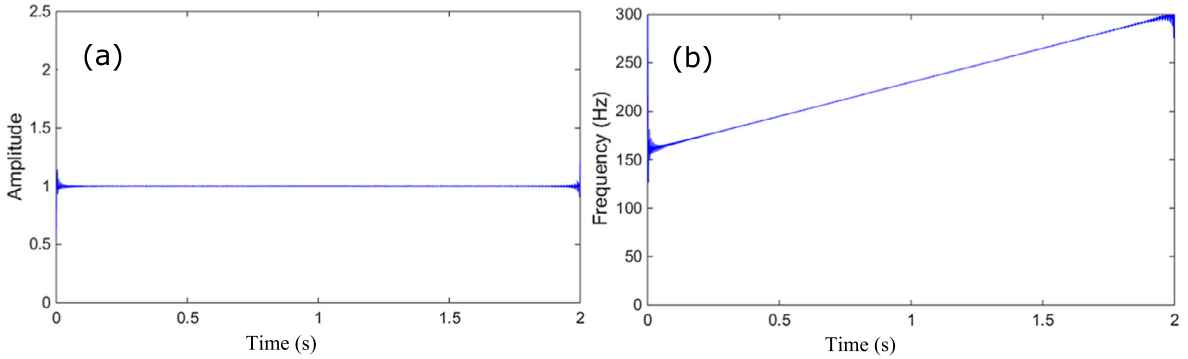

Fig. 3. Amplitude and instantaneous frequency estimates for the 160–300 Hz chirp using the HT (frequency domain) technique [12].

A potentially viable method of separating slowly varying multi-component signals, which overlap in frequency over longer records, is to use the Hilbert-Huang transform (HHT) in conjunction with empirical mode decomposition (EMD) [15] or its variants. This might for example apply to gear vibration signals with widely varying speeds, but where the individual mono-components, such as gearmesh frequencies, always maintain the same relative spacing, even if they do overlap over a long period. However, EMD cannot be done in real-time, and is always a post-processing method. In theory, this might allow longer records to be processed than the HT methods recommended in this paper (where a long record might have to be divided into shorter sections, each with speed range <2:1), but EMD suffers much more than the HT methods from end effects, with an unknown number of intrinsic mode functions (IMFs) required purely to compensate for end effects, and thus with no physical meaning. This is a completely separate problem from that of mode mixing, usually due to additive noise, which can also lead to incorrect decomposition into mono-components, even though a number of techniques such as EEMD have been developed to reduce mode mixing [16]. It should be noted that a considerable amount of computational effort is devoted to ensuring that the resulting mono-components have continuous phase, so it is very inefficient to use this decomposition in cases where the instantaneous phase and frequency of the mono-components are not to be used, just the amplitude or envelope signals. This is even more the case when the signals cannot be validly decomposed into mono-components, with continuous phase, as for the majority of bearing signals, as will be shown in this paper. In any case, as this paper shows, if EMD is used, there is no point in combining it with TKEO, whose only advantage is its real-time capability, and in fact the Hilbert-Huang transform normally uses frequency domain HT methods.

Thus, the main purpose of this paper is to show that there can be considerable advantage in replacing the time domain TKEO with its approximate equivalent (i.e. the squared envelope of the derivative of a signal) using non-causal techniques applied in the frequency domain (with truncation of end effects) for the vast majority of applications in machine diagnostics. The full advantage of this approach was not realised at the time of publishing Ref. [12], so it was referred to as "a TKEO equivalent via the frequency domain". However, it is now suggested that this should be changed to "frequency domain energy operator", or FDEO, which gives particular advantages in consideration of the fact that Eq. (10) can be used directly, without having to resort to Eqs. (13) and (14) as required by the TKEO. Not only can the band to be demodulated be isolated by an ideal filter in the (1-sided) frequency spectrum corresponding to an analytic signal, it can also be differentiated "exactly" by multiplication by $j\omega$ over this same frequency band. The spectra of the original and differentiated signals are then inverse transformed to their analytic signals, which can be operated on using Eq. (10) (or alternatively using Eq. (17), outlined in Section 4.3.2), giving improved instantaneous frequency estimates.

On doing some research it was realised that this relationship had been recognised earlier, leading to the definition of the "Frequency Weighted Energy Operator", or FWEO [17], though we believe it is something of a misnomer, since the frequency weighting is a direct result of the differentiation, which is an intrinsic part of the definition of the TKEO. In any case, it was not suggested in [17] that it should be estimated via the frequency domain, and they proposed use of an "envelope-derivative operator", still enacted in the time domain. They do remark that this is less "real-time" than the TKEO, but that their application to EEG signals does not require real-time processing. The FWEO is used in Ref. [9] for bearing diagnostics, and is therefore discussed in Section 4.2.

4. Applications in machine diagnostics

4.1. Gear diagnostics

It has already been mentioned that Ref. [4] applied the TKEO as an alternative means of performing amplitude and frequency demodulation for gear diagnostics. Despite demonstrating that the TKEO gave a better result for demodulation of a single carrier chirp (Fig. 1) the actual results applied to the diagnostics of a wind turbine gearbox (at constant speed) were virtually indistinguishable, and this can be explained by the fact that for constant speed, the end effects of the frequency domain HT method are very small, but so are the frequency errors in the time domain TKEO (see Section 4.3).

Ref. [5] applied the TKEO to gear diagnostics, but without regard to the requirement of it being a mono-component carrier. On the contrary, it was applied to the time synchronous average (TSA) of a signal encompassing several harmonics of the gearmesh frequency. Fig. 4 shows some results from [5] where the use of the TKEO appears to increase the kurtosis (impulsiveness) in the case of a fault. On the other hand, there is a much greater spread of the results, and even some data sets where the values of the healthy and faulty cases were not differentiated. It is possible that this is due to the amplification of high frequency "noise" even where random noise as such should have been removed by the TSA operation.

This is a case where it seems that the mono-component requirement of the TKEO can be dispensed with, as the phase information was not used, just the waveform of the multi-component signal. This is to be discussed in more detail below, as it is equally relevant to bearing diagnostics.

In contrast to [5], Ref. [6] does make the claim that both the amplitude and frequency components (Eq. (6)) of the TKEO are important in gear diagnostics. This is based on the type of gear fault which gives high frequency impulse responses (IRs) which cannot be captured in the TSA signal for either gear, partly because their high (constant) resonance frequency is independent of shaft speed, so that even small speed variations mean that they are not phase-locked to either gear. This type of problem was discussed in [18]. Because this type of signal is very similar to those from rolling element bearing faults, and because a very similar analysis is made by the same authors in [7], this question is taken up in detail in the next section.

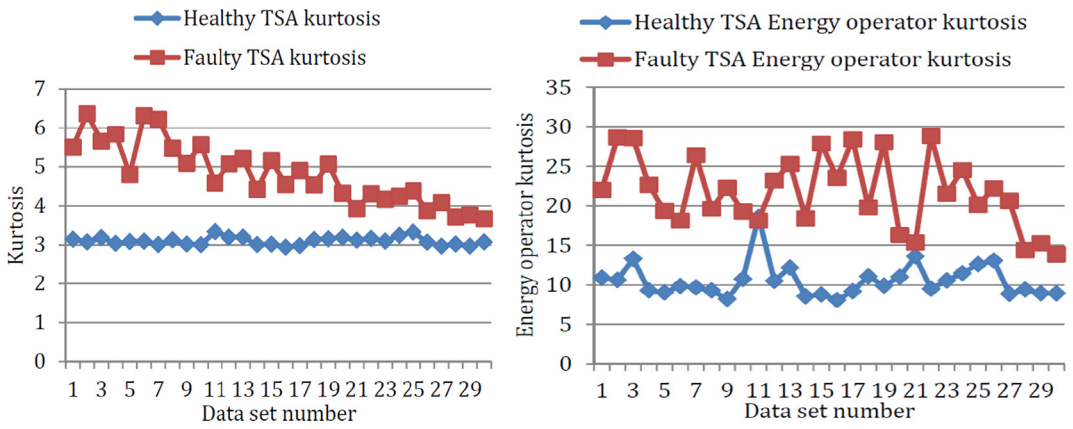

Fig. 4. Comparison of gear diagnostic indicators, for healthy and faulty data sets, with and without TKEO (from [5]).

4.2. Bearing diagnostics

The most comprehensive paper on the use of the TKEO for bearing diagnostics is [7]. It makes the claim that both the amplitude modulation and frequency modulation effects of bearing faults are detected by the TKEO, and therefore it gives benefits over the conventional diagnostic methods involving only amplitude demodulation (envelope analysis). However, just the fact that the TKEO (squared envelope of the derivative) for a displacement signal is (approximately) equal to the direct squared envelope of the velocity signal negates this interpretation. The squared frequency term in Eq. (6) only occurs because the signal is differentiated, and this applies independent of the type of original signal, as demonstrated above. Any benefit coming from the TKEO can be attributed to the differentiation of the signal, giving a weighting of the spectrum proportional to frequency, before it is demodulated for squared envelope analysis.

However, Ref. [7] does present an argument to justify the claim about the additional benefit of the instantaneous frequency term, and this must be considered in detail (essentially the same argument is presented in [6]). The paper does present an excellent analysis of a hypothetical (outer race) bearing fault model, which is basically the same as that illustrated in Fig. 5. This assumes that the series of impulse responses (IRs) which result from contact between the fault and the various bearing elements can be modelled as a single frequency (the resonance frequency of a single response mode) amplitude modulated by a series of exponentially decaying envelopes, with the amplitude steps corresponding to the initiating of each new IR. As the resonance frequency and repetition frequency of the IRs will not in general be commensurate, the phase has to update to zero at the beginning of each new IR, giving a series of phase jumps at the repetition frequency. In principle, since the instantaneous frequency is obtained by differentiation of the phase, the steps should transform to impulses, and [6,7] claim that these will also add into the TKEO. It is indeed claimed that because the phase signal is independent of the signal amplitude, it will be even more sensitive in cases where the bearing signal is weak.

There is an error in the way the model is implemented in Refs. [6,7], in that it is stated that the "settling time" of the exponentially decaying sinusoidal IRs is shorter than the spacing between them, and that after this settling time they become "negligible", and both the amplitude and phase can be set to zero. This is not the case in fact, in that the frequency of an exponentially decaying sinusoid remains constant to infinity, and "negligible" actually means that the amplitude falls below that of whatever noise is in the signal. Fig. 6 illustrates this for a series of uniformly spaced IRs, with the ratio of resonance frequency to repetition frequency arranged to give the maximum phase jump of 180° at the transition where each

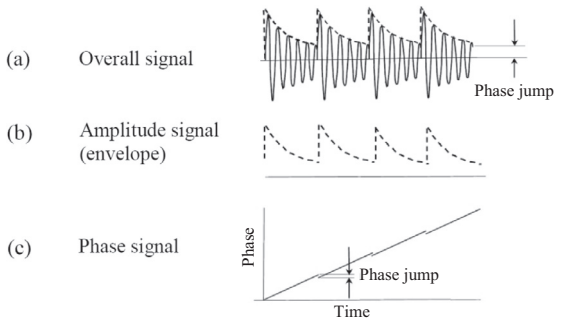

Fig. 5. Modelling a bearing fault signal as amplitude and phase modulation of a single resonance frequency.

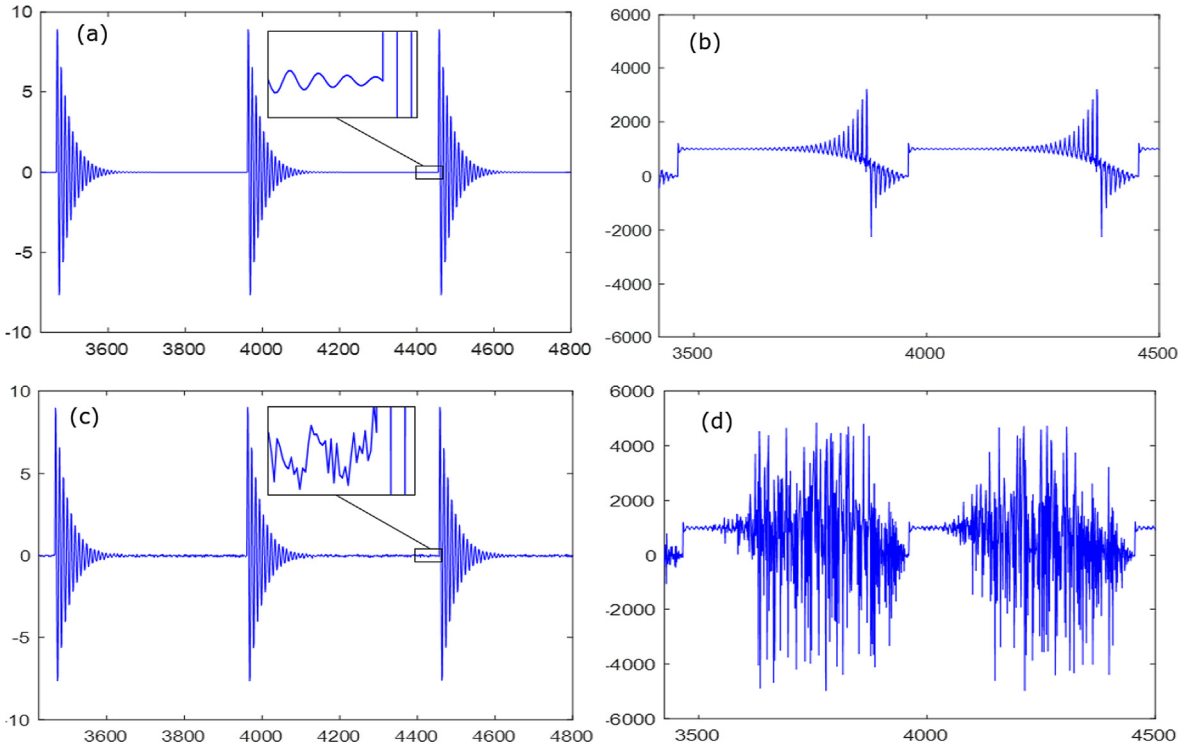

Fig. 6. Simulated bearing signal with and without added noise (a, b) without noise (c, d) with noise (a, c) time signal (inset, transition at new IR) (b, d) estimated instantaneous frequency.

new IR occurs. In Fig. 6(a), with no added noise, it can be seen in the inset, that this phase jump does occur. However, when the instantaneous phase is estimated by unwrapping the phase of the corresponding analytic signal (obtained by frequency domain HT techniques), it does not give this local step, and neither does it give the corresponding impulse when the phase is differentiated to give instantaneous frequency. The result of this operation is shown in Fig. 6(b), where it is seen that rather than maintaining the constant value (1 kHz) of the resonance frequency until the transition, the mean value actually falls to zero, and there is a step to 1 kHz at the transition. This is an estimation error, presumably due to the fact that the condition for being able to carry out these HT operations in the frequency domain is violated because this signal, with a step in phase, cannot be validly represented as a mono-component analytic signal. One of the requirements of the latter is that the frequency fluctuation around the carrier must always be less than the carrier frequency, to give a one-sided spectrum, and it is clearly seen that even without added noise, this is not the case, and results in an aliasing phenomenon.

When a small amount of noise is added to the signal, as in Fig. 6(c), the phase jump at the transition can no longer be seen, and the instantaneous frequency becomes dominated by the noise as soon as the amplitude of the IR falls below it. Interestingly, the step in frequency at the transition, caused by aliasing, is still visible in Fig. 6(d).

Fig. 7 shows the squared envelope spectrum of Fig. 6(c), and the spectrum of the squared frequency of Fig. 6(d). The repetition frequency is 20.2 Hz. Whereas the effect of the noise is negligible in Fig. 7(a), in particular because the envelope is squared, it is quite strong in Fig. 7(b), and made worse by the squaring. This negates the claim that the instantaneous frequency (squared in the TKEO) is more sensitive than the squared envelope. It also justifies dropping the requirement that mono-components should be extracted, since even a single resonance IR cannot meet this requirement, and as will be shown in the following, the true benefit of taking the squared envelope of the derivative of a signal lies in the weighting by frequency over a wide frequency range. This then allows not only for multiple resonances to be excited, but also overlapping resonances, which are very common in the high frequency range, where modal density is high.

The above analysis is reasonably valid for both gear and bearing diagnostics, at least for the situations discussed in Refs. [6,7], i.e. high frequency IRs, even though the simulation is actually based on IRs in terms of displacement, although in practice they are more likely to be measured in terms of acceleration. The latter have a quite different spectrum, in that it is weighted by ω^2, and thus dominated by a constant mass-line at high frequencies. There is very little visible difference in the time domain signals, however, except that the sign is reversed, but that can also correspond to accelerometer mounting in the opposite direction. There is another difference, more important for bearing signals, in that the IRs are not periodic, but have a small random variation in spacing because of slip between the components, and the random placement of the rolling elements in the clearance of the cage. This means that the instantaneous frequency signal would be even less periodic than

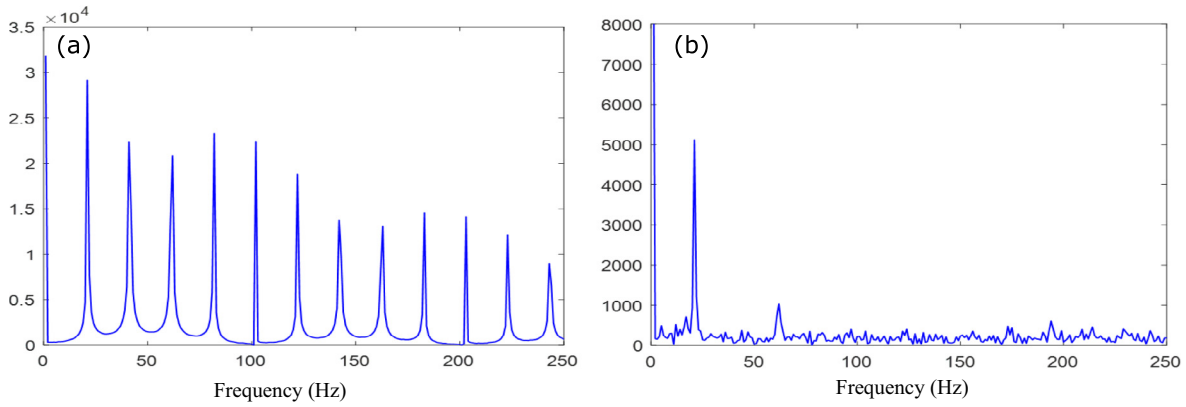

Fig. 7. (a) Spectrum of squared envelope of Fig. 6(c) (b) Spectrum of squared frequency in Fig. 6(d).

indicated in Fig. 7(b), since the phase jumps at the start of each new IR would not only have random spacing, but also random value anywhere in the range ±180°.

It has been shown that even in the case of outer race fault signals, which tend to be uniformly weighted, they cannot be validly represented as mono-component signals because of phase discontinuities. The situation is even worse for inner race and rolling element faults, which rotate through the load zone, since they are unloaded for at least half the time, meaning that the amplitude is then zero, and the phase meaningless.

Ref. [7] does present results from actual bearing faults, and claims that the spectrum of the TKEO is better than that of the envelope of a bandpass filtered signal for the same case, by exhibiting a greater number of harmonics of the fault frequency. Since the current paper shows that the TKEO is basically the squared envelope of the derivative of the signal, it seems most likely that the cited advantage is due primarily to the fact that the TKEO approach was compared to the envelope, rather than the squared envelope, of the signal. Ref. [19] showed in 2000 the benefits of analysing the squared envelope, which are due partly to the fact that the square of an exponential envelope has a spectrum with twice the 3 dB bandwidth, giving twice as many harmonics of a given repetition frequency. The results presented below in Fig. 9 illustrate this, as well as the fact that the linear frequency weighting given by the TKEO (or equally by the FDEO) is only of advantage for differentiation over a wide frequency band, where the requirement of a mono-component cannot be maintained, but it has just been seen that there is no need for the latter in most diagnostic applications.

An example of bearing diagnostics on a signal of this type is now given to illustrate a number of these points, in particular the extent to which it might be an advantage to analyse the squared envelope of the signal or that of its derivative. Fig. 8

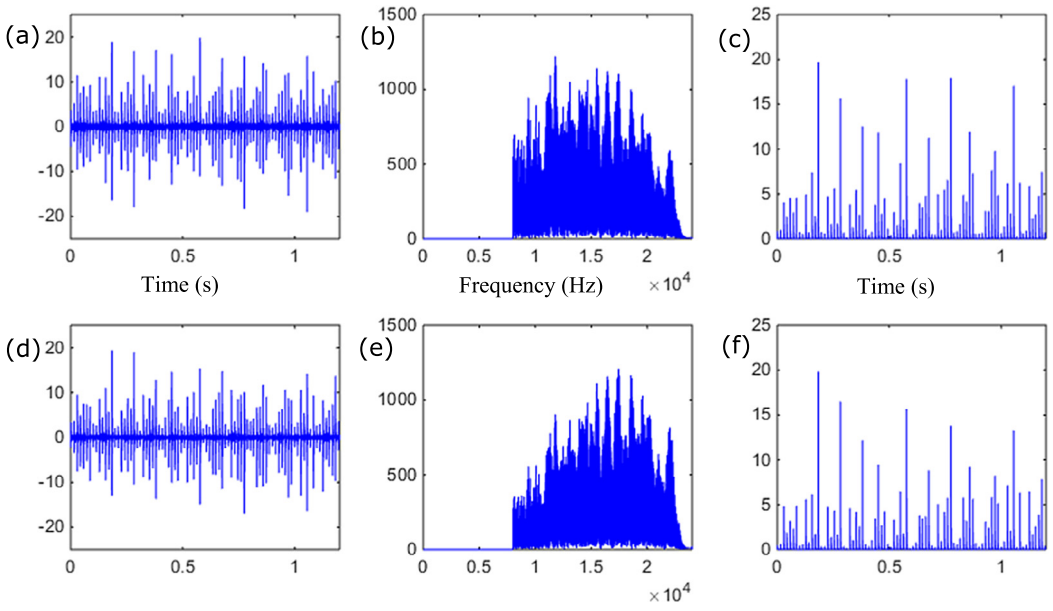

Fig. 8. Comparison of acceleration and its derivative (a, b, c) Acceleration (d, e, f) Jerk (a, d) Time signals (b, e) Spectra (c, f) Squared envelopes.

shows an example comparing a (bandpass filtered) acceleration time signal, for a gearbox bearing with an inner race fault, with its derivative (jerk). The amplitude scales are of course not directly comparable, but have been adjusted to be similar. Fig. 8(a) is the original acceleration signal, and 8(d) is its derivative. This was achieved by multiplication by jω in the frequency domain, which can be seen by comparison of the spectra in Fig. 8(b, e) by the frequency weighting of the amplitude spectrum in 8(e). The bandpass filtration from 8 to 24 kHz was applied to remove masking by gear signals below 8 kHz. It was implemented using an ideal filter in the frequency domain, by setting all frequency components from zero to 8 kHz to zero (with unchanged phase at all retained frequencies). Because the FFT operation is non-causal (the second half of the time record also represents negative time) this explains how the zero phase shift is possible. It is not possible to get a causal filter with such a sharp cut-off.

The upper frequency of 24 kHz corresponds to the Nyquist frequency, and the effects of the antialiasing filter can be clearly seen from 22 to 24 kHz. Fig. 8(c, f) are the squared envelope signals obtained by using HT methods in the frequency domain, inverse transforming the (complex) values in the indicated frequency band. It is interesting that despite the 3:1 frequency range of the spectra, the acceleration and jerk signals are not very different from each other, at least in terms of impulsiveness, and neither are their squared envelopes. This is of course data dependent, and the enhancement of higher frequency components could be expected in general to make the signals more impulsive, but only for broadband signals. In this case, the "centre of gravity" of the spectrum has only moved from about 14 kHz to 18 kHz, or <30%, and is seen to have little practical effect.

The envelope spectra (in various forms) are compared with that of the TKEO (of the ideal bandpass filtered acceleration signal) in Fig. 9. The envelope spectra show a series of harmonics of fault frequency (BPFI) surrounded by sidebands spaced at shaft speed. This confirms that the TKEO gives an almost identical result to the squared envelope of the derivative of the signal, though the former was calculated by the time domain formula. In this case, it is not very different from the squared envelope of the signal, because of the limited effect of frequency weighting. Note that the spectrum of the envelope has reduced harmonics compared with that of the squared envelope, because the squared exponential has half the time constant and double the bandwidth, as stated above. Note also that the squared envelope spectrum has enhanced sidebands, which is the opposite of the effect of squaring the amplitude of the spectrum as opposed to the squaring of the envelope before obtaining the spectrum, even though the units would be the same.

Ref. [7] makes the claim that the TKEO tends to eliminate low frequency interference from other components such as gears, and this is true to some extent but of course it is purely because differentiation gives a highpass filter effect, but is very data dependent. Fig. 10 illustrates this using the same data as Figs. 8 and 9, but analysed using conventional HT techniques via the frequency domain. Fig. 10(a) is the (fast) kurtogram obtained from the original acceleration signal, and indicates the passbands with the highest kurtosis (impulsivity) usually dominated by bearing faults. The highest is actually given by the band centred on 23 kHz, with bandwidth 2 kHz, but for demonstration purposes the analysis is done on the next highest band, centred on 13.5 kHz with bandwidth 3 kHz. Fig. 10(b) is the kurtogram obtained from the signal differentiated twice to give a weighting of $-\omega^2$ on the spectrum. The major difference is that the full bandwidth signal now has a high kurtosis, showing that the part of the original spectrum dominated by gears, up to 8 kHz, no longer dominates. It was also checked that a single differentiation, equivalent to the TKEO, also gave almost as good a result. The spectrum was similar to

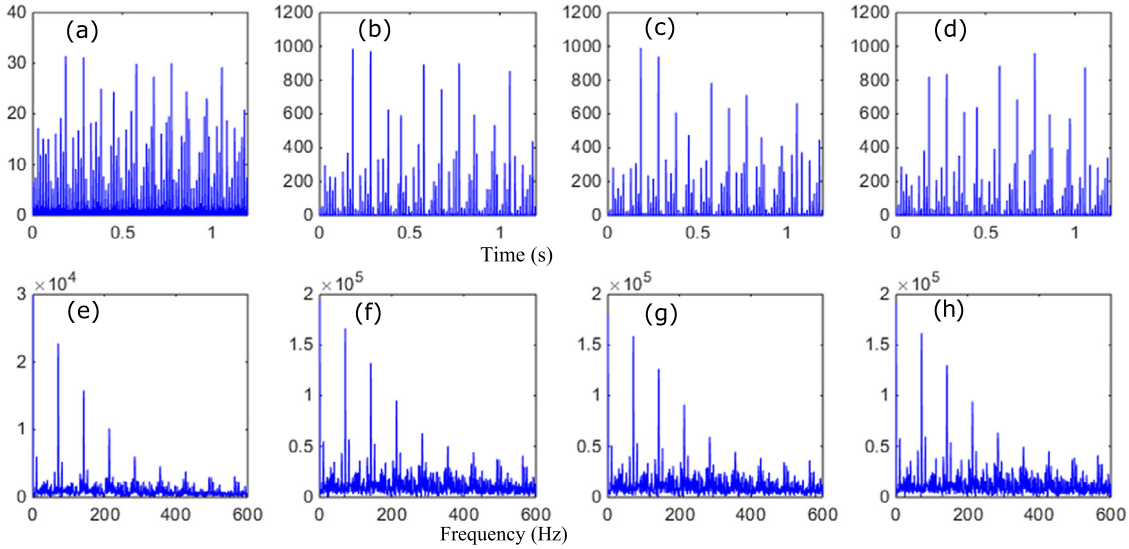

Fig. 9. Comparison of various envelope signals and their spectra (a) Envelope of acceleration (b) Squared envelope of acceleration (c) Squared envelope of jerk (d) TKEO (e-h) Corresponding spectra.

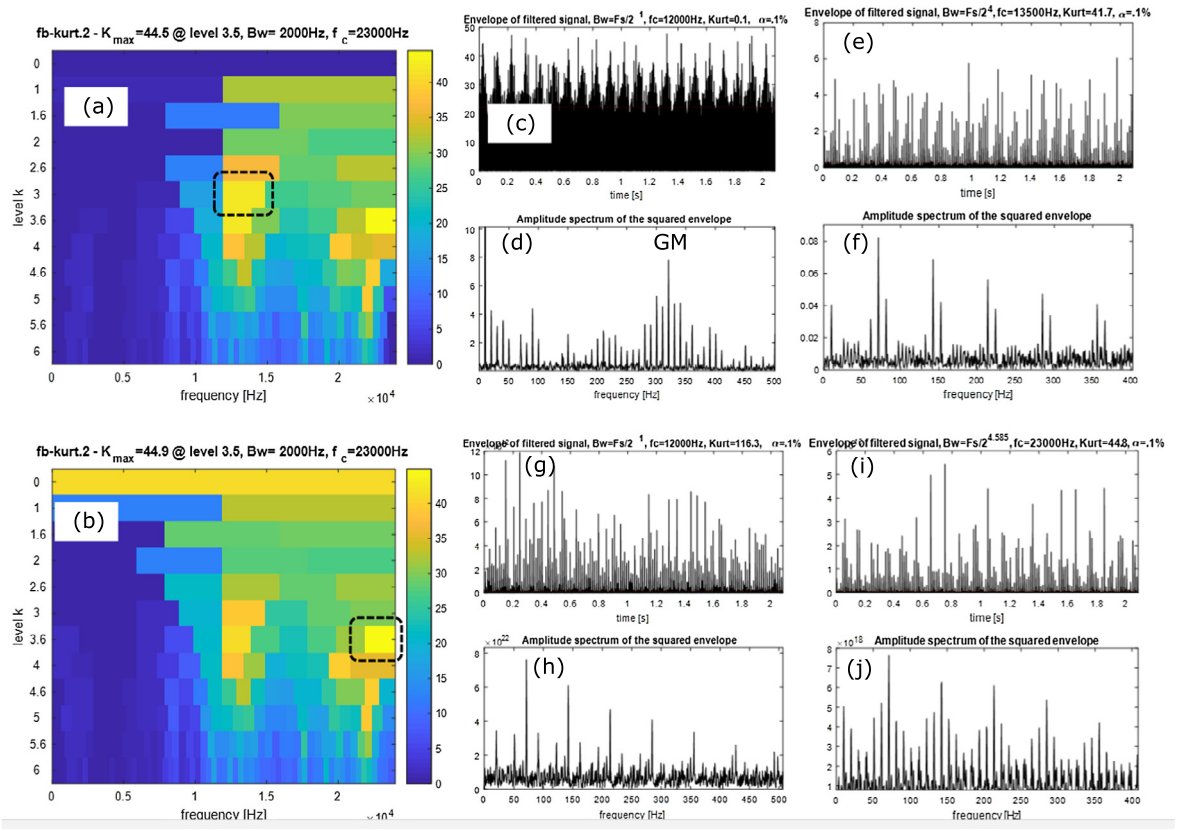

Fig. 10. Effect of differentiation of signal before enveloping. (a) kurtogram of original signal. (b) kurtogram of double differentiated signal. (c) Envelope of original signal. (d) Spectrum of squared envelope in (c). (e) Envelope of selected band in (a). (f) Spectrum of squared envelope in (e). (g) Envelope of double differentiated signal. (h) Spectrum of squared envelope in (g). (i) Envelope of selected band in (b). (j) Spectrum of squared envelope in (i).

that in Fig. 8(e), but extending in a triangular fashion down to zero at zero frequency, instead of the ideal high pass filtration at 8 kHz applied there.

Fig. 10(c) is the envelope of the original full bandwidth signal, and Fig. 10(d) is the spectrum of its square. This is completely dominated by harmonics of the shaft speed, including the gearmesh frequency (GM). Fig. 10(e) is the envelope of the signal, ideal bandpass filtered in the indicated band, and now shows the IRs from the bearing fault. This is confirmed by the corresponding squared envelope spectrum in Fig. 10(f), which has a series of harmonics of BPFI, surrounded by sidebands at shaft speed, as typical of an inner race fault, and agreeing with Fig. 9(f). Fig. 10(g) is the envelope of the full bandwidth double differentiated signal corresponding to the first row (level 0) of the kurtogram in Fig. 10(b). The corresponding squared envelope spectrum, Fig. 10(h), now shows the harmonics of BPFI, but the sideband spacing is twice shaft speed. Fig. 10(i) is the envelope of the signal bandpass filtered in the indicated optimum band, with highest kurtosis, and shows the bearing IRs more clearly than Fig. 10(e) and (g). The squared envelope spectrum of Fig. 10(j) is also clearer, in that the sidebands showing modulation at shaft speed are more prominent. Note that this same band has about the same kurtosis in Fig. 10(a), and so would give a similar result without the double differentiation. This can be explained by the fact that the upper end of the band is only 10% higher than the lower limit, so even double differentiation would only give 20% increase in gain across the band. This, and the difference between Fig. 10(d) and (h), emphasises the fact that the differentiation inherent in the TKEO is only beneficial over a wide frequency range.

Ref. [7] recommends applying the TKEO operation multiple times, testing each time if the bearing frequencies have become visible, but this is a very poor way of applying multiple differentiations, compared with simply multiplying the ideal filtered section of the (1-sided) spectrum by the appropriate power of jω. Firstly, the TKEO is a squared envelope signal, with very different properties than the original signal, so the second application will give a squared envelope of a squared envelope. Ref. [19] points out that the spectrum of a squared envelope is the convolution of the original spectrum with its complex conjugate, doubling the bandwidth each time, so in fact to avoid aliasing, it would be necessary to double the sample rate, or halve the initial bandwidth before each further iteration. Ref. [7] does mention that the number of iterations is limited by the increase in high frequency noise each time, but does not mention the aliasing problem.

In the approach recommended in this paper, the squared envelope of both the signal and its derivative are obtained from signals bandlimited in the frequency domain, but the result is equally valid as a representation of the time domain TKEO if

the full bandwidth (1-sided) spectrum is processed. It would just not be a mono-component signal, and the high frequency part might be greatly contaminated by an unknown amount of noise. There are many machines in which the contamination of the bearing signals by discrete frequency components, such as from gears, as analysed in [7], is limited to a certain maximum frequency, and the higher part of the frequency range is dominated by bearing fault effects, and in that case a simple highpass filtration before processing would be a big advantage, but multiple differentiations might also help as an alternative. However, carrying out these differentiations in the frequency domain, by multiplication by the appropriate power of (jω), is the best way to achieve this, without introducing any aliasing or other problems, and the upper frequency can be limited to avoid enhancing noise.

However, there are many other situations where the high frequency range is not completely dominated by bearing signals, and it is necessary to find a band where the latter dominate, and isolate it, preferably with an ideal filter. A very common example is gearboxes such as from helicopters and wind turbines, with a very high gear ratio of the order of 100:1. It is then very common for the gear frequencies to overlap with the bearing signals. To illustrate this point, the signals used in Figs. 7–10 have been modified by the addition of two high frequency components, to simulate two harmonics of a high speed gearmesh frequency, as shown in Fig. 11. Fig. 11(a) shows the spectrum of the original modified signal, depicted in 11(d). Fig. 11(b) shows the spectrum when filtered by a causal butterworth filter of order 4. It is obvious that this cannot remove the discrete frequencies, in particular that closest to the filter band. A higher order filter could have given somewhat better extraction but at the expense of greater phase distortion, which can modify peak values. Fig. 11(c) shows the resulting squared envelope spectrum, where there is a faint indication of the second, third and fifth harmonics of BPFI, with some sidebands. In contrast, applying an ideal filter in the same optimum band centred on 14.25 kHz, with bandwidth 4.5 kHz, gives the envelope signal of Fig. 11(e), dominated by the bearing fault, as shown by the squared envelope spectrum of Fig. 11(f), which can be compared with Figs. 9(f) and 10(f).

As mentioned above, Ref. [9] uses the FWEO of Ref. [17], and claims that it gives better results than the TKEO, but it is difficult to understand how this arises. The derivations follow exactly those of Ref. [17] and give a discretised time domain version of the operator, including a discretised time domain HT operator. If applied in the time domain it cannot give a more accurate result than the method proposed in this paper where exact differentiation is achieved by a jω operation in the fre-

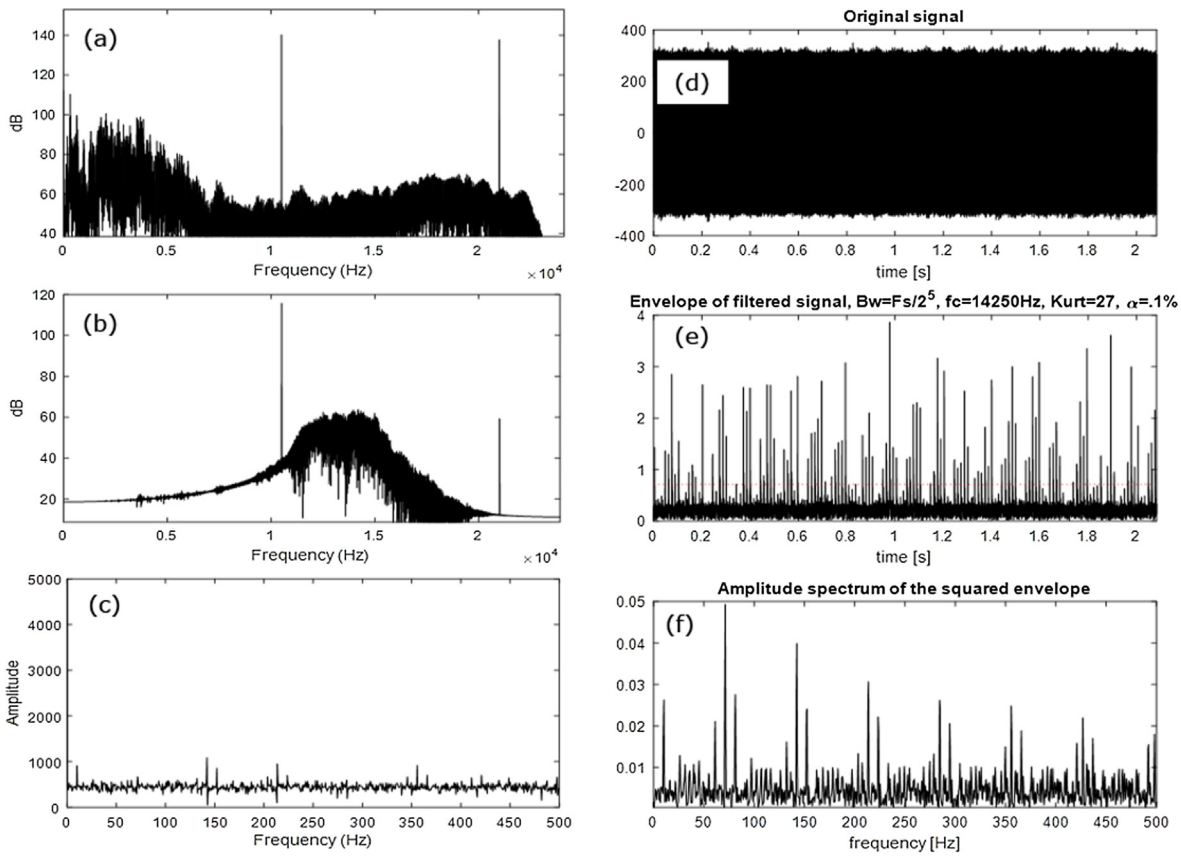

Fig. 11. Masking by discrete high frequency signals. (a) Spectrum of signal with two added high frequency harmonics. (b) Spectrum of signal filtered by a causal bandpass filter in optimum band. (c) Squared envelope spectrum of signal from (b). (d) Original time signal. (e) Envelope of signal filtered by an ideal filter in optimum band. (f) Squared envelope spectrum of signal from (e).

quency domain. Perhaps the frequency weighting is actually done in the frequency domain in [9], because the computation time for their FWEO result is comparable with a normal envelope analysis via the frequency domain and much shorter than the TKEO procedure. There is an error in the paper in that the word "energy" is used to mean two entirely different things. In one case it is taken as the square of the envelope (amplitude) of a signal following Loutridis (Ref. [30] of [9]), which is called "energy density" in the title of that paper, where it is actually defined as the "instantaneous energy per unit time (power)" in Loutridis, and is derived from a Wigner-Ville distribution (WVD), by integrating over a frequency band. The word power there is used in the same sense as in power spectral density, and is simply the square of the instantaneous signal. In [17], the word "energy" is used as in all the literature on energy operators, and is based on the total energy (kinetic plus potential) of a mass/spring system in free vibration. If the system has a damper, the total energy would decay exponentially. If an external force is applied to the system, the total energy in the system will stay the same if the force supplies the same power as the rate of dissipation in the damper(s). If the rate of energy supply from the force is less than or greater than the dissipation in the damper(s) the total energy will decrease or increase accordingly.

Ref. [9] uses EEMD to divide the bearing signals into IMFs, which are supposed to be mono-components, but as discussed in connection with EMD analysis, and Fig. 6, this is not valid for outer race faults, and even less so for inner race faults, where the amplitude is zero for more than half the time. Even if it were possible to get short bursts of IMFs where the signal is non-zero, there is no guarantee they would all be the same, or similar, and they would be dominated by end effects that cannot be eliminated by truncation. In any case, EEMD analysis must be done by post-processing, so there is no benefit to be gained by combining it with causal real-time TKEO, and the FDEO would be preferable.

4.3. Determination of machine speed

4.3.1. FDEO method

Having shown that the application of energy operators to gear and bearing diagnostics does not require them to be limited to mono-components, with correspondingly limited frequency range, an application to the determination of machine speed is now discussed. This application is one where the restriction to a mono-component is absolutely necessary, as it involves frequency demodulation, where the signal being analysed can only contain one carrier frequency, giving information about the speed being sought. This application of an energy operator was first described in [20]. It is based on the FDEO, and in particular Eq. (10), although it was not realised at the time that this gave an advantage with respect to the TKEO, which uses Eq. (14) to estimate instantaneous frequency. As mentioned above, Eq. (14) actually approximates the instantaneous frequency of the *derivative* of the signal, and this is not the same. First, however, the results from [20] will be presented, as they demonstrate that a very accurate estimate can be obtained using the FDEO, both from a tacho signal, and from a vibration response signal.

The main example in [20] is based on data from a gearbox test rig with two parallel shafts, driven by a 4-pole induction motor with variable frequency drive (VFD) as shown in Fig. 12. The 46-tooth (input) and 25-tooth (output) gears are arranged in a speedup ratio to maximise pump torque, and a 2-per-rev tacho signal is taken from the output shaft. A series of vibration signals were taken from the accelerometer mounted on the gearbox casing, and the speed was varied manually using the control buttons for the VFD. Four recordings were made for speed variations around nominal 22 Hz (Signals 1–1 to 1–4), and two for variations around 15 Hz (Signals 2–1 and 2–2). The variable speed part varied by about ± 20% around the nominal speed. Fig. 13 shows the speed profiles of the six signals, which were intended to be used to test a method to remove the effects of speed variation on gear dynamics [21]. Here, only the details of producing the smoothed profiles in Fig. 13 are given.

Fig. 14 shows spectra for the variable speed part of the typical signal 1–1. Fig. 14(a) and (b) show the acceleration spectrum, and Fig. 14(c) and (d) show the spectrum of the tacho signal from the output shaft, with frequency 3.68 times the input

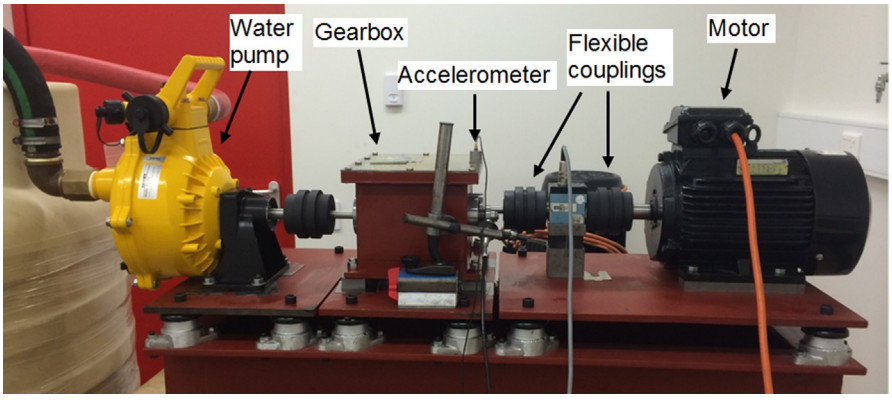

Fig. 12. Layout of gearbox test rig.

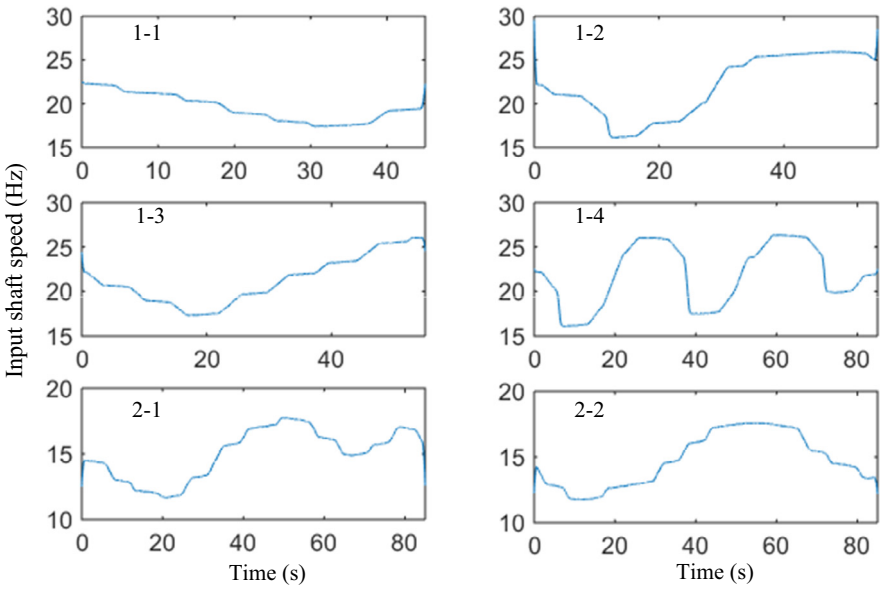

Fig. 13. Smoothed speed profiles for the six signals [20]

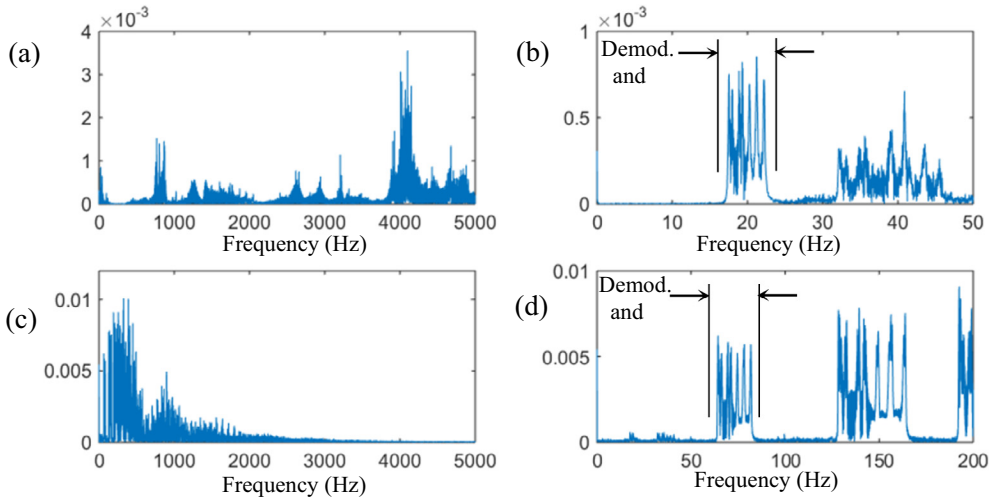

Fig. 14. Spectra of vibration response and tacho signals, analysed for instantaneous speed estimation [20] (a) Acceleration spectrum 0–5 kHz (b) Acceleration 0–50 Hz (c) Tacho 0–5 kHz (d) Tacho 0–200 Hz.

shaft speed (ISS). Even though the overall spectra are quite different, the lowest harmonic of each can be seen in the zoomed spectra of Fig. 14(b, d), with the input shaft speed (ISS) in 14(b) centred on about 20 Hz, and the second harmonic of the output shaft speed (OSS) in 14(d) centred on about 75 Hz. Each of these lowest harmonics is completely separated from adjacent components and noise, allowing for the choice of an uncontaminated demodulation band as indicated in the figure. As might be expected, the second harmonic of the tacho signal (4th harmonic of OSS) in Fig. 14(d) is twice as broad as the first, but the second harmonic of the ISS in Fig. 14(b) is already overlapping with the first harmonic of OSS, and therefore not separable.

Both the indicated demodulation bands were frequency demodulated using Eq. (10) and the results (scaled for the ISS) are shown in Fig. 15(a). Both curve estimates were smoothed as discussed below, but each had about the same amount of noise before smoothing. Fig. 15(a) shows that similar equivalent results can be obtained from both response signal and tacho. There is no absolute measure of the correct result, though the tacho could be expected to be more accurate. Even so, outside the end effect zones, the maximum difference is 0.24% and standard deviation 0.03%. All results shown in Fig. 13 are from the vibration response signals.

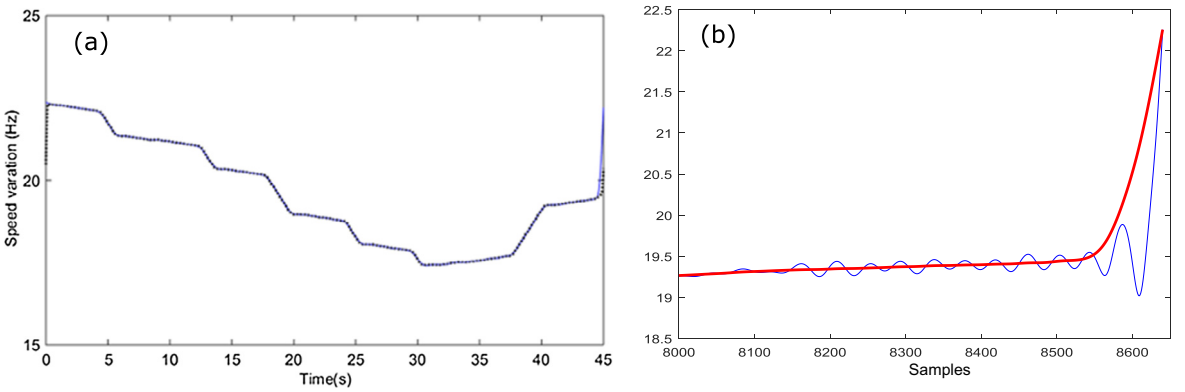

Fig. 15. (a) Comparison of speed estimates from acceleration (blue, solid) and tacho (black, dotted) adjusted for ratio (b) Zoom on end effects for acceleration signal wraparound error (blue) smoothed result (black). (From [20]).

As detailed in [20], there was some noise in the demodulated results, and this was smoothed using a zero phase shift moving average filter (Matlab® function FILTFILT) of length 100 samples. Fig. 15(b) compares the unsmoothed and smoothed results near the right hand end of the speed profile obtained from the acceleration signal. This is scaled in samples, but corresponds to the section from 41 to 45 s in Fig. 15(a). It is seen that the extent of the effects of both the wraparound error and the smoothing filter is of the order of the smoothing filter (100 samples), and could be removed by truncation. It is also seen that the smoothing filter annuls a lot of the increased oscillation amplitude near the end of the record (see for example Fig. 3) meaning that more of the result can be utilised.

4.3.2. "Exact" frequency method and comparison with TKEO and FDEO

Feldman [10] derives the instantaneous frequency of an analytic signal, whose real and imaginary parts are related by a Hilbert transform, by taking the time derivative of the phase angle:

$$\omega(t) = \dot{\phi}(t) = \frac{x(t)\dot{\tilde{x}}(t) - \tilde{x}(t)\dot{x}(t)}{A_x^2(t)} = \mathrm{Im}\left[\frac{\dot{x}_a(t)}{x_a(t)}\right] \tag{17}$$

It can be shown (see Appendix A) that this has an extra term compared with Eq. (10), so that:

$$\omega^2(t) = \frac{A_{\dot{x}}^2(t)}{A_x^2(t)} - \left(\frac{\dot{A}_x(t)}{A_x(t)}\right)^2 \tag{18}$$

where $A_{\dot{x}}^2(t) = \mathrm{Envsq}[\dot{x}(t)]$, so that the first term on the right is the same as Eq. (10). When $\dot{A}_x(t)$ is small, the second term is often almost negligible, but note that in the absence of noise Eq. (10) will give a systematic positive bias compared with the true result. With noise present, the raw speed estimates of Eq. (17) can take on negative values, but the fact that Eq. (10) estimates the squared speed gives an additional positive bias through rectification. Note, too, that Eq. (10) is obtained from Eq. (17) only on the assumption of a slowly varying (squared) envelope (i.e., $\dot{A}_x(t)/A_x(t) \approx 0$), and does not require the additional assumption of Eq. (6) that the instantaneous frequency $\omega(t)$ be slowly varying. It should be kept in mind, however, that the requirement of an analytic signal does place some limits on $\omega(t)$ and its rate of change, as discussed in section 4.2 in connection with Fig. 6.

The errors associated with taking the derivative of a signal, and with the second term in Eq. (18), are discussed in Ref. [22], where it is shown that differentiating a signal changes the mix of amplitude and frequency modulation, and so in general will give a change in them. It can be shown that differentiating a purely amplitude modulated signal will give some frequency modulation (FM), while maintaining the same amplitude modulation (AM), and differentiating a purely frequency modulated signal will give some AM, while maintaining the same FM, but differentiating a mixture will give a change in both.

Fig. 16 shows one example (from [22]) which explains graphically how this can happen for modification by a transfer function, not necessarily a differentiation or integration, at least for amplitude effects. The case illustrated is for a pure AM signal with symmetric sidebands, and is close to that for differentiation (a linear change with frequency) in the relevant frequency range. When this spectrum is multiplied by the transfer function shown, it makes the sidebands uneven, but they can be decomposed into a symmetric pair, giving (the same) AM and an asymmetric pair which in fact primarily give PM/FM, when the sidebands are small with respect to the carrier. This is because a quarter of a period of the modulating frequency after time zero (at which time the AM sidebands are opposed and cancel) the asymmetric pair will align at right angles to the carrier and give a maximum phase deviation. One half period later they will align in the opposite direction, and give a maximum phase deviation in that direction. In between, the phase deviation passes through zero. If the asymmetric sidebands

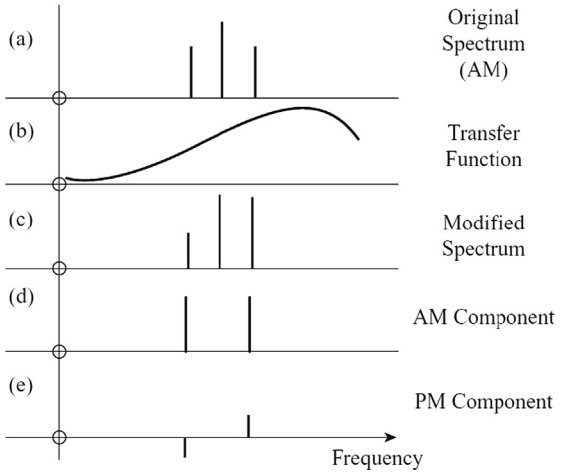

Fig. 16. Illustration of how a transfer function can change the distribution of amplitude and phase modulation.

are small with respect to the carrier, the length of hypotenuse will not change significantly, and a phase modulation at the same frequency will result. Since FM is the derivative of PM (angular velocity vs angular displacement) the FM is scaled by multiplication by ω_m, the modulation frequency in rad/s.

Fig. 17 shows a numerical example from [22] for a case of pure AM (±20%) of a 500 Hz carrier, modulated at 10 Hz. The "exact" result (Eq. (17)) has no error, while the "Ratio of SEs" (squared envelopes), i.e. using the FDEO (Eq. (10)), has the small error given by the second term in Eq. (18), but the "Exact (next deriv)" and "Ratio of SEs (next deriv)" show that the error involved in differentiation is at least an order of magnitude greater, while the error from the additional high frequency ripple in the "TKEO" result (Eq. (14)), with inherent differentiation, is about 50% greater again. The formula for the error purely resulting from differentiation is derived in Appendix A of [22], but is seen to be relatively small when the spread of sidebands is small as in this case. This would seem to indicate that the error could be significantly greater for FM, where the spread of sidebands can be much greater, even for modulation by a single frequency.

For mixtures in general, the AM sidebands would convolve with the FM sidebands across their whole range. However, Eq. (12) shows that the result of a single differentiation does not give an error in the frequency of a pure FM signal (since it depends only on the rate of change of amplitude), but this does then introduce an amplitude modulation, which will lead to errors in subsequent differentiations. In any case, most machine signals will have some AM in addition to FM, and will then give frequency errors for a single differentiation.

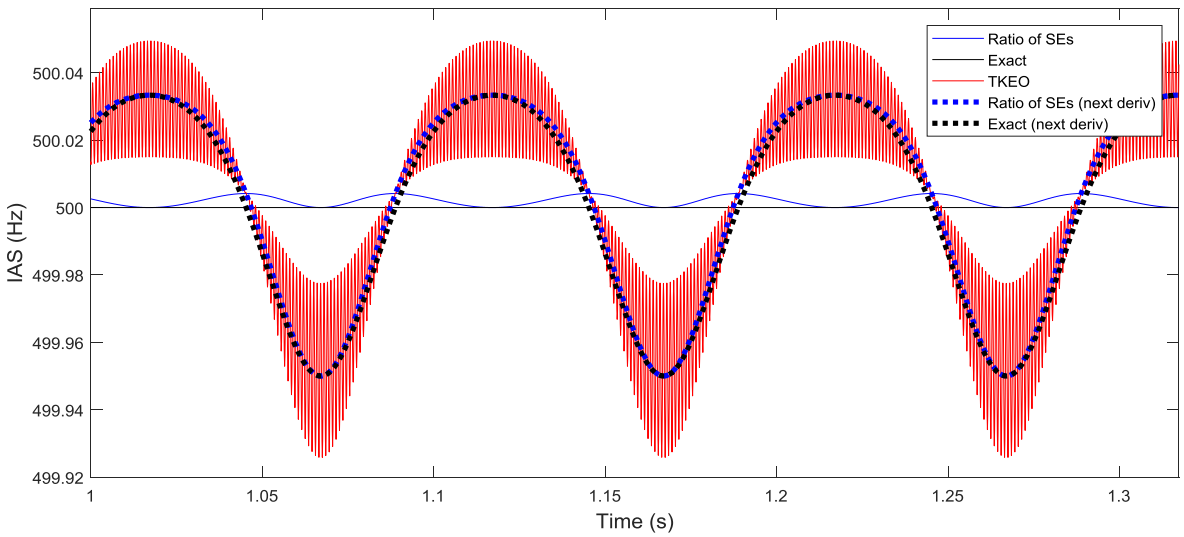

Fig. 17. Effect of differentiation for a signal with pure AM.

Fig. 18(a) shows the estimated speed using the TKEO method (Eq. (14)) on the same data as Fig. 15. To simulate a real-time result, the demodulated band was extracted using an IIR Butterworth filter. The spectrum of the extracted signal is shown in Fig. 18(b). As expected, the TKEO estimate was much noisier than the FDEO result of Fig. 15, and even after smoothing using the same (non-causal) filter, it is still noisier.

Ref. [22] also discusses the question of accuracy of machine speed when obtained from a tacho or vibration response signal. If, as is likely, the frequency of internal forces is most closely related to the machine speed, the frequency of response signals will be affected by passage through a transfer function between the force and the response. If the demodulated frequency band is in the vicinity of a resonance, with phase shift, the instantaneous frequency errors can be quite large because of time delays, as discussed in detail in [23]. However, even if the band lies on a spring line or mass line of the FRF, where response is immediate, the frequency will depend on the measured parameter and whether the band is on a spring line or mass line. This is illustrated in Fig. 19, which shows the resulting errors in instantaneous frequency estimates for a pure FM signal (carrier frequency 500 Hz, frequency sweep ± 100 Hz) when lying on the spring line of a resonance (at 1500 Hz), or on the mass line of a resonance (at 150 Hz). The error given by the exact Eq. (17) is compared with that given by the TKEO, Eq. (14), the latter being much greater. When the demodulated band is on a mass line, which is almost constant for measured acceleration, the error using Eq. (17) is very small, but when it lies on a spring line (effectively double differentiated) the error is almost 0.05%. Note that it could be corrected by double integrating as part of the frequency estimation process, and this is another advantage of non-causal processing, using ideal filters and differentiation/integration by jω operations

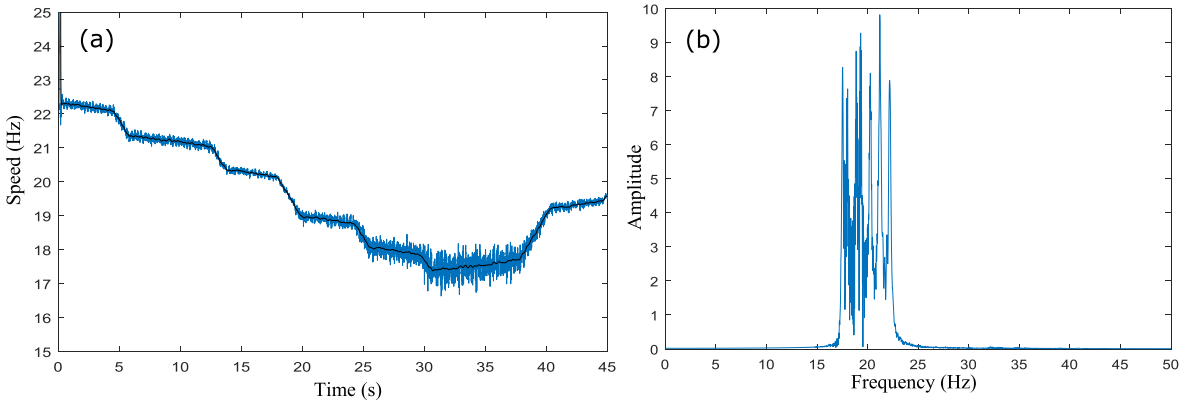

Fig. 18. Speed estimates using TKEO (Eq. (14)) on a bandpass filtered signal encompassing the first harmonic of the acceleration signal. (a) Raw speed estimate and smoothed version (b) Spectrum of bandpass filtered signal.

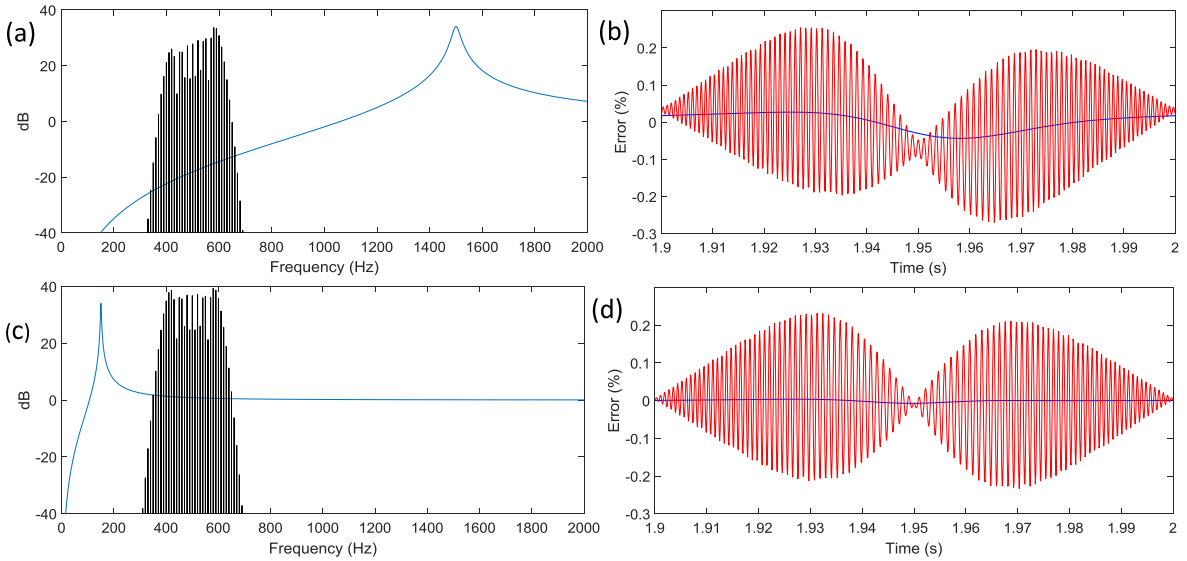

Fig. 19. Effect of transfer function on errors in IAS based on response signal. (a, c) Signal spectrum and FRF (b, d) % Errors (red) TKEO (Eq. (14)), (blue) Exact (Eq. (17)). (a, b) Signal spectrum on spring line (c, d) Signal spectrum on mass line.

in the frequency domain. Ref. [22] also shows how applying an exponential lifter to the real cepstrum of the response signal can indicate whether the band to be demodulated is on a spring or mass line.

5. Conclusion

This paper takes a new look at the TKEO in its application to machine diagnostics. It finds that some dubious claims have been made in the past as to the potential advantages of the TKEO as opposed to other approaches to amplitude and frequency demodulation, in particular in the case represented by machine diagnostics, where real-time operation is not required. In fact, it is a disadvantage, because it imposes causal filtering and other signal processing operations with detrimental effects on the results.

It appears not to have been recognised in all the earlier papers (on machine diagnostics) that the original definition of the TKEO assumes that the signal is dominated by a mono-component carrier frequency. Otherwise any results to do with frequency modulation components are meaningless. On the other hand, for the vast majority of machine vibration signals the maximum frequency range that can encompass a mono-component is 2:1, and so the signal would first have to be bandpass filtered before the application of the TKEO, and if this is to be done in real-time in the time domain, it would have to use causal filters with poor characteristics and phase distortion. It is particularly illogical to combine the potentially real-time TKEO with procedures, such as EMD analysis, which must be done by post-processing.

It appears not to have been realised either that the TKEO is nothing other than (an approximation of) the squared envelope of the derivative of the signal, allowing a direct evaluation of its properties as compared with the squared envelope of the signal directly. The potential advantage is that the weighting of the spectrum proportional to frequency, as given by the differentiation, will increase higher frequency components, but if limited to a maximum frequency range of 2:1, this will have limited effect indeed, not even allowing a second harmonic to be enhanced with respect to the first. On the other hand, if the requirement to be a mono-component is relaxed, then some benefit might be gained by one or more differentiations. The paper shows that, contrary to earlier claims, there is no benefit to be gained by trying to preserve the instantaneous frequency information when performing diagnostics on gears or bearings, at least for those techniques involving high frequency impulse responses, so a wide frequency band can be processed. Multiple differentiations can be achieved at no extra cost, or distortion, by raising the frequency weighting $j\omega$ to any power. Enhancement of high frequency noise is prevented by using ideal filters to isolate only that part of the spectrum to be processed.

The method proposed here generates the squared envelope of both the original signal and its derivative using Hilbert transform techniques via the frequency domain, and the result is called the frequency domain energy operator or FDEO. This means that ideal zero phase shift filters can be used to isolate the single carrier, and differentiations are implemented by $j\omega$ operations on the same section of spectrum as being demodulated, also with zero phase distortion. The disadvantage is the generation of wraparound errors caused by the circularity of the FFT transform, but these can usually be removed by truncation if a slightly longer record is processed.

A promising application is simple estimation of the machine speed, where the requirement of a mono-component, must be adhered to. A harmonic of a tacho signal can be demodulated to obtain the most accurate measure of speed, and if it is the first harmonic it can be done over a speed range up to 2:1. However, a vibration response signal can alternatively be demodulated if a suitable harmonic (preferably the first) can be isolated from other components and noise in the response spectrum. Even though a very accurate result can be obtained using the FDEO, an even more accurate result can be obtained, with the same computational effort, by using the imaginary part of the quotient, rather than the squared envelopes of the analytic signals corresponding to the original signal and its derivative. Both methods are extremely simple to apply, and can normally be smoothed using a zero phase shift smoothing filter.

Acknowledgment

This work was partially funded by the Australian Research Council under Discovery Project DP160103501.

Appendix A Derivation of Eq. (18)

Starting with a slightly modified version of Eq. (12), for an analytic signal, where $\phi(t)$ now includes the carrier component, and $A_x(t)$ is now used to represent the amplitude (envelope) of real signal $x(t)$:

$$x_a(t) = A_x(t)\exp[j\phi(t)] = A_x(t)[\cos\phi(t) + j\sin\phi(t)] = x(t) + j\tilde{x}(t) \tag{A1}$$

where $\tilde{x}$ is the Hilbert transform of x. A_x and ϕ are given by:

$$A_x(t) = \sqrt{x^2(t) + \tilde{x}^2(t)} \tag{A2}$$

and

$$\phi(t) = \tan^{-1}\frac{\tilde{x}(t)}{x(t)} \tag{A3}$$

Now

$$\omega(t) = \dot{\phi}(t) = \frac{d}{dt}\left[\tan^{-1}\frac{\tilde{x}(t)}{x(t)}\right] = \frac{d}{dt}\left[\frac{\tilde{x}(t)}{x(t)}\right] \cdot \frac{1}{1+\left(\frac{\tilde{x}(t)}{x(t)}\right)^2} \tag{A4}$$

which, as shown by Feldman [10] leads to:

$$\omega(t) = \frac{x(t)\dot{\tilde{x}}(t) - \tilde{x}(t)\dot{x}(t)}{A_x^2(t)} \tag{A5}$$

Now:

$$\begin{aligned} x_a(t) &= x(t) + j\tilde{x}(t) \\ \dot{x}_a(t) &= \dot{x}(t) + j\dot{\tilde{x}}(t) \end{aligned} \tag{A6}$$

and if:

$$\begin{aligned} x(t) &= A_x(t)\cos\phi(t) \\ \dot{x}(t) &= -A_x(t)\dot{\phi}(t)\sin\phi(t) + \dot{A}_x(t)\cos\phi(t) \\ \tilde{x}(t) &= A_x(t)\sin\phi(t) \\ \dot{\tilde{x}}(t) &= A_x(t)\dot{\phi}(t)\cos\phi(t) + \dot{A}_x(t)\sin\phi(t) \end{aligned} \tag{A7}$$

then:

$$\begin{aligned} \frac{\dot{x}_a(t)}{x_a(t)} &= \frac{-A_x(t)\dot{\phi}(t)\sin\phi(t) + \dot{A}_x(t)\cos\phi(t) + j\left[A_x(t)\dot{\phi}(t)\cos\phi(t) + \dot{A}_x(t)\sin\phi(t)\right]}{A_x(t)[\cos\phi(t) + j\sin\phi(t)]} \\ &= \frac{\left\{-A_x(t)\dot{\phi}(t)\sin\phi(t) + \dot{A}_x(t)\cos\phi(t) + j\left[A_x(t)\dot{\phi}(t)\cos\phi(t) + \dot{A}_x(t)\sin\phi(t)\right]\right\}\{A_x(t)[\cos\phi(t) - j\sin\phi(t)]\}}{A_x^2(t)} \end{aligned} \tag{A8}$$

of which the real part is:

$$\begin{aligned} \mathrm{Re}\left[\frac{\dot{x}_a(t)}{x_a(t)}\right] &= \frac{-A_x^2(t)\dot{\phi}(t)\cos\phi(t)\sin\phi(t) + A_x(t)\dot{A}_x(t)\cos^2\phi(t) + A_x^2(t)\dot{\phi}(t)\cos\phi(t)\sin\phi(t) + A_x(t)\dot{A}_x(t)\sin^2\phi(t)}{A_x^2(t)} \\ &= \frac{A_x(t)\dot{A}_x(t)\left[\cos^2\phi(t) + \sin^2\phi(t)\right]}{A_x^2(t)} = \frac{\dot{A}_x(t)}{A_x(t)} \end{aligned} \tag{A9}$$

and the imaginary part is:

$$\begin{aligned} \mathrm{Im}\left[\frac{\dot{x}_a(t)}{x_a(t)}\right] &= \frac{\left\{A_x^2(t)\dot{\phi}(t)\sin^2\phi(t) - A_x(t)\dot{A}_x(t)\cos\phi(t)\sin\phi(t) + A_x(t)\cos\phi(t)\left[A_x(t)\dot{\phi}(t)\cos\phi(t) + \dot{A}_x(t)\sin\phi(t)\right]\right\}}{A_x^2(t)} \\ &= \frac{A_x^2(t)\dot{\phi}(t)\sin^2\phi(t) - A_x(t)\dot{A}_x(t)\cos\phi(t)\sin\phi(t) + A_x^2(t)\dot{\phi}(t)\cos^2\phi(t) + A_x\dot{A}_x(t)\sin\phi(t)\cos\phi(t)}{A_x^2(t)} \\ &= \frac{A_x^2(t)\dot{\phi}(t)\sin^2\phi(t) + A_x^2(t)\dot{\phi}(t)\cos^2\phi(t)}{A_x^2(t)} = \dot{\phi}(t) = \omega(t) \end{aligned} \tag{A10}$$

Alternatively:

$$\begin{aligned} \frac{\dot{x}_a(t)}{x_a(t)} &= \frac{\dot{x}(t) + j\dot{\tilde{x}}(t)}{x(t) + j\tilde{x}(t)} = \frac{\left[\dot{x}(t) + j\dot{\tilde{x}}(t)\right][x(t) - j\tilde{x}(t)]}{x^2(t) + \tilde{x}^2(t)} = \frac{\dot{x}(t)x(t) - j\tilde{x}(t)\dot{x}(t) + j\dot{\tilde{x}}(t)x(t) + \dot{\tilde{x}}(t)\tilde{x}(t)}{A_x^2(t)} \\ &= \frac{\left[x(t)\dot{x}(t) + \tilde{x}(t)\dot{\tilde{x}}(t)\right] + j\left[x(t)\dot{\tilde{x}}(t) - \tilde{x}(t)\dot{x}(t)\right]}{A_x^2(t)} \end{aligned} \tag{A11}$$

from which

$$\mathrm{Im}\left[\frac{\dot{x}_a(t)}{x_a(t)}\right] = \frac{x(t)\dot{\tilde{x}}(t) - \tilde{x}(t)\dot{x}(t)}{A_x^2(t)} = \omega(t) \tag{A12}$$

as in (A10), (A5) (Feldman [10]), and repeated in Eq. (17) of the main part.

Taking the square of the absolute value (squared envelope) of the ratio $\frac{\dot{x}_a(t)}{x_a(t)}$ from (A11):

$$\begin{aligned}
\mathrm{Envsq}\left(\frac{\dot{x}_a(t)}{x_a(t)}\right) &= \left\{\frac{\left[x(t)\dot{x}(t)+\tilde{x}(t)\dot{\tilde{x}}(t)\right]}{A_x^2(t)}\right\}^2 + \left\{\frac{\left[x(t)\dot{\tilde{x}}(t)-\tilde{x}(t)\dot{x}(t)\right]}{A_x^2(t)}\right\}^2 \\
&= \frac{x^2(t)\dot{x}^2(t)+2x(t)\dot{x}(t)\tilde{x}(t)\dot{\tilde{x}}(t)+\tilde{x}^2(t)\dot{\tilde{x}}^2(t)}{A_x^4(t)} + \frac{x^2(t)\dot{\tilde{x}}^2(t)-2x(t)\dot{x}(t)\tilde{x}(t)\dot{\tilde{x}}(t)+\tilde{x}^2(t)\dot{x}^2(t)}{A_x^4(t)} \\
&= \frac{x^2(t)\dot{x}^2(t)+\tilde{x}^2(t)\dot{\tilde{x}}^2(t)+x^2(t)(t)+\tilde{x}^2(t)\dot{x}^2(t)}{A_x^4(t)} \\
&= \frac{\left[x^2(t)+\tilde{x}^2(t)\right]\left[\dot{x}^2(t)+\dot{\tilde{x}}^2(t)\right]}{A_x^4(t)} = \frac{A_x^2(t)A_{\dot{x}}^2(t)}{A_x^4(t)} \\
&= \frac{A_{\dot{x}}^2(t)}{A_x^2(t)}
\end{aligned} \tag{A13}$$

i.e., the squared envelope of the ratio is equal to the ratio of the squared envelopes. Using (A9) and (A10) we then have:

$$\mathrm{Envsq}\left(\frac{\dot{x}_a(t)}{x_a(t)}\right) = \frac{A_{\dot{x}}^2(t)}{A_x^2(t)} = \left(\frac{\dot{A}_x(t)}{A_x(t)}\right)^2 + \dot{\phi}^2(t) \tag{A14}$$

or

$$\omega^2(t) = \frac{A_{\dot{x}}^2(t)}{A_x^2(t)} - \left(\frac{\dot{A}_x(t)}{A_x(t)}\right)^2 \tag{A15}$$

proving Eq. (18) of the main part.

References

[1] J.F. Kaiser, On a simple algorithm to calculate the 'energy' of a signal, in: Int. Conf. on Acoustics, Speech, and Signal Process., 1990, pp. 381–384.
[2] P. Maragos, J.F. Kaiser, T.F. Quatieri, On amplitude and frequency demodulation using energy operators, IEEE Trans. Signal Process. 41 (4) (1993) 1532–1550.
[3] P. Maragos, J.F. Kaiser, T.F. Quatieri, Energy separation in signal modulations with application to speech analysis, IEEE Trans. Signal Process. 41 (10) (1993) 3024–3051.
[4] I. Antoniadou, G. Manson, W.J. Staszewski, T. Barszcz, K. Worden, A time–frequency analysis approach for condition monitoring of a wind turbine gearbox under varying load conditions, Mech. Syst. Signal Process. 64–65 (2015) 188–216.
[5] Y. Qu, E. Bechhoefer, D. He, J. Zhu, A new acoustic emission sensor based gear fault detection approach, Int. J. Progn. Health Manag., Special Issue Wind Turbine PHM 4 (2013) 32–45.
[6] M. Liang, I. Soltani Bozchalooi, Teager energy operator for multi-modulation extraction and its application for gearbox fault detection, Smart Mater. Struct. 19 (2010) 18, 075008.
[7] M. Liang, I. Soltani Bozchalooi, An energy operator approach to joint application of amplitude and frequency-demodulations for bearing fault detection, Mech. Systems and Signal Process. 24 (2010) 1473–1494.
[8] Hongmei Liu, Xuan Wang, Lu Chen, Rolling bearing fault diagnosis based on LCD–TEO and multi-fractal detrended fluctuation analysis, Mech. Syst. Signal Process. 60-61 (2015) 273–288.
[9] Y. Imaouchen, M. Kedadouche, R. Alkama, M. Thomas, A Frequency-Weighted Energy Operator and complementary ensemble empirical mode decomposition for bearing fault detection, Mech. Syst. Signal Process. 82 (2017) 103–116.
[10] M. Feldman, Hilbert transform in vibration analysis, Mech. Syst. Signal Process. 25 (2011) 735–802.
[11] M.D. Coats, R.B. Randall, Single and multi-stage phase demodulation based order-tracking, Mech. Syst. Signal Process. 44 (2014) 86–117.
[12] R.B. Randall, A new interpretation of the Teager Kaiser energy operator September, Conference on Vibrations in Rotating Machinery, IMechE, Manchester, 2016.
[13] A.C. Bovik, P. Maragos, T.F. Quatieri, Demodulation of AM-FM signals in noise using multiband energy operators, Proc. IEEE Int. Symp. Inf. Theory (1993) 17–22.
[14] B. Santhanam, P. Maragos, Multicomponent AM–FM Demodulation via Periodicity-Based Algebraic Separation and Energy-Based Demodulation, IEEE Trans. on Communications 48 (3) (2000) 473–490.
[15] N.E. Huang, Z. Shen, S.R. Long, et al, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, Proc. R. Soc. Lond. A: Math., Phys. Eng. Sci. 454 (1971) (1998) 903–995.
[16] Z.A. Wu, N.E. Huang, Ensemble empirical mode decomposition: a noise assisted data analysis method, Adv. Adapt. Data Anal. 1 (2009) 1–41.
[17] J.M. O'Toole, V. Temko, N. Stevenson, Assessing instantaneous energy in the EEG: a non-negative, frequency-weighted energy operator, Proc. IEEE (2014) 3288–3291.
[18] T. Barszcz, R.B. Randall, Application of spectral kurtosis for detection of a tooth crack in the planetary gear of a wind turbine, Mech. Syst. Sig. Process. 23 (4) (2009) 1352–1365.
[19] D. Ho, R.B. Randall, Optimisation of bearing diagnostic techniques using simulated and actual bearing fault signals, Mech. Syst. Signal Process. 14 (5) (2000) 763–788.
[20] R.B. Randall, W.A. Smith, Use of the Teager Kaiser Energy Operator to estimate machine speed 5-8 July, PHM Europe conference, 2016.
[21] R.B. Randall, W.A. Smith, New cepstral methods for the diagnosis of gear and bearing faults under variable speed conditions, ICSV23 Conference, 2016.
[22] R.B. Randall, W.A. Smith, Accuracy of speed determination of a machine from frequency demodulation of response vibration signals, CM-MFPT conference, 2018.
[23] P. Borghesani, P. Pennacchi, R.B. Randall, R. Ricci, Order tracking for discrete-random separation in variable speed conditions, Mech. Syst. Signal Process. 30 (2012) 1–22.