THE HISTORY AND APPLICATION OF THE ENVELOPE DETECTOR

John L. Frarey
JLF Analysis
845 Worcester Drive
Schenectady, NY 12309

Abstract: The envelope detector has become one of the standard techniques used in the detection of defective rolling element bearings. Some forms of the envelope detector are also used in gear box analysis. First uses of the envelope detector seem to date to 1971. This paper will discuss the history of the development. Data from the envelope detector will be analyzed to determine which data characteristics are due to the defective bearing and which are due to the structure and which are due to the characteristics of the envelope detector. Other forms of envelope detectors such as those employing the Hilbert transform will also be covered.

Key Words: Envelope detection; Demodulation; High frequency vibration; Hilbert transform; Machinery diagnosis; Rolling element bearings; Vibration analysis

Introduction: Almost all the early effort to diagnose defective rolling element bearings was based on detecting the classic bearing defect frequencies in low frequency (generally below 500 hz) vibration data. In 1971, with NASA - Huntsville support, a program to investigate early detection of bearing defects showed that this classical approach was not capable of detecting the onset of bearing defects. High frequency vibration and envelope detection was almost accidentally discovered in this program and was shown to be very effective in the early detection of bearing faults. This technique has enjoyed widespread use in the diagnosis of bearing defects particularly in the last five to ten years. Understanding the reason for the failure of the classical diagnostic scheme to detect early failures, which is due to the impact nature of the signal generated by a defect, is also widespread. What is not as well understood is that the amplitude of the envelope detected defect signal from the bearing is a function of the size of the defect but also, unfortunately, is a function of the characteristics of the resonance(s) excited and circuit variables of the envelope detector such as the time constant of the filter employed in the envelope detector. Understanding the role of the analysis circuitry in the defect signal generation is necessary to understand that the data produced by manufacturer A's system may be different from that generated by manufacturer B.

A software envelope detector may be developed by use of the Hilbert transform and analysis of the analytic signal produced by operating on the original and transformed signal. A more brute force approach to a software envelope detector is simply to employ a peak detector with a very fast rise time and slow decay time. A software envelope detector has some advantages, however not enough investigation has been conducted into the response of the Hilbert transform detector when it is presented with multiple resonances excited by the impact. The use of software envelope detectors is not as mature a field as the use of hardware detectors.

Background: The classical approach to the detection of rolling element defects was to examine the low frequency vibration or sonic spectrum for the presence of defect frequencies. These frequencies were calculated by the well known defect frequency formulas that were developed to determine the repetition rate for balls striking an outer race or inner race defect. These equations are developed from the geometry of the bearing and are given below for Ball Bearings:

$$f_o = n/2 * f_r * (1 - BD/PD * \cos\alpha)$$

$$f_i = n/2 * f_r * (1 + BD/PD * \cos\alpha)$$

$$f_s = 1/2 * f_r * PD/BD * (1 - (BD/PD)^2 * \cos^2\alpha)$$

Where:

n = Number of Balls	f_o = Outer race defect frequency
BD = Ball diameter	f_i = Inner race defect frequency
PD = Pitch Diameter	f_s = Ball spin frequency
f_r = rpm/60	α = Contact angle

For roller bearings, α is zero. The equations may also be rewritten in terms of outer and inner race diameters along with the ball or roller diameter.

The diagnostic concept was simply to calculate the defect frequencies and then monitor the vibration or sonic spectrum until these frequencies appeared. While there was some success with this technique, the results were not at all consistent. In order to investigate this, a controlled test was run in the NASA R & D program mentioned above. A small defect was machined into the inner race of a test bearing and then the bearing was operated in a fully instrumented test rig. None if the instrumentation results showed the presence of this defect; these included vibration, sonic, strain torque and temperature readings. Figure 1 shows the comparison of the vibration spectrum for the undamaged and the damaged bearing. These plots are in a logarithmic scale for amplitude which is not usually used. Even in this scale it is not clear if the defect may be detected. A very weak case could be made that the bearing defect signal is just beginning to appear, however there are other spectral regions in which the random variation is as great or greater than the defect region. To this point, the result of the program was virtually nil. One day, he program team was assembled in the instrumentation room while the rig was running trying to decide what to do next. One oscilloscope happened to be set to look at the raw data and one of the team members observed the signal similar to that shown in Figure 2. The time was measured between the impact signals and when converted to frequency, matched the defect frequency for an inner race defect. When a non defective bearing was substituted, the characteristic signal of Figure 2 disappeared.

Why hadn't any of the data processing techniques used detected the presence of a defect?

The first reason could have been found if a paper by H.L. Balderston [1] given in 1968 had been found and read. He noted that defective bearings produced a large amount of energy in the high frequency region of the spectrum, above 10 to 20 kHz. The problem of locating the results of other work in an area of interest was even more difficult in 1971 than it is now. Today vibration related papers would be listed in the Shock and Vibration Digest or a literature search could be requested at a library. Spectrum analyzers used at that time usually only analyzed data to 20 KHz so if the data were at frequencies higher than 20 KHz they would not have been seen by the investigators. To investigate if the energy were above 20 KHz, the data were tape recorded and played back at half speed. When this was done, a large resonant hump was noted in the spectrum at 28 KHz. So the first answer to the question of why no difference had been noted in the data was that its frequency was above the analysis range. It also serves as a good lesson that time data, preferably from an oscilloscope should always be examined along with the spectrum.

In looking at the data shown in Figure 2 again, the information of interest is not really in the high frequency but rather in the shape of the signal. If a line could be drawn around the envelope of the signal as shown in Figure 3, then the spectrum of this envelope should show the defect frequency. In order to produce the envelope signal of Figure 3 a demodulator similar to those used in the detection of amplitude modulated radio waves could be developed for the lower frequencies. A simple demodulator is shown in Figure 4. The value of R and C establishes the cut off frequency or the 3 db down point. The cut off frequency is given by the following equation:

$$f_{co}=1/(2*\pi*R*C)$$

For example, if R is 10 Kohms and C is .08 μF, the cut off frequency is 200 hz. The cut off frequency should be about 3 to 4 times the highest defect frequency and about 1/100 of the resonant frequency. We will see later in this paper how the cut off frequency affects the analysis results.

Employing the demodulator or envelope detector on the data similar to Figure 2 and then doing a low frequency spectrum of the enveloped data, the spectrum shown in Figure 5 results. The defect signal is now clearly evident with a very good signal to noise ratio (note: the amplitude scale is linear, not logarithmic). Commercially available envelope detectors are much more sophisticated than the circuit of Figure 4; using full wave rectification and active detectors, however, satisfactory results can be obtained using simple circuits. The result of this contract work was reported in the final report [2], to MFPG (3) and in NASA Tech Briefs [4&5].

Direct low frequency defect detection insensitivity: It has been shown that the direct analysis of the low frequency spectrum for the defect signal is much less sensitive than the envelope detected signal. In order to understand the difficulty in analyzing the low frequency spectrum for the defect signal, we should examine the energy distribution in a spectrum as a function of the shape of the signal in the time domain. Figure 6 shows this relationship for a sine wave, a square wave and two more cases where the pulse width is greatly reduced. If the nature of the defect signal were a sine wave, then all of the energy would be in the fundamental defect frequency. As the pulse changes from a sine wave to a square wave, one sees that the energy is now shared between the odd harmonics and the amplitude of the fundamental is reduced. In the final example where the defect signal is a very short impulse, one can see that the energy is almost evenly shared over a very wide range of harmonics of the defect frequency. If the manner in which a defect in the race generates a signal is reviewed , it is seen that the most likely signal appearance is the very short pulse shown in the last case of Figure 6. Much as in the case of an automobile hitting a pothole in the road, a very sharp impact results.

The signal that we are looking for in the low frequency spectrum has an energy distribution as shown in the bottom spectrum of Figure 6. In observing normal vibration measurements from machinery, there is always a noise floor due to the machine noise background. In the case of the initiation of a small defect, the harmonics of the defect frequency may easily have amplitudes below the background noise and therefore will not normally be detected. If the amplitude of the impact increases due to more and larger defects, the defect signal will eventually be seen in the low frequency spectrum. In fact the spectrum for the short pulse impact looks very much like the spectrum generated by an impact from the hammer in modal analyses tests. The similarity with modal analysis does not end here. As in modal analysis, the broad distribution of energy will excite structural resonances that may then be studied. For the case of the bearing, the wide distribution of energy will also excite resonances at much higher frequencies. These resonances may be structural or they could be the resonance of the accelerometer. The resonances excited are normally in the high frequency region since at these frequencies, the structure does not have to have much displacement to achieve a high g level. The data in Figure 2 may therefore be viewed as the excitation and decay of some higher frequency resonance due to the impact generated by the ball hitting the surface defect in the race.

The effect of envelope detecting this resonance response is to eliminate the high frequency and greatly increase the pulse width and therefore increases the energy in the defect signal fundamental frequency. In addition to the lengthening of the pulse due to the Q of the resonance, the envelope detector time constant will also play a role. Trying to find a link between the high amplitude of the demodulated spectrum defect signal and the severity of the defect, we find out that it is not only a function of the strength of the impact (or the size of the defect), but also the Q of the resonance and the time constant of the envelope detector.

Factors affecting the amplitude of the demodulated signal: Examine Figure 7 to gain an insight into the relationship between the strength of the impact, the resonance damping and the time constant of the demodulator. The top trace in Figure 7 is a simulated signal in the time domain representing the excitation of the resonance and its decay. Also shown in this time trace is the envelope of the signal produced by the envelope detector. The initial peak amplitude of the resonance is a function of the strength of the impact and the amplification factor of the resonance. If at a later date, the amplitude has increased (say from 1 to 2), then since the characteristics of the resonance are probably the same, the impact strength has doubled (the danger here is that the system is probably not nicely linear). If one compares the bearing peak amplitude for a bearing defect in one machine with a bearing in another machine however, note that the resonance characteristics such as the amplification factor, are probably not the same. In other words, if bearing one shows a peak of 1 and bearing two shows a peak of 2, the impact in bearing two is not necessarily twice that of one since the amplification factor of the resonance may be different. If the damping of the resonance is reduced, then the ring down time will be longer and the trace will spread out. The envelope decay is set not by the ring down time of the resonance but rather by the time constant of the demodulator. In the case shown, the time constant of the decay causes the envelope to be longer than the ring down time of the resonance.

The peak amplitude of the resonance, 1 in the case of Figure 7, is often used to indicate the presence of a defect in the bearing. As stated above, one bearing cannot be compared directly to a different bearing in a dıfferent machine, but on a given machine a good bearing will have a low peak amplitude and a defective bearing will have a significantly higher value. This peak amplitude is often called HFD for High Frequency Detection. The best way to measure the HFD is either by observing the signal on an oscilloscope or on a true peak reading voltmeter. Unfortunately almost none of the walk around data collectors that offer the HFD option measure it in this way. It is much easier to make a calculation in the digital world than to make analog measurements. The middle trace in Figure 7 is the spectrum of the high frequency resonance time trace. In order to calculate the RMS amplitude of a signal from the spectrum, one can simply take the square root of the sum of the squares of all spectrum bins over some frequency. The frequency selected is some frequency that will exclude the low frequency responses due to the once per rev and other low frequency signals. As shown in the center trace, the HFD calculated in this way is 0.172 rather than one. This value is not equal to 1 first because it is an RMS value and secondly because the signal in the time trace is neither a pure tone nor steady state but rather a transient signal, so even multiplying by 1.414 would not yield the peak amplitud

The third trace shows the low frequency spectrum of the envelope of the resonance as seen by the envelope detector with its decay time constant. P_6 is the amplitude of the bin that has the defect frequency fundamental. Note the presence and amplitude of the harmonics of this signal. Do these relative amplitudes have diagnostic significance? The answer is no because the relative harmonic amplitudes is set by the decay shape or time constant of the demodulator and not due to the defect or even the ring down of the resonance.

Figure 8 compares the amplitude of the defect frequency fundamental for three different time constants in the demodulator. The top trace simply shows the high frequency resonance and ring down along with the HFD calculated in the incorrect (or false) manner. The next three traces shows the time constant varying from very short to very long. Note the resonance is the same in all three cases but the amplitude of the fundamental varies by 2:1. If the decay is either too short or too long, the amplitude will be reduced. The optimum time constant is selected so that the envelope decays to zero just before the next impact. The envelope detector however cannot predict the defect frequency so it simply uses a constant value. While the amplitude of the defect signal may be compared for the same bearing in the same machine from one time to another, comparisons may not be directly made for situations where the defect frequency is significantly different or where another manufacturers envelope detector (and probably different time constant) is used.

Figure 9 compares the HFD and the defect frequency amplitude Pn_6. Note that as the damping of the resonance decreases, the ring down time increases but since the envelope time constant is at the optimum setting, the defect amplitude remains essentially constant. The HFD however changes over a range of 3:1 even though the actual peak amplitude of the impact does not change.

To summarize this section, we find that if a defect signal is observed in the demodulated spectrum, then there definitely is a defect present. It is difficult to set limits for either the HFD (true or false method of calculation) or the defect signal since these amplitudes are a function of the bearing defect **AND** the resonance amplification factor **AND** the damping **AND** the envelope detector time constant.

Software envelope detectors: Software envelope detectors may be implemented by use of the Hilbert transform or by implementing a peak detector with a fast rise time and a slower decay time constant (much as in the case of the hardware envelope detector).

The Hilbert filter or transform is a function whose output is exactly the same as the input except that its phase at all frequencies has been shifted by 90 degrees. This is shown in Figure 10. The input and output signals taken as a pair may be considered to be the real and imaginary components of a complex signal. This complex signal is called the analytic signal. In the case of a spectrum, each bin has a real and imaginary component and the magnitude is computed by taking the square root of the sum of the squares of the real and imaginary components. This same process may be done for the analytic signal where each pair of signals are at one instant in time rather than for a spectrum bin. If the input signal is a constant amplitude sine wave, then the output will be a straight line. If the input signal is modulated, then the output from this process will be the envelope of the modulation. This is shown in Figure 11.

It is easy to see that this process will work so long as the input is a well defined signal such as generated by a function generator as in Figure 11 or the gear mesh signal from a gear box. (Applying the Hilbert transform to gear box signals is an entirely different topic. A good explanation of the Hilbert envelope detector and its application to gear boxes is given in reference [6]). How would it work on the resonance excited by a defective bearing. Figure 12 shows the result of this application. The envelope is clearly shown in the center trace and the spectrum of the envelope is shown in the bottom trace. In the case of the data used to generate Figure 12, the data included only one low frequency resonance. In real life, when the resonance is at a much higher frequency, the spectrum of the envelope should be zoomed to the lower frequency portion in order to accurately identify the defect frequency. I have not seen this technique explored for complex cases where several resonances are excited at once. One way of limiting the analysis to a single resonance is to first zoom on one of the resonances, then perform the Hilbert envelope detector on the zoomed time domain signal and then display the envelope spectrum. This is shown in Figure 13. The disadvantage of this zoom process is that the time trace (actually a complex signal) does not look like the modulation. A big advantage of the Hilbert envelope detector is that there is no envelope detector time constant to modify the data. The envelope faithfully follows the shape of the time domain resonance.

A simple peak detector may be implemented in software. The original signal, the envelope of this signal and the spectrum of the envelope are shown in Figure 14. The advantage of this approach is that the envelope will be produced no matter how many resonant frequencies make up the time trace. The disadvantage is that we have reintroduced the envelope detector time constant as a variable. For high frequency resonances, the spectrum of the envelope should be zoomed to the lower frequency region.

Conclusions: By using the envelope detector technology, one can see that the sensitivity to the presence of a small initial bearing defect is many times what it is by observing only the low frequency vibration or sonic data. This added sensitivity allows the detection of the onset of surface defects in rolling element bearings. This advantage is partially offset by the fact that the HFD and defect signal amplitudes are functions of structural and envelope detector characteristics in addition to the size of the defect. One possible approach would be to monitor the envelope spectrum for the presence of defect frequencies and once detected, expand the monitoring to looking for these defect signals in the low frequency spectrum. In any case, determining exactly when to remove a bearing just prior to catastrophic failure remains a crap shoot.

More work should be done investigating the Hilbert envelope detector to very complicated signals made up of several resonances, including the accelerometer resonance. The data reported here for Hilbert envelope detectors is just an initial look at how the Hilbert envelope detector responds to typical bearing defect signals.

References:

1. Balderston, H.L., "The Detection of Incipient Failures in Bearings", Presented at the 28th National Fall Conference of the American Society for Nondestructive Testing, Detroit, Mich. Oct 14-17, 1968

2. Broderick, J.J, Burchill, R.F. and Clark, H.L., "Design and Fabrication of Prototype System for Early Warning of Impending Bearing Failure", Design Report MTI-71TR to NASA under contract NAS8-25706, January 1972

3. Burchill, R.F. "Resonant Structure Technique for Bearing Fault Analysis", Presented at the 18th Meeting of the MFPG, Gathersburg, MD 1972

4. Broderick, J.J., Burchill, R.F. and Clark, H.L. "A system for Early Warning of Bearing Failures", NASA Tech Brief B72-10494, August, 1972

5. Burchill, R.F. and Frarey, J.L. "New Detection Method for Rolling Element and Bearing Defects", NASA Tech Brief B72-10689, October, 1972

6. McFadden, P.D., "Detecting Fatigue Cracks in Gears by Amplitude and Phase Demodulation of the Meshing Frequency", Transactions of the ASME Journal of Vibration, Acoustics, Stress and Reliability in Design, April 1986, Vol. 108

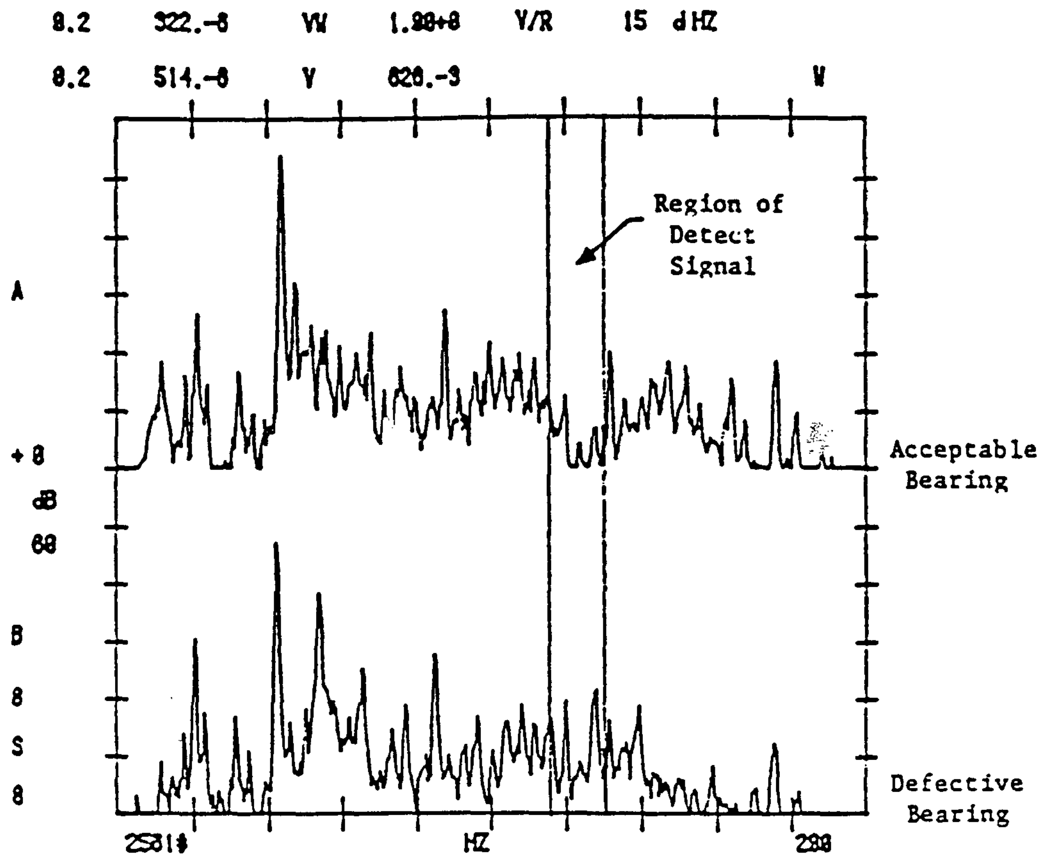

Figure 1 Comparison of the vibration spectrum for a damaged and undamaged bearing.

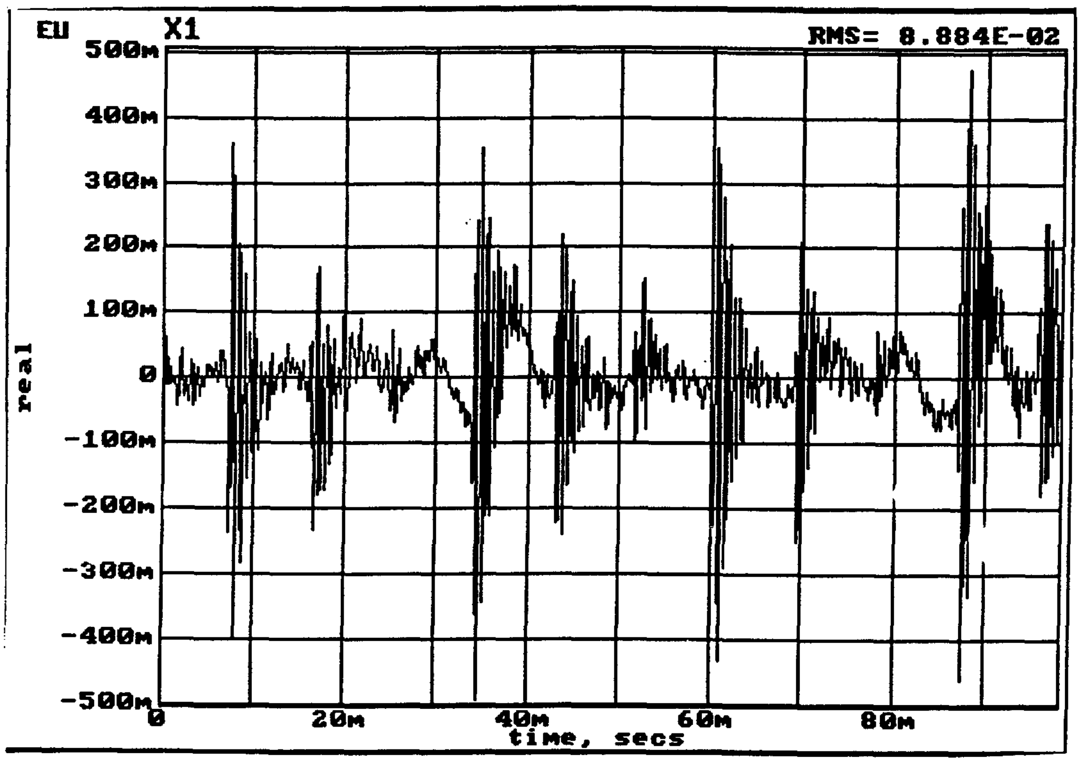

Figure 2 Time domain signal from a damaged bearing.

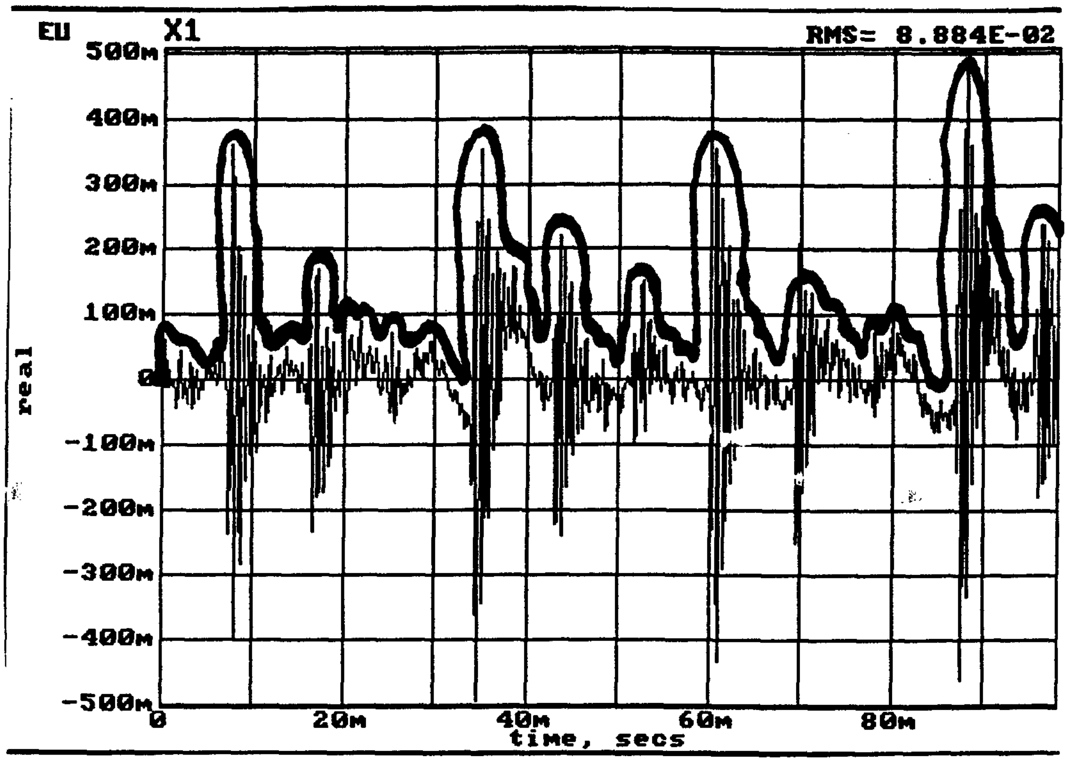

Figure 3 Time trace of damaged bearing with envelope drawn.

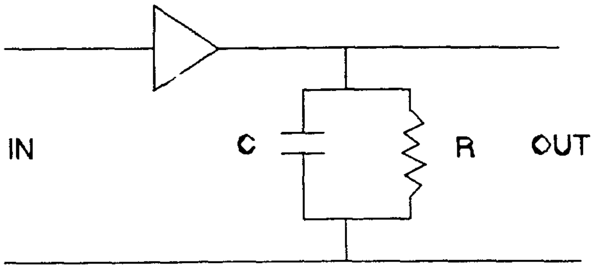

Figure 4 Simple Demodulator

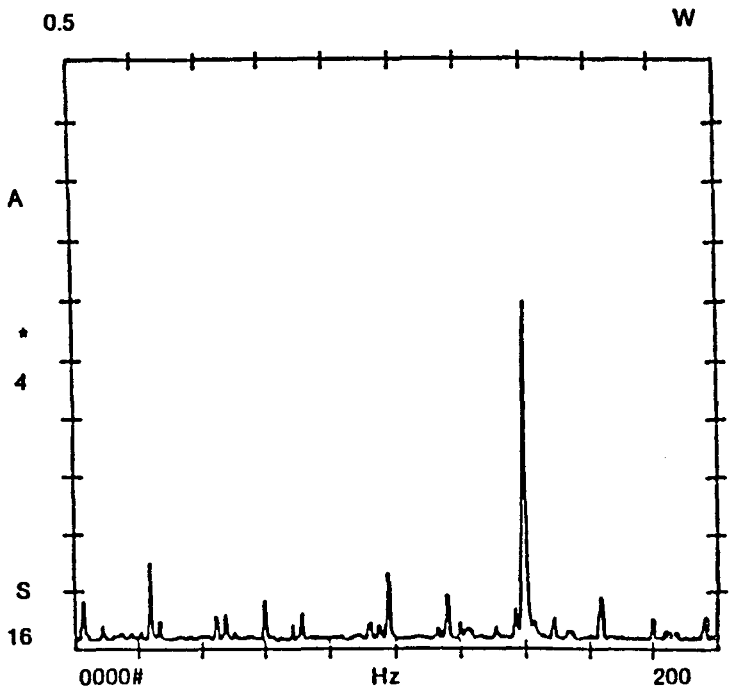

Figure 5 Spectrum of the enveloped data from the same defective bearing as shown in Figure 1.

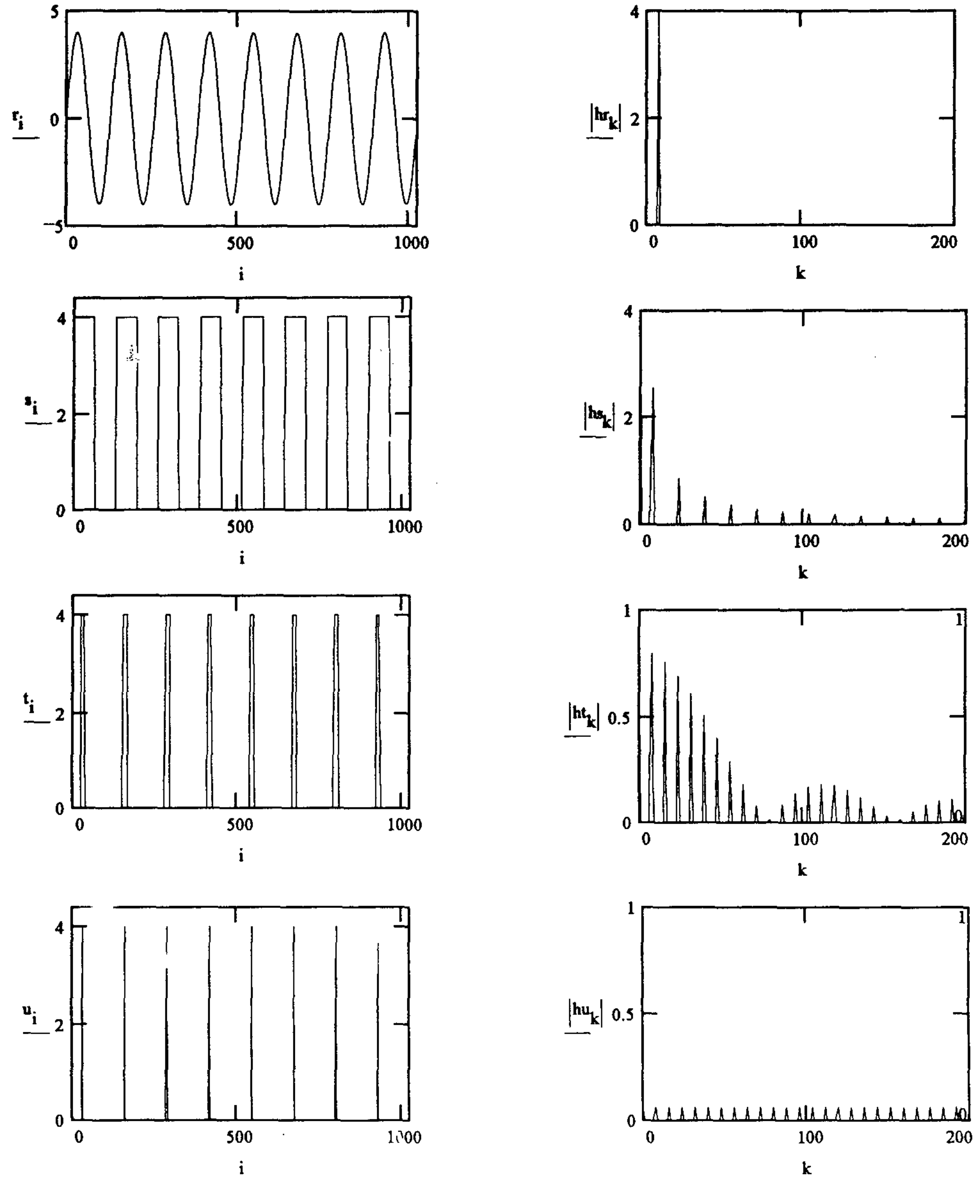

Figure 6 Spectrum energy distribution as a function of pulse shape

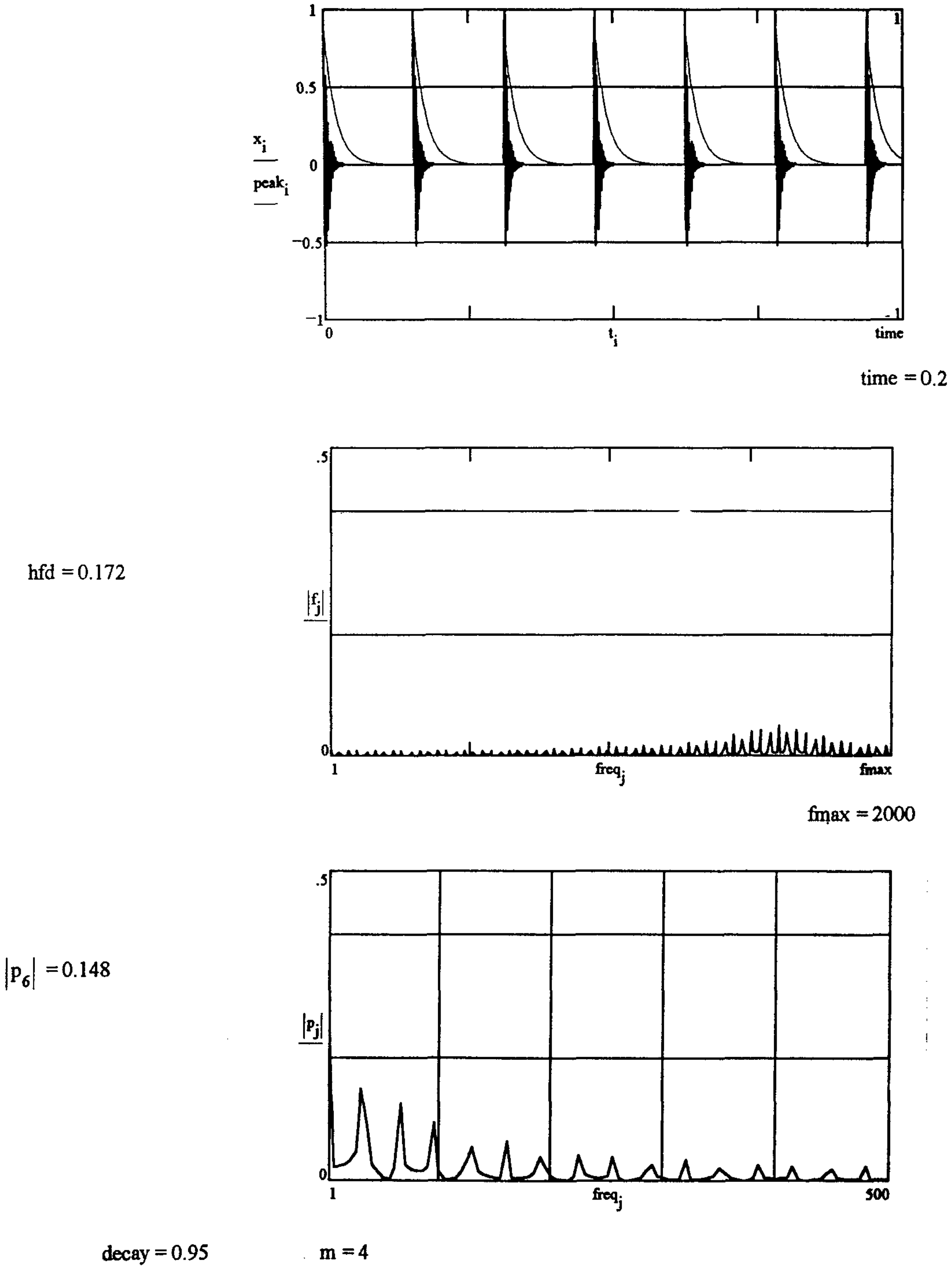

Figure 7. Relationship between HFD (true & false), resonance damping and demodulator time constant

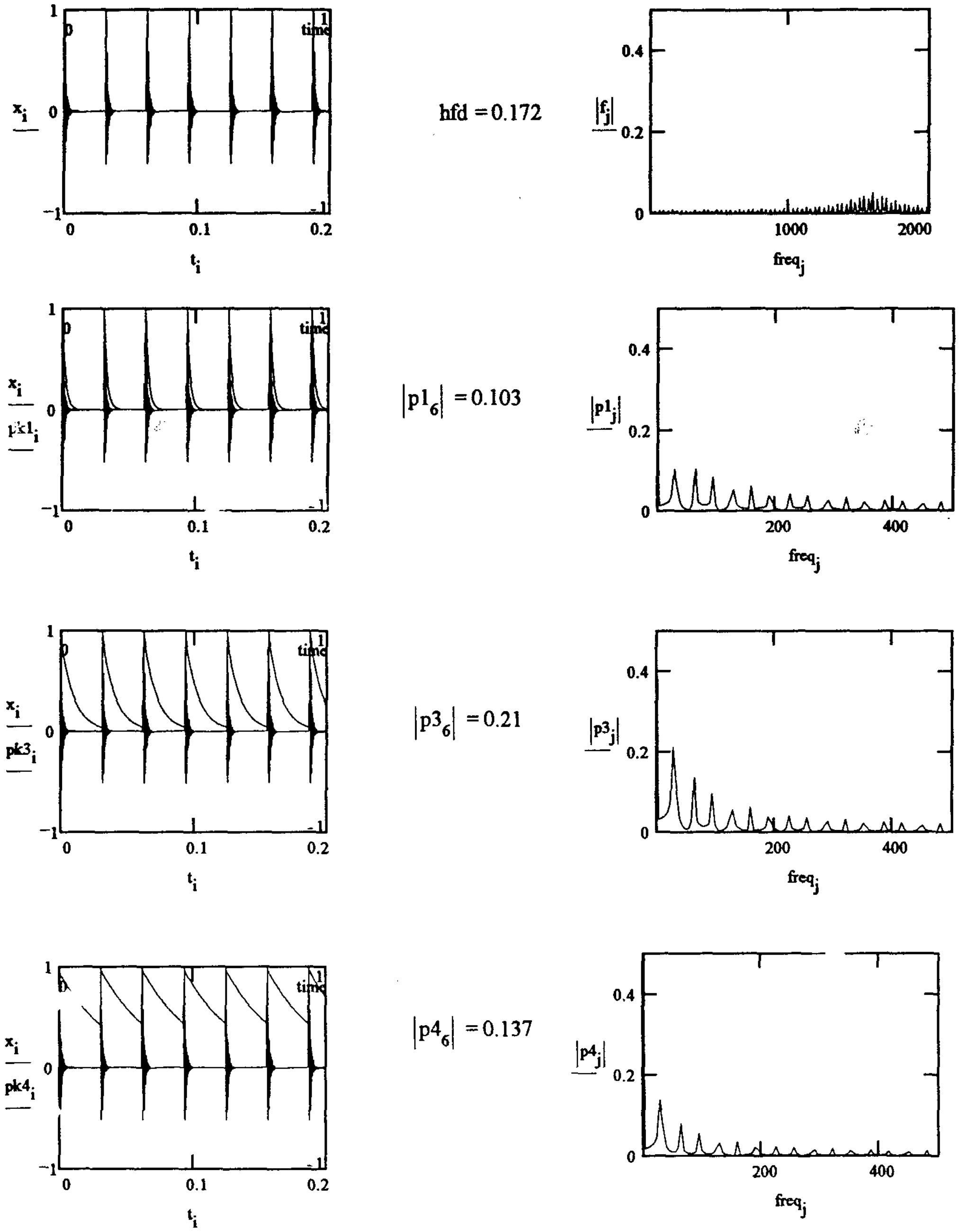

Figure 8. Relationship between the amplitude of the defect spectrum peak and the demodulator time constant

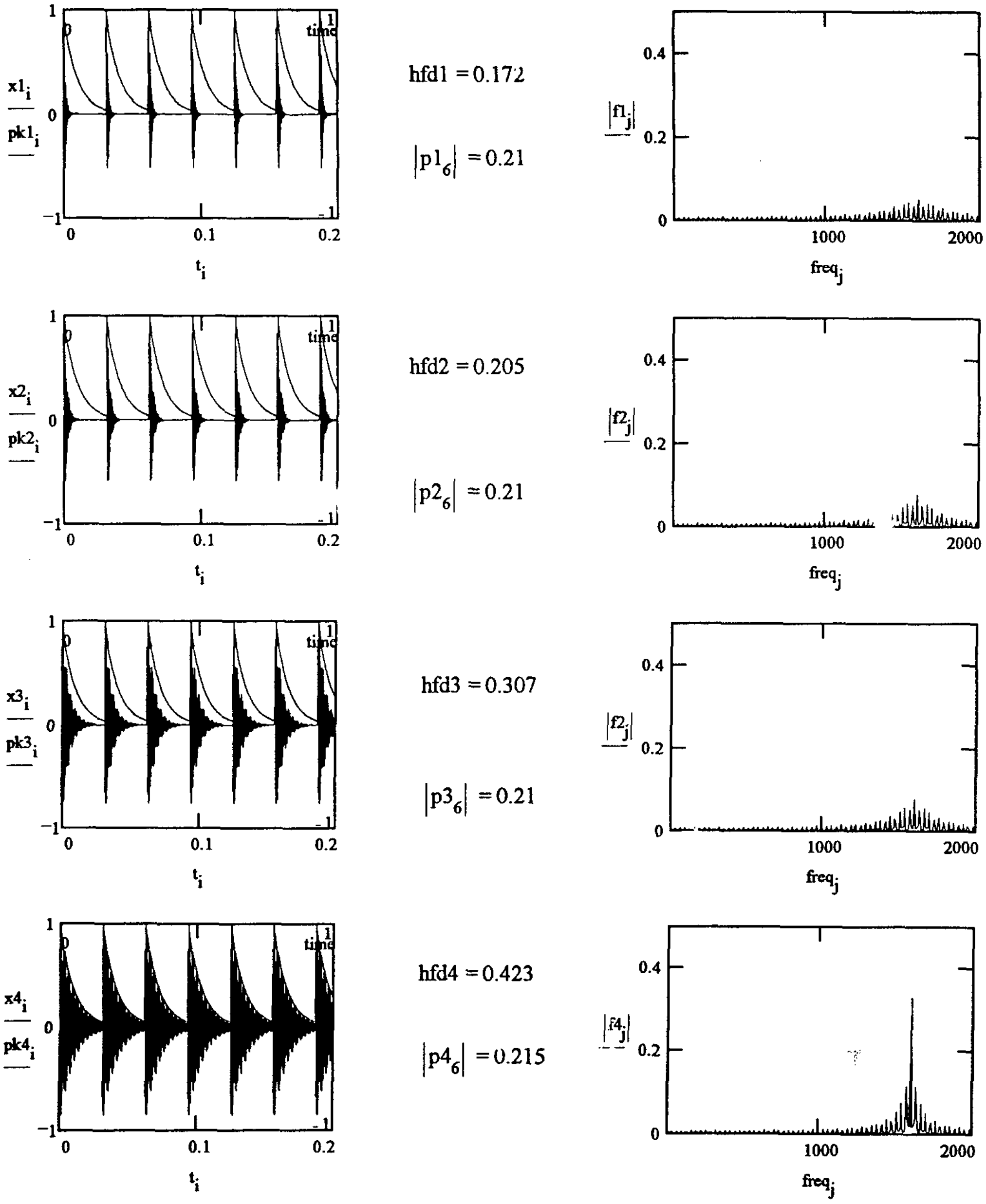

Figure 9. Relationship between the hfd(spec) and the resonance damping

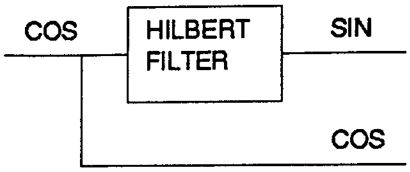

Figure 10 Hilbert filter and analytic signal.

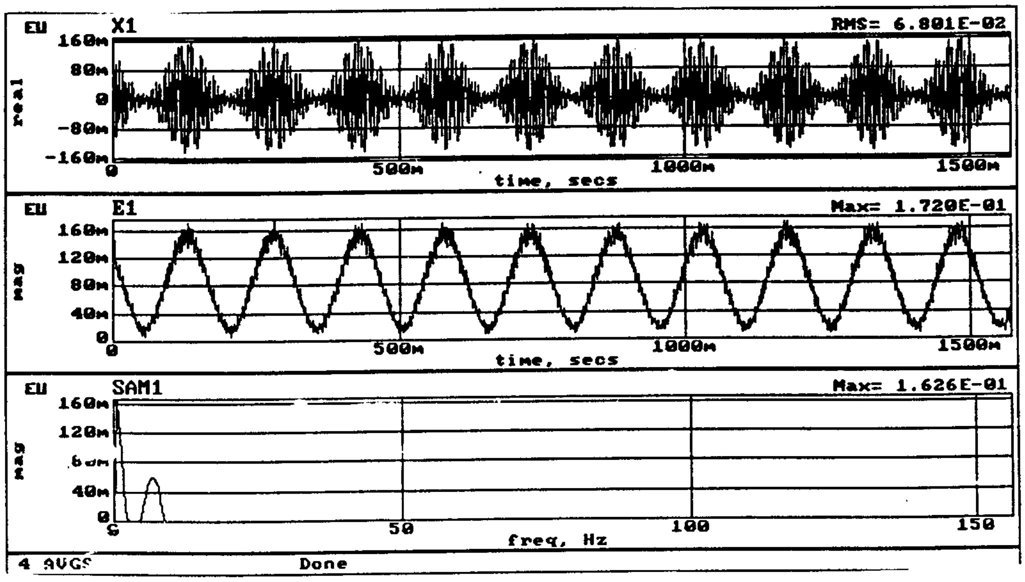

Figure 11 Demodulation by Hilbert transform of a modulated sine wave.

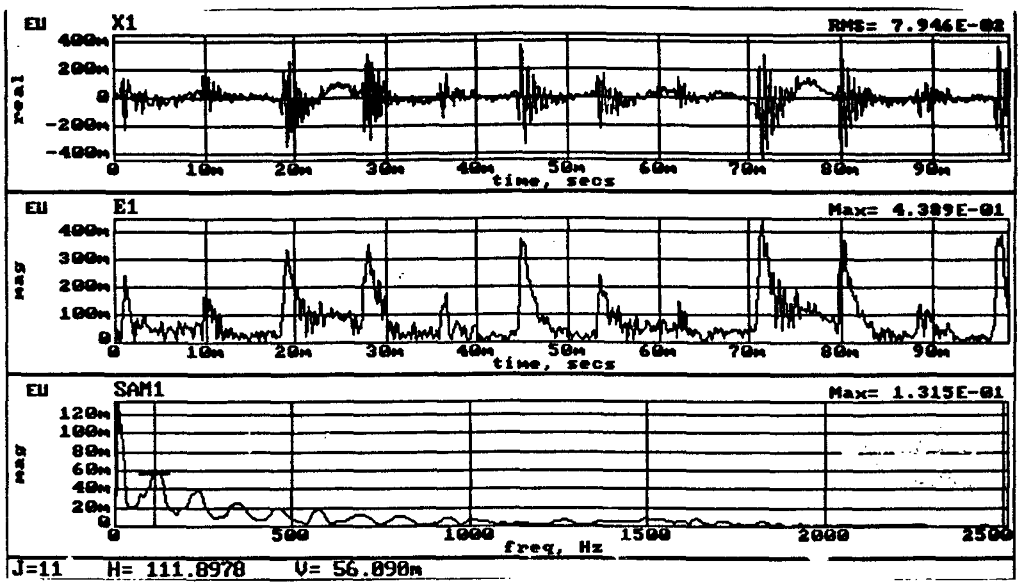

Figure 12 Envelope and spectrum of the envelope for a defective bearing using the Hilbert envelope detector.

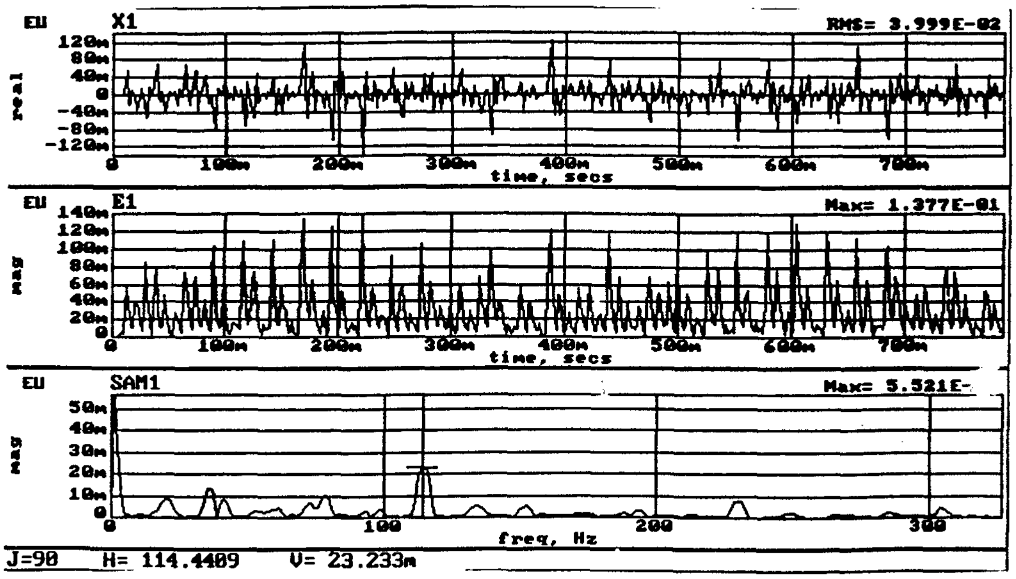

Figure 13 Spectrum formed by zooming on the resonance, then producing the spectrum using the Hilbert envelope detector.

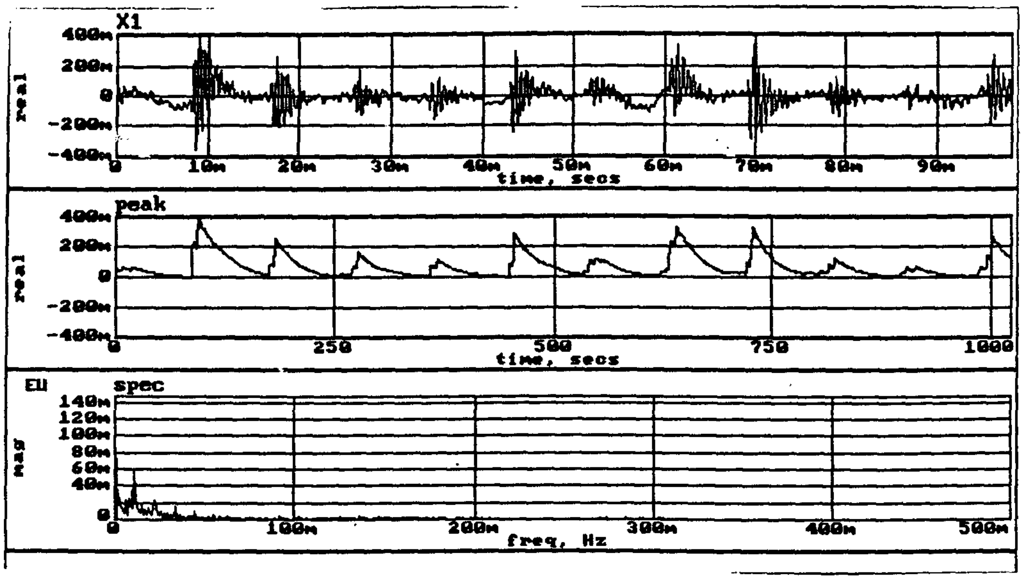

Figure 14 Peak detector used as an envelope detector.