

**THE APPLICATION OF SIGNAL PROCESSING
AND ARTIFICIAL INTELLIGENCE TECHNIQUES IN THE
CONDITION MONITORING OF ROTATING MACHINERY**

N. T. van der Merwe

**Thesis submitted for the degree Philosophiae Doctor
in Electrical and Electronic Engineering
at Potchefstroom University for Christian Higher Education**

Promoter: Prof. Dr A. J. Hoffman

Potchefstroom

2003

Abstract

Condition monitoring of critical machinery has many economic benefits. The primary objective is to detect faults, for example on rolling element bearings, at an early stage to take corrective action prior to the catastrophic failure of a component. In this context, it is important to be able to discriminate between stable and deteriorating fault conditions. A number of conventional vibration analysis techniques exist by which certain faults in rotating machinery may be identified. However, under circumstances involving multiple fault conditions conventional condition monitoring techniques may fail, e.g. by indicating deteriorating fault conditions for stable fault situations or vice versa. Condition monitoring of rotating machinery that may have multiple, possibly simultaneous, fault conditions is investigated in this thesis. Different combinations of interacting fault conditions are studied both through experimental methods and simulated models. Novel signal processing techniques (such as cepstral analysis and equidistant Fourier transforms) and pattern recognition techniques (based on the nearest neighbour algorithm) are applied to vibration problems of this nature. A set of signal processing and pattern recognition techniques is developed for the detection of small incipient mechanical faults in the presence of noise and dynamic load (imbalance). In the case investigated the dynamic loading consisted of varying degrees of imbalance. It is demonstrated that the proposed techniques may be applied successfully to the detection of multiple fault conditions.

Keywords:

Condition monitoring, fault detection, rotating machinery, rolling element bearings, artificial intelligence, pattern recognition, neural networks, multiple simultaneous fault conditions, static and dynamic loading, bearing and imbalance fault conditions

Opsomming

Die kondisiemonitering van kritiese masjinerie het vele ekonomiese voordele. Die belangrikste doelstelling is om foute, soos byvoorbeeld op laers, vroegtydig op te spoor sodat aksie geneem kan word voor 'n katastrofiese faling van 'n komponent. Teen hierdie agtergrond is dit belangrik om te kan onderskei tussen stabiele en verslegtende fouttoestande. 'n Aantal konvensionele vibrasie-analise tegnieke bestaan waarmee sekere foute in roterende masjinerie geïdentifiseer kan word. In die geval van veelvuldige fouttoestande mag konvensionele kondisiemoniteringstegnieke faal, byvoorbeeld deur verslegtende fouttoestande aan te dui vir stabiele fouttoestande of omgekeerd. Hierdie tesis ondersoek die kondisiemonitering van roterende masjinerie wat veelvuldige, selfs gelyktydige, fouttoestande ondervind. Verskillende kombinasies van gekoppelde fouttoestande word bestudeer beide deur eksperimentele metodes en gesimuleerde modelle. Nuwe seinverwerkingstegnieke (soos 'cepstral analysis' en 'equidistant Fourier transforms') en patroonherkenningstegnieke (gebaseer op die naastebuurman-algoritme) word toegepas op sulke vibrasieprobleme. Seinverwerkings- en patroonherkenningstegnieke word ontwikkel om 'n klein wordende meganiese fout in die teenwoordigheid van ruis en dinamiese las (onbalans) op te spoor. Daar word aangetoon dat die voorgestelde tegnieke suksesvol toegepas kan word in die opsporing van veelvuldige fouttoestande.

Sleutelwoorde

Kondisiemonitering, foutopsporing, roterende masjinerie, laers, kunsmatige intelligensie, patroonherkenning, neurale netwerke, veelvuldige gelyktydige fouttoestande, statiese en dinamiese las, laer en onbalans fouttoestande

Afrikaanse titel

DIE TOEPASSING VAN SEINVERWERKINGSTEGNIEKE EN KUNSMATIGE INTELLIGENSIE IN DIE KONDISIEMONITERING VAN ROTERENDE MASJINERIE

TABLE OF CONTENTS

1	Introduction	1
1.1	Overview of condition monitoring	1
1.2	Background	2
1.3	Outline of this thesis	8
2	Literature survey	11
2.1	Introduction	11
2.2	Condition monitoring	12
2.3	Techniques for fault detection	12
2.3.1	<i>Data collection</i>	12
2.3.2	<i>Linear and non-linear processing techniques</i>	13
2.3.3	<i>Feature selection and extraction</i>	13
2.3.3.1	<i>Weibull distribution</i>	14
2.3.3.2	<i>Wavelet transform</i>	14
2.3.3.3	<i>Cepstrum analysis</i>	14
2.3.3.4	<i>Envelope spectrum analysis</i>	14
2.3.4	<i>Pattern recognition</i>	15
2.3.5	<i>Experimental design</i>	16

2.3.6	<i>Signal processing techniques</i>	16
2.3.7	<i>Non-parametric classification</i>	17
2.3.8	<i>Artificial neural networks (MLP)</i>	18
2.4	Vibration metrics	20
2.4.1	<i>Time domain features</i>	21
2.4.2	<i>Frequency domain features</i>	22
2.4.3	<i>Rotational frequency</i>	22
2.4.4	<i>Ball pass frequency of outer race</i>	23
2.4.5	<i>Higher frequency domain components</i>	24
2.4.6	<i>High frequency resonance technique (HFRT)</i>	25
2.4.7	<i>Cepstrum analysis</i>	26
2.5	Fault detection on roller bearings	26
2.5.1	<i>Modelling of roller bearing faults</i>	31
2.5.2	<i>Influence of static load</i>	36
2.5.3	<i>Influence of dynamic load</i>	40
3	Problem statement and contributions of thesis	43
3.1	Introduction	43
3.2	Data set size limitations	45
3.3	Model of bearing defects	47
3.4	Obtaining a normal bearing signal	50
3.5	Problematic nature of the recognition of multiple bearing defects	51

3.6	Problem statement	53
3.7	Contributions	55
4	Experimental setup	59
4.1	Introduction	59
4.2	Influence of dynamic load on bearing fault detection	60
4.3	Laboratory measurements of Vib demo data sets 1 and 2	67
4.4	Laboratory measurements of Vib demo data set 3	69
4.5	Recommendation for bearing fault detection	71
4.6	Conclusion	72
5	Theoretical development and results of pattern recognition techniques	75
5.1	Introduction	75
5.2	Background	75
5.3	Theoretical development	76
5.3.1	<i>Mutual information (MI) based feature selection</i>	76
5.3.2	<i>Improved cepstral analysis</i>	77
5.3.3	<i>Modelling of RMS level of bearing defects</i>	78
5.3.4	<i>Composing validation sets</i>	78
5.4	Results	80
5.4.1	<i>A nearest neighbour rule with class membership (NNRC) for modelling problems</i>	80
5.4.1.1	<i>Introduction</i>	80

5.4.1.2	<i>Problem formulation</i>	81
5.4.1.3	<i>NNRC algorithm</i>	81
5.4.1.3.1	<i>Nearest neighbour rule viewed as a neural network</i>	81
5.4.1.3.2	<i>Nearest neighbour rule as linear interpolator</i>	83
5.4.1.3.3	<i>Nearest neighbour rule with class membership (NNRC)</i>	85
5.4.1.3.4	<i>Gamma test and fractional NNR</i>	87
5.4.1.3.5	<i>Results</i>	88
5.4.1.4	<i>Conclusions</i>	90
5.4.1.5	<i>Future work</i>	90
5.4.2	<i>Developing an efficient cross validation strategy to determine classifier performance (CVCP)</i>	91
5.4.2.1	<i>Introduction</i>	91
5.4.2.2	<i>Problem formulation</i>	93
5.4.2.3	<i>The nearest neighbour rule and validation or classification</i>	95
5.4.2.3.1	<i>Nearest neighbour rule as zero order hold</i>	95
5.4.2.3.2	<i>Validation as fractional NNR</i>	98
5.4.2.3.3	<i>NNR measures the complexity of the output surface</i>	100
5.4.2.3.4	<i>Validation and model verification</i>	101
5.4.2.4	<i>Results</i>	103
5.4.3	<i>Conclusions</i>	111

5.5	Future work	112
5.6	Conclusion	113
6	Results of signal processing techniques	115
6.1	Introduction	115
6.2	Problematic nature of separating the normal and defect vibration signals	116
6.3	Synchronous averaging and demodulation spectrum	117
6.4	Relation between SA on the time and frequency domains (using ESFFT)	123
6.5	Effect of time domain window on the SA	128
6.5.1	<i>Introduction</i>	128
6.5.2	<i>Sidelobe level for specified defect frequency</i>	131
6.5.3	<i>Conclusion</i>	132
6.6	Modelling of speed variations of a vibration signal	132
6.7	Frequency domain modelling of speed variations of a vibration signal	138
6.8	Relation between SA and demodulation spectrum	139
6.9	Kurtosis as fault diagnostic	145
6.9.1	<i>Introduction</i>	145
6.9.2	<i>Maximization of kurtosis (leptokurtic)</i>	145
6.9.3	<i>Minimization of kurtosis (platykurtic)</i>	149
6.9.4	<i>Kurtosis of periodic signal</i>	152

6.9.5	<i>Interpretation of kurtosis in time domain</i>	154
6.9.6	<i>Interpretation of kurtosis in frequency domain</i>	157
6.9.7	<i>Additional interpretation of kurtosis</i>	159
6.9.8	<i>Kurtosis at specific frequency</i>	162
6.9.9	<i>Conclusion</i>	163
6.10	Effect of a non-linear operator on a simulated fault signal	163
6.10.1	<i>Introduction</i>	163
6.10.2	<i>Theory</i>	164
6.10.3	<i>Frequency domain interpretation</i>	171
6.10.3.1	<i>Derivation of an analytical frequency domain expression for non-linear operator with additive noise is problematic</i>	171
6.10.3.2	<i>Heuristic empirical procedure</i>	173
6.10.3.3	<i>Heuristic formula for $N_2(1)$</i>	174
6.10.3.4	<i>Heuristic formula for $N_2(2f_c)$</i>	175
6.10.3.5	<i>Discussion and conclusion</i>	176
6.10.4	<i>Extension to arbitrary signals</i>	178
6.10.5	<i>Amplitude variation</i>	179
6.10.6	<i>Matched filter detection</i>	182
6.10.7	<i>Amplitude modulation in the context of vibration monitoring</i>	183
6.10.8	<i>Conclusion</i>	186

6.11	Amplitude demodulation – relation between the Hilbert transform and square law modulator	186
6.11.1	<i>Spectrum of the square law modulator</i>	186
6.11.2	<i>Spectrum of the absolute value of the analytical signal (derived from the Hilbert transform)</i>	187
6.11.3	<i>Comparison of the spectrum of the absolute value of the analytical signal and the square law modulator</i>	188
6.11.4	<i>Relation between the demodulation spectrum of the positive and negative frequency components</i>	190
6.11.5	<i>Conclusion</i>	191
6.12	Modified cepstral analysis	192
6.13	Applying digital signal processing techniques to improve the signal-to-noise ratio in vibration signals	197
6.13.1	<i>Introduction</i>	197
6.13.2	<i>Problem formulation</i>	198
6.13.2.1	<i>Modelling of multiple fault mechanisms</i>	198
6.13.2.1.1	<i>Modelling of bearing defects (including size of defect)</i>	198
6.13.2.1.2	<i>Modelling of bearing defects (multiple defects)</i>	200
6.13.2.1.3	<i>Influence of environmental conditions</i>	202
6.13.2.1.4	<i>Multiple fault conditions</i>	203
6.13.2.2	<i>Signal processing techniques</i>	203

6.13.2.2.1	Spectral analysis	203
6.13.2.2.2	Time domain analysis	203
6.13.2.2.3	High frequency resonance technique (HFRT)	203
6.13.3	<i>Problem solution</i>	204
6.13.3.1	<i>Influence of the buffer length on demodulation spectrum</i>	204
6.13.3.2	<i>Effect of speed variation on demodulation peak</i>	204
6.13.3.3	<i>Time domain speed compensation</i>	205
6.13.4	<i>Conclusion</i>	209
6.14	Conclusion	210
7	Results of bearing test rig	213
7.1	Introduction	213
7.2	Experimental results of Vib demo data sets 1 and 2	213
7.2.1	<i>Introduction</i>	213
7.2.2	<i>Feature extraction</i>	213
7.2.3	<i>Modelling of Vib demo data set 2</i>	215
7.3	Experimental results of Vib demo data set 3	223
7.3.1	<i>Introduction</i>	223
7.3.2	<i>Application of defect angle (DA)</i>	224
7.3.3	<i>Feature BPFO2_X</i>	224
7.3.4	<i>Feature BPFO_X</i>	228
7.3.5	<i>Cepstral features</i>	228

7.3.6	<i>Feature FFT2_MEAN_X</i>	229
7.3.7	<i>Feature SA_BPFO2_0_X</i>	231
7.3.8	<i>Feature BPFO2_X – influence of buffer length</i>	235
7.3.9	<i>Feature HFD_2_5_kHz</i>	243
7.3.10	<i>Feature KHX_X</i>	244
7.3.11	<i>Feature BPF12_X</i>	246
7.3.12	<i>Feature Fr_X</i>	247
7.4	Future work – position of outer race defect	249
7.5	Conclusion	250
8	Conclusion	253
8.1	Introduction	253
8.2	Bearing fault detection	253
8.3	Signal processing techniques	256
8.4	Modelling of fault signals using pattern recognition techniques	259
8.5	Applying pattern recognition techniques successfully in condition-monitoring problems	262
8.6	Future work	263
8.6.1	<i>Theoretical aspects (pattern recognition and signal processing)</i>	263
8.6.1.1	<i>Feature extraction or selection</i>	263
8.6.1.2	<i>Signal processing techniques</i>	264
8.6.1.3	<i>Pattern recognition techniques</i>	265

8.6.1.4	<i>Remaining theoretical issues</i>	265
8.6.2	<i>Experimental work and modelling</i>	265
8.6.2.1	<i>Experimental work</i>	265
8.6.2.1.1	<i>DA as tool for bearing fault detection</i>	265
8.6.2.1.2	<i>Influence of sensor technology on fault detection</i>	266
8.6.2.1.3	<i>Test rig</i>	267
8.6.2.2	<i>Modelling of defects</i>	269
9	Bibliography	271
10	List of publications	283
11	Appendices	287
11.1	Appendix A: Influence of phase shift on envelope spectrum	287
11.2	Appendix B: Results on Vib demo data set 2 (figures)	291
11.2.1	<i>Feature extraction (Vib demo data set 2)</i>	291
11.2.2	<i>Classifier</i>	306
11.3	Appendix C: Results on Vib demo data set 3 (figures)	313
11.3.1	<i>Feature extraction</i>	313
11.3.1.1	<i>Outer race defects (small, medium and large together)</i>	313
11.3.1.2	<i>Large outer race defect</i>	327
11.3.1.3	<i>Medium outer race defect</i>	341
11.3.1.4	<i>Small outer race defect</i>	355

11.4	Appendix D: SEM analysis of bearing defects	371
11.5	Appendix E: Influence of periodic speed variations	379
11.6	Appendix F: Some condition monitoring data sets in literature for fault identification	381
11.7	Appendix G: Derivation of amplitude of sidelobe levels relative to main lobe	383
11.8	Appendix H: Derivation of amplitude of demodulation spectrum	385
11.8.1	<i>Example 1</i>	385
11.8.2	<i>Example 2</i>	386
11.8.3	<i>Example 3</i>	388
11.9	Appendix I: Experimental results on the effect of a time domain window on SA	391
11.10	Appendix J: Recommendations for the application of pattern recognition techniques in condition-monitoring problems	401
11.11	Appendix K: Examples of simulated bearing signals and the application of a non-linear operator	409
11.11.1	<i>Introduction</i>	409
11.11.2	<i>No noise added, speed variation = 0%</i>	410
11.11.3	<i>No noise added, speed variation = 20%</i>	414
11.11.4	<i>Noise with normal distribution added, speed variation = 20%</i>	418
11.11.5	<i>Noise added, speed variation = 0%</i>	421
11.11.6	<i>More noise added, speed variation = 0%</i>	425

11.11.7	<i>No noise added, linear chirp – speed variation = 0.4084%</i>	429
11.11.8	<i>No noise added, linear chirp – speed variation = 62%</i>	446
11.11.8.1	<i>Results</i>	446
11.11.8.2	<i>Theoretical analysis</i>	448
11.11.8.3	<i>Simplified example of sinusoidal signal with additive noise</i>	450
11.11.9	<i>Simplified vibration model</i>	458
11.11.9.1	<i>Introduction</i>	458
11.11.9.2	<i>Simplified vibration model</i>	458
11.11.9.3	<i>Detection ratio of the simplified vibration model</i>	459
11.11.9.4	<i>Speed variation function</i>	470
11.11.9.5	<i>Results</i>	470
11.11.9.5.1	<i>Small spread of pulse in time domain, no speed variation, $\sigma_n = 2$, $\sigma_x = 0.36$, $p = 4$</i>	470
11.11.9.5.2	<i>Small spread of pulse in time domain, no speed variation, $\sigma_n = 0.1$, $\sigma_x = 0.18$, $p = 256$</i>	474
11.11.9.5.3	<i>Small spread of pulse in time domain, no speed variation, $\sigma_n = 0.1$, $\sigma_x = 0.0477$, $p = 65\ 536$</i>	479
11.11.9.5.4	<i>Small spread of pulse in time domain, no speed variation, $\sigma_n = 20$, $\sigma_x = 3.6455$, $p = 4$, amplitude of sine wave = 10</i>	484

<i>11.11.9.5.5 Multiplicative noise ($\sigma_n = 1$)</i>	488
<i>11.11.9.5.6 Multiplicative noise ($\sigma_n = 5$)</i>	490
<i>11.11.9.5.7 Small spread of pulse in time domain, large speed variation</i>	491
<i>11.11.9.5.8 Large spread of pulse in time domain, large speed variation</i>	494
<i>11.11.9.5.9 Small bearing defect signal ($DA = 270^\circ, 0g$)</i>	496
<i>11.11.9.6 Conclusion</i>	498

LIST OF FIGURES

Figure 1:	Increasing imbalance on channel 1 (0–1kHz) on Vib demo data set 2	23
Figure 2:	Increasing imbalance on channel 1 (0–10kHz) on Vib demo data set 2	25
Figure 3:	Effect of dynamic load on load distribution (drawing is not to scale)	63
Figure 4:	Shift in position of loading zone owing to the pressure plate in the Vib demo test rig	66
Figure 5:	Test rig	68
Figure 6:	Visualization of distance metric $A(X, Y)$	82
Figure 7:	Response of NNRC classifier trained on all six classes (training set samples shown in black)	89
Figure 8:	Response of RBF classifier trained on all six classes (training set samples shown in black)	90
Figure 9:	Output surface corresponding to Voronoi tessellation in [76]; the grey values correspond to class 1 and the white values to class 2 (samples are arbitrarily assigned to classes 1 and 2); the data samples are indicated by four black circles	96
Figure 10:	Scatterplot of normalized features	103
Figure 11:	Response of neural network evaluated on grid	104
Figure 12:	Response of NNR evaluated on grid (NNR output for Iris-setosa indicated by a red triangle, Iris-versicolor by a blue	

	circle and Iris-virginica by a magenta cross; training set data of the NNR is shown by a black circle, black dot and green diamond respectively	105
Figure 13:	Response of neural network evaluated on grid with two neurons in single hidden layer (trainlm, 1600 epochs)	106
Figure 14:	Response of neural network evaluated on grid with 30 neurons in single hidden layer (trainlm, 1600 epochs)	106
Figure 15:	Accuracy of LOO CV on training set	108
Figure 16:	Accuracy of CVCP when applying LOO CV	109
Figure 17:	Accuracy obtained on test set for ANN	109
Figure 18:	Outputs of RBF for normal bearing 0g, 12g and 24g (RF from 0 to 7)	110
Figure 19:	Output of RBF neural network on normal bearing data for 0g (•), 12g (+) and 24g (◊) imbalance respectively	111
Figure 20:	Output of RBF neural network on defective bearing data for 0g (•), 12g (+) and 24g (◊) imbalance respectively	111
Figure 21:	Time domain signals plotted over each other for one period (STD of noise = 0, speed variation = 20%)	169
Figure 22:	Time domain signal of simulated Gaussian pulse with zero speed variation and amplitude variation	180
Figure 23:	Amplitude spectrum of simulated Gaussian pulse with zero speed variation and amplitude variation	181
Figure 24:	Amplitude spectrum of square of simulated Gaussian pulse with zero speed variation and amplitude variation	181
Figure 25:	The convolution of $A^-(f)$ and $A^+(f)$	189

Figure 26:	Gaussian pulse sequence centred at 8.8kHz (fractional bandwidth = 0.035, timing variation = 0%)	193
Figure 27:	Real cepstrum of figure 26 (no speed variation); horizontal axis in quefrency (seconds) and vertical axis in gamnitude; defect quefrency is $\approx 1/(114\text{Hz}) = 0.009$ seconds)	193
Figure 28:	Demodulation spectrum of figure 26 (no speed variation)	194
Figure 29:	Absolute value of log of magnitude spectrum of figure 26 (no speed variation)	194
Figure 30:	LTCEPS of figure 26 (bandwidth = 0.035, threshold = 1%, speed variation = 0%); horizontal axis in quefrency (seconds)	195
Figure 31:	Log of magnitude spectrum (bandwidth = 0.035, threshold = 1%, speed variation = 1%)	195
Figure 32:	LTCEPS (buffer length = 8000, bandwidth = 0.035, threshold = 20%, speed variation = 1%); horizontal axis in quefrency (seconds); defect quefrency is $\approx 1/(114\text{Hz}) = 0.009$ seconds)	196
Figure 33:	LTCEPS (buffer length = 2000, bandwidth = 0.035, threshold = 20%, speed variation = 1%); horizontal axis in quefrency (seconds); defect quefrency is $\approx 1/(114\text{Hz}) = 0.009$ seconds)	196
Figure 34:	Ratio of demodulation peak relative to carpet level (vertical axis in dB) versus length of time domain buffer (horizontal axis in samples)	198
Figure 35:	Compensated time domain signal (horizontal axis in samples)	200

Figure 36:	Magnitude spectrum uncompensated (fractional bandwidth of 0.035 and 1% random speed variation)	205
Figure 37:	Magnitude spectrum compensated (fractional bandwidth of 0.035 and 1% random speed variation)	205
Figure 38:	Plot of energy (square) of time domain signal, bandpass filtered around 8.5kHz (horizontal axis in samples)	208
Figure 39:	Experimental data set uncompensated	208
Figure 40:	Experimental data set compensated	208
Figure 41:	Ratio of demodulation peak relative to carpet level (in dB) versus length of time domain buffer (horizontal axis in samples) [53]	209
Figure 42:	High frequency spectrum (horizontal axis in Hz) [53]	209
Figure 43:	Numbering convention used to indicate proximity of defect angle to centre of load zone	214
Figure 44:	Scatterplot of rotational frequency (RF) features (x and y axes) – Vib demo data set 2	218
Figure 45:	Scatterplot of BPOR feature (x and y axes) – Vib demo data set 2	219
Figure 46:	Schematic representation of feature matrix for the SOM technique – Vib demo data set 2 (refer to [98] for details)	219
Figure 47:	The SOM displaying all four features – Vib demo data set 2 (refer to [98] for details)	220
Figure 48:	Class labels and best matching unit trajectory on the SOM – Vib demo data set 2 (refer to [98] for details)	220

Figure 49:	Outputs of RBF for all six classes on own training set data; bearing defect is shown on the left hand axis, imbalance on the right hand axis – Vib demo data set 2	221
Figure 50:	Outputs of RBF for 18g mass with no bearing defect; bearing defect is shown on the left hand axis, imbalance on the right hand axis – Vib demo data set 2	221
Figure 51:	Outputs of RBF for 18g mass with bearing defect; bearing defect is shown on the left hand axis, imbalance on the right hand axis – Vib demo data set 2	222
Figure 52:	Outputs of the 0g, 12g and 24g RBF for all the training set data (actual measurements on the normal bearing) – Vib demo data set 2	222
Figure 53:	Outputs of the 0g, 12g and 24g RBF for all the training set data (measured on the defective bearing) – Vib demo data set 2 (imbalance on horizontal axis)	223
Figure 54:	Demodulation peak (magnitude spectrum value) versus buffer length for large bearing defect ($DA = 180^\circ$) and 0g (Vib demo data set 2)	236
Figure 55:	Demodulation peak (magnitude spectrum) versus buffer length for large bearing defect ($DA = 180^\circ$) and 0g (Vib demo data set 2) – zoom plot	237
Figure 56:	Demodulation peak (magnitude spectrum) versus buffer length for large bearing defect ($DA = 180^\circ$) and 18g (Vib demo data set 2)	237
Figure 57:	Demodulation peak (magnitude spectrum) versus buffer length for large bearing defect ($DA = 180^\circ$) and 18g (Vib demo data set 2) – zoom plot	238

Figure 58:	Demodulation peak versus buffer length for medium bearing defect and 0g	239
Figure 59:	Demodulation peak versus buffer length for medium bearing defect and 0g (zoom plot)	239
Figure 60:	Demodulation peak versus buffer length for medium bearing defect and 24g	240
Figure 61:	Demodulation peak versus buffer length for medium bearing defect and 24g (zoom plot)	240
Figure 62:	Demodulation peak versus buffer length for large bearing defect and 0g	241
Figure 63:	Demodulation peak versus buffer length for large bearing defect and 0g (zoom plot)	241
Figure 64:	Demodulation peak versus buffer length for large bearing defect and 18g	242
Figure 65:	Demodulation peak versus buffer length for large bearing defect and 18g (zoom plot)	242
Figure 66:	Result of vector addition when vectors add in phase (a) and out of phase (b)	289
Figure 67:	Feature Fr_X (magnitude spectrum value at rotational frequency on x axis) for large defect	293
Figure 68:	Feature $BPFO\_X$ (magnitude spectrum value at ball pass outer race defect frequency on x axis) for large defect	293
Figure 69:	Feature $BPFO2\_X$ (demodulation spectrum value at ball pass outer race defect frequency on x axis) for large defect	294
Figure 70:	Feature Fr_S_X (rotational frequency on x axis) for large defect	294

Figure 71:	Feature <i>BPFO2_S__X</i> (ball pass outer race defect frequency on x axis) for all defects	295
Figure 72:	Feature <i>BPF12_S__X</i> (ball pass inner race defect frequency on x axis) for large defect	295
Figure 73:	Feature <i>BPF12__X</i> (demodulation spectrum value at ball pass inner race defect frequency on x axis) for large defect	296
Figure 74:	Feature <i>KH__X</i> (kurtosis on x axis) for large defect	296
Figure 75:	Feature <i>CH__X</i> (real cepstrum on x axis) for large defect	297
Figure 76:	Feature <i>LH__X</i> (linear cepstrum on x axis) for large defect	297
Figure 77:	Feature <i>CH_p__X</i> (real cepstrum – after highpass filter – on x axis) for large defect	298
Figure 78:	Feature <i>KH_p__X</i> (kurtosis – after highpass filter – on x axis) for large defect	298
Figure 79:	Feature <i>LH_p__X</i> (linear cepstrum – after highpass filter – on x axis) for large defect	299
Figure 80:	Feature <i>FFT4_BPFO__X</i> (magnitude spectrum value of the square of the signal squared at ball pass outer race defect frequency on x axis) for large defect	299
Figure 81:	Feature <i>FFT8_BPFO__X</i> (magnitude spectrum value of the signal to the 8 th power at ball pass outer race defect frequency on x axis) for large defect	300
Figure 82:	Feature <i>SA_BPFO2_0__X</i> (square root of the SOS of synchronous average of the signal at ball pass outer race defect frequency on x axis) for large defect	300
Figure 83:	Feature <i>SA_BPFO2_p__X</i> (maximum value of synchronous average of the signal at ball pass outer race defect frequency on x axis) for large defect	301

Figure 84:	Feature <i>BPFO2_10_Eng__X</i> (square root of the SOS of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect	301
Figure 85:	Feature <i>BPFO2_10_Max__X</i> (maximum value of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect	302
Figure 86:	Feature <i>BPFO2_2_Eng__X</i> (square root of the SOS of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect	302
Figure 87:	Feature <i>BPFO2_2_Max__X</i> (maximum value of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect	303
Figure 88:	Feature <i>HFD_5_10_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 5–10kHz – on x axis) for large defect	303
Figure 89:	Feature <i>HFD_2_5_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 2–5kHz – on x axis) for large defect	304
Figure 90:	Feature <i>HFD_5_6_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 5–6kHz – on x axis) for large defect	304
Figure 91:	Feature <i>HFD_6_7_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 6–7kHz – on x axis) for large defect	305

Figure 92:	Feature <i>HFD_8_10_kHz_X</i> (square root of SOS of HFD components in magnitude spectrum – from 8–10kHz – on x axis) for large defect	305
Figure 93:	Outputs of RBF for 0g normal training set data – Vib demo data set 2	306
Figure 94:	Outputs of RBF for 0g bearing defect training set data – Vib demo data set 2	306
Figure 95:	Outputs of RBF for 12g normal training set data – Vib demo data set 2	307
Figure 96:	Outputs of RBF for 12g bearing defect training set data – Vib demo data set 2	307
Figure 97:	Outputs of RBF for 24g normal training set data – Vib demo data set 2	308
Figure 98:	Outputs of RBF for 24g bearing defect training set data – Vib demo data set 2	308
Figure 99:	Outputs of RBF for 0g training set data (measured on the normal bearing) – Vib demo data set 2	309
Figure 100:	Outputs of RBF for 12g training set data (measured on the normal bearing) – Vib demo data set 2	309
Figure 101:	Outputs of RBF for 24g training set data (measured on the normal bearing) – Vib demo data set 2	310
Figure 102:	Outputs of RBF for 0g test set data (measured on the defective bearing) – Vib demo data set 2	310
Figure 103:	Outputs of RBF for 12g test set data (measured on the defective bearing) – Vib demo data set 2	311
Figure 104:	Outputs of RBF for 24g test set data (measured on the defective bearing) – Vib demo data set 2	311

Figure 105: Feature <i>Fr__X</i> (magnitude spectrum value at rotational frequency on x axis) for all defects	314
Figure 106: Feature <i>BPFO__X</i> (magnitude spectrum value at ball pass outer race defect frequency on x axis) for all defects	314
Figure 107: Feature <i>BPFO2__X</i> (demodulation spectrum value at ball pass outer race defect frequency on x axis) for all defects	315
Figure 108: Feature <i>Fr_S__X</i> (rotational frequency on x axis) for all defects	315
Figure 109: Feature <i>BPFO2_S__X</i> (ball pass outer race defect frequency on x axis) for all defects	316
Figure 110: Feature <i>BPFI2_S__X</i> (ball pass inner race defect frequency on x axis) for all defects	316
Figure 111: Feature <i>BPFI2__X</i> (demodulation spectrum value at ball pass inner race defect frequency on x axis) for all defects	317
Figure 112: Feature <i>KH__X</i> (kurtosis on x axis) for all defects	317
Figure 113: Feature <i>CH__X</i> (real cepstrum on x axis) for all defects	318
Figure 114: Feature <i>LH__X</i> (linear cepstrum on x axis) for all defects	318
Figure 115: Feature <i>CH_p__X</i> (real cepstrum – after highpass filter – on x axis) for all defects	319
Figure 116: Feature <i>KH_p__X</i> (kurtosis – after highpass filter – on x axis) for all defects	319
Figure 117: Feature <i>LH_p__X</i> (linear cepstrum – after highpass filter – on x axis) for all defects	320
Figure 118: Feature <i>FFT4_BPFO__X</i> (magnitude spectrum value of the square of the signal squared at ball pass outer race defect frequency on x axis) for all defects	320

Figure 119:	Feature <i>FFT8_BPFO__X</i> (magnitude spectrum value of the signal to the 8 th power at ball pass outer race defect frequency on x axis) for all defects	321
Figure 120:	Feature <i>SA_BPFO2_0__X</i> (square root of the SOS of synchronous average of the signal at ball pass outer race defect frequency on x axis) for all defects	321
Figure 121:	Feature <i>SA_BPFO2_p__X</i> (maximum value of synchronous average of the signal at ball pass outer race defect frequency on x axis) for all defects	322
Figure 122:	Feature <i>BPFO2_10_Eng__X</i> (square root of the SOS of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for all defects	322
Figure 123:	Feature <i>BPFO2_10_Max__X</i> (maximum value of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for all defects	323
Figure 124:	Feature <i>BPFO2_2_Eng__X</i> (square root of the SOS of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for all defects	323
Figure 125:	Feature <i>BPFO2_2_Max__X</i> (maximum value of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for all defects	324
Figure 126:	Feature <i>HFD_5_10_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 5–10kHz – on x axis) for all defects	324

Figure 127: Feature <i>HFD_2_5_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 2–5kHz – on x axis) for all defects	325
Figure 128: Feature <i>HFD_5_6_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 5–6kHz – on x axis) for all defects	325
Figure 129: Feature <i>HFD_6_7_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 6–7kHz – on x axis) for all defects	326
Figure 130: Feature <i>HFD_8_10_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 8–10kHz – on x axis) for all defects	326
Figure 131: Feature <i>FFT2_MEAN__X</i> (first component of demodulation spectrum on x axis) for all defects	327
Figure 132: Feature <i>Fr__X</i> (magnitude spectrum value at rotational frequency on x axis) for large defect	328
Figure 133: Feature <i>BPFO__X</i> (magnitude spectrum value at ball pass outer race defect frequency on x axis) for large defect	328
Figure 134: Feature <i>BPFO2__X</i> (demodulation spectrum value at ball pass outer race defect frequency on x axis) for large defect	329
Figure 135: Feature <i>Fr_S__X</i> (rotational frequency on x axis) for large defect	329
Figure 136: Feature <i>BPFO2_S__X</i> (ball pass outer race defect frequency on x axis) for large defect	330
Figure 137: Feature <i>BPII2_S__X</i> (ball pass inner race defect frequency on x axis) for large defect	330

Figure 138:	Feature <i>BPF12__X</i> (demodulation spectrum value at ball pass inner race defect frequency on x axis) for large defect	331
Figure 139:	Feature <i>KH__X</i> (kurtosis on x axis) for large defect	331
Figure 140:	Feature <i>CH__X</i> (real cepstrum on x axis) for large defect	332
Figure 141:	Feature <i>LH__X</i> (linear cepstrum on x axis) for large defect	332
Figure 142:	Feature <i>CH_p__X</i> (real cepstrum – after highpass filter – on x axis) for large defect	333
Figure 143:	Feature <i>KH_p__X</i> (kurtosis – after highpass filter – on x axis) for large defect	333
Figure 144:	Feature <i>LH_p__X</i> (linear cepstrum – after highpass filter – on x axis) for large defect	334
Figure 145:	Feature <i>FFT4_BPFO__X</i> (magnitude spectrum value of the square of the signal squared at ball pass outer race defect frequency on x axis) for large defect	334
Figure 146:	Feature <i>FFT8_BPFO__X</i> (magnitude spectrum value of the signal to the 8 th power at ball pass outer race defect frequency on x axis) for large defect	335
Figure 147:	Feature <i>SA_BPFO2_0__X</i> (square root of the SOS of synchronous average of the signal at ball pass outer race defect frequency on x axis) for large defect	335
Figure 148:	Feature <i>SA_BPFO2_p__X</i> (maximum value of synchronous average of the signal at ball pass outer race defect frequency on x axis) for large defect	336
Figure 149:	Feature <i>BPFO2_10_Eng__X</i> (square root of the SOS of the first 10 components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect	336

Figure 150: Feature <i>BPFO2_10_Max__X</i> (maximum value of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect	337
Figure 151: Feature <i>BPFO2_2_Eng__X</i> (square root of the SOS of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect	337
Figure 152: Feature <i>BPFO2_2_Max__X</i> (maximum value of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect	338
Figure 153: Feature <i>HFD_5_10_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 5–10kHz – on x axis) for large defect	338
Figure 154: Feature <i>HFD_2_5_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 2–5kHz – on x axis) for large defect	339
Figure 155: Feature <i>HFD_5_6_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 5–6kHz – on x axis) for large defect	339
Figure 156: Feature <i>HFD_6_7_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 6–7kHz – on x axis) for large defect	340
Figure 157: Feature <i>HFD_8_10_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 8–10kHz – on x axis) for large defect	340

Figure 158: Feature <i>FFT2_MEAN__X</i> (first component of demodulation spectrum on x axis) for large defect	341
Figure 159: Feature <i>Fr__X</i> (magnitude spectrum value at rotational frequency on x axis) for medium defect	342
Figure 160: Feature <i>BPFO__X</i> (magnitude spectrum value at ball pass outer race defect frequency on x axis) for medium defect	342
Figure 161: Feature <i>BPFO2__X</i> (demodulation spectrum value at ball pass outer race defect frequency on x axis) for medium defect	343
Figure 162: Feature <i>Fr_S__X</i> (rotational frequency on x axis) for medium defect	343
Figure 163: Feature <i>BPFO2_S__X</i> (ball pass outer race defect frequency on x axis) for medium defect	344
Figure 164: Feature <i>BPFi2_S__X</i> (ball pass inner race defect frequency on x axis) for medium defect	344
Figure 165: Feature <i>BPFi2__X</i> (demodulation spectrum value at ball pass inner race defect frequency on x axis) for medium defect	345
Figure 166: Feature <i>KH__X</i> (kurtosis on x axis) for medium defect	345
Figure 167: Feature <i>CH__X</i> (real cepstrum on x axis) for medium defect	346
Figure 168: Feature <i>LH__X</i> (linear cepstrum on x axis) for medium defect	346
Figure 169: Feature <i>CH_p__X</i> (real cepstrum – after highpass filter – on x axis) for medium defect	347
Figure 170: Feature <i>KH_p__X</i> (kurtosis – after highpass filter – on x axis) for medium defect	347

Figure 171: Feature LH_p_X (linear cepstrum – after highpass filter – on x axis) for medium defect	348
Figure 172: Feature $FFT4\_BPFO\_X$ (magnitude spectrum value of the square of the signal squared at ball pass outer race defect frequency on x axis) for medium defect	348
Figure 173: Feature $FFT8\_BPFO\_X$ (magnitude spectrum value of the signal to the 8 th power at ball pass outer race defect frequency on x axis) for medium defect	349
Figure 174: Feature $SA\_BPFO2\_0\_X$ (square root of the SOS of synchronous average of the signal at ball pass outer race defect frequency on x axis) for medium defect	349
Figure 175: Feature $SA\_BPFO2\_p\_X$ (maximum value of synchronous average of the signal at ball pass outer race defect frequency on x axis) for medium defect	350
Figure 176: Feature $BPFO2\_10\_Eng\_X$ (square root of the SOS of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for medium defect	350
Figure 177: Feature $BPFO2\_10\_Max\_X$ (maximum value of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for medium defect	351
Figure 178: Feature $BPFO2\_2\_Eng\_X$ (square root of the SOS of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for medium defect	351
Figure 179: Feature $BPFO2\_2\_Max\_X$ (maximum value of the first two components in demodulation spectrum – at multiples	

of the ball pass outer race defect frequency on x axis) for medium defect	352
Figure 180: Feature <i>HFD_5_10_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 5–10kHz – on x axis) for medium defect	352
Figure 181: Feature <i>HFD_2_5_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 2–5kHz – on x axis) for medium defect	353
Figure 182: Feature <i>HFD_5_6_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 5–6kHz – on x axis) for medium defect	353
Figure 183: Feature <i>HFD_6_7_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 6–7kHz – on x axis) for medium defect	354
Figure 184: Feature <i>HFD_8_10_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 8–10kHz – on x axis) for medium defect	354
Figure 185: Feature <i>FFT2_MEAN__X</i> (first component of demodulation spectrum on x axis) for medium defect	355
Figure 186: Feature <i>Fr__X</i> (magnitude spectrum value at rotational frequency on x axis) for small defect	356
Figure 187: Feature <i>BPFO__X</i> (magnitude spectrum value at ball pass outer race defect frequency on x axis) for small defect	356
Figure 188: Feature <i>BPFO2__X</i> (demodulation spectrum value at ball pass outer race defect frequency on x axis) for small defect	357
Figure 189: Feature <i>Fr_S__X</i> (rotational frequency on x axis) for small defect	357

Figure 190:	Feature <i>BPFO2_S__X</i> (ball pass outer race defect frequency on x axis) for small defect	358
Figure 191:	Feature <i>BPFI2_S__X</i> (ball pass inner race defect frequency on x axis) for small defect	358
Figure 192:	Feature <i>BPFI2__X</i> (demodulation spectrum value at ball pass inner race defect frequency on x axis) for small defect	359
Figure 193:	Feature <i>KH__X</i> (kurtosis on x axis) for small defect	359
Figure 194:	Feature <i>CH__X</i> (real cepstrum on x axis) for small defect	360
Figure 195:	Feature <i>LH__X</i> (linear cepstrum on x axis) for small defect	360
Figure 196:	Feature <i>CH_p__X</i> (real cepstrum – after highpass filter – on x axis) for small defect	361
Figure 197:	Feature <i>KH_p__X</i> (kurtosis – after highpass filter – on x axis) for small defect	361
Figure 198:	Feature <i>LH_p__X</i> (linear cepstrum – after highpass filter – on x axis) for small defect	362
Figure 199:	Feature <i>FFT4_BPFO__X</i> (magnitude spectrum value of the square of the signal squared at ball pass outer race defect frequency on x axis) for small defect	362
Figure 200:	Feature <i>FFT8_BPFO__X</i> (magnitude spectrum value of the signal to the 8 th power at ball pass outer race defect frequency on x axis) for small defect	363
Figure 201:	Feature <i>SA_BPFO2_0__X</i> (square root of the SOS of synchronous average of the signal at ball pass outer race defect frequency on x axis) for small defect	363
Figure 202:	Feature <i>SA_BPFO2_p__X</i> (maximum value of synchronous average of the signal at ball pass outer race defect frequency on x axis) for small defect	364

Figure 203: Feature <i>BPFO2_10_Eng__X</i> (square root of the SOS of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for small defect	364
Figure 204: Feature <i>BPFO2_10_Max__X</i> (maximum value of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for small defect	365
Figure 205: Feature <i>BPFO2_2_Eng__X</i> (square root of the SOS of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for small defect	365
Figure 206: Feature <i>BPFO2_2_Max__X</i> (maximum value of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for small defect	366
Figure 207: Feature <i>HFD_5_10_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 5–10kHz – on x axis) for small defect	366
Figure 208: Feature <i>HFD_2_5_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 2–5kHz – on x axis) for small defect	367
Figure 209: Feature <i>HFD_5_6_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 5–6kHz – on x axis) for small defect	367
Figure 210: Feature <i>HFD_6_7_kHz__X</i> (square root of SOS of HFD components in magnitude spectrum – from 6–7kHz – on x axis) for small defect	368

Figure 211: Feature <i>HFD_8_10_kHz_X</i> (square root of SOS of HFD components in magnitude spectrum – from 8–10kHz – on x axis) for small defect	368
Figure 212: Feature <i>FFT2_MEAN_X</i> (first component of demodulation spectrum on x axis) for small defect	369
Figure 213: Inner race defect	372
Figure 214: Inner race defect	372
Figure 215: Roller defect	373
Figure 216: Outer race defect (medium)	373
Figure 217: Outer race defect (medium)	374
Figure 218: Outer race defect (medium)	374
Figure 219: Outer race defect (medium)	375
Figure 220: Rust marks	375
Figure 221: Outer race defect (small)	376
Figure 222: Outer race defect (small)	376
Figure 223: Outer race defect (small)	377
Figure 224: Scratch marks	377
Figure 225: Scratch marks	378
Figure 226: Waterfall plot of signal with Kaiser window ($\beta = 0$)	391
Figure 227: Waterfall plot of signal with Kaiser window ($\beta = 6$)	392
Figure 228: Waterfall plot of signal with Kaiser window ($\beta = 100$)	392
Figure 229: SA of signal with Kaiser window ($\beta = 0$)	393
Figure 230: SA of signal with Kaiser window ($\beta = 6$)	393
Figure 231: SA of signal with Kaiser window ($\beta = 100$)	394

Figure 232: Pulse train corresponding to SA with Kaiser window ($\beta = 0$)	394
Figure 233: Pulse train corresponding to SA with Kaiser window ($\beta = 6$)	395
Figure 234: Pulse train corresponding to SA with Kaiser window ($\beta = 100$)	395
Figure 235: Demodulation spectrum of signal prior to SA for Kaiser window where $\beta = 0$	396
Figure 236: Demodulation spectrum of pulse signal with SA for Kaiser window where $\beta = 0$	396
Figure 237: Demodulation spectrum of signal prior to SA for Kaiser window where $\beta = 6$	397
Figure 238: Demodulation spectrum of signal with SA for Kaiser window where $\beta = 6$	397
Figure 239: Demodulation spectrum of signal prior to SA for Kaiser window where $\beta = 100$	398
Figure 240: Demodulation spectrum of pulse signal with SA for Kaiser window where $\beta = 100$	398
Figure 241: Demodulation spectrum of signal prior to SA for Kaiser window where $\beta = 100$ (zoom plot)	399
Figure 242: Time domain signals plotted over each other for one period (STD of noise = 0, speed variation = 0%)	410
Figure 243: MOS $\Sigma(X^2)/N$ of time domain signal plotted for one period of the synchronous average (STD of noise = 0, speed variation = 0%)	411

Figure 244: Mean squared – $(\Sigma X/N)^2$ – of time domain signal plotted for one period of the synchronous average (STD of noise = 0, speed variation = 0%)	411
Figure 245: Ratio of mean squared to MOS of time domain signal – $(\Sigma X/N)^2/(\Sigma X^2/N)$ – plotted for one period of the synchronous average (STD of noise = 0, speed variation = 0%)	412
Figure 246: Kurtosis of samples of SA – plotted for one period of the synchronous average (STD of noise = 0, speed variation = 0%)	412
Figure 247: FFT magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0, speed variation = 0%)	413
Figure 248: FFT demodulation magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0, speed variation = 0%)	413
Figure 249: Time domain signal plotted for about five periods (STD of noise = 0, speed variation = 20%)	414
Figure 250: Time domain signals plotted over each other for one period (STD of noise = 0, speed variation = 20%)	414
Figure 251: MOS $\Sigma(X^2)/N$ of time domain signal plotted for one period of the synchronous average (STD of noise = 0, speed variation = 20%)	415
Figure 252: Mean squared – $(\Sigma X/N)^2$ – of time domain signal plotted for one period of the synchronous average (STD of noise = 0, speed variation = 20%)	415

Figure 253: Ratio of mean squared to MOS of time domain signal – $(\Sigma X/N)^2/(\Sigma X^2/N)$ – plotted for a period of the SA (STD of noise = 0, speed variation = 20%)	416
Figure 254: Kurtosis of samples of SA – plotted for one period of the synchronous average (STD of noise = 0, speed variation = 20%)	416
Figure 255: FFT magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0, speed variation = 20%)	417
Figure 256: FFT demodulation magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0, speed variation = 20%)	417
Figure 257: Time domain signal plotted for about five periods (STD of noise = 0.02, speed variation = 20%)	418
Figure 258: Time domain signals plotted over each other for one period (STD of noise = 0.02, speed variation = 20%)	418
Figure 259: MOS $\Sigma(X^2)/N$ of time domain signal plotted for one period of the synchronous average (STD of noise = 0.02, speed variation = 20%)	419
Figure 260: Mean squared – $(\Sigma X/N)^2$ – of time domain signal plotted for one period of the synchronous average (STD of noise = 0.02, speed variation = 20%)	419
Figure 261: Ratio of mean squared to MOS of time domain signal – $(\Sigma X/N)^2/(\Sigma X^2/N)$ – plotted for one period of the synchronous average (STD of noise = 0.02, speed variation = 20%)	420

Figure 262: Kurtosis of samples of SA – plotted for one period of the synchronous average (STD of noise = 0.02, speed variation = 20%)	420
Figure 263: Time domain signal plotted for about five periods (STD of noise = 0.1, speed variation = 0%)	421
Figure 264: Time domain signals plotted over each other for one period (STD of noise = 0.1, speed variation = 0%)	421
Figure 265: $MOS \Sigma(X^2)/N$ of time domain signal plotted for one period of the synchronous average (STD of noise = 0.1, speed variation = 0%)	422
Figure 266: Mean squared – $(\Sigma X/N)^2$ – of time domain signal plotted for one period of the synchronous average (STD of noise = 0.1, speed variation = 0%)	422
Figure 267: Ratio of mean squared to MOS of time domain signal – $(\Sigma X/N)^2/(\Sigma X^2/N)$ – plotted for one period of the synchronous average (STD of noise = 0.1, speed variation = 0%)	423
Figure 268: Kurtosis of samples of SA – plotted for one period of the synchronous average (STD of noise = 0.1, speed variation = 0%)	423
Figure 269: FFT magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0.1, speed variation = 0%)	424
Figure 270: FFT demodulation magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0.1, speed variation = 0%)	424
Figure 271: Time domain signal plotted for about five periods (STD of noise = 0.15, speed variation = 0%)	425

Figure 272:	Time domain signals plotted over each other for one period (STD of noise = 0.15, speed variation = 0%)	425
Figure 273:	MOS $\Sigma(X^2)/N$ of time domain signal plotted for one period of the synchronous average (STD of noise = 0.15, speed variation = 0%)	426
Figure 274:	Mean squared – $(\Sigma X/N)^2$ – of time domain signal plotted for one period of the synchronous average (STD of noise = 0.15, speed variation = 0%)	426
Figure 275:	Ratio of mean squared to MOS of time domain signal – $(\Sigma X/N)^2/(\Sigma X^2/N)$ – plotted for one period of the synchronous average (STD of noise = 0.15, speed variation = 0%)	427
Figure 276:	Kurtosis of samples of SA – plotted for one period of the synchronous average (STD of noise = 0.15, speed variation = 0%)	427
Figure 277:	FFT magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0.15, speed variation = 0%)	428
Figure 278:	FFT demodulation magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0.15, speed variation = 0%)	428
Figure 279:	Position of the Gaussian pulse as a function of time (1/defect frequency = 0.008745 seconds)	429
Figure 280:	Amplitude spectrum of signal – linear increase of starting position of pulse in time, maximum variation = 0.41%	430
Figure 281:	Real cepstrum of signal – linear increase of starting position of pulse in time, maximum variation = 0.41%	431

Figure 282: Log amplitude spectrum of signal – linear increase of starting position of pulse in time, maximum variation = 0.41%	431
Figure 283: Complex cepstrum of signal – linear increase of starting position of pulse in time, maximum variation = 0.41%	432
Figure 284: Linear cepstrum (LCEPS) of signal – linear increase of starting position of pulse in time, maximum variation = 0.41%	433
Figure 285: Amplitude spectrum of signal squared – linear increase of starting position of pulse in time, maximum variation = 0.41%	434
Figure 286: Spectrum of sampling sequence (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41%	437
Figure 287: Spectrum of sampling sequence (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41% (zoom plot)	437
Figure 288: Spectrum of sampling sequence (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41% (zoom plot)	438
Figure 289: Amplitude spectrum after bandpass filter (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41% (zoom plot)	440
Figure 290: Amplitude spectrum of squared signal after bandpass filter (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41% (zoom plot)	441

- Figure 291: Amplitude spectrum of squared signal after bandpass filter and with removal of all phase information (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41% (zoom plot) 441
- Figure 292: Phase spectrum of squared signal after bandpass filter (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41% (zoom plot) 442
- Figure 293: Phase spectrum (unwrapped) of squared signal after bandpass filter (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41% (zoom plot) 442
- Figure 294: Combined amplitude and phase diagram of signal after bandpass filter (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41% (zoom plot); the frequency axis is vertical (zoom plot) and the two horizontal axes give the real and imaginary components 443
- Figure 295: Combined amplitude and phase diagram of squared signal after bandpass filter for a frequency of 114.347Hz (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41% (zoom plot); the frequency axis is vertical and the two horizontal axes give the real and imaginary components 443
- Figure 296: Combined amplitude and phase diagram of squared signal after bandpass filter for a frequency of 114.347Hz (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41% (zoom plot); the frequency axis is vertical (zoom plot) and

the two horizontal axes give the real and imaginary components 444

Figure 297: Combined amplitude and phase diagram of squared signal after bandpass filter for a frequency of 112.213Hz (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41%; the frequency axis is vertical and the two horizontal axes give the real and imaginary components 444

Figure 298: Combined amplitude and phase diagram of squared signal after bandpass filter for a frequency of 112.213Hz (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41% (zoom plot); the frequency axis is vertical and the two horizontal axes give the real and imaginary components 445

Figure 299: Combined amplitude and phase diagram (convolutional products) of squared signal after bandpass filter for a frequency of 114.347Hz (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41% (zoom plot); the frequency axis is vertical (zoom plot) and the two horizontal axes give the real and imaginary components 445

Figure 300: Combined amplitude and phase diagram (convolutional products) of squared signal after bandpass filter for a frequency of 112.213Hz (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41% (zoom plot); the frequency axis is vertical and the two horizontal axes give the real and imaginary components 446

Figure 301: Position of the Gaussian pulse (refer to section 3.3) as a function of time (1/defect frequency = 0.008745 seconds)	447
Figure 302: Amplitude spectrum of signal – linear increase of starting position of pulse in time, maximum variation = 62%	447
Figure 303: Amplitude spectrum of signal squared – linear increase of starting position of pulse in time, maximum variation = 62%	449
Figure 304: FFT of sinusoidal signal, $\text{STD}(x) = \sigma_x = 1$	454
Figure 305: $\text{FFT}((x(t) + n(t))^2)$, $\text{STD}(n) = \sigma_n = 2$	455
Figure 306: $\text{FFT}(x(t) + n(t))$, $\text{STD}(n) = \sigma_n = 2$	455
Figure 307: $\text{FFT}((x(t) + n(t))^2)$, $\text{STD}(n) = \sigma_n = 5$	456
Figure 308: $\text{FFT}((x(t) + n(t))^2)$, $\text{STD}(n) = \sigma_n = 5$ (zoom plot)	456
Figure 309: $\text{FFT}(x(t) + n(t))$, $\text{STD}(n) = \sigma_n = 5$	457
Figure 310: Standard deviation of model signal as function of 2^p	459
Figure 311: Ratio of $\text{FFT}(n^2)/\text{FFT}(n)$ (ignore DC component in spectrum) using maximum value function	460
Figure 312: Ratio of $\text{FFT}(n^2)/\text{FFT}(n)$ (ignore DC component in spectrum) using maximum value function – zoom plot	461
Figure 313: Ratio of $\text{FFT}(n^2)/\text{FFT}(n)$ (ignore DC component in spectrum) using mean value function	461
Figure 314: Ratio of $\text{FFT}(n^2)/\text{FFT}(n)$ (ignore DC component in spectrum) using the mean value function – zoom plot	462
Figure 315: Ratio of $\text{FFT}(n^2)/\text{FFT}(n)$ (ignore DC component in spectrum) as a function of model parameter p	462

Figure 316:	The ratio of the (FFT of the) squared signal relative to the (FFT of the) signal as function of the parameter p (the signal $y = K_{sig}(x/\sqrt{2})^p$, $K_{sig} = 1$)	463
Figure 317:	Maximum value of regular magnitude spectrum of model signal $((x/\sqrt{2})^p)$ as function of p	464
Figure 318:	Mean value of spectrum versus noise level	465
Figure 319:	Maximum value of regular magnitude spectrum of model signal $((x/\sqrt{2})^p)$ as function of p – zoom plot	467
Figure 320:	Ratio of $K_{sig}FFT(n^2)/FFT(n)$ (ignore DC component in spectrum) using the mean value function	469
Figure 321:	Ratio of $K_{sig}FFT(n^2)/FFT(n)$ (ignore DC component in spectrum) using the mean value function – zoom plot	469
Figure 322:	Vibration time domain signal $p = 4$	470
Figure 323:	FFT of time domain signal $p = 4$	471
Figure 324:	FFT of additive noise with $STD(n) = \sigma_n = 2$, $p = 4$	471
Figure 325:	FFT of time domain signal plus additive noise with $STD(x) = \sigma_x = 0.36$, $STD(n) = \sigma_n = 2$, $p = 4$	472
Figure 326:	FFT of time domain signal squared using $STD(x) = \sigma_x = 0.36$, $STD(n) = \sigma_n = 2$, $p = 4$	472
Figure 327:	FFT of square of noise signal with $STD(n) = \sigma_n = 2$, $p = 4$	473
Figure 328:	FFT of time domain signal plus additive noise squared using $STD(x) = \sigma_x = 0.36$, $STD(n) = \sigma_n = 2$, $p = 4$	473
Figure 329:	FFT of product of time domain signal and noise signal (multiplicative noise)	474
Figure 330:	Vibration time domain signal $p = 256$	474

Figure 331: FFT of time domain signal $p = 256$	475
Figure 332: FFT of additive noise with $\text{STD}(n) = \sigma_n = 0.1, p = 256$	475
Figure 333: FFT of time domain signal plus additive noise with $\text{STD}(x) = \sigma_x = 0.18, \text{STD}(n) = \sigma_n = 0.1, p = 256$	476
Figure 334: FFT of time domain signal squared with $\text{STD}(x) = \sigma_x = 0.18, \text{STD}(n) = \sigma_n = 0.1, p = 256$	476
Figure 335: FFT of square of noise signal with $\text{STD}(n) = \sigma_n = 0.1, p = 256$	477
Figure 336: FFT of time domain signal plus additive noise squared with $\text{STD}(x) = \sigma_x = 0.18, \text{STD}(n) = \sigma_n = 0.1, p = 256$	478
Figure 337: FFT of product of time domain signal and noise signal (multiplicative noise)	478
Figure 338: Vibration time domain signal $p = 65\ 536$	479
Figure 339: FFT of time domain signal $p = 65\ 536$	479
Figure 340: FFT of additive noise with $\text{STD}(n) = \sigma_n = 0.1, p = 65\ 536$	480
Figure 341: FFT of time domain signal plus additive noise with $\text{STD}(x) = \sigma_x = 0.0477, \text{STD}(n) = \sigma_n = 0.1, p = 65\ 536$	481
Figure 342: FFT of time domain signal squared with $\text{STD}(x) = \sigma_x = 0.0477, \text{STD}(n) = \sigma_n = 0.1, p = 65\ 536$	481
Figure 343: FFT of square of noise signal with $\text{STD}(n) = \sigma_n = 0.1, p = 65\ 536$	482
Figure 344: FFT of time domain signal plus additive noise squared with $\text{STD}(x) = \sigma_x = 0.0477, \text{STD}(n) = \sigma_n = 0.1, p = 65\ 536$	483
Figure 345: FFT of product of time domain signal and noise signal (multiplicative noise)	483

Figure 346: Vibration time domain signal $p = 4$	484
Figure 347: FFT of time domain signal $p = 4$	484
Figure 348: FFT of additive noise with $\text{STD}(n) = \sigma_n = 20, p = 4$	485
Figure 349: FFT of time domain signal plus additive noise with $\text{STD}(x) = \sigma_x = 3.6455, \text{STD}(n) = \sigma_n = 20, p = 4$	486
Figure 350: FFT of time domain signal squared with $\text{STD}(x) = \sigma_x = 3.6455, \text{STD}(n) = \sigma_n = 20, p = 4$	486
Figure 351: FFT of square of noise signal with $\text{STD}(n) = \sigma_n = 20, p = 4$	487
Figure 352: FFT of time domain signal plus additive noise squared with $\text{STD}(x) = \sigma_x = 3.6455, \text{STD}(n) = \sigma_n = 20, p = 65\ 536$	488
Figure 353: FFT of product of time domain signal and noise signal (multiplicative)	488
Figure 354: FFT magnitude spectrum of model signal with multiplicative noise ($\sigma_n = 1$)	489
Figure 355: FFT magnitude spectrum of model signal squared with multiplicative noise ($\sigma_n = 1$)	489
Figure 356: FFT magnitude spectrum of model signal with multiplicative noise ($\sigma_n = 5$)	490
Figure 357: FFT magnitude spectrum of model signal squared with multiplicative noise ($\sigma_n = 5$).	490
Figure 358: Time domain signal, $p = 65\ 536, SV = 45\%, \sigma_n = 0.01$	491
Figure 359: Bandpass filter from 10–20kHz	492
Figure 360: FFT magnitude spectrum of model signal plus noise ($SV = 45\%, p = 65\ 536, \sigma_n = 0.01$)	492

Figure 361: FFT magnitude spectrum of model signal plus noise – zoom plot ($SV = 45\%$, $p = 65\ 536$, $\sigma_n = 0.01$)	493
Figure 362: FFT magnitude spectrum of model signal plus noise squared ($SV = 45\%$, $p = 65\ 536$, $\sigma_n = 0.01$)	493
Figure 363: Time domain signal, $p = 65\ 536$, $SV = 45\%$, $\sigma_n = 0.01$	494
Figure 364: Bandpass filter from 10–20kHz	494
Figure 365: FFT magnitude spectrum of model signal plus noise ($SV = 45\%$, $p = 65\ 536$, $\sigma_n = 0.01$)	495
Figure 366: FFT magnitude spectrum of model signal squared ($SV = 45\%$, $p = 65\ 536$, $\sigma_n = 0.01$)	495
Figure 367: Time domain signal (Vib demo data set 3)	496
Figure 368: Bandpass filter from 10–20kHz	496
Figure 369: FFT magnitude spectrum of measured vibration signal	497
Figure 370: FFT magnitude spectrum of measured vibration signal – zoom plot	497
Figure 371: FFT magnitude spectrum of measured vibration signal squared	498

LIST OF TABLES

Table 1:	Bearing specifications	69
Table 2:	10th sidelobe level of windows	129
Table 3:	Level (dB) of windows at specific frequencies	130
Table 4:	Relation between time domain, autocorrelation and frequency domain after application of a non-linear operator (squaring of the signal)	158
Table 5:	Effect of speed variation on demodulation peak	204
Table 6:	Defect angle (DA in degrees) and value of demodulation peak (<i>BPFO2_X</i> in dB) versus dynamic load for a medium defect	225
Table 7:	Defect angle (DA in degrees) and value of demodulation peak (<i>BPFO2_X</i> in dB) versus dynamic load for a small defect	225
Table 8:	Dynamic load (imbalance in grams) and value of demodulation peak (<i>BPFO2_X</i> in dB) versus defect angle (DA in degrees) and defect size (S, M, L for small, medium and large respectively); the symbol NA means measurement is not available	226
Table 9:	Dynamic load (imbalance in grams) and feature <i>FFT2_MEAN_X</i> (in dB) versus defect angle (DA in degrees) and defect size (S, M, L for small, medium and large respectively); the symbol NA means measurement is not available	229
Table 10:	Dynamic load (imbalance in grams) and feature <i>SA_BPFO2_0__X</i> (in dB) versus defect angle (DA in	

	degrees) and defect size (S, M, L for small, medium and large respectively); the symbol NA means measurement is not available	231
Table 11:	Dynamic load (imbalance in grams) and outer race defect frequency <i>BPFO2_S__X</i> (in Hz) feature versus defect angle (DA in degrees) and defect size (S, M, L for small, medium and large respectively); the symbol NA means measurement is not available	233
Table 12:	Dynamic load (imbalance in grams) and <i>HFD</i> feature <i>HFD_2_5kHz__X</i> (in dB) versus defect angle (DA in degrees) and defect size (S, M, L for small, medium and large respectively); the symbol NA means measurement is not available	243
Table 13:	Dynamic load (imbalance in grams) and kurtosis feature <i>KHX__X</i> (in dB) versus defect angle (DA in degrees) and defect size (S, M, L for small, medium and large respectively); the symbol NA means measurement is not available	244
Table 14:	Dynamic load (imbalance in grams) and rotating speed feature <i>Fr__X</i> (in dB) versus defect angle (DA in degrees) and defect size (S, M, L for small, medium and large respectively); the symbol NA means measurement is not available	248
Table 15:	Some literature references for fault identification	381
Table 16:	Detection ratio (DR) as a function of additive noise σ (the symbol NA means not available)	453

DEDICATION

IN LOVING MEMORY
OF MY LATE BROTHER
JOHANNES PETRUS VAN DER MERWE

ACKNOWLEDGEMENTS

Having reached the end of several years of research on the topic of this thesis, I acknowledge with gratitude that I owe my life and intellectual abilities to God my Creator.

My promoter, Prof. Dr A. J. Hoffman, has given me focused direction and meaningful academic guidance. I thank him sincerely for granting me a generous amount of his precious time.

I appreciate the kind and helpful assistance of library staff at the Ferdinand Postma library of the Potchefstroom University for Christian Higher Education, the library of the Rand Afrikaans University in Johannesburg, the Merensky library of the University of Pretoria and the Samuel Pauw library of the University of South Africa.

My parents and family have encouraged and supported my academic interests over many years. Penetrating discussions with my late brother Johannes, who acted graciously as a soundboard for critical testing of complex problems in the broad area of information technology, have been a special joy in past years.

I have had the privilege of constructive personal discussions or correspondence with a number of experts in the subject area. Although it is impossible to mention everybody who contributed to this thesis, I would like to single out Prof. A. J. Hoffman, Prof. A. K. Nandi, Dr P. D. McFadden, Prof. P. S. Heyns and Mr C. van der Walt.

I am grateful to Prof. Antoniadis for use of the data set in [63] in the framework of collaboration with Aluminum of Greece SA. The mathematical software Matlab (www.mathworks.com) enabled me to prepare the illustrations. I am thankful to Bearing Inspection Services for test results on defective bearings, Mr C. van der Walt

(currently vibration consultant for SASOL) for use of the test rig and discussions as well as Mr R. Tessendorf (at present lecturer at Potchefstroom University for Christian Higher Education). Thanks are also due to Prof. S. Heyns and his students and staff for use of the Goldfields and SASOL laboratories at the University of Pretoria as well as valuable discussions and suggestions in the field of vibration monitoring. I thank Dr L. Tiedt of Potchefstroom University for Christian Higher Education for interpreting the SEM results on the bearings and my sister Mrs I. Stotko for her meticulous language editing of the thesis. Finally, I thank Mr A.D.W. Wolmarans for his encouragement and discussions.

GLOSSARY

ALE	adaptive line enhancement
AM	amplitude modulated
ANC	adaptive noise cancelling
BPFI_X²	value of demodulated (high frequency resonance) x axis vibration signal at BPFI; the signal is bandpass filtered and then demodulated (by squaring of the signal) to retain the high frequency resonance information only
BPFO	the ball pass frequency on the outer race of the defective bearing
BPFO_X	value of x axis vibration signal at BPFO (ball pass frequency outer race)
BPFO_X²	value of demodulated (high frequency resonance) x axis vibration signal at BPFO; the signal is bandpass filtered and then demodulated (by squaring of the signal) to retain the high frequency resonance information only
BPOR	ball pass outer race
CA	cage fault
CBM	condition based maintenance
CVCP	cross validation strategy to determine classifier performance
CVMP	cross validation strategy to determine modelling performance
DA	defect angle (angular position of the outer race defect)
defect 0	'new' bearing (although it already has microscopic defects)
defect 1	small incipient OR defect
defect 2	medium OR defect
defect 5	very severe OR defect

ESFFT	equi-sampled FFT (Fourier series)
FEM	finite element model
FLNN	fuzzy logic neural networks
GA	genetic algorithm
GIGO	(garbage in, garbage out)
HFD	higher frequency domain components (indicative of bearing defect)
HFRT	high frequency resonance technique
IR	inner race (of bearing). Also inner race fault.
Kurt	kurtosis
LCEPS	linear cepstrum
lim	limit as K approaches zero
LOO CV	leave one out cross validation
LOO XVAL	leave one out cross validation
LTCEPS	linear thresholded cepstrum
Mean_X^2	energy of high frequency signal of x axis vibration signal
MI	mutual information
MOS	mean of squares
N-N	bearing in normal condition
N-R	bearing in normal condition with rub fault
NNR	nearest neighbour rule
NNRC	a nearest neighbour rule with class membership
NO	bearing in normal condition

NW	bearing in normal condition (race is worn)
OR	outer race (of bearing). Also outer race fault.
RE	rolling element fault.
RF	rotational frequency
rpm	revolutions per minute
SA	synchronous average
SK	synchronous kurtosis
SNR	signal-to-noise ratio
SOM	self-organizing map
STD	standard deviation
SVM	support vector machines
TCEPS	thresholded cepstrum
VAR	variance
W-N	bearing in normal condition with weight added
W-R	bearing in normal condition with weight added and rub fault
WWW	World Wide Web

SYMBOLS

$\angle$	angle
$\otimes$	convolution
$\mathfrak{F}$	Fourier transform
Σ	summation
$*$	multiplication
$*$	conjugation (always in superscript to distinguish from multiplication), for example $G^*(f)$
$+$	positive part of time function (always in superscript to distinguish from addition), for example $G^+(f)$
$-$	negative part of time function (always in superscript to distinguish from subtraction), for example $G^-(f)$
^{0}	DC (0 Hz) part of time function (always in superscript to distinguish from number 0), for example $G^0(f)$
e^x	exponentiation
$\exp()$	exponentiation
$\cos()$	cosine
$\sin()$	sine
$\log()$	logarithm
$\log_{10}()$	logarithm (to the base 10)
$\log_2()$	logarithm (to the base 2)
$\text{sgn}()$	signum function

mean()	mean value function = $\sum \mathbf{x}/N$, with $\mathbf{x}$ a series of N numbers and mean($\mathbf{x}$) the mean value of $\mathbf{x}$
μ	mean value
σ	standard deviation
std()	standard deviation
var()	variance
<i>kurt()</i>	kurtosis ¹
<i>S</i>	skewness
SA()	synchronous average
<i>Sa()</i>	sinc function
FFT()	FFT (Fast Fourier Transform)
ESFFT()	equi-sampled FFT (Fast Fourier Transform)
E()	expected value function
MOS()	mean of squares
<i>j</i>	imaginary part of complex number
<i>U()</i>	unit step function
δ	delta function
<i>A</i>	impulse of amplitude <i>A</i>
<i>B</i>	$B = A/N$
<i>A_I</i>	$x_0^4 + x_1^4 + \dots x_N^4$
<i>B_I</i>	$2(x_0^2 x_1^2 + \dots x_{N-1}^2 x_N^2)$
<i>A₂</i>	$A_2 = \sum (x - \mu^4)$

B_2	$B_2 = \sum(x-\mu^2)$
$S_{\eta}(\omega)$	power spectrum
z^{-1}	Time delay of Z transform.
Z_n	normalized variable $Z_n = (X-\mu)/\sigma$.

¹ Some handbooks use k or K for kurtosis, but confusion with the variable k seems likely.

Chapter 1

1 Introduction

1.1 Overview of condition monitoring

Complex rotating machinery may be maintained in a periodic, time based fashion. Machinery is regularly inspected and potential faults noted. A disadvantage of this approach is that machinery in perfect working condition may be disassembled unnecessarily (which may induce faults), whereas small incipient faults may not be detected soon enough. By contrast, in condition based maintenance (CBM) the maintenance schedule is based on the actual condition of the machine. Provided that the condition of the machine is quantified via suitable measurements, CBM may be used to prevent expensive downtime and damages owing to unforeseen problems.

Vibration monitoring, which is a particular technology often employed in the context of CBM, has been used successfully in a wide range of condition monitoring applications [75]. Vibration monitoring is particularly useful for the identification of specific types of fault conditions that are commonly experienced in rotating machinery, e.g. imbalance, looseness or bearing defects. However, traditional vibration monitoring techniques may result in faulty diagnostic decisions in multiple or unstable fault conditions. Vibration monitoring allows for variation in speed ([3, 10]) and load ([80]). A secondary fault mechanism (e.g. dynamic load) may also influence the interpretation of vibrations associated with a primary fault mechanism ([25]).

The application of advanced signal processing techniques may reduce or even eliminate some of the environmental disturbances related to vibration signals ([3, 10, 80]), such as external or internal noise in the form of vibration sources, speed and/or amplitude variations. The application of such techniques aims to accentuate the 'signal' while simultaneously reducing the 'noise' or disturbances. The modified and enhanced signals could then be interpreted with a higher level of confidence using pattern recognition techniques.

The diagnosis of complex (e.g. multiple) fault conditions may require the simultaneous interpretation of a number of variables that in combination reflect the true condition of the machine. In such cases pattern recognition techniques may be applied to assist the operator in making consistent decisions regarding the status of a machine. Pattern recognition techniques have traditionally been used to make accurate classification decisions based on a number of extracted features. However, the successful application of pattern recognition techniques relies on the availability of a sufficiently large data set that is representative of all fault conditions that may occur during the operational life of a particular type of machine. In practice, it may often be expensive or even impossible to compile such a data set from experimental data. The use of simulated models may then be considered to expand the available experimental data set artificially, using expert knowledge about the nature of expected faults to construct fault models.

1.2 Background

Vibration analysis is a widely used technique for detecting many kinds of mechanical and structural problems. Faults in rotating machinery in particular may be detected using vibration (and sometimes current²) analysis. Owing to the widespread use of bearings, specialized techniques of bearing fault detection have been developed in the past. Other mechanical faults that may be detected include imbalance, misalignment, rotor and fan defects, resonance, mechanical looseness, gear defects, electrical problems, belt drive problems and others [39]. This thesis examines the detection of an outer race defect in a roller bearing under varying operating conditions (i.e. a dynamic load). The relevance of this problem is also discussed in [6] (although in the case of static loading only). The measurements used to reflect these fault conditions typically include radial and axial vibration measurements using accelerometers that are mechanically mounted as close to the defect as possible.

² In harsh environments in industrial plants sensitive accelerometers may not be robust enough, in which case current measurements may be the only option.

Typical features that are usually extracted on the basis of *a priori* knowledge include rotational frequency (RF) components in the radial x , y and axial z directions. Furthermore, the characteristic defect frequencies corresponding to the inner race and outer race may be calculated, based on bearing geometry. The magnitude of these components may then be used as additional features for fault detection.

In order to compile a set of measurements of a faulted bearing and normal bearing, the usual procedure is to dismantle the machinery to introduce a specific fault condition on a new bearing. Yet, this approach is undesirable, as the dismantling of the machinery changes environmental conditions (hence the vibration signals), impeding direct comparisons of before and after measurements. A novel approach has been proposed in this thesis for this problem, in which the outer race defect is rotated around the loading zone (using a step size of 45°). When the defect is near the top of the bearing, the defect signal is very weak (as described in [55, 84]). As the defect is moved closer to the centre of the loading zone, the defect signal becomes stronger [55, 84]. Consequently, a normal and damaged bearing may be simulated without dismantling the machinery physically. In principle, the same technique could also be used to apply a known defect on the outer race of a bearing, which may be used as a calibration signal subsequently (when the defect is in the loading zone) or as usual (when the defect is out of the loading zone).

Common signal processing techniques that may be used for bearing fault detection include envelope detection and signal averaging.

The high frequency resonance technique (HFRT) of envelope detection is described by McFadden and McFadden in [50]. A bearing defect excites a high frequency resonance at the characteristic defect frequency in the same way that a bell rings when struck by a hammer. Thus the envelope of the high frequency resonance provides information about the (low frequency) modulating function. The signal from the accelerometer is amplified and bandpass filtered around a resonance. An envelope detector consisting of a non-linear element (a half- or fullwave rectifier, or raising the signal to a power such as with a squarer) is subsequently applied to extract the envelope of the signal. The frequency component is then analysed at the characteristic defect frequency.

The technique of signal averaging is also discussed by McFadden *et al.* in [50, 53]. Only frequency components synchronous to the trigger frequency are passed when the number of averages is large [50]. In the frequency domain the technique of signal averaging corresponds to a comb filter at multiples of the trigger frequency (provided of course that enough signal averages are taken). The signal average evaluated at the characteristic defect frequency provides information in the time domain about the defect pulses. However, the real defect frequency may differ from the characteristic defect frequency owing to geometrical errors, for example. For this reason, it is desirable to trigger from a signal synchronized to the characteristic defect frequency (as described by McFadden and Toozy in [53]). Yet, this is often not feasible in an industrial setting, which limits the practical use of this technique.

McFadden and Toozy state that a real defect may be modelled as the superposition of an infinite number of point defects [53]. It is therefore clear that the interpretation of a single frequency component in the envelope spectrum may be misleading [53]. It is shown in this thesis that all the frequency components in the envelope spectrum have to be evaluated together for accurate diagnosis, and the exact relation is elucidated between the envelope spectrum and the signal average.

The early detection of degradation in mechanical equipment may help to reduce potential damage associated with a catastrophic failure, which may even lead to the destruction of a plant [75].

As a result of the shortage of skilled, on-site personnel, much effort has been focused on the automated recognition of vibration signals [75] aimed at developing suitable preprocessing techniques to enable classification of vibration signals by an automated system.

Although positive results have been reported in literature [75], the condition of the machinery is not quantified. This makes it difficult to design a maintenance programme that can identify and correct potential problems before they develop. Furthermore, the generalization performance of the classifier on other machinery is not known from literature and would depend on a number of factors related to the similarity in electrical

and mechanical construction, environment (influencing, for example, speed, static and dynamic load) and operational factors. Pattern recognition techniques typically use a small number of extracted features from the data set to make an accurate classification decision. The feature(s) may be non-linear transformations of the original measurements. The feature(s) should ideally be linearly related to the fault condition. Feature extraction will usually implement a non-linear transformation of the original measurements to establish a more direct relationship between the feature set and the fault condition. In practice, a number of features are extracted, and feature selection is used to limit the number of features in the pattern recognition technique. Feature selection aims to discard irrelevant features (those that are unrelated to the fault condition under consideration) and to retain the informative features that are most important for classification (or modelling). The quality of the feature set has a direct bearing on the classification performance (e.g. [7, 25, 30, 87]).

A data based technique generally requires a number of samples exponentially proportional to the number of input features, assuming mutually independent features. Limiting the number of features, for example, by using the first few harmonics only, reduces the number of samples required, but also reduces the potential modelling capability of the data based techniques. Feature reduction with minimum information loss is therefore required to minimize the data requirements. Alternatively, a model may be used to supplement the set of sparse measurements through simulation.

For meaningful interpretation, the vibration measurements should be acquired under the same conditions. The transfer function from a vibration source to the sensor (e.g. an accelerometer) depends on various factors, which make absolute calibration problematic. Some guidelines on acceptable vibration standards are available, but are often too conservative for industrial use [5]. Hence appropriate vibration levels are often adjusted for each individual case.

Correct interpretation of vibration data is dependent on expert knowledge of the specific machinery and industrial plant. The shortage of skilled personnel has prompted research into the automated on-site processing and diagnosis of vibration measurements [75]. An

automated technique could in principle perform automated diagnosis if the following conditions are met:

- All information, which a human expert requires for a decision, is available in a suitable format. This would typically include all spectrum and time domain information, as well as *a priori* knowledge about the plant. Information about the operating conditions, vibration levels, machinery, and so forth should all be available in a standard database format, independent of specific implementations (e.g. MIMOSA export format).
- Expert knowledge describing the essential characteristics of the database, i.e. the best set of features to use in the automated diagnosis of machinery, is required to construct a relatively small set of features that still contains the primary information that reflects the status of operation of the machinery. A small feature set that reflects the underlying contributors to faulty operation only, while ignoring various sources of noise, would enable better generalization performance.
- Preprocessing of data, involving standard data mining techniques, aids the classifier in extracting relevant information from the multidimensional data set [71], e.g. removing outliers and obvious sources of noise. Such techniques are typically employed as part of feature extraction.
- The diagnosis of the automated technique may be improved by a condition-monitoring programme incorporating closed-loop feedback. Since a specific failure mechanism may occur very rarely, the database would often be unbalanced containing large numbers of observations reflecting normal operation and very few observations indicative of faulty operation. This implies a need to augment certain classes of failure data with additional data, perhaps obtained by experimental design, if appropriate, or by simulation.

A set of vibration measurements, taken under laboratory conditions, enables specific defect mechanisms to be identified and the experimental configuration to be controlled. Owing to the complex nature of an industrial plant, it is more difficult to identify and diagnose the condition of rotating machinery under normal operational conditions.

By trending the vibration measurements over time (establishing a baseline measurement and monitoring deviations) it is possible to establish the onset of failure mechanisms, provided that the measurements are taken under exactly the same operating conditions. Hence the residual life of the machinery may be estimated [101].

To solve the problem of statistical variability in sparse data sets and to enable better interpretation of the available data, it is proposed that a combination of experimental procedures, data based modelling, expert knowledge and automated classification is used.

An automatic pattern recognition technique would typically require a systematic database of previous machine failures to train a classifier in detecting machinery problems. The feature set should contain enough training examples to match the complexity of the classifier; as a rule of thumb 10–100 samples are used for each free parameter in the classifier [18].

The automatic classification of vibration measurements may be implemented using a variety of techniques. Traditionally two types of technique were applied to problems of this nature:

- Rule based techniques are easy to understand. However, translation of the knowledge of a domain expert into a rule set or expert system, sufficient to cover all cases, may be problematic. The rule set is frequently inflexible, as the number of rules grows exponentially with the complexity of the problem. An expert system, while effective in modelling logical structure and relationships, is not flexible in adapting to new data. Furthermore, it is difficult to extract accurate domain knowledge, and changes in the rule set may be problematic. The modelling of environmental conditions often requires many rules.

- Data based techniques use training set data to derive a model, explaining the input-output behaviour of the system. Large data sets are required to cover the multivariate feature space adequately, especially with non-linear relationships. A sufficiently large and representative data set is usually not available in practice.

A vibration database would require many measurements to cover the multidimensional landscape corresponding to the plant parameters adequately. Alternatively, a good model may be used to supplement the data set. The incorporation of *a priori* knowledge helps to reduce the data requirements. However, the classifier also needs to learn from the data instead of being confined to fixed *a priori* relationships. Some techniques may incorporate rules readily, but incorporating quantitative models is not that easy.

1.3 Outline of this thesis

A brief outline of this thesis is as follows:

Chapter	Heading	Pages	Description
Chapter 2	Literature survey	11–42	Brief literature survey on condition monitoring and artificial intelligence techniques. Literature references to the specific problem statement of this thesis are listed in section 2.5.3 on p. 40.
Chapter 3	Problem statement and contributions of thesis	43–58	Brief introduction to and background on the specific problem statement of this thesis. The problem statement of this thesis is given in section 3.6 on p. 53. A brief summary of potential contributions of this thesis is presented in section 3.7 on p. 55.

Chapter 4	Experimental setup	59–74	The experimental setup for use in the problem statement of this thesis is explained.
Chapter 5	Theoretical development and results of pattern recognition techniques	75–114	Application of basic pattern recognition techniques (including simple signal processing techniques for feature extraction) in the Vib demo data set 2. Some novel techniques are proposed.
Chapter 6	Results of signal processing techniques	115–212	This chapter focuses on the application of advanced signal processing techniques in the Vib demo data set 3 (including measurements using very small bearing defect signals in the context of dynamic load) as well as simulated data sets to improve the signal-to-noise ratio (SNR). The investigation of speed and amplitude variations is emphasized in bearing fault detection.
Chapter 7	Results of bearing test rig	213–252	Experimental results obtained on the Vib demo data sets 2 and 3 are examined. A variety of features are extracted for bearing fault detection in the presence of dynamic load.
Chapter 8	Conclusion	253–270	Final conclusions of this thesis are made. Some aspects left for future work are

			discussed in section 8.6 on p. 263.
Chapter 9	Bibliography	271–282	Literature references consulted in writing this thesis are acknowledged.
Chapter 10	List of publications	283–286	A list of publications of the author of this thesis is supplied.
Chapter 11	Appendices	287–500	Appendices of this thesis contain additional detailed information that has been removed from the main text for ease of reading and comprehension.

Chapter 2

2 Literature survey

2.1 Introduction

This literature survey aims to present the extent of knowledge in the specific problem area as accurately as possible. It stands to reason that any literature survey is subject to the following constraints:

- **Time** of literature survey: it is not feasible to redo the literature survey every year (although important new literature should be incorporated where possible). Hence the results of this literature survey reflect literature before the start of the PhD study in 1998 only (although more recent literature has been incorporated where possible). It is inevitable that new literature may replace or update existing knowledge in the problem area.
- **Scope** of literature survey: in the context of a developing country such as South Africa and in a rural area such as Potchefstroom literature references are more difficult to access. Interlibrary loans are often costly and time consuming. Although every effort has been made to obtain information from reliable and authoritative sources, it is inevitable that some omissions or errors could have been made.

The aim of this thesis is certainly not meant to be encyclopaedic. Instead a small **sample** of the complete literature survey is presented in this thesis. Although this selection is perhaps to some degree guided by subjective factors (such as the standing of authorities in the specific problem domain community and relevance to this thesis), it is hoped that the selection provides sufficient background to justify the purpose of this thesis.

2.2 Condition monitoring

Rao [75, 76] discusses and overviews the broad area of condition monitoring and methodologies for fault detection. The signal processing and pattern recognition techniques developed in this thesis, although applied to a very specific problem domain, address a reasonably generic problem (as outlined in the problem statement in section 3.6 on p. 53). Hence the techniques might well be used for an even broader application area, with modifications to suit the characteristics of specific problem domains and signal statistics. It is believed that the methodology outlined in this thesis provides a general framework for the successful application of pattern recognition techniques in condition-monitoring applications. Only vibration-monitoring techniques (chapter 3 of [75]) are considered in this thesis, which are applied to the recognition of multiple fault mechanisms in a bearing test rig³.

The literature related to gearbox defects, more specifically the propagation of bearing defects through a complex set of gears as found in a gearbox, falls outside the scope of this thesis and will not be discussed.

2.3 Techniques for fault detection

2.3.1 Data collection

The collection of a consistent and reliable set of measurements, covering the physical and electrical properties of the machinery, is vital for a reliable diagnosis of faults [24, 76]. A *priori* knowledge regarding the factors influencing vibration measurements is important to define a meaningful and compact feature set. Statistical considerations indicate that increasing the sample size may reduce the variance of results. The sample size is especially important to discover the potential non-linear relationships inherent in the data.

³ Preliminary investigation shows that acoustic measurements close to the bearing test rig may also be used for fault detection if the signal-to-noise ratio is 'high enough'. To date no success has been obtained using the current measurements. One possibility is that the belt drive mechanism from the shaft to the motor alters the transfer function of the defect to such a degree that fault detection is no longer possible, or would require very specialized techniques. Hence these aspects are left for future work.

The discovery of variables with predictive power, using non-linear analysis, facilitates the modelling process.

2.3.2 Linear and non-linear processing techniques

The benefits of non-linear analysis techniques versus standard linear data analysis techniques depend on the complexity of the problem. Neural network classifiers, using non-linear feature mapping and sigmoidal basis functions, may identify multiple faults in vibration signal analysis [31, 32, 33, 42, 43, 44, 47, 49, 60, 76]. The degree of non-linear correlation inherent in a data set is important in the design of a classifier. It may be estimated by an entropy-based algorithm such as the mutual information criterion [90]. This technique has been demonstrated in analysing the non-linear contribution of variables in a large plant.

If small data sets are available but many features are used, overfitting of the model is highly likely. Regularization techniques may be employed to penalize the degree of non-linearity in the initial modelling stages. After progressing through a number of iterative data modelling and experimental verification stages, in which the confidence in the modelling process increases, the complexity and degree of non-linear interactions of the classifier may be increased gradually. The aim is to develop an accurate model for the automatic classifier, while keeping the model as simple as possible.

2.3.3 Feature selection and extraction

The vibration signals form a multivariate feature space. The required number of training samples for a classifier generally increases exponentially as a function of the number of features, assuming uncorrelated data [24]. Furthermore, the performance of the classifier is closely linked to the quality of the features.

The extraction of a compact feature set, which may still capture most of the correlation inherent in the original sample space, is thus crucial in a multivariate setting. Suitable feature extraction methods highlight the important discriminating characteristics of the data, while simultaneously ignoring the irrelevant attributes (i.e. noise) [24].

The frequency domain provides a useful feature set for machine diagnostics [75]. Machinery defects are related to specific frequency domain features [64, 75]. The frequency domain is well suited to the detection of periodic machinery vibrations. Impulsive vibrations are better analysed in the time domain than in the frequency domain. The wavelet transform analyses a signal jointly in the time frequency domain, subject to the uncertainty principle. The uncertainty principle states that an increase in time resolution results in a decrease in frequency resolution and vice versa.

2.3.3.1 Weibull distribution

The Weibull distribution is useful in the statistical analysis of the vibration signals, especially with skewed distributions [64].

2.3.3.2 Wavelet transform

Preprocessing by the wavelet transform has been shown to lead to increased compression ratios, better noise reduction and increased classification performance. Hence the wavelet transform could be used to reduce the dimensionality of the vibration signals.

2.3.3.3 Cepstrum analysis

The aim of cepstrum analysis is to detect periodicity in the spectrum of vibration signals. It is well suited as a tool for the detection of families of harmonics with equal spacing. It is defined as the inverse Fourier transform of the logarithm of the forward Fourier transform of a time signal $x(t)$ [64]

$$C_{\text{cex}}(\tau) = \mathfrak{F}^{-1} \{ \log_{10} \mathfrak{F} \{ x(t) \} \} \quad (1)$$

2.3.3.4 Envelope spectrum analysis

Envelope spectrum analysis is a technique especially suitable for the early detection of damage to rolling element bearings. The technique consists of a bandpass filter that reduces frequency components unrelated to the bearing. The signal is then enveloped by fullwave rectification and lowpass filtration before being spectrum analysed [64]. In [21] the use of the squared envelope and higher powers is examined and conditions derived, as a function of the slip or fluctuation in shaft speed, when the SNR may be improved. Other

techniques that could be useful, but are not discussed in this thesis (refer to section 8.6 'Future work' on p. 263), is adaptive line enhancement (ALE) [106], adaptive noise cancellation (ANC) and self-adaptive noise cancellation (SANC) [3], higher-order spectral (HOS) processing [58, 62] and blind source separation (BSS) [58].

2.3.4 Pattern recognition

The performance of a non-linear classifier, such as a neural network, is closely dependent on the number of training examples relative to the degrees of freedom (complexity) of the classifier [18, 24]. As a rule of thumb, the number of samples should be 10 to 100 for each independent feature. The combination of techniques based on fuzzy logic and neural network techniques may lead to an improved classifier that has the following advantages [99]:

- Fuzzy rule generation and inference provide more transparent understanding of the internal operation of the classifier, compared with purely neural classifiers. *A priori* expert knowledge of the vibration mechanisms may be incorporated in the design phase as a set of fuzzy rules.
- Vagueness and ambiguities may be handled graciously. The degree of membership with regard to the output class provides an indication of the certainty of presence of a mechanism, rather than a crisp 'present' or 'absent' answer.
- Multiple rule sets may be stored, e.g. different machines could use different rule sets.
- The output signal represents the degree of membership with regard to a specific class. By monitoring the confidence of the FLNN in a prediction, the extrapolation of the FLNN may be established beyond training set data. In regions of the input space with sparse data the classifier could request additional training samples.
- Hardware implementation is available for fast online decision-making.

- The number of free parameters that may be adjusted is generally smaller than for the FFNN, reducing the chance of overfitting small data sets with non-linear relationships and reducing the requirement for training samples.

Jack and Nandi [30, 32] examine the use of support vector machines (SVM), a pattern recognition technique, in the detection of bearing faults (five fault conditions) in a test rig. For a detailed discussion of this technique refer to [30, 32] and other texts on pattern recognition techniques. The aim of the SVM technique is to find the largest separating hyperplane (support vectors). The kernel function maps the data (which might be non-linearly separable), using a non-linear transfer function, into another dimension in which the classes could be linearly separable with an appropriate choice of parameters. Jack and Nandi [30, 31] also examine the important question of feature selection using a genetic algorithm (GA). Their results [30, 31] indicate that the SVM comes close to the ANN without GA feature selection. Using GA feature selection the SVM and ANN have comparable performance.

2.3.5 Experimental design

The reduction of the feature set to the minimum required for acceptable modelling is important in the design of structured experiments. The influence of a set of experimental variables on the response variable(s) is determined by conducting a series of experiments. The resulting response surface may be used in a simulation to augment the existing data set. Since the required number of experiments is exponentially proportional to the number of experimental variables to be considered, the smallest possible number of features should be used. An exhaustive evaluation of all possible experiments would be prohibitive. More sophisticated experimental design techniques would therefore be required to minimize the required number of hypercubes to cover the multivariate space.

2.3.6 Signal processing techniques

The direct analysis of spectral analysis information as well as the trending of spectral components have been used for a long time in machinery fault diagnosis [86]. As it is difficult to analyse and interpret the phase information from spectral analysis, which often

contains critical diagnostic information, much interest has recently been expressed in combined time-frequency analysis methods that preserve phase information (e.g. [25]).

A brief literature survey of techniques applicable to fault detection in roller bearings is given in [6]. Signal processing techniques described in [6] include the amplitude spectrum (although the low frequency components are often contaminated by noise components complicating diagnosis); high frequency resonance technique (HFRT) where the noise of low frequency vibrations is filtered out, enhancing the signal-to-noise ratio; enveloping; cepstrum analysis; adaptive line enhancement (ALE); and normalizing. Signal averaging, as described by McFadden and McFadden in [50], may be used to enhance the signal-to-noise ratio by filtering out all asynchronous sources of vibration relative to a trigger signal.

Tandon and Choudbury [85] refer to a number of additional signal processing techniques in their review, including the adaptive noise cancelling (ANC) technique; the self-adaptive noise cancelling (SANC) technique, which has been proposed as a result of difficulties in obtaining a reference signal using the ANC technique; cepstrum analysis; and bispectral analysis.

They describe research with non-contact type displacement or proximity transducers as well as acoustic emission sensors in their article [85]. Research results are reported on the shock pulse method (in which a mechanical filter at high frequency – around 32kHz – filters out most of the noise). They also mention an automated diagnosis system based on pattern recognition and briefly explain the application of neural networks to fault detection.

Combined time-frequency transformations, such as the wavelet transform, may also be used to enhance specific features in the signal.

2.3.7 Non-parametric classification

Only the class of non-parametric classifiers will be considered in this thesis, i.e. no assumptions will be made about the underlying distribution of classes. One of the simplest classifiers is the nearest neighbour rule, in which the class of the closest sample is assigned as the class label. The one nearest neighbour rule (1NNR), despite its simplicity,

achieves an error rate smaller than twice the optimal Bayes error rate [77]. It is also possible to include a 'reject' option in the kNNR rule, in which a classification decision is withheld when a clear majority vote is not obtained [77]. Some of the disadvantages of kNNR include computational and storage requirements, choice of metric and preprocessing requirements (normalization and removal of irrelevant input variables). The classification capabilities of the 1NNR rule will be combined with a feature selection process in this thesis. The heuristic forward feature selection (FFS) [7, 77] algorithm is used with the 1NNR as classifier. This is somewhat similar to the gamma test algorithm [36], although the performance of the gamma test algorithm is still unknown on classification problems.

Artificial neural networks (ANN), consisting of layers of neurons interconnected by a series of weights, may also be used for non-parametric classification. A brief description of ANN, more specifically multilayer perceptrons (MLP) follows.

2.3.8 Artificial neural networks (MLP)

A feedforward network is a network in which vertices may be numbered so that all connections go from one vertex to another with a higher number. In practice, the vertices are arranged in layers, with connections to higher layers only [77]. An MLP consists of layers of perceptrons. The input layer stores the value of the input variables. The hidden layer(s) process the input values sequentially, and the output layer stores the final output values. The general form of the regression function computed by the MLP (for one hidden layer) is [77]

$$y_k = f_o \left(\sum_{i \rightarrow k} w_{ik} x_i + \sum_{j \rightarrow k} w_{jk} f_h \left(\sum_{i \rightarrow j} w_{ij} x_i \right) \right) \quad (2)$$

where x_i are the input variables and y_k are the output variables of interest. The bias terms are represented by the weights w_{0j} . The functions f_o and f_h may be a logistic $f(x) = e^x / (1 + e^x)$ or a threshold function. The MLP with one hidden layer may in principle implement any desired input-output mapping, provided that sufficient capacity is available (i.e. the number of neurons in the hidden layer is sufficient) [7, 77]. In most real-

world problems the size of both the data set and the processor resources is limited. It is therefore important to optimize the parameters of the ANN for best performance.

The learning algorithm presents each training set example and adjusts the parameters of the ANN to minimize the chosen error metric. Online learning (which is often done off-line in practice) adjusts the parameters after presentation of each sample. Batch or off-line learning adjusts the parameters after presentation of all samples. Off-line learning requires the learning parameter h (which determines the overall weighting applied to each weight) to approach zero [77].

The off-line learning algorithm is given by [77]

$$w_{ij} \leftarrow w_{ij} - \eta \left[\left(1 - \alpha \frac{\partial E}{\partial w_{ij}} \right) + \alpha (\Delta w_{ij})_{previous} \right] \quad (3)$$

and the online version by

$$w_{ij} \leftarrow w_{ij} - \eta' y_i^p \delta_j^p + \alpha' (\Delta w_{ij})_{previous} \quad (4)$$

The function to optimize is frequently chosen as a sum of the squares criterion

$$E(w) = \sum_p \|t_p - f(x_p; w)\|^2 \quad (5)$$

Back-propagation may be used to determine the optimal change in weights for one or more hidden layers. The first partial derivative

$$\frac{\partial E}{\partial w_{ij}} = y_i \delta_j \quad (6)$$

with

$$\delta_j = \frac{\partial E}{\partial x_j} = f'_{,j}(x_j) \frac{\partial E}{\partial y_j} \quad (7)$$

supplies the incremental change in the performance criterion relative to the output layer weights. For logistic output units with the least squares criterion

$$f'(x) = y(1 - y) \quad (8)$$

For units in earlier layers

$$\delta_j = f'_j(x_j) \frac{\partial E}{\partial y_j} = f'_j(x_j) \sum_{k: j \rightarrow k} w_{jk} \delta_k \quad (9)$$

where the sum over units k is fed by unit j .

2.4 Vibration metrics

The selection of the smallest possible feature set, which is able to retain the classification accuracy, is of paramount importance before any classifier is trained [18, 24, 77]. Owing to an exponential increase in computational complexity, an exhaustive evaluation of all feature sets is generally not possible. Furthermore, the same feature set might yield different classification results using different classifiers, indicating the dependence of the feature set on the classifier. To achieve better performance, it is possible to adjust parameters relating to the definition of the features in a closed-loop optimization process [96]. Alternatively, a heuristic search strategy may be used, e.g. forward selection (starting with a small number of features and then adding the best feature(s)) or backward selection (starting with a large number of features and then deleting the worst feature(s)) [7, 77].

Since some of the important vibration features are directly related to the speed of rotation, it is important to remove any speed dependence associated with the feature set. This could be done using a multirate filter for sampling rate alteration to normalize the features relative to the rotational frequency. A simpler averaging technique has been used in this thesis, as the speed variation was considered to be sufficiently small to be disregarded. It is also important to normalize each feature with regard to the range of possible values for that feature. Although the neural network may, in principle, perform scaling by adjusting the internal weights linking the input vector to the hidden layer, it nevertheless complicates the training process unnecessarily and may easily result in local minima. Best

results are achieved if the range of input vector values is below one (which is the maximum output a sigmoidal basis function can obtain). Classification results have been obtained using mutual information (MI) and the nearest neighbour rule (NNR) for feature selection. The results obtained from the feature selection process correlate well with the features chosen in this thesis.

2.4.1 Time domain features

The following time domain features were extracted from the data: mean, root mean square (RMS), crest factor, variance, skewness and kurtosis. A brief description of each follows:

The mean value of a function $x(t)$ over an interval T is

$$\bar{x} = \frac{\int_0^T x(t) dt}{T} \quad (10)$$

The RMS value of a function $x(t)$ over an interval of T is

$$X_{rms} = \sqrt{\frac{\int_0^T x(t)^2 dt}{T}} \quad (11)$$

The crest factor is the ratio of the peak level to the RMS level

$$CF = \frac{X_{max}}{X_{rms}} \quad (12)$$

The variance is the mean square value relative to the mean

$$\sigma^2 = \frac{1}{T} \int_0^T [x(t) - \bar{x}]^2 dt \quad (13)$$

The skewness is the third statistical moment of a distribution

$$S = \frac{1}{\sigma^3 T} \int_0^T x^3 dt \quad (14)$$

The kurtosis is the fourth statistical moment of a distribution

$$K = \frac{1}{\sigma^4 T} \int_0^T x^4 dt \quad (15)$$

Although all of these features may not be accurate indicators of fault conditions, it is useful to investigate their performance to reflect stable and deteriorating fault conditions in a system where multiple faults are present.

2.4.2 Frequency domain features

It is well known that defects in rotating machinery may be monitored using vibration frequency domain analysis. In this case, certain features from frequency domain analysis are generated to predict multiple faults in the experimental setup. These features may be calculated using common condition-monitoring techniques. For frequency domain analysis, the frequency band of 5–10kHz was investigated. Time signals were recorded at a sampling rate of 25.6kHz. For spectrum calculation, typically 20 processing averages with a resolution of 1–2Hz were used. The following features were extracted from a frequency domain analysis for the Vib demo data set 2:

- The amplitude of the vibration spectrum at rotational frequency (RF), in a horizontal direction
- The amplitude of the vibration spectrum at RF, in a vertical direction
- The ball pass frequency on the outer race (*BPFO*) of the defective bearing
- Higher frequency domain (HFD) components indicative of bearing defect

2.4.3 Rotational frequency

According to Taylor [86], force imbalance in a rotor may be detected using vibration frequency domain analyses. A peak in the spectrum at the running speed frequency of the shaft will indicate imbalance. The amplitude of acceleration at the running speed frequency may be related to the amount of imbalance in the rotor. The rotational speed of the system was 27.1Hz (1626rpm) for Vib demo data set 2. An increase in vibration amplitude at 27.1Hz therefore indicates an increase in the amount of imbalance. The phenomena are presented in figure 1 where an increase in the amplitude is clearly visible at 27.1Hz. The vibration amplitude at 27.1Hz was taken as a feature for fault identification. The measurements in both the horizontal and vertical directions were processed and used as features.

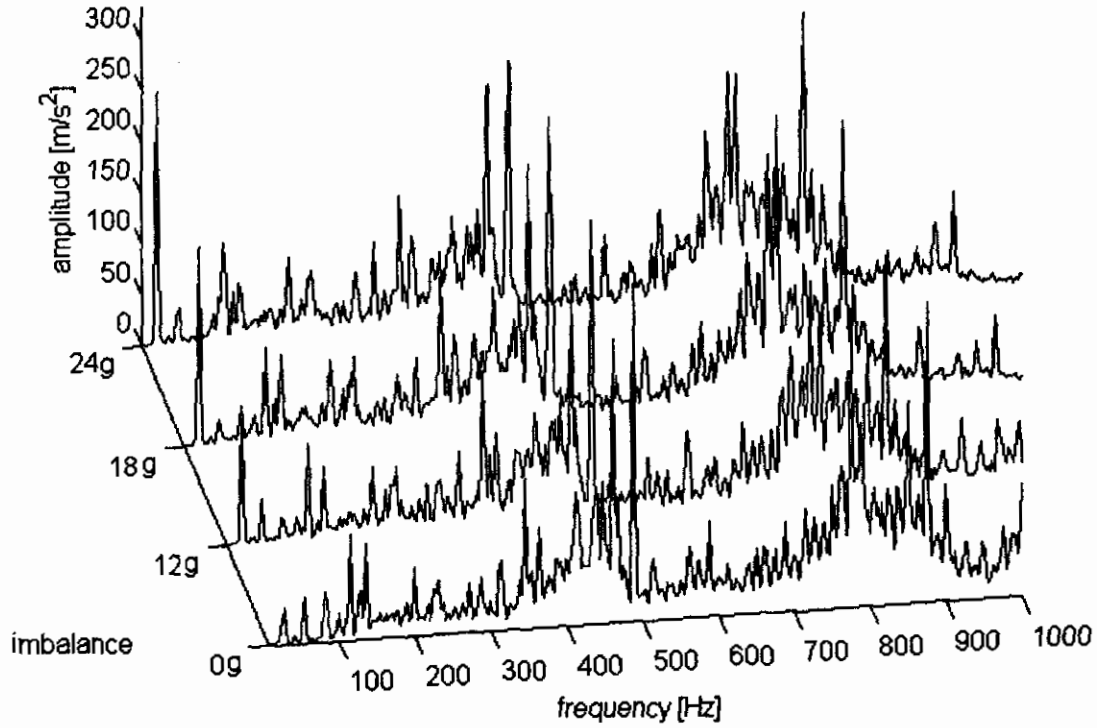


Figure 1: Increasing imbalance on channel 1 (0–1kHz) on Vib demo data set 2

2.4.4 Ball pass frequency of outer race

The ball pass frequency of the stationary outer (*BPFO*) race is defined as the frequency of the balls passing over a single point on the outer race. It is described by the number of balls multiplied by the difference in the rotational frequency between the cage and the outer race. The presence of a peak at this frequency in the spectrum of the vibration signal indicates a defect on the outer race of a bearing. The ball pass frequency on a stationary outer race may be determined as follows [86]:

$$BPFO = \frac{N_b \omega_l}{2} \left(1 - \frac{d}{D} \right) \quad (16)$$

where

BPFO = ball pass frequency on the outer race [rad/s]

N_b = number of rolling elements

d = roller diameter [m]

ω_i = angular velocity of the inner race [rad/s]

D = pitch diameter of the bearing [m]

The *BPFO* of the bearing with the outer race defect was calculated at around 114Hz for the Vib demo data set 3⁴. The amplitude of the vibration spectrum near 114Hz in the envelope spectrum was accordingly taken as a feature in the horizontal direction (a tolerance of a few per cent around the calculated defect frequency ensures that the peak value is captured regardless of minor speed variations).

2.4.5 Higher frequency domain components

The natural frequencies in a rotating machine between 2.5kHz and 20kHz are prone to excitation from the impact of the defects on the elements of the bearings. The energy in the higher frequency domain (HFD) therefore indicates the presence of bearing defects in a machine. Bearing defect frequencies modulate the natural frequencies of the system, which are being rung. The bearing defect frequencies may be recovered by demodulating the vibration signal. A spectrum of the demodulated vibration signal shows peaks at the bearing defect frequencies and their harmonics. The energy in the HFD indicates a bearing defect before the bearing defect frequency appears in the spectrum of the vibration signal.

The HFD components for the experimental setup are typically in the 5–10kHz band. Figure 2 illustrates the influence of the induced imbalance on the defective bearing. Increasing imbalance seems to have the effect of increasing the noise floor of the HFD components. It was decided to generate a number of features, which will be representative of the energy within certain HFD bands. These features are the sum of the vibration spectrum amplitudes between frequency values such as 5–10kHz, 6–8kHz, 7–10kHz, etc.

⁴ The defect frequency showed some variation around 114Hz, seemingly related to the size of the defect and the position of the defect in the loading zone.

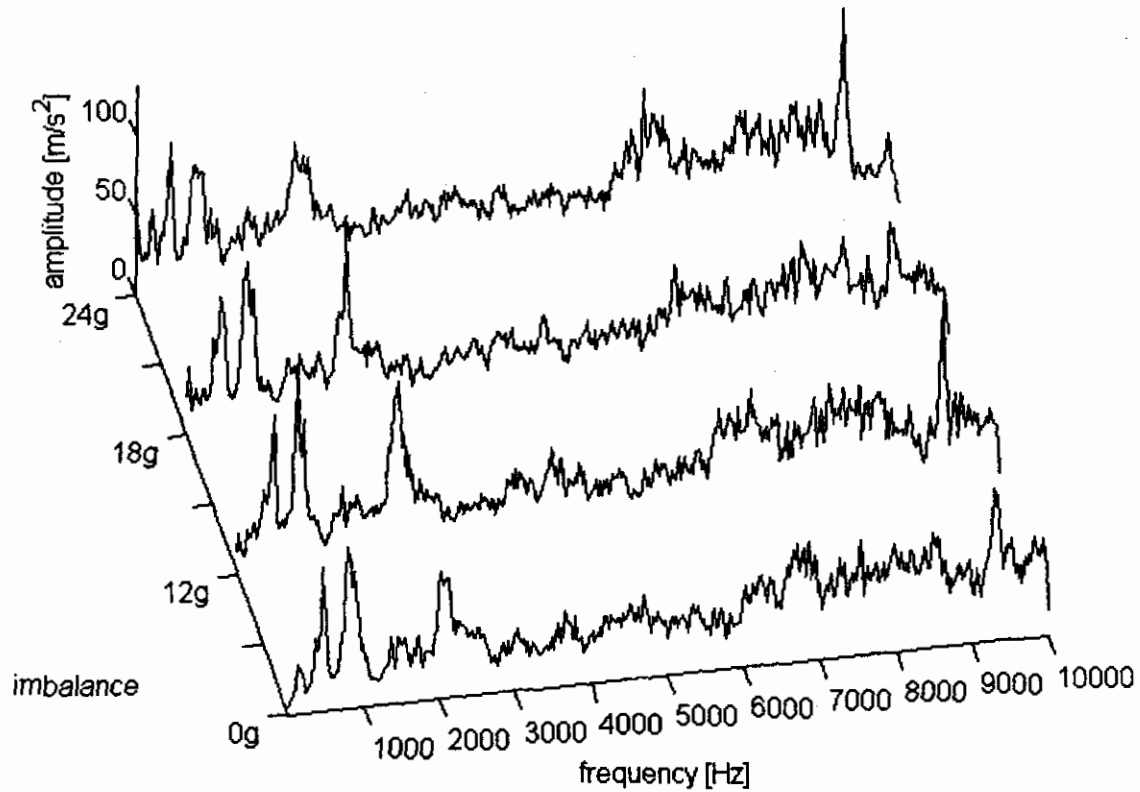


Figure 2: Increasing imbalance on channel 1 (0–10kHz) on Vib demo data set 2

2.4.6 High frequency resonance technique (HFRT)

The HFRT bandpass filters the signal around a suitable demodulation frequency, followed by rectification and lowpass filtering (envelope detection) [72]. The normalized ratio of the demodulation peak in the demodulation spectrum relative to the carpet level provides a measure of the defect growth and is regarded as the best feature of bearing defect evolution [25, 63, 78].

In [2] the effect of non-linear transfer functions on the amplitude density function and power spectrum of band limited ‘white’ noise is derived. On p. 222 of [2] the result of $y = bx^2$ is given as

$$S\eta(\omega) = S_1(\omega) + S_2(\omega) + S_3(\omega)$$

where

$$S_1(\omega) = 4b^2 A^2 (\Delta\omega)^2 \delta(\omega) \text{ for } \omega = 0 \quad (17)$$

$$S_2(\omega) = 4b^2 A^2 (\Delta\omega - |\omega|) \text{ for } |\omega| \leq \Delta\omega$$

$$S_3(\omega) = 2b^2 A^2 (\Delta\omega - (|\omega| - 2\omega_0)) \text{ for } 2\omega_0 - \Delta\omega < |\omega| < 2\omega_0 + \Delta\omega$$

where the power spectrum $S\eta(\omega)$ has a constant value A , centred at ω_0 and bandwidth of $\Delta\omega$. It should be noted that the above expressions are based on the assumption of continuous signals. It is also unclear whether information from the phase spectrum has been incorporated into the derivation of the above expressions. A similar expression for sampled data (for example after application of the FFT) could unfortunately not be obtained in literature. In section 6.10.3 on p. 171 this question is addressed by a heuristic formula proposed in this thesis.

2.4.7 Cepstrum analysis

Cepstrum analysis is used to detect periodicities in the spectral analysis of a signal, as well as to separate the effect of varying transfer functions [72]. It may also be used to enhance the signal, e.g. echo removal by cepstral editing [72]. The real cepstrum is defined as

$$RCEPS(x) = \text{real}(\mathfrak{F}^{-1}(\log|\mathfrak{F}(x)|)) \quad (18)$$

where $\mathfrak{F}$ is the Fourier transform and $\mathfrak{F}^{-1}$ is the inverse Fourier transform of the input signal.

2.5 Fault detection on roller bearings

Condition monitoring of critical machinery has many economic benefits. A number of conventional vibration analysis techniques exist by which certain faults in rotating machinery may be identified and repaired to prevent catastrophic failure of a component. In this context, it is important to be able to discriminate between stable and deteriorating fault conditions. However, conventional techniques used in the condition monitoring of rotating machinery may under certain circumstances indicate deteriorating fault conditions in the case of stable fault situations. This may occur in a system where multiple fault

conditions are present, such as bearing defects, imbalance and looseness. A deteriorating fault condition such as imbalance may influence the interpretation of a rolling element bearing condition as a result of an increase in the radial load on the bearing.

In general, the defect may either remain stationary (such as when the outer race – or inner race – is fixed, the defect is stationary) or start moving (such as when the inner race – or outer race – rotates, the defect also rotates). When the defect rotates, the defect pulses are modulated [55, 84]. If the defect remains stationary, the defect pulses should also remain constant in amplitude [55, 84].

Various authors have investigated multiple fault conditions using vibration monitoring [34, 42, 48]. In [69] Paya *et al.* examine the condition of a model drive-line, consisting of various interconnected rotating parts, including an actual vehicle gearbox, two bearing housings and an electric motor, all connected via flexible couplings and loaded by a disc brake. This model drive-line is run in normal condition after which single and multiple faults are introduced intentionally to the gearbox as well as to one of the bearing housings. These single and multiple faults studied on the drive-line are typical bearing and gear faults, which may develop during normal and continuous operation of this kind of rotating machinery. The real-time domain vibration signals obtained from the drive-line are preprocessed by wavelet transforms for the neural network to perform fault detection and identify the exact kinds of fault occurring in the model drive-line. It is shown that by using multilayer artificial neural networks on the sets of preprocessed data by wavelet transforms, single and multiple faults are successfully detected and classified into distinct groups.

Table 15 on p. 381 in appendix F summarizes the experiments that have been performed by Paya *et al.* as well as some other researchers.

According to McCormick and Nandi [46], vibrations in rotating machinery are produced by a combination of periodic and random processes as a result of the machine's rotation cycle and interaction with the real world. The combination of such components may give rise to signals, which have periodic, time-varying ensemble statistics and are best described as cyclostationary. This study describes second-order cyclic statistical methods

and introduces several applications of these methods to machine vibration analysis. Second-order cyclic statistical methods are used for data collected from a rotating machine subjected to bearing faults. In another paper McCormick and Nandi [42] describe the use of neural networks as a method to classify the machine condition automatically from the vibration time series. Methods for the extraction of features, which may be used as neural network inputs, are described and compared. These methods are based on measuring the zero lag higher-order statistics of the measured vibration time series. The time series for vertical and horizontal vibration signals is considered separately and combined to produce the time series based on the radius of the vibration displacement. The experimental setup used for simulating imbalance and rub faults is also described.

McCormick, Nandi and Jack [45, 48] state that model-based approaches to vibration monitoring may provide a means of detecting machine faults even if data are available from the machine in its normal condition only. The cyclostationary nature of rotating machine vibrations may be exploited by using periodic, time-varying autoregressive models to model the signal better than time-invariant models. Experimental data collected from a rotating machine set subjected to several bearing faults are used to compare time-varying and time-invariant models. A comparison is also made with a simple feature-based neural network fault diagnosis system.

Ypma and Petteer consider several methods for feature extraction [103]:

- One-dimensional projection pursuit [103]
- Linear ICA (fixed-point algorithm with deflation and cubic non-linearity) and second-order ICA (with deflation at 50 steps per component and subspace dimension $L = 25$) [29]
- The measurements were, moreover, linearly combined using principal component analysis
- Raw measurements (normalized logarithmic power spectrum and concatenating the amplitudes)

- Envelope detected data using wavelet bandpass filtering (normalized logarithmic power spectrum and concatenating the amplitudes)
- Envelope detected data using Hilbert-transform demodulation (normalized logarithmic power spectrum and concatenating the amplitudes)

In [102] these researchers conduct experiments on data from a machine running at different speeds (46, 50 and 54Hz) and loads (25, 29 and 33kW). Higher load causes the vibration to be more Gaussian distributed [48]. The authors report results on isolated imbalance and bearing fault conditions, but not on multiple fault mechanisms – the focus of this thesis.

In [41] Loskiewicz-Buczak and Uhrig use neural networks for data compression, classification and fusion. This system is used for the classification of vibration signatures collected from a steel manufacturing mill. The system performs fusion of information from nine sensors. The potential advantages of fusing information from multiple sensors are that the resultant information is more accurate and features that are impossible to detect with individual sensors become available. The probabilistic neural network is used for Bayesian decision fusion. A hybrid of a neural network and a fuzzy logic system for vibration monitoring of rolling element bearings is proposed in another paper. The system takes advantage of the learning and generalization abilities of neural networks. The ambiguity that accompanies fault diagnosis is handled by means of fuzzy membership functions. The combination of neural networks and fuzzy logic supposedly contributes to high speed and flexibility of the system. The Kohonen self-organizing map with categorization was used for the project.

The influence of shaft misalignment on the lifetime of a bearing is investigated by Hines *et al.* in [34] and some guidelines are established. In [40] Loparo *et al.* examine a rub fault in conjunction with raceway faults in rolling element bearings and a rub impact model is proposed to model the resulting non-linear dynamic behaviour. A rolling element bearing model is proposed in [40]. Unfortunately, the proposed model in [40] assumes that the contact surfaces of the healthy bearing are uniform. Furthermore, the dependency of

the excitation forces P_s in [40] on imbalance or rub impact is not given. As a result of the non-linear relation of parameters in [40], it is concluded that the model would be difficult to implement.

In [55, 84] the influence of defect location on vibration signal level is described qualitatively. Figure 5.9 of [55] qualitatively describes the decreasing vibration level as the defect moves out of the loading zone. The effect of multiple defects on the vibration signal is described in [55, 84] in table 5.1 of [55] – the frequency is linearly proportional to the number of defects.

In [78] the authors propose a normalized metric, based on the ratio of the average value of the peak (and harmonics) of the defect frequency (using HFRT) to the average value of the spectrum. Instead of a single frequency component, more than one frequency component is used for the computation of this feature. This is in contrast to the popular ‘trending’ of single frequency components (which may be influenced by environmental factors such as speed and load, thereby complicating the diagnosis procedure).

Although some databases with specific fault mechanisms are presented on the World Wide Web (WWW), [e.g. 102, 103], the databases are often closed (in the sense that the data set is limited to the specific features that the researcher has chosen for research). An open database should have at least the following characteristics for research purposes:

- A set of time series, collected at sufficiently high sampling frequency, with ‘long enough’ time duration is required; alternatively the complete magnitude and phase spectrum of the signal is needed.
- Normal and faulted measurements are required: the influence of environmental factors (such as disassembly, sensor placement) should be discussed.
- Measurements should be taken independently to be able to verify the conclusions reached on the database.

In [85] Tandon and Choudbury write a review article, listing 137 literature references, on the detection of defects in rolling element bearings. They distinguish on p. 470 between

localized defects (cracks, pits and spalls) and distributed defects (surface roughness, waviness, misaligned races, off-size rolling elements). “The dominant mode of failure of rolling element bearings is spalling of the races or the rolling elements, caused when a fatigue crack begins below the surface of the metal and propagates towards the surface until a piece of metal breaks away to leave a small pit or spall.”

In [105] a simple heuristic rule is presented for the determination of the outer race and inner race bearing defect frequencies. The outer race defect frequency is given by $0.4(F_r)N$, where N is the number of balls and F_r is the rotational shaft speed (in Hz). The inner race defect frequency is given by $0.6(F_r)N$. These equations are based on the heuristic that the cage rotates at 40% of the shaft speed F_r . If the number of balls is unknown, a rough estimate may be obtained by assuming $N = 10$ balls (as the majority of bearings has between 9 and 11 balls) [105]. When applied to the Vib demo data set the heuristic provides an estimate of $0.4(27.1)11 = 119\text{Hz}$, which is slightly larger than the actual measured value of 114Hz. Hence the heuristic is very useful to provide a rough estimate of the range of the expected bearing defect frequencies, in the absence of further detailed information (which could also be in error, as described in this thesis; hence the heuristic is also useful for validating the calculated defect frequencies).

Harris [16] mentions on p. 969 in the section ‘Variable elastic compliance’ that the inner and outer raceways may have a periodic relative motion even if the bearing is geometrically perfect. Thus it is concluded from [16] that it is in principle possible to observe a characteristic defect frequency without an induced (local) bearing defect.

2.5.1 Modelling of roller bearing faults

McFadden and Smith describe a model for single and multiple point defects in bearings [51, 52]. The complete model for a single defect may be described mathematically by

$$v(t) = [d(t)q(t)a(t)] \otimes e(t) \quad (19)$$

where $d(t)$ represents a series of impulses at the characteristic defect frequency of strength d_0 , $q(t)$ is the bearing load distribution, $a(t)$ is the effect of transmission of

vibration from bearing to transducer and $e(t)$ is the response of the machine. Hence $v(t)$ is given by the convolution of $e(t)$ with the product of $d(t)q(t)a(t)$. In the spectral domain $V(f)$ is given by the product of $M(f)$ and $E(f)$, where $M(f)$ is the convolution of $D(f)$, $Q(f)$ and $A(f)$. The load distribution $q(t)$ is given by the Stribeck equation [51]

$$q(\theta) = q_0 [1 - (1/2\varepsilon)(1 - \cos \theta)]^n \quad (20)$$

where $n = 10/9$ for roller bearings (or $3/2$ for ball bearings), ε is the load distribution factor and θ is the angular position of the roller. McFadden and Smith show [52] that the resulting spectrum for multiple defects is given by the phasor sum of (the spectrum of) the individual defects. In [53] McFadden *et al.* examine the theoretical effect of synchronous averaging. McFadden *et al.* also discuss a bearing defect model as well as experimental results with multiple spalls on the inner race in [53].

In [84] an analytical model is described for the vibration response of rolling element bearings owing to a localized defect. As a result of the very brief duration of the contact between the roller and the defect, the pulse shape is often approximated by an impulse. Yet, a better description of the severity, extent and age of damage may be obtained by using a pulse representation [84]. The relative velocity v_r between the rolling element and either of the races is given by

$$v_r = \frac{D\omega_s}{4} \left(1 - \frac{d^2}{D^2} \cos^2 \alpha \right) \quad (21)$$

Hence the time between defects is given by $\Delta T = b/v_r$, where b is the width of the defect. However, equation 21 is only dependent on the bearing geometry, not on the defect width.

In the PhD thesis of Zhang [107] an alternative expression for v_s (the dynamic impact speed), which includes the effect of defect width x , is derived

$$v_s = \frac{d_m n_i \pi}{60D} (1 - \gamma) x \sqrt{1 - \frac{x^2}{D^2}} \quad (22)$$

where n_i equals the rotation speed of the inner race (rpm), x is the defect width, d_m signifies the pitch diameter (mm) and α refers to the common contact angle. The geometrical parameter γ is given by

$$\gamma = \frac{D \cos \alpha}{d_m} \quad (23)$$

Chow *et al.* [38] consider the problem of rolling bearing fault diagnosis. Measurements are taken on a test rig (the Machinery Fault Simulator manufactured by SpectraQuest, Richmond, VA, 1997) with three different fault mechanisms: defect on bearing ball, inner raceway and outer raceway. A motor simulation package for MATLAB-SIMULINK, called MotorSim is used to model the effect of different bearing failures by simulating motor system dynamics (allowing the simulation of motor data under different operating and loading conditions)⁵. The measurements are made under five shaft rotational frequencies $F_s = 15, 25, 30, 40, 57.5\text{Hz}$ and five measurements are performed at each shaft frequency (16 384 samples at 20 000 samples per second). No mention of different (static) shaft loads could be found in [38]. The effect of speed deviation is mentioned and a graph of frequency fluctuation versus shaft speed compiled. The frequency deviation generally increases with an increase in shaft speed. These data are used to compile a modified frequency band, enabling a more robust scheme for fault detection under varying shaft speed⁶. The extracted features include the maximum and mean value of the amplitude of vibration, kurtosis, and amplitude of vibration at shaft frequency, outer race and inner race defect frequencies⁷. No mention of high frequency components, such as the HFRT, could be found in [38]. Neural networks are employed for fault detection and a high success rate is reported.

⁵ This package was unfortunately not available during the compilation of this thesis.

⁶ However, the vibration data does not seem to be normalized to the same shaft speed (orders).

⁷ A safety margin of 5Hz was used for each frequency band to cover variations in speed as well as possible differences in bearing geometry. A safety margin of a few per cent of the defect frequency has also been used in this thesis to cover uncertainties.

Randall [72] expands the model proposed by McFadden and Smith in [52] by including a small random variation (e.g. 0.75%) in the shaft speed to accommodate variations in the load angle (hence rolling radius) and tolerances of the cage (presumably assuming that the shaft speed itself is constant). Randall supplies a flow diagram to make a computer-generated bearing fault signal [21]. Unfortunately, no further details were provided on the load modulation function (in order to investigate dynamic loading).

Randall mentions on p.3 of [72] that the effect of the random speed variation may be viewed as a lowpass filter. This correlates well with [10] in which it is shown that the effect of a uniform distribution in speed is given by the convolution of the rectangular function, resulting in a triangular function (sinc^2 in the frequency domain). Figure 1b in [72] shows a line spectrum in the high frequency resonance area, with line spacing equal to the defect frequency.

Randall *et al.* also introduce a statistical model for the vibration produced by a defective rolling element bearing [74]

$$x(t) = \sum A_i s(t - iT - \tau_i) + n(t) \quad (24)$$

where $s(t)$ is the oscillating waveform generated by a single impact and $n(t)$ refers to the additive background noise. According to Randall *et al.* on p. 947 [74]: “In general, the internal damping of the structure will be high enough so that A_i , τ_i and q_i can be considered constant over the effective duration of the i^{th} waveform $s(t)$. In order to introduce some randomness, these are modelled as the random point processes $\{A_i\}_{i \in \mathbb{Z}}$ and $\{\tau_i\}_{i \in \mathbb{Z}}$. Moreover, the point process $\{A_i\}_{i \in \mathbb{Z}}$ is assumed to be periodically correlated with a period Q strictly longer than T in order to encompass amplitude modulation. For instance, this situation would model an inner race fault or a rolling element fault.”

The model of timing variation proposed in [10] implies a convolution of the function $G(\omega)$ with the high frequency excitation of the resonance(s) owing to the low frequency impacting signal. The function $R_{xx}(\omega)$ itself is periodic (refer to appendix E for details). Thus the effect of timing variation in the power spectral domain is given by a convolution of the function $G(\omega)$ (repeated at multiples of the repetition or defect frequency) with the

original signal. This seems to differ somewhat from the interpretation in figure 2 of [72], in which the discrete frequency components are attenuated by the action of the lowpass filter $G(\omega)$. Rather than simply multiplying the spectrum as in figure 2, it is believed that the effect of timing variation is rather a convolution in the frequency domain as derived in [10]⁸. If the effect of timing variation were simply to lowpass filter the periodic sampling signal as indicated in figure 2 of [72], it would be difficult to explain the frequency smearing observed in figure 1e of [72].

In [73] it is held that the cepstrum may give similar and valuable diagnostic information as the demodulation spectrum. The advantage of the cepstrum is that convolutive processes would add in the cepstral domain owing to the logarithm employed in the transform.

If a large peak is found in the cepstral domain (such as with bearing defect signals), it is possible to zero it and then transform the signal from the cepstral domain back to the power spectrum domain. This allows the effect of repetitive frequencies (e.g. gear mesh) to be removed, thus allowing other smaller signals (e.g. bearing signals) to be observed more clearly.

Braun proposes a model for localized bearing defects [8]. Each defect is assumed to give a vibratory response $g_i(t)$ with the total response given by $x_i(t)$

$$x_i(t) = \sum_j g_i(t - jT_i)U(t - jT_i) \quad (25)$$

where $U(t)$ is the unit step function. Given the additional assumption that the structural response to a defect induced impact consists of damped sinusoids

$$g_i(t) = A_i e^{-\alpha_i t} \sin(\omega_0 t) \quad (26)$$

and the assumption that the outer race may be represented by a thin ring, an expression for the multiple modes of frequency f_{0n} may be derived [8]

⁸ Unfortunately, the reference [Randall, 2001A] cited in [72] is not yet available for clarification.

$$f_{on} = \frac{1}{2\pi} \frac{n(n^2 - 1)}{(1 + n^2)^{1/2}} \left(\frac{El}{\mu r^\mu} \right)^{1/2} \quad (27)$$

For many localized defects i , each mode has the form

$$x(t) = \sum_i x_i(t) = \sum_i \sum_j A_i e^{-\alpha_n(t-jT_i)} \sin(\omega_{0n}(t-jT_i) U(t-jT_i)) \quad (28)$$

with spectrum

$$X(f) = 2\pi \sum_i \sum_j X_i(f) \delta(f - kf_i) \quad (29)$$

All integer harmonics of the defect frequency f_i are thus generated with magnitude dictated by the structural frequency response. The effect of speed variations would be to supply spectral smearing and could be incorporated into the model above by replacing $t - jT_i$ with $t - jT_i - \tau_j$, where τ_j is the instantaneous speed variation of defect pulse j . In practice, the bearing invariably forms part of a complex system, hence the observed vibration spectrum could deviate substantially from the above theoretical prediction.

An expression is also derived for the vibrations produced by an ideal bearing [8]. The total deflection at a location on the bearing ring as a function of time is given by the superposition of K equal contact forces of magnitude P_0 , at angles ϕ_k , produced by K rolling elements [8]. The effect of distributed defects may then be incorporated by assuming a contact force distribution along ϕ as the roller rotates along the race [8]. Braun [8] proposed a model that may account for many distributed geometric defects (e.g. misalignment, lack of roundness, rolling elements of unequal diameter and even structural effects). The main (theoretically predicted) effect of the geometrical defects seems to be the generation of additional sidebands [8].

2.5.2 Influence of static load

White analyses the kinematic properties of roller and ball bearings under load on p. 677 [104]:

The effect of vibration transmission through rolling element bearings was investigated in order to aid signal interpretation for use in condition-monitoring studies. The relationship

between bearing non-linear stiffness and frequency response function was derived for bearings of type commonly used in rotating machinery. By considering the contact deformation of individual elements the stiffness characteristic for a complete bearing was evaluated for the case of a radially applied load. Variation in stiffness of the complete bearing as elements moved through the load zone was calculated for bearings of a specified type and size. These values were also used to dynamically model typical rolling element bearing non-linear behaviour. Derived non-linear stiffness coefficients were substituted into the equations of motion for a model bearing. Solution was obtained by digital integration, and the bearing frequency response function was then evaluated for discrete increments of static loading. Rolling element bearing frequency response was found to be very dependent on load. For operation under light load the bearings had especially strong non-linear vibration transfer characteristics. At this operating condition a small increase in radial load produced a significant change in the bearing transfer characteristic (owing to an increase in the resonance frequency).

A two-degree-of-freedom dynamic model of a generalized bearing is proposed in [104], with non-linear bearing stiffness and damping coefficients. In order to solve the differential equations digital integration has to be performed.

Roller bearing stiffness increases with increased load. The effect is more pronounced under light load, where the load is shared between few elements and the effect of non-linear element stiffness is more significant in the bearing transfer function. If it is assumed that the vibration decreases with increasing stiffness, the article seems to support the conclusions reached in [6] and in this thesis that an increase in load (static or dynamic) may lead to reduced levels of vibration in specific circumstances.

Harris briefly describes the influence of load and speed on bearing stiffness. Figure 26.16 on p. 1036 of [16] seems to indicate that normalized bearing stiffness increases with increase in speed and decreases with increase in load. However, on p. 969 of [16] Harris states: "Bearing stiffness increases with increasing load, a characteristic referred to as a 'hardening' spring."

The PhD thesis of Zhang [107] examines the influence of a variation in operating conditions, such as speed and static load, on bearing fault diagnosis. Equation 6 in [108] relates the RMS vibration level to the defect width, the (positive) coupling of defect width and the shaft speed as well as the (negative) coupling of defect width and lateral load.

Billington [6] discusses the influence of sensor location on bearing defect detection in the presence of static loading. Depending on the sensor location, the RMS vibration level either increases or decreases with an increase in the applied (static) load. A vibration dynamic model is proposed on p. 85 of [6] to explain the somewhat counterintuitive decrease in vibration level with increase in static load:

The model consists of three mass elements: the bearing tooling, the housing above the bearing, and the housing above the end cap...The spring between the tooling and the housing above the bearing is modelled as a non-linear spring element...Simulations of impulse magnitude vs. acceleration of points 2 and 3, suggest that there are cases for increasing load in which vibration level decreases at point 2 and increases at point 3.

Padovese [67] discusses the recognition of bearing faults (normal, pit, corrosion, corrosion2) under conditions of varying shaft speed (1000rpm, 1200rpm and 1400rpm) and load (200N, 400N and 600N). Twelve signals of 512 samples (at a sample rate of 5.12kHz the duration is 0.1 seconds only) were collected for the four fault classes, hence the complete database consisted of 432 acoustic signals. The low frequency spectrum (0–600Hz) was used as features, using the Welch PSD to obtain a feature vector of 64 points eventually. The size of the test set for validation was approximately 20% of the data set size. The size of the MLP was 64 for the input layer, 20 for the hidden layer and 3 for the output layer. The authors report excellent fault classification results using MLP SEC, MLP CG and PNN algorithms.

The number of the weights of the middle layer of the MLP in [67] was $64 * 20 = 1280$ weights. If the commonly used heuristic of 10 training sample for each weight is used [18], it is clear that the MLP in [67] is over parameterized (too large). Either the number of features should be reduced, or the complexity of the classifier should be reduced (by reducing the parameters – weights – of the neural network, or stopping training early). Although some results have been validated using a small test set (and care has been taken to make the test set as representative of the training set as possible), the ideal would be to use leave one out cross validation. In [100] it is demonstrated that the selection of the samples for use in the test set may make a large difference in the classification results of the Iris data set. In [83] an accuracy of 100% is claimed on the Iris

data set. This is possible if the problematic data samples (with class overlap) were chosen in the training set and the non-overlapping samples in the test set respectively. Hence the importance of doing proper validation in the building of a model becomes apparent. The novel techniques proposed in this thesis (the CVCP strategy discussed in section 5.4.2 on p. 91) could be applied to the results of the (neural network) model on the training set. The resulting complexity map would have prevented the conclusion of 100% accuracy from being drawn on the training set itself (even if all the samples were used in the training set).

The good results in [67] are attributed to acoustic measurements, but unfortunately no vibration measurements have been performed to enable a fair comparison to be made. The paper also does not discuss why acoustic measurements should perform better in theory (in a laboratory setup the influence of external noise sources should be easier to control). The features extracted do not seem to include any high frequency components (for example, by using HFRT), nor could any theoretical discussion on the merits of the low frequency components be found. No specific bearing fault frequencies have been used as features. This is good in the sense that the results are not dependent on a specific test rig. However, the validity of the results (the generalization performance of the neural networks) is queried owing to the large number of features used during training and the complexity of the neural network. Furthermore, the results would be valid for the specific test rig only and under the specified operating conditions. The small buffer size (512 samples correspond to 0.1 seconds, or only 1.666 shaft revolutions at 1000rpm) makes generalization unlikely, especially when the change in environmental conditions, such as speed and load, is taken into consideration.

The shaft speed variations have been accommodated in [67] using a separate neural network structure for each shaft speed. This is problematic for generalization purposes. An alternative approach could have been to normalize the shaft speed, using the normalized features as input to a single neural network. Finally, the influence of shaft load is unfortunately only mentioned briefly in the paper. The author points out that the shaft load would influence the amplitude of vibration. As the fault mechanisms introduced in [67] were very large, a slight change in the amplitude of vibration would not have a significant

effect on the classification results. However, in an industrial environment one would be interested in detecting very small defects, in which case even a slight change in the amplitude of vibration might have a significant effect on bearing fault detection. Furthermore, one would also be interested in the evolution of the bearing defect, which is not discussed in [67]. In summary, the methodology employed in [67] does not seem to validate the results and does not provide a good framework for this thesis. Low frequency components only, which may be easily corrupted by noise, have been used. Nevertheless, HFRT has been shown to provide the best diagnostic indicator of bearing failure and defect size as discussed by McFadden, Toozhy, Shiroshi, Li, Liang, Kurfess and Danyluk [50, 53, 78].

Tandon and Choudbury [84] derive an expression for the total radial deflection of outer race, inner race and rolling element defects. From this expression the magnitude of the frequency components may be derived at multiples of the characteristic defect frequency. It is shown that the magnitude of the components increases with an increase in radial load. The expressions are derived for the low frequency band owing to a localized defect in a bearing element [84]. Unfortunately, the expressions cannot be used in this study, in which HFRT has been used to examine the high frequency components. A further limitation seems to be that the effect of instantaneous speed variations is not included in the model.

No experimental results on the effect of static load on the frequency components could be found in [84]. A computer program was used to calculate the magnitude of the frequency components numerically. Unfortunately, absolute amplitudes of vibration could not be predicted – only the approximate amplitude ratios among various spectral lines [84].

Braun discusses the influence of loading effects in [8]. The load distribution $q(t)$ is given by the Stribeck equation [8, 51] (refer to p. 32). For a rotating inner ring it is shown that the modulation effect by loading causes sidebands to appear in the frequency spectrum.

2.5.3 *Influence of dynamic load*

In [82] the theoretical effect of imbalance on a defect pulse is described quantitatively on p. 77: “For an outer race defect the eccentric loading due to unbalance rotates at shaft frequency so that $w(t)$ is embedded and has a signature with period $(f_o)^{-1}$ where f_o is the

shaft frequency in Hz and $w(t)$ is a weighting function representing the contact energy.” No reference could be found in [82] to experimental results on the influence of imbalance (dynamic load) on the magnitude of the defect pulses.

On p. 969 of [16] the effects of static and dynamic load on bearing stiffness are described (please note that the concept of bearing stiffness seems to imply the low frequency response at the characteristic defect frequency, hence it cannot be directly compared with the high frequency resonance technique (HFRT)):

Bearing stiffness increases with increasing load, a characteristic referred to as a “hardening” spring. Larger nominal operating loads or built-in preload would result in smaller variations in dynamic bearing deflection when subjected to a particular dynamic load variation. Similarly, increased bearing stiffness raises the value of a resonant frequency associated with this spring, since a resonant frequency is inversely proportional to the square root of stiffness...Since the bearing “spring” is non-linear, it is evident that sinusoidal deviations from the nominal load will not cause sinusoidal bearing deflection. When the load⁹ is greatest, the increase to nominal bearing deflection will be less than the decrease from nominal bearing deflection when the load is at its lowest value. If large dynamic fluctuations in load are experienced, say in a radially loaded bearing, then it is possible for the load zone to alternate from the bottom to the top of the outer raceway. If the bearing has radial internal clearance, there is the possibility of essentially no loading at all on the outer raceway for brief instances. Such conditions would arise because of external loading or conditions within the bearing.

No specific experimental results of a bearing defect in conjunction with an imbalance condition (dynamic load) could therefore be found in the literature surveyed. The focus is accordingly to demonstrate a strategy for the recognition of multiple fault conditions, applied specifically to the abovementioned fault condition.

⁹ Load is assumed to mean the instantaneous dynamic load in this context.

Chapter 3

3 Problem statement and contributions of thesis

3.1 Introduction

This chapter provides a detailed substantiation and description of the primary focus of this thesis, which involves the automated detection of multiple fault conditions in rotating machinery.

Ideally researchers would have access to a single database containing the results of tests performed on a large number of machines experiencing a representative set of fault conditions using a wide variety of transducers. In such circumstances a technique with a high modelling capacity, e.g. neural networks, may be trained to detect and classify any fault condition or combination of fault conditions automatically that could be reasonably foreseen.

Unfortunately, this ideal cannot be achieved in practice. The challenge is therefore to come as close as possible to the ideal, taking into account practical limitations regarding the introductions of faults into machines and economic costs associated with the compilation of a set of measured data. This may be achieved partly by artificially extending the available data set through simulated models, and partly by using techniques that may generalize well on the basis of a small number of training set examples.

The introduction of fault mechanisms for fault detection is frequently problematic in an industrial setting, where every effort is made to keep the machinery functioning normally. In order to obtain useful results, laboratory experiments are frequently substituted for real-world situations. As a bearing failure may take years to be induced, either a localized fault may be induced for the purpose of laboratory experiments by etching a fault on the race or roller of the bearing or an accelerated lifetime test may be performed under conditions of high stress. It should be borne in mind that a real defect will be much more complex (in principle it may be regarded as the superposition of an infinite number of localized defects

as described by McFadden and McFadden [50]). In particular, the evolution of a defect in a real-world situation – once the defect has been detected – is difficult to interpret [6, 50]. However, this information is critical for accurate bearing prognostics [6]. Although it is often not too difficult to detect bearing failures when well developed (provided that the correct set of techniques has been applied), it is desirable to locate the fault at a very early stage. Even when the fault has been located, it is very important to be able to predict the severity of the defect as it progresses from an early to an advanced stage [6]. Furthermore, the influence of secondary fault conditions, such as static and dynamic load (induced with an unbalance mass), may change the interpretation of the frequency component at the characteristic defect frequency [6, 78, 98].

As a result of the extreme difficulty of obtaining access to a suitable database of sufficient size for the application of pattern recognition and signal processing techniques, a bearing test rig has been used to perform certain experiments for the purpose of this thesis. The set of practical measurements was extended by building simplified models to represent the relationships between observed fault conditions (e.g. a vibration peak at a particular frequency) and underlying causes (e.g. a bearing that has been damaged to a certain level of severity). The aim of this thesis is neither to present the best model of vibration nor to derive a theoretical model explaining the dynamics of rolling bearing defects, but rather to illustrate how signal processing and pattern recognition techniques should be applied for successful detection of multiple fault conditions. As each measurement also has an implicit cost element associated with it, it is of interest to examine techniques that minimize the number of measurements to be taken for fault detection. The end user of such techniques would ideally like to use the smallest number of sensors (as well as the most robust, cheapest, easiest to install, etc.), and the least expensive instrumentation with the lowest possible sampling frequency and the smallest possible buffer length. In order to determine the minimum set of information that still supports reliable diagnostics and prognostics, it is first necessary to collect more extensive data in a few representative locations in the plant. From the observations made on this smaller data set, it is hoped that inferences may be made on the rest of the plant on a statistically sound basis.

Techniques that can generalize well on the basis of a small number of training set examples may be constructed by combining conventional modelling and classification techniques. The nearest neighbour rule may be used for statistical cross validation of results, but at a high computational expense [97]. However, the proposed new cross validation strategy to determine classifier performance (CVCP) technique may be used for efficient cross validation [97]. Furthermore, it may also be applied to modelling problems as shown by the novel nearest neighbour rule with class membership (NNRC) technique in [94].

3.2 Data set size limitations

Owing to practical limitations in obtaining a data set of sufficient size for the training of a purely data based technique such as neural networks in condition monitoring, a number of alternatives have to be considered for the successful application of pattern recognition techniques:

- The size of the data set may be increased. As a result of the cost associated with data collection, it is desirable to minimize the amount of data collected. Owing to cost constraints, the number of measuring channels as well as the cost of transducers should be minimized. The design of experiments on the basis of statistical theory may help to minimize the size of the data set. In this way, more data may be collected in the areas of large uncertainty, hence covering the feature space more effectively with fewer measurements.
- Existing knowledge may be used. If *a priori* knowledge is available regarding the condition of the machine or type of fault (for example by means of an expert system or mathematical relationships) then such knowledge may be incorporated into the training process as a set of rules. A fuzzy neural network may combine an existing rule base with the ability to learn from new data as they become available [100].

- The complexity of the relationship that must be modelled may be reduced by selecting more appropriate features. While the relationship between a specific fault condition and the raw data measurements may be complex, a much simpler relationship may exist with optimally selected features that are closely related to the fault condition.
- The number of features may be reduced without discarding important information. A smaller feature set reduces the size of the database required and enables a classifier or model of lower complexity to be used. The ideal would be to obtain a single feature that would be able to predict the severity of the fault mechanism.
- The complexity of the classifier or model may be reduced. Regularization of a neural network may improve its capability to generalize. However, oversimplification should be avoided if the non-linear modelling capabilities of the neural network are to be realized. In the worst case of a single neuron in the hidden layer no advantage is obtained over a linear regressor. Reducing the complexity of the relationships that may be learnt, for example by stopping training early, is a possibility for avoiding overfitting. If the full non-linear capabilities of the neural network for problem solving were to be realized, this would be undesirable. The existing data set may be augmented on the basis of expert knowledge regarding the influence of certain parameters. Owing to the complexity of real-world mechanical systems and the non-linear relationships inherent in such systems, this approach is often not feasible in practice.
- The mechanical system may be modelled, either through mathematical models or a finite element model (FEM). Developing a detailed FEM is very costly and time consuming and would only be done for very expensive machinery or perhaps in mass production. A simplified mathematical model for bearing faults will be used, building on previous literature in the field.

3.3 Model of bearing defects

The model proposed for bearing defects by McFadden and Smith in [51, 52] has been extended by other researchers to incorporate additional environmental conditions. For example, Brie [10] discusses the effect of period variations caused by changing contact angle. It is shown in [78] that the size of the demodulation peak in the vibration spectrum is linearly related to the defect size. The normalized ratio of demodulation peak to the ‘carpet level’ (noise in the demodulation spectrum) hence provides a quantitative measure of the bearing defect condition. Bearing replacement is suggested when this ratio is above 15dB. It should be noted that determining the carpet level is no longer trivial when no trending measurements are available and when speed variation is present.

McFadden and Smith propose a model based on an exponentially damped sinusoid as the response of the structure to an impulse $d(t)$ caused by a ball contacting a single point (inner race) defect [51, 52]. The impulses $d(t)$ are modulated by the load function $q(t)$ and the transfer function to the sensor $a(t)$. In [10] alternative models are proposed, firstly using time-frequency analysis, allowing instantaneous changes in frequency (chirp signal). The second model, using time-varying parameters (LTV system), is based on a first degree of freedom system only. It is also shown in [10] that variations in instantaneous speed may be modelled using a quasi-periodic impulse train, and analytical expressions are derived for the autocorrelation and power spectral density (PSD).

It is clear that instantaneous variation in the contact angle causes timing jitter in the response of the system to a bearing defect, even if no speed variation in the rotating speed takes place. The timing jitter causes spectral smearing [10], complicating spectral and cepstral analysis.

As stated before, the goal of this research is not to develop the best vibration model for defect detection, but rather to evaluate the effectiveness of various signal processing operations, usually implemented in a preprocessing stage, to support the extraction of characteristic features. A simple vibration model for defect detection, with reasonable correspondence to the measured data set, is therefore required.

A system response model (with defect pulse excitation) is proposed in this thesis as a Gaussian modulated sinusoidal signal with fractional bandwidth BW and resonance frequency f_c . Observation of the time domain waveforms suggests that the above choice could be suitable (as indicated by McFadden *et al.* in [50, 53], the amplitude modulation function is of primary interest, hence the exact choice of modulating function is not that critical as far as the interpretation of the demodulation spectrum is concerned). The bandwidth of the pulse is $100 * BW\%$ measured at a level of BWR dB relative to the normalized peak of the signal. The signal may be expressed analytically as

$$Y_b = A \exp \left(\left(\frac{-(t - \tau)}{\sqrt{2} t_v} \right)^2 \right) \exp(-j2\pi f_c (t - \tau)) \quad (30)$$

with amplitude A , Gaussian mean f_c , Gaussian variances $t_{v2} = 1/(4\pi^2 f_v^2) = t_v^2$, $t_v = 1/(2\pi f_v)$, $f_v = \sqrt{f_{v2}}$, $f_{v2} = -BW^2 f_c^2 / (8 \log(r))$ with reference level $r = 10^{BWR/20}$ and time delay of τ . The effect of multiple defects may be accommodated by the model proposed in [52] (refer to section 2.5.1 on p. 31 for details).

A general model for bearing defect signals may also be proposed, based on the concepts of communication theory. The existence of high frequency bearing defect pulses in the high frequency resonance area may be modelled as a sinusoidal carrier wave (frequency ω_c) with amplitude ($A(t)$), phase ($p(t)$) and possibly frequency modulation¹⁰ ($\omega_{mod}(t)$)

$$y(t) = A(t) e^{j(\omega_c + \omega_{mod}(t))t + p(t)} \quad (31)$$

The amplitude modulation $A(t)$ is the product of an exponential (or RF modulated) function $g(t)$ (which models the shape of the defect pulses) and a modulation function $m(t)$, or $A(t) = g(t)m(t)$. The modulation function $m(t)$ is usually a periodic function that repeats at shaft speed. The term ω_c , which is the carrier wave frequency, is caused by a mechanical impact (such as a bearing defect) exciting mechanical resonances in the system. Although more than one vibration mode (or resonance) would be excited, it

¹⁰ An example would be sinusoidal modulation: $\omega_{mod}(t) = \beta \sin(\omega_m t)$.

is often sufficient to consider one specific vibration mode only using a suitable bandpass filter. The objective is to choose the most responsive frequency range to bearing defects (optimizing the SNR) for demodulation purposes [10]¹¹. In the case of sinusoidal modulation, sidebands (spaced at the modulating frequency) may be observed next to the carrier frequency.

Phase modulation $p(t)$ of the carrier is caused by periodic variation in speed. The magnitude of the impact may cause frequency modulation of the carrier (larger impact causes lower frequency). The effect of phase modulation on the frequency domain is a smearing of frequency components [95], which may complicate recognition of bearing defects (relying on periodicity in the frequency domain)¹². From communication theory it is also well known that phase modulation and frequency modulation are closely linked [81]. Hence phase modulation of the carrier causes the frequency bandwidth to increase. This may be understood at an intuitive level by considering a simple carrier wave of frequency f_c . If the phase is slowly increased over time, the effective time between, for example, the crests of the carrier also increases over time. Thus the effective frequency content of the carrier would also increase over time, even though the frequency of the carrier remains the same. So the bandwidth would increase in proportion to the amount of phase modulation. (For a more rigorous derivation refer to standard textbooks on communication theory, for example [81].)

Hence the bearing defect may be regarded as a communication system, in which the signal from a small bearing defect is transmitted over a communication channel (the mechanical transfer function of the bearing defect to the sensor, for example an accelerometer) in the presence of interfering signals (other vibration sources as well as noise). The aim is to recognize the bearing defect reliably at the receiver, using the smallest number of samples (buffer length), in the presence of noise components. From a signal processing perspective

¹¹ Mechanical impact tests with an instrumented hammer (as described by, for example, McFadden *et al.*) could be used to detect system resonances and may provide a suitable starting point for choosing the best demodulation frequency range(s).

¹² An improved series of cepstrum techniques has been proposed in [93] for bearing fault detection.

the objective is to optimize the signal-to-noise ratio. Hence signal processing techniques, which are commonly used for communication systems, may be used with success in the recognition of bearing defects. If a trigger signal on the shaft (or rollers) is available, it may be used in the computation of synchronous averages [50]. This situation is analogous to the presence of an exact replica of the sender carrier wave at the receiver. If the trigger signal is not available, an approximate trigger signal may be derived on the basis of the vibration signals [95]. However, deriving an appropriate trigger signal becomes problematic under conditions of low SNR. Thus a trigger signal may enhance the diagnostic process¹³, but is unfortunately not that easy to obtain in a plant, as described by McFadden and McFadden [50].

3.4 Obtaining a normal bearing signal

In the initial set of Vib demo 2 measurements the vibration signals were measured on both a normal and a defective bearing. Unfortunately, a number of complications developed:

- The normal bearing had different geometrical properties compared with the defective bearing, making direct comparison difficult.
- The vibration from the defective bearing seemed to be transmitted to the normal bearing that was part of the same measurements setup (although much reduced in amplitude). For the purpose of fault detection this would make the pattern recognition technique more robust and is therefore beneficial. Yet, this is problematic from the point of view of obtaining a vibration signal of a 'normal' bearing.

To eliminate these problems an additional set of Vib demo 3 measurements of vibration signals, corresponding to a bearing before and after damage, was generated. A general

¹³ It is possible that the trigger signal, even on its own (without vibration signals), could also be used as a diagnostic indicator. Studies of gearbox monitoring have shown that torsional vibration signals may be used for diagnosing specific fault conditions in gears. As demonstrated in this thesis, a bearing defect should also cause an instantaneous change in shaft angle; hence monitoring the instantaneous shaft angle should provide valuable diagnostic information on the condition of bearings.

problem with this method is the disassembly of the test rig, introducing environmental factors that could be unrelated to the defect. To avoid such problems the defect was moved around the load zone to simulate the effect of a normal (and defective) bearing. It is proposed in this thesis that this method is inherently superior for fault detection of bearing defects (such as outer race defects¹⁴).

The question arises as to whether it is possible to derive a good or normal vibration signal from a measured vibration defect signal. In order to do this it would be necessary to filter out the defect components in the vibration signal as well as possible. It also seems to require an invertible transform (hence the HFRT, using the square of the signal, does not seem suitable¹⁵).

3.5 Problematic nature of the recognition of multiple bearing defects

The interpretation of the envelope spectrum, which is obtained by applying the HFRT, may be problematic for fault detection in a real bearing defect (which may be regarded as the superposition of a number of single point defects) as discussed by McFadden and McFadden [50]. It will be shown in this thesis that the envelope spectrum may be interpreted as a special kind of synchronous average. Frequency components at $kBPFO$ (where $BPFO$ is the outer race bearing defect frequency and $k = 1, 2, 3...$) correspond to the Fourier transform of the synchronous average (SA) of the defect signal. The SA is simply the time domain average resulting from stacking N consecutive slices of the time domain signal (each slice of period T , related to the trigger frequency of the SA by $T_r = 1/T$) on top of each other and subsequently computing the average of the slices. The SA is discussed in more detail later on in the thesis (particularly in chapter 6 ('Results

¹⁴ To apply the same technique for inner race defects it would be necessary for the shaft (corresponding to the inner race) to be stationary (or rotating very slowly) relative to the casing (outer race), which would probably require substantial mechanical modifications. In a new design this would be easier to do.

¹⁵ If the interest is primarily in the low frequency amplitude modulating function (and not the detailed time domain behaviour), the asynchronous AM demodulated signal could be modulated to high frequencies by carrier wave(s) to the vibration modes of the system. Although the time domain behaviour would differ substantially, the demodulated spectrum should be nearly identical. If it is assumed that the demodulation spectrum is sensitive towards mechanical faults such as bearing failures, the above model could be used to model mechanical failures.

of signal processing techniques' on p. 115)). Define the equi-sampled Fourier series ESFFT as¹⁶

$$\text{ESFFT}(k, m) = \text{FFT}(k * m), \text{ where } k \text{ is an integer} \quad (32)$$

and m is an integer corresponding to the samples of the FFT (m is the width of the comb filter in the frequency domain). The integer k ranges from 0 to $N/m - 1$, where N is the number of samples in the FFT. It is inversely related to the period (in samples) over which the synchronous average will be taken (in the time domain). Hence long periods in the time domain give a small value of m in the frequency domain (the combs of the comb filter spaced close together) and vice versa.

If the defects are widely distributed (and large), the fundamental component of the envelope spectrum at *BPFO* will be large (with the harmonics of *BPFO* becoming smaller), and vice versa. It is very important to consider all the demodulation peaks (harmonics) of *BPFO* together, rather than considering a single frequency peak (such as *BPFO*) in isolation. This is especially true when many point defects contribute to the envelope spectrum, as in a real bearing defect (which may be widely distributed, especially in the final stages of failure, as opposed to very localized failures that are frequently employed in the laboratory). As a simple example, if two identical bearing point defects are shifted in phase by 180° (i.e. exactly opposite each other on the outer race) the demodulation peak at frequency *BPFO* in the envelope spectrum will be zero (which obviously complicates bearing defect diagnosis), but the component at frequency $2 * \text{BPFO}$ in the envelope spectrum will have a maximum¹⁷ value. Thus it is clear that an incorrect diagnostic decision will not be made if all the frequency components of the envelope spectrum are considered together¹⁸. It is well known that the

¹⁶ $\text{FFT}(k * m)$ refers to samples $k * m$ of the FFT of a signal. If $Y(f) = \text{FFT}(y(n))$ then $\text{FFT}(k * m) = Y(k * m)$. So $\text{ESFFT}(k, m) = Y(k * m)$. In a similar way $\text{ESFFT}(F_d, m)$ would refer to the (discrete) frequency components $m * F_d$.

¹⁷ The possible effects of the load distribution have been ignored.

¹⁸ Hence the total energy in the *BPFO* (and the higher harmonics of *BPFO*) should provide a better feature for accurate bearing diagnosis (for example as described in [78]), especially for real distributed bearing defects (as opposed to well localized bearing defects introduced in a laboratory setting).

synchronous average may reduce noise by a factor of $1/\sqrt{N}$ [86] (where N is the number of samples in the buffer). In particular, any vibration not synchronous with the trigger period T will be filtered out (if a sufficient number of averages are taken). As the ESFFT, defined in equation 32, actually computes a synchronous average it should also reduce the noise by a factor of $1/\sqrt{N}$ [86]. Hence the good performance of the synchronous average in reducing non-synchronous noise is also apparent in the ESFFT (equi-sampled demodulation spectrum).

3.6 Problem statement

An attempt is made in this thesis to establish a sound methodology for the application of pattern recognition and signal processing techniques in the detection of multiple fault conditions in condition monitoring. More specifically, the influence of dynamic load (small, medium and large) on the detection of bearing faults (small, medium and large) is illustrated, which is believed to be a novel contribution.

The ability of the following classification techniques to discriminate accurately between different types of fault conditions is evaluated, using features extracted from vibration measurements:

- The self-organizing map (SOM) technique for data visualization
- Suitable novel pattern recognition techniques proposed in this thesis include nearest neighbour rule with class membership (NNRC) and cross validation strategy to determine classifier performance (CVCP)

The effectiveness of the following signal processing techniques to extract fault indicators from vibration measurements will be evaluated:

- The standard high frequency resonance technique (HFRT) using bandpass filtering and non-linear square law amplitude demodulation
- The standard low frequency characteristic defect frequency (e.g. the outer race defect frequency)

- The novel thresholded cepstrum (TCEPS), linear cepstrum (LCEPS) and linear thresholded cepstrum (LTCEPS), which are modified cepstral analysis techniques
- The equi-sampled FFT (Fourier series) technique (ESFFT), which is compared with the standard high frequency resonance technique (HFRT). The HFRT is regarded as a benchmark

The influence of buffer length on fault detection is described briefly. The technique of synchronous averaging is also applied to the data sets and conceptually related to the high frequency resonance technique (HFRT), which is also believed to be a novel contribution. Speed variations may also be viewed as a noise source, resulting in a lowering of the signal-to-noise ratio (SNR), and an algorithm to reduce the impact of speed variations is presented.

In order to establish the influence of multiple fault conditions a test rig was used for experimental investigation, as a literature survey revealed no suitable mathematical or computer model. The test rig allows for the influence of a secondary fault mechanism on a primary fault condition to be determined by applying a load, a well localized outer race bearing fault of small, medium and large severity. A novel aspect is the variation of the defect angle (DA) in the test rig, allowing the effective static loading on the defect to be changed in increments of 45° ¹⁹. A simple model of bearing faults is also proposed and used to evaluate various techniques²⁰.

Various pattern recognition and signal processing techniques are applied to this data set and evaluated. The influence of speed variations on simulated and experimental data sets is demonstrated.

¹⁹ This in effect allows a 'normal' bearing to be simulated without using a 'normal' bearing, which would require disassembly of the test rig.

²⁰ A model helps to extend the size of the database effectively. However, developing a suitable finite element model for very accurate modelling of bearing faults is very time consuming and expensive and falls outside the scope of this thesis. Instead a simple model, based on established literature, is used in this thesis.

3.7 Contributions

This thesis aims to build upon established literature to make a number of novel contributions. Unique contributions made by this thesis are as follows:

1. Experimental setup. The detection of multiple fault conditions was investigated in [31, 32, 33, 42, 43, 44, 47, 49, 60, 76]. The series of measurements performed on this test rig, incorporating multiple fault conditions – especially the influence of dynamic load (imbalance) on bearing faults – is believed to be unique [98]²¹. Further innovations concern the measurement of defect versus angle position (which is mentioned in [55, 84], but no experimental results are presented) and the generation of a trigger signal on the roller via an eddy current sensor (in [50] a similar result is achieved by triggering separately on the shaft and cage, but it is mentioned that this triggering technique would be difficult to apply in an industrial setting). The measurement of defect versus angle position allows the measurement of a ‘normal’ bearing without introducing external environmental conditions owing to physical disassembly. The database is open in the sense that time buffers of sufficient duration are available. Hence future research is not limited by a specific set of extracted features. Measurements on the test rig include a wide variety of sensor technologies including vibration measurements (mounted permanently as well as handheld measurements at various positions on the test rig), acoustic measurements and current measurements. In an industrial setting certain measurements might be difficult (or even impossible) to obtain, so it is of interest to establish whether similar conclusions regarding fault condition may be inferred on the basis of measurements that are easy to make.
2. Although some experiments have been performed on the effect of static load on bearing diagnostics (see the section on the literature survey on p. 36), no similar

²¹ In the literature survey no mention could be found of experimental data sets relating the influence of dynamic load on bearing defects of varying severity. In [47] multiple simultaneous fault conditions, including Rub and Imbalance, are investigated (but not with bearing defects).

experimental results for dynamic loads are presented in the literature. Thus it is believed that the results presented in this thesis on the relation between dynamic load and bearing fault detection make a unique contribution to literature.

3. The relationship between the techniques of signal averaging and HFRT (envelope spectrum) is demonstrated. More particularly, it is shown that trending a single frequency component (using HFRT) is best when a single, well localized defect is present (as, for example, when an artificial defect is introduced in a laboratory setting). For widely distributed defects, in contrast, it is better to consider all the frequency components simultaneously (as a result of phase shifts between the defects, some frequency components will add out of phase, hence potentially cancelling each other and thereby complicating bearing diagnosis – refer to McFadden and McFadden [50] as well as appendix A). More specifically, it is shown that the frequency components at the characteristic defect frequency (and harmonics) using HFRT (more precisely ESFFT²² as defined in equation 32 on p. 52) actually correspond to the spectral information of a signal average. Thus the good performance of the demodulation spectrum (more specifically the ESFFT) is partly explained by the implicit synchronous averaging that is applied²³. The influence of the synchronous average (SA) in the time and frequency domains is shown in section 6.4 on p. 123. Although not specifically covered by the literature survey, the influence of the time domain window on the computation of the SA in section 6.5 on p. 128 is also believed to be a novel contribution. The influence of speed variations on the spectrum of a signal is derived in section 6.6 on p. 132 from first principles. In section 6.8 on p. 139 it is shown that any signal may be decomposed as the summation of a

²² To date no reference to ESFFT could be found in literature. However, the interpretation of synchronous averaging as a comb filter, with the teeth of the comb becoming narrower as the number of averages increases, is well known (refer to McFadden *et al.* [53]). The relationship between the demodulation spectrum (ESFFT or comb filter) and the synchronous average is explored in this thesis. For a large number of averages in the time domain, only frequency components are kept at multiples of the trigger frequency.

²³ Bandpass filtering and non-linear operation (for example squaring the input signal) also contribute to the success of HFRT.

number of suitable synchronous averages, which is subsequently used to illustrate the operation of the demodulation spectrum. In section 6.9 on p. 145 various aspects, believed to be novel, are elucidated related to the kurtosis statistic.

4. Novel signal processing and pattern recognition techniques presented in this thesis:

- The cross validation strategy to determine classifier performance (CVCP) [97] and nearest neighbour rule with class membership (NNRC) [94] pattern recognition techniques
- Thresholded cepstrum (TCEPS), linear cepstrum (LCEPS), linear thresholded cepstrum (LTCEPS) [93] and equi-sampled FFT (Fourier series) (ESFFT) signal processing techniques

The cross validation strategy to determine classifier performance (CCVP) technique applies the well-known nearest neighbour rule (NNR) technique to classifiers such as neural networks and is believed to be novel. The main advantages over, for example, neural networks are that no lengthy and complex mathematical optimization is required for validation of results (assuming the model has already been trained) and better use of the data set. The nearest neighbour rule with class membership (NNRC) technique extends the cross validation strategy to determine classifier performance (CCVP) to modelling problems (where the class membership is a continuous variable, instead of a discrete variable as in the nearest neighbour rule – NNR – technique). The linear cepstrum (LCEPS), thresholded cepstrum (TCEPS) and linear thresholded cepstrum (LTCEPS) techniques are specifically designed for detecting periodicities in a time domain signal. The absence of the logarithmic term, as defined in the normal cepstrum, seems to give better visual results on the vibration data sets. The equi-sampled FFT technique (ESFFT) is simply a synchronous average (SA) in the time domain, but computed from the FFT itself (instead of a comb filter, for example).

The effect of speed variation, while acknowledged in literature as a problem [3, 10, 21, 43, 44, 47, 49, 51, 60, 72, 73, 76], reduces the signal-to-noise ratio (SNR). Techniques have been proposed in this thesis aimed at reducing the effect of speed variations [95]. The unexpected diagnostic capabilities of the *BPF12_X* feature (value of demodulation spectrum at inner race defect frequency) in the outer race bearing defect will be analysed and explored in future work (refer to section 8.6 'Future work' on p. 263 for further details). Furthermore, a relation is found between the absolute value of the analytical signal derived from the Hilbert transform (the amplitude modulation signal) and the square law detector (refer to section 6.11 on p. 186).

5. Modelling: proposal of a simple model to explain the change in vibration of a primary fault condition in the presence of environmental conditions such as speed variations.
6. Standard techniques, as described in previous literature references, are used in order to validate new techniques against standard benchmark results. The combination of existing techniques, applied in a new context, may also be viewed as a unique contribution.

4 Experimental setup

4.1 Introduction

The problem statement of this thesis is the analysis of multiple fault conditions using artificial intelligence and signal processing techniques. In this chapter the experimental setup, which is believed to illustrate several innovations, is described. It is shown that the test rig allows for easy introduction of multiple fault conditions (a bearing defect in the presence of a dynamic load). It is not feasible (or even desirable) to build an exhaustive database with all possible combinations of environmental and fault conditions, hence a simple model of bearing defects will also be used in subsequent chapters. The Vib demo bearing test rig has been designed especially to facilitate the introduction of dynamic load in the form of unbalance mass(es). This enables easy experimentation with various levels of dynamic load, which is believed to be a novel contribution. It is also straightforward to introduce additional fault mechanisms on the bearing, as a result of the special, separable bearing used in the test rig.

- The first series of measurements, which is called the Vib demo data sets 1 and 2 and discussed in section 4.3 on p. 67, illustrates the influence of dynamic loading on a large bearing defect. Measurements are also made on a normal bearing. The measurements are performed with the DA (defect angle) not in the centre of the loading zone. As will be shown later, recognition of the defect signal in the time domain signal itself is difficult, but is easily performed using the demodulation spectrum.
- The second series of measurements, which is called the Vib demo data set 3 and discussed in section 4.4 on p. 69, illustrates the influence of dynamic loading on

large, medium and small bearing defects. Only results obtained on the outer race with permanently mounted accelerometers are discussed in this thesis.²⁴

The impact of the defect angle (DA) relative to the loading zone for a (stationary) outer race defect is analysed using the above test setup. This enables the static loading on the defect to be varied. Proving this hypothesis (a change of the DA is similar to a change in static loading) is left for future work. The maximum static loading for the specific test rig, using vibration measurements described in chapter 7 from p. 213, is obtained when $DA = 225^\circ$ (corresponding to the largest defect signal) and the minimum static loading when $DA = 360^\circ$ (corresponding to the smallest defect signal, hence postulating that it is similar to a 'normal' bearing signal). For this reason, it is possible to vary the amplitude of the defect signal by varying the DA in increments²⁵.

4.2 Influence of dynamic load on bearing fault detection

As mentioned in the literature survey on p. 40 the influence of a dynamic load on bearing fault detection is described quantitatively only in [82]. Consider a large outer race defect placed in the centre of the loading zone. Each time a roller (or ball) contacts the defect, the defect will excite one or more system resonances. If the shaft rotates at a constant speed F_s , the roller would also hit the defect at a constant rate of *BPFO* times per second²⁶. If the roller makes firmer contact with the defect, the intuitive expectation would be that the magnitude of the defect pulse would also increase. Thus an increase in radial load should theoretically also increase the vibration levels.

As mentioned in section 2.5.2 on p. 36, Harris briefly describes the influence of load and speed on bearing stiffness. Figure 26.16 (the text in [16] actually refers to figure 26.17 instead of figure 26.16) on p. 1036 of [16] seems to indicate that normalized bearing

²⁴ Defects on the inner race and rollers have also been introduced. Measurements using hand-mounted accelerometers in various positions, current measurements and acoustic measurements have also been made, but will not be discussed in this thesis (this is left for future work).

²⁵ Increments of about 45° have been used in this thesis, but even better results could be obtained using a smaller spacing. This is left for future work.

²⁶ This is done in a somewhat analogous way to a hammer striking a (high frequency) bell at regular intervals.

stiffness increases with increase in speed and decreases with increase in load. However, the text in [16] on p. 1036 mentions: “As a result of kinematic changes, the increase in α_i [contact angle] causes a decrease in stiffness with increasing speed.” The text on p. 1035 also supports the conclusion that increasing speed leads to reduced stiffness: “As speed increases, there is an additional effect that occurs simultaneously with the contact load effects already described; the outer raceway contact angle decreases and outer raceway contact stiffness increases while the inner sees an opposite effect”. An expression for K_r , the combined radial stiffness (springs in series) is derived in [16] by evaluating $k_r = 1/(1/k_{or} + 1/k_{ir})$ (where $k_{or} = k_o \cos \alpha_o$ and $k_{ir} = k_i \cos \alpha_i$)

$$k_r = \frac{k_i \cos \alpha_i}{1 + \frac{k_i}{k_o} \cos \alpha_i / \cos \alpha_o} \quad (33)$$

As the numerator in equation 33 makes a larger contribution relative to the denominator, an increase in α_i causes a decrease in the numerator $k_i \cos(\alpha_i)$, hence k_r decreases with increasing speed. However, figure 26.16 of [16] shows an increase in normalized bearing stiffness with an increase in speed. It is assumed that the stiffness K_r at static conditions would be the same for the speeds of 12 000, 24 000 and 36 000rpm shown in figure 26.16. Using this assumption the normalized stiffness increases with speed for a specific load value in figure 26.16. It is assumed in this thesis that figure 26.16 of [16] actually intended to show $1/k_r$ and thus that the bearing stiffness decreases with an increase in bearing speed (which is supported by the text of [16] and equation 33). Therefore figure 26.16 of [16] is interpreted as the inverse of the bearing stiffness ($1/k_r$) for the purposes of this thesis. This explanation seems to be supported by figure 26.17 on p. 1036²⁷ of [16], in which an increase in speed leads to a decrease in normalized bearing stiffness, when the contact angle is much larger than zero. Nevertheless, for very small values of contact angle the opposite effect occurs (the normalized bearing stiffness increases with an increase in speed

²⁷ The caption of figure 26.17 of [16] does not seem to match the text on p. 1036.

– refer to figure 26.17 of [16]). The effect of contact angle and load on bearing stiffness is derived in equation 26.66 of [16]

$$k_r = \frac{\partial Q}{\partial r} = \frac{3}{2} K^{2/3} Q^{1/3} \cos^2 \alpha \quad (34)$$

The above equation predicts that an increase of the contact angle α would decrease the bearing stiffness k_r . In the text it is not stated whether the contact angle α of equation 26.66 refers to the outer race contact angle α_o or the inner race contact angle α_i . In the text of [16] it is mentioned on p. 1035 that the outer raceway contact angle decreases as the speed increases (the opposite for the inner race). Hence it is assumed that the contact angle α is in fact the outer race contact angle α_o . This assumption helps to correlate equation 26.66 well with figure 26.17 of [16], but not with figure 26.16 (in which an increase in load results in an increase of the bearing stiffness). However, increasing the load Q in equation 26.66 increases the bearing stiffness k_r . The shape of the graph of normalized bearing stiffness against normal load in figure 26.16 seems to correspond well with a $1/(Q^{1/3})$ relationship (for example as Q approaches 0 a large value would be expected from a $1/(Q^{1/3})$ relationship, correlating well with the existing figure 26.16). If it is therefore assumed that figure 26.16 in fact shows $1/k_r$ instead of k_r , the text agrees better.

The conclusion for this thesis from [16] is that that the normalized bearing stiffness decreases with an increase in speed, which agrees with [6] and the results of this thesis²⁸. Additionally, the bearing stiffness increases with an increase in bearing load. This is supported on p. 969 of [16]: “Bearing stiffness increases with increasing load, a characteristic referred to as a ‘hardening’ spring.” From the preceding discussion the theoretical prediction of vibration would be to increase with increasing speed and to decrease with increasing load. Here it is assumed that low frequency and high frequency behaviour is identical. Yet, high frequency behaviour only is analysed in this thesis (as the low frequency components are subject to contamination by noise). Further exploration on

²⁸ Unfortunately, no experimental results could be found on the effect of load and speed on bearing stiffness.

the effect of static and dynamic load on low and high frequency vibration spectrum is left for future work.

The effect of a dynamic load – such as an unbalance mass – on vibration signals differs from a static load. In theory, the additional mass of the unbalance mass should increase the vibration levels (if one were to assume a linear spring, which is not actually the case – a simple spring model of the bearing should actually be non-linear to achieve accurate modelling [16]). However, the applied unbalance mass is generally much smaller than the mass of the shaft. The rotating unbalance mass causes the roller closest to the unbalance mass to make contact more firmly with the outer race. Hence the load distribution of the bearing is no longer statically distributed, but rather changes dynamically with the position of the unbalance mass.

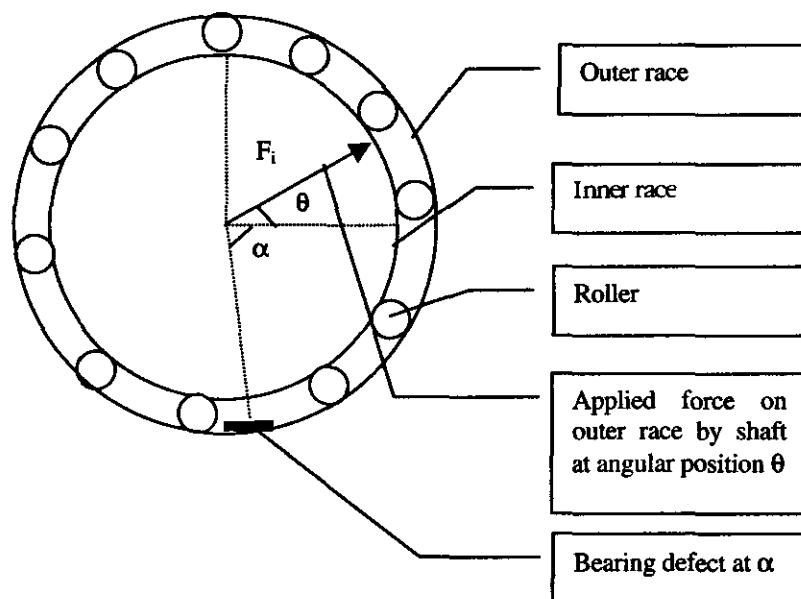


Figure 3: Effect of dynamic load on load distribution (drawing is not to scale²⁹)

²⁹ In the figure the balls are not touching, although they would be close together in an actual bearing (the balls are held together in the cage).

Figure 3 shows the effect of a dynamic load in the form of an unbalance mass on the loading distribution. The angular position of the unbalance mass is given by θ (theta). As the unbalance mass rotates at frequency F_s , a force is applied from the shaft to the inner race at frequency F_s (rotational frequency of the shaft). The inner race transmits the force F_i to the outer race through the roller (closest to the unbalance mass on the outer race at angular position θ). Hence it is clear that the load distribution on the rollers will no longer be symmetrical; instead the load will be increased on the roller at position θ . When the position of the unbalance mass corresponds to the position of the bearing defect (when $\theta = \alpha$) the load on the roller should increase slightly. However, as soon as the unbalance mass rotates away from angle α , the load at angular position θ increases slightly and the load at α decreases slightly. Thus it is clear that the dynamic load causes the load on the roller closest to the bearing defect (at angular position α) to decrease for a significant proportion of the time that the unbalance mass takes to travel around the circumference of the shaft. So while the load on the roller closest to the bearing defect could increase momentarily, the load on the roller closest to the bearing defect would also decrease for a larger proportion of the time.

This hypothesis proposed in this thesis seems to be supported qualitatively by [82], in which it is shown that the loading rotates at shaft frequency f_o as a result of imbalance, hence the vibration also has a period f_o . Unfortunately, the effect of increased dynamic load is not discussed in more detail in [82]. In [55, 84] it is also indicated that a defect out of the loading zone would decrease vibrations. As the bearing defect stays constant in the specific test rig, any variation in the observed demodulation peaks must be attributed to a change in the dynamic loading acting on the roller closest to the bearing defect (refer to chapter 7 for further discussion). When the load on the roller closest to the defect stays constant, the demodulation peak as a function of buffer length could also be expected to increase linearly with buffer length (refer to section 7.3.8 on p. 235 for further discussion). Any instantaneous deviation in the load on the roller closest to the defect would change the value of the demodulation peak (as a function of buffer length). As shown in section 7.3.8 virtually no increase is evident in the value of the demodulation peak for

some portions of the buffer, implying that the instantaneous load on the roller (closest to the defect) is correspondingly small. As indicated in [6] an increase in load may lead to a decrease in vibration in specific circumstances (owing to the complicated transfer function from the applied load to the bearing). It is hypothesized that a similar effect may be observed in the test rig used in this thesis: as the applied load is increased, the load (applied unbalance mass) circulates around the shaft and increases the instantaneous load on the rollers. But the load on the roller closest to the bearing defect increases only once per shaft revolution and may actually decrease for the rest of the revolution. Consequently, the load on the roller closest to the bearing defect may actually decrease for a large part of a shaft revolution. An inspection of the instantaneous value of the vibration demodulation peak in section 7.3.8 – in which the vibration sometimes stays constant for almost a shaft revolution – seems to support this hypothesis (at least for the medium and small bearing defects).

It might also be possible to model the defective bearing as a bearing with time dependent bearing stiffness. As soon as the roller contacts the bearing defect, the bearing stiffness rapidly decreases, resulting in a sudden increase in vibration (so the bearing stiffness would change at the characteristic bearing defect frequency). An increase in the total dynamic load (applied unbalance mass) could lead to a decrease in the dynamic load on the roller closest to the defect (with a corresponding increase in the dynamic loading on the other rollers not associated with the defect). So the vibration could in principle decrease with an increase in dynamic load (unbalance mass).

Usually the centre of the loading zone would be at an angular position of 180° from the vertical as a result of gravity. Figure 4 illustrates the effect of the pressure plate, applying pressure on the bearing housing. The resultant of the vector forces is no longer at 180° (which would result from the gravity force F_g). Owing to the horizontal pressure of the plate at 270° (pressing on the outer race of the bearing) it is rather closer to 225° as a result of vector addition. The exact resultant vector would depend on the plate pressure (with no plate pressure the angle should be close to 180° , while with maximum plate pressure the

rollers will be forced motionless into the races at 270°). Hence the maximum value of vibration should be found closer to 225° , rather than at 180° ³⁰.

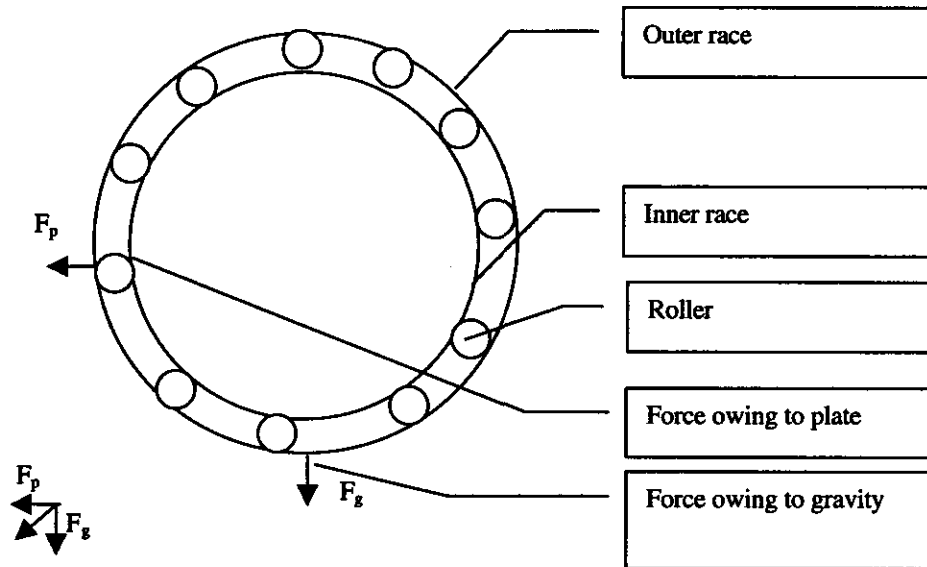


Figure 4: Shift in position of loading zone owing to the pressure plate in theVib demo test rig

As mentioned in section 2.7 on p. 34 of [16], Harris briefly described the influence of dynamic load on bearing stiffness. However, the concept of bearing stiffness seems to imply the low frequency response at the characteristic defect frequency and cannot be directly compared with the high frequency resonance technique (HFRT), in which the high frequency defect pulses are examined. The information in [16] seems to support the hypothesis of a change in the loading zone with dynamic load (an unbalance mass would correspond to the ‘radially loaded bearing’ of [16]), in particular a rotation of the loading zone with dynamic load as proposed in this section. The possibility of very small (or even zero) instantaneous loading on the outer raceway for brief periods is also mentioned in [16]. This will be discussed in more detail in the light of experimental results in sections 7.3.3 and 7.3.8 on pp. 224 and 235 (headings are: ‘*Feature BPFO2_X*’ and ‘*Feature BPFO2_X – influence of buffer length*’).

³⁰ It will be shown later that this is consistent with the experimental results.

4.3 Laboratory measurements of Vib demo data sets 1 and 2

Multiple fault conditions were induced on a Vib demo VIB 2.100 Pruftechnik AG test rig (see figure 5). The test rig consisted of three plummer block bearings, which support a shaft. An outer race defect was induced on the centre bearing of the test rig. Residual imbalance (dynamic loading) was induced onto the shaft of the system, using weights of 0, 12, 18 and 24 grams respectively, to simulate a deteriorating secondary fault condition. The 0 gram weight represents a very low imbalance condition (in practice it is not possible to eliminate all vibrations completely at shaft speed), the 12 gram weight signifies a medium imbalance condition and the 24 gram weight depicts a high imbalance condition. Acceleration measurements were taken on the centre bearing and on the bearing closest to the induced imbalance. Vibration measurements were taken using 100 mV/g ICP accelerometers in both the horizontal and vertical directions during the four tests.

The accelerometers consist of a seismic mass, which may vibrate in sympathy with the monitored object, when the accelerometer is fixed firmly to the surface of the monitored object. The seismic mass is attached to a piezoelectric transducer, converting the change in vibration level to an electric signal. This signal is amplified and then sampled for further processing in a computer. The sampling frequency used in this thesis is 25.6kHz (oversampling by 2.56) for a bandwidth of 10kHz, which is necessary to preserve the high frequency defect pulses³¹. In addition to the vibration signals, a current signal and trigger signal have also been measured. In some measurements an acoustic signal has been measured instead of the current signal. A series of handheld vibration measurements has also been made at different locations on the test rig. The vibration signals were then processed in the frequency domain and components evaluated at the characteristic defect frequencies (depending on the geometry of the bearing).

³¹ Some of the last measurements on the small bearing defect have been made at 51 200Hz by accident. Where possible these data have been decimated by a factor of 2 (with appropriate filtering) before the extraction of features in order to have a single sample rate. However, the total buffer length after decimation is obviously shorter using a higher sampling rate.

A DSP Siglab analyser model 20-42 was used to collect the data. The rotational speed of the system at 1626rpm was measured using an ONO SOKKI HT-4100 digital tachometer. The frequency band was set between 5Hz and 10kHz for the analysis. The test rig is shown in figure 5 and the bearing details are presented in table 1. The Vib demo data set 1 consisted entirely of measurements made by a magnet-mounted accelerometer and will not be discussed in the thesis (however, the measurements obtained do support the general conclusions of this thesis). Vib demo data set 2 consisted entirely of permanently mounted accelerometer measurements (on at least one measurement channel). Great care has been taken in the setup of the Vib demo data set 2 to optimize all aspects related to vibration analysis.

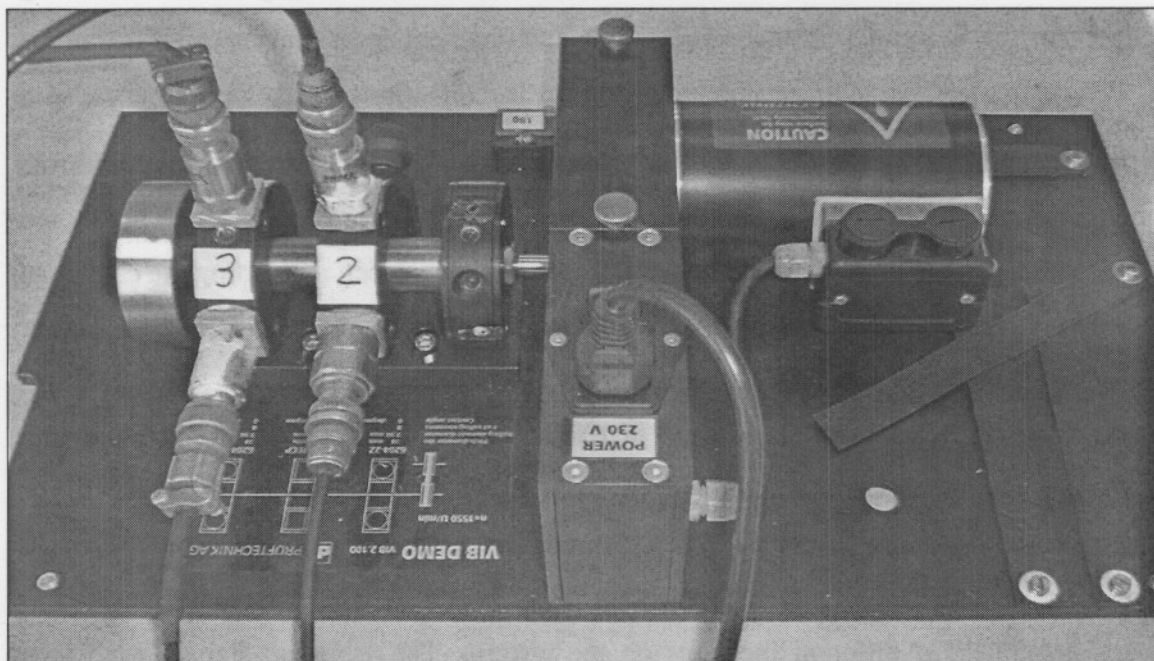


Figure 5: Test rig

	Normal Bearings	Defective Bearing (nameplate)	Defective Bearing (measured)
Model Number	6 204-2 z	NJ204ECP	NJ204ECP
Pitch Diameter	34mm	34mm	34mm
Rotating Element Diameter	7.94mm	6.5mm	7.5mm
Number of Rolling Elements	8	10	11

Table 1: Bearing specifications

4.4 Laboratory measurements of Vib demo data set 3

An additional series of measurements has been made on the Vib demo test rig to address the following unresolved questions associated with Vib demo data sets 1 and 2:

- What bearing defect has been induced?
- What is the effect of speed variation on the vibration signals?
- How do increasing bearing defects, in the presence of a secondary fault condition (imbalance), affect vibration signals? In particular, what is the characteristic of a 'normal' bearing signal (compared with a defective bearing)³²?
- Is it possible to use acoustic, current and handheld vibration measurements (in various positions around the test rig) for fault detection in this data set?
- A clear demodulation peak has been observed for the defective bearing, yet the time domain signal did not clearly show repetitive energy spaced at *BPFO*. What is the explanation?

³² The 'normal' measurements, obtained in the Vib demo data set 2, have been made on a different bearing than the defective bearing. The validity of comparing measurements on bearings of different construction has been questioned. Also, a vibration peak (of much smaller amplitude) has been observed at the *BPFO* defect frequency in the demodulated spectrum, hence the inference that vibration from the defective bearing has been transmitted to the 'normal' bearing.

The information printed on the nameplate for the NJ204ECP bearing listed the roller diameter as 6.5mm. This information was inconsistent with the observed defect frequencies in the Vib demo data set 2. A physical measurement of the roller diameter showed that the diameter was 7.5mm. (In chapter 7 it will be shown that this measurement is consistent with an outer race defect. The outer race defect was confirmed visually.)

The following measurements have been made in the Vib demo data set 3:

- Channel 1 is always a speed reference signal (triggered either on the shaft or rollers, using an eddy current vibration sensor [19])³³.
- Channel 2 is measured, always using a permanently mounted accelerometer, in the x axis radial direction.
- Channel 3 is measured, using either a permanently mounted accelerometer or handheld accelerometer, in the y axis radial direction (or other positions on the test rig in the case of the handheld accelerometer).
- Channel 4 is measured, using either a current sensor or acoustic measurements.

Handheld measurements have been made in different physical positions around the test rig. Acoustic measurements have been made in three physical positions (i.e. very close to the defective bearing, about 30cm from the defective bearing and far from the defective bearing (about 1m) using a Panasonic DS38 digital video recorder. Environmental conditions affecting the experiments include the looseness of the knurl knob applying pressure to the bearing plate (in this thesis only results with the knob turned tight have been included) as well as disassembly of the bearing blocks.

The introduction of additional bearing defects required physical disassembly and reassembly of the test rig. Care has been taken to place the bearing block always back in the same physical position (otherwise shaft misalignment could result). The disassembly process also required the loosening of a knurled knob. The purpose of this knob is to apply pressure to the bearing. The knob has always been tightened to the point that it could not

³³ Probably as a result of poor mechanical contact, some measurements do not have a valid speed reference signal. One of the bearing cover plates has been removed in the test rig to gain physical access.

be moved further by hand. It will be shown that the load angle, which is normally centred at 180° , may be influenced by the tension of the screw. The largest defect peak has been observed when the outer race defect is placed at 225° (and not at 180° as in Vib demo data set 2)³⁴. Owing to time constraints increments of 45° have been used in the defect position.

The measurements have been obtained in a large laboratory, with plenty of external noise sources. As such the vibration measurements have some additional harmonic components, probably related to the increased noise sources. Some of the measurements (at the end, for a small bearing defect) have been made at 20kHz instead of 10kHz, which is usually used³⁵. This data set is consequently perhaps more representative of typical measurements that would be obtained in an industrial setting, whereas the Vib demo data set 2 represents the best that can be obtained in the laboratory. In both cases no external noise, such as might be obtained from a gearbox, has been added (which would make the SNR much smaller).

4.5 Recommendation for bearing fault detection

By varying the position of the outer race the defect angle (DA) may be adjusted with respect to the loading zone. A 'normal' or new bearing should display little variation in vibration at the defect frequency (demodulation spectrum) with variation in DA (although a 'normal' bearing always has small discontinuities – incipient faults – on the surface). However, a bearing with localized defect will show a large variation of vibration with the defect angle (DA) – please refer to section 7.3. Hence each area of the bearing surface (outer race) may be placed in the loading zone to examine the influence of the defect angle (DA). Although DA is varied in steps of 45° in this thesis, in principle it could be varied in much smaller increments if required (e.g. by using a stepper motor). The mechanical details required to implement this are outside the scope of this thesis. Another potential

³⁴ The demodulated BPFO vibration signal with defect at 270° is also larger than at 180° , suggesting that the load zone is closer to 270° than 180° . It is easy to understand the shift in load angle in terms of a vectorial sum of forces (gravity working downwards and pressure applied at 270°), resulting in a maximum force somewhere between 270° and 180° .

³⁵ A correction has been made to the feature extraction program, otherwise the second harmonic of RF and BPFO would have been used.

advantage could be that the inner race (usually rotating at close to shaft speed) could be held stationary if the outer race is rotated at a speed close to the shaft speed (as the inner race and outer race both rotate close to shaft speed, the inner race would effectively rotate at a very low frequency corresponding to the difference in frequency of the outer race and inner race). This would enable the loading effect associated with an inner race defect to be removed. The surface of the inner race could also be examined as a function of the defect angle (DA) by controlling the speed of the outer race.

Another possibility would be to mount the test rig on a table that can rotate, allowing the force of gravity to move the loading zone away from the defect. This would, unfortunately, only seem practicable on smaller test rigs (such as the Vib demo test rig). The potential advantage of this technique is that no mechanical modifications are required to the bearings.

In practice, this would require considerable mechanical changes to the system, which would be justified only if the cost of the machinery (or results of failure) is high enough. In future it would be better to incorporate this idea into the design stage itself. Yet, the basic idea is simple: allow the outer race to rotate (either slowly to investigate outer race defects or fast in order to freeze inner race defects).

4.6 Conclusion

The experimental setup and measurements, believed to illustrate several innovations, have been described. The first requirement of separating a fault or no fault condition may be achieved easily with the proposed DA (defect angle) technique without requiring physical disassembly of the bearing (which would introduce undesirable environmental factors in the measurements). The second requirement of applying different fault conditions (such as outer race, inner race and roller defects) may also be achieved simply owing to the construction of the test rig (however, physical disassembly of the bearing housing is required). The third requirement of varying the degree of dynamic loading may be achieved simply by adding unbalance masses to the setup. The static loading itself may

also be changed by adjusting the DA³⁶. Hence it is concluded that the test rig is well suited to the experimental demands.

³⁶ Formal proof of this statement is left for future work.

Chapter 5

5 Theoretical development and results of pattern recognition techniques

5.1 Introduction

In this chapter artificial intelligence techniques will be applied to the problem statement of this thesis, which is the analysis of multiple fault conditions in condition monitoring applications. Firstly, the MIFS technique of Battiti [4] is described theoretically on p. 76. Secondly, it will be shown that the nearest neighbour rule (NNR) technique may be extended to modelling problems in a test rig (Vib demo data set 2) where multiple fault conditions are present (instead of classification problems only). This is believed to be a novel contribution. Finally, it will be demonstrated how efficient cross validation can be implemented on the standard Iris data set (implementing leave one out cross validation is generally not feasible in neural networks owing to the time-consuming parameter optimization that has to be performed). Validation of any model is vital to ensure that the model may generalize adequately.

5.2 Background

In classification problems the number of classification options (or classes) is finite, in contrast to modelling problems in which the output signal could be continuous in principle (even if sampled digitally with a finite precision). The quality of the extracted features has a major impact on the classification performance that can be achieved. Hence expert knowledge and feature selection or transformation can be invaluable in optimizing the overall performance. A neural network architecture with one hidden layer and sufficient complexity (number of neurons in the hidden layer) can, in principle³⁷, be applied to map any input signal(s) to any output signal(s). Owing to the non-linear mapping of the neural network it is also possible to easily overfit the data set, resulting in poor generalization

³⁷ It is assumed, of course, that the right features are chosen, correct preprocessing is applied, the correct learning strategy is used and sufficient time is available for convergence (i.e. the parameter optimization is performed correctly).

performance ('training to the noise'). Thus special care needs to be taken to employ these techniques successfully in a condition-monitoring context where multiple and complex fault conditions are present.

5.3 Theoretical development

5.3.1 Mutual information (MI) based feature selection

The mutual information feature selection (MIFS) algorithm of Battiti [4] is based on the concept of mutual information derived from the field of Information Theory, which supplies the amount of information that a random variable Y conveys about X , or vice versa

$$\begin{aligned} MI(X; Y) &= K(p(x, y), p(x)p(y)) \\ &= \sum_{x \in \mathbf{X}} \sum_{y \in \mathbf{Y}} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \end{aligned} \quad (35)$$

Equation 35 is the Kullback-Leibler distance between the joint probability and the factorized probability. It is equal to zero for independent X and Y . Since it is in general not computationally efficient to compute the mutual information between all the features, owing to the curse of dimensionality, a greedy selection algorithm has been developed [4].

The MIFS algorithm contains a class dependent component and a feature dependent component. The MIFS algorithm firstly computes the mutual information of the features with respect to the class variable, i.e. $MI(f, C)$. It is likely that the best set of features thus selected would still contain mutual correlation components. The second step of the MIFS algorithm computes the correlation between the candidate feature and the best feature set found so far, i.e. $MI(f_1, f_2)$. Candidate features with good class membership correlation, but highly correlated with previous features selected, are penalized by an adjustable parameter α . In this way, a feature set is obtained with good class discrimination properties as well as minimal correlation between the selected features. Further implementation details are available in [89].

5.3.2 Improved cepstral analysis

The complex cepstrum preserves phase information and allows reconstruction of the input signal (after editing in the cepstral domain). Owing to the log operation applied to the magnitude spectrum of the signal, very small components (as a result of spectral leakage) are magnified and influence the resulting cepstrum. The following modified cepstrum algorithms are proposed:

- Linear cepstrum (LCEPS) does not apply the usual log operation (refer to equation 18) when applying the Fourier transform inherent in the (real) cepstrum³⁸.
- Thresholded cepstrum (TCEPS) thresholds small FFT components (the threshold is relative to the maximum value of the FFT). It uses the usual log operation of the cepstrum (refer to equation 18). The thresholding operation is a non-linear operation. The input signal cannot be perfectly reconstructed after application of thresholding.
- Linear, thresholded cepstrum (LTCEPS) combines the linear and thresholded cepstrum.

The modified cepstrum algorithms can be helpful when:

- the dynamic range of the magnitude spectrum is very large
- the magnitude spectrum has very small values that are amplified after applying the log operation to the FFT; when the IFFT operation of the cepstrum is subsequently applied, the amplified values interfere (refer to figure 29)

Signal processing techniques, such as the cepstrum, can be influenced by the noise caused by angular speed variations of the shaft. Although a reference signal related to the angular position of the shafts can reduce this problem, this signal is unfortunately not available in many industrial data sets.

³⁸ The complex cepstrum can be defined as [13] $\chi(n) = \mathcal{F}^{-1}\{\log \mathcal{F}\{s(n)\}\}$ in which log now denotes the complex logarithm $\log z = \log |z| + j\arg(z)$. The complex cepstrum preserves phase information, unlike the real cepstrum. To compute $\arg(z)$ phase unwrapping is required, which can be difficult in practice [13]. A linear complex cepstrum (LCCEPS) could be defined analogous to the linear real cepstrum (LCEPS) by using $\log z = |z| + j\arg(z)$ instead of $\log z = \log |z| + j\arg(z)$. However, further research on the properties of the proposed LCCEPS is left for future work.

5.3.3 Modelling of RMS level of bearing defects

Experimental results of the defect width versus RMS level described in [107] are in agreement with equation 22 (although the effect of static load is not made explicit in equation 22, requiring calibration for each load condition, which may be problematic in an industrial setting). If the defect width x is much smaller than D (which would be true for very small bearing defects), equation 22 can be further simplified (not discussed in [107]) to

$$v_s = \frac{d_m n_i \pi}{60D} (1 - \gamma) x \quad (36)$$

Equation 36 indicates that the impact velocity v_s is proportional to the defect width x , at least for small defects ($x/D \ll 1$). If the contact angle is zero, equation 36 simplifies to

$$v_s = \frac{n_i \pi}{60} (d_m / D - 1) x \quad (37)$$

The RMS level of vibration is then related to the square of the impact velocity v_s [107], or the square of the defect width (at least for small defects and zero contact angle).

5.3.4 Composing validation sets

When building a model the ideal would be to achieve approximately the same accuracy on the training set as on the test set. (Hence unless the accuracy on the training set is close to that of the CVCP algorithm, overfitting could be expected).

It stands to reason that any model (which could be obtained from regression, neural networks, fuzzy logic, etc.) has to be validated on a data set different from the data set used to train it³⁹. If the data set is sufficiently dense (for example by using a very fine

³⁹ The data set used for testing during training is called the 'validation set', whereas the data set used for testing after training is called the 'testing set'. The performance on the test set is used as a measure of the generalization performance of the classifier. Ideally the test set should be as large as possible. Unfortunately, increasing the test set size reduces the size of the training set (and hence the complexity of the neural models that could be safely deployed).

mesh or grid⁴⁰), a picture of the 'smoothness' of the model can be envisaged. The data set for validation⁴¹ may of course be obtained in many ways:

- A testing set is kept separately. The obvious disadvantage is that fewer samples are available for training, which might lead to a poorer model. For large data sets this would likely be the simplest option (in the real world large data sets are, unfortunately, scarce). The test set should also be representative of the relationship to be learnt. If the test set is too small, misleading results could be obtained.
- k fold cross validation is used with a number of test sets. The obvious disadvantage is that the model optimization has to be performed k times. For LOO XVAL (leave one out cross validation) $k = N$ (N is the number of data samples) and the validation set consists of one sample only, often leading to prohibitive computational complexity.
- The original data set is augmented with additional samples. The additional samples are used to augment the validation set⁴². This could be based on expert knowledge about the specific problem domain (for example, by asking an expert what the value of the model should be, given the feature set).

An alternative would be to use interpolation between data samples to generate additional data samples. The simplest option would be to use zero order interpolation, as with the NNR⁴³. Although the NNR uses N validation samples only (where N is the data set size) for validation, it would be possible in principle to construct a grid with any number of data samples. Any model with too many free parameters (such as a neural network) would then be penalized for any deviations from the target values on the grid. Hence the local behaviour of the model could be regularized to a certain extent (deviations from the piecewise constant values in the localized grid would increase the overall error). A

⁴⁰ Evaluating the output of the classifier on a fine mesh would pose severe computational problems as the number of features increases ('curse of dimensionality'). Furthermore, obtaining the class membership values on the fine grid would also be time consuming.

⁴¹ Validation here refers to a test set after training of the model itself is completed.

⁴² They could also be used to augment the training data set, at the expense of a smaller test set.

⁴³ It is assumed that the samples are dense enough (which is also a general requirement for application of the NNR).

mesh or grid⁴⁰), a picture of the ‘smoothness’ of the model can be envisaged. The data set for validation⁴¹ may of course be obtained in many ways:

- A testing set is kept separately. The obvious disadvantage is that fewer samples are available for training, which might lead to a poorer model. For large data sets this would likely be the simplest option (in the real world large data sets are, unfortunately, scarce). The test set should also be representative of the relationship to be learnt. If the test set is too small, misleading results could be obtained.
- k fold cross validation is used with a number of test sets. The obvious disadvantage is that the model optimization has to be performed k times. For LOO XVAL (leave one out cross validation) $k = N$ (N is the number of data samples) and the validation set consists of one sample only, often leading to prohibitive computational complexity.
- The original data set is augmented with additional samples. The additional samples are used to augment the validation set⁴². This could be based on expert knowledge about the specific problem domain (for example, by asking an expert what the value of the model should be, given the feature set).

An alternative would be to use interpolation between data samples to generate additional data samples. The simplest option would be to use zero order interpolation, as with the NNR⁴³. Although the NNR uses N validation samples only (where N is the data set size) for validation, it would be possible in principle to construct a grid with any number of data samples. Any model with too many free parameters (such as a neural network) would then be penalized for any deviations from the target values on the grid. Hence the local behaviour of the model could be regularized to a certain extent (deviations from the piecewise constant values in the localized grid would increase the overall error). A

⁴⁰ Evaluating the output of the classifier on a fine mesh would pose severe computational problems as the number of features increases (‘curse of dimensionality’). Furthermore, obtaining the class membership values on the fine grid would also be time consuming.

⁴¹ Validation here refers to a test set after training of the model itself is completed.

⁴² They could also be used to augment the training data set, at the expense of a smaller test set.

⁴³ It is assumed that the samples are dense enough (which is also a general requirement for application of the NNR).

problem would still be the relative sparseness of data samples in multidimensional space, which would require further research into the role of the grid for validation purposes (uniform versus non-uniform, size and shape of the grid elements). Intuitively it is clear that the grid should be reasonably ‘close’ to the data samples, as the behaviour of the model far from the data samples is unknown. The NNR is one example of a grid conforming to this requirement.

5.4 Results

5.4.1 A nearest neighbour rule with class membership (NNRC) for modelling problems

5.4.1.1 Introduction

The nearest neighbour rule (NNR) has been used widely to determine a bound on the performance of classifiers. It has been shown that the error rate of the nearest neighbour classifier bounds the optimal Bayes error rate by a factor of at most two [7, 77]. NNRC, a nearest neighbour rule with class membership, is presented in this thesis to model the multiple fault conditions on a test rig. NNR can be used for classification problems only. Thus we extend the NNR with NNRC to allow the use of continuous class labels as well.

The nearest neighbour rule has been successfully applied in the past to a variety of real-world classification problems. As the NNR effectively stores or memorizes the training set, the computational resources required can become a problem with large data sets (e.g. in data mining applications). Various efficient alternatives to the NNR have been proposed in the past, e.g. CNN, RNN, ICA, ENN and SNN [79]. The aim is to find the minimally consistent subset, which classifies the original training set correctly. Other approaches include learning vector quantization and clustering, in which a vector or cluster centre attempts to represent a number of samples.

However, the focus of this section is not on the problem of computational complexity, but rather on extending the definition of the NNR to modelling problems. In this section NNR is extended as a modelling tool in the form of NNRC.

5.4.1.2 Problem formulation

A technique is presented in this thesis that allows continuous class membership values, instead of the binary class labels usually associated with the NNR. The gamma test has been used to determine the data set size necessary to build a reasonable model, as well as to select the important features [36]. However, the gamma test is still based on the NNR. The NNR selects the closest sample, irrespective of the class label. The approach here is to select the closest class (or cluster) to the sample. The output value of the NNRC can vary continuously, while the NNR always assigns a discrete class label (based on the distance to the closest neighbour). This enables the NNRC to be applied in the modelling of a primary fault mechanism (an outer race defect) in the presence of a secondary fault mechanism (an unbalance mass) [98]. The advantage of the NNRC over other modelling techniques, such as neural networks (multilayer perceptrons or radial basis functions), is that no lengthy training time is necessary.

The specific application, in which a primary fault mechanism is obscured by the presence of a secondary fault mechanism, is also novel. Although various authors have investigated multiple fault mechanisms in the past [31, 32, 33, 34, 42, 43, 44, 46, 47, 49, 60, 69, 76], no specific reference could be found to a stable primary fault condition (a bearing defect) in the presence of a deteriorating secondary fault mechanism (an induced imbalance). The NNR always assigns the class label of the closest sample, using a suitable metric such as the Euclidean metric, to the sample to be classified. The decision surface associated with the NNR is consequently always discrete. No use is made of the distance of the sample to the nearest sample(s) or class(es). The NNRC algorithm gives high class membership values if the sample is close to a nearby class, and low membership values if it is not. By allowing continuous class membership values, the NNRC algorithm can be used for modelling problems.

5.4.1.3 NNRC algorithm

5.4.1.3.1 Nearest neighbour rule viewed as a neural network

The kNNR, which is a classification rule, uses a metric first to compute the distance from a new sample to every other sample in the training set. The class, determined by voting

amongst the k nearest neighbours, is then assigned as the best estimate of the class label of the new sample. The distance metric used may be important in the performance of the kNNR classifier. The Euclidean distance is often used as a distance metric

$$D(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^k (x_i - y_i)^2} \quad (38)$$

where $\mathbf{X}$ is the sample vector to be classified, and $\mathbf{Y}$ a vector from the training set [7, 77, 87]. The class labels of the k nearest neighbours, i.e. the set of k vectors closest to the sample (with minimum value of D), determine the predicted class label of the kNNR algorithm. The functional form of the distance metric determines the shape of the basis function. Suppose the distance metric is defined as the L_1 norm

$$A(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^k |x_i - y_i| \quad (39)$$

The distance metric corresponding to $A(\mathbf{X}, \mathbf{Y})$ may be visualized as a (triangular) straight line segment with slope of one, centred at vector $\mathbf{Y}$.

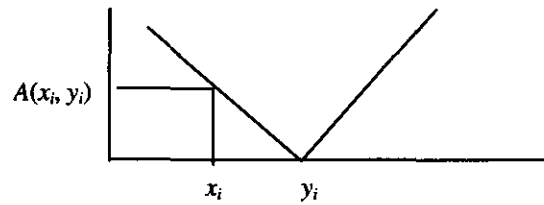


Figure 6: Visualization of distance metric $A(\mathbf{X}, \mathbf{Y})$

Figure 6 shows that the value of $A(\mathbf{X}, \mathbf{Y})$ increases linearly to the distance between the current sample x_i and the training set y_i . The training set samples y_i tile the axis corresponding to each dimension of x_i .

Application of the L_2 norm results in the same shape of graph when considering a single dimension only. In two dimensions a circle is formed with the radius proportional to the distance from the centre of the basis function. In more dimensions a series of hypercircles is formed. As far as the NNR (classification rule) is concerned, adding a constant, or scaling all the variables with a constant, has no effect on determining the minimum value.

Zero mean variables with unity variance or norm may therefore be used in the NNR without changing the classification results.

The NNR therefore consists simply of a series of basis functions, centred at the training vectors, with increasing activation values proportional to the distance from the basis centres. When the Euclidean metric is used, the activation values increase quadratically in proportion to the values of the features. The output, using the Euclidean metric, is given by the sum of the activation values. In two dimensions, the result would be a circle, whereas in higher dimensions the result would be a hypercircle. More generally, the surface associated with two dimensions (features) would be a cone, whereas the surface associated with higher dimensions would be a hypercone.

It should be noted that the NNR could be sensitive towards the scaling of individual variables.

5.4.1.3.2 Nearest neighbour rule as linear interpolator

The previous paragraph described the NNR as a special form of RBF neural network, which uses a distance metric and the complete training set as basis vectors. As it is well known that increasing the number of free parameters in a classifier can lead to overfitting problems [7, 77, 87], the question arises as to why the NNR has been found to give good results in general, with an error rate of at most twice the optimal Bayes error rate [7, 77], when using such a large number of basis functions.

Firstly, it should be noted that the basis functions of the NNR are fixed in position and width (extent) once the ‘training’ is completed⁴⁴. Hence no ‘free parameters’ exist in the NNR as can be found in other classifiers (such as MLP and RBF). Rather, the basis functions are used as is with a suitable metric, eliminating the risk of overfitting. Secondly, note the good local behaviour of the NNR classifier. In one dimension, the activation value is proportional to the scalar variable. With feature values of (x_1, x_2) and output values of (y_1, y_2) respectively, the NNR would give an activation value of

⁴⁴ In fact, no training is performed in the NNR – the basis functions simply correspond to the training set data, while the classification rule is the nearest neighbour algorithm.

$$y_{NNR} = \frac{x - x_1}{x_2 - x_1} * (y_2 - y_1) + y_1 \text{ with} \quad (40)$$

$$x_1 \leq x \leq x_2 \quad (41)$$

when performing linear interpolation between the two output values. The feature values x_1 and x_2 are observations from the training set with corresponding target values (class labels) of y_1 and y_2 . The activation value would be the mean value of y_1 and y_2 precisely when the variable x is halfway between x_1 and x_2 . Also, when x is close to x_1 , then y_{NNR} is close to y_1 . The NNR is thus well behaved when evaluated close to the basis functions, i.e. small changes in the input space x lead to small changes in the output space y_{NNR} . The NNR algorithm usually chooses the output vector (or class label) corresponding to y_1 or y_2 of the closest vector (x_1 or x_2) to x as the decision of the classifier.

In most other classifiers, especially including neural network classifiers, choosing a suitable architecture (number of layers, neurons, basis functions, etc.) involves a large part of the design effort. The NNR, in contrast, requires no explicit training time. However, the recall performance of the NNR is in general considerably longer.

The usual interpretation of the NNR algorithm is to assign the class label of the closest basis function as the output value⁴⁵, leading to a set of finite output values (corresponding to the set of values contained in the output vectors or class labels). For classification problems the usual NNR algorithm always has discrete class labels, with discrete values of 0 and 1 respectively. Let us consider a modified NNR with output value proportional to the weighted average (linear interpolation) of the outputs associated with the two closest samples. Hence the modified NNR would have an activation value of $\frac{1}{2}$ if the class labels⁴⁶ of the two closest samples are 0 and 1 respectively (and vice versa). From the

⁴⁵ A modification of the NNR algorithm may be considered: the distance to the nearest neighbour must be greater than zero. This can eliminate problems with some pathological examples, such as choosing the test set (of the NNR) to be the training set. The nearest neighbour of any sample in the test set would then always be the training set sample, hence the class labels would (always) be identical. Hence the result on the test set would always be 100%, instead of the desired LOO XVAL score obtained on the test set. The proposed modification to the NNR rule (i.e. the distance to nearest neighbour must be non-zero) could eliminate this undesirable effect.

⁴⁶ For a classification problem the class labels would be discrete and could be encoded with 0, meaning no class membership and 1, meaning full class membership. For example, to model the 12g imbalance a class membership value

preceding discussion and equation 40 it is evident that the NNR can also be given a probabilistic interpretation, in which the output of the NNR varies continuously between the class labels of the two closest basis functions (x_1, x_2). When viewed in this way, the good local behaviour of the NNR becomes evident and may partially help to explain the good performance results of the NNR obtained up to date.

The outputs of other classifiers, in particular neural network classifiers, are not constrained outside of the training set vectors. If the classifier has sufficient degrees of freedom, it might interpolate wildly between the training set vectors. A validation set is frequently used to detect the start of excessive interpolation (overfitting) by a neural classifier during training [7] (however, a test set is used to determine the generalization performance of a classifier after training). The idea of a fractional NNR algorithm, and its relation to cross validation, is examined in section 5.4.2.3.2.

5.4.1.3.3 Nearest neighbour rule with class membership (NNRC)

From the preceding subsections it is evident that the NNR represents a non-linear transformation of input space into output space. The output space consists of a number of discrete valued classes (output values) owing to the application of the minimum value function in the NNR. The definition of the class label (or output value for a modelling problem) is now extended in this thesis by introducing the distance from the nearest class. More specifically, the k classes are assumed to be represented by the variables c_k and each variable c_k is assumed to be encoded as follows:

$$c_k = \begin{cases} 1 & \text{for class } k \\ 0 & \text{for all other classes} \end{cases} \quad (42)$$

Let the distance d_1 from the vector x with class label c_k , using the L_2 norm, to the nearest class c_k (where $c_k = 1$) that is to be classified (or modelled) be given by

$$d_1(x, w) = \min \sqrt{\sum_{i=1}^n (x_i - w_k)^2} \quad (43)$$

and let the distance d_2 of the nearest class c_k (where $c_k = 1$) to the nearest other class c_k (where $c_k = 0$) be given by

$$d_2(w, z) = \min \sqrt{\sum_{\substack{i=1 \\ i \neq k}}^k (w_i - z_k)^2} \quad (44)$$

The variable w_i represents the centre of the nearest class c_k (where $c_k = 1$) to vector x . The variable z_k represents the centre of the nearest other class c_k (where $c_k = 0$) to vector x . The centre may be defined in various ways: for example, as the mean value of the samples belonging to class c_k , or as the closest sample from class c_k , or as the k -means vector obtained by the k -means algorithm (nearest cluster interpretation). In the present section only the mean value of the samples of class c_k , as defined in equation 44 will be considered. Hence it is assumed that the vectors belonging to class c_k are well clustered. If this condition is not satisfied, equations 43 and 44 can be readily modified by using the nearest cluster (obtained by a suitable clustering algorithm, or an edited NNR algorithm [77, 79]). Each class is thus divided into a number of clusters. An example of a standard machine learning data set that would require a number of clusters to represent accurately is the two spiral data set [91].

The NNRC decision rule can now be formulated as

$$y_{NNRC} = \begin{cases} 1 - d_1/d_2, & d_1 \leq d_2 \\ 0, & d_1 \geq d_2 \end{cases} \quad (45)$$

The following special cases are of interest:

$$d_1 \rightarrow 0 \Rightarrow y_{NNRC} \rightarrow 1 \text{ as } d_1 \rightarrow 0, d_2 \rightarrow \infty \quad (46)$$

$$d_2 \rightarrow 0 \Rightarrow y_{NNRC} \rightarrow 0 \text{ as } d_2 \rightarrow 0, d_1 \rightarrow \infty \quad (47)$$

If d_1 is much smaller than d_2 , the class label will approach one as expected. Also, if d_1 is much larger than d_2 , the class label will approach zero. The effective width corresponds to the distance d_2 between the two closest class centres (or more generally cluster centres). Class membership values decrease as the distance from the class (cluster) centre increases.

Provided that d_1 and d_2 are always positive (which is guaranteed when the L_2 norm is used), the output value y_{NNRC} is always a fractional value between 0 and 1, representing a probabilistic measure of the class membership of class c_k .

5.4.1.3.4 *Gamma test and fractional NNR*

The gamma test is used to measure the complexity of the input to output mapping before any model is trained [36]. For the special case of the NNR (nearest neighbour $k = 1$) and a normally distributed output with standard deviation of σ , the average distance between near neighbours would be σ . The asymptotic value of the gamma test would then approach σ for a sufficiently large data set (k is small enough). Classification problems may be represented as multiple output signals with discrete values of either 0 (no class membership) or 1 (full class membership). As the nearest neighbour rule becomes more accurate with an increase in nearest neighbours k (eventually decreasing in accuracy as the neighbour width becomes too large), no clear asymptote would be formed when applying the gamma test, as would be found with modelling problems. The discrete binary nature of the output values also invalidates the assumption of continuous first order derivatives required by the gamma test.

From the discussion in section 5.4.1.3.1 on p. 81 it is clear that the NNR is a special form of classifier with basis functions based on distance. Hence it may also be applied directly to modelling. Instead of a time-consuming process of mathematical parameter optimization for the training of neural network architectures, no lengthy training times are required using the NNRC. The NNRC algorithm can be efficiently evaluated in order $M \log_2(M)$ operations, where M is the sample size.

As the gamma test consists of an application of the NNR rule with decreasing values of k (nearest neighbours) to the input and output vector spaces, the NNR ($k = 1$) may be viewed as an application of leave one out cross validation (LOO CV) in the context of modelling problems [36]. The distance between neighbouring output values, when applying the NNR, gives an indication of the inherent complexity of the problem. Whenever the output values, corresponding to samples close together in the input space, differ greatly it will be difficult for any classifier to separate them [36].

Although the gamma test provides a suitable metric for evaluating model complexity before training a model, it still requires a model to be implemented. Evaluating the performance of the model can be done using LOO CV, but with high computational complexity.

An extension of the gamma test CVMP (cross validation strategy to determine modelling performance) is proposed in this thesis, in which the implemented model is evaluated using the NNR. The technique is an extension of the CVCP technique to modelling problems [97]. The gamma test, which is based on evaluating the nearest neighbours in the input and output spaces using the k nearest neighbour rule (kNNR), may hence also be applied to the implemented model. The performance of the model, derived from mathematical optimization procedures, can be evaluated in $M\log_2(M)$ operations only. In particular, any overfitting of the implemented model would result in a lower value of CVMP, owing to wild interpolation between the training set samples.

Extrapolation of the regression line in the gamma test, in order to determine the gamma estimate, corresponds to a fractional NNR algorithm [97]. Fractional values of k may be regarded as increasing the size of the training set. By extending the size of the training set, for example by interpolating between training sample vectors using linear interpolation, the performance of a trained model may be evaluated without expensive cross validation (CV) [97].

5.4.1.3.5 Results

The performance of the NNRC is compared with neural models on a vibration monitoring data set containing multiple fault mechanisms [98]. The amplitude at RF is commonly used to detect imbalance, whereas the *BPFO* component is indicative of a defective bearing. Figure 44 shows a scatterplot of the X and Y channels at RF. It is not possible to separate all the classes completely using the RF components only. The frequency spectrum of a bearing, with increasing imbalance imposed by adding weights to the setup, are shown in figure 1 and figure 2.

Figure 7 and figure 8 show respectively the output of NNRC and RBF classifiers trained on observations from all six classes. The *BPFO* feature indicating a bearing defect

condition increases from front to back, whereas the RF feature indicating an imbalance condition increases from left to right. The RBF classifier has been trained using the RBF training algorithm of Matlab [61]. The data for the NNRC technique has been normalized with respect to length (norm of one).

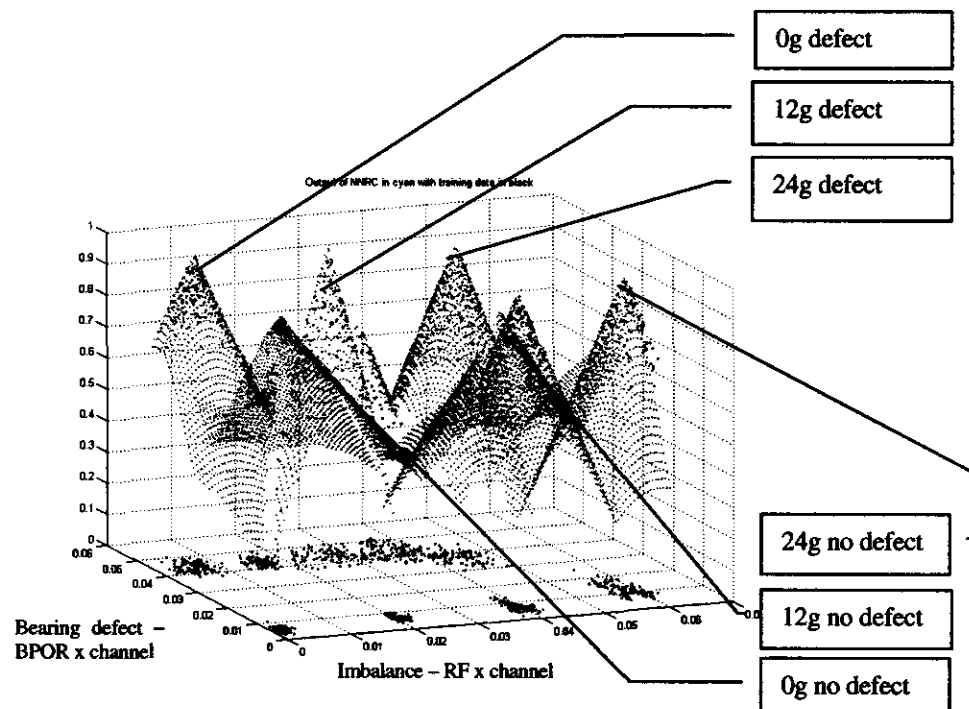


Figure 7: Response of NNRC classifier trained on all six classes (training set samples shown in black)

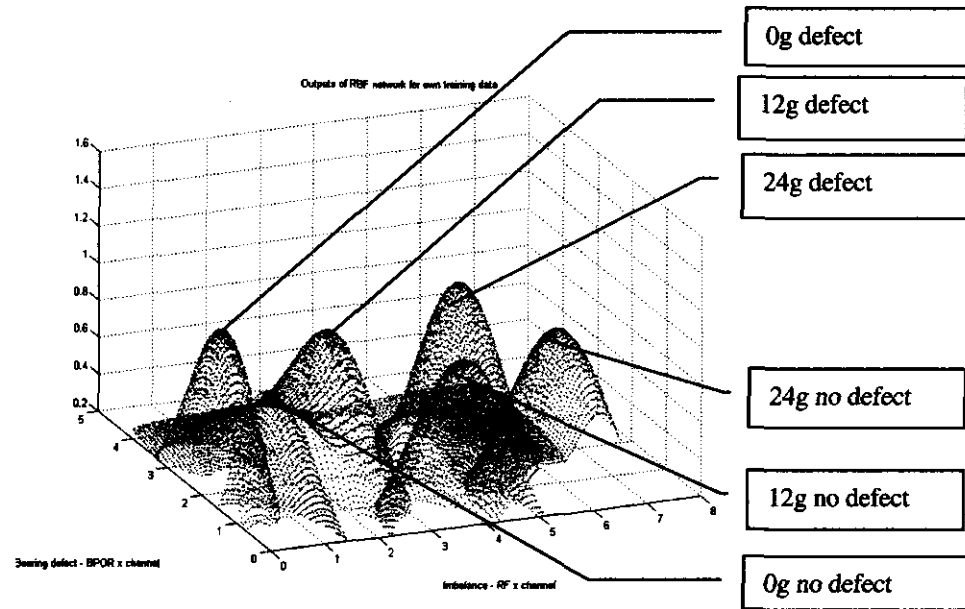


Figure 8: Response of RBF classifier trained on all six classes (training set samples shown in black)

5.4.1.4 Conclusions

The NNR is extended in this thesis to allow continuous class membership values. The NNRC technique is applied to a data set containing multiple fault mechanisms. It is then compared with neural modelling techniques. It is shown that the NNRC technique can perform as well as neural modelling techniques without requiring any training or complex parameter optimization.

5.4.1.5 Future work

The NNRC technique applies also to classification problems. A classification problem is a special case of a modelling problem in which the output values are limited to a discrete set of values (the output classes). The specific choice of output encoding may influence the training of the classifier [7]. The advantage of NNRC over NNR in classification problems would be that a degree of class membership can be obtained, as is possible, for example, in fuzzy logic neural networks (FLNN).

Feature selection for modelling problems has been successfully implemented using the gamma test [36]. The NNRC algorithm, as proposed in this section, may be used to implement a modelling problem rapidly based on a reduced feature set, with no lengthy training times and without complex and time-consuming mathematical optimization procedures.

When a model has been implemented, the concept of LOO CV may be applied to the model to determine the success of the model and the smoothness of the output surface, as implemented by the model. When applying LOO CV with a sample size of M , it is usually necessary to train M separate models. When M is large, this results in high computational complexity. By evaluating the distance between nearest neighbours of the model, as in the proposed CVCP and CVMP techniques, it would be possible to determine the degree of overfitting by a model (resulting from an excessive number of free parameters) in modelling problems in order $M\log_2(M)$ operations [97].

The CVMP technique may be applied directly during training of a model or classifier. The objective function to be minimized during training would be the difference between the CVCP or CVMP scores and the optimal Bayes error rate (or NNR score for modelling problems). Large differences would be an indication of either overfitting (excessive number of free parameters) or underfitting (too few parameters available to fit the complexity of the problem).

5.4.2 Developing an efficient cross validation strategy to determine classifier performance (CVCP)

5.4.2.1 Introduction

An efficient cross validation strategy to determine classifier performance (CVCP) is developed in this thesis. Previous techniques for leave one out cross validation (LOO CV) required the evaluation of up to M models, where M is the sample size, making it prohibitively expensive for many practical applications. The CVCP technique is presented in this thesis, requiring only $O(M\log_2(M))$ operations, and the performance of the classifier is shown to be evaluated accurately with this new technique. To combat overfitting of neural classifiers cross validation (CV) is often used to determine the performance on an

independent test set. An extreme form of validation is LOO CV, in which each sample forms an independent test set. In this way very efficient use is made of the data set. Unfortunately, this implies that M models need to be built. Owing to the slow convergence of neural network architectures, it may not be practical to use LOO CV in many applications. An efficient algorithm CVCP is proposed in this thesis to determine CV performance. The resulting error rate compares favourably with the Bayes error rate. CCVP can thus be used to determine classifier performance without the associated computational complexity of LOO CV.

The one nearest neighbour rule (abbreviated as NNR) has been successfully applied in the past to a variety of real-world classification problems [7, 77, 79, 87]. Since the NNR effectively stores or memorizes the training set, the computational resources required may become a problem with large data sets (e.g. in data mining applications). Various computationally efficient alternatives to the NNR have been proposed in the past, e.g. CNN, RNN, ICA, ENN and SNN [79]. The aim is to find the minimally consistent subset that classifies the original training set correctly. Other approaches include learning vector quantization and clustering, in which a vector or cluster centre attempts to represent a number of samples.

The focus of this section is on efficient cross validation strategies. The technique proposed in this thesis is applicable to any classifier and requires $M \log_2(M)$ operations only. Furthermore, no lengthy retraining is required for cross validation.

The cross validation strategy for evaluation of classifier performance (CVCP) technique, based on the well-known NNR algorithm, is proposed in this thesis. It is shown that application of the CVCP algorithm yields results consistent with the Bayes error rate. The CVCP algorithm may also be interpreted as a measure of the output surface complexity of the classifier. As such, it provides a valuable metric to the designer of a classifier.

5.4.2.2 Problem formulation

The performance of any classifier has to be assessed using some form of validation strategy [7, 77]. The training population, consisting of feature values and class labels, is often split into a training, validation and test set (often chosen to be the same size). Neural network training uses the error on a validation set to stop training when the classifier starts to overfit (also called ‘training on the noise’) [7]. This technique has some obvious disadvantages:

- The size of the training set data is decreased.
- The specific choice of training set data can influence classifier results.

Alternative validation strategies include cross validation (CV), jack knifing and bootstrapping. They are all computationally demanding. A particular choice of cross validation, in which one sample only is kept in the test set, is known as leave one out cross validation (LOO CV). This technique requires M repetitions of classifier training, where M is the sample size. When M is large, the computational expense becomes prohibitive.

This prompts research into algorithms for cross validation with lower computational complexity. For certain classifiers, notably linear classifiers, it is possible to evaluate LOO CV with little increase in computational complexity [77]. In most real-world problems and data sets, non-linear relations unfortunately preclude the use of linear classifiers.

The paper by Monari and Dreyfus presents an interesting approach [56]. The influence of removing a single example from the training set is estimated by a Taylor series expansion of the model parameters. The authors introduce a parameter h_{ii} , referred to as the leverage of a training set sample. The closer h_{ii} is to 1, the larger the influence or significance of a training set sample on the model. It may be hypothesized that the greatest influence on the Taylor expansion would be given by the sample (or samples) closest in feature space to the removed training set sample. This would seem to suggest the idea of using the class

label of the prototype closest to the sample to be removed as a representative of the class label of the removed sample.

The CVCP algorithm is proposed in this thesis as an efficient algorithm for cross validation, based on the idea of using the prototype closest to the sample to be removed, as a prototype for classification. No computationally demanding Taylor expansions, as required in the paper of Monari and Dreyfus [56], are necessary. The NNR algorithm, which can be evaluated in $M \log_2(M)$ time, is used instead to determine the closest sample from the removed sample in feature space. The class label of the closest sample to the removed sample is used as an estimation of the actual output of the classifier at the removed sample. The estimated output value of the CVCP algorithm is then compared with the actual class label to give an error e_k . The sum squared error $\sum e_k^2$ may be used as a performance metric. The assumption made is that the training set should be well populated, or equivalently that the average distance between nearest neighbours should be small enough.

The CVCP algorithm determines the smoothness of the output surface as fitted by the classifier. If the class label of the closest prototype (or even, more generally, the output of the classifier) differs frequently from the class label of the prototype itself, it will be difficult to fit a model with low complexity (i.e. a model with few adjustable parameters). To avoid overfitting by models, excessive over-parameterization of models should be avoided [7]. The CVCP model therefore supplies valuable information during the model building process. It may also be used during training of the classifier to limit overfitting of the model.

5.4.2.3 *The nearest neighbour rule and validation or classification*

5.4.2.3.1 *Nearest neighbour rule as zero order hold*

As the final output values of the NNR are always constrained to the range of the output values⁴⁷, the output values of the NNR are discrete whenever the given output values are discrete (corresponding to specific classes).

For the 1NNR (the nearest neighbour rule where $k = 1$) it is possible to define areas or regions corresponding to the points that are closest to the sample x_i (refer, for example, to [87]). When the one nearest neighbour technique is used the decision surface is called the Voronoi tessellation and consists of a number of regions separated by straight lines [87]. The resulting Voronoi tessellation is shown in [87] for a two-dimensional space and the Euclidean distance norm. Each region of the Voronoi tessellation would then have a constant output value corresponding to the output value of the nearest sample x_j to the sample x_i . The output surface corresponding to the Voronoi tessellation⁴⁸ of [87] is shown in figure 9 for a simple two-class problem. The output of the 1NNR corresponds to the output value of the nearest sample to the sample x_i . If the four samples in figure 9 were the only samples in the data set (or the distance to the other neighbours is much larger than the largest distance between the four samples), the actual output values corresponding to the 1NNR would be the inverse of the values in figure 9 (i.e. the data samples in the white region would belong to class 1 and vice versa owing to the 1NNR).

⁴⁷ Let the set of class labels be given by Y_k . (Y_k could also be continuous for modelling problems.) As the NNR always selects the closest neighbour(s), it is clear that the output values of the NNR would also be given by the set Y_k .

⁴⁸ The Voronoi tessellation is the output surface of the NNR.

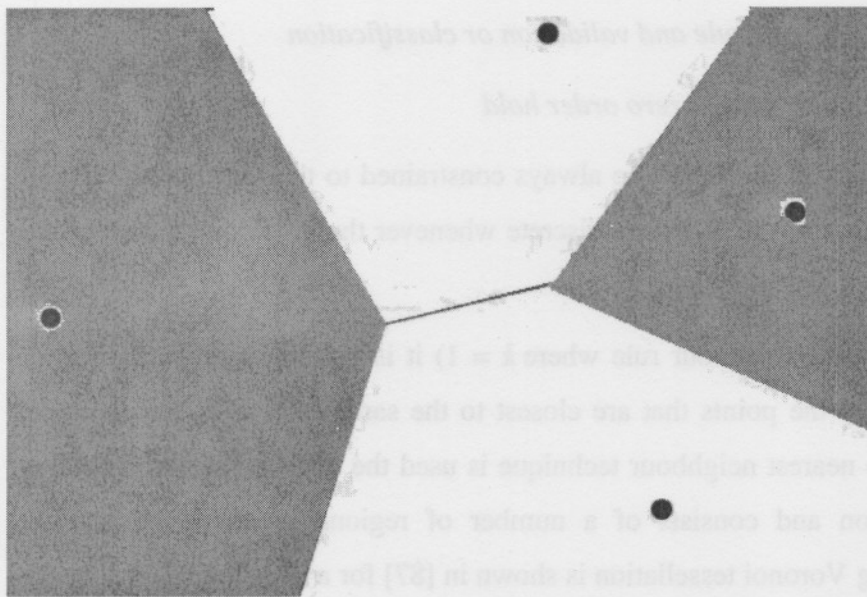


Figure 9: Output surface corresponding to Voronoi tessellation in [87]; the grey values correspond to class 1 and the white values to class 2 (samples are arbitrarily assigned to classes 1 and 2); the data samples are indicated by four black circles

From the preceding discussion it is evident that the output surface of the 1NNR is piecewise constant. An analogy may be considered with the zero order hold in signal processing (time sampling). If the nearest neighbour (spatial sampling) is compared with a delay of one sample (time sampling) and the 1NNR itself with the action of the zero order hold (sampling the input value and keeping it constant for the sample period), the analogy is complete. As it is possible to do linear interpolation between two data samples instead of simply keeping the value constant, the next question arises as to whether a similar possibility could be obtained by a modified NNR. This was examined in section 5.4.1.3.2 on p. 83, where it was shown that the NNR can be modified to use linear interpolation (corresponding to a first order hold).

As the class labels are discrete (with values of either 0 or 1), the output of the normal NNR would also be discrete. Instead of simply using the minimum value of the activation function to compute the output value of the NNR, the output value of the NNR may also

be computed using linear interpolation of the nearest neighbours (or clusters)⁴⁹ using a modified NNR. The modified NNR may also be applied to modelling problems. The resulting output of the modified NNR could have a fractional value between 0 and 1. An output value of $\frac{1}{2}$ on a 24g imbalance class output, for example, could be interpreted as 12g imbalance.

If the NNR is to be extended to higher order interpolation, the minimum distance function of the NNR is clearly no longer suitable (as the output surface is piecewise constant). However, simply using the two nearest neighbours could present a problem with linear interpolation between two data points x_1 and x_2 (assume x_1 and x_2 belong to the training set). Assume, for example, that x_1 represents a 12g imbalance and x_2 a 24g imbalance. If it is assumed that more samples are located close to the data point x_1 and that the distance from the desired interpolated value x to x_1 is much smaller than to x_2 , the nearest neighbours would always be x_1 . The output values for the previous example would be either 12g or 24g, depending on the value of x . Hence no linear interpolation has been performed and no improvement on the 1NNR could be expected.

What is desired, instead, is taking the influence of the sample(s) close to x_2 into account. Thus the requirement of the minimum distance should be replaced by a function that can take into account the distances to all samples in the training set. Samples very far removed from x should still be considered, although heavily penalized. A possibility would be to employ a metric, using the distances d_k to the sample x_k and the class label c_{kc} (assuming that c_{kc} is normalized between 0 and 1 to denote the class membership to class c). This is left for future work.

As the output surface of the NNR is piecewise constant, it is evident that no extreme interpolation is possible (as with an over-parameterized neural network). Hence the good

⁴⁹ The output of a neural network varies continuously (for example between 0 and 1 for a sigmoid transfer function in the output layer), even if the target values (class labels) are discrete. Hence the output of the neural network could be interpreted as a probability of class membership. In an analogous way the modified NNR (with linear interpolation) could be given a probabilistic interpretation.

local behaviour of the NNR is evident and partly explains its success in pattern recognition applications.

5.4.2.3.2 Validation as fractional NNR

In order to determine the performance of a trained classifier, as well as to aid in the training process itself, the sample population may be divided into K subsets of size M/K each (where M is the total number of samples in the population). The training of the classifier then proceeds on the $K - 1$ subsets [77]. An extreme form of cross validation occurs when $K = M$, also referred to as leave one out cross validation (LOO CV). If no duplicate vectors occur in the population, it is clear that the vector values of the validation set may be written as a perturbed training set. Additional data points are located close to the original data set in the perturbed training set

$$x_v = x_t + p_v \quad (48)$$

The perturbation vector p_v would in general be small, otherwise the classifier will have to perform extrapolation (which is a difficult problem). Equation 48 shows that x_v (corresponding to the validation set) corresponds to a perturbation p_v of x_t (corresponding to the training set). The notion of a fractional NNR algorithm, and its relation to cross validation, is subsequently introduced.

The gamma test has been introduced in the context of modelling problems to determine [36]:

- whether a model can in fact be built, before training any classifier
- how complex the model should be
- how many data points are required to build a model

The gamma test has been applied primarily to modelling problems, as a classification problem often does not meet the requirements for application of the gamma test (in particular continuous derivatives of the output vectors). In the gamma test, the average distance between the output and input vectors is computed as a function of k (the number

of nearest neighbours). A straight regression line is often obtained when graphed on a logarithmic plot. The best regression line is then extrapolated to obtain the gamma estimate, which represents the best any classifier can hope to achieve on the data set. The interpretation of the extrapolation of the regression line beyond $k = 1$ in this thesis is that of a fractional NNR algorithm. The extension of the regression line to $k = \frac{1}{2}, \frac{1}{4}$, etc. implies a corresponding increase in the size of the population. However, extension of the regression line to $k = 0$ would imply access to an infinite population. Extending the regression line to values smaller than $k = 1$ implies knowledge about data not yet available, or perhaps knowledge of an extrapolated larger data set based on the current population. In both cases *a priori* assumptions about the distribution of data are made.

To illustrate, assume that the population contains M samples. The NNR corresponds to the choice $k = 1$. Let us divide the population into two subsets of size $M/2$ each. The NNR on the subset of size $M/2$ now corresponds to the 2NNR (second nearest neighbour) of the population with M samples. Yet, the NNR of the population with M samples would correspond to the $\frac{1}{2}$ NNR ($k = \frac{1}{2}$) of the subset with $M/2$ samples.

It is thus clear that the population with M samples represents a perturbation of the subset with $M/2$ samples, as represented by equation 48. It can also be represented as a fractional kNNR algorithm, in which the number of nearest neighbours k is a fractional value. The meaning of the fractional value of k is that $1/k$ additional samples would be generated based on extrapolation from the current population. The sample values should be close to the current population.

If the extrapolated sample values are obtained by linear interpolation of the sample values x_1 and x_2 (with fractional values of k)

$$x_{kNNR} = (x_2 - x_1)kr + x_1 \text{ and } r = 1 : \left(\frac{1}{k} - 1\right) \quad (49)$$

it is possible to evaluate the overfitting of a classifier with increasing accuracy as the value of k is decreased towards zero. The class labels of the interpolated training set vectors may also be interpolated.

Any wild interpolation between the training set would result in a corresponding increase in the fractional kNNR, resulting in overfitting detection of the classifier. If the error rate of the kNNR classifier does not increase as k approaches zero, one would have reason to suspect that the classifier is not overfitting between the training set samples.

It is clear that equation 49 is just a special case of equation 48 in which the perturbed vectors are equidistant from each other and all in the same direction. Other validation strategies, such as bootstrapping and jack kniving, may also be represented in the form of equation 48. Even though the perturbed vectors have no apparent pattern as in the case of equation 49, one may still like to think of the validation strategies, such as cross validation and bootstrapping, as a fractional kNNR approach. They also involve the concept of a local neighbourhood between training set vectors x_1 and x_2 , even though the distance between the interpolated vectors does not need to have a fixed pattern (and could be arbitrarily small, unlike equation 49). If the average distance between the interpolated vectors becomes very small, the number of interpolated vectors should also increase to ensure adequate coverage of the feature space. This is unfortunately often not possible in real world data sets. However, by using equation 49, it is at least possible to verify that the classifier has not learnt extraneous relationships not inherent in the original training set, i.e. that the classifier has not overfitted beyond the original training set. It should be noted that the computational complexity increases proportional to $1/k$.

5.4.2.3.3 NNR measures the complexity of the output surface

Let the set of training set examples be denoted by the set of vectors x_i, y_i . The vectors x_i represent the features and the vectors y_i represent the class labels (or output values). Let the vector c_i represent the closest neighbour to x_i and z_i the class label of the vector c_i . The error rate of the NNR is given by

$$E_{NNR} = \sum_{i=1}^N (z_i - y_i)^2 \quad (50)$$

Equation 50 therefore represents the total number of times the labels of the vector and its nearest neighbour differ. It is assumed that the output vectors y_i are encoded in a suitable

binary format and that each component of the vector is 0 or 1 for a classification problem. For a modelling problem continuous values would be used and equation 50 would be interpreted as the sum square error between the output value of the sample and its nearest neighbour.

5.4.2.3.4 Validation and model verification

As indicated in section 5.3.4 on p. 78 a validation set may be composed in a number of ways. Using the CVCP algorithm may alleviate the problem with computational complexity of LOO XVAL. The CVCP algorithm could be viewed as a selection of a validation set (repeated N times with one sample each time, where N is the number of samples in the training set). The model is evaluated by comparing the output of the model (for example a neural classifier) on sample x_i to the output of the model on the nearest neighbour to sample x_i in the training set. The process is then repeated on all samples in the training set. Let us consider two possibilities:

- If the model has simply memorized the data set, the CVCP algorithm would essentially compute the NNR. As the NNR is known to approximate the Bayes error rate asymptotically, the CVCP algorithm would give a good estimate of the inherent complexity of the data set (and clearly indicate the overfitting of the model). The benefit of using the CVCP algorithm is that no separate test is required and that it can be computed using $O(M \log_2(M))$ computations only. This would seem to make it an ideal performance metric when evaluating the performance of different models. However, it should be noted that the distance metric employed could have a crucial impact on the CVCP algorithm, which is based on the NNR algorithm [87]. This is especially likely in high-dimensional feature space, again highlighting the importance of proper feature selection and extraction in pattern recognition techniques (the aim should always be to minimize the number of neurons in the neural network, while of course retaining enough capacity to model the input–output relationships).
- If the model has not converged at all on the data set, the CVCP algorithm would give a value close to that of the model accuracy (which would obviously not be a good

value). Hence the CVCP algorithm would clearly indicate the discrepancy between the model output and the desired output.

From the above discussion it is clear that the NNR may be viewed as a special case of the CVCP algorithm, in which the model output is given by the training set vectors. If X represents a sample of the training set and Y the corresponding target value (for classification problems it would be the class label), the model output Y_m is given by Y for input of X . It is clear that $Y_m = Y$ represents overfitting of the model. But using the 1NNR a meaningful performance metric may be readily derived, even though the model overfits. The 1NNR uses LOO XVAL to evaluate the nearest neighbour X_n to X . If Y_n is the corresponding target value of sample X_n , the 1NNR uses $E_n = (Y_n - Y)^2$ as the performance metric. (The process is repeated for all samples belonging to the training set and the resulting error determined by summation of the individual errors.) It is clear that any model evaluated with CVCP in which $Y_m = Y$ would give the same performance as the 1NNR. But for good generalization performance the performance of the model on the training set should also be close to that of the CVCP algorithm.

An aspect that is not considered in this thesis is the optimum number of neighbours to be used for the kNNR (only $k = 1$ is considered for the nearest neighbour in this thesis). As the value of k is increased a larger area of the feature space is considered. Consider two samples of class 1 that are close together in feature space, so the nearest neighbour belongs to class 1. If these samples were also in the feature space of class 2, the 1NNR would always assign them to class 1 (the nearest neighbour). However, the 3NNR would select class 2 if majority voting is implemented. In general, consider M samples of class 1 that are close together in feature space, so the $M - 1$ nearest neighbour selects class 1. If these samples were also in the feature space of class 2, they would be assigned to class 1 by the kNNR, where $k = M - 1$, and to class 2 by the mNNR (where $m = M + 1$). Hence the membership of the samples of class 1 would be determined by the choice of parameter k of the kNNR. Further exploration is left for future work. The NNR could be evaluated more efficiently [14] (overcoming to a certain extent the limitation of slow

classification of the NNR). This would be especially important in a feature set of high dimension.

5.4.2.4 Results

The performance of the CVCP algorithm will be shown on the well-known Iris data set [91]. Two of the features, the petal length and width, have been selected to facilitate a graphical representation. Before training a classifier, one would inspect features individually, then two features pairwise with the aid of scatterplots, to establish the degree of intra-class and inter-class scatter. Figure 10 shows a scatterplot of the petal length and width with normalized features.

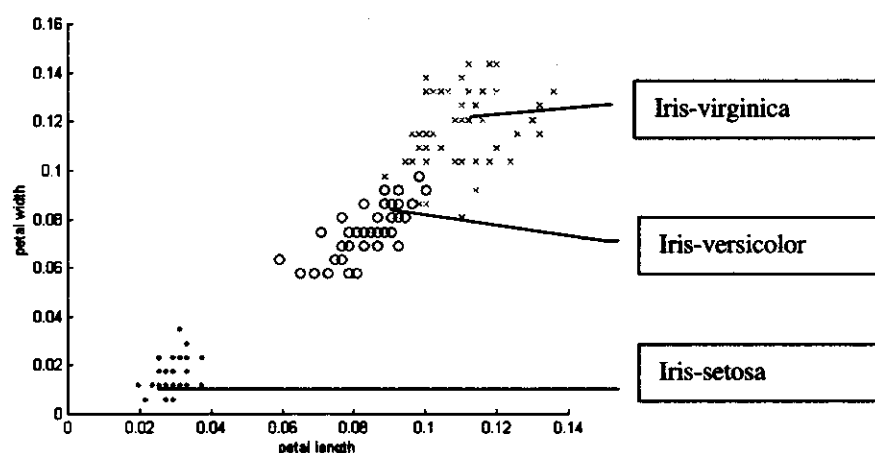


Figure 10: Scatterplot of normalized features

From figure 10 it is evident that only a few samples from different classes are close in feature space and that a high accuracy on the training set can be expected. A validation set, commonly used to combat overfitting [76], was not used for these experiments – as the aim of the experiments is to show overfitting of the neural network⁵⁰. A neural network with a single hidden layer of 10 neurons has been trained using the Levenberg-Marquardt training algorithm of Matlab [61]. The number of training epochs is limited to 800. The

⁵⁰ The likely result of the use of a validation set would be similar to an 'early stopping' training algorithm. As the training is stopped prematurely, similar results to a network with fewer parameters (neurons in the hidden layer) could be anticipated. The disadvantage of this training strategy would be that the full non-linear capabilities of the neural network cannot be fully realized due to early termination of the training algorithm.

relatively high number of neurons permits one to observe the effect of overfitting in figure 11. The output of the neural network has been evaluated on a two-dimensional grid to visualize the effects of overfitting. Owing to the curse of dimensionality such a strategy would not be feasible using a large number of features. It is clear that the grid represents a set of perturbed vectors corresponding to the training set of figure 10 and equation 48.

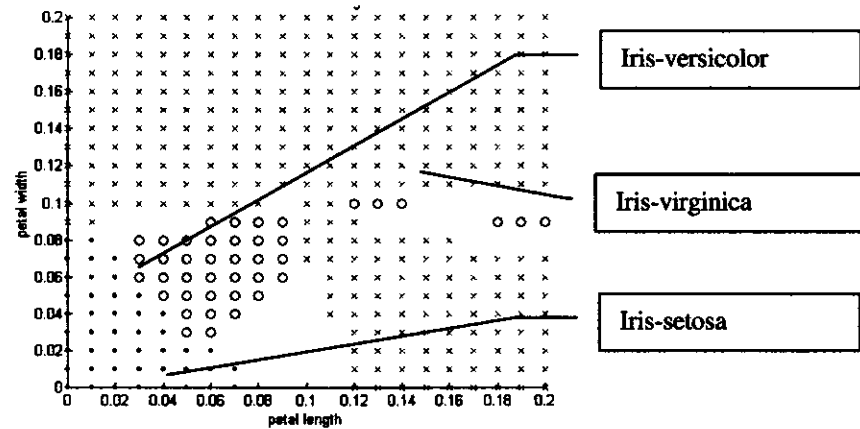


Figure 11: Response of neural network evaluated on grid

From figure 11 it is evident that the neural network has learnt (extrapolated) extraneous relationships owing to the extra parameters (neurons) available. This notion is confirmed by the application of the CVCP algorithm to the trained model. The result of $96.8\% \pm 0.85\%$ by the CVCP algorithm on the neural network (using the training data set), which is much lower than the 98% of the NNR (no more than twice the optimal Bayes error rate) and the 99% of the trained neural network on the training data set (ignoring the contribution of the four training sessions, which failed to converge and can hence be considered as outliers) suggests that the class labels of a number of training set vectors differ. The neural network has therefore overfitted between the training set vectors, instead of fitting a smooth output surface. For good performance the CVCP and NNR should almost be the same. While the difference of 2% between the CVCP and NNR would not in general be considered to be a large difference, it is in the context of the Iris data set in which only three of the samples (representing 2%) cause problems as far as overlap in feature space is concerned.

Figure 12 shows the response of the NNR evaluated on a grid. The output surface consists of three regions (for the three classes) separated by straight lines. Three samples of class Iris-versicolor are assigned to Iris-virginica instead by the NNR (including an overlapping sample), representing an error of $3/150 = 2\%$.

Figure 13 shows the response of the ANN with only two neurons evaluated on a grid. The output surface also consists of three regions (for the three classes) separated by straight lines (almost identical to the NNR of figure 12). Three samples of class Iris-versicolor are assigned to Iris-virginica instead by the ANN (including an overlapping sample), representing an error of $3/150 = 2\%$ on the training set.

Figure 14 shows the response of the ANN with 30 neurons evaluated on a grid. The output surface consists of at least seven regions (for the three classes). Two overlapping samples of class Iris-versicolor are assigned to Iris-virginica instead by the ANN (including an overlapping sample), representing an error of $2/150 = 1.3\%$ on the training set. However, the output surface is much more complex (by visual examination).

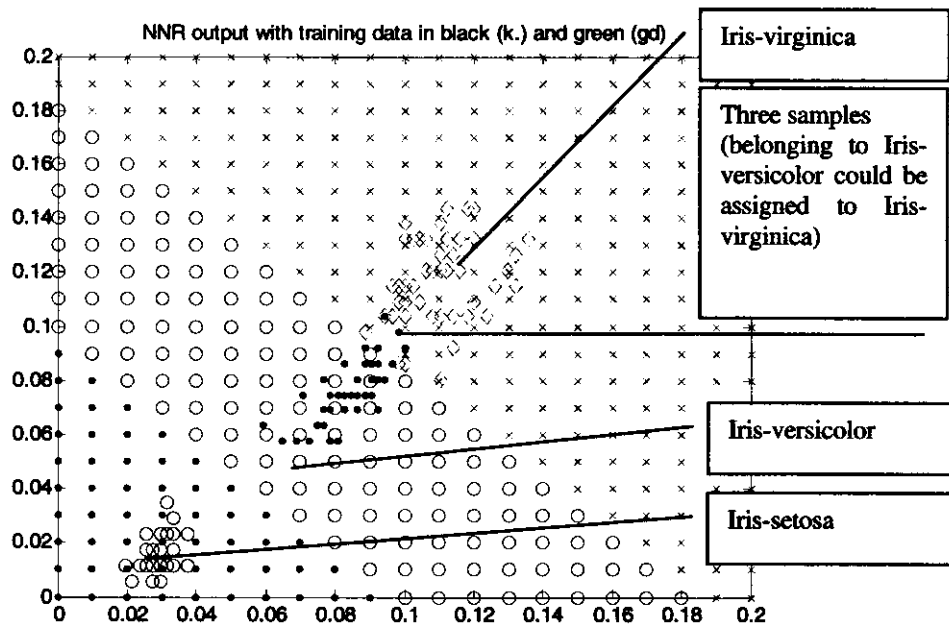


Figure 12: Response of NNR evaluated on grid (NNR output for Iris-setosa indicated by a red triangle, Iris-versicolor by a blue circle and Iris-virginica by a magenta cross; training set data of the NNR is shown by a black circle, black dot and green diamond respectively)

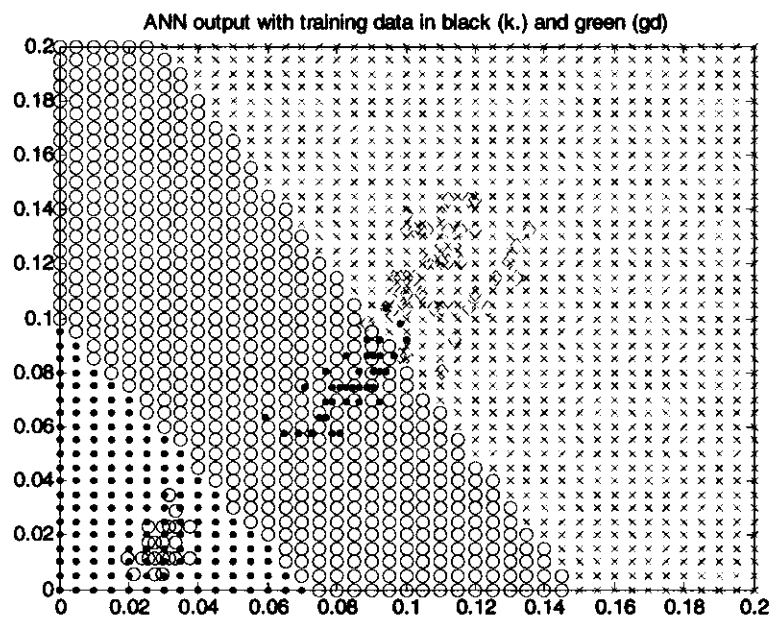


Figure 13: Response of neural network evaluated on grid with two neurons in single hidden layer (trainlm, 1600 epochs)

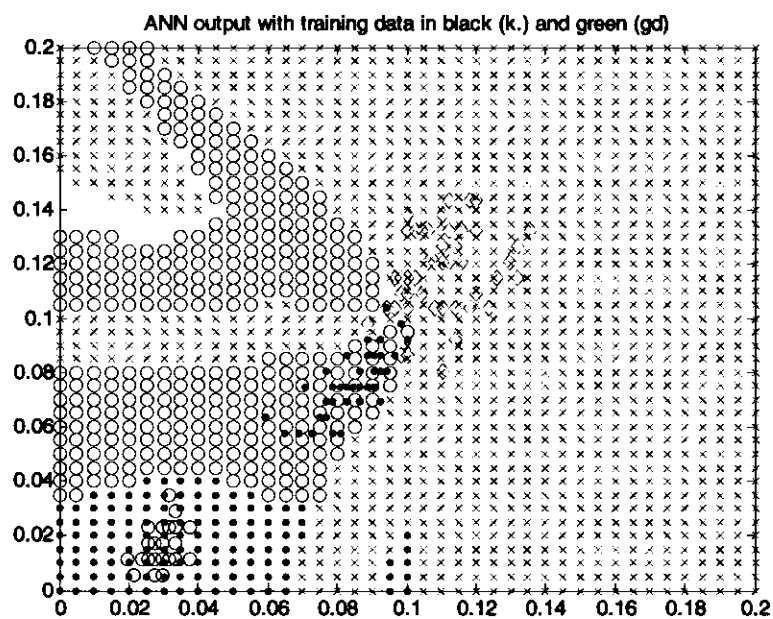


Figure 14: Response of neural network evaluated on grid with 30 neurons in single hidden layer (trainlm, 1600 epochs)

In summary, it can be seen that the NNR and ANN with two neurons have comparable output surface complexity, whereas the ANN with 30 neurons has a much greater output surface complexity. The overlapping sample belonging to classes *Iris-versicolor* and *Iris-virginica* (very close together in feature space) should not be artificially separated by a classifier using a complex decision surface. The CVCP evaluated on the training data set would provide results approaching the training set of the NNR for the ANN with two and 30 neurons. This is immediately evident for the ANN with 30 neurons, as the model overfits and obtains an accuracy close to 100% on the training set. Hence the output of the ANN is very close to the true class labels. The NNR (of the class labels) should be very close to the NNR (of the ANN output with 30 neurons). Consequently, the CVCP should be close to that of the NNR. For the ANN with two neurons the output surface of the ANN is close to that of the NNR, thus the CVCP and NNR may be expected to be close.

If the CVCP is evaluated on the fine grid itself (which would of course require expanding the training set considerably), a measure of the output surface complexity is possible. Figure 13 would supply a CVCP of $14 + 19 = 33$ and figure 14 gives a CVCP of about $55 + 8 + 38 + 18 = 119$. Consequently, it can be seen that over-parameterization supplies a much larger CVCP score (where class labels are different) on the grid. In practice it would unfortunately be difficult to populate a high dimensional feature space with a fine grid owing to the 'curse of dimensionality'⁵¹.

The question arises as to what the lowest possible CVCP achievable on the grid is. It could be hypothesized that the lowest CVCP would be obtained by making the fewest assumptions about the output of the model far from the training set. The NNR of figure 12 makes assumptions about the class membership of samples very far from the actual training set. A modification to the NNR could be proposed, in which the output of the NNR is 'uncertain' when the sample to be classified is 'far' from the training set. The concept of 'far' could be quantified in a number of ways, but a simple way would be to use the rule: output of the NNR is class 'uncertain' if $X_m < \min(X_m)$ or $X_m > \max(X_m)$,

⁵¹ A good starting point would be the output of the NNR on the grid, assuming that the grid is 'close enough' to the training data set.

and X_m is feature m of variable X . Another possibility to quantify the concept of 'far' is if the distance between X_m and the nearest neighbour is large relative to the average distance between samples in the training set. This aspect is left for future work.

Figure 15 shows the accuracy when LOO CV is applied to the complete data set of 150 samples. Apart from the four samples failing to converge, the mean average accuracy obtained was $99.11\% \pm 0.69\%$ (only accuracies greater than 90% were considered for computation). Each sample was removed from the training data set, a neural network was then trained and the performance stored. The process was then repeated for all 150 samples and the average value obtained (ignoring the samples in which the neural network failed to converge).

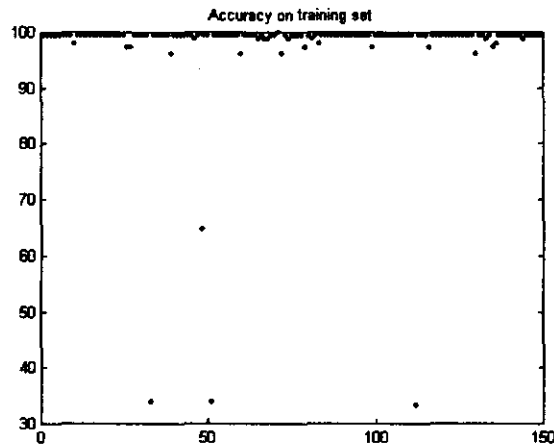


Figure 15: Accuracy of LOO CV on training set

Figure 16 shows the accuracy of the CVCP method when applied using LOO CV. The shape of the graph bears a striking resemblance to that of figure 15. However, the graph is much lower than that of figure 15, which suggests that the output surface of the neural network is overly complex. In order to generalize well, the number of free parameters in the classifier should therefore be adjusted to bring the results of the classification techniques of CVCP (using ANN as a modelling technique) and NNR closer together.

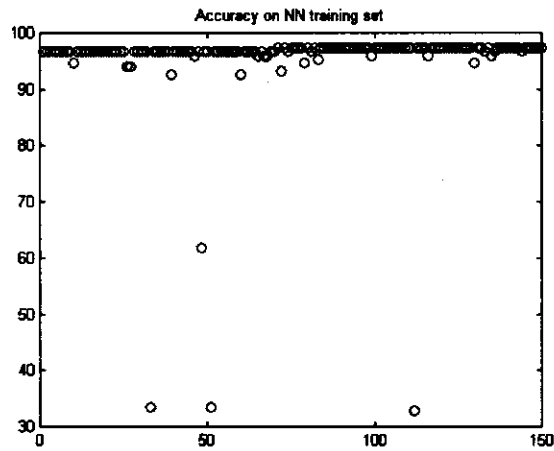


Figure 16: Accuracy of CVCP when applying LOO CV

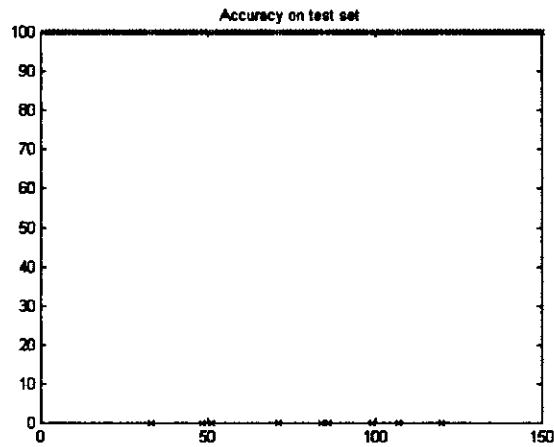


Figure 17: Accuracy obtained on test set for ANN

Finally, figure 17 shows the accuracy obtained on the single test sample. The prediction on the test set fails for a number of samples, including of course the samples in figure 16 in which the neural network failed to converge. The mean accuracy on the test set was 94% (9 test samples failed, 4 samples failed as a result of convergence problems in training the neural network).

Pattern recognition techniques have been applied to the Vib demo data set 2⁵². An RBF network has been trained on the normal bearing data as indicated in figure 19. The outputs (0g, 12g and 24g) of the RBF network are shown in figure 18. The testing data from the defective bearing are shown in figure 20. It is clear that the 24g output of the RBF on the 24g bearing defect data is incorrect. Assuming that the threshold value of the RBF outputs is set at 0.5, the accuracy of the CVCP is $100 - 111 / (94 + 98 + 93 + 101 + 99 + 111) * 100 = 81\%$.

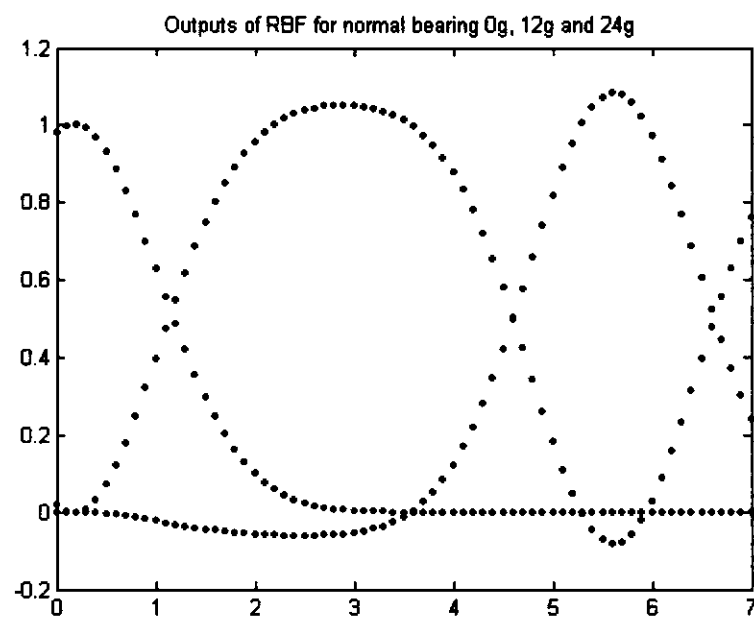


Figure 18: Outputs of RBF for normal bearing 0g, 12g and 24g (RF from 0 to 7)

⁵² A more comprehensive training set is required for a meaningful application of pattern recognition techniques and is left for future work.

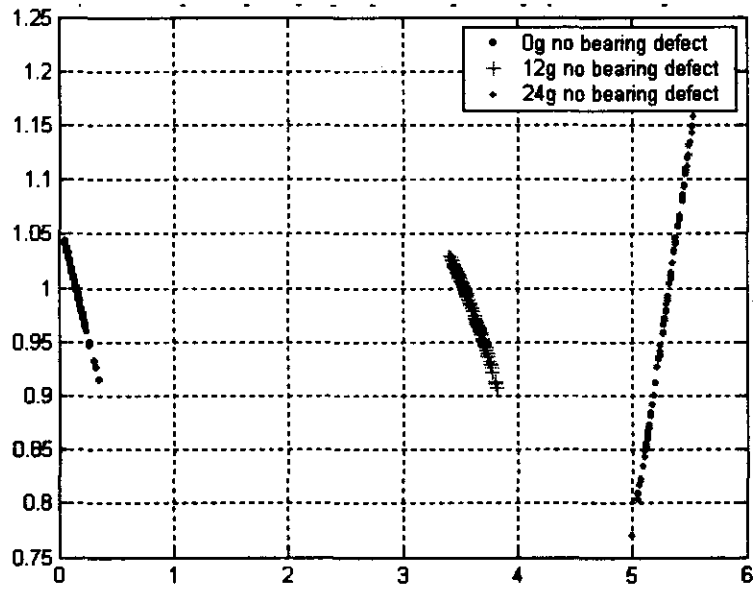


Figure 19: Output of RBF neural network on normal bearing data for 0g (●), 12g (+) and 24g (◇) imbalance respectively

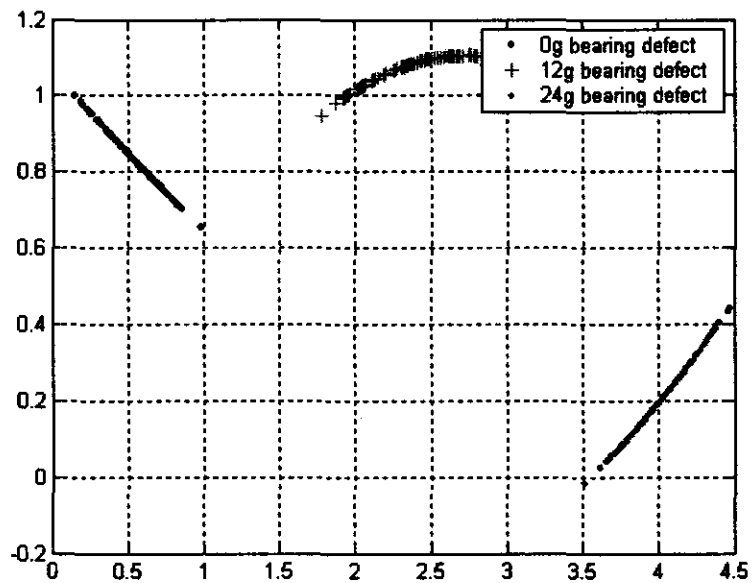


Figure 20: Output of RBF neural network on defective bearing data for 0g (●), 12g (+) and 24g (◇) imbalance respectively

5.4.3 Conclusions

In order to generalize well, the performance of a classifier has to be tested on an independent test set. Cross validation techniques require many models to be evaluated on

separate test sets, thereby greatly increasing the computational complexity. A new technique CVCP, which can be used to determine classifier performance with $O(M \log_2(M))$ complexity only, is presented in this thesis. By using LOO CV or NNRC the Bayes error rate may be determined using classifiers with a large number of basis functions (NNR and NNRC).

5.5 Future work

The incorporation of the CVCP technique into the design and training of classifiers would be a computationally efficient way to combine the advantages of the NNR technique with that of traditional classifiers, to combat overfitting and to apply regularization. The output surface complexity map may give a good indication of the complexity of the model and may be compared with the expected complexity (e.g. an NNR approach) for validation of the model complexity. The fractional NNR algorithm may be used to probe the local neighbourhood properties of various models. The CVCP algorithm may be further refined by taking into account the output vectors (class labels) of the closest neighbours and then interpolating the value of the removed sample from these values. The result of the interpolation process may be a continuous value (instead of a discrete class label) and may be given a probabilistic interpretation.

The CVCP technique currently provides an indication only of the amount of overfitting, but not any direct guidelines on the choice of architecture and the number of free parameters (type of network, layers, neurons) to be used. As an upper bound the number of misclassified samples of the NNR may be used to initialize hyperplanes (e.g. multilayer perceptron) or basis functions (e.g. radial basis function neural network). However, increasing the number of nearest neighbours k should improve the generalization performance until the neighbours of a different class are encountered with increasing values of k (larger neighbourhood size).

The NNRC technique will be investigated in more detail in the context of other standard pattern recognition techniques, for example mutual information based feature selection (MIFS), decision trees, self-organizing maps (SOM) and vector quantization (VQ).

5.6 Conclusion

The application of various novel pattern recognition techniques to data sets (the detection of multiple fault conditions and the Iris data set) was investigated.

In order to generalize well, the performance of a classifier has to be tested on an independent test set. Cross validation techniques require many models to be evaluated on separate test sets, thereby greatly increasing the computational complexity. A new technique CVCP, which can be used to determine classifier performance with $O(M \log_2(M))$ complexity only, is presented in this thesis. It is shown that classifiers with a large number of basis functions (NNR and NNRC) may be used to determine the Bayes error rate, provided that LOO CV or NNRC is applied.

In section 5.4.1 on p. 80 the NNR was extended to allow continuous class membership values. The NNRC technique was applied to a data set containing multiple fault mechanisms and was compared with neural modelling techniques. It was shown that the NNRC technique can perform as well as neural modelling techniques on this data set without requiring any training or complex parameter optimization that neural networks, for example, would require. Provided that both the RF and BPFO features are available, the NNRC technique can achieve perfect (100%) separation of all the 6 classes.

In section 5.4.2 on p. 91 a technique called CVCP was presented and applied to the Iris data set, which determines classifier performance with $O(M \log_2(M))$ complexity only. Training M neural networks, by way of contrast, would pose a formidable parameter optimization problem (with M samples, K features, L neurons in a single hidden layer, E epochs and a single output neuron the complexity⁵³ would be in the order of $M * K * L * E$ ⁵⁴. If the rule of thumb of $K * L = M/10$ is employed for good

⁵³ The complexity refers to the number of summations and multiplications required. Evaluation of the transfer function would also pose a problem, for example $M * L * E$ evaluations of a sigmoid (exponential function) would be required. Owing to the assumption of a single output neuron the computational complexity associated with the hidden layer to output layer as well as the backpropagation has been disregarded.

⁵⁴ The biases of the neurons are ignored in the computation.

generalization performance [18], the complexity is $O(M^2E)$. As $E > 1$ the complexity of the ANN would always be worse than the CVCP technique⁵⁵.

⁵⁵ Unless special techniques for efficient evaluation of LOO CV (such as proposed in [56]) are used, in which case a new comparison would have to be made.

6 Results of signal processing techniques

6.1 Introduction

The application of various signal processing techniques to the detection of multiple fault conditions is illustrated in this chapter. Artificial data sets, in which the parameters can be easily adjusted, are also used in conjunction with the experimental data sets to evaluate the signal processing techniques objectively.

The outline of this chapter is as follows. The problems involved in obtaining a normal vibration signal from a defect signal are outlined in section 6.2 on p. 116. The relation between synchronous averaging and demodulation spectrum is shown in section 6.3 on p. 117. The relation between the SA on the time and frequency domains, using the ESFFT, is described in section 6.4 on p. 123. The influence of the time window on the SA is shown in section 6.5. In sections 6.6 and 6.7 the influence of speed variations is discussed in the context of modelling. In section 6.8 the relation between the SA and demodulation spectrum is examined. Section 6.9 highlights various aspects related to the kurtosis statistic for fault diagnosis (including minimization and maximization of the kurtosis statistic). The theoretical effect of non-linear operators, such as squaring, that are used in HFRT is discussed in section 6.10. The influence of noise and speed variation (also a form of noise) is analysed. The detection of bearing defects requires the detection of periodicities in the signal, especially under low signal-to-noise ratios. The relation between the absolute value of the analytical signal derived from the Hilbert transform (the amplitude modulation signal) and the square law detector is examined in section 6.11. Novel cepstrum techniques, described in section 6.12, can be useful in the detection of multiple fault conditions. Finally, the influence of speed variations (which can be regarded as a source of noise) is investigated and a speed compensating algorithm developed in section 6.13. If a trigger signal is available, it could be used to perform speed compensation. Unfortunately, a trigger signal is frequently not available in a plant, hence

the speed compensation technique can be very useful in removing the noise associated with speed variations.

6.2 Problematic nature of separating the normal and defect vibration signals

In this section the problem of obtaining a ‘normal’ vibration signal from measurements of a defective bearing is considered. This question is especially relevant in the Vib demo data set 2 in which the measurements on the normal bearing are contaminated with the defect signal from the defective bearing.

The cepstrum can possibly be used to remove the effect of the repetitive bearing defect impact signal, which should then supply the normal signal. As the normal bearing signal has been measured on a bearing other than the defective bearing, it is not easy to develop a suitable model of the defect itself [28]. The cepstrum is known to supply peaks at the defect quefrency for large bearing defects [21]. Using the cepstrum it should therefore be possible to obtain a normal signal and a defect signal (various degrees of defect size) from a single model (actually the power spectrum, not the raw time signals).

The effect of noise (broadband and discrete frequency as in a gearbox signal) can then be investigated. However, the initial set of vibration signals measured in [28] (Vib demo data sets 1 and 2) does not have a tachometer signal, which could be important if timing variations are to be investigated. Nonetheless, a trigger signal could be obtained from the imbalance measurements (using bandpass filtering near the defect frequency).

A technique (perhaps similar to the ANC and SANC techniques of Randall [72]) could be developed to remove certain repetitive components from the raw time domain signal. Unfortunately, the components from the time domain signal cannot be removed easily using the demodulated spectrum or the real cepstrum (cepstral transformations are not invertible transformations and they destroy the phase information).

One possibility might be to evaluate the power spectrum $D(\omega)$ of the defective bearing and then estimate the normal power spectrum $N(\omega)$ by using the previous technique (i.e. using the cepstrum to remove defect frequencies and harmonics or sidebands). One could then evaluate $N(\omega) = |X(\omega)|^2$, where $|X(\omega)|$ represents the amplitude spectrum. The time

domain signal $n(t)$ may then be estimated by using the inverse FFT of $|X(\omega)|\angle(X(\omega))$. The accuracy obviously depends on whether $\angle(X(\omega))$ is close to $\angle(N(\omega))$ and whether $N(\omega)$ is representative of the normal condition (i.e. whether the cepstrum and liftering⁵⁶ removed the defect signal).

The same technique could also be applied in principle to remove gear defect signal(s), allowing the bearing defect signal to be analysed.

Another possibility would be to apply load demodulation normalization as described in [80] to remove the effect of load on the vibration signals. Unfortunately, the technique is applied in the context of gear fault detection, in which a clear and large amplitude modulated signal is available owing to the meshing of the gears. By contrast, a bearing fault supplies a large defect signal only when the roller contacts the defect. Since load normalization of the signal would increase the amplitude of the noise signal rather than that of the defect signal, these techniques have not been considered in this thesis.

6.3 Synchronous averaging and demodulation spectrum

McFadden *et al.* show [50, 53] that the operation of synchronous averaging in the time domain corresponds to a comb filter in the frequency domain (as a special case the trigger frequency F_t of the SA can also be the rotational frequency F_r , so $F_t = F_r$). The width of the comb filter narrows as N , where N is the number of averages, increases, reaching an impulse in the limit as the number of averages tends to infinity.

It is easy to understand the reduction of noise as a function of the number of averages N in the frequency domain. Assume that white noise is added to a periodic signal of period T , giving additive noise, and that the averaging period is T . The comb filter keeps all frequency components that are a multiple of the trigger frequency $F_t = 1/T$ and removes all frequency components (hence noise components) that are not a multiple of F_t . As the width of the teeth in the comb filter decreases, the amount of noise excluded will also

⁵⁶ The term 'liftering' is used to describe a filtering operation in the cepstral domain (another example would be the reversal of the first characters in spectrum to give cepstrum).

increase. Taylor [86] shows that the signal-to-noise ratio $(\text{SNR}) = \sqrt{N}$, where N is the number of averages. Thus to obtain a tenfold increase in the SNR, $N = 100$ averages should be taken. Any frequency component in the signal that is not a multiple of the trigger frequency F_t will also be removed. Consequently, only synchronous components in the signal are kept. If a sufficient number of averages N are taken, the non-synchronous components will be removed.

The effect of synchronous averaging is perhaps not that easy to visualize in the time domain. However, assume that the noise signal is decomposed into its Fourier components (theoretically all of the same amplitude). Now consider the frequency component with frequency equal to the trigger frequency F_t . Application of the synchronous average to this frequency component will effectively 'freeze' this component⁵⁷, hence passing the frequency component as previously shown. The same will happen using frequency components at multiples kF_t of the trigger frequency F_t . Therefore, the operation of synchronous averaging $\text{SA}(k, F_t)$ may be visualized as a stroboscope at trigger frequency F_t pulsing at a rate of F_t pulses per second and averaging all the amplitudes of the signal (or noise) at a specific value of k (where k represents an advance in the phase of the stroboscope). Just as the stroboscope will freeze mechanical vibrations synchronous to the trigger frequency F_t , the synchronous average will freeze components in the signal (or noise) synchronous with the frequency F_t . This may also help to explain the popularity of the stroboscope in vibration monitoring applications, as the eye effectively performs a synchronous averaging when using the stroboscope. If the width of the pulses radiated by the stroboscope is small enough, then the stroboscope provides the synchronous average for a single value of k . By changing the trigger frequency slightly, it is possible to obtain any value of k (representing the phase). In practice, the limitation of a single value of k in the stroboscope is not a problem, as the phase can be readily adjusted by varying the trigger frequency slightly. It stands to reason that very small vibration levels would be difficult to detect with the naked eye, hence improved optical methods of vibration

⁵⁷ Again assuming the number of averages N is large enough.

detection are being developed. Better results with the stroboscope could be obtained if an external trigger signal were available for the stroboscope.

Next, consider frequency components with frequency F_c different from F_t . In this case beat notes proportional to the difference in frequency $F_c - F_t$ may be seen in the time domain signal. Thus the average value will be zero and the synchronous average will be zero, provided of course that enough averages N are taken. A window, such as the Hanning or Blackman-Harris, may also be applied if the rectangular window is not suitable. In general, the frequency components that are analysed with the FFT should be a multiple of the frequency resolution. The presence of strong sinusoidal components can mask a low level sinusoidal signal owing to the high sidelobe level. The application of a window function has the following disadvantages:

- The frequency resolution is decreased (implying some of the information in the time signal is not used).
- The signal is multiplied by a window signal with very small values at the endpoints. Hence only part of the signal (generally in the middle of the buffer) is used. This helps to explain the poorer frequency resolution.

The application of advanced windows functions generally increases the main lobe width but reduces the sidelobe level. No window function has been applied prior to the FFT function in this thesis, unless explicitly mentioned.

To summarize, decomposing the noise signal into its Fourier components shows that only the noise components synchronous to the trigger frequency F_t will be passed, whereas all other components will be removed. If the signal is periodic with period $= 1/F_t$, then all the signal components will be passed by the synchronous average, while most of the noise components will be blocked.

Any speed variation in the signal will make the signal non-periodic. Consequently, the frequency components of the signal will no longer be spaced exactly at multiples of a defect frequency. Spectral smearing will result if the buffer is long enough. However, applying the synchronous average at trigger frequency F_t will still freeze the frequency

components that are synchronous to the trigger frequency F_t . The components that are not synchronous to the trigger frequency F_t (resulting from speed variations) will be reduced in amplitude by the operation of synchronous averaging, depending on the width of the teeth in the comb filter (hence on the number of averages N). So any speed variation around the trigger frequency F_t will be reduced in amplitude by the operation of synchronous averaging.

The effect of synchronous averaging is generally derived in the frequency domain, but it stands to reason that the process is reversible. Consequently, a comb filter in the frequency domain will correspond to a synchronous average, provided that enough averages N are used (for this reason, the width of the teeth should decrease as the buffer length increases). The 'picket fence' effect occurs when a signal contains sinusoidal components that are not an exact multiple of the sampling frequency. This effect predicts that sinusoidal frequency components that are not a multiple of the basic frequency resolution of the FFT will not be removed⁵⁸. It is well known that each frequency component of the FFT corresponds to a bandpass filter. The frequency resolution increases with an increase in buffer length. Next, consider sampling the FFT at regular intervals of kF_d (where F_d is corresponding to the defect frequency) in the frequency domain (after applying the bandpass filter and non-linear operator such as squaring⁵⁹) and call the resulting transform the ESFFT(F_d, n) (equi-sampled FFT). Hence it can be stated that the ESFFT will correspond to a synchronous averaging of the signal. Theoretically only frequency components that are multiples of the basic defect frequency F_d (corresponding to the trigger frequency F_t of the synchronous average) will be kept. Owing to finite buffer lengths (non-ideal filtering) some of the adjacent frequency components, as a result of speed variations, will also be kept – though at reduced level.

⁵⁸ A suitable time domain window other than the standard rectangular window of the FFT can reduce (but not eliminate) the amplitude of the sidelobes.

⁵⁹ The bandpass filter and non-linear operator are applied, as the high frequency components of the vibration signal are believed to indicate bearing faults more reliably (see the literature survey in chapter 2 from p. 11 as well as the results on feature extraction in chapter 7). The low frequency vibration components are frequently corrupted by noise, making fault diagnosis unreliable.

If the frequency resolution is much smaller than the variation F_v in defect frequency F_d owing to speed variations, then it will be possible to observe the effect of frequency components $F_d + F_v$ in the ESFFT($F_d + F_v, n$) or the demodulation spectrum itself. Thus the peak in the demodulation spectrum will not be a single well-isolated peak in most cases, but rather the width of the peak should give an indication as to the amount of speed variation present. It should be noted that the non-linear operator (more specifically the square of the signal⁶⁰) itself clearly reduces the effect of speed variation (this will be discussed at length in section 6.4 on p. 123). If no speed variation or noise is present (the bearing defect signal is perfectly periodic) then the non-linear operator will have no beneficial aspect. Yet, if speed variation and/or noise is present, the non-linear operator can help to reduce the effect of noise (this will be discussed in more detail in section 6.4 on p. 123).

The technique of synchronous averaging can therefore be efficiently computed in the frequency domain, and the trigger frequency F_t can be varied simply by changing the trigger frequency F_t of the ESFFT. Owing to the reduced number of samples of the ESFFT, computing the IFFT (inverse FFT) is computationally very efficient. The ESFFT is a novel approach to explain the successful results already achieved with the demodulation spectrum (HFRT), relating the concepts of synchronous averaging and demodulation spectrum. In particular, the demodulation spectrum actually shows the frequency content of a number of synchronous averages. If only the first component in the ESFFT is of interest, examination of the (low) frequency content of the demodulation spectrum already provides valuable information on repetitive frequency components. In general, the additional components of the ESFFT provide even more valuable information for diagnostic decisions (for example two outer race bearing defects spaced 180° apart would theoretically have a first ESFFT component of zero, but would have a second ESFFT component of maximum value [50, 53]).

⁶⁰ The non-linear operator can be chosen from any function that would isolate the 'envelope' of the time domain signal (equivalently supply a convolution in the frequency domain). Suitable examples would be the Hilbert transform, absolute value and signum function and higher powers of the signal.

Hence computing the ESFFT of the signal supplies the frequency components related to the synchronous average of the signal at trigger frequency $F_t = F_d$ (F_d = defect frequency). The first component of the ESFFT, corresponding to the first demodulation peak at the defect frequency F_d of the demodulation spectrum, corresponds to the low frequency detail of the synchronous average. As more components in the ESFFT are added, higher frequency detail of the synchronous average becomes evident. When all components in the ESFFT are added, the IFFT of the ESFFT will supply the synchronous average of the signal at the trigger frequency F_t . For the demodulation spectrum it is necessary, of course, to apply bandpass filtering and a non-linear operator (for example squaring) prior to the computation of the synchronous average. Consequently, it is important generally to consider all the components of the ESFFT together, rather than a single component (the first component of the ESFFT corresponding to the first demodulation peak). This is especially true if the bearing surface is composed of a number of widely distributed defects, as is characteristic of a bearing defect that is well developed. Using a single frequency component in this case is bound to be misleading (50, 53). Hence all components of the ESFFT should be considered in the final diagnosis decision. A distributed bearing defect, such as geometrical imperfections, will be distributed over many frequency components in the ESFFT. Therefore, it is important to consider all the components of the ESFFT together when making a diagnostic decision. Considering a single demodulation peak (first component of the ESFFT) is sufficient for a single, well localized bearing defect).

The influence of bandpass filtering prior to application of the synchronous average may be shown as follows. Bandpass filtering of the signal will of course remove some of the low frequency and high frequency components. Thus the ESFFT, obtained by equi-sampling of the resulting FFT, would also show the removal of some components⁶¹. Furthermore, the effect of a bandpass filter on the input signal is to change the frequency components participating in the FFT (and ESFFT). The computation of the synchronous average is also

⁶¹ The non-linear operator, such as squaring, would convolve the high frequency spectral information resulting in low frequency components.

correspondingly changed in the time domain, by convolving the input signal with the impulse response of the bandpass filter prior to application of the synchronous average (and non-linear operator).

It should be noted that bandpass filtering could also be performed using the FFT, by keeping the desired frequency components only (of course prior to application of the non-linear operator or synchronous average). The demodulation spectrum can then be obtained either by convolution in the frequency domain (assuming a squaring operation), or by taking the IFFT and applying the non-linear operator. The ESFFT can then be obtained from the resulting demodulated spectrum. The resulting computational complexity would be dominated by the IFFT if the time domain approach were followed (hence a $n \log_2 n$ complexity for each bandpass filter⁶²).

Finally, note that the ESFFT inherits all the desirable properties of the FFT itself.

6.4 Relation between SA on the time and frequency domains (using ESFFT)

The novel ESFFT algorithm has been introduced in the previous section. The ESFFT is a comb filter, as described in section 6.3. This section investigates this aspect in more detail in addition to the relation between the SA computed in the time domain and the ESFFT computed in the frequency domain.

In [50, 53] McFadden *et al.* derive an expression of the frequency domain response of the SA operator in the time domain. If $c(r)$ is defined as a train of N impulses with weight $1/N$ [53]

$$c(t) = \frac{1}{N} \sum_{n=0}^{N-1} \delta(t + nT_i) \quad (51)$$

where $T_i = 1/F_i$ is the defect period (F_i is the trigger frequency) and N is the number of synchronous averages.

⁶² As a result of the many zero or small frequency components present, it might be possible to improve on the $n \log_2 n$ complexity of the FFT.

In [68] it is proved that the spectrum $C(f)$ of $c(t)$ is

$$\begin{aligned} C(f) &= \frac{1}{N} \sum_{n=0}^{N-1} e^{j2\pi nTf} \\ &= \frac{1}{N} \frac{\sin(\pi NTf)}{\sin(\pi Tf)} \end{aligned} \quad (52)$$

The SA of $y(t)$ is given by the convolution between $c(t)$ and the signal $x(t)$. In the frequency domain the equivalent expression is given by a multiplication of the corresponding spectrum

$$\begin{aligned} Y(f) &= C(f)X(f) \\ &= \frac{1}{N} \frac{\sin(\pi NTf)}{\sin(\pi Tf)} X(f) \end{aligned} \quad (53)$$

In [92] the response of a DFT filter bank is derived in the Z domain. The Z transform $X(z)$ of the signal $x(n)$ is

$$X_k(z) = \sum_{i=0}^{M-1} (zW^k)^{-i} X(z) \quad (54)$$

Hence $X_k(z) = H_k(z)X(z)$ with [92]

$$\begin{aligned} H_k(z) &= H_0(zW^k) \\ H_0(z) &= 1 + z^{-1} + \dots + z^{-(M-1)} \end{aligned} \quad (55)$$

The term $H_k(z)$ above represents the Z transform of the prototype filter H_0 modulated by W^k , whereas $X_k(z)$ is the Z transform of the signal after application of the filter $H_k(z)$. The frequency response of the above equation is shown on p. 115 of [92] and is a sinc function with zeros at $k(2\pi/M)$ – where k is an integer (1, 2, 3...). An analytical expression for $H_0(z)$ is supplied without derivation in [92]. Thus the expression for the frequency response $H_0(\omega)$ of the prototype DFT filter may be derived as follows in this thesis. Start with the prototype DFT filter $H_0(z)$

$$\begin{aligned}
H_o(z) &= 1 + z^{-1} + \dots + z^{-(M-1)} \\
&= 1 + \left(\frac{1}{z}\right)^1 + \left(\frac{1}{z}\right)^2 + \dots + \left(\frac{1}{z}\right)^{M-1} \\
&= \frac{1 - \left(\frac{1}{z}\right)^{M-1+1}}{1 - \left(\frac{1}{z}\right)}
\end{aligned} \tag{56}$$

Where $z = e^{j\omega}$ is substituted

$$\begin{aligned}
H_o(e^{j\omega}) &= \frac{1 - e^{-j\omega M}}{1 - e^{-j\omega}} \\
&= \frac{1 - (\cos(M\omega) - j \sin(M\omega))}{1 - (\cos(\omega) - j \sin(\omega))}
\end{aligned} \tag{57}$$

Consequently, the amplitude frequency response of the prototype filter (DFT filterbank) is given by

$$\begin{aligned}
|H_o(e^{j\omega})| &= \left| \frac{1 - e^{-j\omega M}}{1 - e^{-j\omega}} \right| \\
&= \frac{|1 - (\cos(M\omega) - j \sin(M\omega))|}{|1 - (\cos(\omega) - j \sin(\omega))|} \\
&= \sqrt{\frac{(1 - \cos(M\omega))^2 + \sin(M\omega)^2}{(1 - \cos(\omega))^2 + \sin(\omega)^2}} \\
&= \sqrt{\frac{1 - 2\cos(M\omega) + \cos(M\omega)^2 + \sin(M\omega)^2}{1 - 2\cos(\omega) + \cos(\omega)^2 + \sin(\omega)^2}} \\
&= \sqrt{\frac{2(1 - \cos(M\omega))}{2(1 - \cos(\omega))}} \\
&= \sqrt{\frac{2(\sin(M\omega/2))^2}{2(\sin(\omega/2))^2}} \\
&= \frac{\sin(M\omega/2)}{\sin(\omega/2)} \\
&= \frac{\sin(M\pi f)}{\sin(\pi f)}
\end{aligned} \tag{58}$$

This is the same expression as on p. 116 in [92]. The above expression may be compared with the expression for $C(f)$ given by equation 52 on p. 124. The same functional

$\sin(Nx)/\sin(x)$ form may be observed in both equations. The expressions for $C(f)$ and $H(f)$ differ as follows:

1. The expression for $C(f)$ has a constant $1/N$ factor. As the scaling factor of $1/N$ in the DFT filterbank may be distributed in the forward or backward transform (or equally between both transforms using a factor of $1/(\sqrt{N})$, this poses no problem.
2. The expression for $C(f)$ has an extra factor of T_f in the frequency domain expression resulting from the continuous Fourier transform of $c(t)$. By contrast, the expression of $H(e^{j\omega})$ results from a periodic sampling $x(n)$ of the signal $x(t)$.

Furthermore, the ESFFT retains the frequency components only corresponding to integer multiples of the trigger frequency T_f . As the ESFFT keeps the prototype filters corresponding to an integer multiple of the sampling frequency only (corresponding to the trigger frequency), the frequency domain transfer function of the ESFFT is⁶³

$$\text{ESFFT}(j\omega) = |H_o(e^{j\omega})| + \dots + |H_{mL}(e^{j\omega})| \quad (59)$$

where $m = 1, 2, 3, \dots$ and L is an integer corresponding to $L = F_s/F_t$ (F_s is the sampling frequency and F_t is the trigger frequency) and let $J = 1/L = F_t/F_s$. Furthermore, the prototype DFT filter bank [92] is a shifted version of the prototype filter $H_o(e^{j\omega})$

$$|H_{mL}(e^{j\omega})| = |H_o(e^{j\omega - (2\pi k/M)})|, k = mL \quad (60)$$

$$k = mM/L$$

hence

$$\begin{aligned} &= |H_o(e^{j\omega})| + \dots + |H_o(e^{j\omega - 2\pi mL/M})| \\ &= \sin(M\omega/2)/\sin(\omega/2) + \dots + \sin(m(\omega - 2\pi mL/M)) \end{aligned} \quad (61)$$

⁶³ Instead of the magnitude (absolute value) the energy (square of the signal) could also have been used.

where $(M/L)/M = J = 1/L = F_s/F_s$. Although an exact analytical expression for equation 61 is not available, the following observation can be made:

As the number of synchronous averages M increases to infinity, the frequency selectivity increases. In the limit as M approaches infinity the frequency response of the prototype filter $H_0(e^{j\omega})$ becomes a dirac delta impulse. Thus the frequency response of the prototype filter $H_m(e^{j\omega})$ becomes a dirac delta impulse centred at frequency $2\pi(m/L)$. A dirac delta impulse sequence results when M approaches infinity, having a repetition frequency of $2\pi(m/L)$. The corresponding analog frequency is: $f = mJ * F_s = m/L * F_s = mF_s$.

This corresponds well with equation 53 (where the comb filter is unity at frequencies of $f = m/T_s$).

Another way to derive the same result would be to examine the contribution of the shifted prototype filters $H_m(e^{j\omega})$ based on the prototype filter $H_0(e^{j\omega})$. As the prototype filter $H_0(e^{j\omega})$ becomes more localized, the response at frequencies far from $\omega = 0$ starts to decrease. In the limit, as the number of averages M approaches infinity, the prototype filter $H_0(e^{j\omega})$ responds around $\omega = 0$ only (becoming a dirac delta pulse at $\omega = 0$). In a similar way the shifted prototype filters $H_m(e^{j\omega})$ respond to frequencies around $\omega = 2\pi(mJ)$ only. Hence according to McFadden and Toozy [53]

$$\begin{aligned} |H_0(e^{j\omega})| &= \frac{\sin(M\omega/2)}{\sin(\omega/2)} \\ &= \begin{cases} M & \text{for } \omega \rightarrow 0 \text{ and } \omega = \pi k / M \text{ where } k = 0, 2, 4, \dots \\ 0 & \text{for } \omega = \pi k / M \text{ where } k = 1, 3, 5, \dots \end{cases} \\ &= \delta(\omega) \text{ as } M \rightarrow \infty \end{aligned} \quad (62)$$

and

$$\begin{aligned} |H_k(e^{j\omega})| &= \frac{\sin(M\omega - (2\pi(k/M))/2)}{\sin(\omega - (2\pi(k/M))/2)} \\ &= \delta(\omega - (2\pi(k/M))) \text{ as } M \rightarrow \infty \end{aligned} \quad (63)$$

It stands to reason that the normalized value of $1/M H_m(e^{j\omega})$ would be one (McFadden and Toozy [53]) and not infinity as required for the impulse signal, but the characteristic feature of the (idealized) impulse signal, which is that all the energy is concentrated

around the origin, remains the same. In addition, $M(1/M H_{mJ}(e^{j\omega})) = H_{mJ}(e^{j\omega})$ would be an impulse signal as M approaches infinity.

6.5 Effect of time domain window on the SA

6.5.1 Introduction

Implicit in the computation of the SA is the choice of window. This aspect was not explored initially in the literature survey, but subsequent experiments (refer to appendix I on p. 391) indicated that the choice of the time domain window could be important. No reference could be found to the influence of time domain window for the computation of the SA. Hence a few exploratory comments will be made on this subject.

It is well known that an optimal time domain window always involves a trade-off between the sidelobe level and the width of the main lobe [66]. Decreasing the amplitude of the first sidelobe (usually the largest sidelobe) increases the main lobe (transition) width and vice versa. Windows achieving high sidelobe attenuation (such as the Blackman-Harris window) invariably have a wider transition width.

The first sidelobe level is fixed at 13dB for the **rectangular window** regardless of the frequency resolution [70]. An expression for the amplitude of sidelobe k of the rectangular window is given by (refer to appendix G for further details)

$$S(k) = \frac{2}{\pi(2k+1)} \quad (64)$$

On a dB scale it is $-20\log(k) + C$ when k is large enough ($2k > 1$), where $C = -20\log(2)$.

The first sidelobe level is fixed at 27dB for the **triangular window** regardless of the frequency resolution [70]. An expression for the amplitude of sidelobe k is given by

$$S(k) = \left(\frac{2}{\pi(2k+1)} \right)^2 \quad (65)$$

On a dB scale it is $-40\log(k) + C$ when k is large enough ($2k > 1$), where $C = -40\log(2)$.

The sidelobe levels of the triangular window are clearly significantly lower than those of the rectangular window, although the frequencies corresponding to the sidelobe levels do not have to be the same for different windows. The following table lists the sidelobe levels corresponding to the 10th sidelobe level of the windows [66] for fair comparison:

Table 2: 10th sidelobe level of windows

Window function	Level of 10 th sidelobe (dB)	Frequency of 10 th sidelobe
Rectangular	-30	0.45π
Bartlett	-55	0.85π
Hanning	-70	0.45π
Hamming	-45	0.45π
Blackman	-80	0.5π

Alternatively the level of the window (attenuation in dB) may be considered at fixed frequencies for window functions ([66]):

Table 3: Level (dB) of windows at specific frequencies

Window function	0.1π	0.2π	0.3π	0.4π	0.5π
Rectangular	-20	-30	-38	-38	-40
Bartlett	-40	-35	-43	-60	-60
Hanning	-30	-58	-65	-80	-82
Hamming	-50	-52	-42	-45	-60
Blackman	-20	-62	-72	-80	-82

It should be noted that the values in table 3 above are obtained by inspection from figure 7.30 on pp. 448–449 of [66] and have not been calculated, hence they serve as approximate values only as a guideline for the influence of window functions.

It is also important to note that a time domain window would have no effect if the defect repeats periodically at a frequency exactly corresponding to a multiple of the frequency resolution of the DFT. The high sidelobe levels of the rectangular window may be disregarded in this event, as the defect frequency is always a multiple of the inherent DFT frequency bins (no picket fence effect). By contrast, if the defect repeats periodically at a frequency corresponding to a non-integer multiple of the frequency resolution of the DFT then the sidelobe levels of the time domain window would become relevant. Increasing the sampling frequency would increase the frequency resolution. Consequently, the first sidelobe would occur at a lower frequency and the attenuation should be greater at a specified frequency. As the sampling frequency increases, it is also more likely that the periodic defect frequency would align perfectly with the inherent frequency bins of the DFT. However, in most practical systems the sampling frequency cannot be increased arbitrarily. By increasing the buffer length (for example by zero padding the signal) the frequency resolution may be increased, but this does not add any new information to the

spectrum [92]. The process of zero padding may simply be regarded as an interpolation procedure in the frequency domain [92].

Another comment about the introduction of time domain windows is as follows. A flat frequency response is usually desirable for filter design, even if the frequency is not an exact multiple of the inherent DFT frequency resolution. However, by attenuating frequencies not corresponding to an exact multiple of the inherent DFT frequency resolution the energy not synchronous with the inherent DFT frequency resolution is also filtered out. This may be desirable in some cases, but it should be borne in mind that the time domain signal is modified by the application of the time domain window function. In particular, appropriate choice of the time domain window function can mask or hide any amplitude or speed variations in the signal (which may be an advantage or disadvantage, depending on the application).

6.5.2 Sidelobe level for specified defect frequency

Both the influence of the sidelobe level for a specified defect frequency as well as the number of SAs (indicated by the symbol M) are of interest. The frequency spectrum corresponding to the SA (using a rectangular window) is given by [50]

$$G(\omega) = \frac{1}{M} \frac{\sin(\omega M / 2)}{\sin(\omega / 2)} \quad (66)$$

The number of sidebands k corresponding to a defect frequency ω_k is

$$k = \frac{\omega_k}{3\pi/M} \quad (67)$$

Hence the dB values of the amplitude spectrum $A(\omega)$ for sidelobe level k are given by⁶⁴

$$\begin{aligned} A(\omega) &= -20 \log \omega_k + 20 \log \left(\frac{3\pi}{M} \right) \\ &= -20 \log \omega_k - 20 \log(M) + 20 \log(3\pi) \end{aligned} \quad (68)$$

⁶⁴ The frequency attenuation corresponding to a sidelobe level k is given by $2/(\pi(2k+1))$ for the rectangular window – refer to section 6.5.1 on p. 128.

It is clear from equation 68 that increasing ω_k or M would result in decreased sidelobe levels. As it may not be possible or practicable to increase the defect frequency ω_k , the alternative is to increase the buffer length M (i.e. to increase the number of SAs). It is also evident from equation 68 that a 6dB decrease in the sidelobe level may be obtained by a doubling of the defect frequency ω_k or buffer length M . It is well known from literature that large, slow-moving machinery presents special problems for fault detection. Equation 68 supports this knowledge and indicates that it would be necessary to use a larger buffer length M to obtain the same sidelobe level for machinery with low rotation speed.

6.5.3 Conclusion

Some aspects have been discussed related to the choice of window function when computing the SA. It has been shown that either increasing the defect frequency or buffer length can decrease the sidelobe levels for the rectangular window. The results of experiments (appendix I on p. 391) suggest that the SA becomes narrower (more selective) as the sidelobe levels decrease. This would seem to imply that a time window with decreased sidelobe levels could reduce the effect of speed variations in a signal (speed variations corresponding to instantaneous changes in frequency – a time window with high sidelobe level attenuation can attenuate all frequencies not synchronous with the trigger frequency). This is left for future work (section 8.6 on p. 263).

6.6 Modelling of speed variations of a vibration signal

A simplified model may be proposed to model the speed variations of a vibration signal based on the following simplifying assumptions:

1. All the defect pulses $g(t - kT)$, where T is the repetition period of the pulses, have the same shape (are identical).
2. The defect pulses $g(t)$ are zero outside the repetition period T of the pulse (it is assumed the ringing of the defect pulses decays fast enough when excited

by the impact of the defect so that the response of $g(t)$ is zero or negligible for $t > T$).

3. No amplitude modulation of the basic defect pulses is considered.

The effect of assumption 1 is that the synchronous average (SA) of the defect pulses⁶⁵ equals the basic defect pulse shape $g(t - kT)$. Let the basic defect pulse be represented by a function $g(t)$, which could, for example, be a Gaussian modulated signal as proposed in this thesis. With no speed variation present and a basic defect frequency of f_I the fault signal can be represented by the convolution of the basic pulse shape $g(t)$ with a delta comb sequence on the time domain

$$\begin{aligned} d(t) &= g(t) + g(t - T) + \dots g(t - kT) \\ &= g(t) \otimes \sum_{k=-\infty}^{\infty} \delta(t - kT) \end{aligned} \quad (69)$$

The frequency domain interpretation would be a multiplication of $G(\omega)$ by a comb sequence (in the frequency domain), resulting in a series of regularly spaced impulses in the frequency domain, as would be expected from a periodic time domain sequence.

Now let us assume that some defect pulses have speed variations (equivalent to a phase offset). The equation for the signal $d_I(t)$ with no speed variations and basic repetition period of $T_I = 1/f_I$ is

$$\begin{aligned} d_I(t) &= g(t) + g(t - T_I) + \dots g(t - kT_I) \\ &= g(t) \otimes \sum_{k \in F_I} \delta(t - kT_I) \end{aligned} \quad (70)$$

where F_I denotes the set of all integers corresponding to no speed variations.

If we assume for the moment that the speed variation is limited to a frequency f_2 only, the equation for the signal $d_2(t)$ with speed variation is

⁶⁵ In a practical system where amplitude and speed variations are present the signal would be filtered and demodulated (for example by using squaring) before applying the SA.

$$\begin{aligned}
d_2(t) &= g(t) + g(t - T_2) + \dots g(t - mT_2) \\
&= g(t) \otimes \sum_{k \in F_2} \delta(t - kT_2)
\end{aligned} \tag{71}$$

$$d(t) = d_1(t) + d_2(t) = g(t) \otimes \sum_{k \in F_1} \delta(t - kT_1) + g(t) \otimes \sum_{k \in F_2} \delta(t - kT_2) \tag{72}$$

The above equation may be extended in an obvious way to accommodate arbitrary speed variations

$$d(t) = d_1(t) + d_2(t) + \dots = g(t) \otimes \sum_{k \in F_1} \delta(t - kT_1) + g(t) \otimes \sum_{k \in F_2} \delta(t - kT_2) + \dots \tag{73}$$

Unfortunately, the Fourier transform (or series) of a delta sequence containing an arbitrary number of terms is not available in literature as far as could be determined. It is therefore necessary to rewrite the equations for easier mathematical analysis, using the rectangular function $r(t)$ as defined below

$$\begin{aligned}
d_k(t) &= g(t) + g(t - T_1) + \dots g(t - kT_1) \\
&= \left[g(t) \otimes \sum_{k=-\infty}^{\infty} \delta(t - kT_1) \right] \sum_{m \in F_1} r(t - mT)
\end{aligned} \tag{74}$$

where $r(t) = 1$ for $0 \leq t \leq T$

and $r(t) = 0$ for $t > T$ or $t < 0$

The equation above simplifies to equation 69 when m varies from $-\infty$ to ∞ instead of over the limited set F_1 . Hence $(\sum r(t - mT)) = 1$ for all t , and the Fourier transform is a delta pulse at the origin and the convolution with the infinite delta sequence remains an infinite delta sequence.

Firstly, consider the effect of two rectangular pulses next to each other

$$f(t) = r(t) + r(t - T) \tag{75}$$

$$\begin{aligned}
F(\omega) &= TSa(\omega T / 2) + e^{-j\omega T} TSa(\omega T / 2) \\
&= (1 + e^{-j\omega T}) TSa(\omega T / 2) \\
&= e^{-j\omega T / 2} (e^{j\omega T / 2} + e^{-j\omega T / 2}) TSa(\omega T / 2) \\
&= e^{-j\omega T / 2} (\cos(\omega T / 2)) T \frac{\sin(\omega T / 2)}{\omega T / 2} \\
&= e^{-j\omega T / 2} T \frac{2 \cos(\omega T / 2) \sin(\omega T / 2)}{\omega T} \\
&= e^{-j\omega T / 2} T \frac{\sin(\omega T)}{\omega T} \\
&= e^{-j\omega T / 2} TSa(\omega T)
\end{aligned} \tag{76}$$

As $f(t)$ can be rewritten as $f(t) = \text{rect}((t - T/2)/(2T))$ the Fourier transform can be written as

$$F(\omega) = e^{-j\omega T / 2} TSa(\omega T) \tag{77}$$

which confirms the previous derivation.

Now consider the effect of two rectangular pulses spaced mT apart

$$f(t) = r(t) + r(t - mT) \tag{78}$$

$$\begin{aligned}
F(\omega) &= TSa(\omega T / 2) + e^{-j\omega mT} TSa(\omega T / 2) \\
&= (1 + e^{-j\omega mT}) TSa(\omega T / 2) \\
&= e^{-j\omega mT / 2} (e^{j\omega mT / 2} + e^{-j\omega mT / 2}) TSa(\omega T / 2) \\
&= e^{-j\omega mT / 2} (\cos(\omega mT / 2)) T \frac{\sin(\omega T / 2)}{\omega T / 2} \\
&= A(\omega) e^{-j\omega mT / 2} TSa(\omega T / 2), \text{ where } A(\omega) = \cos(\omega mT / 2)
\end{aligned} \tag{79}$$

Hence $A(\omega)$ represents an amplitude modulation term in the frequency domain that modifies the basic $Sa(\omega T / 2)$ waveform. As the second rectangular pulse becomes spaced further from the first rectangular pulse the position of the first null (at $\omega mT / 2 = \pi$ or $\omega = \pi / (mT)$) in the frequency domain decreases (nulls would occur when the frequency of the sine wave is such that it fits exactly between the rectangular pulses, including higher order harmonics of course). However, the amplitude of the first sidelobe of $F(\omega)$ at $\omega = 2\pi / (mT)$ would be much larger than the first sidelobe of the $Sa(\omega T / 2)$ waveform. Thus it may be concluded that the best situation as far as reduction of sidelobe

levels is concerned would occur when the rectangular pulses are right next to each other (or $m = 1$). In this case the frequency of the first sidelobe is halved.

In general, the frequency of the first sidelobe would occur at $1/k$ of the basic $Sa(\omega T/2)$ waveform for k rectangular pulses next to each other. Therefore, as the number of consecutive rectangular pulses is increased ($k \rightarrow \infty$) the first sidelobe becomes closer and closer to the origin (in the limit of $k \rightarrow \infty$ a dirac delta pulse is obtained at the origin). When the k rectangular pulses are no longer spaced together the amplitude modulation term $A(\omega)$ determines the frequency response of $F(\omega)$ in a non-trivial way. A lower limit (the best that could be hoped for when the pulses are all next to each other) on the frequency of the first sidelobe level would be $\omega = 2\pi/(kT)$.

In the worst case the basic $Sa(\omega T/2)$ waveform would be modulated by a high frequency carrier resulting in a high amplitude of the first sidelobe at $\omega = 2\pi/(kT)$ when the pulses are spaced far apart⁶⁶. Alternatively, it can be seen that the basic $Sa(\omega T/2)$ waveform is not changed significantly by the amplitude modulation term $A(\omega)$ when the pulses are spaced far apart⁶⁷. Hence the amplitude and frequency of the first sidelobe level of the basic $Sa(\omega T/2)$ waveform remain the same even with k rectangular pulses.

In any dynamic system operating in an industrial setting it is extremely unlikely that the speed variations would follow a simple mathematical formula (such as a simple ramp in frequency). A more likely assumption would be that the speed variations follow, for example, a Gaussian distribution. According to this assumption, the number of rectangular pulses decreases with increasing speed variation. Thus the frequency of the first sidelobe could be expected to increase with increasing speed variation⁶⁸. For a Gaussian distribution of rectangular pulses it is also unlikely that a sine wave of particular frequency would fit exactly between all the rectangular pulses. Consequently, no nulls could be

⁶⁶ Refer to the previous result for two rectangular pulses spaced far apart.

⁶⁷ The basic $Sa(\omega T/2)$ waveform is simply modulated by a high frequency carrier wave when the pulses are spaced far apart.

⁶⁸ It stands to reason that with increasing speed variation the defect frequency also increases.

expected to occur in the frequency domain for a Gaussian distribution of rectangular pulses.

It is clear that equation 74 represents the convolution of $g(t)$ with an infinite dirac delta impulse sequence, which is then multiplied by the rectangular series. Therefore, the Fourier domain expression would consist of the multiplication of $G(\omega)$ by an infinite dirac delta impulse sequence, which is then convolved with the spectrum $F(\omega)$ of the rectangular pulses. Although it seems difficult at present to obtain an analytical expression for $F(\omega)$, the following qualitative comment may be made. The width of $F(\omega)$ may be related to the frequency of the first sidelobe. Fewer rectangular pulses would occur with increasing speed variation, hence the frequency of the first sidelobe would increase, resulting in a wider frequency response. When $F(\omega)$ is convolved with the regularly spaced harmonic lines it is evident that the spectrum would be smeared more with increasing speed variation. The frequency response $D_I(\omega)$ of $d_I(t)$ would be given by

$$\begin{aligned}
 D_I(\omega) &= G(j\omega) + e^{j\omega 2T_1} G(\omega) + \dots e^{j\omega 2T_1} g(\omega) \\
 &= \left[G(\omega) \sum_{k=-\infty}^{\infty} \delta(\omega - k\omega_{T_1}) \right] \otimes \mathfrak{S} \left(\sum_{m \in F_1} r(t - mT_1) \right) \\
 &= [G(k\omega_{T_1})] \otimes \mathfrak{S} \left(\sum_{m \in F_1} r(t - mT_1) \right)
 \end{aligned} \tag{80}$$

The total frequency response $D(\omega)$ of $d(t) = d_I(t) + \dots + d_k(t)$ would be given by $D(\omega) = D_I(\omega) + \dots + D_k(\omega)$.

In a practical system, where speed and amplitude variations are present, it would not be trivial to segment the defect signal as proposed above. One approach could be to represent the defect signal as a convoluted (not linear) mixture of elementary defect signals as proposed above, using the framework of blind source separation (BSS). The aim would be to determine the mixtures (each representing a different frequency) based on some suitable performance metric. This is left for future work. An alternative model, based on the frequency domain interpretation of the SA as a comb filter, may be considered instead (refer to section 6.7 on p. 138).

A model may be proposed based on the simplified assumption of identical defect pulses. Let the basic defect pulse be given by the average value (SA) of the filtered and demodulated measured signal⁶⁹. A more representative defect signal may be obtained by distorting the basic defect pulse $g(t)$ and adding noise. Speed and amplitude variations can also be incorporated into the model. The amplitude variations could be based on the second order statistic of the standard deviation (instead of the first order SA only). This is also left for future work.

6.7 Frequency domain modelling of speed variations of a vibration signal

Consider the effect of the SA in the frequency domain. The frequency response of the SA is given by a comb filter in the frequency domain [50, 53]. However, it would not be possible to represent any arbitrary signal using a single SA only (unless of course the signal happened to be exactly periodic at the defect frequency). In order to obtain the property of perfect reconstruction it would be necessary to increase the 'coverage' of the comb filter in order to cover the complete frequency range (eventually resulting in a unity transfer function in the frequency domain).

One possibility of achieving the above requirement for perfect reconstruction would be to simply shift the comb filter associated with the defect frequency ω_d from frequencies 0 to ω_d . The corresponding effect on the time domain would be multiplication by the complex exponential $\exp(j\omega_m\omega)$. Hence all the SAs are computed in the time domain after first applying a modulating function $\exp(j\omega_m\omega)$. If the original time domain is periodic then modulation by $\exp(j\omega_m\omega)$ would have little effect other than at $\omega_m = 0$. Consequently, if the frequency domain signals corresponding to the comb filter with frequency shift F_m are represented by $G(\omega, \omega_m)$ then

$$G(\omega, \omega_m) = 0 \text{ for all } \omega_m \neq 0 \quad (81)$$

⁶⁹ A suitable window function could also be applied. The demodulated signal representing the 'envelope' of the signal could be modulated again by a high frequency carrier.

When a large bearing defect is present the above equation would be valid to a certain degree, as most energy in the signal would be concentrated in the SA at the defect frequency ($\omega_m = 0$). However, for smaller bearing defects amplitude and speed variations would contribute to a 'haystack' in which the presence of a single defect frequency is no longer clear⁷⁰. Furthermore, it is evident that the sum of $G(\omega, \omega_m)$ should be unity in the frequency domain⁷¹

$$\sum_{\omega_m=0}^{\omega_b} G(\omega, \omega_m) \cong 1 \quad (82)$$

Hence multiplication of the spectrum $X(\omega)$ by $G(\omega, \omega_m)$ yields the perfect reconstruction property

$$\sum_{\omega_m=0}^{\omega_b} G(\omega, \omega_m) X(\omega) = X(\omega) \quad (83)$$

6.8 Relation between SA and demodulation spectrum

To obtain the demodulated spectrum the signal is bandpass filtered around system resonance(s) prior to application of a non-linear operator (such as squaring). Let $X(\omega)$ represent the filtered signal before demodulation. After applying the squaring operator the demodulated signal $X_2(\omega)$ is given by the convolution between $X(\omega)$ and $X(\omega)$ (for example [66])

$$X_2(\omega) = \int_{\varpi=-\infty}^{\infty} X(\varpi) X(\varpi - \omega) d\varpi \quad (84)$$

⁷⁰ Assuming that the non-linear operator corresponding to squaring is not applied.

⁷¹ The effect of the window function on the summation is disregarded to simplify the discussion.

As

$$Y_{\omega_m}(\omega) = X(\omega) \sum_{\omega_m=0}^{\omega_d} G(\omega, \omega_m) \quad (85)$$

then $X(\omega) = \sum_{\omega_m=0}^{\omega_d} Y_{\omega_m}(\omega)$

where $Y(\omega_m, \omega)$ is the signal corresponding to the synchronous average at repetition frequency ω_d and shift of frequency ω_m . Hence it is clear that the demodulated spectrum $X_2(\omega_d)$ at the defect frequency ω_d or $X_2(\omega_d)$, where $\omega_d = 2\pi F_d$, is given by

$$\begin{aligned} X_2(\omega_d) &= \int_{\omega=-\infty}^{\infty} \sum_{\omega_m=0}^{\omega_d} Y_{\omega_m}(\omega) \sum_{\omega_m=0}^{\omega_d} Y_{\omega_m}(\omega - \omega_d) d\omega \\ &= \int_{\omega=-\infty}^{\infty} \sum_{\omega_m=0}^{\omega_d} Y_{\omega_m}(\omega) \sum_{\omega_m=0}^{\omega_d} Y_{\omega_m + \omega_d}(\omega) d\omega \\ &= \int_{\omega=-\infty}^{\infty} \sum_{\omega_m=0}^{\omega_d} Y_{\omega_m}(\omega) Y_{\omega_m + \omega_d}(\omega) d\omega \end{aligned} \quad (86)$$

Thus the demodulated spectrum $X_2(\omega_d)$ is obtained by computing (in the frequency domain) the signal (corresponding to all shifts ω_m) multiplied by the same signal with a shift of ω_d in frequency and summing over all frequencies ω . In the time domain this corresponds to the SA of the first signal after multiplication by the complex exponential $\exp(j\omega_m t)$ for all frequencies ω_m between 0 and ω_d and multiplication of the second signal by the complex exponential $\exp(j\omega_{m2} t)$ for all frequencies ω_{m2} between ω_m and $\omega_m + \omega_d$ and summing the result.

The demodulated spectrum $X_2(\omega_d)$ is obtained by computing (in the frequency domain) the following for a periodic signal $\omega_m = 0$

$$\begin{aligned}
X_2(\omega_d) &= \int_{-\infty}^{\infty} \left[\sum_{\omega_m=0}^{\omega_d} Y_{\omega_m}(\omega) \sum_{\omega_n=0}^{\omega_d} Y_{\omega_n}(\omega - \omega_d) \right] d\omega \\
&= \int_{-\infty}^{\infty} \sum_{\omega_m=0}^{\omega_d} Y_{\omega_m}(\omega) \sum_{\omega_n=0}^{\omega_d} Y_{\omega_n + \omega_d}(\omega) d\omega \\
&= \int_{-\infty}^{\infty} \sum_{\omega_m=0}^{\omega_d} Y_0(\omega) Y_{\omega_d}(\omega) d\omega \\
&= \int_{-\infty}^{\infty} Y_0(\omega) Y_{\omega_d}(\omega) d\omega
\end{aligned} \tag{87}$$

When $\omega_d = 0$ then $X_2(0)/2\pi$ is simply the energy in the SA of the signal. Otherwise $X_2(\omega_d)$ is obtained by shifting $Y_0(\omega)$ with a frequency ω_d in the frequency domain and then multiplying by $Y_0(\omega)$ and summing the result. Note that $Y_0(\omega)$ consists of harmonic lines at multiples of the defect frequency ω_d . In the time domain the interpretation of a shift of a frequency $\omega_m = \omega_d$ in $Y_0(\omega)$ (in the frequency domain) is a multiplication of the time domain signal $y_0(t)$, repeating at defect frequency ω_d , by a complex exponential $\exp(j\omega_m t)$. Hence the time domain signal $y_0(t)$ is multiplied by $\exp(j\omega_m t)$ (equivalently bandpass filtered around frequency ω_m) and then multiplied by $y_0(t)$ and summed⁷².

As $\exp(j\omega_m t) = \cos(\omega_m t) + j\sin(\omega_m t)$ the time domain signal $y_0(t)$ is modulated with $\cos(\omega_m t)$ and the result is added to $jy_0(t)\sin(\omega_m t)$. In the frequency domain the spectrum of $Y_0(\omega)$ is convolved with the spectrum of $\exp(j\omega_m t)$ – a dirac delta impulse at ω_m in the frequency domain. The result is a shift of ω_m in frequency.

It follows from the previous discussion and the definition of the ESFFT that the time domain signal $y_0(t)$ is bandpass filtered around ω_m and subsequently multiplied by the time domain signal $y_0(t)$ and the result summed to obtain the energy at a shift of ω_m in frequency. The resulting signal is related to $X_2(\omega_m)$

⁷² In the product only the component at $\omega = 0$ contributes, hence the convolution changes to a multiplication (Parseval's formula).

$$\begin{aligned}
X_2(\omega_m) &= 2\pi \int y_0(t) e^{j\omega_m t} y_0(t) dt \\
&= 2\pi \int y_0(\tau) e^{j\omega_m \tau} y_0(\tau) d\tau \\
\text{Hence } X_2(0)/2\pi &= \int y_0(\tau) y_0(\tau) d\tau
\end{aligned} \tag{88}$$

is simply the energy of the signal as expected. However, when $\omega_m > 0$ the time domain signal $y_0(t)$ is first bandpass filtered around ω_m and then multiplied by $y_0(t)$ and integrated. The resulting energy (with the shift of frequency ω_m in the frequency domain) should always be smaller than or equal to the energy of the original time domain signal $y_0(t)$, as the filtering process always removes energy from a frequency band (bandpass filter): $|X_2(\omega_m)| \leq |X_2(0)|$. This result may be established by analogy with a theorem of the autocorrelation function ([68])

$$|R(\tau)| \leq R(0) \tag{89}$$

This result is derived based on the information that the power spectrum is real and positive ([68]). As $R(\tau)$ represents a correlation (or equivalently a convolution if the signal is swapped around on the time axis) between the signal $x(t)$ and the signal $x(t)$ delayed by τ , a similar result should also apply to a frequency domain convolution between $X(\omega)$ and $X(\omega)$ shifted (delayed) by a frequency. The square of the signal $x(t)$ is known to be always real and positive, hence

$$|X_2(\omega)| = \frac{1}{2\pi} \left| \int_{-\infty}^{\infty} x(t)^2 e^{j\omega t} dt \right| \leq \frac{1}{2\pi} \int_{-\infty}^{\infty} x(t)^2 dt = X_2(0) \tag{90}$$

In appendix H on p. 385 the above result is illustrated by considering k sinusoidal sequences in the most general case.

For a non-periodic signal $\omega_m \neq 0$. Thus $X(\omega)$ would consist of the superposition of many SAs, with each SA bandpass filtered around ω_{m2} (where ω_{m2} varies from 0 to $\omega_m = \omega_d$). The interpretation essentially remains the same, except that $X_2(\omega_m)$ is the sum of all the individual comb filter (ESFFT) contributions. The previous equations could, of course, all have been written using the ESFFT definition.

To illustrate the above statement, consider the expansion of a signal $x(t)$ in terms of the summation of periodic signals (each of course corresponding to a SA) $p_k(t)$: $x(t) = \sum p_k(t)$, where $p_0(t) = SA(x(t))$, $p_1(t) = SA(e^{j\omega_d t} x(t))$, ... $p_k(t) = SA(e^{j\omega_k t} x(t))$. The signal $x(t)$ has been modulated in frequency by ω_k (where $0 \leq \omega_k < \omega_d$). If the signal $x(t)$ is periodic (for example a large bearing defect in the loading zone) then $x(t) = SA(x(t))$ ⁷³ and the Fourier transform of (a periodic expansion of) $x(t)$ would consist of a set of lines spaced harmonically at the defect frequency. Otherwise $x(t)$ is given by the sum of a finite number of periodic signals $p_k(t)$.

If it is assumed that the signal $x(t)$ has already been bandpass filtered, the influence of the squaring operation $x(t)^2$ in the time domain may be determined: $x(t)^2 = (\sum p_k(t))^2 = \sum p_k(t)^2 + 2\sum p_m(t)p_n(t)$ (where $m \neq n$). The interpretation is straightforward: the squared signal consists of the summation of each squared individual periodic signal $p_k(t)^2$ (the summation of $x(t)^2$ would of course be related to the energy of the signal) and (twice) the summation of each individual periodic signal $p_m(t)$ multiplied by the other periodic signal $p_n(t)$. For a periodic signal $x(t)$ only $p_0(t) \neq 0$, hence $x(t)^2 = p_0(t)^2 = SA(x(t))^2$.

In the frequency domain the expansion of $p_0(t)$ consists of harmonic lines $P_0(\omega)$ spaced at the defect frequency ω_d . The demodulated spectrum of $x(t)^2 = p_0(t)^2$ consists of the convolution of $P_0(\omega)$ with $P_0(\omega)$, which also consists of harmonic lines spaced at the defect frequency ω_d . When $x(t)$ is no longer periodic, the demodulated spectrum of $x(t)^2$ consists of

$$X_2(\omega) = P_0(\omega) \otimes P_0(\omega) + P_1(\omega) \otimes P_1(\omega) + \dots P_k(\omega) \otimes P_k(\omega) + 2\sum P_m(\omega) \otimes P_n(\omega) \\ = \sum P_k(\omega) \otimes P_k(\omega) + 2\sum P_m(\omega) \otimes P_n(\omega).$$

The term $\sum P_k(\omega) \otimes P_k(\omega)$ represents the contribution of the periodic signals $p_k(t)$ and supplies a series of harmonic lines at the defect frequency ω_d . The term $P_m(\omega) \otimes P_n(\omega)$ supplies the non-periodic contribution to the demodulation spectrum. It is important to note that the non-periodic components may have a number of causes:

⁷³ Assuming that the SA is periodically extended (alternatively confining the analysis to the time period of the SA).

1. Periodic speed variations of a periodic defect signal
2. Instantaneous amplitude variations of a periodic defect signal
3. Additive and multiplicative noise

Multiple defects at the defect frequency contribute to the periodic components only (assuming of course that no speed and amplitude variations occur), although the relative magnitudes of the harmonic components are influenced.

Although it is tempting to use normalized features based on the contribution of the periodic components relative to the non-periodic components [25, 63, 78], these could be misleading (as shown above) if the amplitude and speed variations cannot be eliminated. The normalized ratio measures the contribution of periodic components at the defect frequency (usually related to a defect) relative to the non-periodic components (which are related to the ‘carpet level’, but also to amplitude and speed variations in the defect signal itself). For small bearing defects it can be difficult to separate the random contributions (representing the ‘haystack’) from the very small bearing defect signal itself (as shown in this thesis). The individual convolutional components $P_k(\omega)$ of the demodulated spectrum could be analysed in this regard, but this is left for future work.

The demodulated spectrum contains contributions both from periodic and non-periodic signal components. However, the synchronous average (SA) considers periodic signal components at a specific frequency only. Consequently, the SA would provide good results when the defect signal is nearly periodic (or the periodic component is dominant). This is unlikely to be the case for small incipient bearing defects and better results can be expected with the demodulated spectrum (using bandpass filtering and squaring). In this section it was shown that the demodulated spectrum may be regarded as a summation of a number of SAs (consisting of the periodic and non-periodic signal components), which helps to explain the success of the demodulated spectrum relative to the SA (for small bearing defects) in this thesis.

6.9 Kurtosis as fault diagnostic

6.9.1 Introduction

Various aspects related to the kurtosis fault diagnostic are investigated in this section. The literature survey already mentioned the kurtosis statistic for fault detection. The kurtosis statistic is generally related to the ‘impulsiveness’ of a time domain signal and could provide fault detection under these circumstances. For a small bearing defect the ‘impulsiveness’ of the vibration signal decreases and the kurtosis statistic could be expected to be unsuitable (this will still be explored in more detail in chapter 7). General consensus has been reached on the interpretation of the first three central moments (mean, standard deviation and skewness) ([12, 57]). The interpretation of the fourth order statistic, kurtosis, is the subject of discussion in this section. Darlington [12] suggests that the kurtosis is a measure of the degree to which a distribution’s z-scores⁷⁴ cluster around +1 and -1. In [57] it is suggested that the interpretation of kurtosis as a measure of the bimodality of a distribution should be modified so that the kurtosis measures the dispersion around two values $\mu \pm \sigma$ (it is an inverse measure of the concentration in these two points).

6.9.2 Maximization of kurtosis (leptokurtic)

What is the signal that maximizes the kurtosis statistic? As the kurtosis is a measure of the ‘spikiness’ of a signal, it could be anticipated that a well localized signal such as the dirac delta impulse would maximize the kurtosis statistic. A distribution with a high peak is called leptokurtic, a flat-topped curve is called platykurtic, and the normal distribution is called mesokurtic [1]. Alternatively, the kurtosis may be viewed as a measure of the deviation from a normal Gaussian distribution that has a kurtosis of 3 (or normalized kurtosis of $3 - 3 = 0$). Again a Gaussian distribution would have instantaneous amplitude values widely varying in amplitude. So an impulse on the time domain would be most

⁷⁴ The z-scores are normalized by the standard deviation. In [57] this is explicitly stated as the transformation formula $Z_n = (X - \mu)/\sigma$, where X is the random variable, μ is the mean and σ is the standard deviation. The kurtosis k is then simply $k = E(Z_n^4) = \text{var}(Z_n^2) + (E(Z_n^2))^2 = \text{var}(Z_n^2) + 1$ where $\text{var}(Z_n^2)$ is the variance of Z_n^2 . The minimum value is $k = 1$ when $\text{var}(Z_n^2) = 0$ and the maximum value is infinity (when $\text{var}(Z_n^2)$ approaches infinity).

unlike the characteristics of the Gaussian distribution and hence a likely candidate to maximize the kurtosis. To proceed formally, start with the definition of kurtosis (where N is the buffer length of the finite time domain signal and assuming a zero mean signal x for the moment)

$$\begin{aligned}
kurt(x) &= \frac{\sum_{x=-\infty}^{\infty} x^4 / N}{\left(\sum_{x=-\infty}^{\infty} x^2 / N \right)^2} \\
&= N \frac{\sum_{x=-\infty}^{\infty} x^4}{\left(\sum_{x=-\infty}^{\infty} x^2 \right)^2} \\
&= N \frac{x_0^4 + x_1^4 + \dots x_N^4}{(x_0^2 + x_1^2 + \dots x_N^2)^2} \\
&= N \frac{x_0^4 + x_1^4 + \dots x_N^4}{x_0^4 + x_1^4 + \dots x_N^4 + 2(x_0^2 x_1^2 + \dots x_{N-1}^2 x_N^2)} \quad (91) \\
&= N \frac{A_1}{A_1 + B_1} \\
&= N \frac{1}{1 + B_1 / A_1}
\end{aligned}$$

where $A_1 = x_0^4 + x_1^4 + \dots x_N^4$ and $B_1 = 2(x_0^2 x_1^2 + \dots x_{N-1}^2 x_N^2)$

The above equation for the kurtosis will be maximized (with maximum value N) when $B_1 = 0$, or equivalently when $x_0^2 x_1^2 + \dots x_{N-1}^2 x_N^2 = 0$. As $x_0^2 \geq 0 \dots x_N^2 \geq 0$, the minimum value of zero will be obtained only when $x_0 = 0 \dots x_N = 0$ except for $x_k \neq 0$. Hence all the values of x are zero except for an impulse where $x_k \neq 0$. The resulting kurtosis would be

$$\begin{aligned}
kurt(x) &= \frac{\sum_{x=-\infty}^{\infty} x^4 / N}{\left(\sum_{x=-\infty}^{\infty} x^2 / N \right)^2} \\
&= N \frac{\sum_{x=-\infty}^{\infty} x^4}{\left(\sum_{x=-\infty}^{\infty} x^2 \right)^2} \\
&= N \frac{x_0^4 + x_1^4 + \dots x_N^4}{(x_0^2 + x_1^2 + \dots x_N^2)^2} \\
&= N \frac{1}{1 + B_1 / A_1} \\
&= N
\end{aligned} \tag{92}$$

and is independent of the signal amplitude.

As the mean value of a single impulse with amplitude A is A/N , the kurtosis of $x - A/N$ is (the mean value of $x - A/N$ is zero)

$$\begin{aligned}
kurt(x) &= \frac{\sum_{x=-\infty}^{\infty} (x - \mu)^4 / N}{\left(\sum_{x=-\infty}^{\infty} (x - \mu)^2 / N \right)^2} \\
&= N \frac{A^4 + NB^4}{(A^2 + NB^2)^2} \\
&= N \frac{A^4(1 + 1/N^3)}{(A^2(1 + 1/N))^2} \\
&= N \text{ when } N \text{ is large}
\end{aligned} \tag{93}$$

In [57] it is suggested that high kurtosis may arise from two factors:

1. A concentration of probability mass near μ
2. A concentration of probability mass in the tails

For a constant signal the first condition would apply, hence one would intuitively expect high kurtosis. However, the standard deviation σ is zero. Thus computation of the kurtosis involves the limit of numbers approaching zero over zero⁷⁵. It would therefore seem as though the above two conditions should be modified when σ becomes very small (approaches zero):

1. A concentration of probability mass near μ where $\sigma \neq 0$ (peaked unimodal distribution)
2. A concentration of probability mass in the tails of the distribution where $\sigma \neq 0$ (bimodal distributions can also have high kurtosis [57] when the modes are not close to $Z_n = \pm 1$). Hence kurtosis does not have to be associated with bimodality
3. When the standard deviation σ approaches zero the kurtosis is one, not infinity (section 6.9.3 on p. 149)

In [20] an example of a solid bimodal distribution is presented with excess kurtosis (3 is subtracted from the kurtosis value) ranging from -2 to $+3$. The following points are observed in this thesis:

- The kurtosis can never be smaller than one. The kurtosis k of the transformed (normalized and zero mean) variable $Z_n = (X - \mu)/\sigma$ is given by [57]:

$$kurt = E(Z_n^4) = \text{var}(Z_n^2) + (E(Z_n^2))^2 = \text{var}(Z_n^2) + 1$$
where $\text{var}(Z_n^2)$ is the variance of Z_n^2 . Thus the minimum value is $kurt = k = 1$ when $\text{var}(Z_n^2) = 0$.
- According to the previous point, the kurtosis of the proposed distribution actually varies between 1 and 3 for $\alpha > 1$.

⁷⁵ The rule of L'Hospital may be used to evaluate this limit, as will be shown later.

- According to [20] on p. 43 $kurt = k > 3$ for $\alpha < 1$ (“For $\alpha < 1$, $k > 3$ since k is decreasing in α ; when $\alpha < 1$ the density blows up at 0 but still should be called unimodal.”). Hence for $\alpha < 1$ the distribution is actually unimodal, not bimodal.
- The conclusion drawn in [20] is that the kurtosis statistic $kurt = k$ varies from -1.2 to 3 (in this thesis it is held that it should rather be between 1 and 3) for $\alpha > 1$ in the ‘two-tailed gamma’ distribution. However, this simply confirms the postulate in [12] that low kurtosis is an indication of bimodality in the signal (distribution). It is maintained in this thesis that [20] has not indicated a bimodal distribution with high kurtosis, as might be suggested by $kurt = k > 3$ for $\alpha < 1$ (which is actually a unimodal distribution with expected high value of kurtosis).
- When σ approaches zero (for example for a constant signal – one-point distribution) the transformed variable $Z_n = (X - \mu)/\sigma$ tends to infinity. Yet, the variable $X - \mu = 0$ for a constant signal, thus the limit of zero over zero has to be evaluated (for example using the rule of L’Hospital). Hence great care has to be taken when computing the kurtosis $kurt = k = E(Z_n^4) = \text{var}(Z_n^2) + (E(Z_n^2))^2 = \text{var}(Z_n^2) + 1$, where $\text{var}(Z_n^2)$ is the variance of Z_n^2 [57]. This will be explored in more detail in section 6.9.3 using some assumptions and the rule of L’Hospital to evaluate the problematic limit of zero over zero.

6.9.3 Minimization of kurtosis (platykurtic)

What is the signal that minimizes the kurtosis statistic? As the kurtosis is a measure of the ‘spikiness’ of a signal, it could be anticipated that a distributed signal in the time domain, such as a normal Gaussian distribution or constant signal, would minimize the kurtosis statistic. Again a Gaussian distribution would have instantaneous amplitude values widely varying in amplitude. So a normal Gaussian distribution or constant signal would be most unlike the characteristics of the impulse signal and consequently a likely candidate to

minimize the kurtosis. Instead of a constant signal $x = K$ with a mean value of K , consider a rectangular waveform with mean value of zero (this corresponds to a bimodal distribution as discussed in [12])

$$x = \begin{cases} K & \text{for } 0 \leq k \leq N/2 \\ -K & \text{for } N \geq k > N/2 \end{cases} \quad (94)$$

In [57] the transformation $Z_n = (X - \mu)/\sigma$ is used, where X is the random variable, μ is the mean and σ is the standard deviation. The kurtosis $kurt = k$ with the transformation $Z_n = (X - \mu)/\sigma$ as defined in [57] is $kurt = k = E(Z_n^4) = \text{var}(Z_n^2) + (E(Z_n^2))^2 = \text{var}(Z_n^2) + 1$ where $\text{var}(Z_n^2)$ is the variance of Z_n^2 . For the bimodal distribution x as defined above the standard deviation σ would be one. But $\text{var}(Z_n^2)$, the variance of Z_n^2 , is $\text{var}(X^2) = \text{var}(K^2) = 0$. Hence the kurtosis is a minimum at $kurt = k = 1$.

To proceed formally, start with the definition of kurtosis (where N is the buffer length of the finite time domain signal and assuming a zero mean signal x as defined above).

$$\begin{aligned} kurt(x) &= \frac{\sum_{x=-\infty}^{\infty} (x - \mu)^4 / N}{\left(\sum_{x=-\infty}^{\infty} (x - \mu)^2 / N \right)^2} \\ &= N \frac{NK^4}{(NK^2)^2} \\ &= 1 \end{aligned} \quad (95)$$

This interpretation of the kurtosis is entirely consistent with [12, 57], in which it is suggested that a two-point distribution has the kurtosis $kurt = k = 1$. For a constant signal the kurtosis involves the limit of zero over zero.

$$kurt(x) = \frac{\sum_{x=-\infty}^{\infty} (x_k - \mu)^4 / N}{\left(\sum_{x=-\infty}^{\infty} (x_k - \mu)^2 / N \right)^2}$$

but $x_1 = x_2 = \dots = x_k = A$

let $y_k = x_k - A$

(96)

$$\begin{aligned} kurt(x) &= \lim_{y_k \rightarrow 0} \frac{N^2 y_k^4}{(N y_k^2)^2} \\ &= \frac{\lim_{y_k \rightarrow 0} 4y_k^3}{\lim_{y_k \rightarrow 0} 4y_k^3} \text{ (L'Hospital)} \\ &= 1 \end{aligned}$$

Thus the constant signal also has minimum kurtosis in addition to the bimodal distribution [12].

As a high kurtosis value is frequently associated with 'spikiness' or 'impulsive behaviour' in the time domain, a constant signal would seem to have low 'spikiness', hence low kurtosis. As will be explored in section 6.9.5 on p. 154 and section 6.9.6 on p. 157 the kurtosis may also be interpreted as a measure of the change resulting in a shift of the squared signal. From this perspective the autocorrelation function of the square of the bimodal distribution (signal defined above as $-K$ and K) would be a constant signal, hence the kurtosis should be small (as expected). The autocorrelation function of the (squared) constant signal would be a constant signal of zero amplitude owing to the subtraction of the mean value. As this autocorrelation function is somewhat similar to a constant autocorrelation function with very small amplitude (for which the kurtosis should be small), the definition of one makes sense for the kurtosis of the constant signal. However, for an impulse signal the autocorrelation function of the (squared) impulse signal would also be an impulse signal. Thus the kurtosis would be expected to be large (shifting the squared signal results in a significant change).

A further anomaly seems to be that the kurtosis of a normal Gaussian distribution with standard deviation σ is defined as 3. When the standard deviation σ tends to zero, the distribution becomes increasingly unimodal (with a high kurtosis value expected). Yet, the

kurtosis of the normal Gaussian distribution is not a function of σ . This investigation is left for future work.

6.9.4 Kurtosis of periodic signal

For a periodic signal, the kurtosis of the signal may be computed by starting with the definition of kurtosis for a non-periodic signal with buffer length N

$$\begin{aligned}
 kurt(x) &= \frac{\sum_{x=-\infty}^{\infty} (x - \mu)^4 / N}{\left(\sum_{x=-\infty}^{\infty} (x - \mu)^2 / N \right)^2} \\
 &= N \frac{\sum_{x=-\infty}^{\infty} (x - \mu)^4}{\left(\sum_{x=-\infty}^{\infty} (x - \mu)^2 \right)^2} \\
 &= N \frac{A_2}{B_2^2}
 \end{aligned} \tag{97}$$

Now assume the signal x repeats once and let us define the new signal as x_2 (hence the buffer length becomes $2N$)

$$\begin{aligned}
kurt(x_2) &= \frac{\sum_{x=-\infty}^{\infty} (x_2 - \mu)^4 / 2N}{\left(\sum_{x=-\infty}^{\infty} (x_2 - \mu)^2 / 2N \right)^2} \\
&= 2N \frac{\sum_{x=-\infty}^{\infty} (x_2 - \mu)^4}{\left(\sum_{x=-\infty}^{\infty} (x_2 - \mu)^2 \right)^2} \\
&= 2N \frac{2A_2}{(2B_2)^2} \\
&= N \frac{A_2}{B_2^2}
\end{aligned} \tag{98}$$

It is clear that the kurtosis statistic is unchanged by the repetition of the signal x . In general with k repetitions of the signal x , resulting in the signal x_k (where the buffer length is kN), the kurtosis would be

$$\begin{aligned}
kurt(x_k) &= \frac{\sum_{x=-\infty}^{\infty} (x_k - \mu)^4 / kN}{\left(\sum_{x=-\infty}^{\infty} (x_k - \mu)^2 / kN \right)^2} \\
&= kN \frac{\sum_{x=-\infty}^{\infty} (x_k - \mu)^4}{\left(\sum_{x=-\infty}^{\infty} (x_k - \mu)^2 \right)^2} \\
&= kN \frac{kA_2}{(kB_2)^2} \\
&= N \frac{A_2}{B_2^2}
\end{aligned} \tag{99}$$

The kurtosis statistic is unchanged by the repetition of the signal x for the general case of k periods of the signal x . Furthermore, it is also not sensitive to the signal amplitude (as was illustrated by the kurtosis of the constant signal). The above characteristics may be an advantage for fault diagnosis ⁷⁶.

6.9.5 Interpretation of kurtosis in time domain

Consider the definition of kurtosis and assume a buffer length of N samples and signal x with zero mean

$$\begin{aligned} kurt(x) &= \frac{\sum_{x=-\infty}^{\infty} (x - \mu)^4 / N}{\left(\sum_{x=-\infty}^{\infty} (x - \mu)^2 / N \right)^2} \\ &= N \frac{\sum_{x=-\infty}^{\infty} (x)^4}{\left(\sum_{x=-\infty}^{\infty} (x)^2 \right)^2} \end{aligned}$$

⁷⁶ The vibration amplitudes would also be very low for very small bearing defects. The kurtosis statistic would amplify this low level signal effectively. Hence it is possible that impulsive signals corresponding to, for example, geometric bearing defects could be mistaken for localized bearing defects. The kurtosis statistic is also a global statistic and not sensitive towards defects at specific frequency (if the defect at a specific frequency happened to dominate the demodulation spectrum, as frequently occurs in large bearing defects, the kurtosis would of course be a suitable metric).

$$\begin{aligned}
&= N \frac{x_0^4 + x_1^4 + \dots + x_k^4}{(x_0^2 + x_1^2 + \dots + x_k^2)^2} \\
&= N \frac{x_0^2 x_0^2 + x_1^2 x_1^2 + \dots + x_k^2 x_k^2}{(x_0^2 + x_1^2 + \dots + x_k^2)(x_0^2 + x_1^2 + \dots + x_k^2)} \\
&= N \frac{y_0 y_0 + y_1 y_1 + \dots + y_k y_k}{(y_0 + y_1 + \dots + y_k)(y_0 + y_1 + \dots + y_k)} \\
&\text{where } y_k = x_k^2 \text{ and } R_y(\tau) \text{ is the autocorrelation of } y_k \\
&= N \frac{R_y(0)}{R_y(0) + 2(R_y(1) + R_y(2) + \dots + R_y(k))} \\
&= N \frac{R_y(0)}{\sum_{\tau=-\infty}^{\infty} R_y(\tau)} \\
&= N \frac{1}{1 + \sum_{\tau \neq 0} R_y(\tau)/R_y(0)} \tag{100}
\end{aligned}$$

The above equation may be interpreted as follows. When the autocorrelation of the squared signal is peaked at the origin (for example an impulse at $\tau=0$) then $\sum R_y(\tau) = 0$ and the kurtosis $kurt = N$ is maximized as discussed previously. An impulse in the time domain is one example of a signal that would give an impulse in the autocorrelation function.

However, when the autocorrelation of the squared signal is 'flat' in the sense that a shift in the squared signal makes no difference to the value of the autocorrelation (so the autocorrelation is a constant K for all τ), then $R_y(0)/\sum R_y(\tau)$ (for $\tau \neq 0$) is $1/N$ and the kurtosis is one as discussed previously.

It is well known for the signal x that the autocorrelation function is related via the Fourier transform to the power spectrum (for example [68]). A similar relation of the autocorrelation function to the power spectrum could be anticipated for the squared signal x .

As the kurtosis measures the 'spikiness' of a signal it is often used as a fault diagnostic. The assumption is that a localized fault causes impulsive behaviour in the time domain signal, which can often be quantified by the kurtosis statistic. However, the kurtosis might not be a good description for low level signals, as the very low level signals are greatly amplified by the computation of the kurtosis. Any localized geometric bearing defect could also display impulsive behaviour without necessarily being cause for immediate concern. Hence trending of (absolute) vibration amplitude levels remains an important diagnostic tool, provided that enough background information can be ascertained on the factors related to the measured vibration levels (for example operating speed, static and dynamic load, bearing condition, lubrication levels). The advantage of the kurtosis statistic is that it is normalized, no prior knowledge about defect frequencies is required and it is sensitive towards impulsive behaviour (frequently associated with localized bearing faults). The disadvantages include lower sensitivity compared with other techniques (for example trending of the level of a spectral component in the demodulation spectrum) and perhaps lower sensitivity to very small incipient faults of very low amplitude.

6.9.6 Interpretation of kurtosis in frequency domain

$$\begin{aligned}
 kurt(x) &= \frac{\sum_{x=-\infty}^{\infty} (x - \mu)^4 / N}{\left(\sum_{x=-\infty}^{\infty} (x - \mu)^2 / N \right)^2} \\
 &= N \frac{X_4(0)}{(X_2(0))^2} \\
 \text{where } X_2(0) &= \frac{1}{N} \sum_{k=-\infty}^{\infty} |X(k)|^2 \text{ and } X(k) \text{ is the FFT of } x \\
 \text{and } X_4(0) &= \frac{1}{N} \sum_{k=-\infty}^{\infty} |X_2(k)|^2 \text{ and } X_2(k) \text{ is the FFT of } x^2 \\
 &= N \frac{\frac{1}{N} \sum_{k=-\infty}^{\infty} |X_2(k)|^2}{\left(\frac{1}{N} \sum_{k=-\infty}^{\infty} |X(k)|^2 \right)^2} \\
 &= \frac{\sum_{k=-\infty}^{\infty} |X_2(k)|^2}{(X_2(0))^2} = \left(1 + \frac{\sum_{k \neq 0} |X_2(k)|^2}{(X_2(0))^2} \right) \tag{101}
 \end{aligned}$$

For a constant signal $y(k) = A$ the FFT of $x(k)^2$ is $X_2(k) = 0$ (where $x = y - \text{mean}(y)$), resulting in the problematic limit of zero over zero. However, if the limit of $\epsilon \rightarrow 0$ $X_2(0) = \epsilon$ is evaluated, the result is one (which confirms the derivation of section 6.9.3 on p. 149). For an impulse signal $x = \delta(n)$ the FFT of $x(k)^2$ is $X_2(k) = 1$ (where $x = y - \text{mean}(y)$)⁷⁷. Hence the kurtosis is $\sum X_2(k)^2 / X_2(0)^2 = N$ (which confirms the derivation of section 6.9.2 on p. 145).

It should be noted that the above equation in the frequency domain bears a striking resemblance to the previous equation 100 in section 6.9.5 (assuming a zero mean signal x).

⁷⁷ The Matlab convention for scaling factors of the FFT/IFFT is applied (only the IFFT is scaled by a factor of $1/N$).

Table 4: Relation between time domain, autocorrelation and frequency domain after application of a non-linear operator (squaring of the signal)

Time	Autocorrelation	Frequency
$(\sum x^2)^2$	$R_x(0)^2$	$(1/N \sum X(k) ^2)^2$
$\sum u^2 = \sum x^4$ (where $u = x^2$)	$R_u(0)$	$1/N \sum X_2(k) ^2$ (where $X_2(k) = \text{FFT}(x^2)$)

The Wiener-Khintchine theorem (connecting the power spectral density and the autocorrelation function) relates the denominator $(\sum |X(k)|^2)^2$ to $(\sum R_x(\tau))^2$. This is simply an energy relationship that can, of course, be expressed in the time domain, as an autocorrelation, and in the frequency domain. A similar relation exists in the numerator, but with the squared time domain signal. The relation is between $R_u(0)$ and $1/N \sum |X_2(k)|^2$ (where $X_2(k) = \text{FFT}(x^2)$). This is also an energy relationship (in the time domain, frequency domain and as an autocorrelation function) but using the squared time domain signal. In the frequency domain the FFT of the squared signal results in the demodulated spectrum $X_2(k)$.

The kurtosis can then simply be interpreted as the ratio of the energy in the demodulated spectrum $X_2(k)$ versus the square of the energy in the regular spectrum $X(k)$ multiplied by a constant N^2 . Alternatively, it may be interpreted as the ratio of the energy in the demodulated spectrum $X_2(k)$ versus the energy in the first component of the demodulated spectrum $X_2(0)$. It is clear that $X_2(0)$ is related to the energy in the regular spectrum $X(k)$ multiplied by a constant $(1/N)$.

The kurtosis may also be expressed in terms of the autocorrelation function as the ratio of the first autocorrelation coefficient (of the squared signal) $R_u(0)$ to the squared energy of the signal (in the first autocorrelation coefficient of the signal) $R_u(0)^2 = \sum R_u(\tau)$.

The differences are as follows:

- The $X_2(0)$ term is in the denominator in the frequency domain, whereas the $R_y(0)$ term of equation 100 is in the numerator. The same comment applies to the numerator of $X_2(k)$ and the denominator of $R_y(k)$.
- The $X_2(0)$ term is squared (computing the power in $X_2(k)$ in the frequency domain, whereas the $R_y(0)$ term is not. However, $R_y(k)$ actually computes the autocorrelation of the squared signal x^2 , hence the terms are comparable to a certain extent.
- An extra factor of N appears in the autocorrelation of the squared signal x^2 (owing to the inverse FFT in $1/N \sum |X_2(k)|^2$).

The above discussion illustrates that the demodulation spectrum $X_2(k)$ and autocorrelation function $R_y(\tau)$ (where $y = x^2$) may be regarded as extensions of the autocorrelation function $R_x(\tau)$ and power spectrum $X_2(f)$ of $R_x(\tau)$ based on higher (second) order signal processing.

6.9.7 Additional interpretation of kurtosis

In addition to the standard interpretation of the kurtosis in terms of the ‘spikiness’ or ‘Gaussianity’ of a signal, the following interpretations are suggested:

- In the **time domain** the kurtosis is a measure of the ‘flatness’ of the autocorrelation of the squared signal. A constant signal has a flat autocorrelation function, thus the ratio $R_y(0)/\sum R_y(\tau)$ is small ($1/N$ where N is the buffer length) and the kurtosis is small ($N/N = 1$). By contrast, an impulsive signal has a very peaked autocorrelation function, hence the ratio $R_y(0)/\sum R_y(\tau)$ is one (large) and the kurtosis is large (N). For a Gaussian distributed noise signal the autocorrelation function of the squared signal is also flat for $\tau > 0$, but $R_y(0)$ is in general substantially larger than $R_y(\tau)$. Another way of interpreting the kurtosis would be that it provides a measure of the change resulting from a shift in the squared (‘envelope’) signal, as the kurtosis can also be written as $N/(1 + \sum R_y(\tau)/R_y(0))$, where $\tau \neq 0$. If no change results from a shift of the squared signal ($\sum R_y(\tau) = (N-1)R_y(0)$ where $\tau \neq 0$), the kurtosis is one, whereas a maximum change ($\sum R_y(\tau) = 0$ where $\tau \neq 0$) corresponds to a kurtosis of N . This is equivalent to interpreting the kurtosis as a measure of the ‘spread in time’ of a signal. Well localized

signals (such as impulse signals) would have a peaked autocorrelation function, hence high kurtosis ($\sum R_y(\tau) = 0$ – where $\tau \neq 0$). Flat signals, such as a bimodal distribution and the constant signal, are distributed globally and would tend to have low kurtosis values.

- In the **frequency domain** the kurtosis is a measure of the ‘flatness’ of the (power) spectrum of the squared signal. A constant signal in the time domain has an impulse in the demodulated power spectrum, thus the ratio $\sum X_2(k)^2/X_2(0)^2$ is one and the kurtosis is small (one). An impulsive signal in the time domain has a constant demodulated power spectrum, consequently the ratio $\sum X_2(k)^2/X_2(0)^2$ is large (N) and the kurtosis is N (large). For a Gaussian noise signal the power spectrum of the squared signal is given by the spectrum of the convolution of the noise signal with itself. When applying the convolution both the amplitude and (random) phase should be considered (in [2] it would seem as though only the amplitude function of the spectrum has been considered). Even if the original Gaussian noise signal has a constant (white) spectrum, it is unlikely that the power spectrum of the squared signal would also have a constant spectrum. This might help to explain the kurtosis value of around 3 for the Gaussian noise signal.

The kurtosis may be interpreted as the ratio of the energy in the demodulation spectrum relative to the regular spectrum. As indicated in section 6.8 on p. 139, the demodulation spectrum may also be interpreted in the time domain as a modulation by $\exp(j\omega_m t)$ (bandpass filtering). Thus a specific component of the demodulation spectrum is determined by the degree of correlation between the original signal and all the bandpass filtered signals. Fault signals that repeat periodically supply a delta impulse sequence in the demodulation spectrum as well as the regular spectrum. However, the demodulation spectrum may reduce the effect of amplitude and speed variations in the fault signal. Hence the kurtosis, measuring the energy in the demodulation spectrum, may be a suitable fault indicator (at least for large bearing defects with clearly defined periodically repeating fault signals). The kurtosis is given by the total energy in the demodulation spectrum (spectrum of the squared signal after bandpass filtering) relative to the total energy in the

regular spectrum. For small bearing defects, in which no periodic fault signal is clearly visible, the kurtosis might no longer be a suitable indicator (this is discussed in more detail in chapter 7 on p. 213). This is primarily owing to the poor performance of the demodulation spectrum in very low signal-to-noise ratios (although the defect frequency component of the demodulation spectrum is still suitable).

An alternative interpretation of the kurtosis in the frequency domain is that it provides a measure of the change resulting in a shift (in frequency) of the spectrum of the signal. Each frequency component in $X_2(f)$, where $f \neq 0$, provides an indication of the influence of a shift in frequency of the spectrum of the signal. In the time domain this frequency shift (delay in the frequency domain) corresponds to a modulation of the time domain signal. The subsequent summation in the time domain corresponds to the component at DC (0Hz) in the frequency domain. Hence the composite operation of frequency shifting and integration, implicit in the frequency domain convolution, may be viewed as evaluating the output of a bandpass filter. In the time domain the original signal is convolved with the signal modulated around the frequency shift of $\omega_m = \omega_t$ and the component at $t = 0$ is retained (in the frequency domain the spectrum of the signal and the shifted spectrum of the signal are multiplied and then summed, after the inverse FFT to the time domain the multiplication changes to a convolution, the frequency shift to a modulation by ω_m and the summation in the frequency domain corresponds to the component at $t = 0$). When $\omega_m = 0$ no frequency shift occurs and the energy in the signal is computed. Otherwise the correlation between the signal and a modulated version (with centre frequency at the frequency shift ω_m) of the signal is computed for all frequency shifts ω_m from 0 to ∞ . Alternatively, the operation in the time domain may be visualized simply as the summation of the signal squared modulated by a frequency ω_m .

According to the DFT filter bank interpretation, the signal is modulated by a complex exponential $\exp(j\omega_m t)$ in the time domain and then integrated. The frequency domain interpretation is a shift in frequency by ω_m and the summation in the time domain is performed by retaining the frequency component at DC (0Hz). In analogy with the DFT filter bank interpretation, the frequency domain interpretation of the component in the

demodulated spectrum is that of a summation of the spectrum of the signal multiplied by the spectrum shifted by ω_m . The simple time domain interpretation of the component in the demodulated spectrum is that of a bandpass filter at ω_m applied to the squared signal and summation of the resulting signal.

6.9.8 Kurtosis at specific frequency

From the preceding discussion it is clear that the kurtosis statistic provides a global metric indicating the influence of a shift in the time (the autocorrelation of the squared signal after filtering) or frequency domains (the demodulation spectrum resulting from the spectrum of the squared signal after filtering). It may be assumed that the component associated with the (large) bearing defect would dominate in, for example, the demodulation spectrum. Thus the kurtosis statistic would essentially measure the contribution of the large bearing defect, as most of the energy would be concentrated in the bearing defect component and harmonics; hence it would be a suitable metric with the added advantage that no prior knowledge is required of bearing geometry.

However, in a real-world situation the focus is frequently on detecting very small incipient bearing defects. The bearing defects could also have to propagate through a very complex mechanical structure in order to reach the sensor. Furthermore, additional fault conditions such as imbalance, misalignment and gear meshing could mask the low level signal to some degree. For this reason, trending of frequency components at a specific frequency is often employed in industrial settings. The disadvantage of this strategy is that the trending is specific to a particular machine (hence no generalization can be expected) and that environmental conditions such as speed, load and geometry could cause misdiagnosis.

As mentioned in the literature survey on p. 25 the normalized ratio of the demodulation peak in the demodulation spectrum relative to the carpet level can provide a normalized diagnostic. An alternative is a normalized metric, based on the ratio of the average value of the peak (and harmonics of) the defect frequency (using the HFRT) to the average value of the spectrum (p. 30 in the literature survey). In contrast to the trending of specific frequency components, more than one frequency component is considered in this normalized metric. This metric may be visualized simply as the application of a comb

filter to the demodulated spectrum, removing all the energy in the signal except those at multiples of the trigger (defect) frequency. Hence the relation to the ALE technique is clear. Applying the kurtosis statistic to the signal resulting from the application of a comb filter to the demodulated spectrum should consequently be of value in providing diagnostic information at a specific defect frequency. As the kurtosis statistic is normalized it could provide valuable information on a variety of machinery. This is left for future work (section 8.6 on p. 263).

6.9.9 Conclusion

It is postulated that the minimization (maximum platykurtic and minimum leptokurtic) and maximization (maximum leptokurtic and minimum platykurtic) of kurtosis are achieved by a signal (such as a signal with bimodal distribution) where the autocorrelation function of the squared signal is a constant and an impulse signal respectively, with kurtosis values of one and N respectively. The kurtosis of a periodic signal is the same as the kurtosis of the basic pulse shape, which is a desirable characteristic for fault diagnosis. The relation between the demodulated spectrum of the squared signal and the autocorrelation function of the squared signal is illustrated in the context of the computation of the kurtosis. Finally, it is suggested that the normalized kurtosis statistic should be applied at a specific bearing defect frequency in order to improve the sensitivity of the kurtosis statistic.

6.10 Effect of a non-linear operator on a simulated fault signal

6.10.1 Introduction

The demodulation spectrum, which is obtained using HFRT, consists of the operation of a bandpass filter followed by a non-linear operator (such as squaring the time domain signal). The non-linearity complicates mathematical analysis on the effect of noise (additive noise, amplitude and speed variations). The aim of this section is simply to present the effect of a non-linear operator, such as a squaring operation, on a simulated fault signal in the presence of speed variations and additive noise in a few examples. In section 6.10.2 some theory is presented relating the non-linear operator, SA, autocorrelation and kurtosis. In section 6.10.3 the effect of additive noise on the non-linear

operator is considered and a heuristic rule derived. Comparison is made between the demodulated spectrum and existing results on ALE detectors (with and without non-linearities). On the one hand, it is known that additive noise deteriorates the performance of the ALE detector. On the other hand, it is known that the demodulation spectrum is effective in reducing the effect of speed variations. In section 6.10.5 the effect of amplitude variations (owing to instantaneous variations in loading on the roller) on the non-linear operator is considered. In sections 6.10.6 and 6.10.7 the relation is shown to existing signal processing communications theory.

6.10.2 Theory

Define the signal x_k , where k is an integer from 0 to N . The effect of the operator $(\Sigma x)^2$ is

$$\begin{aligned}
 & (x_0 + x_1 + x_2 + x_3 + \dots)^2 \\
 &= (x_0^2 + x_1^2 + x_2^2 + x_3^2 + \dots) + x_0(x_1 + x_2 + x_3 + \dots) \\
 &+ x_1(x_0 + x_2 + x_3 + \dots) + x_2(x_0 + x_1 + x_3 + \dots) \\
 &+ x_3(x_0 + x_1 + x_2 + \dots) + \dots \\
 &= (x_0^2 + x_1^2 + x_2^2 + x_3^2 + \dots) \\
 &+ 2(x_0x_1 + x_1x_2 + x_2x_3 + \dots x_0x_2 + x_1x_3 + \dots x_0x_3 + \dots)
 \end{aligned} \tag{102}$$

Define the autocorrelation $R_x(\tau)$ of the signal x

$$R_x(\tau) = \sum_{\tau=-\infty}^{\infty} x_k x_{k+\tau} \tag{103}$$

Hence the effect of the operator $(\Sigma x)^2$ in relation to the autocorrelation $R_x(\tau)$ is

$$\begin{aligned}
 & (x_0 + x_1 + x_2 + x_3 + \dots)^2 \\
 &= (x_0^2 + x_1^2 + x_2^2 + x_3^2 + \dots) \\
 &+ 2(x_0x_1 + x_1x_2 + x_2x_3 + \dots x_0x_2 + x_1x_3 + \dots x_0x_3 + \dots) \\
 &= R_x(0) + R_x(1) + R_x(2) + \dots \\
 &= \sum_{k=-\infty}^{\infty} R_x(k)
 \end{aligned} \tag{104}$$

The term $R_x(0)$ is the sum of squares (Σx^2) , or variance, of the signal x . The operator $(\Sigma x)^2$, or the square of the sum of the signal x , may be calculated by obtaining the area underneath the autocorrelation function of x . If the signal x happened to be a constant $x_k = A$, the variance would be zero and the ratio $(\Sigma x)^2/(\Sigma x^2)$ would then tend to infinity.

In the context of signal averaging, the synchronous average is defined by McFadden *et al.* [53] as

$$\begin{aligned} u &= [u_0 \ u_1 \ u_2 \ \dots u_r] = \text{SA}(y, T) \\ &= [(y_0 + y_m + y_{2m} + y_{3m} + \dots)(y_1 + y_{1+m} + y_{1+2m} + y_{1+3m} + \dots) \dots (y_r + y_{r+m} + \dots)] \\ \text{hence } u(r) &= \sum_{k=0}^{\infty} y_{r+km} \end{aligned} \quad (105)$$

where r is an integer denoting the sample r of the synchronous average of the signal y over a period m (where $m = TF_s$ – where T refers to the repetition period of the synchronous average and F_s refers to the sampling frequency). Let us examine the square of the synchronous average of the signal y , $(\text{SA}(y, T))^2$

$$\begin{aligned} u^2 &= [u_0 \ u_1 \ u_2 \ \dots u_r]^2 = (\text{SA}(y, T))^2 \\ u(r)^2 &= \left(\frac{1}{N} \sum_{k=0}^N y_{r+km} \right)^2 \\ &= u_r^2 \\ \left(\sum_{r=0}^{\infty} u_r \right)^2 &= \frac{1}{N^2} \sum_{\tau=-\infty}^{\infty} R_u(\tau) \end{aligned} \quad (106)$$

Now $\Sigma(\text{SA}(y, T)^2) = R_u(0)$ is the mean square value of the signal

$$\begin{aligned} &\text{SA}(y^2, T) \\ &= \frac{1}{N} \sum_{k=0}^N y_{r+km}^2 \\ &= q_r^2 \\ \sum_{r=0}^{\infty} q_r^2 &= \frac{1}{N} R_u(0) \end{aligned} \quad (107)$$

The ratio of $((\Sigma u_r)^2)/(\Sigma q_r^2) = (\Sigma SA(y, T))^2/\Sigma(SA(y, T)^2) = 1/N\Sigma R_u(k)/R_u(0)$ is large when the mean of squares (MOS) $R_u(0)$ is small. If the signal y is exactly periodic with a period of m samples, then the synchronous average u_r would equal one period of the signal y , or equivalently $u_r = y_r$. Hence the MOS of a specific synchronous average u_r would be given by y_r^2 . In general, it is evident that the standard deviation σ_x of a signal x is given by $\text{var}(x) = \sigma_x^2 = \text{MOS}(x) - \text{mean}(x)^2$, where $\text{var}(x)$ is the variance of x with $\text{var}(x) = \sigma_x^2 = \Sigma(x - \text{mean}(x))^2/N$, N is the number of samples in x , $\text{mean}(x)$ is the mean value of x and $\text{MOS}(x)$ is the mean of x squared. Thus $\sigma_r^2 = \text{MOS}(y_{r+km}) - \text{mean}(y_{r+km})^2$ - where $m = T * F_s$, T is the repetition period of the synchronous average, F_s is the sampling frequency and k is an integer from 0 to infinity. For a periodic signal $\sigma_r = 0$ as $\sigma_r^2 = \Sigma y_r^2/N - (\Sigma y_r/N)^2 = y_r^2 - y_r^2 = 0$. But the MOS of a synchronous average u_r (and the signal y) is of course not zero (unless $y=0$). Hence the ratio of $\text{MOS}(y_{r+km})/\text{mean}(y_{r+km})^2 = 1$. The total MOS of the periodic signal y is given by Σy_r^2 . If the signal y is no longer periodic (perhaps owing to amplitude or speed variations), the MOS is given by $q_r^2 = \Sigma y_{r+km}^2$ and the total MOS by $\Sigma q_r^2 = \Sigma y_r^2$. The standard deviation σ_r is given by $\sigma_r^2 = \text{var}(y) = \text{MOS}(y) - \text{mean}(y)^2$. With speed variations present the samples of a specific synchronous average would no longer be the same (with no speed variations the samples would be a constant signal with minimal kurtosis). Thus the kurtosis of the samples of a specific SA could be expected to be high (showing impulsive behaviour). The kurtosis of each SA may also be a useful performance metric (in addition to other statistical moments) to establish the statistical variation inherent in the samples of the SA. Define the synchronous kurtosis as

$$\begin{aligned}
 \text{Let } u &= [u_0 \ u_1 \ u_2 \ \dots \ u_r] = SA(y, T) \\
 &= [(y_0 + y_m + y_{2m} + y_{3m} + \dots)(y_1 + y_{1+m} + y_{1+2m} + y_{1+3m} + \dots) \dots (y_r + y_{r+m} + \dots)] \\
 k(x) &= \frac{\frac{1}{N} \sum_{n=0}^N (x(n) - \bar{x})^4}{\left(\frac{1}{N} \sum_{n=0}^N (x(n) - \bar{x})^2 \right)^2} \tag{108} \\
 k_s &= [k_0, k_1, \dots, k_r]
 \end{aligned}$$

where $\mathbf{k}_s$ is the synchronous kurtosis ($\mathbf{k}_s$ is a vector with the number of entries corresponding to the number of samples $m = TF_s$ of the SA), $k_r = k(y_r + nm)$ where n is an integer, $m = TF_s$ (T is the period of the SA and F_s is the sampling frequency in Hz) and $k(x)$ is the kurtosis function.

The synchronous kurtosis supplies additional information about the higher order statistics of the samples of a specific SA. However, the kurtosis may also be written as the ratio of a second order moment over the square of the first order moment (instead of the 4th order moment over the square of the second order moment)

$$kurt(x) = \frac{\frac{1}{N} \sum_{n=0}^N (x(n) - \bar{x})^4}{\left(\frac{1}{N} \sum_{n=0}^N (x(n) - \bar{x})^2 \right)^2}$$

where

$$y(n) = (x(n) - \bar{x})^2$$

$$kurt(x) = \frac{\frac{1}{N} \sum_{n=0}^N (y(n))^2}{\left(\frac{1}{N} \sum_{n=0}^N y(n) \right)^2}$$

$$= N \frac{\sum_{n=0}^N (y(n))^2}{\left(\sum_{n=0}^N y(n) \right)^2} \quad (109)$$

by using the substitution of $y(n) = (x(n) - \text{mean}(x(n)))^2$. The kurtosis may therefore be interpreted as the ratio of $\text{MOS}(y)$ over the square of $\text{mean}(y)$, where $y = (x - \text{mean}(x))^2$. Care should of course be taken in the computation of the kurtosis when the $\text{mean}(y)$ is close to zero (which is likely when the amplitude of the signal y is very low). Equivalently the kurtosis may be written as the ratio of $\text{var}(y) + \text{mean}(y)^2$ over $\text{mean}(y)^2$, or $kurt = 1 + \text{var}(y)/\text{mean}(y)^2$.

The ratio of $(\sum u_r)^2 / \sum q_r^2$ may also be related to the kurtosis statistic of section 6.9 on p. 145. The kurtosis $kurt = k = (\sum g^4 / N) / (\sum g^2 / N)^2$ where g is a zero mean signal. The kurtosis may also be written as $kurt = k = (\sum u^2 / N) / (\sum u / N)^2 = N(\sum u^2) / (\sum u)^2$ where $u = g^2$. Hence $(\sum u_r)^2 / \sum q_r^2 = N/k$. If $g^2 = Kg$, where K is a constant, the kurtosis would be unchanged. Consider the signals that minimize and maximize kurtosis respectively:

- For a constant signal $x = A$ the kurtosis is one (refer to section 6.9.3). But $(\sum b^2 / N) / (\sum b / N)^2 = N/N = 1$ where $b = x - \text{mean}(x)$. Hence $(\sum u_r)^2 / \sum q_r^2 = N/k = N$ and $1/N(\sum u_r)^2 / \sum q_r^2 = 1$.
- For an impulse signal $x = A\delta(n)$ the kurtosis is N for large N (refer to section 6.9.2). However $(\sum b^2 / N) / (\sum b / N)^2 = N$ where $b = x - \text{mean}(x)$. But $(\sum u_r)^2 / \sum q_r^2 = N/k = 1$ and $1/N(\sum u_r)^2 / \sum q_r^2 = 1/N$.

A few examples, using a simulated bearing fault signal as described in equation 124 on p. 199, are demonstrated in appendix K.

Note that the time domain signal becomes totally obscured by the noise, yet the signal average using the non-linear operator of squaring still shows the outline of the pulse sequence clearly, in contrast with the square of the (normal) signal average, which does not show the pulse sequence when speed variation of 20% and noise are added. Also note that the addition of speed variation (without adding noise) increases the MOS, hence the ratio $u_r^2 / q_r^2 = (\sum x)^2 / (\sum x^2)$ becomes smaller. Thus it becomes more difficult to distinguish the pulse sequence when noise is also added. The standard deviation σ (STD) of the simulated fault signal is about 0.01.

The above results seem to correlate well with literature. The normal SA can be expected to perform well with additive noise, but not with speed variations (which would smear the spectrum). The non-linear operator (such as squaring) can be expected to perform well in

removing speed variations (when the frequency component at the defect frequency is examined), but not necessarily in removing additive noise in the signal⁷⁸.

From figure 21 it is clear that any speed variation spreads the time domain signal (compare with figure 242, for example).

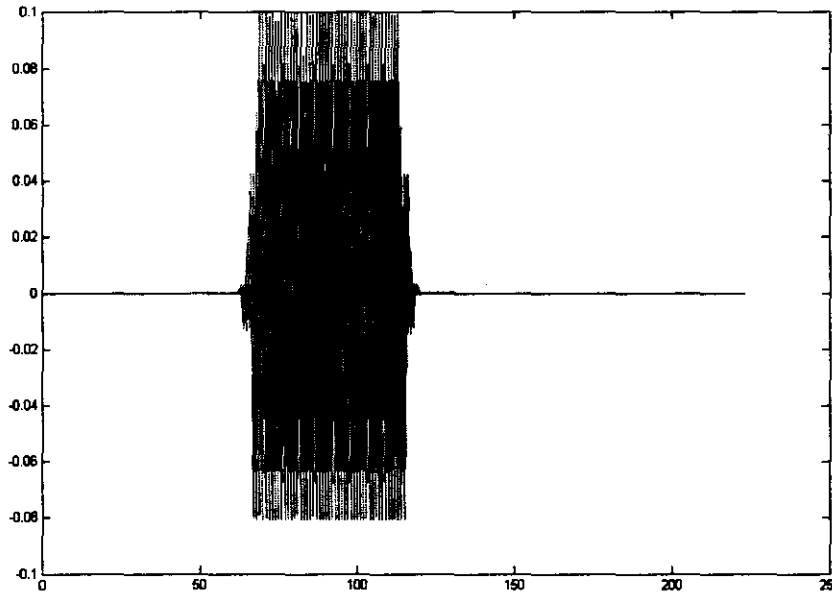


**Figure 21: Time domain signals plotted over each other for one period
(STD of noise = 0, speed variation = 20 %)**

Hence the synchronous average $SA(y(r))$, at a particular value of r , will in general have contributions from many other (unrelated) samples. This may be viewed as a noise source and increases with increasing speed variation. If a once per revolution trigger signal is available then the noise from speed variation may be reduced substantially. A signal from a shaft encoder would be even better, but is unfortunately often not available. One possibility, based on this thesis, is to deliberately introduce a (known, calibrated) discontinuity (i.e. a small defect) in the surface of the outer race of the bearing. This may then be used as a (once per revolution) trigger signal, and as an estimation of the influence

⁷⁸ The demodulation spectrum with a non-linear operator also computes a kind of SA (which should help to remove noise) but with the square of the noise signal. Hence the normal SA could be more effective in removing purely additive noise, whereas the demodulation spectrum would possibly be better in removing amplitude and speed variations.

of loading. The discontinuity can be placed out of the loading zone for normal operation. It is even possible to consider introducing multiple discontinuities on the surface of the outer race of the bearing to supply a more continuous trigger signal. This resembles a geometrical bearing defect. The potential advantages of this technique are as follows:

1. No mechanical modifications are required to the assembly. However, physical access to the bearing is required to introduce the defect, as well as to shift the defect out of the loading zone. In future designs the possibility of automatically shifting the defect by remote control may be investigated.
2. Calibration of the vibration monitoring equipment may be easily performed, as a known defect of specified dimensions is available. The performance of handheld accelerometers (e.g. to determine optimal placement of transducers in a noisy environment) and current and flux measurements may be evaluated objectively in an industrial plant, without requiring complex FEM models to be built.
3. A trigger signal for accurate computation of, for example, synchronous averages may be readily obtained, provided that the defect is small enough. As mentioned previously, multiple discontinuities (defects) may provide even more accurate tracking, but the effect needs to be investigated on the lifetime of the bearing (especially the influence of the position of the defect relative to the loading zone).
4. By introducing a defect of size that would normally be unacceptable in practice, it is easy to estimate from the vibration signals when the actual defects on the bearing surface exceed this size. The analytical prediction of bearing defect growth (bearing prognostics, for example [6]) is very difficult.

Hence this empirical technique may be of benefit in estimating the size of the defect and in preventing further damage to equipment⁷⁹.

6.10.3 Frequency domain interpretation

6.10.3.1 Derivation of an analytical frequency domain expression for non-linear operator with additive noise is problematic

Multiplying a real-time domain signal with itself corresponds to a convolution of the corresponding spectrum⁸⁰ [for example refer to McFadden *et al.* in 50, 53]. If no speed variation is present and the signal is exactly periodic with period T , then spectral lines will be found at multiples of the trigger frequency $F_t = 1/T$. Speed variation will cause the signal to be non-periodic. Let us assume for the moment that the speed variation is in the form of a decreasing linear chirp for simplicity (i.e. the period decreases linearly with time, or the frequency increases). The instantaneous spectral lines⁸¹ would then shift to the left as time increases. If the speed variation is large enough, the spectral lines (of the Fourier analysis performed over the complete buffer) at higher frequency would then start to merge (or smear together), so the total spectrum no longer consists of harmonic lines spaced at a single trigger frequency.

If the speed variation occurs randomly, the effect on the total spectrum will still be the same, resulting in a smearing of the frequency components. The effect of the non-linear operator, such as squaring, is to convolve the spectrum with itself. Hence the demodulation peak (or first component of the ESFFT) consists of the product of all frequency components spaced by the defect frequency (owing to the convolution in the frequency domain as a result of the squared signal): Let $X_2(f)$ be the demodulation spectrum⁸²

⁷⁹ Assuming, of course, that easy physical access to the bearing is possible to rotate the defect applied on the outer race of the bearing.

⁸⁰ As the spectrum is complex in general, the complex conjugate should be used.

⁸¹ Evaluated using, for example, the spectrogram with overlapping windows, or perhaps a time frequency transform or wavelet.

⁸² Complex conjugation should be used if $X(f)$ is complex.

$$X_2(f) = \sum_{k=-\infty}^{\infty} X(k)X(f-k) \quad (110)$$

So if f_d represents the defect frequency, $X_2(f_d)$ represents the demodulation peak (first component of ESFFT) and equation 110 consists of the (sum of the) product of all frequency components $X(k)X(f_d - k)$ spaced by the defect frequency. Thus it can be seen that even with speed variation a reasonably clear demodulation peak $X_2(f_d)$ will still be obtained (owing to the effect of the non-linear operator in convolving the spectrum).

Let the additive Gaussian noise signal of zero mean be represented by $n(t)$ and the Fourier transform by $N(f)$. The Fourier transform of the signal plus noise is given by $N(f) + X(f)$, whereas the Fourier transform of the square of the signal plus noise is given by $X(f) \otimes X(f) + 2 * X(f) \otimes N(f) + N(f) \otimes N(f)$, where the symbol $\otimes$ denotes convolution. The time domain signal is given by $y_1(t) = x(t)^2 + 2x(t)n(t) + n(t)^2$. Assume that the signal and noise are independent, the mean of the signal is zero and that the STD of the noise is A . As the noise distribution is nearly white (constant over frequency), the STD of $n(t)^2$ will be A^2 .

$$\begin{aligned} \text{Hence } \Sigma y_1(t)^2/N &= \Sigma(x(t)^4 + 2(2x(t)^3n(t) + x(t)^2n(t)^2 + 2x(t)n(t)^3) + n(t)^4)/N \\ &= \Sigma(x(t)^4 + 4x(t)^3n(t) + 2x(t)^2n(t)^2 + 4x(t)n(t)^3 + n(t)^4)/N \\ &= (\Sigma x(t)^4 + 4\Sigma(x(t)^3n(t)) + 2\Sigma(x(t)^2n(t)^2) + 4\Sigma(x(t)n(t)^3) + \Sigma n(t)^4)/N \\ &= (\Sigma x(t)^4 + \Sigma n(t)^4 + 2\Sigma(x(t)^2n(t)^2) + 4\Sigma(x(t)^3n(t)) + 4\Sigma(x(t)n(t)^3))/N \end{aligned}$$

If the length of X is N , the variance of Y_1 is $\text{var}(Y_1) = \Sigma y_1(t)^2/N$.

$$\begin{aligned} \text{The mean square value of } y_2(t) &= (x(t) + n(t))^2 \text{ is } (\Sigma y_1(t)/N)^2 \\ &= (\Sigma((x(t) + n(t))^2) / N)^2 \\ &= (\Sigma(x(t)^2 + 2x(t)n(t)n(t)^2) / N)^2 \\ &= (\Sigma(x(t)^2 + 2x(t)n(t) + n(t)^2))^2 / N^2 \\ &= (\Sigma x(t)^2 + 2\Sigma x(t)n(t) + \Sigma n(t)^2)^2 / N^2 \\ &= ((\Sigma x(t)^2)^2 + 2(2\Sigma x(t)n(t)\Sigma x(t)^2 + 2\Sigma x(t)n(t)\Sigma n(t)^2 + \Sigma x(t)^2\Sigma n(t)^2) + (\Sigma n(t)^2)^2) / N^2 \end{aligned} \quad (111)$$

Unfortunately, the above expressions cannot be simplified any further, thus it is not possible to derive the theoretical effect of the non-linear (squaring) operator.

In section 11.11.8.3 on p. 450 (appendix K) a simplified example of a sinusoidal signal with additive noise is considered and the detection ratio is investigated for different noise parameters.

6.10.3.2 Heuristic empirical procedure

However, the following empirical procedure may be employed (to derive an analytical frequency domain expression for a non-linear operator in the presence of additive noise). Measure $R_2(2f_c)$ for a convenient value of the noise standard deviation K_b and call the resulting measurement value $R_{2b}(2f_c)$. If the standard deviation of the noise increases by K , the value of $R_{2b}(2f_c)$ will increase by K^2 . Now $R_{2b}(2f_c)/R_b(f_c) = R_{2b}(2f_c)/(\sqrt{MA_x}/(2K_b))$. Call the desired value $R_2(2f_c)$ when $R_2(2f_c) = 1$. So $R_2(2f_c)/R_b(f_c) = 1/K R_{2b}(2f_c)/R_b(f_c)$. But $R_b(f_c) = \sqrt{MA_x}/(2K_b * K)$. Consequently, $(2K_b * K)/(\sqrt{MA_x}) = (1/K * R_{2b}(2f_c))/(\sqrt{MA_x}/(2K_b))$ or $K^2 = R_{2b}(2f_c)$. Thus $K = \sqrt{R_{2b}(2f_c)}$, so it is possible to compute the correction factor K once a measurement $R_{2b}(2f_c)$ is available. Note that the expression for K does not depend on the value of K_b .

For instance, consider the previous example of $K_b = 2$ and $R_{2b}(2f_c) = 7.1$. Consequently, $K = \sqrt{7.1} = 2.6646$ and the standard deviation of the noise that supplies $R_2(2f_c) = 1$ will be given by $K * K_b = 5.3292$, which is very close to the value of 5 in the previous example. As another example, consider $K_b = 10$ and $R_{2b}(2f_c) = 0.2828$. Hence $K = \sqrt{0.2828} = 0.5318$ and the standard deviation of the noise that supplies $R_2(2f_c) = 1$ will be given by $K * K_b = 5.3179$, which is also very close to the value of 5 in the previous example.

In summary, it is possible to compute a correction factor K to the standard deviation of the noise source, which enables computation of the condition $R_2(2f_c) = 1$, provided that a measurement of $R_2(2f_c)$ is available (and the standard deviation of the noise source K_b is known, of course). The correction factor is given by $KK_b = K_b\sqrt{R_2(2f_c)}$ or $K = \sqrt{R_{2b}(2f_c)}$.

6.10.3.3 Heuristic formula for $N_2(1)$

In order to solve the problems associated with the previous procedure, a heuristic formula for $N_2(1)$, which is the second component of the FFT of the noise signal squared, will now be considered. From simulation results the ratio of $N_2(0)/N_2(1) = K_I$ seems to be related to $K_I = 2\sqrt{M}$. But $N_2(0) = \sum n(t)^2 = (B/2)^2/M$. Hence $N_2(1) = N_2(0)/K_I$

$$= ((B/2)^2/M)/(2\sqrt{M})$$

but $B/2 = M * \text{std}(n)$.

$$\text{Thus } N_2(1) = ((M * \text{std}(n))^2/M)/(2\sqrt{M}).$$

$$N_2(1) = ((M^2 \text{var}(n))/M)/(2\sqrt{M})$$

$$N_2(1) = (M \text{var}(n))/(2\sqrt{M})$$

$$N_2(1) = \frac{1}{2}\sqrt{M} \text{var}(n) \quad (112)$$

In order to explain the heuristic formulation of equation 112, consider the following argument. It is clear from simulation results using the FFT that $N_2(0)^2 = \frac{1}{2} \sum |N_2(f)|^2$, $f \neq 0$, excluding $N_2(0)$. Thus the power of the DC frequency component is half that of the total power under $N_2(f)$, excluding the DC frequency component of course. It is also clear from simulation results using the FFT that $(N_2(0)/2)^2 = N_2(1)^2 M$. Hence if it is assumed that the spectrum $N_2(f)$ is white (which is actually not the case even if $N(f)$ is white), the frequency component $N_2(1)^2$ contains $1/M$ times the value of $(N_2(0)/2)^2$. If the spectrum of $N_2(f)$ is assumed to be white, then a factor of $\sqrt{M}$ for $N_2(1)$ is necessary, which explains the factor of $\sqrt{M}$ in the formula for $N_2(1)$. However, the average value of $|N(f)|$ for $f \neq 0$ (or $\sum |N_2(f)|/M$) is more than the value of $|N_2(1)|$. Simulation results indicate that the factor is 2.5496⁸³. This factor is likely to be related to the spectrum of the white noise signal after squaring, which is no longer white (even if the original spectrum $N(f) = B/2$ is white).

⁸³ The factor also seems reasonably close to the oversampling factor of 2.56 when digitally sampling the signal at 25 600Hz to obtain a 10 000Hz bandwidth.

Hence the values of the mean and standard deviation will no longer be the same as the values without the square law modulator.

As a test of this formula, for $\text{var}(n) = 2^2 = 4$ the value of $N_2(1) = \text{var}(n)\sqrt{M/2} = 4 * 71.554/2 = 143.1084$, which compares very well with the measured value of $N_2(1) = 142.0073$ (the previous average value of $N_2(f)$, excluding $N_2(0)$, is 362.1 – the ratio to $N_2(1)$ is 2.5496). Similarly, for $\text{var}(n) = 5^2$, $N_2(1) = \text{var}(n)\sqrt{M/2} = 894.4272$, which compares very well with the measured value of $N_2(1) = 887.5457$ (the previous average value of $N_2(f)$, excluding $N_2(0)$, is 2262.9 – the ratio to $N_2(1)$ is 2.5496). For $\text{var}(n) = 10^2$, $N_2(1) = \text{var}(n)\sqrt{M/2} = 3577.7$, which compares very well with the measured value of $N_2(1) = 3550.2$ (the previous average value of $N_2(f)$, excluding $N_2(0)$, is 9051.6 – the ratio to $N_2(1)$ is 2.5496).

6.10.3.4 Heuristic formula for $N_2(2f_c)$

Based on the previous heuristic arguments, the following heuristic formula for $N_2(2f_c)$ is proposed:

$$N_2(2f_c)N_2(1)2.5496 = 1.2748(\text{var}(n)\sqrt{M}) \quad (113)$$

Equation 113 shows that increasing the standard deviation of the noise by K increases the value of $N_2(2f_c)$ by K^2 , confirming the earlier predictions. In order to double the value of $N_2(2f_c)$ it is necessary to use four times the number of samples, according to equation 113. The reason for the heuristic scaling factor of 2.5496 in equation 113 is not entirely clear at present.

$$R(f_c) = X(f_c)/N(f_c) = (A/2)/(B/2) = (MA_x/2)/(\sqrt{M} * \text{std}(n)) = \sqrt{MA_x}/(2\text{std}(n))$$

$$N_2(2f_c) = 1.2748\sqrt{M\text{var}(n)}$$

$$X_2(2f_c) = A_x^2 M/4$$

$$R_2(2f_c) = X_2(2f_c)/N_2(2f_c) = (1/4)/1.2748 * A_x^2 \sqrt{M/\text{var}(n)}$$

$$= 0.1961 A_x^2 \sqrt{M/\text{var}(n)}$$

$$R(f_c) = X(f_c)/N(f_c) = 1/2 \sqrt{MA_x}/\text{std}(n) \quad (114)$$

Hence the overall detection ratio is $R_2(2f_c)/R(f_c) = 0.3922A_x/\text{std}(n)$. The detection ratio will be the same when $\text{std}(n)/A_x = 0.3922$. Application of the square law modulator can potentially provide an increase when $R_2(2f_c)/R(f_c) > 1$ or $\text{std}(n)/A_x < 0.3922$ (alternatively in a high signal-to-noise ratio). However, in low signal-to-noise ratios ($\text{std}(n)/A_x > 0.3922$) the detection ratio $R_2(2f_c)/R(f_c)$ will be smaller than one (A_x is the signal amplitude and $\text{std}(n)$ is the standard deviation of the noise signal). From simulation results $R(f_c) = 91.231$ and $R_2(2f_c) = 91.951$ for $\text{std}(n) = 0.5546$ and $A_x = \sqrt{2}$, which is in agreement.

The condition $R_2(2f_c) = 1$, implying that it is no longer possible to distinguish the signal and noise after the square law modulator, will be reached when

$$R_2(2f_c) = X_2(2f_c)/N_2(2f_c) = 0.1961 A_x^2 \sqrt{M}/\text{var}(n) = 1$$

or $\text{var}(n) = 0.1961 A_x^2 \sqrt{M}$. Hence $R_2(2f_c) = 1$ implies $\text{std}(n) = 0.4428A_x \sqrt{\sqrt{M}} = 0.4428A_x * M^{1/4} = 5.2975$ ($A_x = \sqrt{2}$ and $M = 5 * 1024$). This also correlates well with the previous results.

By contrast, figure 309 shows that the FFT of the signal plus noise still shows a clear peak at the frequency of the sinusoidal waveform $x(t)$.

Hence it is concluded from this simple example that the non-linear operator does not improve the detection capability when broadband noise of sufficient magnitude is added to the signal. This is in agreement with literature, where it is shown that the performance of a square law demodulator will always be worse compared with a synchronous detector in low signal-to-noise ratios [81]⁸⁴.

6.10.3.5 Discussion and conclusion

In [15] it is also shown that an adaptive line enhancer (ALE) with multiplicative noise and square law modulation (termed ALE2 in [15]) is better than the ALE alone (termed ALE1

⁸⁴ In communication systems a bandpass filter is usually applied prior to the square law demodulator, hence the noise is not 'white' over all frequencies. This minor difference does not seem to have a major influence on the conclusion reached in literature [81]. As a matter of fact, a bandpass filter (similar to that used in a communication system) would also be applied in order to select the most 'resonant' frequency band for detection of the weak defect signals. Hence the conclusions in literature [81] seem to be equally applicable to the problem statement of this thesis.

in 15]) when $K\sigma_b^2 \geq A^2$ (where σ_b is the standard deviation of the multiplicative noise, A is the amplitude of the complex exponential and $K = 0.6$) and the additive noise is small in comparison with the multiplicative noise (also refer to section 6.10.7 ‘Amplitude modulation in the context of vibration monitoring’ on p. 183). As the additive noise starts to dominate the multiplicative noise, the normal ALE1 detector outperforms the ALE2 detector [15] from which it may be deduced that the presence of additive noise deteriorates the performance of the square law modulator ALE2 (conversely the presence of multiplicative noise tends to improve the performance of ALE2).

In practice, the signal $x(t)$ would be more complex than the simple sinusoidal waveform $x(t)$. The results presented in sections 11.11.1 to 11.11.6 (also section 11.11.9) with a simulated vibration signal and additive noise confirm the hypothesis that the non-linear operator performs quite well in the removal of speed variations, but does not perform as well in the removal of broadband noise.

Ho and Randall [21] show that raising the envelope spectrum to a higher power can be beneficial if the SNR is greater than one, otherwise the result will be worse. The envelope spectrum, obtained by applying the Hilbert transform [21], already squares the input signal effectively. Hence squaring of the envelope spectrum would be similar to raising the input signal to the fourth power. Furthermore, the noise discussed in [21] refers to discrete frequency components resulting from interference with gear mesh signals (although random noise could of course also be considered). Although the propagation of bearing signals through a gearbox presents a very interesting and challenging problem in itself, this falls outside the scope of this thesis. The important point to note is that the SNR should be high enough for squaring (or higher powers) to be of benefit⁸⁵. It is concluded that the non-linear operator will prove beneficial, provided that the SNR is ‘high enough’ (Ho and Randall [21] propose $\text{SNR} \geq 1$). Too much background noise (or conversely a bearing signal of too low an amplitude) may cause poor performance in the demodulation

⁸⁵ Ho and Randall state that an SNR of greater than unity proves beneficial for squaring the envelope signal [21]. However, the mean square ratio (MSR) between the bearing and noise components does not need to correspond with a SNR of one, owing to smearing [21]. Speed variations effectively add another noise component, complicating the analysis of the SNR.

spectrum. However, the convolution of the spectrum with itself (implied from the squaring of the signal plus noise) has the beneficial effect of removing small speed deviations, thus removing another noise source, which might be dominant in certain circumstances. The instantaneous load on the rollers (hence the defect) could vary substantially for a small bearing defect.

6.10.4 Extension to arbitrary signals

Although the results in the previous section 6.10.3 have been derived using a single sinusoidal waveform, it can (in principle) be extended to arbitrary signals provided that the FFT of the square of the signal and the FFT of the signal multiplied by the noise can be computed.

The FFT of the signal multiplied by the noise will introduce a convolution of the signal spectrum and noise spectrum in the frequency domain. A random phase term will be added to all the frequency components. Hence the noise will be coloured instead of white in general.

The FFT of the square of the signal may be computed in the frequency domain using autoconvolution of the spectrum of the signal $X(f)$ with itself⁸⁶. If it is further assumed that $X(f)$ contains a line spectrum at some defect frequency F_d then $X_2(f = kF_d) = X(f = kF_d) \otimes X(f = kF_d)$

If $X(kF_d) = \mathbf{X}_k$ then

$$X_2(k) = \mathbf{X}_k \otimes \mathbf{X}_k$$

$$= \sum \mathbf{X}_j \mathbf{X}_{k-j} \quad (j = -\infty \text{ to } \infty)$$

$$\text{so } X_2(0) = X(0)X(0) + X(1)X(1) + \dots X(M)X(M)$$

$$X_2(1) = X(0)X(1) + X(1)X(2) + \dots X(M-1)X(M) + \dots \quad (115)$$

⁸⁶ A periodic convolution should be used instead of linear convolution.

The values may be readily computed by an autocorrelation function $R_x(\omega)$ in the frequency domain (instead of in the time domain). The maximum correlation would occur for $R_x(0) = X(0)X(0) + X(1)X(1) + \dots X(M)X(M)$, where

$R_x(1) = X(0)X(1) + X(1)X(2) + \dots X(M-1)X(M) + \dots$ having a smaller value. Eventually the tails of $R_x(\omega)$ would approach zero.

If the values of $X_2(1)$ and the additive noise $N(f)$ are known, $X_2(1)$ may be viewed as the effective contribution of all the complex exponentials (separated by the defect frequency F_d). Thus $X_2(1)$ may be regarded as a single complex exponential that represents the contribution of all the complex exponentials and may be substituted into the equations of section 6.10.3 to determine the detection ratio.

6.10.5 Amplitude variation

Any amplitude variation of the signal $X(f)$ will also manifest in spectral smearing. For example, if the signal $x(t)$ is multiplied by the modulating function $m(t)$, the spectrum of $x(t)m(t)$ is $X(f) \otimes M(f)$. As a simple example, if $m(t)$ is a sinusoidal signal of frequency f_m , all components of $X(f)$ will be shifted by f_m in the frequency domain. If $M(f) = K$ for $0 \leq f \leq f_m$, then $X(f)$ will be convolved with a rectangular function $M(f)$ of width f_m . If the signal $X(f)$ is given by $P(f)D_1(f)$ where $D_1(f) = 1/T_1 \sum \delta(f - k/T_1)$ then $X(f) \otimes M(f)$ would be a series of rectangular pulses in the frequency domain of width f_m spaced at frequency $1/T_1$. If the highest frequency of the modulation function $M(f)$ exceeds the defect frequency then spectral smearing is likely to result.

The demodulation spectrum $Y_2(f)$ is given by $Y_2(f) = (X(f) \otimes M(f)) \otimes (X(f) \otimes M(f)) = (X(f) \otimes X(f)) \otimes (M(f) \otimes M(f)) = X_2(f) \otimes M_2(f)$ owing to the associative property of convolution (where $M_2(f) = M(f) \otimes M(f)$). As an example, if $M(f)$ is a rectangular function with value K for $0 \leq f < f_m$, then $M_2(f)$ would be a triangular function of width $2f_m$. If the frequency f_m is quite close to the defect frequency, then the convolution of $M(f)$ with $X(f)$ would make it difficult to identify repetitive impulses in the frequency domain signal $X(f) \otimes M(f)$. However, the function $Y_2(f)$ would show repetitive energy in the frequency domain owing to the function $M_2(f)$.

Figure 22 shows an example of a simulated Gaussian signal in the time domain with no speed variation and amplitude variation. As the amplitude variation function is given by multiplicative noise with standard deviation of one, the multiplicative noise may be close to zero. This explains why the time domain signal is not clearly periodic at the defect frequency. Figure 23 shows the spectrum of the signal in which some spectral smearing is present owing to the amplitude variations. The peaks at multiples of the defect frequency are harmonics of the defect frequency. Figure 24 shows the demodulation spectrum of the signal squared in which a clear demodulation peak at the defect frequency is present, although the noise level is also relatively high.

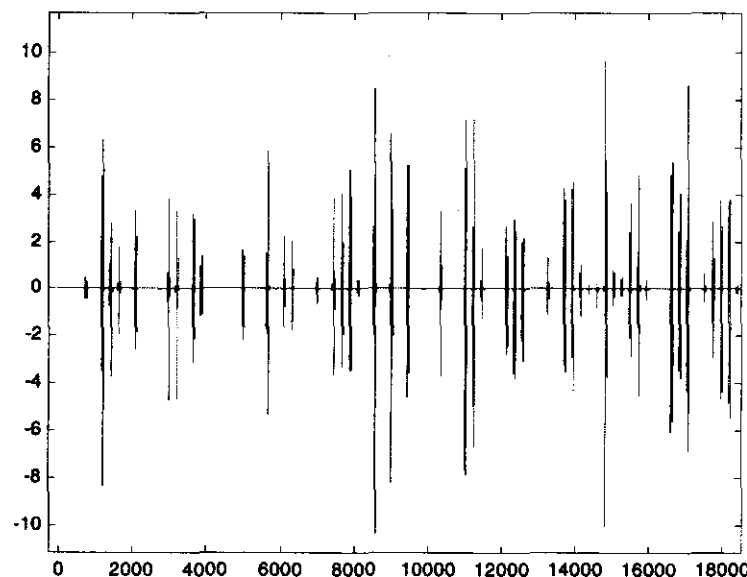


Figure 22: Time domain signal of simulated Gaussian pulse with zero speed variation and amplitude variation

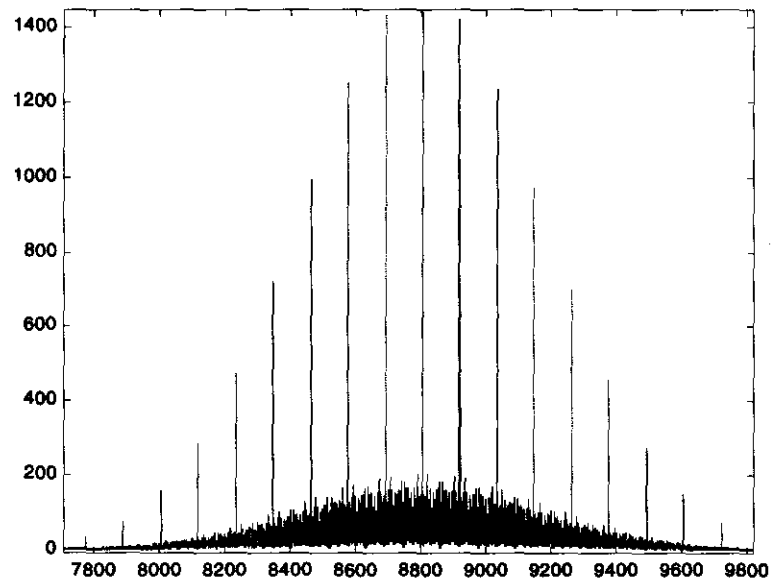


Figure 23: Amplitude spectrum of simulated Gaussian pulse with zero speed variation and amplitude variation

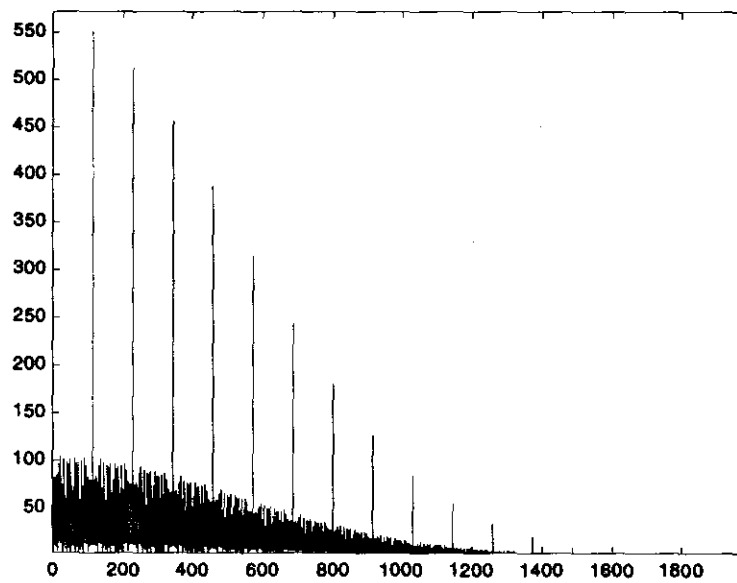


Figure 24: Amplitude spectrum of square of simulated Gaussian pulse with zero speed variation and amplitude variation

6.10.6 Matched filter detection

It is also possible to use principles of communication system theory in the present problem of the detection of multiple fault conditions (refer to section 3.3 on p. 47). Knowledge about the transmitted pulse over a communication channel allows the signal-to-noise ratio to be optimized in a communication system [81]. The optimal filter for the detection of the pulse (matched filter) will be related to the transmitted pulse [81]. When the received signal is subsequently filtered with the matched filter the signal-to-noise ratio will be optimized. In the context of bearing fault detection the waveform of the received pulse is changed by the loading, speed variations (phase modulation) and transfer function to the sensor. Hence it is not that easy to determine the source pulse corresponding to the transmitted pulse in a communication system. One approach would be to use the synchronous average of the vibration signal and extract one cycle only for use as the transmitted pulse. This could then be used as a matched filter detector. In the frequency domain the effect will be to interpolate between the equally spaced samples of the synchronous average in the frequency domain, resulting in a continuous spectrum (instead of the line spectrum associated with the synchronous averaging). The transmitted pulse may then be convolved in the time domain with the received signal (the vibration signal). This corresponds to a multiplication of the respective spectrum in the frequency domain. Unfortunately, the application of the matched filter poses a number of problems:

- Knowledge about the transmitted pulse is required. This may be influenced by the shape (geometrically distributed or well localized) and size of the bearing defect (large, medium or small defects would have different waveforms for a specific bearing), the loading on the roller (static or dynamic), and the sensor transfer function. Hence further work would be required before the influence of all of the above factors may be determined reliably.
- Even if the transmitted pulse is known exactly, for example in a simulated vibration signal consisting of Gaussian pulses with timing jitter, the convolution of the transmitted pulse with the received signal will not supply

a delta impulse at each defect location. If the delta impulse sequence is $D(f)$ and the elementary Gaussian pulse is $S(f)$, then the spectrum of the simulated Gaussian pulses is $G(f) = D(f)S(f)$. In principle, the delta impulse sequence $D(f)$ may be obtained by $D(f) = G(f)/S(f)$ (provided that $S(f) \neq 0$ for all f). The spectrum $S(f)$ could be estimated from one cycle of a synchronous average. However, if $G(f) = 0$ for some frequencies (for example $G(f)$ could be an ideal bandpass filter with rectangular frequency window), then it will no longer be possible to reconstruct the sampling signal $D(f)$ perfectly owing to the rectangular function applied in the frequency domain (equivalent to convolution in the time domain with a sinc function, which is the IFFT of the rectangular function).

- In high noise level scenarios, which typically result from small signal-to-noise ratios combined with amplitude and speed variations, it might be difficult to estimate the transmitted pulse accurately enough.

Hence the application of the matched filter approach will no longer be investigated in this thesis, although it seems to be a potentially novel application in vibration monitoring.

6.10.7 Amplitude modulation in the context of vibration monitoring

The amplitude, phase or frequency of a carrier may be changed (modulated) to convey information in a communication system. If a lowpass filtered baseband signal $m(t)$ with maximum frequency f_m is multiplied by a carrier waveform $f_c(t) = A\cos(\omega_c t)$ the spectrum of the information signal $m(t)$ will be convolved with the spectrum of the carrier $f_c(t)$. If the carrier frequency $f_c = \omega_c/(2\pi)$ is much higher than f_m , the baseband spectrum $M(\omega) = \mathfrak{F}(m(t))$ – where $\mathfrak{F}$ is the Fourier transform of $m(t)$ – will be shifted up by f_c Hz⁸⁷ (e.g. [81]).

⁸⁷ If a double-sided spectral representation with negative frequencies is used, the baseband spectrum $M(\omega)$ would also be shifted down to $-f_c$ Hz and the amplitude of the resulting carrier wave halved.

In order to demodulate the high frequency amplitude modulated (AM) signal $y(t) = m(t)f_c(t)$ (where the spectrum is given by $Y(\omega) = M(\omega) \otimes F_c(\omega)$) synchronous demodulation may be used if a copy of the carrier wave is available. By multiplying the signal $y(t)$ with the carrier wave $f_c(t)$ it is therefore possible to recover the baseband information signal $m(t)$ – the spectrum $Y(\omega)$ shifts down by f_c Hz back to baseband [81]. Synchronous demodulation would work well when there are neither speed variations nor amplitude modulation of the individual defect pulses (provided that the carrier frequency is a multiple of the defect frequency⁸⁸). Unfortunately, speed variations (which are present even if the shaft speed is constant [10, 21]) change the instantaneous frequency of the carrier wave(s). The effect is more pronounced the higher the demodulated frequencies are (for details please refer to section 6.10.3 starting on p. 171). Hence synchronous AM demodulation cannot provide a clear demodulation peak in general as is possible with asynchronous AM demodulation (even if the theoretical defect frequencies may be computed, small computational errors – or slight differences in geometry – would propagate rapidly in the high frequency range). However, for a large bearing defect the energy (or amplitude) levels in the high frequency range do increase with an increase in defect angle (DA). Thus it is in principle still possible to use synchronous AM demodulation for bearing fault detection⁸⁹, even though no clear demodulation peak might be visible in the baseband demodulated signal (of the synchronous AM demodulator⁹⁰). The disadvantage is that the exact defect frequency has to be known, while the asynchronous AM demodulator directly supplies the defect frequency (refer also to the discussion in appendix K).

⁸⁸ As the time domain signal of interest is typically a bandpass filtered signal, with a frequency range much larger than the defect frequency, the carrier frequency to demodulate the high frequency signal back to baseband has to be a multiple of the fundamental defect frequency.

⁸⁹ Assuming, of course, that other vibration signals, such as gear noise, do not swamp the weak bearing signal.

⁹⁰ However, the asynchronous AM demodulator should provide a clear demodulation peak if the signal-to-noise ratio is 'high enough' (refer also to the discussion in section 11.11). The asynchronous AM demodulator implies a frequency domain convolution which could, in specific circumstances, improve the demodulation peaks relative to the synchronous AM demodulator.

When speed variations are present asynchronous AM demodulation (which does not require an accurate copy of the carrier signal) is more appropriate. In a communication system a large carrier (which represents 'wasted' power when transmitted) is added to enable easy demodulation of the signal $y(t)$ without requiring an accurate carrier (hence enabling a cheaper receiver). In the context of vibration monitoring a carrier wave is unfortunately not available⁹¹, thus asynchronous amplitude demodulation (by applying fullwave rectification by squaring the bandpass filtered signal or the Hilbert transform) may be used to obtain the (low frequency) AM (amplitude modulation) function representing the (almost) periodic impact of the defect. It is postulated that better results would be obtained using asynchronous AM demodulation in the presence of amplitude and speed variations than using synchronous AM demodulation. This aspect is left for future work. Yet, it is clear that spectral smearing would result with amplitude or speed variations in the case of synchronous AM demodulation. Simply shifting the spectrum back to baseband by synchronous AM demodulation would not at all change the spectral smearing (assuming amplitude and/or speed variations). However, asynchronous AM demodulation (HFRT) changes the spectrum by convolution. Hence some improvement may be expected relative to the synchronous AM demodulation. Some justification for this statement may be found in [15], in which it is shown that an ALE with multiplicative noise and square law modulation (termed ALE2 in [15]) is better than the ALE alone (termed ALE1 in [15]) when $K\sigma_b^2 \geq A^2$ (where σ_b is the standard deviation of the multiplicative noise, A is the amplitude of the complex exponential and $K = 0.6$) and the additive noise is small compared with the multiplicative noise [15]. The above result is valid until the additive noise and multiplicative noise powers are about equal [15], after which a very rapid decrease in the value of K is evident with an increase in the additive noise. The implication is that as the additive noise starts to dominate the multiplicative noise, the normal ALE1 detector outperforms the ALE2 detector. This result is in complete agreement with the results presented in this chapter, in which it is shown that the

⁹¹ One might possibly associate the harmonic peaks resulting from a large bearing defect as a series of carrier waves and then add the resulting contribution of each carrier wave using the principle of superposition. The result is likely to be similar to HFRT for zero speed variation.

square law modulator does not perform as well in the presence of additive noise. Furthermore, any amplitude variation in the signal (represented by the multiplicative noise term in [15]) would give the ALE2 detector an advantage compared with the ALE1 detector, thus implying the superiority of the square law modulator in this situation.

6.10.8 Conclusion

In general, it is evident from the examples discussed that the non-linear operator (squaring) applied to the signal gives an improvement over the usual synchronous average in the presence of speed variation. This is in agreement with literature, in which the beneficial aspects of non-linear operators are discussed by McFadden *et al.* [50, 53]. Randall also indicates that in certain circumstances of signal-to-noise ratios, even higher order powers of the signal may be beneficial. As the synchronous average is closely linked to the ESFFT, hence to the demodulation spectrum, using, for example, the HFRT it is clear that at least some of the success of the demodulation spectrum in bearing fault detection may be ascribed to the presence of the non-linear operator⁹². When broadband noise of sufficient magnitude was added to the signal, the demodulation spectrum (non-linear operator with squaring of the signal) did not seem to perform as well as the regular magnitude spectrum (i.e. usual FFT of the signal plus noise – refer also to the discussion in appendix K). A heuristic formula is derived in section 6.10.3 of this thesis for the detection ratio of the square law modulator and conditions are stipulated for the detection of a signal in the presence of white noise.

6.11 Amplitude demodulation – relation between the Hilbert transform and square law modulator

6.11.1 Spectrum of the square law modulator

The amplitude modulation function is given by the absolute value of the analytical signal derived from the Hilbert transform [73], $\hat{a}(t) = a(t) + j\tilde{a}(t)$, where $\tilde{a}(t)$ is the Hilbert transform of the signal $a(t)$ and $\hat{a}(t)$ is the analytical signal corresponding to $a(t)$. It is of

⁹² The implicit calculation of the synchronous average using the ESFFT, as described in this thesis, is another obvious factor.

interest to examine the relationship between the spectrum of the square law modulator and the Hilbert transform.

The spectra of the square law modulator are given by $A_2(f) = A(f) \otimes A(f)$ (where the symbol $\otimes$ denotes convolution in the frequency domain⁹³). Divide the amplitude spectra $A(f)$ into positive, zero and negative components

$A(f) = A^+(f) + A^0(f) + A^-(f)$, where $A^+(f) = A(f)$ for $f > 0$ (0 otherwise), $A^-(f) = A(f)$ for $f < 0$ (0 otherwise) and $A^0(f) = A(f)$ for $f = 0$ (0 otherwise). Alternatively $A^+(f) = A(f)U(f)$, where $U(f)$ is the unit step function (similarly $A^-(f) = A(f)U(-f)$). Then

$$\begin{aligned} A_2(f) &= A(f) \otimes A(f) \\ &= (A^+(f) + A^0(f) + A^-(f)) \otimes (A^+(f) + A^0(f) + A^-(f)) \\ &= A^+(f) \otimes A^+(f) + A^-(f) \otimes A^-(f) + 2A^-(f) \otimes A^+(f) \\ &\quad + 2A^0(f) \otimes (A^+(f) + A^0(f) + A^-(f)) \end{aligned} \quad (116)$$

Convolution with the constant $A^0(f)$ in equation 116 above is simply multiplication by $A^0(f)$ and integration over just one point in the frequency domain, giving a result of 0. If it is assumed that $A^0(f) = 0$ (zero mean value), the equation simplifies to

$$\begin{aligned} A_2(f) &= A(f) \otimes A(f) \\ &= A^+(f) \otimes A^+(f) + A^-(f) \otimes A^-(f) + 2A^-(f) \otimes A^+(f) \end{aligned} \quad (117)$$

6.11.2 Spectrum of the absolute value of the analytical signal (derived from the Hilbert transform)

The spectrum of the analytical signal is given by

$$\hat{A}(f) = A(f) + j\tilde{A}(f) \quad (118)$$

where $\hat{A}(f)$ is the spectrum of $\hat{a}(t)$, $A(f)$ is the spectrum of $a(t)$ and $\tilde{A}(f)$ is the spectrum of $\tilde{a}(t)$.

⁹³ The periodic convolution using the FFT (instead of the linear convolution) should be used for sampled signals.

But the spectrum of the Hilbert transform $\tilde{a}(t)$ of $a(t)$ is given by $\tilde{A}(f) = j * A(f)\text{sgn}(f)$ [73]. Hence the spectrum of the square of the absolute value of the analytical signal $|\hat{a}(t)|^2 = a(t)^2 + \tilde{a}(t)^2$ is given by the following:

$$\begin{aligned} W(f) &= A(f) \otimes A(f) + \tilde{A}(f) \otimes \tilde{A}(f) = A(f) \otimes A(f) - A(f)\text{sgn}(f) \otimes A(f)\text{sgn}(f) \\ &= A_2(f) - A(f)\text{sgn}(f) \otimes A(f)\text{sgn}(f) = A_2(f) - \hat{A}_2(f) \end{aligned}$$

where

$$\begin{aligned} \hat{A}_2(f) &= A(f)\text{sgn}(f) \otimes A(f)\text{sgn}(f) \\ &= (A^+(f) + 0 * A^0(f) - A^-(f)) \otimes (A^+(f) + 0 * A^0(f) - A^-(f)) \\ &= A^+(f) \otimes A^+(f) + A^-(f) \otimes A^-(f) - 2A^-(f) \otimes A^+(f) \end{aligned} \quad (119)$$

Hence $W(f) = A_2(f) - \hat{A}_2(f) = 4A^-(f) \otimes A^+(f)$, as the other terms from the square law modulator cancel. Thus the spectrum of $w(t)$ is $4A^-(f) \otimes A^+(f)$, and the absolute value of the analytical signal is given by $|\hat{a}(t)| = \text{sqrt}(\mathfrak{F}^{-1}(4A^-(f) \otimes A^+(f)))$, where $\mathfrak{F}^{-1}$ is the inverse Fourier transform.

6.11.3 Comparison of the spectrum of the absolute value of the analytical signal and the square law modulator

A graphical illustration may help to visualize the convolution of $A^-(f)$ and $A^+(f)$.

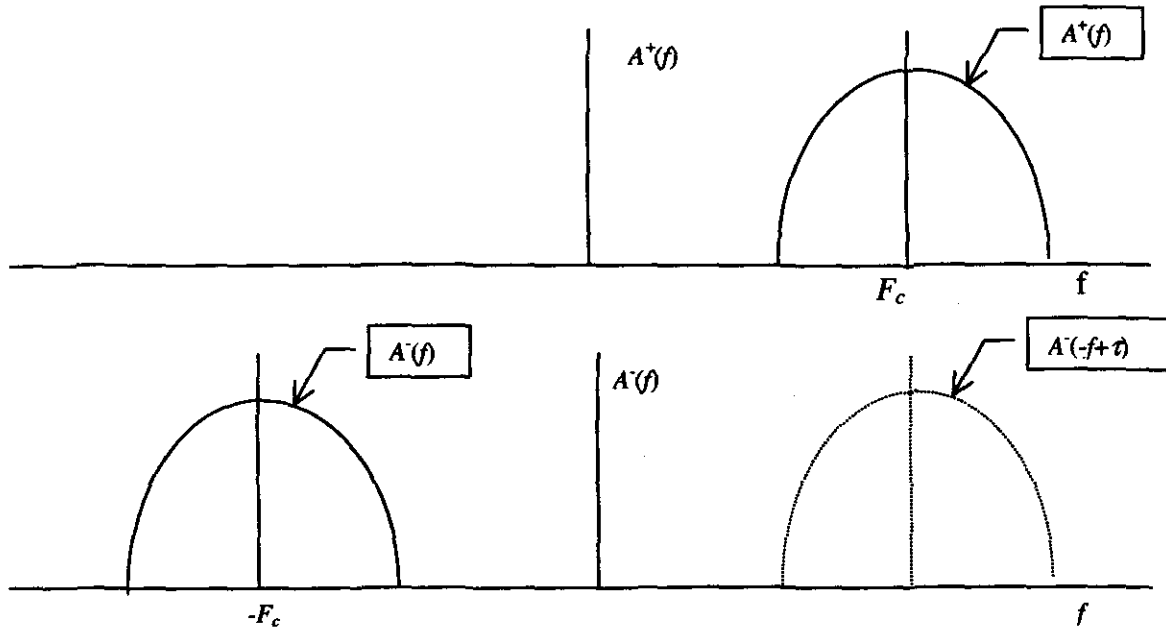


Figure 25: The convolution of $A^-(f)$ and $A^+(f)$

As can be seen from figure 25, the convolution of $A^-(f)$ and $A^+(f)$ results in a reflection of $A^-(f)$ around the frequency axis $A^-(f)$. A symmetrical reflection around the $f=0$ axis is used in figure 25, although this does not have to be the case in general. The area underneath $A^+(f)A^-(f)$ is related to the mean square value of the signal. It is clear that as the signal $A^-(f+\tau)$ is shifted from $\tau=0$, the value of $W(\tau)/4$ is the area underneath $A^+(f)A^-(f+\tau)$. It should be evident from figure 25 that $W(f)$ would be a lowpass response if $A(f)$ is a bandpass response. Hence the bandpass response of $A(f)$ is amplitude demodulated to baseband (refer to figure 25) by application of the analytical signal derived from the Hilbert transform.

A similar observation may be made for the square law modulator. Equation 116 also has a term $2A^-(f) \otimes A^+(f)$ that would amplitude demodulate the signal. Furthermore, it has two terms related to $A^+(f) \otimes A^+(f)$ and $A^-(f) \otimes A^-(f)$. These terms represent the high frequency response (if it is assumed that $A(f)$ has a bandpass response) and would be filtered out after application of the square law modulator to give the envelope of the signal.

6.11.4 Relation between the demodulation spectrum of the positive and negative frequency components

It may be shown that $G^+_2(f) = A^+(f) \otimes A^+(f)$ and $G^-_2(f) = A^-(f) \otimes A^-(f)$ are related as $G^-_2(f) = G^{+*}_2(-f)$ (also equivalently that $G^+_2(f) = G^{-*}_2(-f)$), where $G^+_2(f)$ are the demodulation spectra resulting from a convolution of the positive frequency components in the spectrum of $A(f)$. For a real valued time signal $g(t)$, the following relations are known [73]:

$$\begin{aligned}
 G(f) &= G^*(-f) \\
 R(f) &= R(-f) \\
 I(f) &= -I(-f) \\
 G(f) &= R(f) + jI(f) \\
 &= R(-f) - jI(-f) \\
 &= G^*(-f)
 \end{aligned}
 \quad \begin{array}{l} \nearrow \\ \text{The spectrum is conjugate even (real parts} \\ \text{even, imaginary parts odd)} \end{array}
 \quad (120)$$

and

$$\begin{aligned}
 G(-f) &= R(-f) + jI(-f) \\
 &= R(f) - jI(f) \\
 &= G^*(f)
 \end{aligned}
 \quad (121)$$

But

$$\begin{aligned}
 G^+_2(f) &= A(f)U(f) \otimes A(f)U(f) \\
 &= (R(f) + jI(f))U(f) \otimes (R(f) + jI(f))U(f) \\
 &= R(f)U(f) \otimes R(f)U(f) + R(f)U(f) \otimes jI(f)U(f) \\
 &\quad + jI(f)U(f) \otimes R(f)U(f) + jI(f)U(f) \otimes jI(f)U(f) \\
 &= R(f)U(f) \otimes R(f)U(f) - I(f)U(f) \otimes I(f)U(f) \\
 &\quad + j2R(f)U(f) \otimes I(f)U(f)
 \end{aligned}
 \quad (122)$$

and

$$\begin{aligned}
G_2(f) &= A(f)U(-f) \otimes A(f)U(-f) \\
&= (R(f) + jI(f))U(-f) \otimes (R(f) + jI(f))U(-f) \\
&= R(-f)U(-f) \otimes R(-f)U(-f) \\
&\quad - jI(-f)U(-f) \otimes (-jI(-f)U(-f)) \\
&\quad + R(-f)U(-f) \otimes (-jI(-f)U(-f)) \\
&\quad - jI(-f)U(-f) \otimes R(-f)U(-f) \\
&= R(-f)U(-f) \otimes R(-f)U(-f) \\
&\quad - I(-f)U(-f) \otimes I(-f)U(-f) \\
&\quad - 2jR(-f)U(-f) \otimes I(-f)U(-f) \\
&= (G_2^+)^*(-f)
\end{aligned} \tag{123}$$

Hence $G_2(f) = G_2^{+*}(-f)$ as stated. A consequence of this relation is that the convolution of the negative frequency components may be obtained easily from the convolution of the positive frequency components. Thus the number of computations may be reduced.

6.11.5 Conclusion

If the spectrum $A(f)$ of a signal $a(t)$ has a bandpass response (which would usually be the case, as the signal is bandpass filtered around a resonance prior to the application of the square law operation), then $A^+(f) \otimes A^+(f)$ and $A^-(f) \otimes A^-(f)$ would have very high frequency components (around $2F_c$, where F_c is the carrier frequency of the resonance), which may be removed easily with lowpass filtering⁹⁴. Furthermore, $A^+(f) \otimes A^-(f)$ would be a lowpass response for the amplitude demodulator. Hence the absolute value of the analytical signal (derived from the Hilbert transform) and the square law modulator would give similar results owing to the $A^+(f) \otimes A^-(f)$ term in both spectra (if the output of the amplitude modulator is lowpass filtered after application of the non-linear operator). Thus it is concluded that both non-linear operators, the square law modulator as well as the analytical signal derived from the Hilbert transform, provide the desired result of

⁹⁴ If the spectrum $A(f)$ of a signal $a(t)$ has a lowpass response (which usually would not be the case, as the signal is bandpass filtered around a resonance prior to the application of the square law operation and hence is a bandpass signal), then $A^+(f) \otimes A^+(f)$ and $A^-(f) \otimes A^-(f)$ would also have lowpass frequency components (where $A^+(f)$ and $A^-(f)$ are the positive and negative frequency components of $A(f)$ respectively).

amplitude modulation using a bandpass signal as input. The analytical signal derived from the Hilbert transform has the potential advantage of giving a lowpass response only (the amplitude modulation of the signal) using a bandpass signal as input⁹⁵. However, if the interest is simply in the value of the low frequency components (as a measure of the vibration at a defect frequency using a non-linear operator, for instance) a lowpass filter is not strictly necessary (provided, of course, that the bandwidth of the bandpass signal is small enough). Yet, for a proper time domain analysis of the envelope of the bandpass signal, a lowpass filter is required to do the envelope detection. It has also been shown that the spectra $G_2^+(f) = A^+(f) \otimes A^+(f)$ and $G_2^-(f) = A^-(f) \otimes A^-(f)$ of the signal $a(t)$ are related by frequency reflection and conjugation ($G_2^-(f) = G_2^{+*}(-f)$ and $G_2^+(f) = G_2^{-*}(-f)$), which is similar to the result $G(f) = G^*(-f)$ developed in literature [73].

6.12 Modified cepstral analysis

In section 5.3.2 the theory of cepstral analysis has been described and the LCEPS, TCEPS, and LTCEPS algorithms have been introduced. Figure 26 shows a typical simulated bearing fault signal – refer to equation 30 on p. 48.

⁹⁵ In principle, it is easy enough to filter out the high frequency components resulting from the square law modulator using a lowpass filter (which could be used to filter the absolute value of the analytical signal as well).

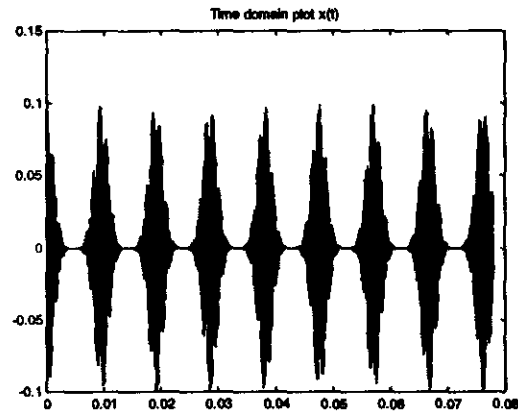


Figure 26: Gaussian pulse sequence centred at 8.8kHz (fractional bandwidth = 0.035, timing variation = 0%)

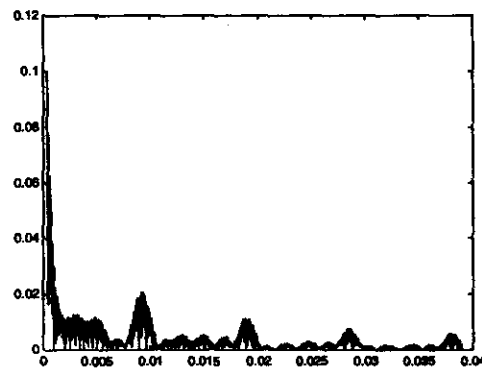


Figure 27: Real cepstrum of figure 26 (no speed variation); horizontal axis in quefrency (seconds) and vertical axis in magnitude; defect quefrency is $\approx 1/(114\text{Hz}) = 0.009$ seconds)

The real cepstrum of figure 26, shown in figure 27, does not have a clear peak, in contrast with the well-defined peak in the demodulation spectrum of figure 28.

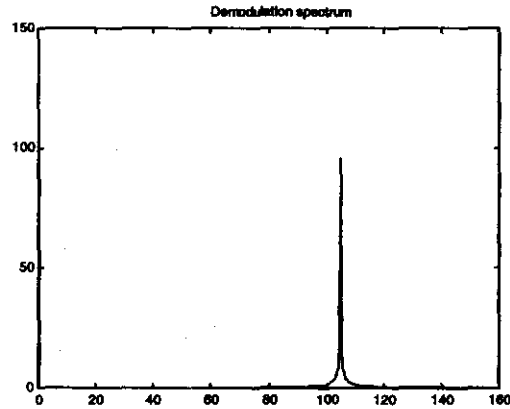


Figure 28: Demodulation spectrum of figure 26 (no speed variation)

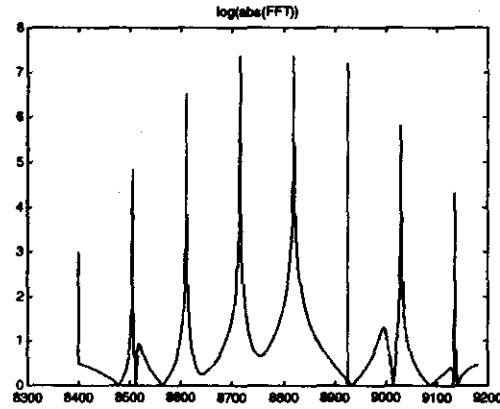


Figure 29: Absolute value of log of magnitude spectrum of figure 26 (no speed variation)

The reason for the unclear peak at the defect frequency is interference of small frequency components amplified by the log operation, as shown in figure 29. Small FFT components are amplified by the log operation and interfere when applying the IFFT of the cepstrum. Furthermore, spectral leakage of the FFT causes a broad peak in the FFT of figure 29, instead of a series of impulses (ideally the dirac delta) spaced with the defect frequency. When applying the FFT again to obtain the cepstrum, a series of broad peaks is evident (refer to figure 27 and figure 29) instead of a single, well-defined peak as in figure 28.

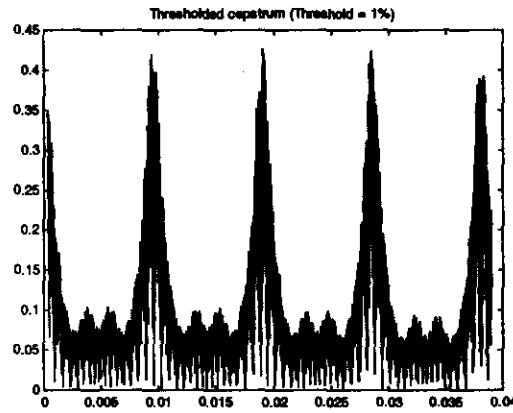


Figure 30: LTCEPS of figure 26 (bandwidth = 0.035, threshold = 1%, speed variation = 0%); horizontal axis in quefrency (seconds)

Figure 30 shows that the LTCEPS algorithm, with 1% threshold, supplies a much-improved series of impulses in the cepstral domain.

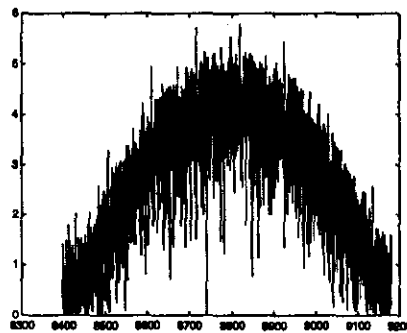


Figure 31: Log of magnitude spectrum (bandwidth = 0.035, threshold = 1%, speed variation = 1%)

The dramatic effect of instantaneous speed variation is evident in figure 31, where a speed variation of just 1% causes significant spectral smearing.

Figure 32 and figure 33 show the effect of buffer length on the LTCEPS algorithm. A smaller buffer length improves the definition of the cepstrum peaks in figure 33 in comparison with figure 32. It may be argued that the longer the time domain buffer, the greater the spectral smearing would be owing to instantaneous speed variations. Unfortunately, smaller buffer lengths could lead to reduced frequency resolution.

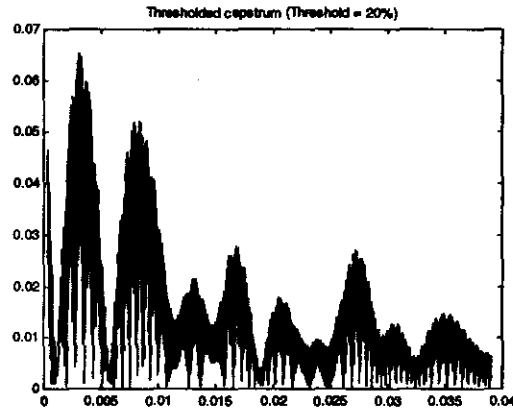


Figure 32: LTCEPS (buffer length = 8000, bandwidth = 0.035, threshold = 20%, speed variation = 1%); horizontal axis in quefrency (seconds); defect quefrency is $\approx 1/(114\text{Hz}) = 0.009$ seconds)

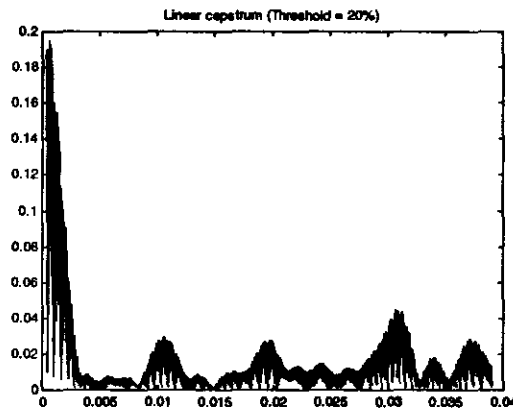


Figure 33: LTCEPS (buffer length = 2000, bandwidth = 0.035, threshold = 20%, speed variation = 1%); horizontal axis in quefrency (seconds); defect quefrency is $\approx 1/(114\text{Hz}) = 0.009$ seconds)

In summary, the conclusion of this section is that cepstral analysis is problematic when speed variations are present. An improved cepstral analysis was proposed and illustrated for different buffer lengths. It is concluded that the proposed improved cepstral analysis supplies more sharply defined peaks at the defect frequency than the regular real cepstrum, although still not as sharp as the demodulation spectrum peaks.

6.13 Applying digital signal processing techniques to improve the signal-to-noise ratio in vibration signals

6.13.1 Introduction

This section investigates the application of signal processing techniques to improve the signal-to-noise ratio in vibration signals used in condition monitoring. However, environmental conditions such as instantaneous speed variations as well as the presence of multiple fault conditions may obscure the defect signals that are required for reliable diagnostics and can lead to faulty diagnostic decisions [98]. While these problems can be solved with the right combination of techniques, the difficulty remains of obtaining sufficiently large measured data sets on which to train these techniques. Artificially generated training sets, by empirical modelling of defects, are hence investigated. A simple vibration model that includes the effect of period variation is proposed for the bearing defect data set in [98].

The vibration signals caused by bearing and gear defects both lead to instantaneous variations that are observable in the time and frequency domains. Yet, various practical effects, e.g. the simultaneous presence of more than one fault mechanism, may obscure the defect signals that are required for reliable diagnostics. It has, for example, been shown that a deteriorating secondary fault condition can lead to faulty diagnostic decisions [98]. Correct preprocessing, extraction of appropriate signal features and the application of automated classification may overcome these problems in principle.

Owing to the small available data sets, empirical modelling of vibration phenomena related to such defects is hence investigated to generate training sets artificially. McFadden and Smith in [51, 52] propose a simple vibration model for inner race defects, which is extended by Brie in [10] to account for the effect of speed variations (which occur even if the shaft speed is constant). A simple vibration model, which includes the effect of period variation, is proposed for the bearing defect data set in [98]. Signal processing techniques that can reduce the noise caused by angular speed variations of the shaft and provide better features for fault recognition are subsequently investigated. Although a reference signal related to the angular position of the shafts can reduce this

problem, this signal is unfortunately not available in many industrial data sets. A speed compensation algorithm is proposed in section 6.13.3.3 on p. 205 and successfully applied to a synthetic vibration signal with speed variation.

6.13.2 Problem formulation

6.13.2.1 Modelling of multiple fault mechanisms

6.13.2.1.1 Modelling of bearing defects (including size of defect)

It has been shown in [78] that the size of the demodulation peak in the vibration spectrum is linearly related to the defect size using vibration measurements. Thus the normalized ratio of demodulation peak to the ‘carpet level’ noise (noise in the demodulation spectrum) provides a quantitative measure of the bearing defect condition. Bearing replacement is suggested when this ratio is above 15dB. Figure 34 shows the influence of buffer length on the demodulation peak of the bearing data set of [98]. Increased averaging leads to higher signal-to-noise levels. It should be noted that determining the carpet level is no longer easy with no trending measurements available and when speed variation is present.

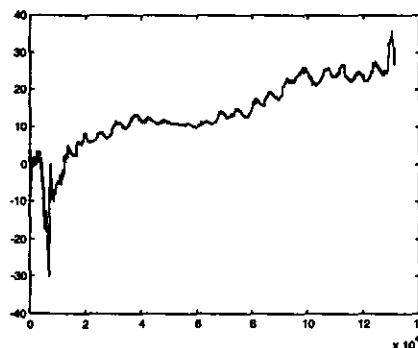


Figure 34: Ratio of demodulation peak relative to carpet level (vertical axis in dB) versus length of time domain buffer (horizontal axis in samples)

McFadden and Smith proposed a model of an exponentially damped sinusoid as the response of the structure to an impulse $d(t)$ caused by contact of a ball with a single point (inner race) defect [51]. The impulses $d(t)$ are modulated by the load function $q(t)$ and the

transfer function to the sensor $a(t)$. In [10] alternative models are proposed, firstly using time-frequency analysis, allowing instantaneous changes in frequency (chirp signal). The second model, while based on a first degree of freedom system only, uses time-varying parameters (LTV system). It is also shown in [10] that variations in instantaneous speed may be modelled using a quasi-periodic impulse train and analytical expressions are derived for the autocorrelation and power spectral density (PSD).

It is clear that instantaneous variation in the contact angle causes timing jitter in the response of the system to a bearing defect, even if no speed variation occurs in the rotating speed. The timing jitter causes spectral smearing [10], complicating spectral and cepstral analysis.

The research conducted in this thesis is not aimed at developing the best vibration model for defect detection, but rather at evaluating the effectiveness of various signal processing operations, usually implemented in a preprocessing stage, to support the extraction of characteristic features. A simple vibration model for defect detection, with reasonable correspondence to the measured data set, is therefore required.

A model of the response of the system to a defect impulse is proposed in this thesis as a Gaussian modulated sinusoidal signal with fractional bandwidth BW and resonance frequency f_c . The bandwidth of the pulse is $100 * BW\%$ measured at a level of BWR dB relative to the normalized peak of the signal. The signal may be expressed analytically as (also refer to equation 30 on p. 48)

$$Y_b = A \exp \left(\left(\frac{-(t-\tau)}{\sqrt{2}t_v} \right)^2 \right) \exp(-j2\pi f_c(t-\tau)) \quad (124)$$

with amplitude A (which can also, of course, be amplitude modulated over time to simulate the effect of loading), Gaussian mean f_c , Gaussian variances $t_{v2} = 1/(4\pi^2 f_v^2) = t_v^2$, $t_v = 1/(2\pi f_v)$, $f_v = \sqrt{f_{v2}}$, $f_{v2} = BW^2 f_c^2 / (8 \log(r))$ and reference level $r = 10^{BWR/20}$ and time delay τ .

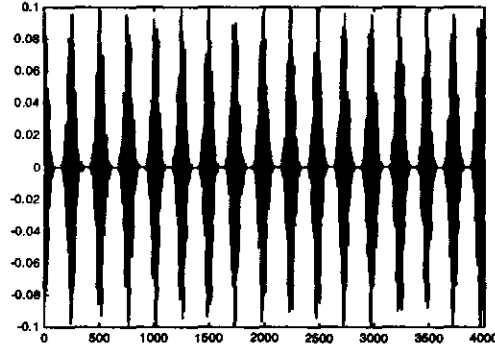


Figure 35: Compensated time domain signal (horizontal axis in samples)

An example of a time domain signal, repeating at 105Hz intervals, is given in figure 35 where $BW = 0.035$.

6.13.2.1.2 Modelling of bearing defects (multiple defects)

The model proposed in [52] by McFadden and Smith can accommodate the effect of multiple defects. In [52] the effect of M multiple defects is modelled as the superposition of M single point defects. It is assumed that the M defects occur at regularly spaced intervals corresponding to the defect time (inverse of the fundamental defect frequency).

$\mathfrak{I}(x(t))$ is the Fourier transform of the signal $x(t)$. For the special case of two defects of equal strength, $M = 2$ and the demodulated spectrum is given by

$$\begin{aligned} H_{M=2}(f) &= \mathfrak{I}\left((h(t) + h(t - \tau))^2\right) \\ &= \mathfrak{I}\left((h(t))^2\right) + \mathfrak{I}\left((h(t - \tau))^2\right) + 2\mathfrak{I}(h(t)h(t - \tau)) \\ &= H_{M=1} + I_1 + I_2 \end{aligned} \tag{125}$$

$$\begin{aligned} I_1 &= \int_{-\infty}^{\infty} h(t - \tau)^2 e^{-j2\pi ft} dt \\ &= \int_{-\infty}^{\infty} h(\alpha)^2 e^{-j2\pi f\alpha} d\alpha e^{-j2\pi f\tau} \end{aligned}$$

$$H_{M=2}(f) = H_{M=1}(f)(1 + e^{-j2\pi f\tau}) + I_2$$

assuming that $h(t)$ is the response of the system to a bearing defect and $H(f)$ denotes the demodulation spectrum of the fullwave rectified signal $h(t)^2$. If it is assumed that the response of the system $h(t)$ to an impact caused by the bearing defect is zero before the time delay τ , say at time t_1 , and that $\tau < T/M$, the integral I_2 is zero. Given these assumptions the expression simplifies to

$$H_{M=2}(f) = H_{M=1}(f)(1 + e^{-j2\pi f\tau}) + I_2 \quad (126)$$

The above expression indicates that the minimum value (zero) of the demodulation peak will be reached when $f = 0.5/\tau$, i.e. when the second defect is delayed by half the defect period. This is obvious, as the time between defects is now halved. The defect frequency is consequently doubled. When $\tau = 0$ the demodulation peak is doubled, as expected.

A general expression for M defects may now be postulated given the following assumptions:

- The amplitude and time delay of defect k are given by A_k and τ_k respectively.
- The response $h(t - \tau_k)$ of defect k ends before the start of defect $k + 1$. This ensures that the cross products in equation 125 are zero.
- The response of the system to the defects are all determined by scaled and shifted replicas of $h(t)$.

The desired expression may now be determined by using a binomial expansion given the assumptions listed above

$$\begin{aligned} H_M(f) &= \Im((A_0 h(t - \tau_0) + \dots A_M h(t - \tau_M))^2) \\ &= \Im(A_0^2 (h(t - \tau_0))^2) + \dots \Im(A_M^2 (h(t - \tau_M))^2) + \dots \\ &= H_{M=1}(f)(A_0^2 e^{-j2\pi f\tau_0} + \dots A_M^2 e^{-j2\pi f\tau_M}) \end{aligned} \quad (127)$$

Equation 127 indicates that the amplitude of the demodulation peak owing to M defects will be given by the vector sum of the squared amplitude values $A_k^2 \angle \theta_k$, where phase

angle $\angle \theta_k$ is determined by the ratio of time delay τ_k to period $T = 1/f$ (k varies from 1 to defect M).

It is evident from equation 127 that the combined effect of M defects on the demodulation spectrum is given by a single point defect with impulse response

$$h_M(t) = A_0 h(t - \tau_0) + \dots A_M h(t - \tau_M) \quad (128)$$

The demodulation spectrum of equation 128 would be identical to that of equation 127. Hence it may be concluded that an approximation to the impulse response of equation 128 would approximate the demodulation spectrum well. More specifically, the proposed model given by equation 30 on p. 48 could be tuned in amplitude to best fit the demodulation spectrum against time (refer to figure 34 for an example of the demodulation spectrum against time for a defect signal). The idea would be to determine the parameters of the model (such as the amplitude of a specific defect pulse, perhaps owing to a change in the dynamic loading) so that the demodulation spectrum of the model matches the observed demodulation spectrum against time. By adjusting the time position of the pulse, the effect of small speed variations could also be incorporated.

6.13.2.1.3 Influence of environmental conditions

The introduction of fault conditions by dismantling machinery always changes the vibration levels owing to the changing transfer function of the small defect signal through the complex mechanical interface leading to the vibration sensor. It is therefore preferable to trend vibration measurements during fault evolution (with no dismantling). Unfortunately, it is very difficult to detect small incipient defects in a complex industrial setting. The component would also usually not be replaced at this early stage, causing production losses. Thus it is difficult to establish a machinery database of small defects. Smaller experimental setups, in which environmental factors (such as speed, load and temperature) are controlled, may be used to discover certain general trends and relationships. The influence of period variation may be modelled as a smoothing of the spectrum by a function related to the probability density function of the signal [10]. As a result, high frequency harmonics will tend to smear into each other, as shown in figure 36.

The effect of time domain speed compensation, described in the section 'Time domain speed compensation' on p. 205, is shown in figure 37, in which a repeating series of harmonics is clearly visible.

6.13.2.1.4 Multiple fault conditions

A bearing frequently acts as a support mechanism for a complex mechanical assembly consisting of gearboxes, etc. As it is often impossible to obtain access to the bearings, vibration measurements are often made on the gearbox casing. In [65] the total response is described as a linear superposition of the bearing and gear excitations. The small bearing defect signal could hence be swamped by the large meshing frequencies of the gearbox, making detection very difficult. Furthermore, a deteriorating secondary fault mechanism, e.g. increasing imbalance, can lead to faulty diagnostic decisions [98].

6.13.2.2 Signal processing techniques

6.13.2.2.1 Spectral analysis

Spectral analysis and trending of spectral components have been used extensively in machinery fault diagnosis [86, 98]. As it is difficult to analyse and interpret the phase information from spectral analysis (which often contains critical diagnostic information), much interest has recently been expressed in combined time-frequency analysis methods that preserve phase information (e.g. [65]).

6.13.2.2.2 Time domain analysis

Indicators such as RMS, crest factor, kurtosis and statistical moments are useful for bearing diagnosis [98].

6.13.2.2.3 High frequency resonance technique (HFRT)

In HFRT the signal is bandpass filtered around a suitable demodulation frequency and then rectified and lowpass filtered (envelope detection) [98]. The normalized ratio of the demodulation peak in the demodulation spectrum relative to the carpet level provides a measure of the defect growth. As this ratio is regarded as the best feature of bearing defect evolution [25, 63, 78], no time domain and spectral features have been used in this section.

6.13.3 Problem solution

6.13.3.1 Influence of the buffer length on demodulation spectrum

Figure 34 shows the influence of buffer length on the demodulation peak relative to the carpet level. It is evident that a small buffer length does not allow a sufficient threshold for detection of the demodulation peak above the carpet level. From samples 40 000 to 80 000 very little contribution is made to this ratio in dB. By accumulating the individual contributions over a substantial buffer length reliable detection of a defect becomes possible (refer to figure 34).

6.13.3.2 Effect of speed variation on demodulation peak

Table 5 shows that small amounts of speed variation on a simulated data set have little effect on the magnitude of the demodulation peak, although speed variation does smear the high frequency harmonics, making direct frequency analysis difficult.

Table 5: Effect of speed variation on demodulation peak

BW ⁹⁶	Speed variation %	Demodulation peak amplitude
0.035	0	9604.8
0.035	2	9603.3
0.035	20	9041.1
0.2	0	1963.6
0.2	2	1961.7

⁹⁶ The parameter choice $BW = 0.2$ corresponds to an impulsive source, whereas $BW = 0.035$ seems to match the experimental signal more closely.

6.13.3.3 Time domain speed compensation

Figure 36 and figure 37 show the magnitude spectrum of the simulated time domain signal before and after compensation. The smearing of spectral peaks can be clearly seen in figure 36.

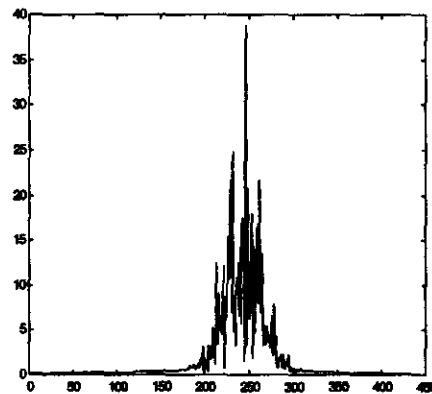


Figure 36: Magnitude spectrum uncompensated (fractional bandwidth of 0.035 and 1% random speed variation)

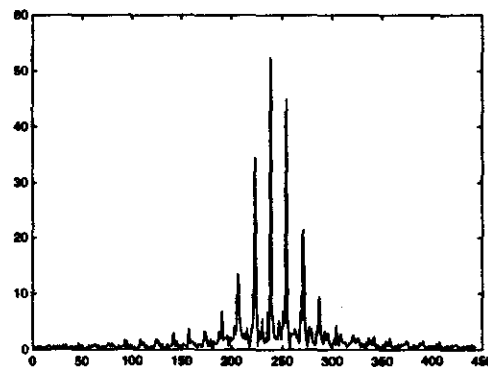


Figure 37: Magnitude spectrum compensated (fractional bandwidth of 0.035 and 1% random speed variation)

The compensation algorithm is described as follows:

Algorithm for time domain speed compensation (pseudo code)

1.	Peaks = vector with sorted peak values ⁹⁷ in time domain signal
2.	Determine SC (Samples to Compensate) $SC = \text{cumsum}^{98}(\text{diff}^{99}(\text{diff}(\mathbf{Peaks})))$ $SC = \sum_{k=1}^N x(k) - 2x(k-1) + x(k-2)$
3.	For $lk = 2$ to $\text{length}(\mathbf{Peaks}) - 1$ $dT = \mathbf{Peaks}(lk + 1) - \mathbf{Peaks}(lk)$ <i>%dT refers to samples between consecutive peaks</i> <i>%K1 is constant defining offset of inserted/</i> <i>%deleted samples (e.g. $K_1 = 0.3$)</i> If $SC > 0$ then delete SC samples at position $\mathbf{Peaks}(lk) + dT * K_1$ else insert SC samples (zeros) at position $\mathbf{Peaks}(lk) + dT * K_1$ end if end lk

The compensation algorithm computes the number of samples to compensate during each period and then inserts or deletes the required samples at a suitable position. As the algorithm uses a peak-finding algorithm to locate significant repeating energy, it is often difficult in experimental data sets to locate the peaks separated by the defect frequency.

⁹⁷ The pickpeak function has been used in Matlab®.

⁹⁸ cumsum – cumulative summation of series $x(k)$.

⁹⁹ diff – difference between consecutive samples of series $x(k)$.

Figure 38 shows this problem in an experimental data set [25]. It can be seen that it is difficult to determine the exact time between peaks, even though a clear demodulation peak is visible in the demodulation spectrum. The compensation algorithm has been applied to the experimental data set in figure 40 (figure 39 shows the uncompensated magnitude spectrum for comparison). Unfortunately, a clear line spectrum, as in figure 37, is not obtained. It is evident from figure 38 that the amplitude modulation, implying a large dynamic range variation as well as multiple responses (peaks) per defect time, contributes to the performance of the simple compensation algorithm on the experimental data set. The spectrum of the compensated signal is shown in figure 40.

Figure 41 shows the ratio of demodulation peak to carpet level on an industrial motor bearing as described in [63], with an outer race bearing defect. Figure 42 indicates that repetitive frequency components are already visible in the raw high frequency spectrum (which is not the case in the data set of figure 39). This would seem to indicate that instantaneous speed variations, owing to instantaneous changes in the contact angle, are smaller in the data set described in [63]. As the demodulation peak is only about 10dB above adjacent frequency components in figure 41 at the end of the buffer, it is not that easy to form a clear diagnostic decision. The carpet level itself was estimated from small frequency components close to the demodulation peak.

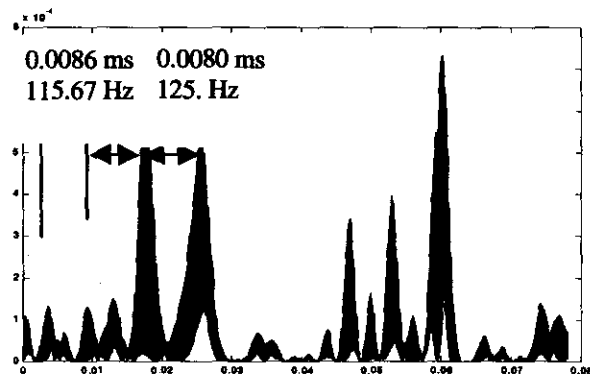


Figure 38: Plot of energy (square) of time domain signal, bandpass filtered around 8.5kHz (horizontal axis in samples)

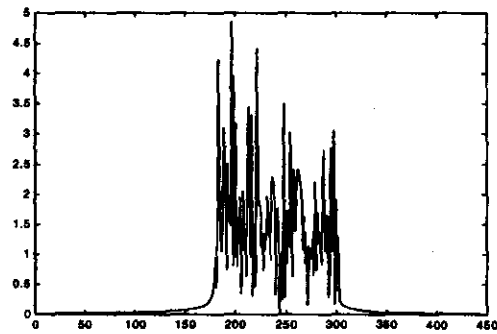


Figure 39: Experimental data set uncompensated

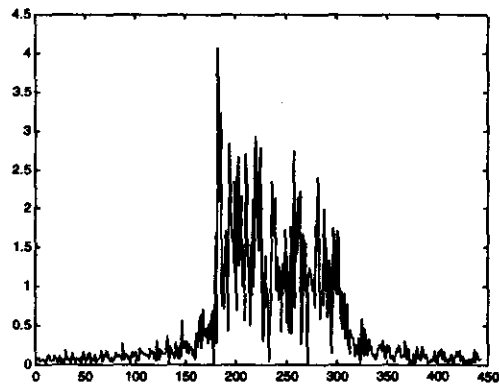


Figure 40: Experimental data set compensated

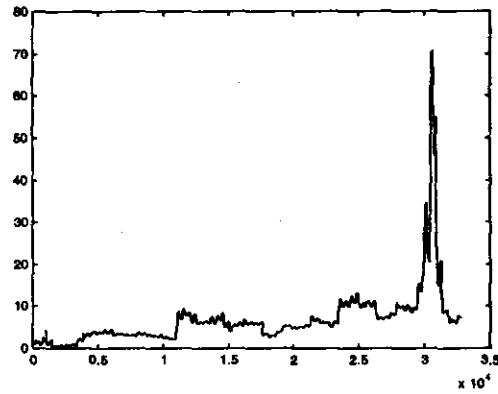


Figure 41: Ratio of demodulation peak relative to carpet level (in dB) versus length of time domain buffer (horizontal axis in samples) [63]

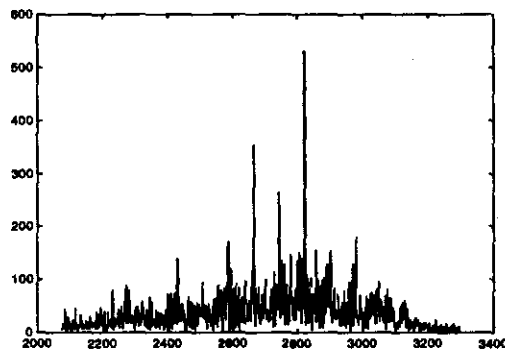


Figure 42: High frequency spectrum (horizontal axis in Hz) [63]

6.13.4 Conclusion

A simple vibration model of a bearing defect has been proposed, building on previous work in this field. The model incorporates a Gaussian modulated sinusoidal source, which seems to mimic the observed vibration signal of the measured data set more closely than an exponentially decaying sinusoid. The model also includes the effect of instantaneous speed variations¹⁰⁰ as proposed in [65], which explains the observed spectral smearing. This model is used to illustrate the effect of speed variations on the demodulation peak, which is the primary feature indicating bearing defect. The influence of multiple defects and buffer length on the demodulation peak is also investigated. An algorithm is proposed to compensate for speed variations, and the ability of this algorithm to resolve the periodic

¹⁰⁰ Speed variations correspond to a perturbation of the starting position of the pulse.

frequency spectra of simulated signals is demonstrated. However, it is shown that the improvement obtained with experimental data sets is not as dramatic. This is likely as a result of the difficulty in obtaining a suitable trigger signal in the time domain. The demodulation spectrum, in contrast, contains a clear demodulation peak. It is hypothesized in this thesis that this is owing to the implicit synchronous averaging that is performed (in addition to the effect of the non-linear operator).

6.14 Conclusion

Various signal processing techniques have been proposed and examined in this chapter for the detection of multiple fault conditions in the presence of noise.

The relation between the SA and demodulation spectrum, using the ESFFT (section 6.3 on p. 117), and the influence of the time window on the SA (section 6.4 on p. 123) were discussed using the proposed ESFFT. Aspects related to the modelling of vibration signals with speed variation were discussed in section 6.6 on p. 132 and section 6.7 on p. 138. The relation between the SA and demodulation spectrum was demonstrated in section 6.8 on p. 139. The minimization and maximization of the kurtosis, as well as interpretation of the kurtosis in the time and frequency domains, were investigated in section 6.9 on p. 145. Section 6.10 on p. 163 used examples to illustrate the effect of the non-linear squaring operator on a faulted vibration signal. A relation was found between the absolute value of the analytical signal derived from the Hilbert transform (the amplitude modulation signal) and the square law detector (section 6.11 on p. 186). The effect of buffer length on the detection of bearing faults was investigated using HFRT. It is concluded that for medium and large bearing defects the buffer length can be quite small (although the frequency resolution would also decrease), but for a small bearing defect with low SNR the buffer length should be as large as possible for reliable diagnosis. Hence small incipient bearing defects would be most reliably detected using a larger buffer length. From the perspective of relating the demodulation peak and harmonics in the frequency domain to the computation of synchronous averages in the time domain, it is clear that increasing the buffer length also increases the number of synchronous averages. For large and medium bearing defects the variation in amplitude of defect pulses is not that large, thus the

synchronous average would not differ much from an individual defect pulse (consequently a smaller number of averages and buffer length could be used). For a small bearing defect the converse applies – the number of averages and buffer length should rather be increased.

To avoid spectral smearing owing to timing variations (e.g. shaft speed) the demodulated resonance should have as low a frequency as possible. However, the system is often most responsive in the high frequency range. Hence a compromise might be necessary if regular spectral (or cepstral) techniques are used. The HFRT (squaring of the bandpass filtered signal or using the Hilbert transform to supply the amplitude modulation function) is not as sensitive to speed variations. Experimental results related to the Vib demo data sets 1, 2 and 3 will be presented in the next chapter.

7 Results of bearing test rig

7.1 Introduction

In this section results of measurements made on a test rig will be presented. The novel aspects regarding the experimental setup and the detection of multiple fault conditions have already been discussed in chapter 4 on p. 59. Firstly, the results on a test rig with normal and faulted bearings (present on the same test rig) will be presented (Vib demo data sets 1 and 2). Secondly, results on the same test rig, but with measurements on the faulted bearing only, will be presented (Vib demo data set 3). The novel technique of moving the defect out of the loading zone for the outer race defect (presented in this thesis) allows the operation of a normal bearing to be simulated without physically dismantling the test rig (or by introducing further uncertainties by using a bearing of different construction as in the Vib demo data sets 1 and 2).

7.2 Experimental results of Vib demo data sets 1 and 2

7.2.1 Introduction

In this section the Vib demo data sets 1 and 2 will be analysed, and relevant features will be extracted for pattern recognition algorithms.

7.2.2 Feature extraction

The influence of dynamic load on the extracted features is shown in the set of figures in appendix B from p. 291, in which the dynamic load (unbalance mass in grams) is shown on the horizontal axis, whereas the feature values are shown on the vertical axis using a logarithmic scale. Each data point is indicated by a (blue) cross and some text next to the data point. The text displayed next to the data point is the defect angle (in degrees), the bearing identification number (1 means the original bearing of the Vib demo test rig with outer race defect, 4 is a smaller outer race defect, etc.), the bearing defect severity on a scale of 10 (the severity of the original bearing of the Vib demo test rig with outer race

defect has been judged to be 5) and finally the position of the defect relative to 225° (owing to the pressure applied by the knurled knob to the bearing plate, the centre of the loading zone is not at 180°, but rather closer to 270°). The following diagram illustrates the numbering convention used for the fourth number, indicating the position of the defect relative to 225° (which is approximately the centre of the loading zone, as the maximum vibration amplitude on the envelope spectrum is found at a defect angle of 225°).

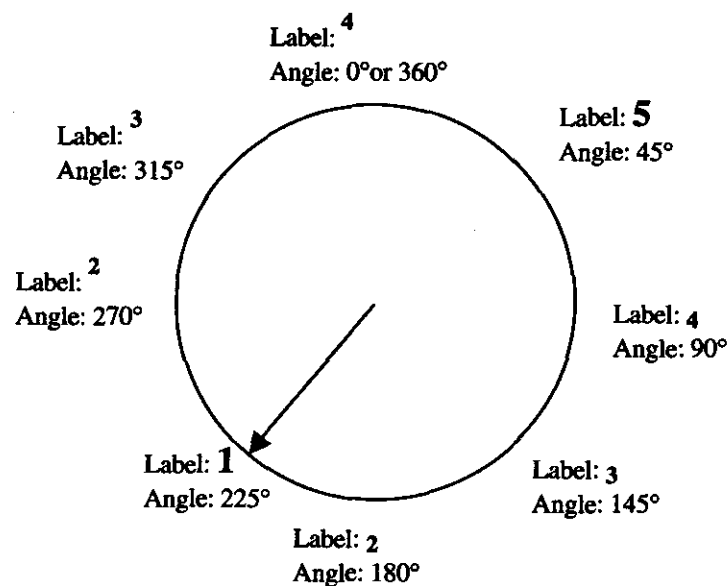


Figure 43: Numbering convention used to indicate proximity of defect angle to centre of load zone

The defect angle of 180°, used in this set of experiments, is far from the centre of the loading zone. Although a clear peak may be observed in the frequency domain component at *BPFO*, it is difficult to observe a regularly repeating series of pulses in the time domain signal itself. The effect of various defect angles is shown in the results on the Vib demo data set 3 in section 7.3 from p. 223, including 225° (when the defect is close to the centre of the loading zone and the defect is clearly visible in the time domain signal itself).

The values of the data points are also tabulated on the right hand side of the figures in appendix B from p. 291 (between the 18g and 24g unbalance masses), starting with an index number (1, 2, 3...), followed by the defect angle in degrees, the bearing number, bearing defect severity, unbalance mass (in grams), the number of applied weights (one or

two), the position of the defect according to the numbering convention of figure 43 and finally the value of the feature in dB ($20\log_{10}(x)$, where x is the feature value).

From the figures in appendix B on p. 291 it is apparent that the feature values related to bearing failure generally decrease slightly with an increase in the dynamic load. The feature Fr_X (refer to appendix B from p. 291 for details about the extracted features) shows an increase with an increase in imbalance, as would be expected. The $BPFO\_X$ feature decreases monotonically with an increase in load. The $BPFO2\_X$ feature also decreases with an increase in load, although the relationship is much more complex. The synchronous average and the HFD features also decrease monotonically with an increase in load. The other features show much more complex behaviour, although the general trend is a decrease with an increase in load. The shaft speed increases slightly when imbalance is increased. The $BPFO$ and $BPMI$ frequencies also increase slightly when imbalance is increased.

In summary, the rotational speed feature Fr_X behaves as expected theoretically. The $BPFO2\_X$ feature (indicative of an outer race bearing fault) generally decreases with an increase in (dynamic) load. Although at first this may seem slightly counter-intuitive, it is in line with other observations in literature. The other features display more complex behaviour and will be discussed in more detail in the Vib demo data set 3 in section 7.3 from p. 223.

7.2.3 Modelling of Vib demo data set 2

A classifier is typically used to detect a fault condition from a data set containing normal and fault data. Each class represents a specific state (for example, class 1 could be the normal data and class 2 the outer race bearing defect data). The classes are usually represented by discrete values (e.g. 0 and 1 to indicate degree of membership to a class). Usually the fault condition occurs rarely and hence the size of the normal data is much larger than the fault data. Consequently, the first preprocessing step is to balance the data set so that each class contains approximately the same number of observations. A simple linear regressor may be tried as a classifier. If not successful, a non-linear regressor (such as a neural network) has to be used. However, the data requirements also increase as the

complexity of the neural network classifier increases (as a rule of thumb about 10 data points for each free parameter in the neural network). A classifier is typically useful when a large data set (with enough observations from each class) is available and a simple yes/no diagnostic decision (fault/no fault) needs to be made.

An alternative is to model the fault condition(s) using feature values as input to a neural network model. In this case the output of the neural network varies continuously (for example from 0 to 1) instead of discretely (as with a classifier). In this way, it might be possible to augment the original data set using the neural network model.

Several different approaches may be used for modelling of the data set with ANN. Firstly, a fixed number of known classes may be used (typically identified through an exploratory analysis using SOM techniques). Alternatively, a classification problem may be used, with each class representing a well-defined fault condition, and then the number of classes may be increased as additional data become available.

Secondly, class membership may be indicated either by an associated output value of 1 (whereas a 0 indicates the opposite), or the outputs of the ANN may be used to indicate, on a continuous scale, the severity of the fault mechanism present.

Results based on the second approach are shown in the next set of figures in appendix B from p. 306. A radial basis function (RBF) neural network has been trained to distinguish the following six classes:

- Unbalance masses of 0g, 12g and 24g (with no bearing defect) – classes 1 to 3
- Unbalance masses of 0g, 12g and 24g (with an outer race bearing defect) – classes 4 to 6

Figure 49 shows the output of the RBF for all six classes. It may be observed that the imbalance condition increases from left to right on the right hand axis. The outputs of the RBF also increase from left to right on the right hand axis. Furthermore, the imbalance condition increases on the left hand axis. The outputs of the RBF also increase on the left hand axis. The same information is also presented in figure 93 to figure 98 as separate

graphs for each combination of primary and secondary fault mechanism. It is also evident that the RBF had generalized, to some extent, beyond the original training set data. Thus it has interpolated between the original training set data. This is evident from figure 52 by observing that the output of the 12g RBF output is approximately $\frac{1}{2}$ when presented with an imbalance feature halfway between 12g and 24g (representing 18g). The same is true using a 24g RBF output. Presenting an imbalance feature of 18g (with no defect) therefore results in an output of approximately $\frac{1}{2}$ for the 12g and 24g outputs, representing the respective degree of membership towards each class (i.e. approximately 18g).

Figure 50 and figure 51 show the outputs of the trained RBF network for the 18g test set data. It can be seen that the 24g RBF output only is activated. This is also evident from the scatterplot of figure 44. Figure 52 shows the outputs of the trained RBF network on the data from the normal bearing only. Figure 53 shows the outputs of the trained RBF network on the test set data from the defective bearing. It is evident that the 24g data from the defective bearing have been misclassified by the RBF when trained on the normal bearing data only. The requirement for data from both the normal and defective bearings is thus clearly established. Figure 99 to figure 104 show all the outputs of the RBF network individually for training set data belonging to each fault condition, again trained on the data from the normal bearing only. Figure 104 again confirms the requirement for comprehensive training set data.

A multilayer perceptron (MLP) has also been trained on the same data, but using fewer classes. The ability of the MLP neural network to converge has been observed to decrease as the number of classes increases. The MLP also experienced problems distinguishing between the 18g and 24g masses (the feature values are quite close together, resulting in an ill-posed inverse problem). Further experiments may be done in future on this type of network using different parameters, e.g. a linear transfer function at the output. At present the RBF seems to train better with multiple outputs than the MLP.

Initial experiments with the MLP indicate that it is able to model the degree of imbalance with a continuous output (representing the second approach mentioned earlier), although

the MLP often had trouble distinguishing between closely separated masses during training.

The RBF classifier achieved perfect separation of classes with all measurements present. As can be seen from figure 104, it is no longer possible to separate the classes using the measurements from the normal bearing only. The average value of the 0g RBF network output for the measurements on the normal bearing is around 0.85, as shown in figure 102. If the 0g RBF network output represents a 'no or small imbalance' condition, the average membership value of 0.85 represents a small amount of imbalance, while it is known that no imbalance has indeed been applied. Similarly, if the 24g RBF network output of figure 104 represents a 'high imbalance' condition, the average membership value of 0.25 represents a small amount of 'high imbalance', while it is known that a high imbalance has indeed been applied. The influence of the secondary fault mechanism is therefore clearly evident.

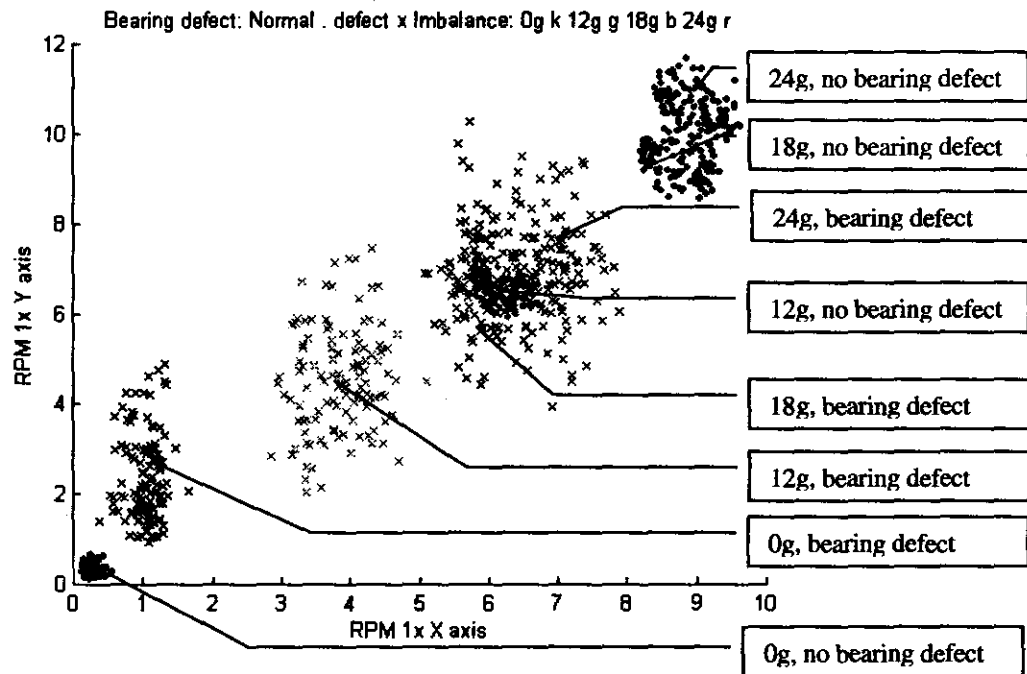


**Figure 44: Scatterplot of rotational frequency (RF) features (x and y axes)
– Vib demo data set 2**

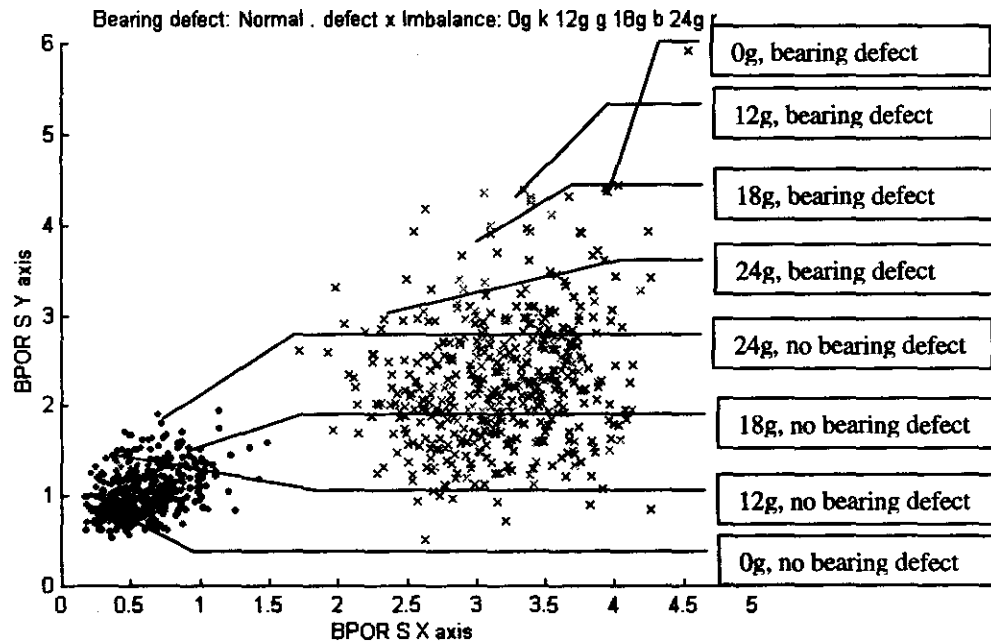


Figure 45: Scatterplot of BPOR feature (x and y axes) – Vib demo data set 2

	Feature 1	Feature 2	Feature 3	Feature 4
0g				
12g	Normal bearing			
18g				
24g				
0g				
12g	Defect bearing			
18g				
24g				

Figure 46: Schematic representation of feature matrix for the SOM technique – Vib demo data set 2 (refer to [98] for details)

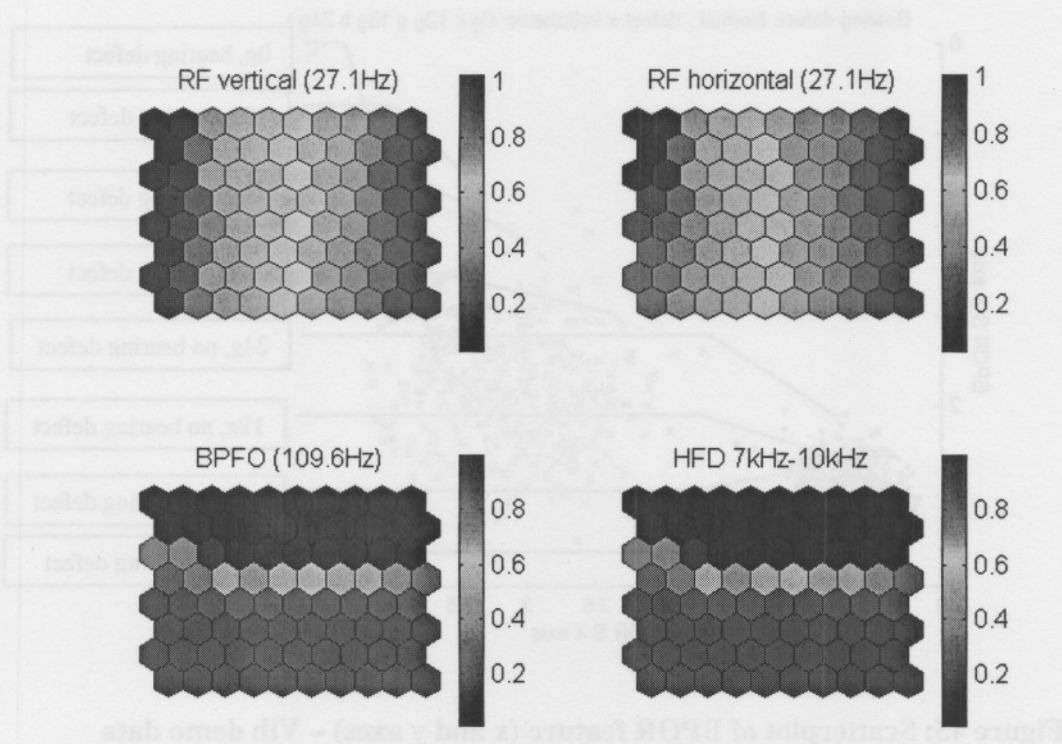


Figure 47: The SOM displaying all four features – Vib demo data set 2 (refer to [98] for details)

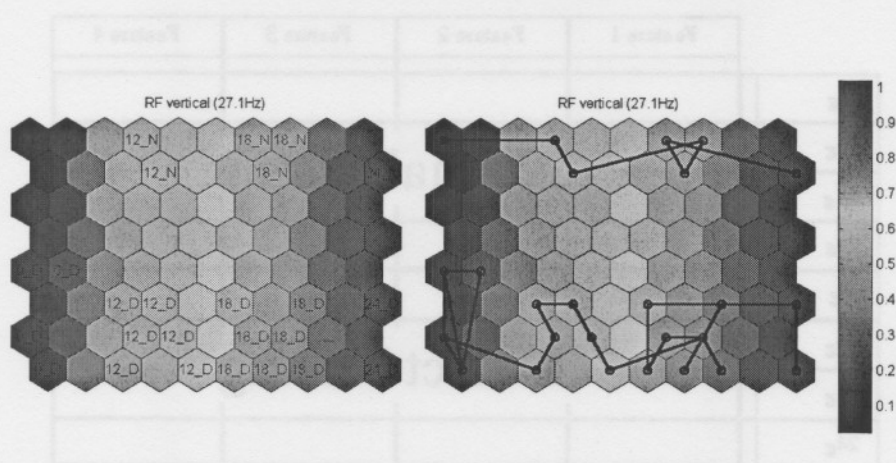


Figure 48: Class labels and best matching unit trajectory on the SOM – Vib demo data set 2 (refer to [98] for details)

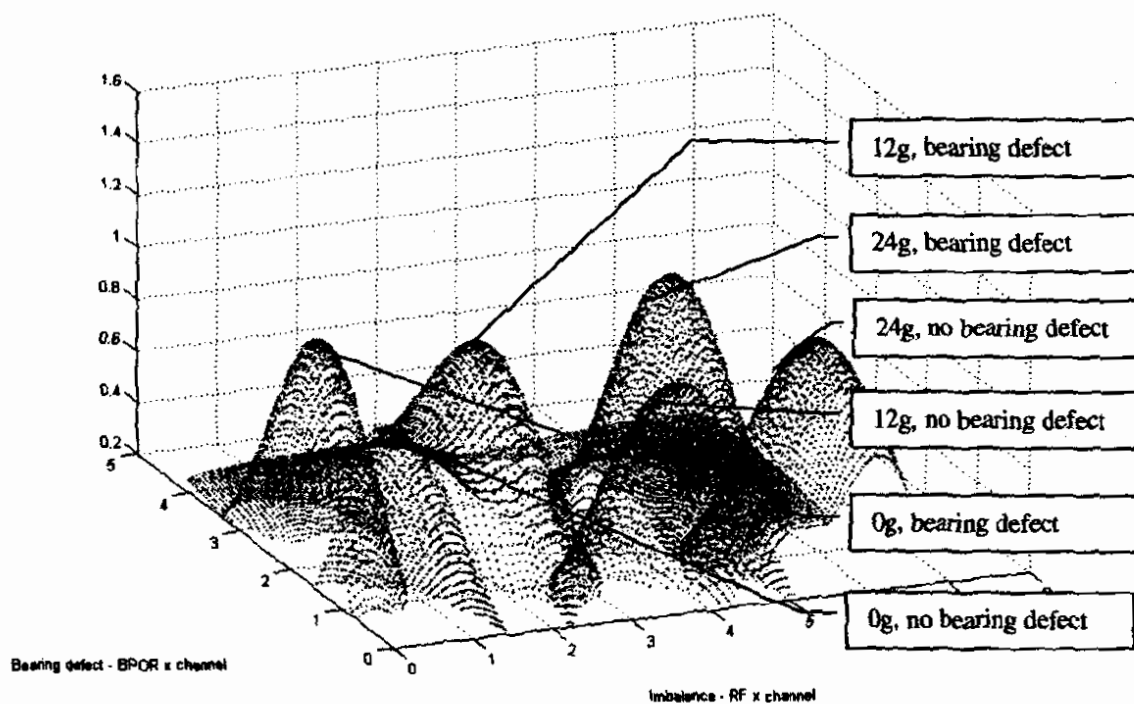


Figure 49: Outputs of RBF for all six classes on own training set data; bearing defect is shown on the left hand axis, imbalance on the right hand axis – Vib demo data set 2

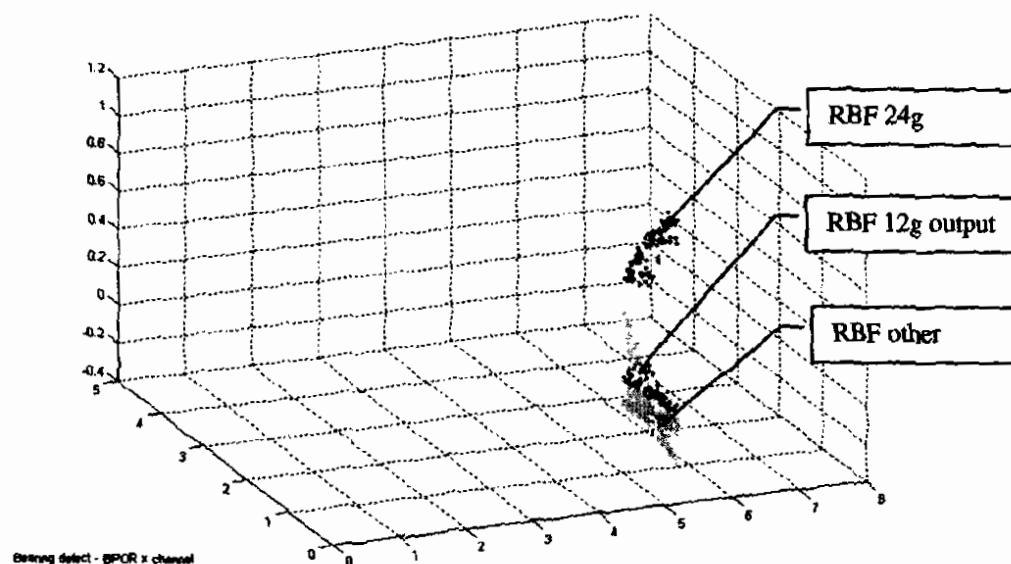


Figure 50: Outputs of RBF for 18g mass with no bearing defect; bearing defect is shown on the left hand axis, imbalance on the right hand axis – Vib demo data set 2

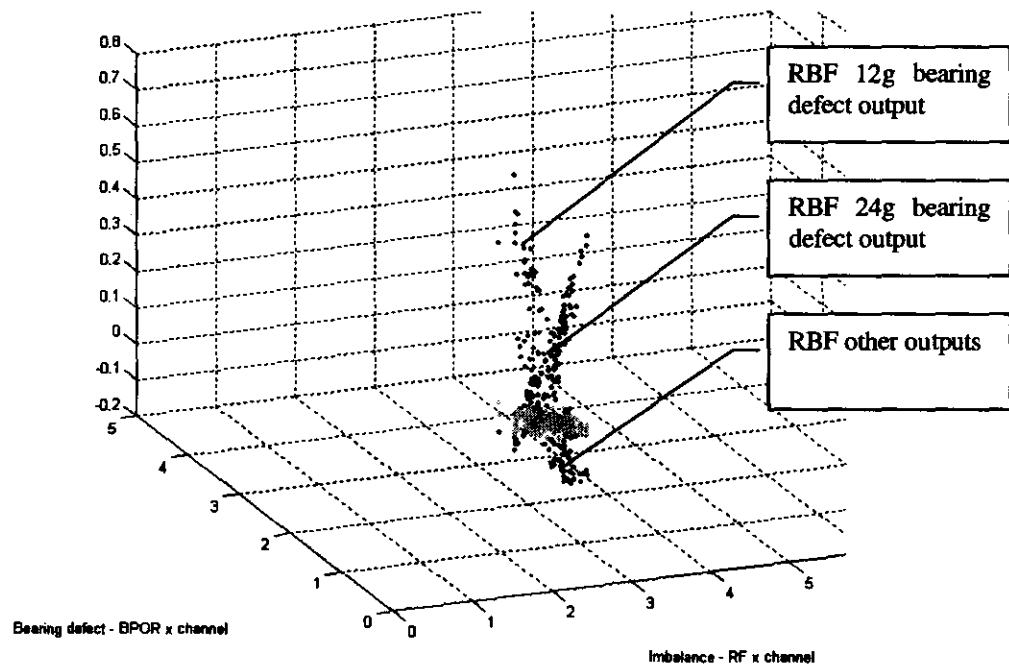


Figure 51: Outputs of RBF for 18g mass with bearing defect; bearing defect is shown on the left hand axis, imbalance on the right hand axis – Vib demo data set 2

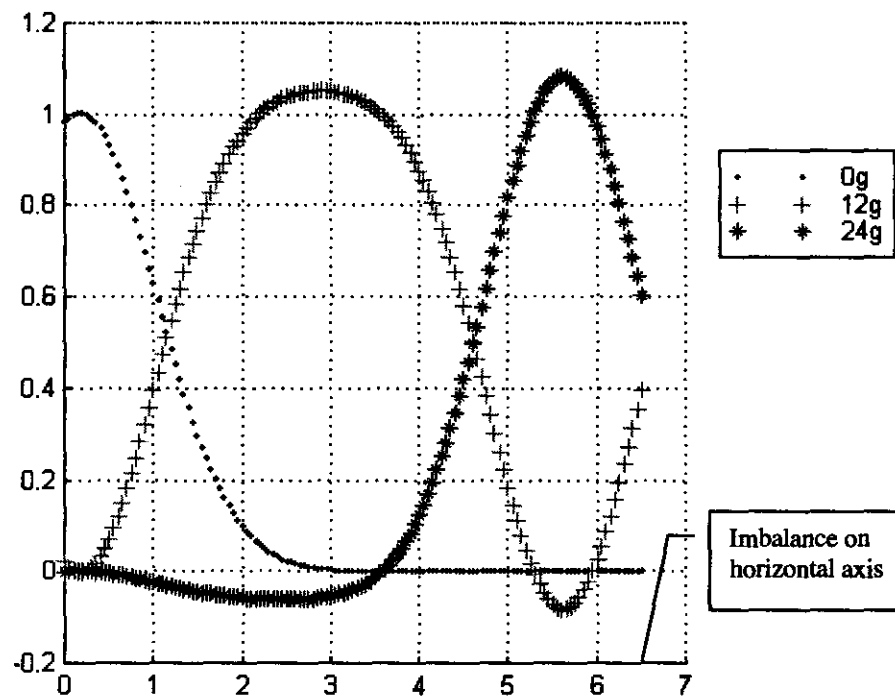


Figure 52: Outputs of the 0g, 12g and 24g RBF for all the training set data (actual measurements on the normal bearing) – Vib demo data set 2

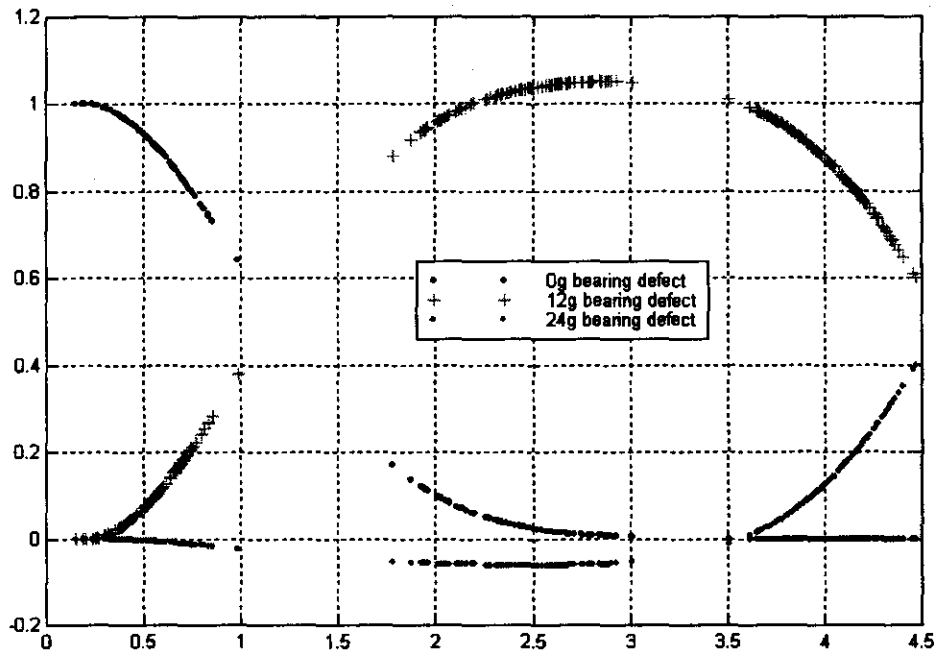


Figure 53: Outputs of the 0g, 12g and 24g RBF for all the training set data (measured on the defective bearing) – Vib demo data set 2 (imbalance on horizontal axis)

7.3 Experimental results of Vib demo data set 3

7.3.1 Introduction

In this section brief experimental results will be presented related to the problem statement of this thesis, which is the recognition of multiple fault conditions. A number of features will be extracted and the influence of bearing defect (large, medium and small) and dynamic load (unbalance masses of 0g, 12g, 18g and 24g) will be shown. The novel technique of varying the defect angle (DA) is applied to all the features. A buffer length of 120 000 samples (about 5 seconds at 25.6kHz) is used throughout the experiments in this section, unless explicitly stated otherwise (section 7.3.8 examines the influence of buffer length).

Note that some of the measurements of the small bearing defect at defect angles of 315°, 360°, 45°, 90° and 135° have been made using a sampling rate of 51 200Hz (20kHz bandwidth) instead of the usual 25 600Hz (10kHz bandwidth). In this section a sample rate converter has been used to resample the signals at 51 200Hz to 25 600Hz.

Please note that the resampling might therefore have some unknown influence on the results for the small bearing defect where $DA = [315^\circ, 360^\circ, 45^\circ, 90^\circ, 135^\circ]$.

7.3.2 Application of defect angle (DA)

The change in defect angle may be used also to predict an inner race time domain defect signal, as can be seen for a large outer race defect from feature *BPFO2_X* (the suffix *_X* means the vibration measurement on the x axis) on p. 329 in which the maximum to minimum values are $DA = [225^\circ, 270^\circ, 315^\circ, 90^\circ, 135^\circ, 360^\circ]$ ¹⁰¹. The difference between smallest (defect angle = 360°) and largest (defect angle = 225°) defect signal values is huge, about 50dB (also from feature *BPFO2_X* on p. 329). When the defect angle = 360° the outer race bearing defect is far from the loading zone. Since the effect of the large bearing defect is effectively removed, it can also be used to predict the time domain defect signal of a 'new' bearing (when the defect angle is close to 360° , i.e. the defect is out of the loading zone). It is not necessary to open the bearing physically, although it is necessary to loosen the screw in the test rig.

7.3.3 Feature *BPFO2_X*

For the medium bearing defect, the results are given in table 6¹⁰². There are no real surprises. Note that the demodulation peak at defect angle $DA = 270^\circ$ decreases by about 10dB on the application of a dynamic load of 24g. The difference between maximum and minimum values is 30dB for a dynamic load of 0g ($DA = 270^\circ$) and 20dB for a dynamic load of 24g ($DA = 270^\circ$). The difference between maximum and minimum values is 40dB for a dynamic load of 0g ($DA = 225^\circ$), slightly smaller compared with the value of 50dB for the large defect.

¹⁰¹ The influence of dynamic load is reasonably small (a few dB).

¹⁰² Each defect is shown in a separate table for ease of reference – in table 8 all the defects are shown together.

Table 6: Defect angle (DA in degrees) and value of demodulation peak (BPFO2_X in dB) versus dynamic load for a medium defect

0g DA	0g BPFO2_X	12g DA	12g BPFO2_X	18g DA	18g BPFO2_X	24g DA	24g BPFO2_X
225°	46	270°	37	315°	34	270°	27
270°	37			90°	17	45°	16
315°	34			45°	13	135°	7
90°	16			135°	12	360°	6
45°	13			360°	7		
135°	9						
360°	6						

For the small bearing defect, the results are given in table 7. Note that the demodulation peak at DA = 225° stays constant with an increase in dynamic load. The difference between maximum and minimum values is 19dB only for dynamic loads of 0g and 24g. Hence it becomes much more difficult to distinguish the bearing defect signal from the background noise.

Table 7: Defect angle (DA in degrees) and value of demodulation peak (BPFO2_X in dB) versus dynamic load for a small defect

0g DA	0g BPFO2_X	12g DA	12g BPFO2_X	18g DA	18g BPFO2_X	24g DA	24g BPFO2_X
225°	20	90°	5	135°	9	225°	21
270°	13	360°	3	45°	4	270°	12
315°	13	135°	2	90°	2	360°	10
360°	11					315°	2

45°	7						
90°	2						
135°	1						

The influence of load and defect size versus defect angle is shown in table 8.

Table 8: Dynamic load (imbalance in grams) and value of demodulation peak (*BPF02_X* in dB) versus defect angle (DA in degrees) and defect size (S, M, L for small, medium and large respectively); the symbol NA means measurement is not available

DA	225°	225°	225°	270°	270°	270°
S, M, L	L	M	S	L	M	S
0g	60	46	20	58	37	13
12g	NA	NA	NA	NA	37	NA
18g	62	NA	NA	NA	NA	NA
24g	NA	NA	21	NA	27	12
DA	90°	90°	90°	45°	45°	45°
S, M, L	L	M	S	L	M	S
0g	21	16	2	NA	13	7
12g	NA	NA	5	NA	NA	NA
18g	26	17	2	15	13	4
24g	NA	NA	NA	NA	16	NA

DA	360°	360°	360°	315°	315°	315°
S, M, L	L	M	S	L	M	S
0g	13	6	11	48	34	13, 4
12g	NA	NA	3	NA	NA	NA
18g	8, 3	7	NA	48	34	NA
24g	NA	6	10	NA	NA	2
DA	135°	135°	135°			
S, M, L	L	M	S			
0g	19	9	0.6			
12g	NA	NA	1.5			
18g	21	12	9			
24g	NA	7	NA			

The feature *BPFO2_X* can separate all the defects at $DA = 225^\circ$, $DA = 270^\circ$, $DA = 315^\circ$ and 0g. At $DA = 360^\circ$ it is no longer possible to distinguish any faults reliably (as expected, placing the defect out of the loading zone makes it more difficult to recognize faults).

In equation 36 derived by Zhang, the RMS value was shown to increase with the square of the defect width under simplifying assumptions (valid for small defects). Hence a doubling of the defect width implies an increase of $20\log(2^2) = 12\text{dB}$ in the RMS value. However, the feature *BPFO2_X* uses the square of the RMS value, so doubling of the defect width implies an increase of $20\log(4^2) = 26\text{dB}$ in the square of the RMS value. From the SEM photos of figure 216 and figure 221 it can be seen that the medium outer race defect is almost twice the width of the small outer race defect, which is in agreement with the difference of 26dB in the *BPFO2_X* value for $DA = 225^\circ$ from table 8. Thus the

experimental results seem to correlate well with literature. Nevertheless, the aim of this thesis is not to present a detailed analysis of the dynamic forces involved and this is left for future work.

When the defect is in the centre of the loading zone for a medium bearing defect at $DA = 270^\circ$, a decrease in the value of $BPFO2\_X$ with increasing dynamic load (imbalance) is evident (about 10dB down from imbalance of 0g to 24g). A possible explanation for this would be that the large dynamic load effectively changes the load distribution from the static load distribution. Hence the load will increase slightly when the unbalance mass is close to the defect, but during the rest of the shaft revolution the load will decrease (the roller will contact the outer race defect about four times during one revolution of the shaft, and during three of the four contacts the dynamic load will be smaller). Hence the $BPFO2\_X$ feature may decrease with increasing unbalance mass, which is also documented (for a static load) in literature [6].

The difference of about 10dB when $DA = 225^\circ$ (small defect in the loading zone) versus $DA = 360^\circ$ (small defect out of the loading zone) and 20dB for $DA = [90^\circ, 135^\circ]$ is less than the 25dB of the small to medium defect. Thus it could be postulated that the defect size of the 'normal' bearing signal resulting from $DA = 360^\circ$ is about half the size (in the order of $10\mu\text{m}$) of the small defect according to Zhang's calculations [107]. For $DA = [90^\circ, 135^\circ]$ the estimated defect size would be $\frac{1}{4}$ (or $5\mu\text{m}$).

7.3.4 Feature $BPFO\_X$

The $BPFO\_X$ feature is not as reliable as the feature $BPFO2\_X$ in defect detection, compare, for example, feature $BPFO\_X$ on p. 328 with feature $BPFO2\_X$ on p. 329. It is clear that the value of $BPFO\_X$ (for example at $DA = 225^\circ$) is not always the maximum value using the feature $BPFO\_X$.

7.3.5 Cepstral features

The linear cepstrum is almost as effective as $BPFO2\_X$ in detecting faults (the maximum to minimum values are $DA = [270^\circ, 225^\circ, 315^\circ, 90^\circ, 360^\circ, 135^\circ]$ for a large defect and 0g imbalance – refer to p. 332). The real cepstrum is slightly worse owing to the $\log_{10}$

operation (the maximum to minimum values are $DA = [270^\circ, 315^\circ, 225^\circ, 90^\circ, 360^\circ, 135^\circ]$ for a large defect and 0g imbalance – refer to p. 332).

7.3.6 Feature *FFT2_MEAN_X*

The feature *FFT2_MEAN_X* calculates the power of the signal (the mean square value). For the medium and large bearing defects the bearing defect signal contributes significantly to the total signal power (hence the feature may be used for fault detection). However, any secondary fault mechanism such as dynamic load (imbalance in table 9) will also influence this feature. Hence it is not as selective as the *BPFO2_X* feature.

Table 9 shows that the feature *FFT2_MEAN_X* can separate all the defects at $DA = 225^\circ$, $DA = 270^\circ$, $DA = 315^\circ$ and 0g. At $DA = 90^\circ$, $DA = 135^\circ$, $DA = 45^\circ$ and $DA = 360^\circ$ it is no longer possible to distinguish any faults reliably. The feature *FFT2_MEAN_X* can separate the large defect at $DA = 225^\circ$ and 0g (refer to p. 341), but is not as sensitive as *BPFO2_X* for smaller defects.

Table 9: Dynamic load (imbalance in grams) and feature *FFT2_MEAN_X* (in dB) versus defect angle (DA in degrees) and defect size (S, M, L for small, medium and large respectively); the symbol NA means measurement is not available

DA	225°	225°	225°	270°	270°	270°
S, M, L	L	M	S	L	M	S
0g	64	48	35	61	41	34
12g	NA	NA	NA	NA	41	NA
18g	64	NA	NA	NA	NA	NA
24g	NA	NA	36	NA	38	33
DA	90°	90°	90°	45°	45°	45°
S, M, L	L	M	S	L	M	S
0g	34	32	33	NA	33	33

12g	NA	NA	33	NA	NA	NA
18g	35	33	33	31	33	34
24g	NA	NA	NA	NA	33	NA
DA	360°	360°	360°	315°	315°	315°
S, M, L	L	M	S	L	M	S
0g	32	30	34	50	40	34
12g	NA	NA	33	NA	NA	NA
18g	30, 32	31	NA	50	40	NA
24g	NA	31	35	NA	NA	33
DA	135°	135°	135°			
S, M, L	L	M	S			
0g	33	35	33			
12g	NA	NA	33			
18g	33	35	34			
24g	NA	35	NA			

For the small bearing defect the maximum to minimum values for the defect angle (using feature *FFT2_MEAN_X*) $DA = [225^\circ, 315^\circ, 270^\circ, 45^\circ, 360^\circ, 90^\circ, 135^\circ, 315^\circ]$ is given by $[35.0539, 34.4332, 34.0434, 33.5746, 33.5478, 33.311, 32.8744, 32.7057]$. Compare this with the $DA = [225^\circ, 270^\circ, 315^\circ, 90^\circ, 135^\circ, 360^\circ]$ for the large bearing defect with feature *SA_BPFO2_X*. The $DA = 315^\circ$ in the small bearing defect using the feature *FFT2_MEAN_X* seems too large. Note the very small dynamic range (2dB) of the measurements in the feature *FFT2_MEAN_X*, which is a consequence of the small bearing

defect (in comparison the large bearing defect has a dynamic range of 30dB and the medium bearing defect has a range of 13dB).

7.3.7 Feature SA_BPFO2_0_X

Table 10: Dynamic load (imbalance in grams) and feature SA_BPFO2_0_X (in dB) versus defect angle (DA in degrees) and defect size (S, M, L for small, medium and large respectively); the symbol NA means measurement is not available

DA	225°	225°	225°	270°	270°	270°
S, M, L	L	M	S	L	M	S
0g	-31.6	-47.1	-54.7	-34.4	-52.0	-48.8
12g	NA	NA	NA	NA	49.5	NA
18g	38.7	NA	NA	NA	NA	NA
24g	NA	NA	-54.4	NA	-49.1	-53.9
DA	90°	90°	90°	45°	45°	45°
S, M, L	L	M	S	L	M	S
0g	-52.2	-50.8	-53.0	NA	-54.1	-52.5
12g	NA	NA	-42.4	NA	NA	NA
18g	-49.3	-44.6	-53.1	-51.6	-44.5	-53.6
24g	NA	NA	NA	NA	43.5	NA
DA	360°	360°	360°	315°	315°	315°
S, M, L	L	M	S	L	M	S
0g	-45.2, -50.7	-50.6	-40.9	-42.3	-44.9	-55.3, -54.8
12g	NA	NA	-57.3	NA	NA	NA

18g	-50.5, -53.3	-53.4	NA	-42.3	-50.3	NA
24g	NA	-48.1	-46.3	NA	NA	-54.6
DA	135°	135°	135°			
S, M, L	L	M	S			
0g	-49.1	-46.6	-53.7			
12g	NA	NA	-53.2			
18g	-50.6	-49.4	-52.7			
24g	NA	-42.6	NA			

Table 10 shows that the synchronous average feature $SA\_BPFO2\_0\_X^{103}$ performs almost as well as $BPFO2\_X$ for large bearing defects. However, for small bearing defects, this feature performs very poorly for 0g imbalance (dynamic load) and with $DA = [360^\circ, 270^\circ, 45^\circ, 90^\circ, 135^\circ, 225^\circ, 315^\circ]$. For the medium bearing defect the performance of the feature is also degraded relative to the $BPFO2\_X$ feature: $DA = [315^\circ, 135^\circ, 225^\circ, 360^\circ, 90^\circ, 270^\circ, 45^\circ]$.

A possible explanation for this could be found in the variation of the trigger frequency with defect size in table 11. When the defect is near the centre of the loading zone ($DA = [225^\circ, 270^\circ, 315^\circ, 135^\circ, 90^\circ, 45^\circ]$) and 0g imbalance, the defect frequency generally increases with increasing bearing defect size. When the defect is out of the loading zone ($DA = [360^\circ]$) and 0g imbalance, the defect frequency (generally) decreases with increasing bearing defect size¹⁰⁴. The computation of the synchronous average is dependent on the trigger frequency. Hence when the defect is out of the loading zone, the

¹⁰³ The same results are generally valid for the other synchronous average features, even though they are not shown.

¹⁰⁴ The values for the medium and large bearing defects are about the same.

trigger frequency can increase for the small defect (refer to table 11). This may cause an increase in the value of the synchronous average feature for small defects out of the loading zone, which would help to explain the poor performance of the synchronous average feature on the small defect. Another possibility would be speed or amplitude variations as well as multiple defects (resulting from surface irregularities). All would contribute to the non-periodic components, which the SA would not detect. In contrast to the SA, the demodulated spectrum can detect the non-periodic components, thus allowing successful detection of the small defect.

Table 11 also indicates that it is possible to determine the bearing condition (on this test rig) for some defect angles from an examination of the outer race defect frequency (feature *BPFO2_S_X* in Hz) when the defect is in the loading zone. The expectation would be that an examination of the instantaneous shaft angle would provide additional information about the condition of the bearing (and would change with bearing defect but also with environmental conditions such as speed and load)¹⁰⁵. Unfortunately, a trigger signal is not easy to obtain in an industrial setting. A further discussion is beyond the scope of this thesis and may be examined in future work.

Table 11: Dynamic load (imbalance in grams) and outer race defect frequency *BPFO2_S_X* (in Hz) feature versus defect angle (DA in degrees) and defect size (S, M, L for small, medium and large respectively); the symbol NA means measurement is not available

DA	225°	225°	225°	270°	270°	270°
S, M, L	L	M	S	L	M	S
0g	115.0	112.6	113.5	114.3	113.9	113.3
12g	NA	NA	NA	NA	113.9	NA
18g	115.2	NA	NA	NA	NA	NA

¹⁰⁵ More specifically, any local discontinuity, such as a bearing defect, would lead to an instantaneous change in the shaft angle. Hence an examination of the changes in shaft angle could also provide diagnostic information about the surface of the bearing (and possible fault conditions).

24g	NA	NA	113.3	NA	112.2	113.3
DA	90°	90°	90°	45°	45°	45°
S, M, L	L	M	S	L	M	S
0g	114.1	113.9	113.5	NA	113.5	113.1
12g	NA	NA	113.3	NA	NA	NA
18g	114.3	113.7	117.1	114.3	113.5	113.1
24g	NA	NA	NA	NA	113.7	113.3
DA	360°	360°	360°	315°	315°	315°
S, M, L	L	M	S	L	M	S
0g	114.3, 114.6	114.1	115.4	113.7	112.9	113.7, 116.3
12g	NA	NA	115.0	NA	NA	NA
18g	114.3, 114.1	114.1	NA	113.7	112.9	NA
24g	NA	113.5	115.0	NA	NA	113.3
DA	135°	135°	135°			
S, M, L	L	M	S			
0g	114.6	112.2	113.3			
12g	NA	NA	113.1			
18g	114.8	112.6	113.1			
24g	NA	112.6	NA			

Note also that the margin between large and medium defects for 0g imbalance in feature *SA_BPFO2_0_X* of table 10 is about 16dB, but only about 6dB between the medium and small defects (which would make the recognition of a small bearing defect much more difficult).

7.3.8 Feature *BPFO2_X* – influence of buffer length

The influence of the dynamic load on the value of the demodulation spectrum *BPFO2_X* for the medium bearing defect will now be illustrated by examining the demodulation peak as a function of buffer length. Intuitively one would expect a smooth, continuous progression of the demodulation peak versus buffer length if the instantaneous load on the roller (hence on the outer race defect) remains constant. Figure 58 on p. 239 shows the demodulation peak versus buffer length for a **medium** bearing defect and 0g dynamic load, whereas figure 59 shows a zoom plot. The demodulation peak progresses smoothly versus buffer length. The introduction of a 24g imbalance (dynamic load) produces a more discontinuous graph as can be seen in figure 60. In particular, note that the demodulation peak tends to stay constant for a few cycles of the bearing defect, and then increases in value for the next few cycles – repeating the same trend over and over – as shown in the zoom plot of figure 61¹⁰⁶. This would seem to support the hypothesis that the introduction of a dynamic load changes the load distribution. With no dynamic load the progression on each defect is much smoother, as was shown in figure 59. Note also the much lower final value of the demodulation peak with dynamic load (figure 60).

For a **large** bearing defect, figure 62 shows the demodulation peak versus buffer length for a large bearing defect and 0g dynamic load, whereas figure 63 shows a zoom plot. The demodulation peak increases smoothly (nearly linearly) with time (buffer length). The introduction of an 18g imbalance (dynamic load) produces a rather smooth graph as can be seen in figure 64 and in the zoom plot of figure 65. Hence the dynamic load seems to

¹⁰⁶ By contrast, the progression is much smoother for a large bearing defect. As the buffer length increases, a new cycle of the defect signal is included. If the defect signal is modelled as an ideal impulse sequence, the progression would occur at the position of the defect pulses (which is almost what happens for the large bearing defect).

have little influence on the smoothness of the load distribution for a very large bearing defect. The exact reason is not clear and is left for future work.

As stated in section 3.6 the problem statement does not primarily involve the detailed kinematic analysis of the mechanical system associated with the test rig. However, one possibility would be that the (very large) ridge for a very large bearing defect (refer to SEM photo in section 11.4) contacts the roller regardless of the instantaneous load on the roller. Thus the roller always contacts the defect (the ridge) regardless of the load on the roller. For a medium bearing defect (where the bearing defect is near the centre of the loading zone, i.e. $DA = 225^\circ$ for the test rig) the ridge is much smaller, hence the load could have a larger effect.

Another example would be a large bearing defect with bearing defect close to the centre of the loading zone (for example $DA = 180^\circ$ in the Vib demo data set 2). In this case the progression of the demodulation peak versus buffer length is less smooth (refer to figure 54).

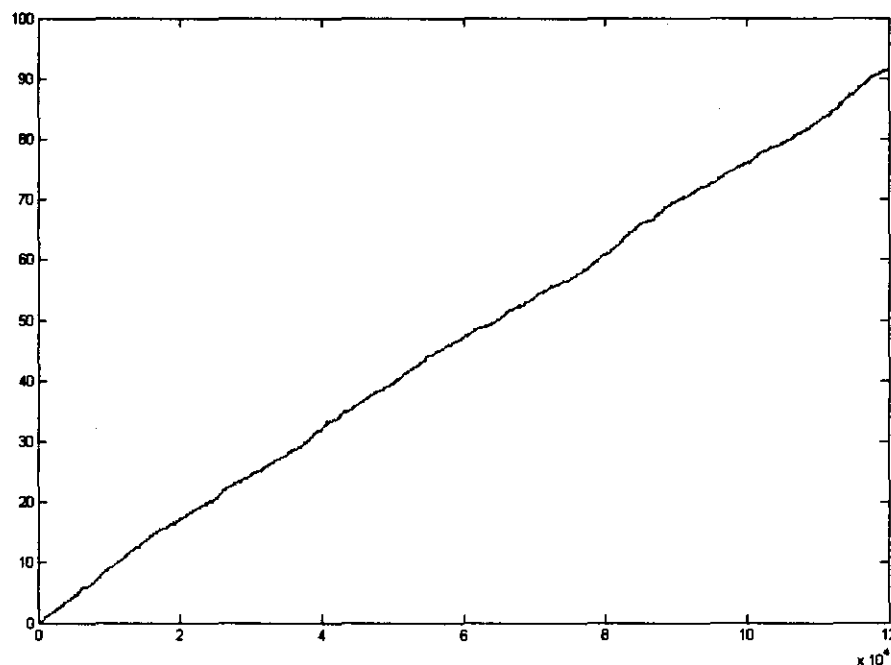


Figure 54: Demodulation peak (magnitude spectrum value) versus buffer length for large bearing defect ($DA = 180^\circ$) and 0g (Vib demo data set 2)

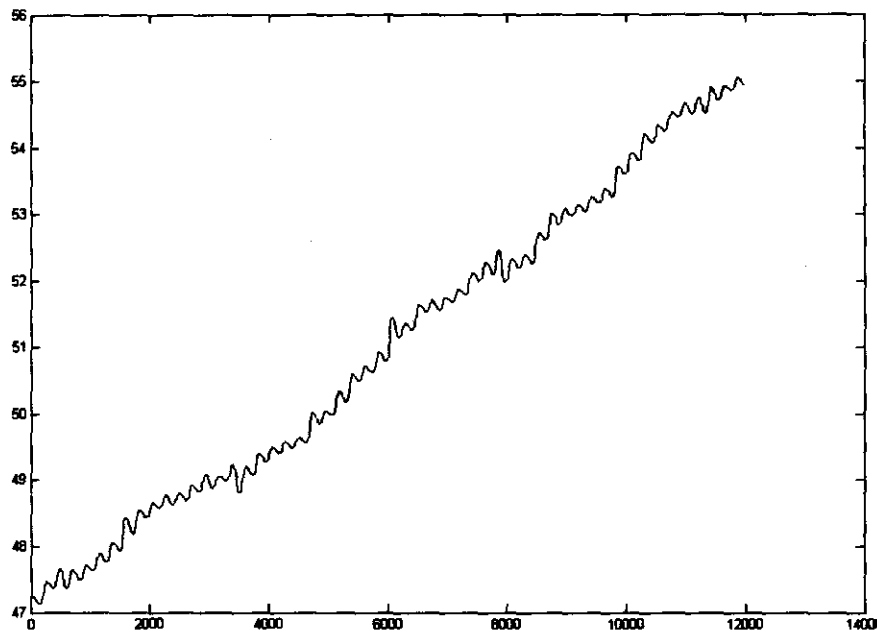


Figure 55: Demodulation peak (magnitude spectrum) versus buffer length for large bearing defect (DA = 180°) and 0g (Vib demo data set 2) – zoom plot

For a larger dynamic load (18g applied imbalance) the final demodulation peak value is somewhat smaller (refer to figure 56 and figure 64).

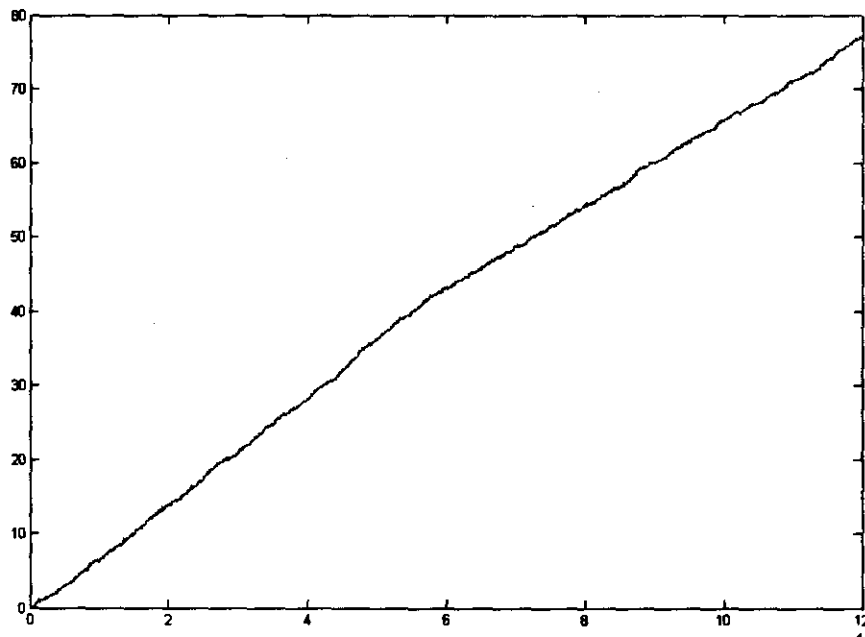


Figure 56: Demodulation peak (magnitude spectrum) versus buffer length for large bearing defect (DA = 180°) and 18g (Vib demo data set 2)

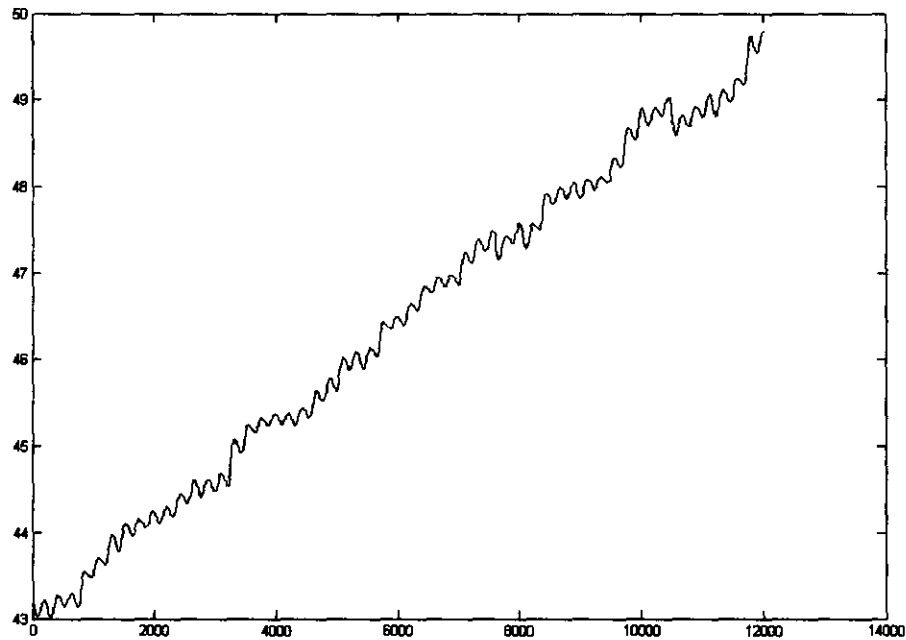


Figure 57: Demodulation peak (magnitude spectrum) versus buffer length for large bearing defect (DA = 180°) and 18g (Vib demo data set 2) – zoom plot

It is evident that the results on the Vib demo data set 2 seem to support the hypothesis of the interaction between the ridge (representing the localized outer race bearing defect) and the load on the roller closest to the bearing defect (which is coupled in a complex, non-linear way to the dynamic load, i.e. applied imbalance).

The results above also seem to confirm the information presented by Harris [16] on the effect of dynamic load in the literature survey in section 2.5.3 on p. 40 'Influence of dynamic load'. The effect of instantaneous variations in the dynamic load is clearly evident in the zoom plots (refer to figure 55, figure 61, figure 63 and figure 65).

In summary, the buffer length required for detection of smaller bearing defects with increased dynamic load would be larger, as the progression of the demodulation peak versus buffer length is not as smooth as for large defects.

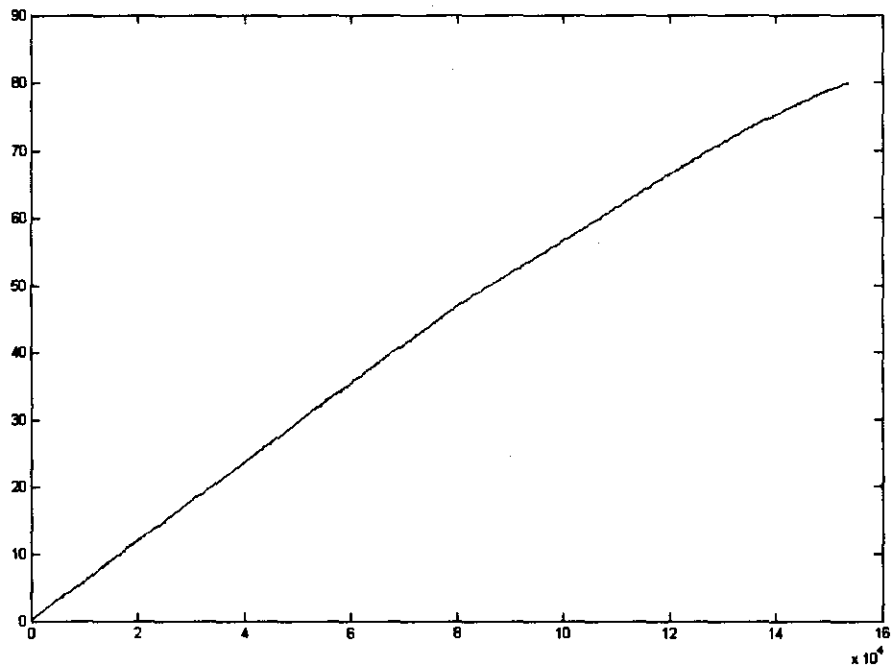


Figure 58: Demodulation peak versus buffer length for medium bearing defect and 0g

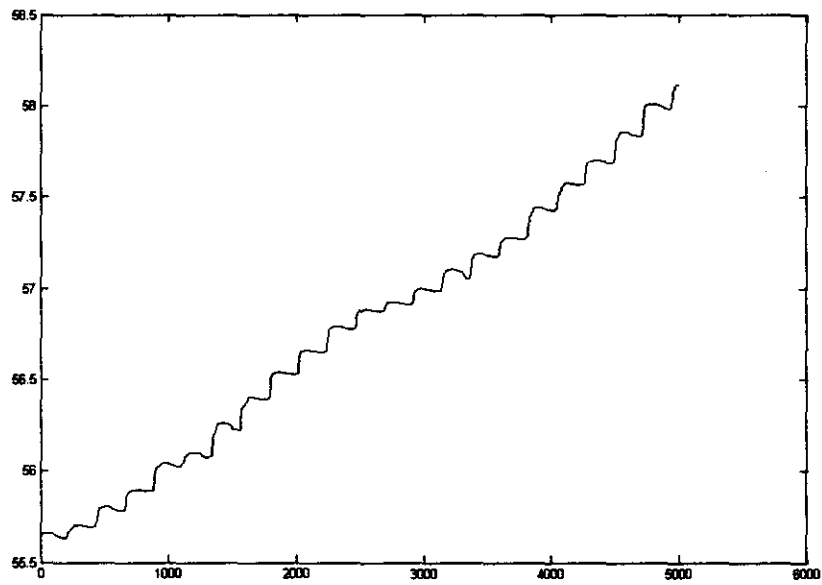


Figure 59: Demodulation peak versus buffer length for medium bearing defect and 0g (zoom plot)

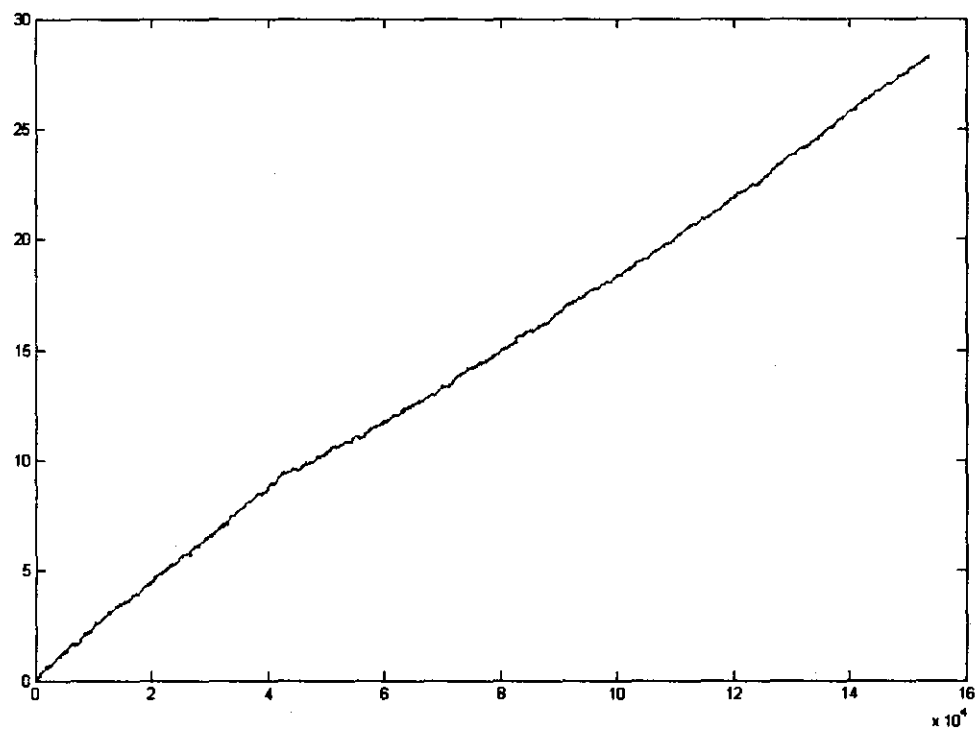


Figure 60: Demodulation peak versus buffer length for medium bearing defect and 24g

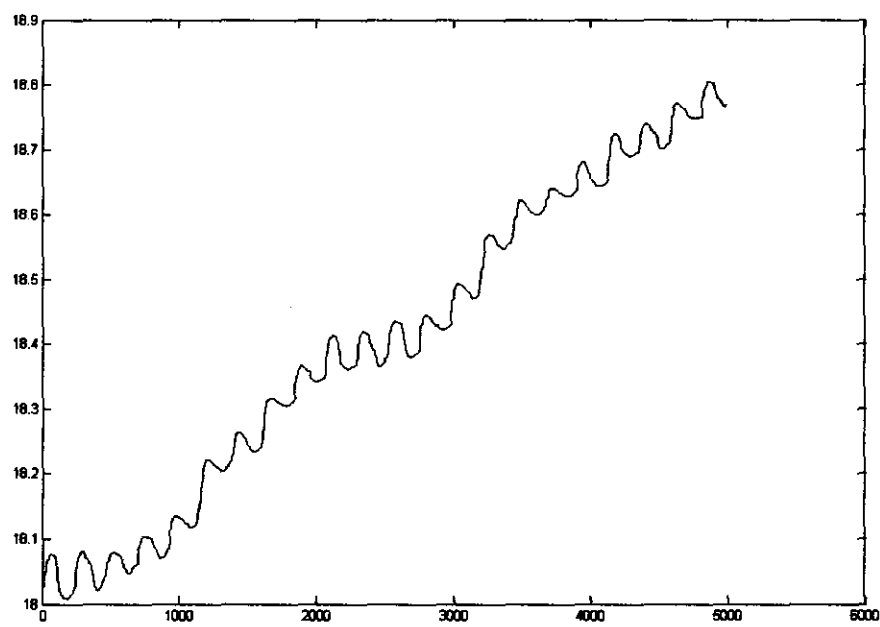


Figure 61: Demodulation peak versus buffer length for medium bearing defect and 24g (zoom plot)

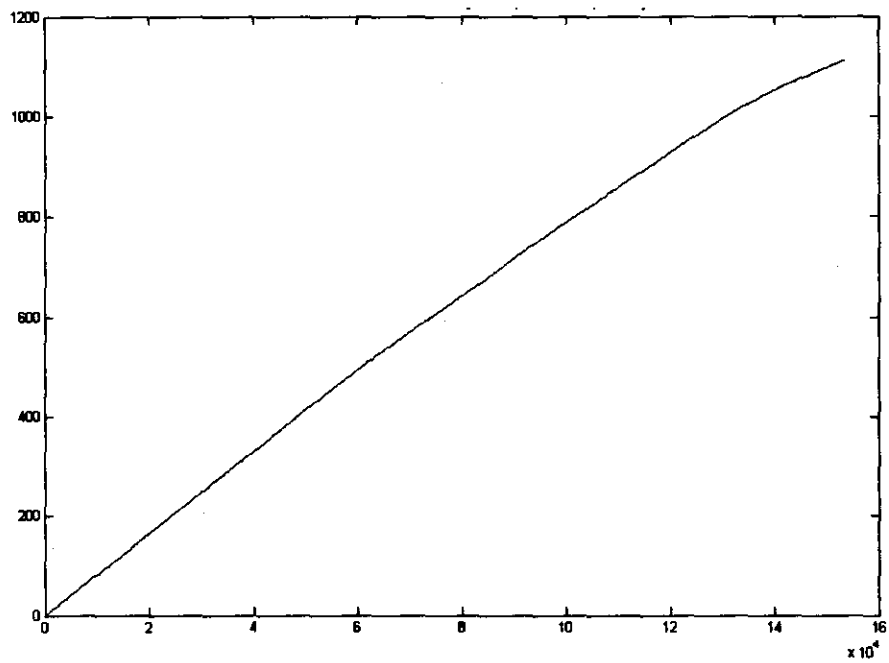


Figure 62: Demodulation peak versus buffer length for large bearing defect and 0g

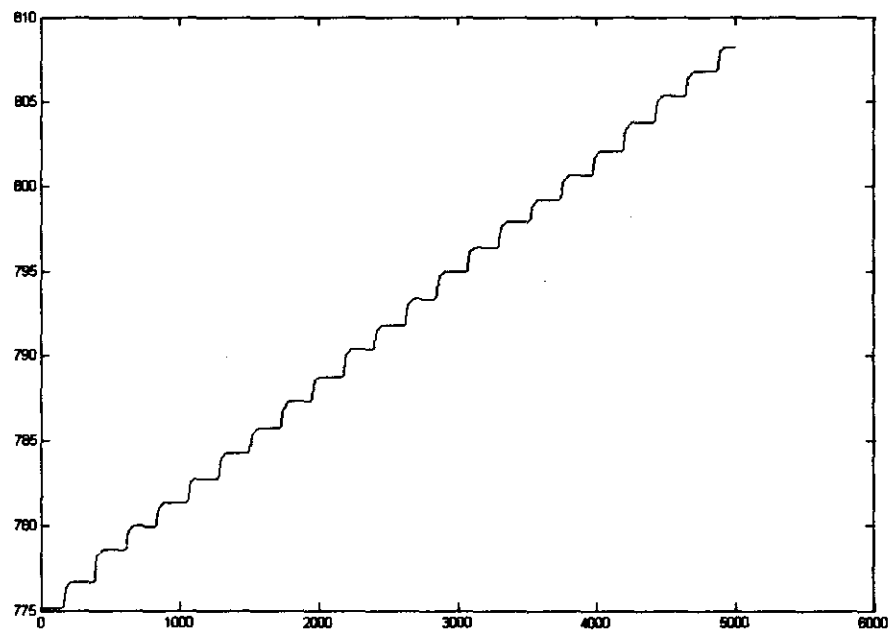


Figure 63: Demodulation peak versus buffer length for large bearing defect and 0g (zoom plot)

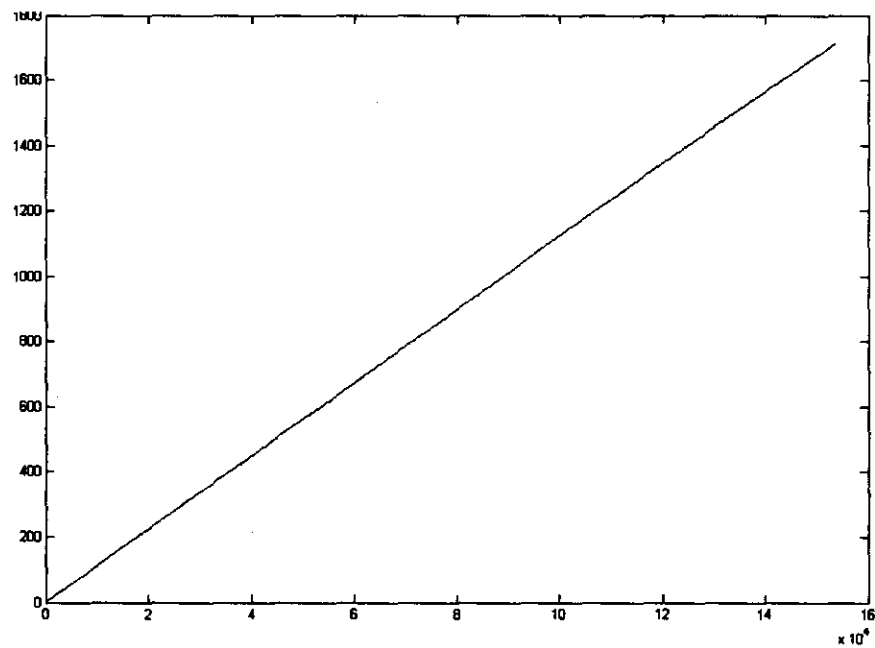


Figure 64: Demodulation peak versus buffer length for large bearing defect and 18g

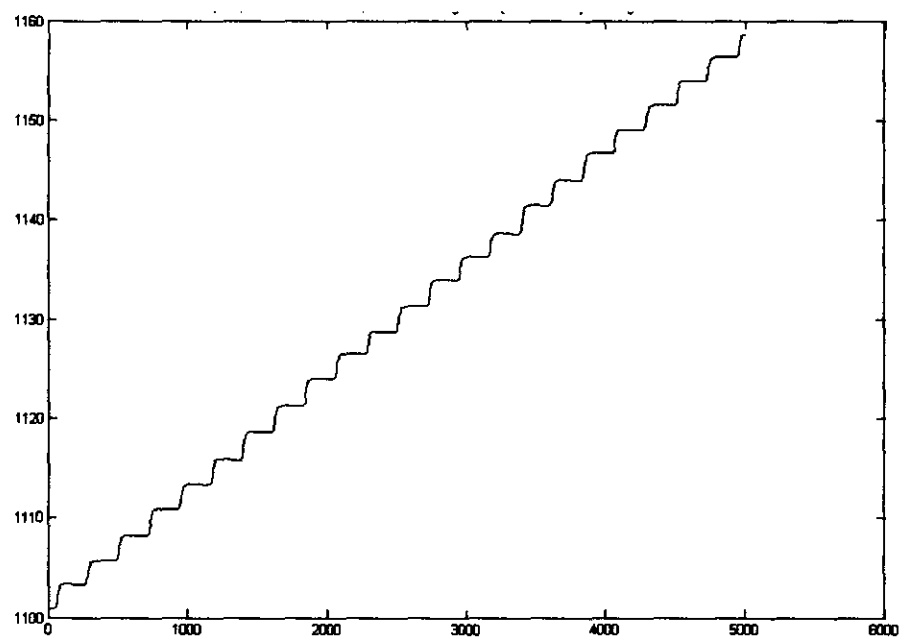


Figure 65: Demodulation peak versus buffer length for large bearing defect and 18g (zoom plot)

7.3.9 Feature *HFD_2_5_kHz*

The feature *HFD_2_5_kHz_X* analyses the frequency content in the high frequency domain (*HFD*) from 2–5kHz on the x channel.

Table 12: Dynamic load (imbalance in grams) and *HFD* feature *HFD_2_5kHz_X* (in dB) versus defect angle (DA in degrees) and defect size (S, M, L for small, medium and large respectively); the symbol NA means measurement is not available

DA	225°	225°	225°	270°	270°	270°
S, M, L	L	M	S	L	M	S
0g	64.5	47.8	32.8	65.0	40.9	31.8
12g	NA	NA	NA	NA	41.4	NA
18g	64.6	NA	NA	NA	NA	NA
24g	NA	NA	33.7	NA	36.3	30.9
DA	90°	90°	90°	45°	45°	45°
S, M, L	L	M	S	L	M	S
0g	32.7	30.8	31.3	NA	30.8	32.3
12g	NA	NA	31.0	NA	NA	NA
18g	34.1	31.3	31.5	29.1	30.8	32.2
24g	NA	NA	NA	NA	30.5	NA
DA	360°	360°	360°	315°	315°	315°
S, M, L	L	M	S	L	M	S
0g	28.0, 27.0	27.6	31.7	54.2	39.5	33.1, 31.2
12g	NA	NA	30.3	NA	NA	NA

18g	28.3, 27.1	28.3	NA	54.8	41.5	NA
24g	NA	29.0	34.3	NA	NA	31.2
DA	135°	135°	135°			
S, M, L	L	M	S			
0g	31.8	31.8	31.1			
12g	NA	NA	31.0			
18g	31.6	31.3	32.1			
24g	NA	31.6	NA			

Table 12 shows that the *HFD* performs almost as well as *BPFO2_X* for 0g imbalance, where $DA = [270^\circ, 225^\circ, 315^\circ, 90^\circ, 135^\circ, 360^\circ]$ for the large bearing defect, $DA = [225^\circ, 270^\circ, 315^\circ, 135^\circ, 90^\circ, 45^\circ, 360^\circ]$ for the medium bearing defect and $DA = [315^\circ, 225^\circ, 45^\circ, 270^\circ, 360^\circ, 135^\circ, 90^\circ, 315^\circ]$ for the small bearing defect. When the defect is out of the loading zone (for example when $DA = 90^\circ$ or 360°) the feature value *HFD_2_5_kHz_X* is much smaller, as expected (the same applies to the feature *BPFO2_X*).

7.3.10 Feature *KHX_X*

Table 13: Dynamic load (imbalance in grams) and kurtosis feature *KHX_X* (in dB) versus defect angle (DA in degrees) and defect size (S, M, L for small, medium and large respectively); the symbol NA means measurement is not available

DA	225°	225°	225°	270°	270°	270°
S, M, L	L	M	S	L	M	S
0g	21.4	25.0	13.9	22.5	22.2	12.4

12g	NA	NA	NA	NA	21.7	NA
18g	21.5	NA	NA	NA	NA	NA
24g	NA	NA	15.6	NA	13.9	12.3
DA	90°	90°	90°	45°	45°	45°
S, M, L	L	M	S	L	M	S
0g	16.5	16.6	11.6	NA	14.8	11.7
12g	NA	NA	11.2	NA	NA	NA
18g	16.5	16.6	11.6	NA	14.8	11.7
24g	NA	NA	11.2	NA	NA	NA
DA	360°	360°	360°	315°	315°	315°
S, M, L	L	M	S	L	M	S
0g	13.2, 11.8	12.7	14.9	24.1	19.7	14.3, 11.6
12g	NA	NA	12.4	NA	NA	NA
18g	12.9, 11.9	13.2	NA	24.4	23.7	NA
24g	NA	12.9	16.9	NA	NA	12.0
DA	135°	135°	135°			
S, M, L	L	M	S			
0g	16.8	12.2	12.1			
12g	NA	NA	12.2			
18g	16.4	11.7	12.2			

24g	NA	11.8	NA			

Table 13 shows that the kurtosis feature performs reasonably compared with the *BPFO2_X* for 0g imbalance, where $DA = [315^\circ, 270^\circ, 225^\circ, 135^\circ, 90^\circ, 360^\circ]$ for the large defect, $DA = [225^\circ, 270^\circ, 315^\circ, 90^\circ, 45^\circ, 360^\circ, 135^\circ]$ for the medium defect and $DA = [360^\circ, 315^\circ, 225^\circ, 270^\circ, 135^\circ, 45^\circ, 315^\circ, 90^\circ]$ for the small defect (the large value of $DA = 360^\circ$ is problematic). The values for the large and medium defects are closer to each other at 0g (making recognition problematic). The values for the large and medium defects are about 5dB higher than the small defect, except for $DA = [45^\circ, 135^\circ, 360^\circ]$. Hence it is still possible to isolate the small bearing defect even if the defect is not in the centre of the loading zone.

The kurtosis feature KH_p_X (feature values are not shown in a table – refer to the figure on p. 333 in section 11.3) with bandpass filtering performs much better than the raw kurtosis feature for 0g imbalance where $DA = [315^\circ, 225^\circ, 270^\circ, 90^\circ, 135^\circ, 360^\circ]$ for the large bearing defect (it is not clear why $DA = 315^\circ$ is the maximum value, instead of $DA = 225^\circ$), $DA = [225^\circ, 270^\circ, 315^\circ, 135^\circ, 360^\circ, 90^\circ, 45^\circ]$ for the medium bearing defect and $DA = [360^\circ, 315^\circ, 225^\circ, 135^\circ, 270^\circ, 90^\circ, 45^\circ]$ for the small bearing defect.

7.3.11 Feature *BPFI2_X*

The feature *BPFI2_X* can separate the large defect (refer to p. 331), but performs worse on the medium and small defects. The explanation for the sensitivity of the inner race defect frequency to the defect is not clear at present¹⁰⁷.

The feature *BPFI2_X* (inner race defect frequency of the demodulated spectrum) performs surprisingly well for the outer race defect and 0g imbalance (dynamic load)¹⁰⁸

¹⁰⁷ This investigation into the sensitivity of feature *BPFI2_X* is part of future work.

¹⁰⁸ A very surprising result is that the feature *BPFI2_X* can separate the large and medium defects, but not the small defect at 0g imbalance. The explanation is not clear at present. It seems unlikely to be a higher harmonic of running speed. The following explanation is proposed. The inner race also has some microscopic defects (along with the roller). The very severe OR defect causes the roller to ‘jump’ every time the roller contacts the OR defect. This causes the inner ring to make contact more firmly with the roller, thus enhancing any existing microscopic defect on the inner race. At the

where $DA = [270^\circ, 225^\circ, 315^\circ, 360^\circ, 90^\circ, 135^\circ]$ for the large bearing defect, $DA = [225^\circ, 270^\circ, 315^\circ, 45^\circ, 135^\circ, 90^\circ, 360^\circ]$ for the medium bearing defect and $DA = [315^\circ, 90^\circ, 45^\circ, 135^\circ, 360^\circ, 270^\circ, 225^\circ]$ for the small bearing defect. Future work can investigate the relevance of this feature for outer race fault detection in more detail (please refer to section 8.6 on p. 263 'Future work' for further discussion).

7.3.12 Feature Fr_X

The feature Fr_X for $DA = [225^\circ, 270^\circ]$ and 0g imbalance in table 14 shows that the frequency component at the rotational shaft frequency is not significantly changed with bearing defect size¹⁰⁹. The influence of the dynamic loading on Fr_X is clearly evident¹¹⁰. Theoretically, an increase of $20\log_{10}(2) = 6\text{dB}$ would be expected (with the assumption of linearity) on doubling the imbalance.

rate of $BPFO$ the inner ring receives shock pulses, which cause the (very small) IR defect to make contact more firmly with the roller. Usually any IR defect would have a very small effect at an angle of 0° (top) and a maximum effect at 180° (at the bottom). Hence a modulation at shaft frequency results. However, in the measurements in this thesis $BPFI2\_X$ has been used directly (i.e. the possible modulation effect has not been considered). A possible hypothesis would be that the presence of the very severe OR defect causes the modulation effect to disappear and increases the vibration signal at $BPFI2\_X$. Another possibility is that of 'Variable elastic compliance' mentioned in the literature survey in section 2.5 'Fault detection on roller bearings' on p. 26 – also refer to section 8.6 on p. 263 'Future work.'

¹⁰⁹ However, some variation in the small defect for specific values of DA is evident.

¹¹⁰ It is clear from table 14 that some variability in the values of feature Fr_X occur even for 0g imbalance (for example, a 20dB change from 45° to 90° for the small bearing defect and 0g imbalance). However, the absolute vibration levels are quite small, hence the relatively large dB reference is no immediate cause for concern. Slight misalignment, when loosening the screw on changing the defect angle between measurements, would seem the most likely cause for the very slight difference in absolute vibration levels. In future work this issue could be addressed using more measurements.

Table 14: Dynamic load (imbalance in grams) and rotating speed feature Fr_X (in dB) versus defect angle (DA in degrees) and defect size (S, M, L for small, medium and large respectively); the symbol NA means measurement is not available

DA	225°	225°	225°	270°	270°	270°
S, M, L	L	M	S	L	M	S
0g	27.4	26.0	17.2	27.6	23.0	22.4
12g	NA	NA	NA	NA	44.8	NA
18g	49.1	NA	NA	NA	NA	NA
24g	NA	NA	50.2	NA	49.3	50.0
DA	90°	90°	90°	45°	45°	45°
S, M, L	L	M	S	L	M	S
0g	25.5	23.6	11.0	NA	17.6	33.6
12g	NA	NA	44.8	NA	NA	NA
18g	49.7	47.3	49.8	48.8	47.9	47.7
24g	NA	NA	NA	NA	50.5	NA
DA	360°	360°	360°	315°	315°	315°
S, M, L	L	M	S	L	M	S
0g	32.2, 28.1	20.2	21.9	27.2	25.0	26.8, 28.1
12g	NA	NA	45.9	NA	NA	NA
18g	48.5, 48.0	44.9	NA	48.4	47.8	NA
24g	NA	49.8	51.9	NA	NA	NA

DA	135°	135°	135°			
S, M, L	L	M	S			
0g	31.1	14.1	32.2			
12g	NA	NA	47.8			
18g	47.5	47.6	48.7			
24g	NA	50.1	NA			

7.4 Future work – position of outer race defect

As the outer race of a bearing is usually permanently fixed, the loading zone concentrates a large part of the load on a relatively small surface area of the bearing. It is proposed in this thesis to vary the position of the outer race to remove the effect of a bearing defect. If the outer race of the bearing is mounted in a circular ring of the same dimensions and fixed by a tensioning screw, it would be possible to move the outer race by loosening the screw. This would provide the following benefits:

- Examination of the surface of the outer race is possible (which could be used, for example, to confirm the presence of a localized outer race bearing defect).
- If a localized bearing defect has already developed it is possible to rotate the position of the outer race to an area with no defect, hence prolonging the lifetime of the bearing.
- If it is possible (by suitable mechanical construction) to rotate the outer race at close to the inner race (shaft) speed, it would be possible to remove the loading effect on the inner race to a large degree. Thus it might be easier to detect small, incipient bearing failures on the inner race. The outer race would then be experiencing a periodic change in the loading zone at shaft speed, similar to the inner race when the outer race is stationary. If the outer race also has a defect, then it would of course be more difficult to make an accurate diagnosis

(sidebands spaced at multiples of the shaft frequency around the outer race defect frequency may be expected).

- As mentioned previously in this thesis, a localized outer race bearing defect may be introduced in a suitable position of the bearing for calibration purposes of instruments. A further advantage is that an alarm threshold can be set on the vibration by introducing the largest acceptable bearing defect on the outer race. Rotating the outer race, so that the bearing defect is out of the loading zone, permits normal operation of the bearing to resume. An additional advantage of this technique is that no knowledge of the shaft speed or bearing construction (to determine fault frequencies) is required for fault location (which may often be misleading, as illustrated in this thesis). Furthermore, the effect of operating conditions for a specific plant (such as shaft speed, static and dynamic loading) may be determined readily by suitable measurements, without requiring complicated mechanical simulation models.

7.5 Conclusion

In this chapter experimental results (Vib demo data sets 1, 2 and 3) related to the detection of multiple fault conditions were presented. It was shown that the RBF classifier can separate the multiple fault conditions (Vib demo data sets 1 and 2). The SOM technique is useful for visualization of the data. Each fault condition may be represented as a class and a classifier may then be trained to distinguish the fault conditions. However, it is necessary to have a complete training set available.

The results on the same test rig, but with measurements on the faulted bearing only, were presented (Vib demo data set 3). The influence of the defect angle (DA) was investigated in relation to the loading zone for the outer race defect using various features. The demodulation peak corresponded well to the bearing defect size, although the dynamic load did influence the absolute level of vibration (making accurate diagnosis of defect size more difficult). The best feature was found to be the *BPFO2_X* feature (as expected), although some of the other features were also useful for low noise levels (large bearing

defect). The simulation study of appendix K on p. 409 seems to indicate that the small bearing defect signal has amplitude and speed variations (as the demodulation spectrum provides better performance than the regular spectrum under these conditions according to the results of appendix K¹¹¹). The resonance frequency for demodulation has not been optimized in this thesis (as good results have been obtained with the selected frequency range from the large bearing defect).

¹¹¹ And, of course, literature references as indicated in this thesis.

8 Conclusion

8.1 Introduction

In this chapter the most important conclusions related to the problem statement of the recognition of multiple fault conditions will be summarized briefly.

8.2 Bearing fault detection

It has been shown in this thesis that the demodulation spectrum, obtained by bandpass filtering the signal and applying a non-linear operation (such as squaring) - followed by a lowpass filter and application of the FFT, actually computes the Fourier components of a synchronous average of the signal (after applying the bandpass filtering and non-linear operation, of course)¹¹², provided that the frequency components are retained at multiples of the defect frequency¹¹³. If the resulting frequency components of the demodulation spectrum are called the ESFFT (equi-sampled demodulation spectrum), then the ESFFT supplies a picture of the spectral content of the signal, averaged over a large number of cycles (corresponding to a synchronous average in the time domain). The first peak at the defect frequency of the demodulation spectrum (for example at $BPFO$ Hz) is an indication of the presence of a single, well localized bearing defect (the first peak corresponds to the first component of the ESFFT). The second peak at (twice) the defect frequency of the demodulation spectrum is an indication of the presence of two, well localized bearing defects spaced 180° out of phase. The k^{th} peak at (k times) the defect frequency of the demodulation spectrum (for example at $kBPFO$ Hz) is an indication of the presence of k well localized bearing defects spaced $360/k^\circ$ out of phase¹¹⁴.

¹¹² Assuming that a sufficient number of averages are taken.

¹¹³ This operation applies a bandpass filter at multiples of the defect frequency (comb filter). This is equivalent to the computation of a signal average.

¹¹⁴ In general the k^{th} peak corresponds to the k^{th} component of the ESFFT. The probability that k defects would be spaced as specified above is certain for $k = 1$. For $k = 2$ it is unlikely that the second defect would be spaced exactly $360^\circ/2 =$

Consequently, it is important to consider all the demodulation peaks together when diagnosing the bearing defect, rather than in isolation. Normalization of feature values, such as the demodulation peak at $BPFO$ Hz, should also be considered¹¹⁵.

The signal average, computed with the defect frequency as the trigger frequency, supplies a picture of the interactions of the roller and surfaces of the bearing (inner race and outer race) over a large number of revolutions. Any vibration that is not synchronous with the trigger frequency will be filtered out if a sufficient number of averages are taken. It is clear that localized faults introduced on the surface of the bearing, as in this thesis, will modify the signal average by showing a large peak (ideally an impulse) at the location of the defect on the surface (assuming a defect on the outer race). However, the signal average will also be influenced by any other factor (not necessarily a fault mechanism) related to the interaction of roller and surface (examples would be lubrication, geometrical errors, surface irregularities, and load and speed variations). Hence the demodulation spectrum, obtained by averaging the signal over a large number of cycles, can also be changed by factors other than real bearing defects.

The real challenge lies in detecting the very small signal of an incipient bearing defect from the background chatter of surface irregularities in the presence of environmental conditions, such as speed and load variations, that cause a reduction in the SNR. The signal average may also be interpreted as the periodic signal that best matches the observed vibration waveform (the inverse FFT of the demodulation spectrum supplies the signal average) and may therefore be useful in modelling (especially when combined with

180° from the first defect. However, an angular position of the second defect of θ relative to the first defect would supply a peak at $mBPFO$ Hz, where $m = 360^\circ/\theta$. Hence the m^{th} component of the ESFFT would have a peak for $k = 2$. For $k > 2$ a number of peaks in the ESFFT would be activated, dependent on the vector sum of all individual defect pulses ([50, 53]).

¹¹⁵ One example of a normalized feature would be the kurtosis. The kurtosis is a very useful feature as no prior knowledge is required about the rotational speed or machine construction. Normalization on a feature can be performed with respect to rotation speed (to reduce speed variations) as well as amplitude levels (to reduce the effect of dynamic load). Normalizing, for example, the demodulation peak – representing the periodic components of the observed signal – would seem to involve using the non-periodic components of the observed signal. However, the non-periodic components of the observed signal would in general be contaminated with amplitude and speed variations. Hence the validity of such a normalized feature could be questioned. The kurtosis of small (bearing defect) signals would be an example of where the (normalized) kurtosis feature could be misleading, as the small signal is effectively amplified by the computation of the kurtosis statistic. Other normalized features might well experience the same kind of difficulty.

ALE/SANC techniques that are able to filter out some of the noise). Another challenge is to give an accurate indication of the progression of a bearing defect, once it has been detected (bearing prognostics). As the defect advances, the fundamental defect frequencies may even disappear, complicating diagnosis.

It has been shown in this thesis that quite a large and well localized bearing defect can be detected reliably, even when it is not in the centre of the loading zone. However, in advanced stages of failure, the defect may become more distributed and the resulting vibration signal more stochastic. Hence clear defect frequencies may no longer be present and other more advanced signal processing techniques may be required (having a disciplined condition-monitoring programme should help to prevent this).

It has been illustrated in this thesis that the demodulation spectrum (and the corresponding signal average) may be influenced by a secondary fault mechanism (in this case a dynamic load in the form of an unbalance mass). The vibration level of the demodulation peak at *BPFO* Hz generally decreases with increasing dynamic load. For this reason, trending of vibration levels at a defect frequency, based solely on absolute vibration levels, may lead to potential misdiagnosis in the presence of load variations.

It is very important to take a set of baseline readings on installation of any new bearing before any fault mechanism develops. Comparing the signal average, as well as demodulation spectrum, over time (trending) may help to pick up failure mechanisms early. Apart from helping to plan timely replacement of bearings and preventing additional damage, the knowledge gained may be useful in preventing future problems (owing to the precision manufacturing of a bearing, a defective bearing is usually a symptom only – a warning message – and not the root cause of failure). Improper installation, incorrect preload, shaft misalignment and incorrect lubrication may all contribute to early bearing failure.

As the signal average is also influenced by variations in dynamic load, a calibration signal with a defect of known geometry – as proposed in this thesis – could be used in a plant to help remove the effect of load and sensor placement on the measurements (as well as to ascertain the proper functioning of the system). The idea is to introduce a bearing defect

intentionally on a bearing to be monitored in the plant for calibration purposes. Usually extreme care is taken to avoid any fault conditions in a plant. Yet, the induced bearing defect simply corresponds to a change in the surface characteristics of the outer race and normal operation may be restored by simply rotating the outer race defect out of the loading zone (assuming that the outer race is stationary). The potential advantages of the proposed technique would of course have to be weighed against the potential for decreasing reliable system operation. In a periodic maintenance schedule the disassembly of machinery at specific time intervals would seem to provide an ideal opportunity for testing this technique. By inducing a small fault condition the influence of environmental and secondary fault conditions should be much easier to remove (the fault condition may of course be removed in principle simply by rotating the outer race).

In summary, the problem statement of multiple fault conditions may be addressed successfully using a combination of signal processing and pattern recognition techniques. The main aim of this thesis is not simply to solve a vibration problem, but rather to provide some guidelines on the successful implementation of artificial intelligence (pattern recognition) and signal processing techniques in condition-monitoring applications. Owing to the specific focus of the problem statement very specialized techniques have been used in this thesis, which might not be needed in other problem domains. However, it is hoped that the proposed vibration model for fault detection and subsequent proposed techniques are sufficiently generic for the same concepts to find application also in other (at present unrelated) areas of condition monitoring.

8.3 Signal processing techniques

It has been shown that proper preprocessing and feature extraction may result in features (Fr_X and $BPFO2_X$) that supply very good diagnostic decisions. Dynamic loading does influence the feature $BPFO2_X$ (demodulation peak) of the medium bearing defect significantly, complicating accurate diagnosis. Further experimentation and mathematical modelling would be required to establish rigorous guidelines for the influence of dynamic (and static loading) on bearing fault detection, which is left for future work.

Modified cepstral analysis proposed in this thesis, while improving on the usual cepstral analysis, does not perform quite as well as the *BPFO2_X* feature. It is shown mathematically that the Hilbert transform and square law demodulator are very closely related for a bandpass filtered input signal. Hence the amplitude demodulation of high frequency defect signals may be accomplished using a variety of (non-linear) transforms. The essential characteristics for successful bearing fault detection seem to be the following:

- The best frequency band(s) (also called modes) have to be isolated for fault detection (filtering in signal processing terms). The effect of speed variation increases with frequency, so a compromise on the frequency band is often dictated for best overall performance¹¹⁶.
- A non-linearity has to be applied to isolate the ‘envelope’ of the signal (a low frequency signal that modulates a high frequency carrier). The relation between squaring and the Hilbert transform was investigated in this thesis. For narrowband modes it is concluded theoretically in the thesis that the above techniques perform equally as far as the extraction of the envelope signal is concerned. Other possibilities for non-linearities, not discussed in this thesis, include raising to a higher power [21] and halfwave rectification.
- Suitable features (frequency components) may be extracted from the envelope signal (after the application of a suitable non-linearity). For distributed defects it would be important to incorporate the harmonics of the characteristic defect frequency as part of the feature set. Usually the low frequency components (representing the envelope of the signal) would be of primary interest for fault detection. If only the first harmonic of the defect frequency is used, a lowpass filter on the envelope signal is effectively obtained.

¹¹⁶ The most resonant modes are often high frequency, but to reduce the effect of speed variation the frequency band should have as low a frequency as possible. Hence a trade-off is required.

When speed and/or amplitude variations are present, the technique of asynchronous AM demodulation (HFRT) provides better results than synchronous AM demodulation (as was demonstrated in this thesis and in literature [15]). It may be concluded that proper signal processing techniques alone are not sufficient to make correct diagnostic decisions for the data sets considered in this thesis in the presence of a small dynamic load owing to the influence of the secondary fault condition (dynamic load) on the primary fault condition (bearing defect). Additional load normalization techniques should consequently be developed to compensate for the amplitude modulation function introduced by the presence of the dynamic load. Alternatively, the additional application of pattern recognition techniques in conjunction with better feature extraction (incorporating the effect of the load modulation function) may be considered, provided that enough data are available to be able to exploit the inherent non-linear modelling capabilities of, for example, neural network techniques fully.

It has been shown in this thesis that the demodulated spectrum may be regarded as a superposition of a number of SAs (consisting of the periodic and non-periodic signal components), which helps to explain the success of the demodulated spectrum relative to the SA (for small bearing defects non-periodic components are also generated), as the SA considers periodic components only¹¹⁷.

In appendix K (section 11.11.9) a simplified vibration model $y = \sin(\omega t)^p$ is proposed and it is shown that the demodulated spectrum is better than the regular spectrum when the SNR is greater than $\sqrt{2K_p^{-1}}$. Expressions for the detection ratio of the regular and demodulated spectrum using the FFT with additive noise are derived and illustrated with some examples. The demodulated spectrum performs better than the regular spectrum with multiplicative noise (the opposite applies with additive noise). The effect of speed variations is considered using a simple step change in frequency. The demodulated spectrum directly provides the fundamental defect frequency, which is more difficult to obtain with the regular spectrum under conditions of low SNR (for example a small

¹¹⁷ For larger bearing defects the periodic signal components would dominate, hence good results could be expected using the SA (or indeed any other signal processing technique that can detect periodic components, such as the cepstrum).

bearing defect). Closer inspection of the (high frequency components of the) regular spectrum reveals that (for the examples considered) the high frequency response is quite similar in shape to that of the (demodulated low frequency components of the) demodulated spectrum.

8.4 Modelling of fault signals using pattern recognition techniques

It has been shown in this thesis that pattern recognition techniques may be applied successfully both to classification problems (the number of outcomes – classes – are finite) as well as modelling problems (the number of outcomes – classes – are potentially infinite). It should be emphasized that the models developed are valid for the training set samples obtained in the Vib demo data sets 1, 2 and 3 only. Any attempt to extract knowledge about samples some distance in the feature space from the training set data would tax either the extrapolation or interpolation capabilities of the pattern recognition technique (for example a neural network). Hence while it is possible to derive a working model without deriving a mathematical model from first principles, this model will in general be valid around the training set samples ('operating conditions') only. It is possible to impose smooth interpolation performance through careful attention to detail, but the extrapolation performance cannot be regulated easily¹¹⁸.

Very simple pattern recognition techniques have been applied successfully to classification problems. For the data set in [102, 103] it is possible to obtain perfect separation using a single feature (variable). The feature set is computed by binning (averaging) the spectral components of the signal. It is found, using an NNR technique, that several high frequency components achieve separation of the classes on their own¹¹⁹. The same observation may also be made on the Vib demo data sets 1 and 2, although the dynamic load does influence the interpretation of the medium bearing defect in the Vib

¹¹⁸ In a single dimension with a non-linear sigmoidal transfer function the output value will approach one and stay nearly constant as the input value is increased in value. Hence the output saturates (stays the same) when extrapolating in a single dimension.

¹¹⁹ Results on exploratory feature selection are not shown in this thesis, but the authors of [102, 103] reported similar results (private communication).

demo data set 3 to some degree. The SOM technique was found to be useful for visualization purposes and has been applied successfully to the Vib demo data set 2. Thus it is clear that the full power and value of the pattern recognition techniques in non-linear modelling cannot at present be fully used owing to constraints in the size and completeness of the training sets.

The application of pattern recognition techniques to modelling problems, in which the number of classes can be very large (infinite), represents a bigger challenge. Building a physics based finite element model (FEM) of the vibration problem is of course the best option, but may be very difficult to implement (often requiring 3-D modelling techniques and very small mesh sizes for accurate modelling). A simplified model of bearing defects has been proposed in this thesis. Although the effect of speed variations may be accommodated readily, the basic assumption is still that the shape of the individual defect pulses remains the same. To enhance the performance of the model locally, an amplitude variation function, a model with multiple defect pulses (possibly of different shape) or time-varying systems ([10, 65]) could be used. The effect of dynamic load could be incorporated into this model by, for example, an amplitude variation function and is left for future work.

It should be clear from this thesis that pattern recognition techniques provide no 'magic bullet'. Rather, they are useful tools for complicated modelling tasks in which the relationship between input and output variables may be non-linear and multidimensional. The role of the expert is vital in providing good features to the model. Since the neural network usually performs in a 'black-box' way, it cannot provide insight into the relationship between input and output variables (features). Hence the aim should always be to reduce the feature set to the minimum (usually by using transformations of the data to capture suitable 'features' of the data, suggested by an expert). Furthermore, although the pattern recognition technique (for example neural network) can make a prediction for any data sample, good results can of course be expected only if the interrogated sample is 'close enough' to the training set data. If the interrogated sample is between two known samples the neural network would tend to perform a linear interpolation, which could be

acceptable in certain circumstances. If the interrogated sample is far from other samples the neural network would tend to saturate (modelling the value of the closest sample)¹²⁰. This behaviour of the neural network when operating far from the training data set (i.e. to do nothing) might again be acceptable in certain circumstances. However, it is clear that the neural network has to be exposed to all the relevant environmental conditions if good modelling performance is to be ensured. In this sense the neural network is operating in a similar way to the NNR (nearest neighbour rule), which may be viewed as a lookup table. Hence the importance of the completeness of the training set in modelling performance cannot be stressed enough.

The results obtained in this thesis (refer specifically to section 7.3.8) indicate that the effect of an applied dynamic load (in the form of imbalance) is most pronounced either for small/medium bearing defects (where DA is close to 225°), or for a large bearing defect where DA is further from 225° (for example the Vib demo data set 2). In summary, the actual load experienced by the roller closest to the bearing defect (representing the induced vibration owing to the defect) is dependent on at least the following factors:

- The applied dynamic load in the form of an imbalance. However, the relationship is not simply linear but rather highly non-linear and complex. In particular, it has been shown in this thesis that the vibration may actually decrease with an increase in dynamic load in specific circumstances.
- The size of the bearing defect (large, medium or small). Moving a large bearing defect out of the loading zone (where DA is far from 225°) has the effect of decreasing vibrations (effectively decreasing the load on the roller closest to the bearing defect). If the loading zone of the test rig had been placed at 180° the functional form of the load would likely follow the Stribeck equation (equation 20 on p. 32).

¹²⁰ Assuming, of course, that the neural network has in fact converged to a reasonable approximation of the training data set. It is also assumed that the number of free parameters in the neural network is related to the size of the training data set.

- The position of the DA (angular position of the defect) relative to the loading zone. In the Vib demo test rig used in this thesis the loading zone is not at 180° (as would have been expected owing to gravity), but rather closer to 225° (as a result of an additional pressure plate acting horizontally).

8.5 Applying pattern recognition techniques successfully in condition-monitoring problems

In appendix J on p. 401 some recommendations are made for the successful application of pattern recognition techniques in condition-monitoring problems. The main points may be summarized as follows:

1. Understand the data set completely and apply standard linear regression techniques¹²¹. Balance the data set¹²². Proceed to step 2 only if the results are not acceptable, otherwise stop.
2. Extract and select salient features (using simple classifiers, for example MIFS or NNR). Rules from experts may also be used profitably.
3. As a double check on the neural network learning process consider applying NNR, CVCP or the Gamma test to the same data set.
4. Preprocess correctly for neural network.

¹²¹ Understand the data set completely and apply standard linear regression techniques. The neural network can effectively do linear regression in a two-class problem if the number of neurons in the hidden layer(s) is set to one. Hence a good test of the degree of non-linearity in a problem is to compare the accuracies with and without linear regression (for a neural network, compare the best performance of the neural network to the performance constrained on a single neuron, which thereby supplies the effective 'gain' of the non-linear neural network structure). If the difference is small (effective 'gain' close to one) then the conclusion would be that a neural network (with many free parameters) is not required for the problem. In this case a further test would be to examine which feature(s) are correlated with the outputs (feature selection).

¹²² To ensure that each class has the same number of samples, otherwise excellent results may be achieved by a naïve classifier simply outputting the majority class irrespective of the feature values. As failure data are usually very scarce, balancing of the classes would almost always be required.

5. Select appropriate neural network topology, parameters and learning strategies for the problem.
6. Train the neural network model and validate it on an independent test set.
7. The result of the step 6 should be close to that of the NNR (step 3).

8.6 Future work

As indicated in this thesis, a number of problems have been left for future experimentation and investigation.

8.6.1 Theoretical aspects (pattern recognition and signal processing)

8.6.1.1 Feature extraction or selection

- The trending of measurements over time may be sensitive towards the absolute level of vibration (which may be influenced significantly by a number of operating conditions). Hence reliable normalized features should be developed for bearing fault detection in the presence of static and dynamic loading. The ideal is to reflect the operating condition of the machinery as reliably as possible (using the minimum number of features).
- Future work may investigate the relevance of the feature *BPF12__X* (inner race defect frequency of the demodulated spectrum) for outer race fault detection in more detail. A related question would be the separation of geometrical bearing defects from any local bearing defects using advanced signal processing techniques, perhaps focusing on aspects of the dynamic time domain signal [16]¹²³.

¹²³ A localized bearing defect may in principle also be described as a geometrical (manufacturing) defect, although the term seems to be used in the context of distributed geometrical defects (as opposed to localized defects). However, owing to the presence of the loading zone, some of the distributed geometrical defects on the surface of the outer (or inner) race would not be easily detectable (or at least reduced greatly in amplitude).

8.6.1.2 Signal processing techniques

- The techniques of adaptive line enhancement (ALE) and self-adaptive noise cancelling (SANC) were not included in the literature survey, but could potentially prove very useful particularly in low SNR [3, 106]. The exact behaviour of the ALE in the presence of speed and amplitude variations is not known at present and is therefore left for future work. In [3] the ALE is applied before the non-linear operator to separate the gear and bearing signals, whereas in [78] it is applied after the non-linear operator to remove the noise components (increasing the SNR).
- The technique of adaptive noise cancelling (ANC) could be applied by measuring a reference signal far from the test rig (with the signal severely attenuated). Furthermore, measurements at different points on the test rig could be used to investigate BSS techniques (refer to Nandi [58]). This would unfortunately require additional instrumentation and associated costs.
- The techniques of higher order spectral (HOS) processing and blind source separation (BSS) could be very useful for the specific problem addressed in this thesis [58,59]. Nandi *et al.* discuss the application of HOS (in contrast to single order spectral (SOS) processing, which are well known) in the context of HOS statistics, blind signal estimation, blind system identification, blind source separation and robust cumulant estimation.
- The matched filter approach (or other related techniques from communication systems) could be applied to vibration monitoring. For instance, in [11] the analogy is shown between the vibrations and acoustic waves of rotating machinery and the techniques of speech processing.
- It has been postulated that better results would be obtained using asynchronous AM demodulation than synchronous AM demodulation in the presence of amplitude and speed variations. Further theoretical investigation is required.

8.6.1.3 *Pattern recognition techniques*

- The NNRC technique should be investigated more closely in the context of other standard pattern recognition techniques, for example mutual information based feature selection (MIFS), decision trees, self-organizing maps (SOM) and vector quantization (VQ).

8.6.1.4 *Remaining theoretical issues*

- As any signal may be decomposed into a number of SAs (section 6.8 on p. 139) it would be of interest to determine how many SAs would be required for performance equivalent to the demodulation spectrum.
- The poor performance of the synchronous average feature on the small bearing defect has to be investigated in more detail (refer to section 7.3.7 'Feature SA_BPFO2_0_X' on p. 231 for further details). It would be of interest to examine whether the ALE/SANC techniques could be of benefit in the problem domain [3].
- The influence of a time domain window when computing the SA (as in the ESFFT of the demodulation spectrum) could be examined in reducing the effect of speed variations in the signal.

8.6.2 *Experimental work and modelling*

8.6.2.1 *Experimental work*

8.6.2.1.1 *DA as tool for bearing fault detection*

The focus of this thesis was to detect small incipient faults with multiple fault conditions (bearing outer race defect in the presence of dynamic loading in the form of imbalance). In a broader context the same problem might also be viewed as a control problem, in which the transmitted vibrations of the bearing to a sensor should be minimized (which may include non-invasive sensors such as current and acoustic monitoring). The ideal would be not only to detect changes in the vibration (owing to, for example, non-uniformities in the

surface of the bearing), but also to control (and minimize) the vibrations if a suitable actuator can be installed (one possibility would be ‘magnetic bearings’). If control is done at the right time significant reduction could be expected in vibration. Hence it might be possible to prolong the bearing life in critical production processes without the necessity of shutting down the plant, causing production losses. In this context, the results of this thesis on dynamic loading seem to be especially relevant if a suitable control strategy is to be realized.

In section 7.4 from p. 249 the possible advantages of varying the outer race defect position were mentioned.

Although no results have been reported in this thesis on accelerated life-time testing of bearings, the results of this thesis seem to indicate the possibility of interesting research into the effects of dynamic loading on bearing life¹²⁴. Further investigation into this complex aspect is left for future work.

8.6.2.1.2 Influence of sensor technology on fault detection

Only results based on permanently mounted vibration sensors have been reported in this thesis. The measurements on the test rig included the following additional sensor technologies:

1. Acoustic monitoring (close, intermediate and far locations)
2. Current monitoring

¹²⁴ As proposed in this thesis, the vibration may in specific circumstances actually decrease with an increase in dynamic loading. Hence it would be of interest to determine if the bearing life could perhaps be extended with increased dynamic loading. If the proposed hypothesis of decreased dynamic loading on the roller (closest to the bearing defect) proves to be correct, the intuitive expectation would be an increased lifetime owing to the reduced load on the roller closest to the bearing defect. If the proposed hypothesis of decreased dynamic loading on the roller (closest to the bearing defect) proves to be correct, the intuitive expectation would be an increased lifetime owing to the reduced load on the roller closest to the bearing defect (assuming that the dynamic load – or imbalance mass – is far from the roller closest to the bearing defect). However, when the imbalance mass is closest to the bearing defect the momentary load on the roller would be increased significantly. Hence it is possible that this increased (instantaneous) loading might counter the decreased loading experienced when the imbalance mass is moving away from the roller (closest to the bearing defect), hence reducing the potential advantages of increased lifetime (obtainable by the decreased dynamic loading when the imbalance mass is further from the roller closest to the bearing).

3. Handheld vibration measurements at various locations around the test rig

It would be of interest to investigate the utility of other sensor technologies for fault detection¹²⁵.

In general, the high frequency response of the sensor should be of high enough fidelity to enable the high frequency resonances (for HFRT) to be of sufficient magnitude above the noise floor. However, the linearity of the transducer might not be a requirement for bearing fault detection¹²⁶. Non-invasive techniques such as current, acoustic and optical monitoring would sometimes present the only practical option in an industrial plant (vibration sensors have problems in extreme environmental conditions). Hence further research should be done to exploit alternative sensor technologies in future and to determine the optimal balancing of sensors in a plant. The technique of varying the DA, discussed in this thesis, could also be used in a plant to determine the suitability of sensors without requiring elaborate mathematical modelling of the mechanical system.

8.6.2.1.3 Test rig

The test rig proved suitable for the purposes of the problem statement of this thesis. A number of modifications to the test rig could be investigated in future.

The main points may be summarized as follows:

¹²⁵ An initial investigation by the author already confirmed the suitability of acoustic and magnetically mounted vibration measurements for bearing fault detection. As the material on the test rig did not permit fixing by magnetism, the sensors for this particular experiment had been held in place by hand – hence it would be more accurate to describe the measurements as handheld measurements. For handheld measurements a large bearing defect is used out of the loading zone (DA = 180°), which is a medium defect – further experiments would be required using small incipient faults. Unfortunately, no success has been had yet using the current measurements (this might be owing to insufficient preprocessing of the current signal or to the fact that vibrations from the bearing are not transmitted sufficiently through the coupling to the shaft). In particular, it might be necessary to remove the large observed current peak to improve the signal-to-noise ratio of the smaller signals transmitted from the bearing to the shaft.

¹²⁶ Inspection of the transfer functions of different sensors (for example permanently mounted, magnetically mounted and handheld) shows that the high frequency response of the handheld vibration sensor is poor compared with the permanently mounted vibration sensor. It might still be possible to do fault detection on the handheld vibration sensor in spite of the distortion. The exact shape of the defect pulses is not as important as the envelope for fault detection [50, 53]. Hence it is possible that the handheld vibration sensor could still provide adequate information on the envelope, enabling fault detection.

1. A static load may be applied on the shaft (either with a permanent coupling or a belt-driven arrangement). This would enable direct verification of the hypothesis that the DA is related to the static loading. The application of the idea of changing the DA for better diagnostic results in an industrial setting may be investigated.
2. Increased resolution on the DA and dynamic load (imbalance) could be useful for vibration measurements¹²⁷.
3. The speed of the motor should be controlled (allowing the influence of operating speed to be determined on vibration measurements). Furthermore, the influence of speed and DA may be determined together.
4. The angular position of the shaft, outer/inner race or rollers (for example by using a shaft encoder¹²⁸) may be measured. The torsional vibration might supply even better fault diagnostics (for example, in [80] it is shown that torsional vibration features may be helpful in a gearbox operating under dynamic load, hence a similar improvement could be expected in the context of bearing fault detection¹²⁹). This measurement would enable speed compensation to be applied to the measurements and the position of a defect to be accurately determined.
5. A direct coupling to the shaft could be considered for current measurements.

¹²⁷ More measurements should be made on small defects (with low SNR) using the DA as a parameter. Experiments with multiple defects could also provide additional information. Further experiments could be performed with defects on the inner race and rollers.

¹²⁸ An eddy current sensor has been used in this thesis.

¹²⁹ The 'ideal' gear would transmit no vibration, assuming that the number of teeth approaches infinity, which is not possible in practice. Two rollers of different diameters meshing together approach this ideal condition, but in practice considerable problems are experienced with transmission of torque using belt-driven arrangements. In practice, the meshing of the teeth always generates vibration at a characteristic meshing frequency. An ideal bearing, in contrast, is supposed to transmit little vibration. A localized gear fault might be represented as an instantaneous change in stiffness, which could be compared with the instantaneous change in bearing stiffness owing to a localized bearing fault.

6. Additional sensors (acoustic, handheld and current monitoring) would enable BSS techniques to be investigated (more research is required especially for small defects).
7. The effects of various dynamic loading functions as a function of the DA may be performed using a suitable mechanical actuator and the activation functions programmed with the SIGLAB© vibration analyser.

8.6.2.2 *Modelling of defects*

- The dynamic properties associated with the outer race bearing defect (relating the vibration amplitude to the defect size and loading level using a dynamic model such as in [108]) could be analysed. This would enable a better theoretical understanding of the results.
- Additional load normalization techniques should be developed to compensate for the amplitude modulation function introduced by the presence of the dynamic load.
- A simplified model of bearing defects has been proposed in this thesis. The effect of dynamic load could be incorporated into this model by an amplitude variation function¹³⁰.

¹³⁰ With possible speed variation (phase modulation) as well.

9 Bibliography

- [1] M. Abramowitz and I. A. Stegun (Eds.). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, 9th printing, 1972.
- [2] A. Ambrózy. *Electronic Noise*. McGraw-Hill and Akadémiai Kiadó, 1982.
- [3] J. Antoni and R. B. Randall. Optimisation of SANC for separating gear and bearing signals. In *Proceedings of the Condition Monitoring and Diagnostic Engineering Management (COMADEM) Conference*, pages 89–96, Manchester, September 2001.
- [4] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4): 537–550, July 1994.
- [5] J. E. Berry. How to specify machinery vibration spectral alarm bands. *Journal of Sound and Vibration*, September 1990.
- [6] S. A. Billington. *Sensor and machine condition effects in roller bearing diagnostics*. MsC thesis in Mechanical Engineering, Georgia Institute of Technology, 1997.
- [7] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
- [8] S. Braun. *Mechanical Signature Analysis*. Academic Press, 1986.
- [9] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.
- [10] D. Brie. Spalled bearing vibration signal modelling. *Mechanical Systems and Signal Processing*, 14(3):353–369, May 2000.

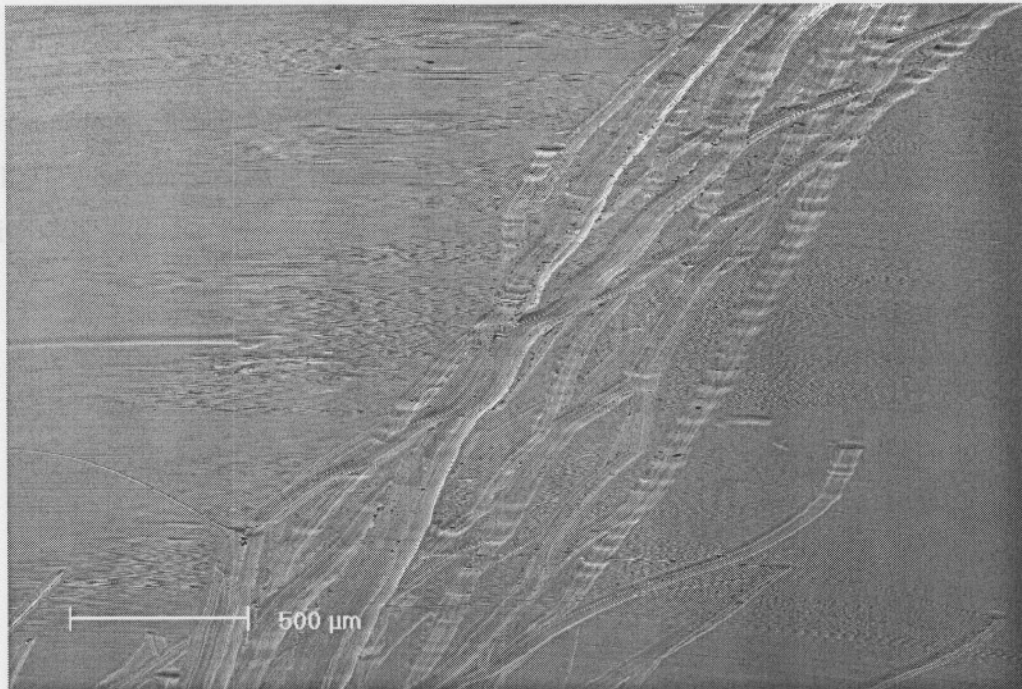


Figure 213: Inner race defect

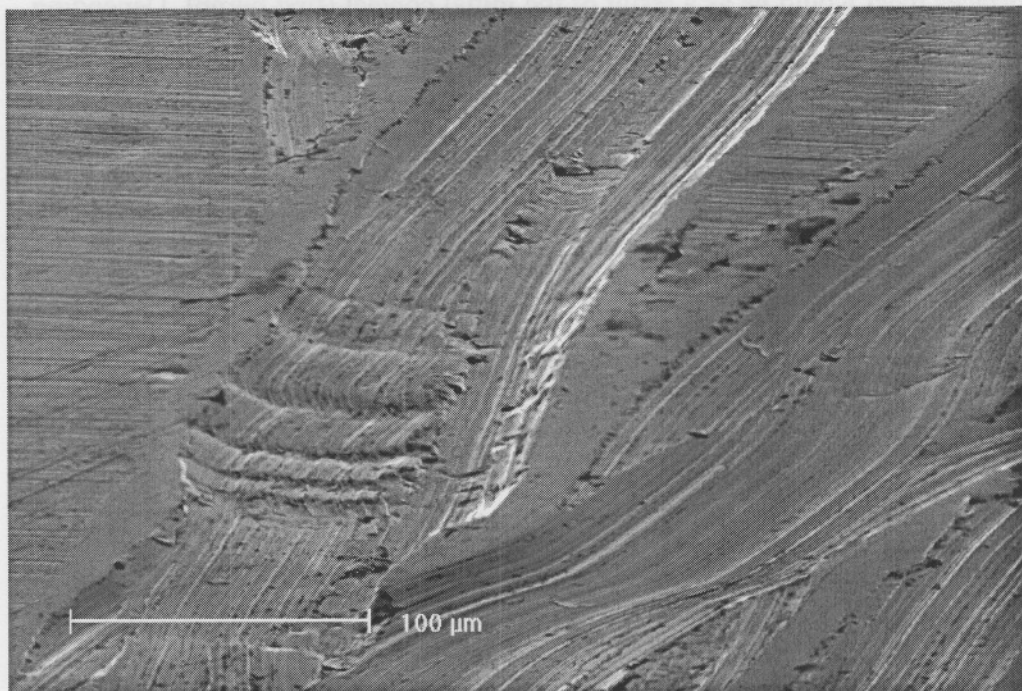


Figure 214: Inner race defect

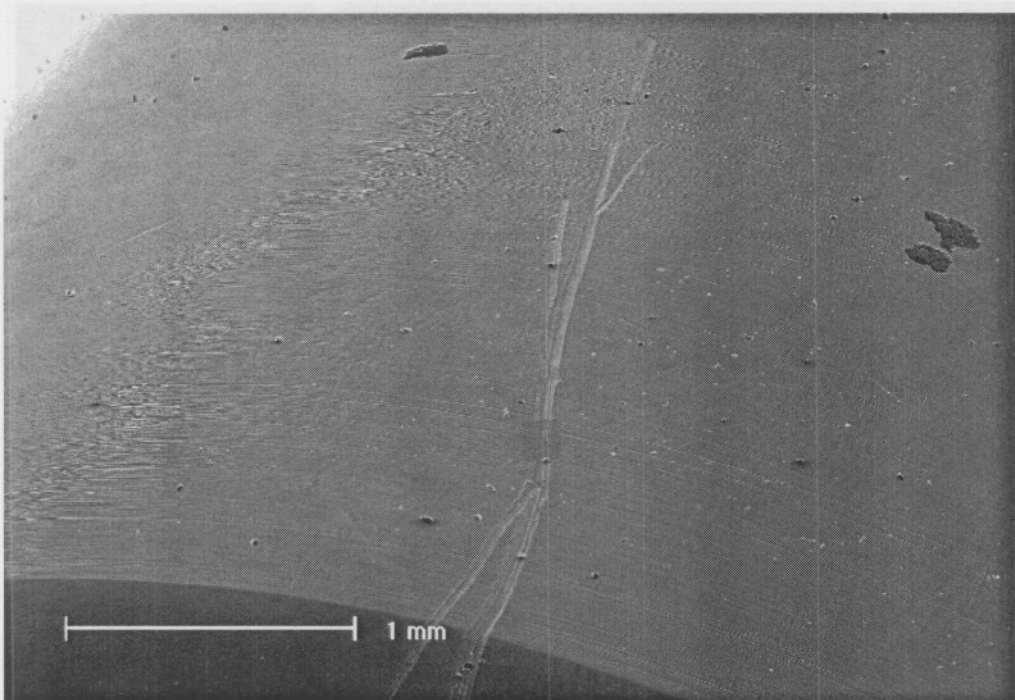


Figure 215: Roller defect

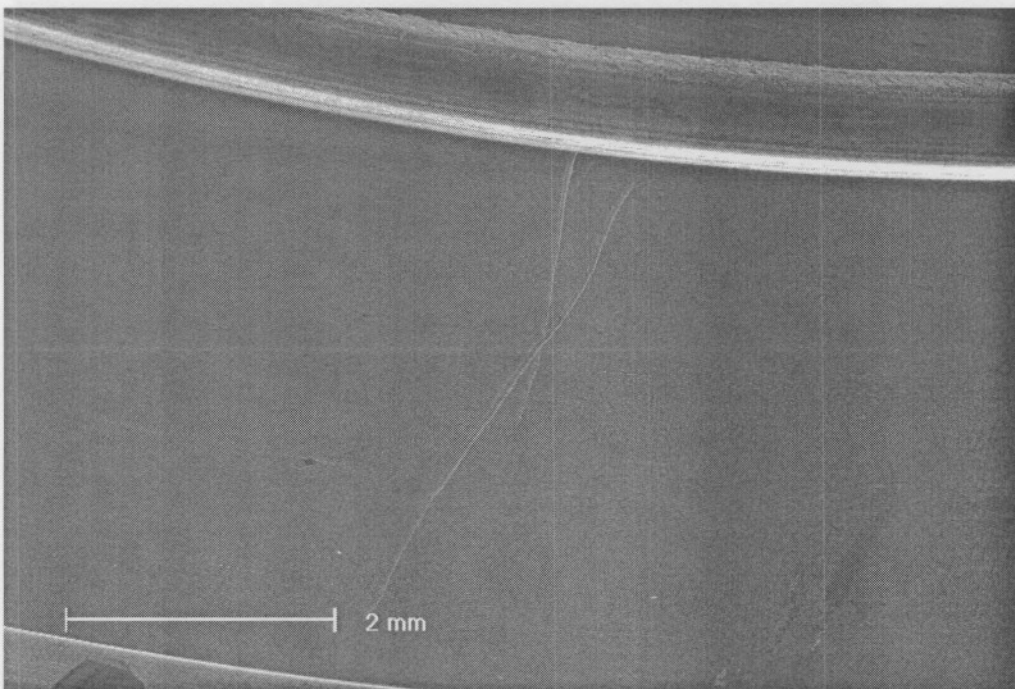


Figure 216: Outer race defect (medium)

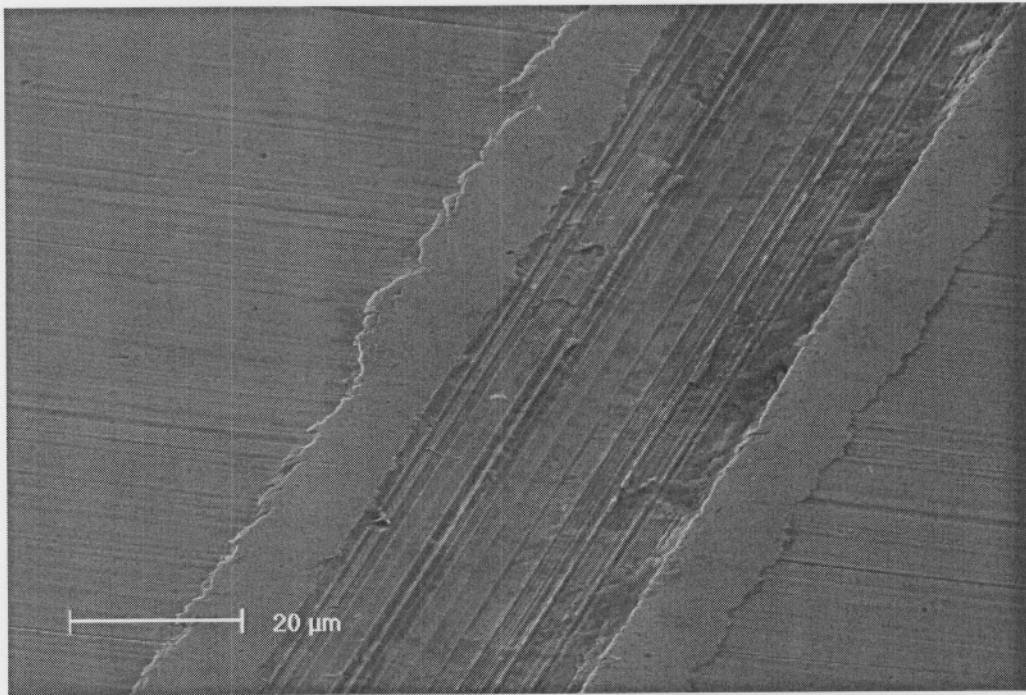


Figure 217: Outer race defect (medium)

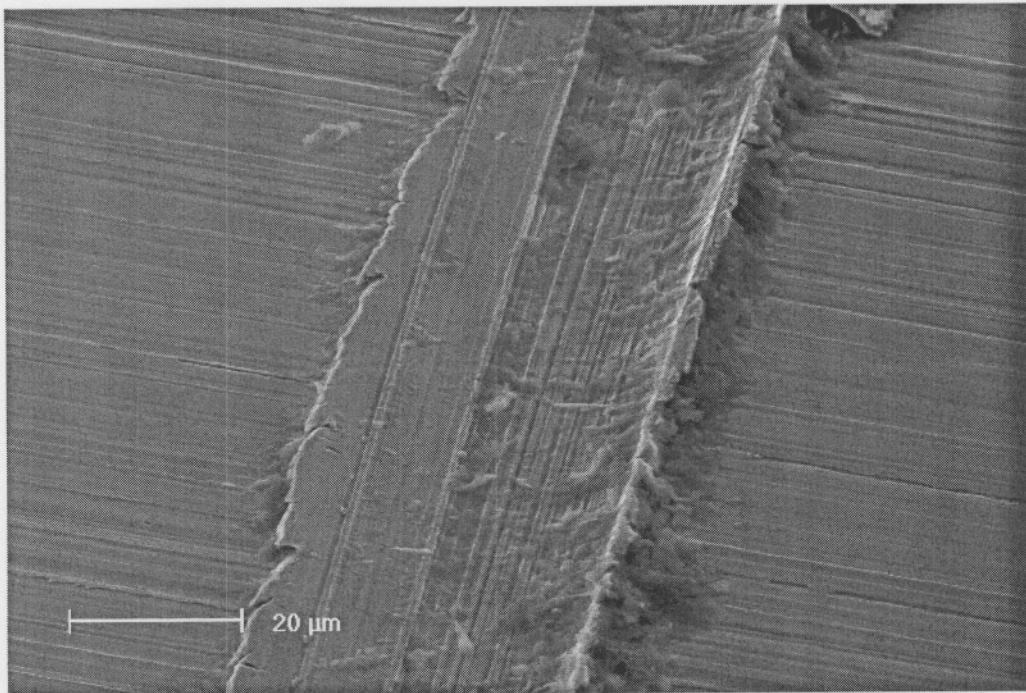


Figure 218: Outer race defect (medium)

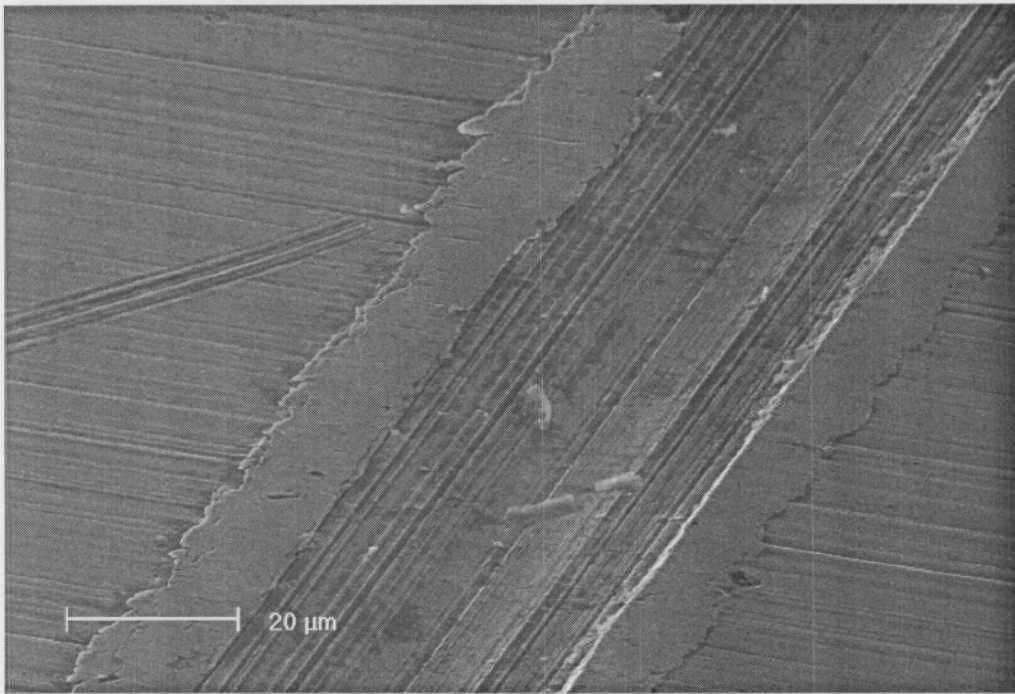


Figure 219: Outer race defect (medium)

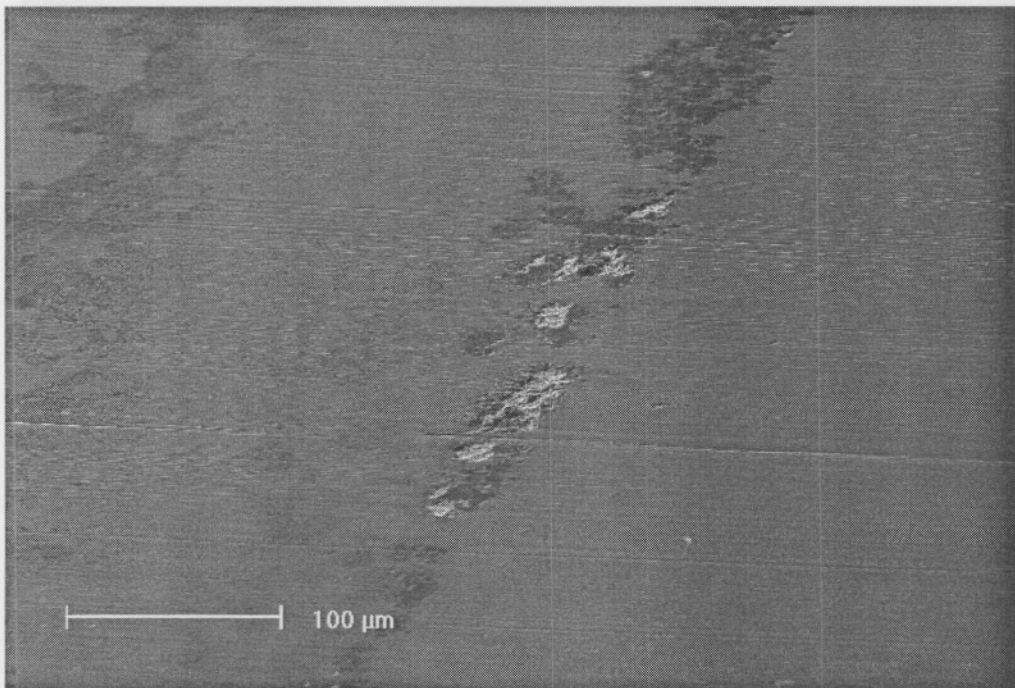


Figure 220: Rust marks

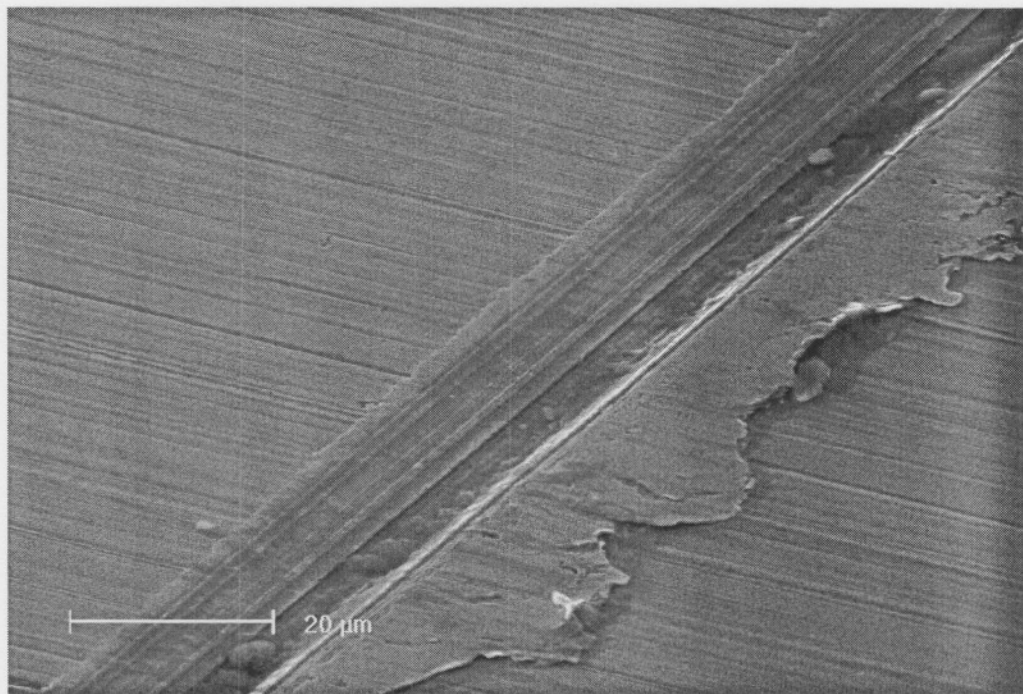


Figure 221: Outer race defect (small)

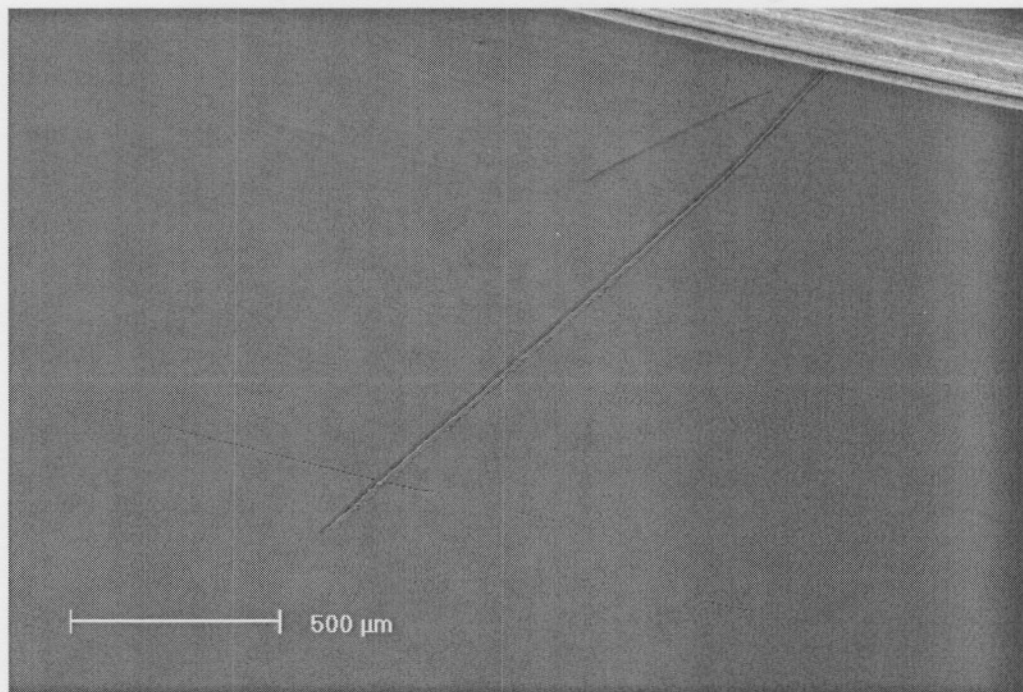


Figure 222: Outer race defect (small)

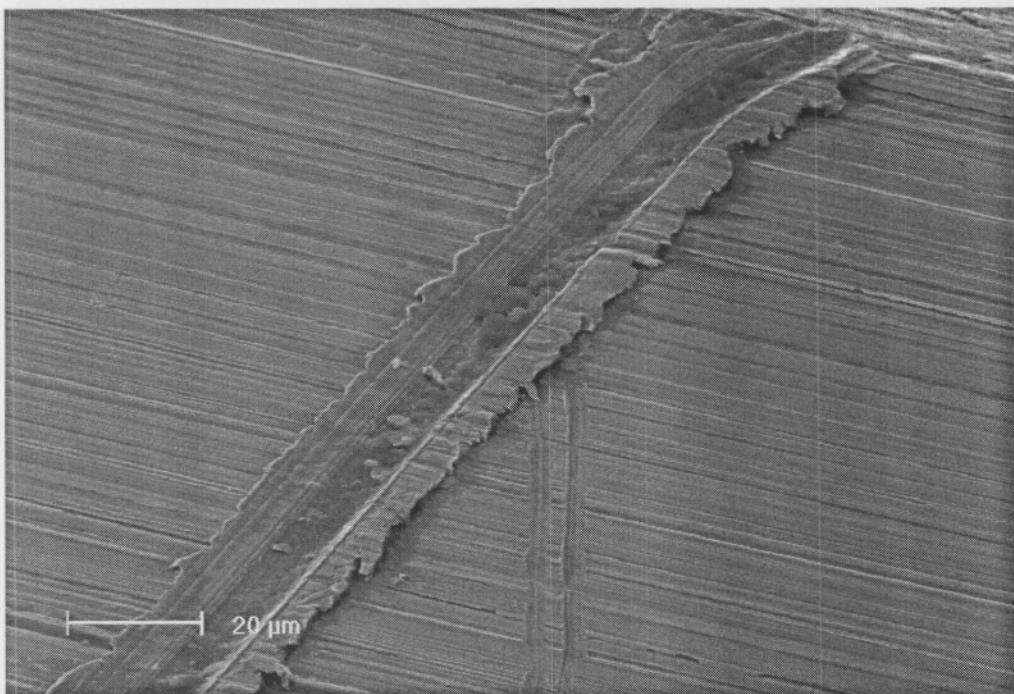


Figure 223: Outer race defect (small)

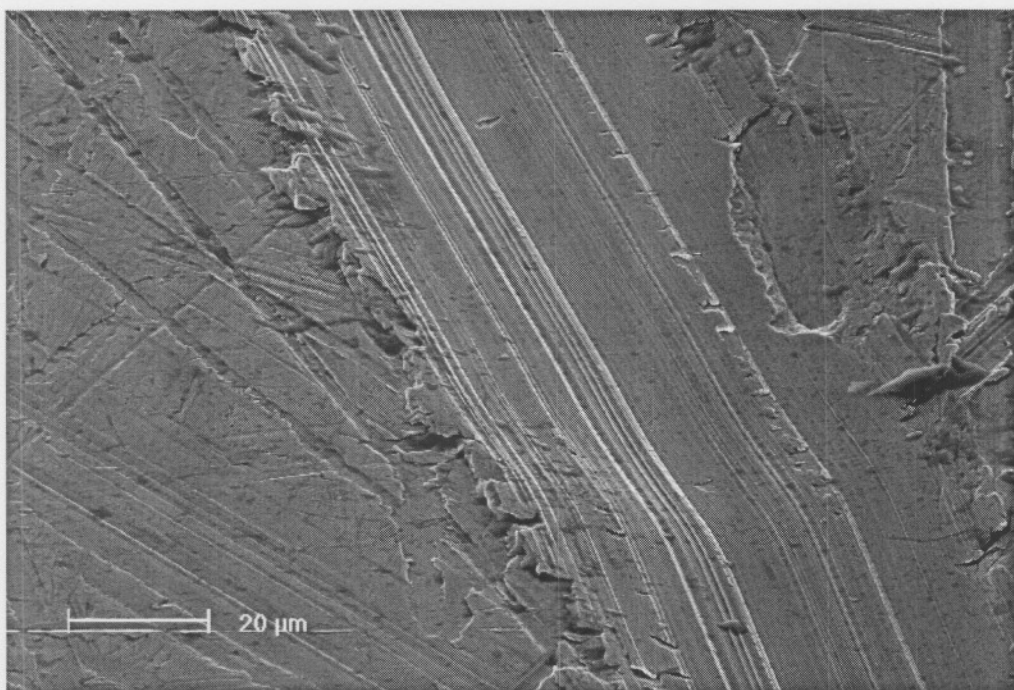


Figure 224: Scratch marks

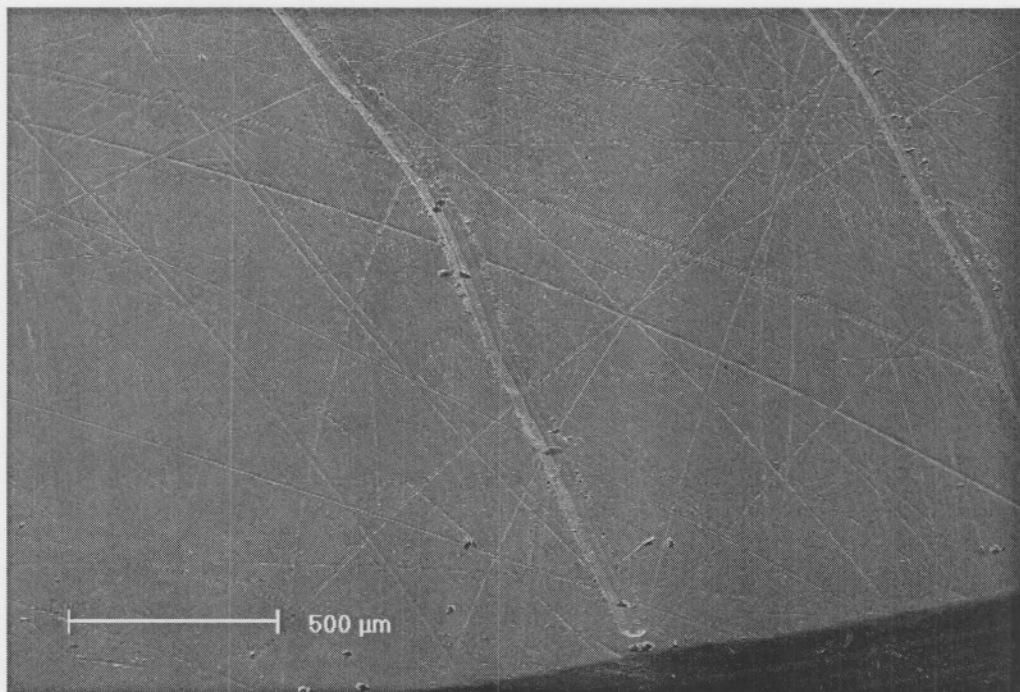
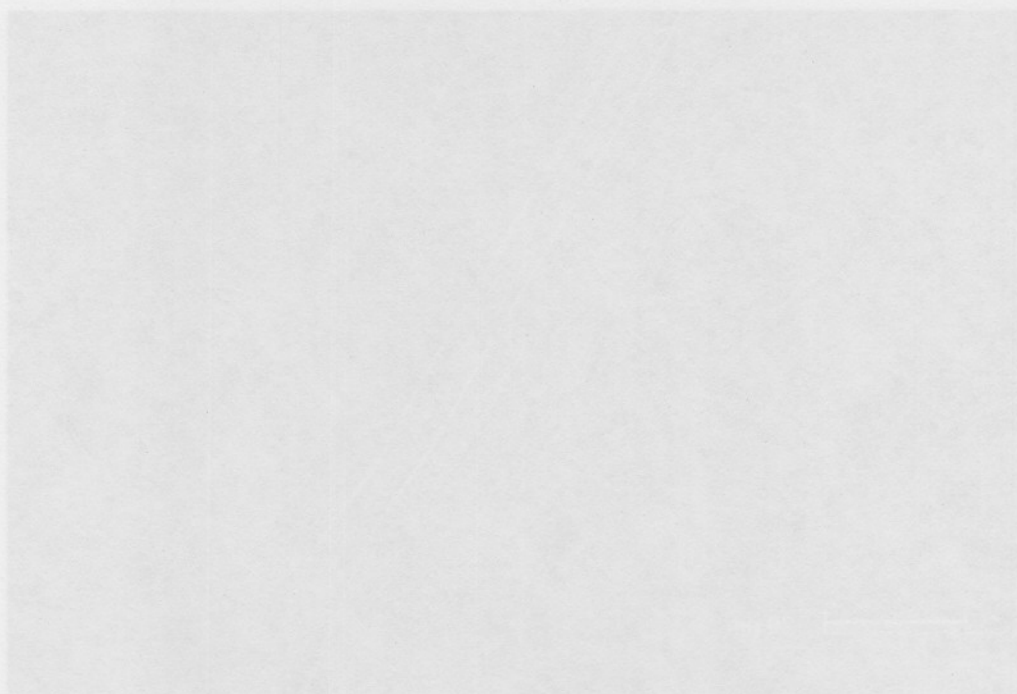


Figure 225: Scratch marks



- [11] G. K. Chaturvedi. *Adaptive signal processing techniques and condition monitoring of rotating machines*. A thesis submitted for the degree of Doctor of Philosophy, University of Southampton, September 1981.
- [12] R. B. Darlington. Is kurtosis really peakedness? *American Statistician*, 24:19–22, 1970.
- [13] J. R. Deller, J. G. Proakis and J. H. L. Hansen. *Discrete-Time Processing of Speech Signals*. Prentice-Hall, 1987.
- [14] J. H. Friedman, F. Baskett and L. J. Shustek. An algorithm for finding nearest neighbours. *IEEE Transactions on Computers*, October: 1000–1006, 1975.
- [15] M. Ghogho and M. Ibnkahla. Analytic behavior of the LMS adaptive line enhancer for sinusoids corrupted by multiplicative and additive noise. *IEEE Transactions on Signal Processing*, 46:(9): 2386–2393, September 1998.
- [16] T. A. Harris. *Rolling Bearing Analysis*. Wiley: 4th edition, 2001.
- [17] E. Hattiris, P. A. Orton, J. F. Poliakoff and P. D. Thomas. Analysis and calibration of an incremental motion encoder. In *Proceedings of the International Conference on Condition Monitoring*, pages 201–211, University of Wales, April 1999. Coxmoor.
- [18] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Macmillan, 1994.
- [19] P. S. Heyns. Private communication.
- [20] D. K. Hildebrand. Kurtosis measures bimodality? *American Statistician*, 25:42. February 1971.
- [21] D. Ho and R. B. Randall. Optimisation of bearing diagnostic techniques using simulated and actual bearing fault signals. *Mechanical Systems and Signal Processing*, 14(5):763–788, 2000.

- [22] A. J. Hoffman, C. Hoogenboezem, N. T. van der Merwe and C. J. A. Tollig. Seismic buffer recognition using mutual information for selecting wavelet based features. In *Proceedings of the International Conference on Signal Processing Applications & Technology (ICSPAT'98)*, pages 728–732, Toronto, September 1998.
- [23] A. J. Hoffman, R. E. Tessendorf and N. T. van der Merwe. Neuro-fuzzy network for condition monitoring of bearings in rotating machines. In *Proceedings of the Twelfth South African Universities Power Engineering Conference (SAUPEC'2003)*, Technikon Pretoria, South Africa. Chosen as plenary session paper.
- [24] A. J. Hoffman and C. J. A. Tollig. Neural network recognition of partial discharge signals. In *Proceedings of the Seventh South African Universities Power Engineering Conference*, University of Stellenbosch, South Africa, January 1997.
- [25] A. J. Hoffman and N. T. van der Merwe. A comparative evaluation of neural classification techniques for identifying multiple fault conditions. In *Proceedings of the 5th WSEAS International Conference on Circuits, Systems, Communications and Computers (CSCC 2001)*, pages 209–214, Greece, July 2001.
- [26] A. J. Hoffman and N. T. van der Merwe. Novel signal processing techniques applied to vibrational signals. In *Proceedings of the Twelfth Southern African Universities Power Engineering Conference (SAUPEC'2003)*, Technikon Pretoria, South Africa, January 2003.
- [27] A. J. Hoffman and N. T. van der Merwe. The application of neural networks to vibrational diagnostics for multiple fault conditions. *Computer Standards & Interfaces*, 24(2):139–149, June 2002.
- [28] A. J. Hoffman, N. T. van der Merwe, P. S. Heyns, C. Scheffer and C. Stander. The application of neural networks to vibrational diagnostics for

- multiple fault conditions. In *Proceedings of the Condition Monitoring and Diagnostic Engineering Management (COMADEM) Conference*, pages 537–544, Manchester, September 2001.
- [29] A. Hyvarinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.
 - [30] L. B. Jack and A. K. Nandi. Fault detection using support vector machines and artificial neural networks, augmented by genetic algorithms. *Mechanical Systems and Signal Processing*, 16(2–3):373–390, 2002.
 - [31] L. B. Jack and A. K. Nandi. Genetic algorithms for feature selection in machine condition monitoring with vibration signals. *IEE Proceedings-Part VIS*, 147(3):205–212, 2000.
 - [32] L. B. Jack and A. K. Nandi. Support vector machines for detection and characterization of rolling element bearing faults. *Proceedings Institution of Mechanical Engineers, Part C*, 215(C9):1065–1074, 2001.
 - [33] L. B. Jack, A. K. Nandi and A. C. McCormick. Diagnosis of rolling element bearing faults using radial basis functions. *Applied Signal Processing*, 6(1):25–32, 1999.
 - [34] J. S. Jesse, W. Hines and A. Edmondson. Motor shaft misalignment bearing load analysis. In *Proceedings of the Maintenance and Reliability Conference (MARCON 99)*, Tennessee, May 1999.
 - [35] T. Kohonen. The self-organizing map. *Neurocomputing*, 21:1–6, 1998.
 - [36] N. Končar. *Optimisation methodologies for direct inverse neurocontrol*. Master's thesis, University of London, 1997.
 - [37] R. Krishnan, G. Sivakumar and P. Bhattacharya. Extracting decision trees from trained neural networks. *Pattern Recognition*, 32:1999–2009, 1999.
 - [38] B. Li, M.-Y. Chow, Y. Tipsuwan and J.C. Hung. Neural-network-based motor rolling bearing fault diagnosis. *IEEE Transactions on Industrial*

Electronics, 47(5):1060–1069, October 2000.

- [39] G. Lipovszky, K. Sólymovári and G. Varga. *Vibration testing of machines and their maintenance*. Elsevier Science, 1990.
- [40] K. A. Loparo, M. L. Adams, W. Lin, M. Farouk Abdel-Magied and N. Afshari. Fault detection and diagnosis of rotating machinery. *IEEE Transactions on Industrial Electronics*, 47(5):1005–1013, October 2000.
- [41] A. Loskiewicz-Buczak and R. E. Uhrig. Hybrid neural network – fuzzy logic diagnosis system for vibration monitoring. In *Proceedings of the 1993 Artificial Neural Networking Engineers*, 3:757–762, 1993.
- [42] A. C. McCormick and A. K. Nandi. A comparison of artificial neural networks and other statistical methods for rotating machine classification. In *Proceedings of the 1997 IEE Colloquium on Modelling and Signal Processing for Fault Diagnosis*, (260):1–6, 1996.
- [43] A. C. McCormick and A. K. Nandi. Bispectral and trispectral features for machine condition diagnosis. *IEE Proceedings – Part VIS*, 146(5):229–234, 1999.
- [44] A. C. McCormick and A. K. Nandi. Classification of rotating machine condition using artificial neural networks. *Proceedings Institution of Mechanical Engineers*, Part C, 211(C6):439–450, 1997.
- [45] A. C. McCormick and A. K. Nandi. *Condition monitoring using periodic time-varying AR models*. In *Signal analysis and prediction (edited by A Prochazka, J Uhler, P J W Rayner and N G Kingsbury)*, pages 197-204, Birkhauser, Boston, USA, 1998.
- [46] A. C. McCormick and A. K. Nandi. Cyclostationarity in rotating machine vibrations. *Mechanical Systems and Signal Processing*, 12(2):225–242, 1998.

- [47] A. C. McCormick and A. K. Nandi. Real time classification of rotating shaft loading conditions. *IEEE Transactions on Neural Networks*, 8(3):748–757, 1997.
- [48] A. C. McCormick, A. K. Nandi and L. B. Jack. Application of periodic time-varying autoregressive models to the detection of bearing faults. In *Proceedings of the Institution of Mechanical Engineers*, 212(6):417–428, September 1998.
- [49] A. C. McCormick, A. K. Nandi and L.B. Jack. Digital signal processing algorithms in condition monitoring. *International Journal of COMADEM*, 1(3):5–14, 1998.
- [50] P. D. McFadden and P. L. McFadden. Application of signal averaging to vibration monitoring of rolling element bearings (I) Introduction and pilot study. Report no. OUEL 2146/97, University of Oxford, Department of Engineering Science, Oxford, 1997.
- [51] P. D. McFadden and J. D. Smith. Model for the vibration produced by a single point defect in a rolling element bearing. *Journal of Sound and Vibration*, 96(1):69–82, 1984.
- [52] P. D. McFadden and J. D. Smith. Model for the vibration produced by multiple point defects in a rolling element bearing. *Journal of Sound and Vibration*, 98(2):263–272, 1985.
- [53] P. D. McFadden and M. M. Toozhy. Application of signal averaging to vibration monitoring of rolling element bearings (II) Determining damage distribution on the inner race. Report no. OUEL 2216/99, University of Oxford, Department of Engineering Science, Oxford.
- [54] P. D. McFadden and M. M. Toozhy. Application of synchronous averaging to vibration monitoring of rolling element bearings. *Mechanical Systems and Signal Processing*, 16(2–3):373–390, 2002.

- [55] John S. Mitchell. *An Introduction to Machinery Analysis and Monitoring*. Pennwell, 2nd edition, 1993.
- [56] G. Monari and G. Dreyfus. Withdrawing an example from the training set: An analytical estimation of its effect on a non-linear parameterised model. *Neurocomputing*, 35:195–201, 2000.
- [57] J. J. A. Moors. The meaning of kurtosis: Darlington reexamined. *American Statistician*, 40:283–284, 1986.
- [58] A. K. Nandi (Ed.) *Blind Estimation using Higher-Order Statistics*. Kluwer, 1999.
- [59] A. K. Nandi (Ed.) *Higher order statistics in signal processing*. In *Signal analysis and prediction (edited by A Prochazka, J Uhler, P J W Rayner and N G Kingsbury)*, pages 57-78, Birkhauser, Boston, USA, 1998.
- [60] A. K. Nandi and L. Jack *Advanced digital vibration signal processing for condition monitoring*. In *International Journal of COMADEM*, 7(1):3-12, 2004.
- [61] Neural Network Toolbox 2.0. The Mathworks: Available from <http://www.mathworks.com/products/neuralnet/>. Accessed September 13 2003.
- [62] C. L. Nikias and A. P. Petropulu. *Higher-Order Spectra Analysis – a Nonlinear Signal Processing Framework*. Prentice Hall, 1993.
- [63] N. G. Nikolaou and I. A. Antoniadis. Application of wavelet packets in bearing fault diagnosis. In *Proceedings of the 5th WSEAS International Conference on Circuits, Systems, Communications and Computers (CSCC 2001)*, pages 12–19, Greece, July 2001.
- [64] M. P. Norton. *Fundamentals of Noise and Vibration Analysis for Engineers*. Cambridge University Press, 1989.

- [65] H. Oehlman, D. Brie, M. Tomczak and A. Richard. A method for analysing gearbox faults using time frequency representations, *Mechanical systems and signal processing*, 11(4):529–545, 1997.
- [66] A. V. Oppenheim and R. W. Schaffer. *Digital signal processing*, Prentice-Hall, 1975.
- [67] L. R. Padovese. Comparison between probabilistic and multilayer perceptron neural networks for rolling bearing fault classification. *International Journal of Modelling and Simulation*, 22(2), 2002.
- [68] A. Papoulis. *Signal Analysis*. McGraw-Hill, 1977.
- [69] B. A. Paya, I. I. Esat and M. N. M. Badi. Artificial neural network based fault diagnostics of rotating machinery using wavelet transforms as a preprocessor. *Mechanical Systems and Signal Processing*, 11(5):751–765, 1997.
- [70] B. Porat. *A Course in Digital Signal Processing*. Wiley, 1997.
- [71] D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999.
- [72] R. B. Randall. Bearing diagnostics in helicopter gearboxes. In *Proceedings of the Condition Monitoring and Diagnostic Engineering Management (COMADEM) Conference*, pages 1–11, Manchester, September 2001.
- [73] R. B. Randall. *Frequency Analysis*. Bruel and Kjaer, 1987.
- [74] R. B. Randall, J. Antoni and S. Chobsaard. The relationship between spectral correlation and envelope analysis in the diagnostics of bearing faults and other cyclostationary machine signals. *Mechanical Systems and Signal Processing*, 15(5):945–962, 2001.
- [75] B. K. N. Rao. *Handbook of Condition Monitoring*. Elsevier Science, 1996.
- [76] Raj B. K. N. Rao and Asoke K Nandi (ed.), . *Condition Monitoring, Diagnostic and Engineering Management (COMADEM) Conference*.

COMADEM International (UK), 2002.

- [77] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.
- [78] J. Shirosi, Y. Li, S. Liang, T. Kurfess and S. Danyluk. Bearing condition diagnostics via vibration and acoustic emission measurements. *Mechanical Systems and Signal Processing*, 11(5):693–705, 1997.
- [79] D. B. Skalak. *Prototype selection for composite nearest neighbour classifiers*. A dissertation in the Department of Computer Science, University of Massachusetts, Amherst, 1997.
- [80] C. J. Stander, P. S. Heyns and W. Schoombie. Fault detection on gearboxes operating under fluctuating load conditions. In *Proceedings of the Condition Monitoring and Diagnostic Engineering Management (COMADEM) Conference*, pages 537–544, Manchester, September 2001.
- [81] F. G. Stremler. *Introduction to Communication Systems*. Addison-Wesley, 1982.
- [82] Y.-T. Su and S.-J. Lin. On initial fault detection of a tapered roller bearing: frequency domain analysis. *Journal of Sound and Vibration*, 155(1):75–84, 1992.
- [83] C.-T. Sun and J.-S. Jang. A neuro-fuzzy classifier and its applications. In *Proceedings of the 1993 IEEE International Conference on Fuzzy Systems*, pages 94–98, San Francisco, March 1993.
- [84] N. Tandon and A. Choudbury. An analytical model for the prediction of the vibration response of rolling element bearings due to a localized defect. *Journal of Sound and Vibration*, 205(3):275–292, 1997.
- [85] N. Tandon and A. Choudbury. A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings. *Tribology International*, 32:469–480, 1999.

- [86] J. I. Taylor, *The Vibration Analysis Handbook*. Vibration Consultants, 1994.
- [87] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 1988.
- [88] L. Tiedt. Private correspondence. Laboratory for electron microscopy, Potchefstroom University for Christian Higher Education.
- [89] C. J. A. Tollig. *Combining wavelets and neural networks in classifying and detecting transient signals*. M.Eng dissertation at Potchefstroom University for Christian Higher Education, 1998.
- [90] R. E. Uhrig. Integration of artificial intelligence systems into a monitoring and diagnostic system for nuclear power plants. In *The Special Meeting on Instrumentation and Control of the Halden Research Centre*, Lillhammer, 1998.
- [91] University of California at Irvine Repository of Machine Learning Databases. For information contact ml-repository@ics.uci.edu.
- [92] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice-Hall, 1993.
- [93] N. T. van der Merwe and A. J. Hoffman. A modified cepstrum analysis applied to vibrational signals. In *Proceedings of the Conference on Digital Signal Processing Proceedings (DSP2002)*, pages 873–8, Greece, July 2002.
- [94] N. T. van der Merwe and A. J. Hoffman. A nearest neighbour rule with class membership (NNRC) for modelling problems. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'01)*, pages 2104–2109, Washington D.C., July 2001. Co-sponsored by IEEE Neural Network Society, IEEE and International Neural Network Society.
- [95] N. T. van der Merwe and A. J. Hoffman. Applying digital signal processing techniques to improve the signal to noise ratio in vibrational signals. In

Proceedings of the 5th WSEAS International Conference on Circuits, Systems, Communications and Computers (CSCC 2002), Greece, July 2002.

- [96] N. T. van der Merwe and A. J. Hoffman. Closed loop optimisation of features for neural classifiers. In *Proceedings of IEEE South African Symposium on Communications and Signal Processing (COMSIG'98)*, pages 79–82, University of Cape Town, South Africa, September 1998.
- [97] N. T. van der Merwe and A. J. Hoffman. Developing an efficient cross validation strategy to determine classifier performance (CVCP). In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'01)*, pages 1663–1668, Washington D.C., July 2001. Co-sponsored by IEEE Neural Networks Society, IEEE and International Neural Network Society.
- [98] N. T. van der Merwe, A. J. Hoffman, C. Stander and S. P. Heyns. Identifying multiple faults in rotating machinery. In *Proceedings of the Tenth South African Universities Power Engineering Conference (SAUPEC'2001)*, pages 36–39, University of Cape Town, South Africa, January 2001. Chosen as plenary session paper.
- [99] P. A. van Vuuren. Neuro-fuzzy techniques in pattern recognition. Internal report, Potchefstroom University for Christian Higher Education, 1998.
- [100] P. A. van Vuuren. *Pattern recognition with the aid of a neuro-fuzzy network*. M.Eng dissertation at Potchefstroom University for Christian Higher Education, Potchefstroom, 1999.
- [101] E. M. P. van Wyk. *The application of neural networks and digital signal processing techniques for condition monitoring in a power station*. M.Eng dissertation at Potchefstroom University for Christian Higher Education, 1998.

- [102] A. Ypma, R. Ligteringen, E. E. E. Friedman and R. P. W. Duin. Recognition of bearing failures using wavelets and neural networks. In *Proceedings of the UK Symposium on Applications of Time-Frequency and Time-Scale Methods*, Coventry, UK, August 1997.
- [103] A. Ypma and P. Petheri. Second-order ICA in machine vibration analysis, Pattern Recognition Group, Faculty of Applied Science, Delft University of Technology, July 1998. The Web address of this paper is: http://www.ph.tn.tudelft.nl/~ypma/papers/PostScript/report_hut.ps
- [104] M. F. White. Rolling element bearing vibration transfer characteristics: effect of stiffness. *Journal of Applied Mechanics*, 46:677–684, September 1979.
- [105] W. Wowk. *Machinery Vibration: Measurement and Analysis*. McGraw-Hill, 1995.
- [106] J. R. Zeidler, E. H. Satorius, D. M. Chabries and H. T. Wexler. Adaptive enhancement of multiple sinusoids in uncorrelated noise. *IEEE Transactions on Signal Processing*, 26(3):240–254, June 1978.
- [107] C. Zhang. *Defect prediction and life prediction of rolling element bearings*. PhD thesis in Mechanical Engineering, Georgia Institute of Technology, March 2001.
- [108] C. Zhang, Y. Li, T. Kurfess, S. Danyluk and S. Liang. Minimizing the influence of operating condition variations in bearing diagnostics. *International Journal of Condition Monitoring and Diagnostic Engineering Management(COMADEM)*, 2(1): 23–29, 1999. Elsevier Science.

10 List of publications

- 1 A. J. Hoffman, R. E. Tessendorf and N. T. van der Merwe. Neuro-fuzzy network for condition monitoring of bearings in rotating machines. In *Proceedings of the Twelfth South African Universities Power Engineering Conference (SAUPEC'2003)*, Technikon Pretoria, South Africa, January 2003. Chosen as plenary session paper.
- 2 A. J. Hoffman and N. T. van der Merwe. Novel signal processing techniques applied to vibrational signals. In *Proceedings of the Twelfth South African Universities Power Engineering Conference (SAUPEC'2003)*, Technikon Pretoria, South Africa, January 2003.
- 3 A. J. Hoffman, R. Hoogenboezem, N. T. van der Merwe and T. Tollig. Seismic buffer recognition using wavelet-based features and neural classification. *Geophysical Prospecting*, 50(4):361–372, July 2002.
- 4 A. J. Hoffman and N. T. van der Merwe. The application of neural networks to vibrational diagnostics for multiple fault conditions. *Computer Standards & Interfaces*, 24(2):139–149, June 2002.
- 5 N. T. van der Merwe and A. J. Hoffman. Applying digital signal processing techniques to improve the signal to noise ratio in vibrational signals. In *Proceedings of the 5th WSEAS International Conference on Circuits, Systems, Communications and Computers (CSCC 2002)*, Greece, July 2002.
- 6 N. T. van der Merwe and A. J. Hoffman. A modified cepstrum analysis applied to vibrational signals. In *Proceedings of the Conference on Digital Signal Processing Proceedings (DSP 2002)*, pages 873–878, Greece, July 2002.

- 7 A. J. Hoffman and N. T. van der Merwe. A comparative evaluation of neural classification techniques for identifying multiple fault conditions. In *Proceedings of the 5th WSEAS International Conference on Circuits, Systems, Communications and Computers (CSCC 2001)*, pages 209–214, Greece, July 2001.
- 8 A. J. Hoffman, N. T. van der Merwe, P. S. Heyns, C. Scheffer and C. Stander. The application of neural networks to vibrational diagnostics for multiple fault conditions. In *Proceedings of the Condition Monitoring and Diagnostic Engineering Management (COMADEM) Conference*, pages 537–544, Manchester, September 2001.
- 9 N. T. van der Merwe and A. J. Hoffman. A nearest neighbour rule with class membership (NNRC) for modelling problems. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'01)*, pages 2104–2109, Washington D.C., July 2001.
- 10 N. T. van der Merwe and C. P. Bodenstein. A new variable wordlength direct form filter structure and finite wordlength analysis with SARON. In *Proceedings of the 5th WSEAS International Conference on Circuits, Systems, Communications and Computers (CSCC 2001)*, pages 37–41, Greece, July 2001.
- 11 N. T. van der Merwe and A. J. Hoffman. Developing an efficient cross validation strategy to determine classifier performance (CVCP). In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'01)*, pages 1663–1668, Washington D.C., July 2001.
- 12 N. T. van der Merwe, A. J. Hoffman, C. Stander and S. P. Heyns. Identifying multiple faults in rotating machinery. In *Proceedings of the Tenth South African Universities Power Engineering Conference (SAUPEC'2001)*, pages 36–39, University of Cape Town, South Africa, January 2001. Chosen as plenary session paper.

- 13 N. T. van der Merwe and A. J. Hoffman. Automated pattern recognition techniques applied to condition monitoring of electrical machines using vibration signals. In *Proceedings of the Ninth South African Universities Power Engineering Conference (SAUPEC'2000)*, University of Natal, South Africa, January 2000.
- 14 N. T. van der Merwe and A. J. Hoffman. Automated pattern recognition techniques applied to PD monitoring in large electrical machines. In *Proceedings of the Eighth South African Universities Power Engineering Conference (SAUPEC'99)*, Potchefstroom University for Christian Higher Education, South Africa, January 1999.
- 15 N. T. van der Merwe and A. J. Hoffman. Closed loop optimisation of features for neural classifiers. In *IEEE Proceedings of the 1998 South African Symposium on Communications and Signal Processing (COMSIG'98)*, pages 79–82, University of Cape Town, South Africa, September 1998. IEEE (South African Section).
- 16 N. T. van der Merwe. 'n Vergelyking van digitale filterstrukture deur wisselende woordlengte en 'n stuksgewyse reguitlynmodel vir eindige woordlengte te gebruik (*A comparison of digital filter structures using variable wordlength and a piecewise straight-line model for finite wordlength*). M.Eng dissertation at Potchefstroom University for Christian Higher Education, Potchefstroom, South Africa, 1998.
- 17 A. J. Hoffman, C. Hoogenboezem, N. T. van der Merwe and C. J. A. Tollig. Seismic buffer recognition using mutual information for selecting wavelet based features. In *IEEE International Symposium on Industrial Electronics (ISIE'98)*, pages 663–667, University of Pretoria, Pretoria, July 1998.
- 18 A. J. Hoffman, C. Hoogenboezem, N. T. van der Merwe and C. J. A. Tollig. Seismic buffer recognition using mutual information for selecting

wavelet based features. In *Proceedings of the International Conference on Signal Processing Applications & Technology (ICSPAT'98)*, pages 728–732, Toronto, September 1998.

- 19 N. T. van der Merwe. Local and global aspects of product quantization. *IEEE Proceedings of the South African Symposium on Communications and Signal Processing (COMSIG'97)*, pages 151–155, Rhodes University, Grahamstown, South Africa, 1997. IEEE (South African Section).

11 Appendices

11.1 Appendix A: Influence of phase shift on envelope spectrum

Denote the instantaneous value of the envelope spectrum by the vector quantity $H_e(f) = |H_e(f)|\angle H_e(f)$.

Suppose the point defect has a phase shift of $\angle Q$ degrees relative to $\angle H_e(f)$. If the vector representing the point defect Q at frequency f is $Q(f) = |Q(f)|\angle Q(f)$, the vector sum of $H_e(f)$ and Q is

$$\begin{aligned}
 Y &= H_e(f) + Q(f) \\
 &= |H_e(f)|\angle H_e(f) + |Q(f)|\angle Q(f) \\
 &= H_r + jH_i + Q_r + jQ_i \\
 \text{where } H_r &= |H_e(f)|\cos(\angle H_e(f)) \\
 H_i &= |H_e(f)|\sin(\angle H_e(f)) \\
 Q_r &= |Q(f)|\cos(\angle Q(f)) \\
 Q_i &= |Q(f)|\sin(\angle Q(f)) \\
 \text{Now} \\
 |Y|^2 &= H_r^2 + Q_r^2 \\
 &= |H_e(f)|^2 \cos^2(\angle H_e(f)) + |Q(f)|^2 \cos^2(\angle Q(f)) \\
 &\quad + 2|H_e(f)||Q(f)|\cos(\angle H_e(f))\cos(\angle Q(f)) \\
 &\quad + |H_e(f)|^2 \sin^2(\angle H_e(f)) + |Q(f)|^2 \sin^2(\angle Q(f)) \\
 &\quad + 2|H_e(f)||Q(f)|\sin(\angle H_e(f))\sin(\angle Q(f)) \\
 &= |H_e(f)|^2 + |Q(f)|^2 + 2|H_e(f)||Q(f)|\cos(\angle H_e(f) - \angle Q(f))
 \end{aligned}$$

Hence the maximum value of the expression for $|Y|$ will be reached when $\angle Q(f) = \angle H_e(f) + k * 360$ (where k is an integer and the angle is represented in degrees, not radians).

The minimum value of the expression for $|Y|$ will be reached when $\angle Q(f) = \angle H_e(f) + k * 180$ (where k is an integer and the angle is represented in degrees, not radians).

If $|Q(f)| = |H_e(f)|$ and Q is shifted in phase by 180° relative to H_e , complete cancellation occurs and $|Y(f)| = 0$.

However,

$$\begin{aligned} |Y(2f)|^2 &= |H_e(2f)|^2 + |Q(2f)|^2 + 2|H_e(2f)||Q(2f)|\cos(2k180) \\ |Y(2f)|^2 &= |H_e(2f)|^2 + |Q(2f)|^2 + 2|H_e(2f)||Q(2f)| \\ |Y(2f)|^2 &= (|H_e(2f)| + |Q(2f)|)^2 \end{aligned}$$

Thus the value of $|Y(2f)|$ is a maximum at frequency $2f$, as expected [55].

$$\begin{aligned} |Y(mf)|^2 &= |H_e(mf)|^2 + |Q(mf)|^2 + 2|H_e(mf)||Q(mf)|\cos(mkP) \\ |Y(mf)|^2 &= |H_e(mf)|^2 + |Q(mf)|^2 + 2|H_e(mf)||Q(mf)| \\ |Y(mf)|^2 &= (|H_e(mf)| + |Q(mf)|)^2 \end{aligned}$$

In general, the maximum value for $|Y(mf)|$ will be reached when $\cos(mkP) = 1$ (or $mkP = \pm 360r$, where r is any integer, P is the phase shift in degrees: $P = 360/m$ and m is the number of defects. Hence the value of $|Y(mf)|$ is a maximum at frequency mf , as expected [55].

From geometrical considerations the same result can be visualized readily with the aid of vector addition, as the maximum value will be obtained when the angle of $Q(mf)$ is in the same direction as the angle of $H_e(mf)$. As the vector $H_e(mf)$ represents the value of the envelope spectrum at frequency f , it is obvious that the maximum value will be reached when the vector representing the new defect $Q(mf)$ is aligned with vector $H_e(mf)$. Similarly, the minimum value will be obtained (from geometrical considerations) when the angle of $Q(mf)$ is opposite the angle of $H_e(mf)$, in other words shifted by 180° .

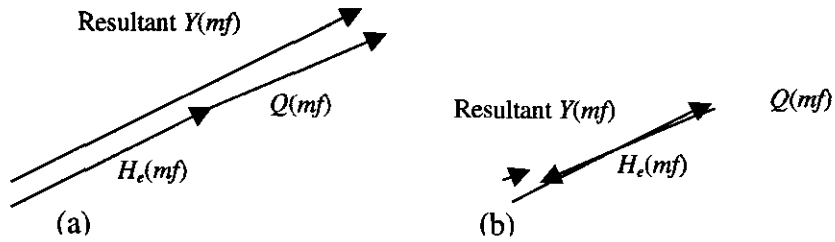


Figure 66: Result of vector addition when vectors add in phase (a) and out of phase (b)

It is clear that the maximum contribution to $H_e(mf)$ will be obtained when the vector to be added has exactly the same phase as the current phase angle of $H_e(mf)$ (i.e. when it adds in phase). Thus the defect vector should have (as far as possible) the same phase as $H_e(mf)$ if the maximum value is to be obtained. Any speed variations in the defect pulses would also change the contribution to $H_e(mf)$ owing to the vectors adding slightly out of phase, hence complicating the diagnosis.

11.2 Appendix B: Results on Vib demo data set 2 (figures)

11.2.1 Feature extraction (Vib demo data set 2)

The following features are shown sequentially in the following figures:

Fr_X: X axis vibration frequency component at the rotational frequency *Fr*

BPFO_X: X axis vibration frequency component at the *BPFO* frequency

BPFO2_X: X axis vibration frequency component in the envelope spectrum at the *BPFO* frequency

Fr_S_X: X axis frequency of *Fr*

BPFO2_S_X: X axis frequency of *BPFO*

BPFI2_S_X: X axis frequency of *BPFI*

BPFI2_X: X axis vibration frequency component in the envelope spectrum at the *BPFI* frequency

KH_X: Kurtosis of x axis signal

CH_X: Cepstrum of x axis signal

LH_X: Linear cepstrum of x axis signal

CHp_X: Cepstrum of x axis signal after bandpass filtering

KHp_X: Kurtosis of x axis signal after bandpass filtering

LHp_X: Linear cepstrum of x axis signal after bandpass filtering

FFT4_BPFO_X: X axis vibration frequency component at the *BPFO* frequency (signal raised to 4th power)

FFT8_BPFO_X: X axis vibration frequency component at the *BPFO* frequency (signal raised to 8th power)

SA_BPFO2_0_X: Square root of sum of squares of the synchronous average of the x axis signal at the *BPFO* frequency; the trigger signal is fixed at frequency *BPFO* (period $1/BPFO$)

SA_BPFO2_0_p_X: The maximum value of the synchronous average of the x axis signal at the *BPFO* frequency; the trigger signal is fixed at frequency *BPFO* (period $1/BPFO$)

BPFO2_10_Eng_X: Square root of the sum of squares of the first ten components in the envelope spectrum (at multiples of the *BPFO* frequency)

BPFO2_10_Max_X: Maximum value of the first ten components in the envelope spectrum (at multiples of the *BPFO* frequency)

BPFO2_2_Eng_X: Square root of the sum of squares of the first two components in the envelope spectrum (at multiples of the *BPFO* frequency)

BPFO2_2_Max_X: Maximum value of the first two components in the envelope spectrum (at multiples of the *BPFO* frequency)

HFD_5_10kHz_X: Square root of the sum of squares of the *HFD* components in the frequency band from 5–10kHz

HFD_2_5kHz_X: Square root of the sum of squares of the *HFD* components in the frequency band from 2–5kHz

HFD_5_6kHz_X: Square root of the sum of squares of the *HFD* components in the frequency band from 5–6kHz

HFD_6_7kHz_X: Square root of the sum of squares of the *HFD* components in the frequency band from 6–7kHz

HFD_8_10kHz_X: Square root of the sum of squares of the *HFD* components in the frequency band from 8–10kHz

FFT2_MEAN_X: First component of demodulation spectrum

For each data point marked by 'x' the defect angle (DA) (180° for the Vib demo data set 2), the bearing number (1 for the Vib demo data set 2), the bearing defect severity (5 for the Vib demo data set 2) and the distance of the defect from 225° are shown (refer to figure 43 on p. 214 for notational details). The dynamic load (unbalance mass in grams) is shown on the horizontal axis to demonstrate the influence of dynamic loading on bearing defects of decreasing severity and defect angle.

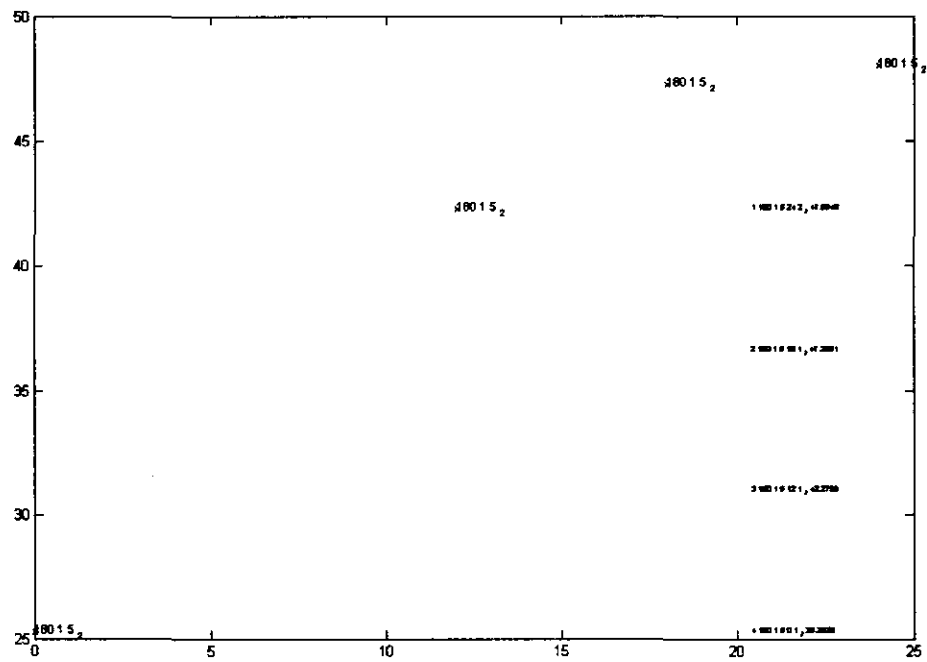


Figure 67: Feature *Fr_X* (magnitude spectrum value at rotational frequency on x axis) for large defect

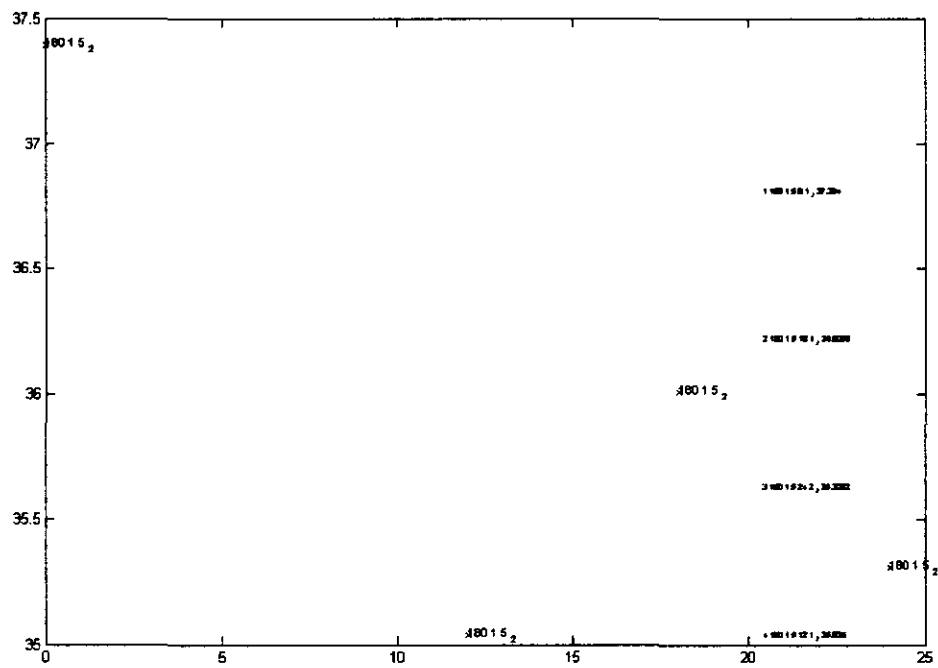
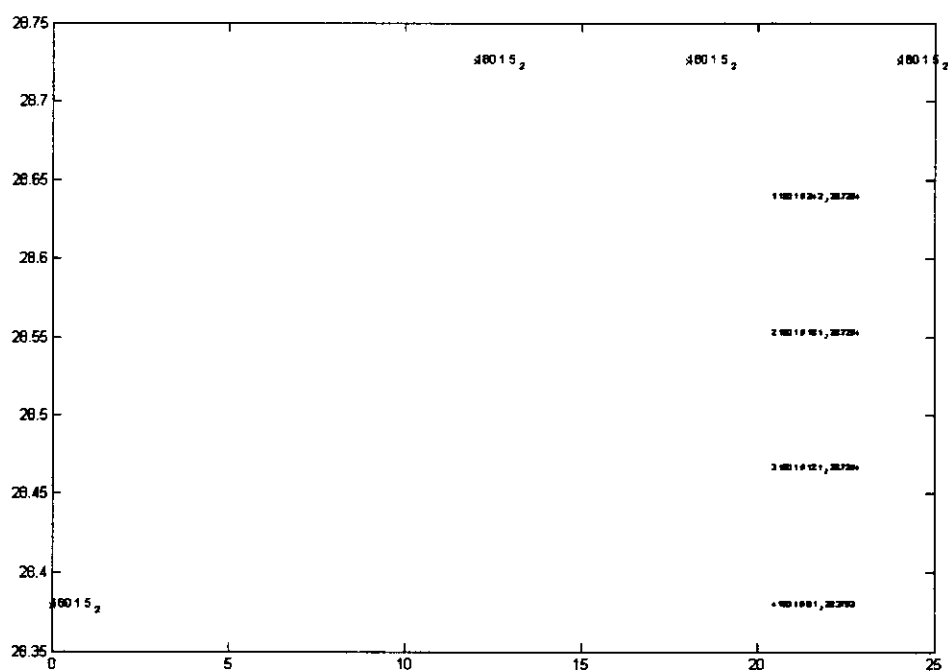
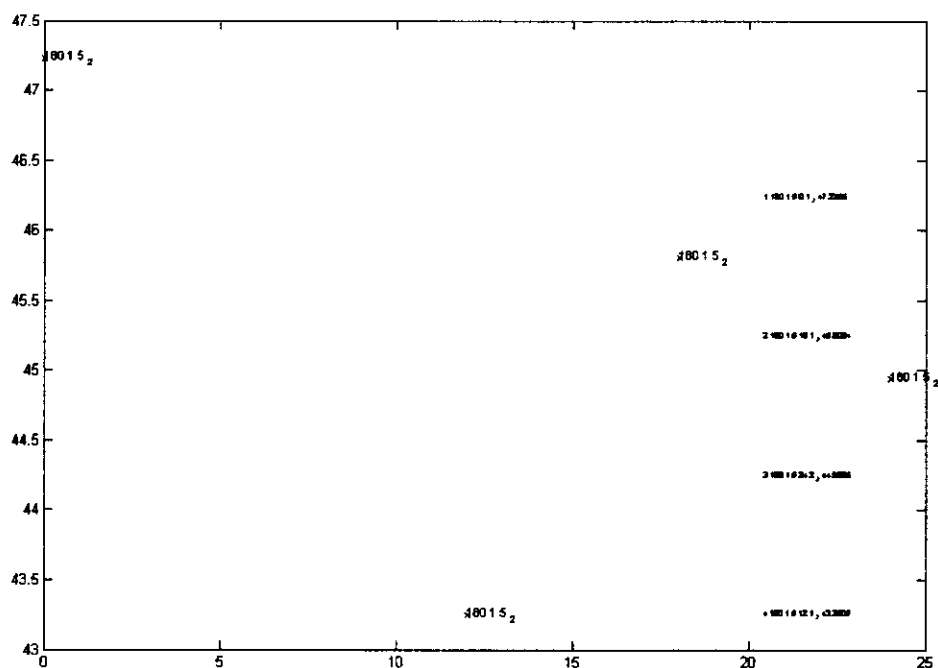
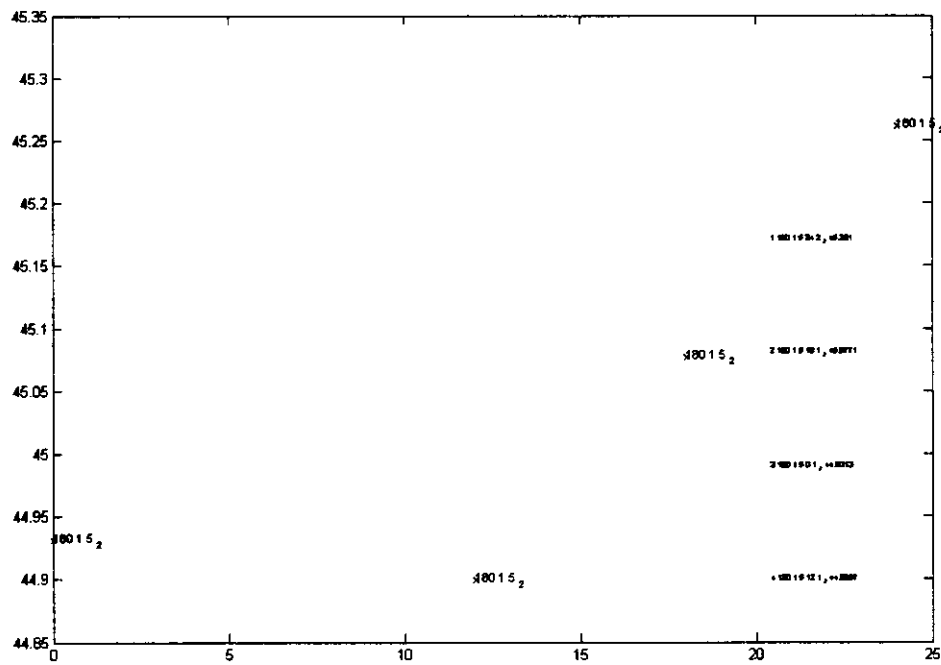
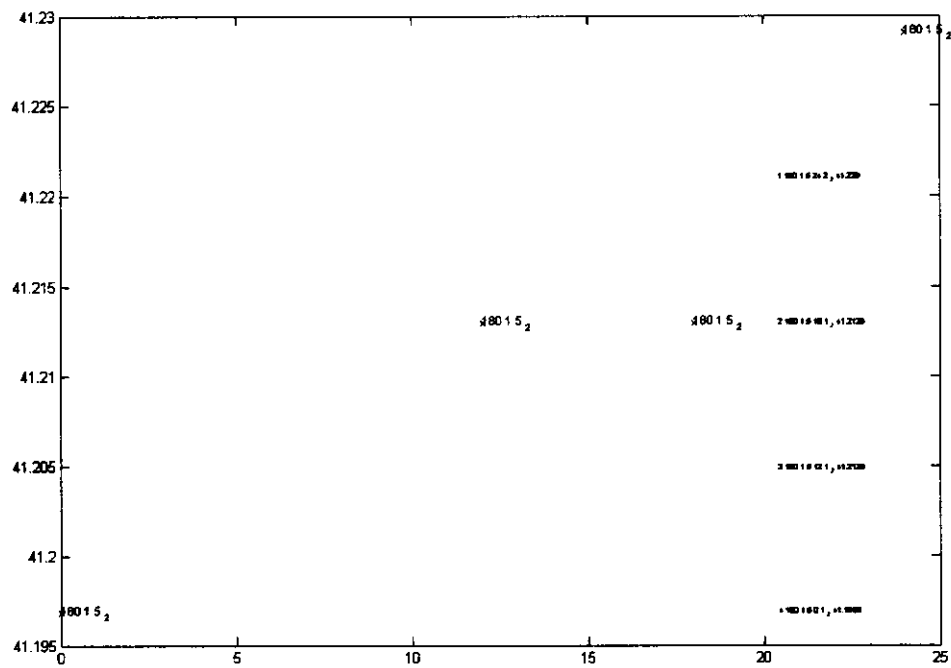


Figure 68: Feature *BPFO_X* (magnitude spectrum value at ball pass outer race defect frequency on x axis) for large defect





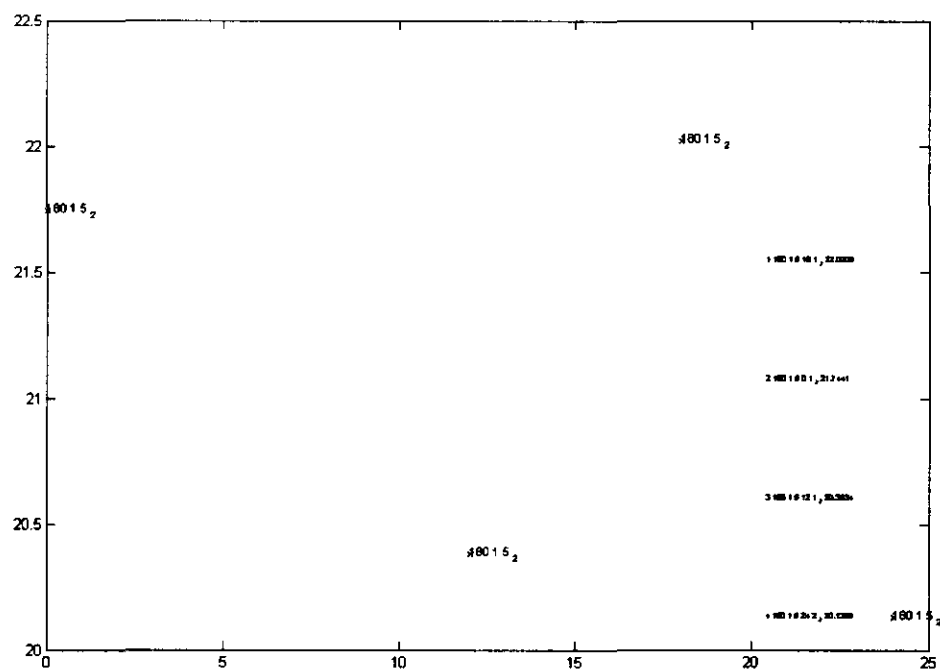


Figure 73: Feature *BPF12_X* (demodulation spectrum value at ball pass inner race defect frequency on x axis) for large defect

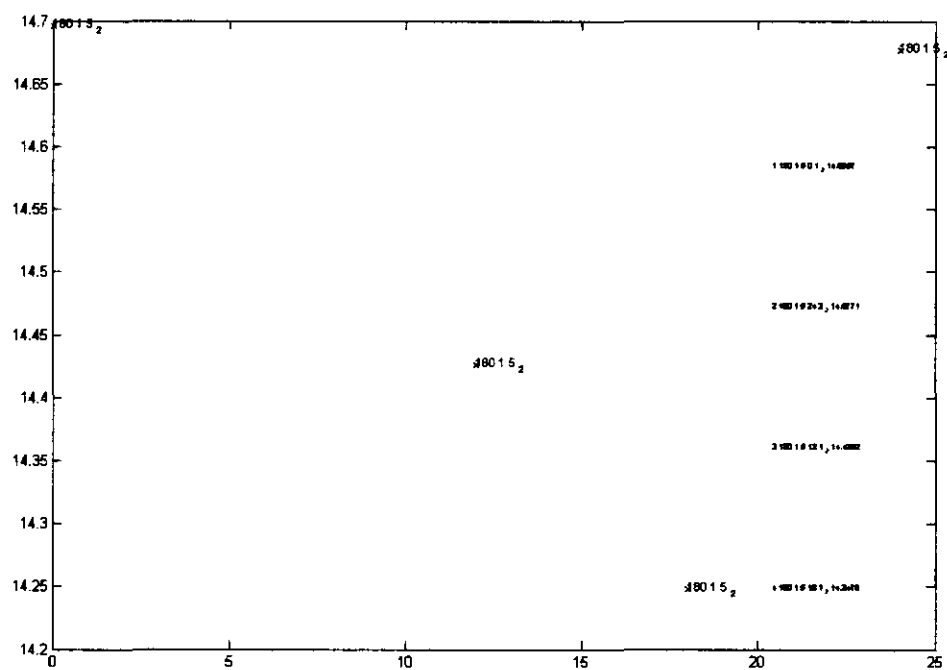


Figure 74: Feature *KH_X* (kurtosis on x axis) for large defect

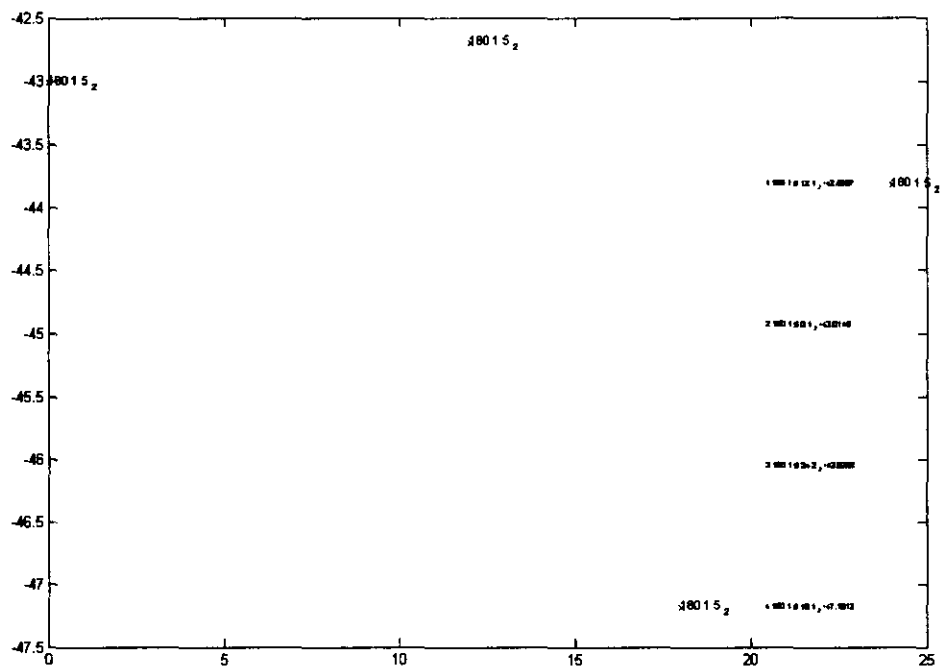


Figure 75: Feature CH_X (real cepstrum on x axis) for large defect

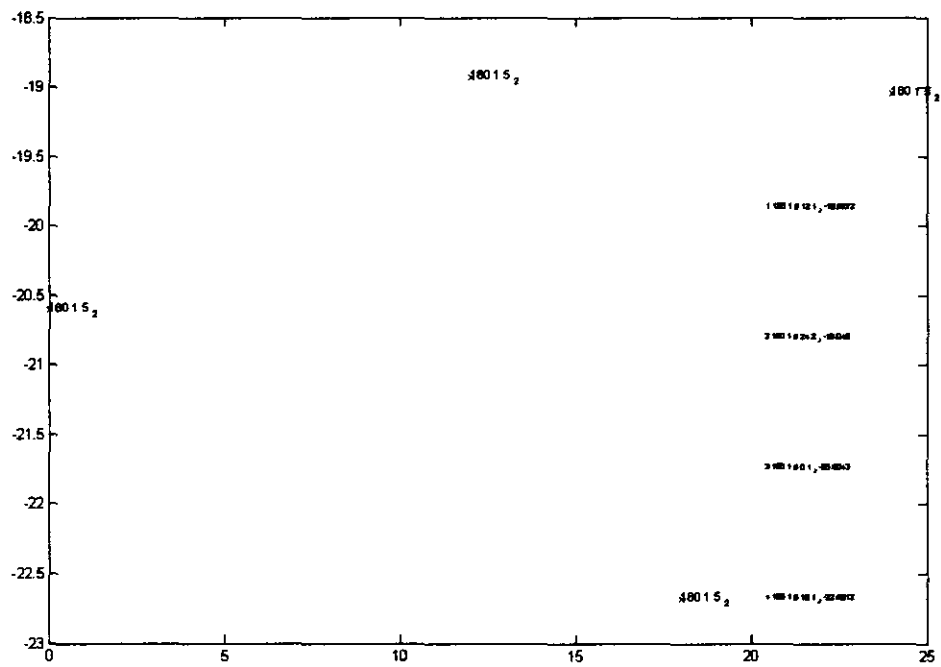


Figure 76: Feature LH_X (linear cepstrum on x axis) for large defect

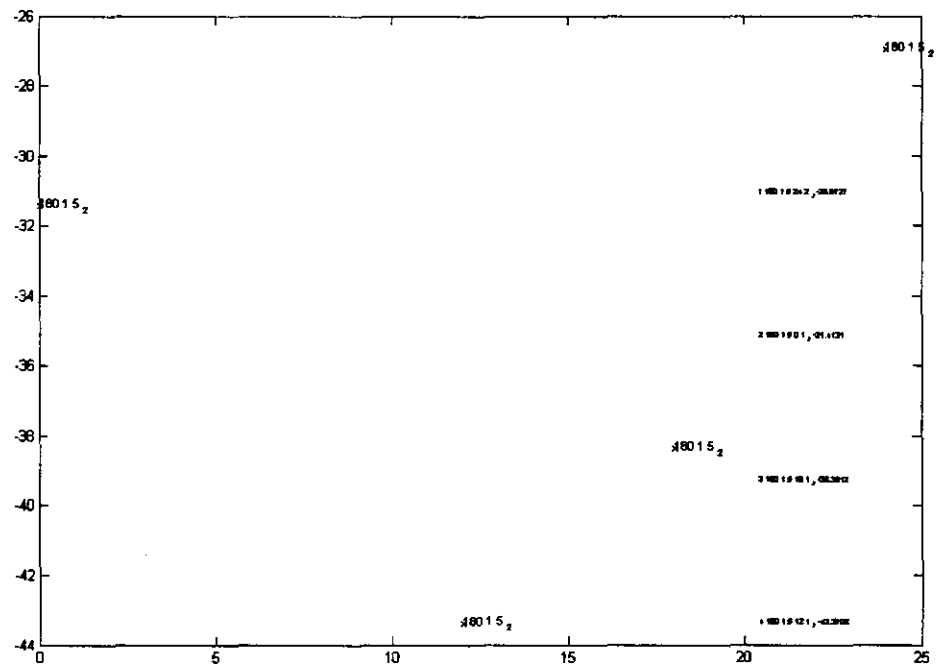


Figure 77: Feature CH_p_X (real cepstrum – after highpass filter – on x axis) for large defect

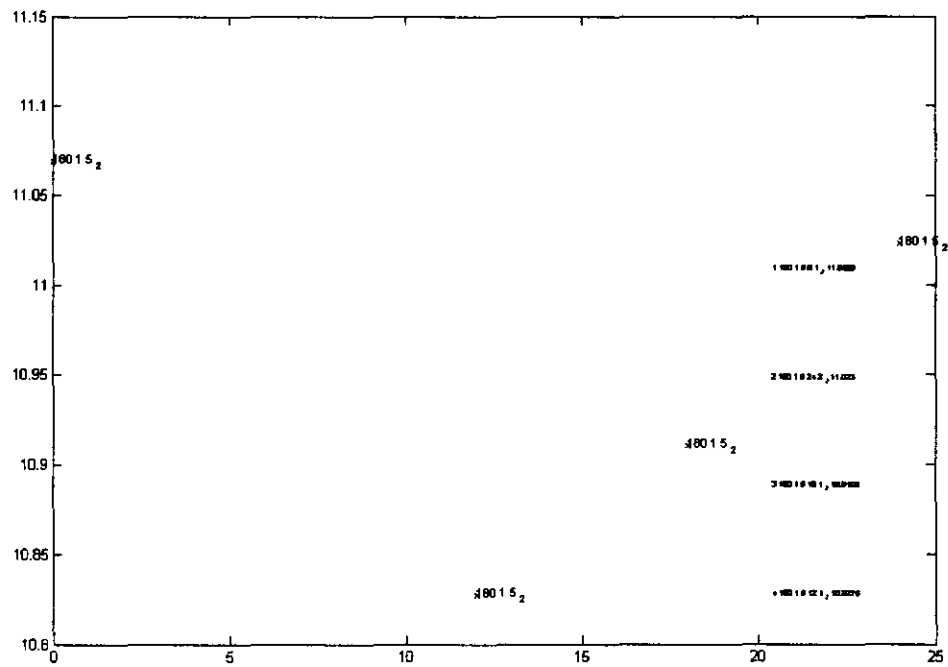


Figure 78: Feature KH_p_X (kurtosis – after highpass filter – on x axis) for large defect

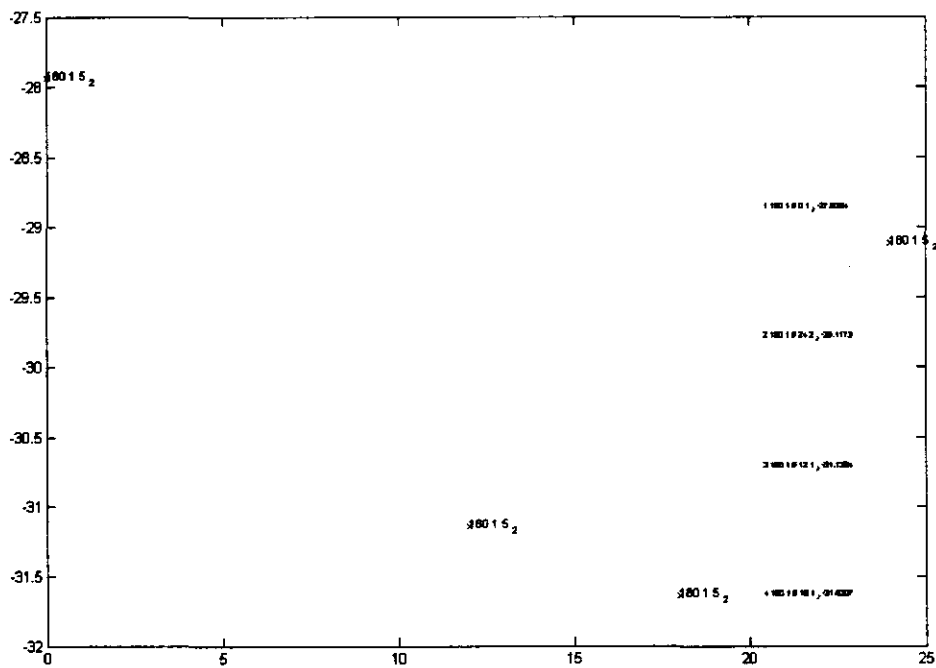


Figure 79: Feature LH_p_X (linear cepstrum – after highpass filter – on x axis) for large defect

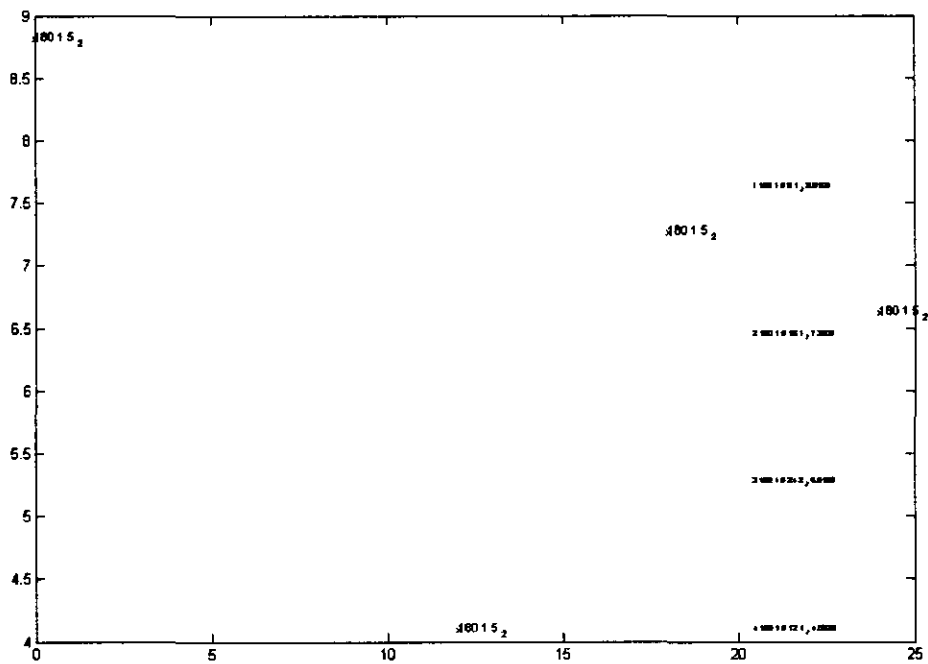


Figure 80: Feature $FFT4\_BPFO\_X$ (magnitude spectrum value of the square of the signal squared at ball pass outer race defect frequency on x axis) for large defect

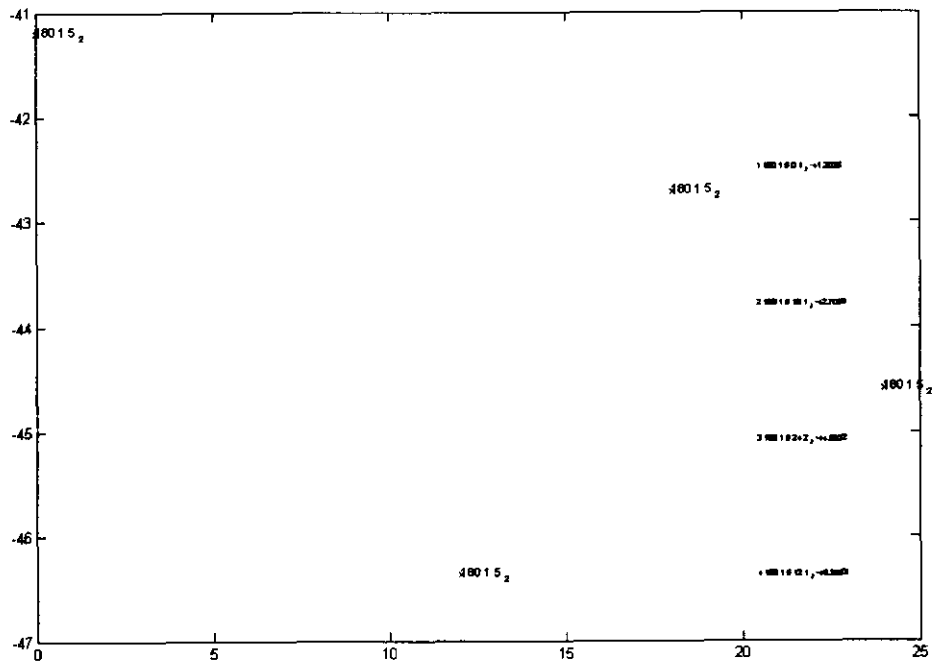


Figure 81: Feature *FFT8_BPFO_X* (magnitude spectrum value of the signal to the 8th power at ball pass outer race defect frequency on x axis) for large defect

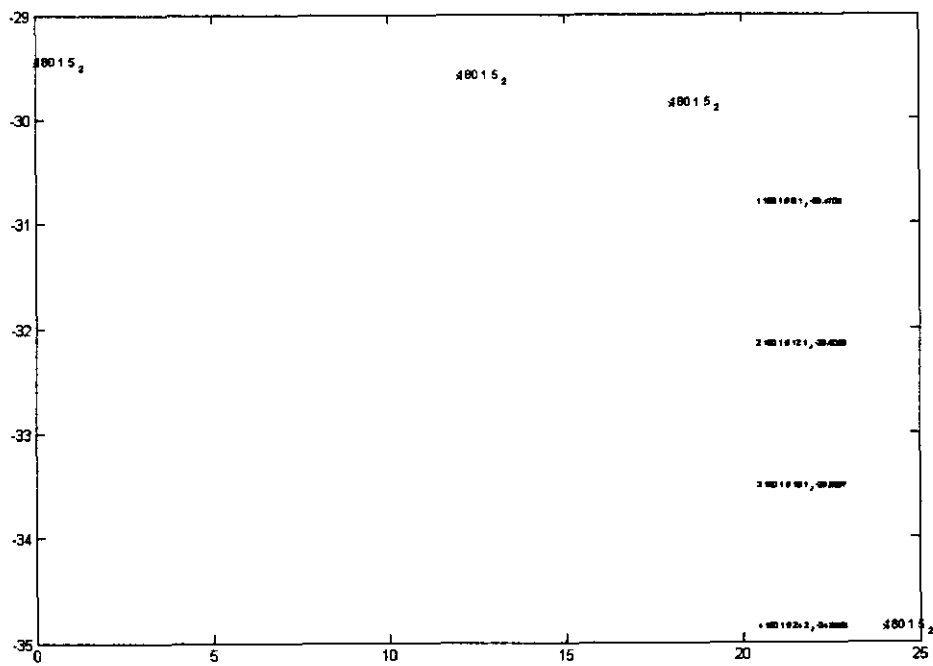


Figure 82: Feature *SA_BPFO2_0_X* (square root of the SOS of synchronous average of the signal at ball pass outer race defect frequency on x axis) for large defect

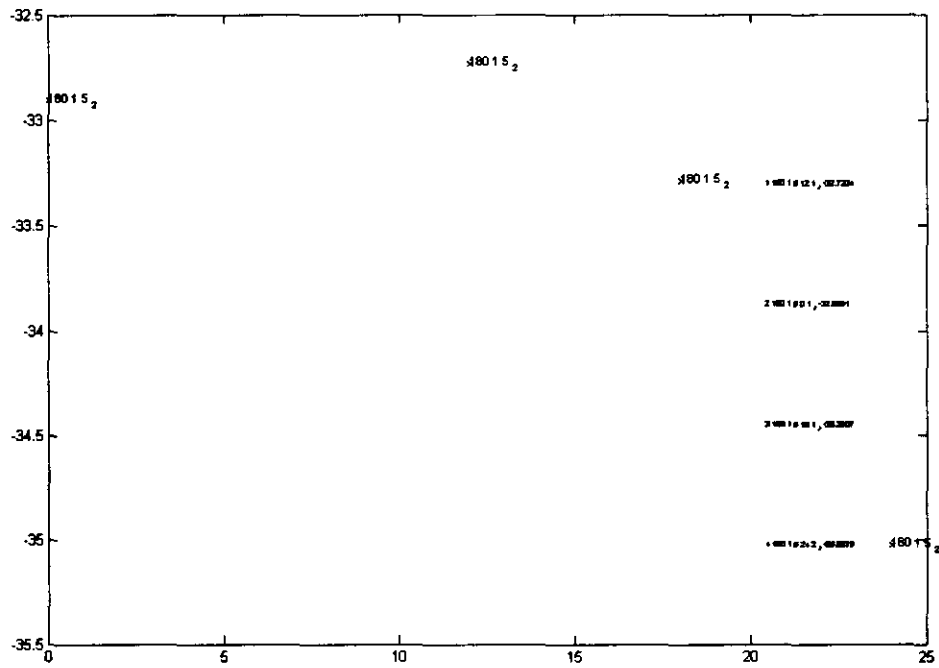


Figure 83: Feature *SA_BPFO2_p_X* (maximum value of synchronous average of the signal at ball pass outer race defect frequency on x axis) for large defect

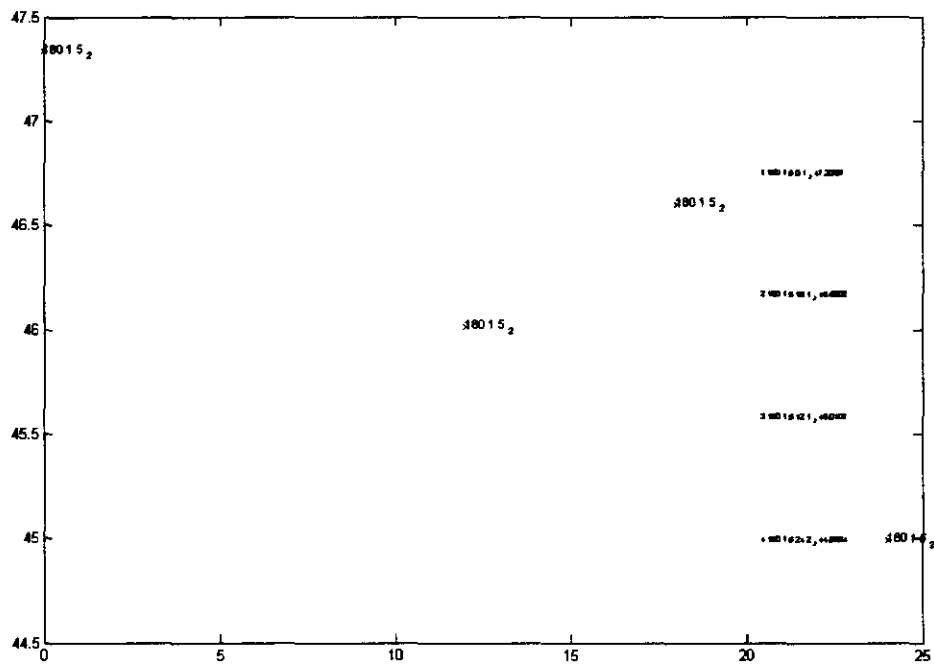


Figure 84: Feature *BPFO2_10_Eng_X* (square root of the SOS of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect

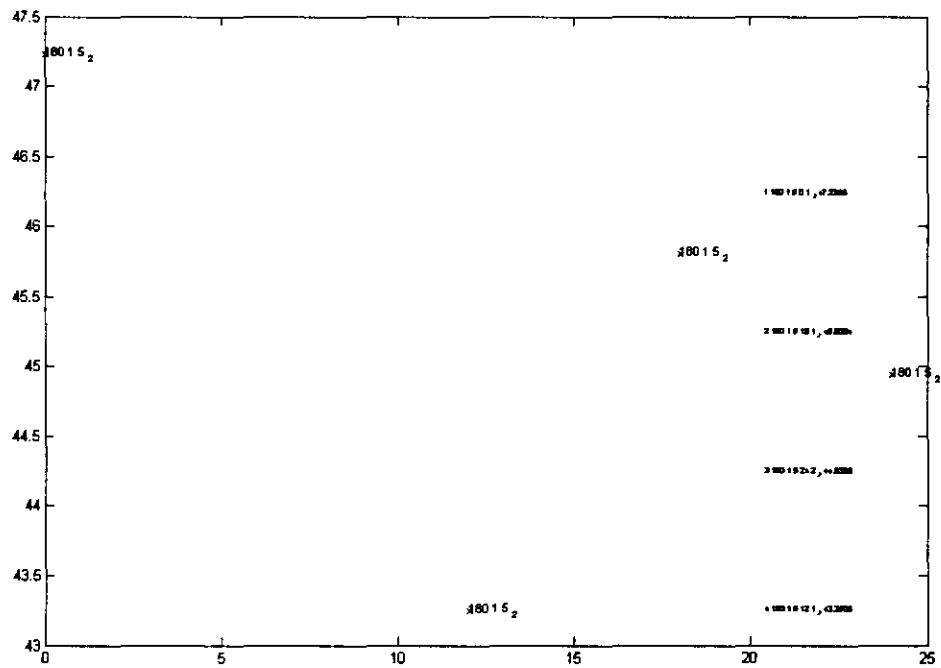


Figure 85: Feature *BPFO2_10_Max_X* (maximum value of first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect

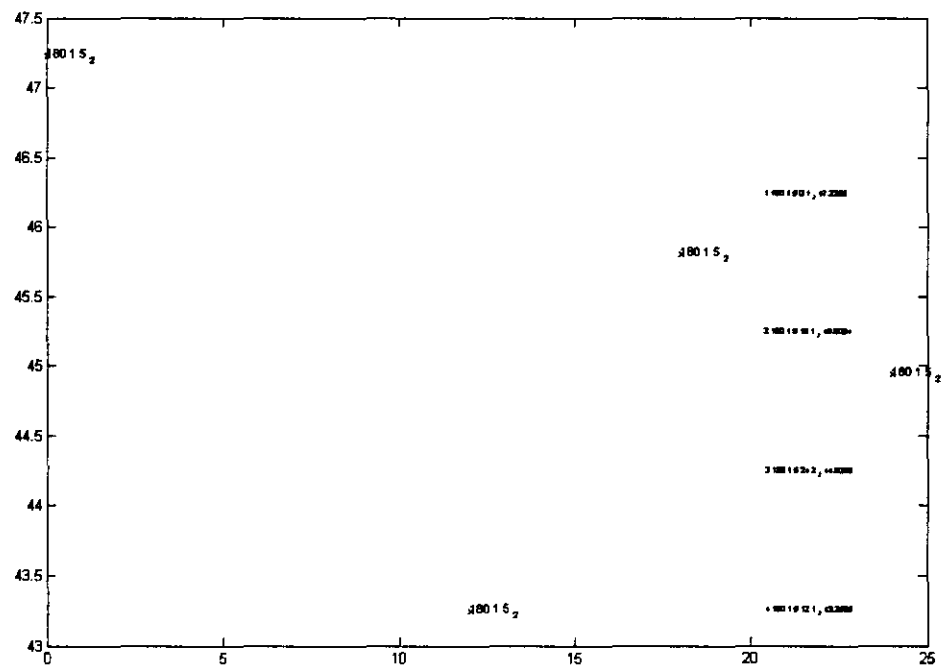


Figure 86: Feature *BPFO2_2_Eng_X* (square root of the SOS of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect

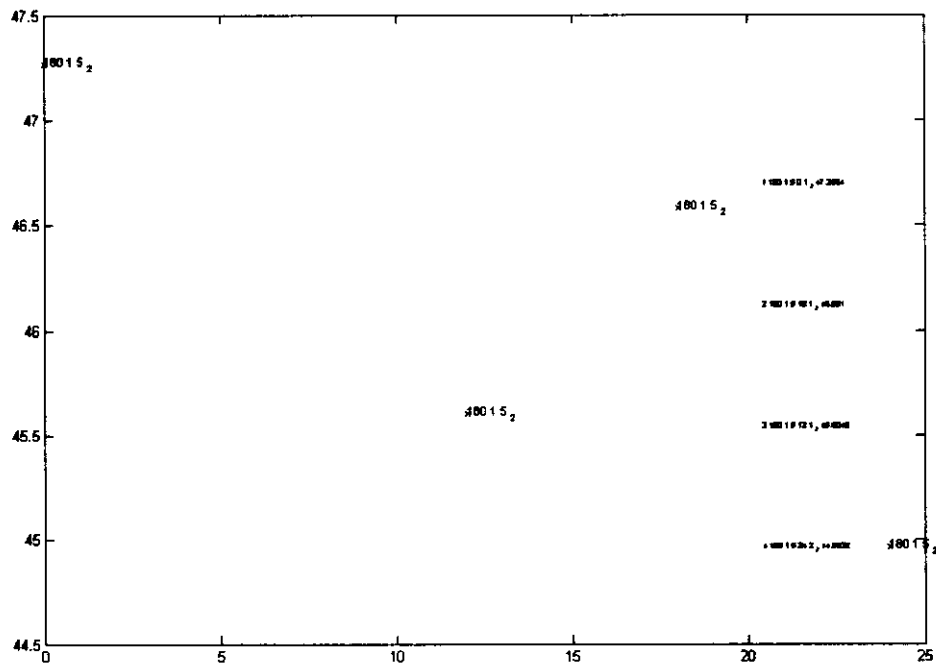


Figure 87: Feature *BPFO2_2_Max_X* (maximum value of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect

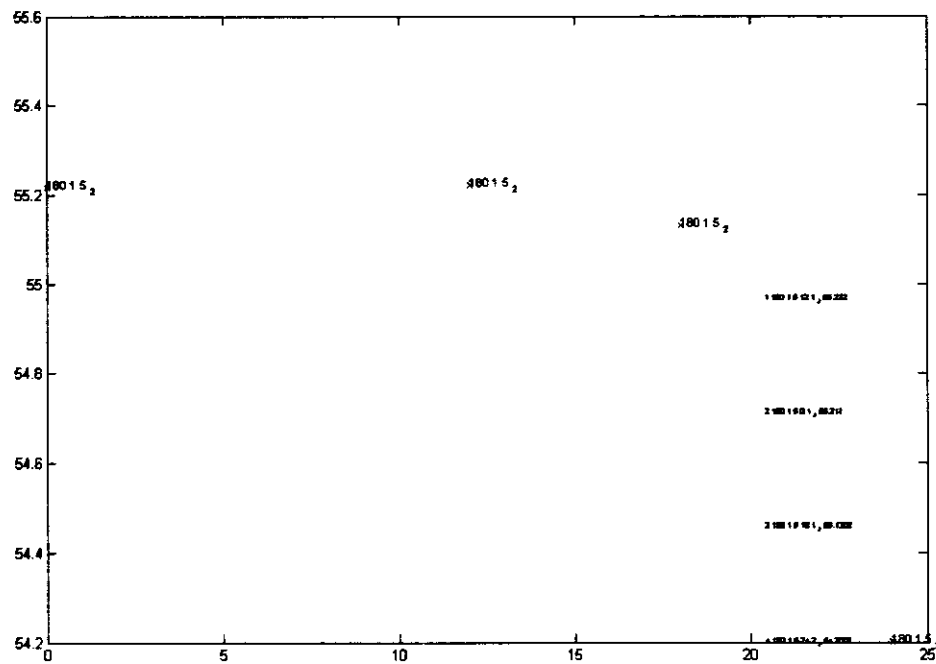


Figure 88: Feature *HFD_5_10_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 5–10kHz – on x axis) for large defect

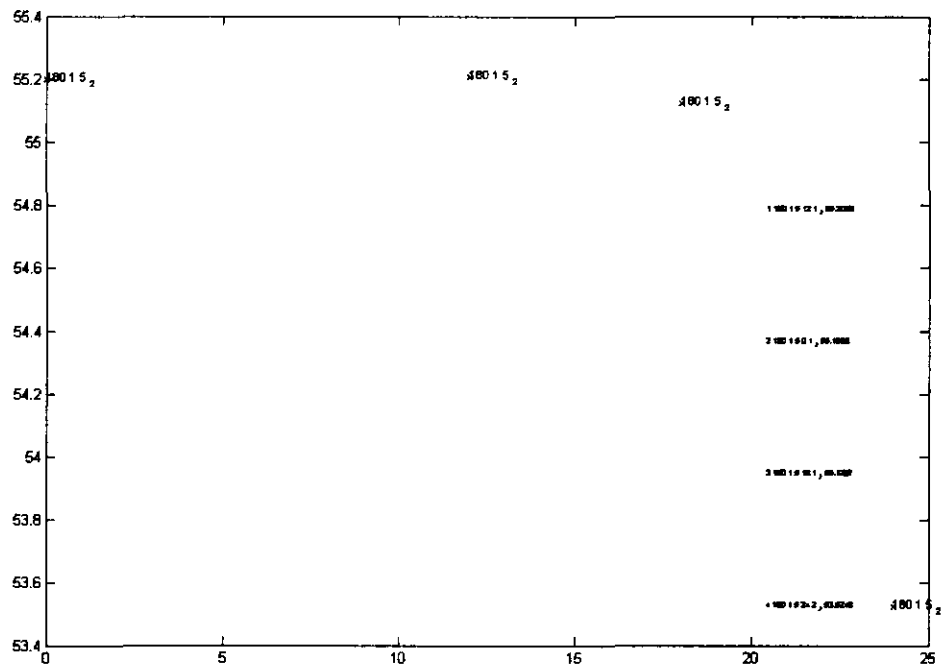


Figure 89: Feature $HFD\_2\_5\_kHz\_X$ (square root of SOS of HFD components in magnitude spectrum – from 2–5kHz – on x axis) for large defect

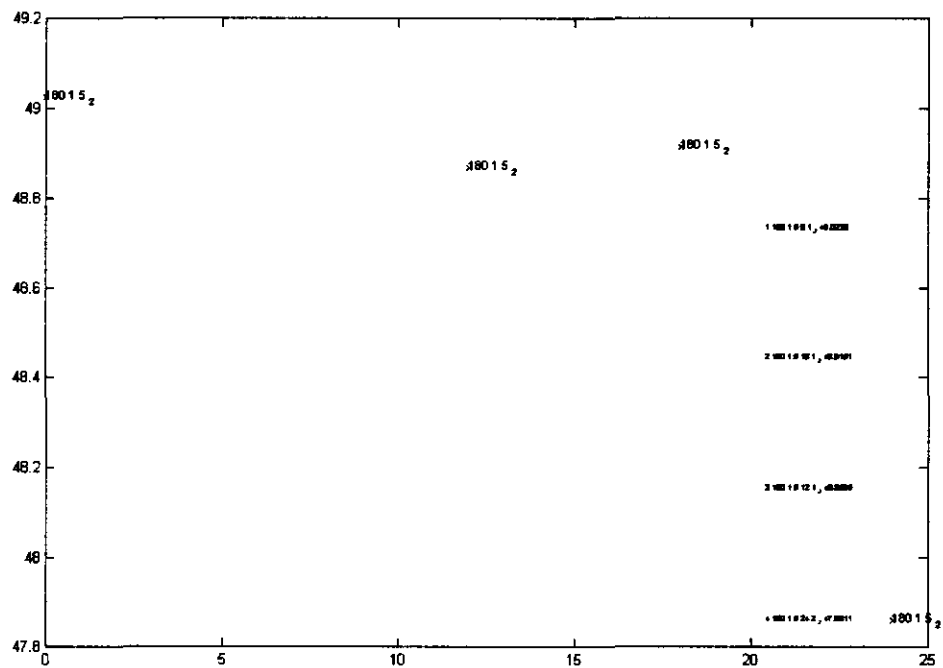


Figure 90: Feature $HFD\_5\_6\_kHz\_X$ (square root of SOS of HFD components in magnitude spectrum – from 5–6kHz – on x axis) for large defect

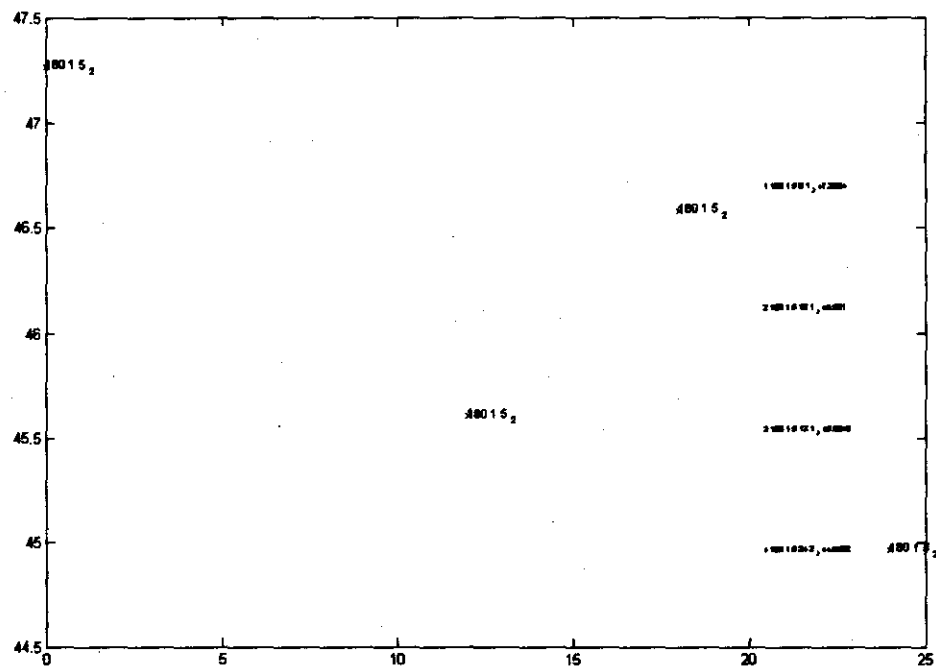


Figure 87: Feature *BPF02_2_Max_X* (maximum value of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect

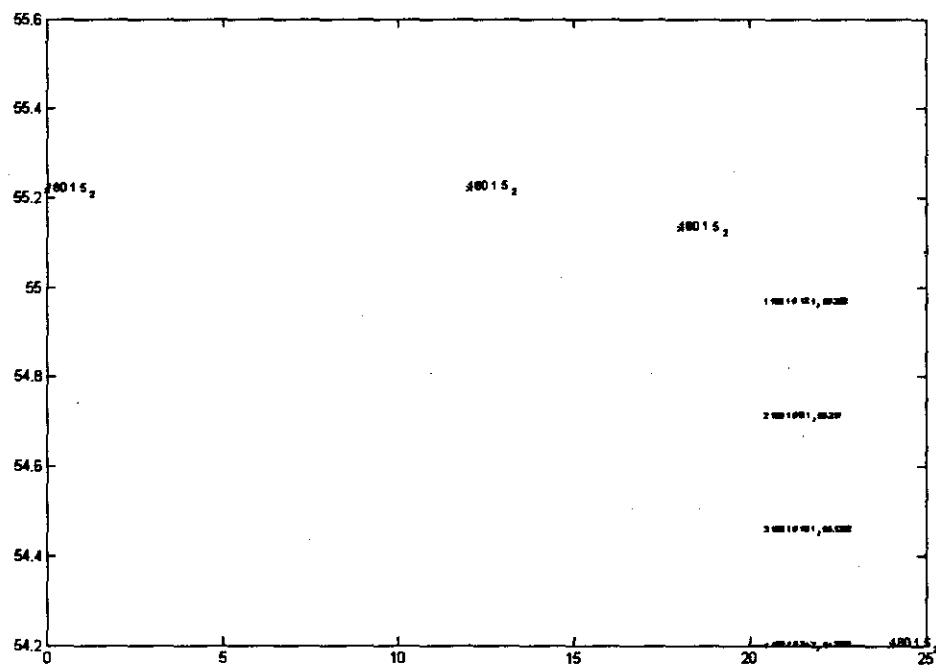


Figure 88: Feature *HFD_5_10_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 5–10kHz – on x axis) for large defect

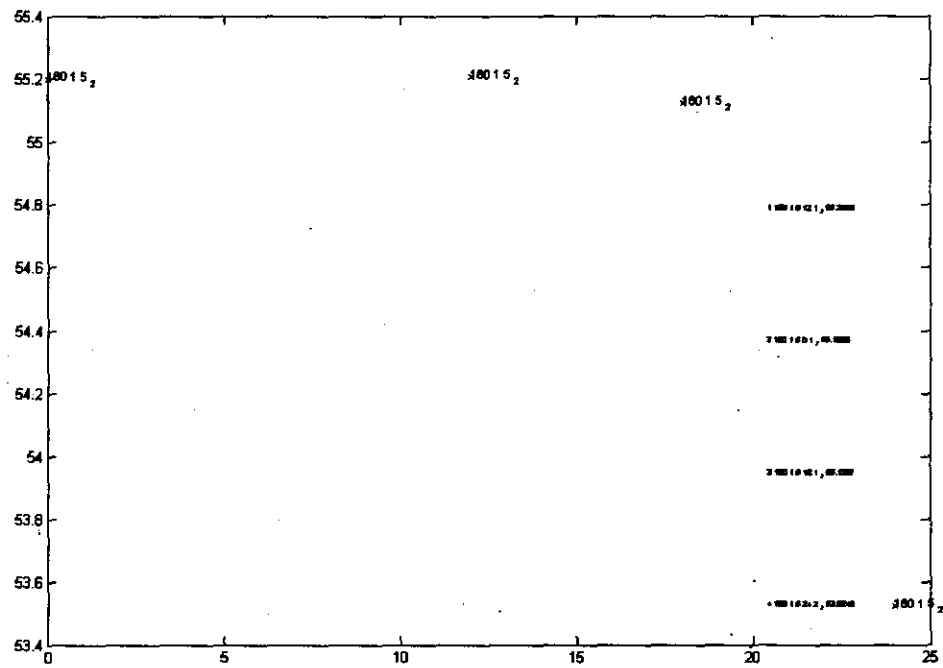


Figure 89: Feature $HFD\_2\_5\_kHz\_X$ (square root of SOS of HFD components in magnitude spectrum – from 2–5kHz – on x axis) for large defect

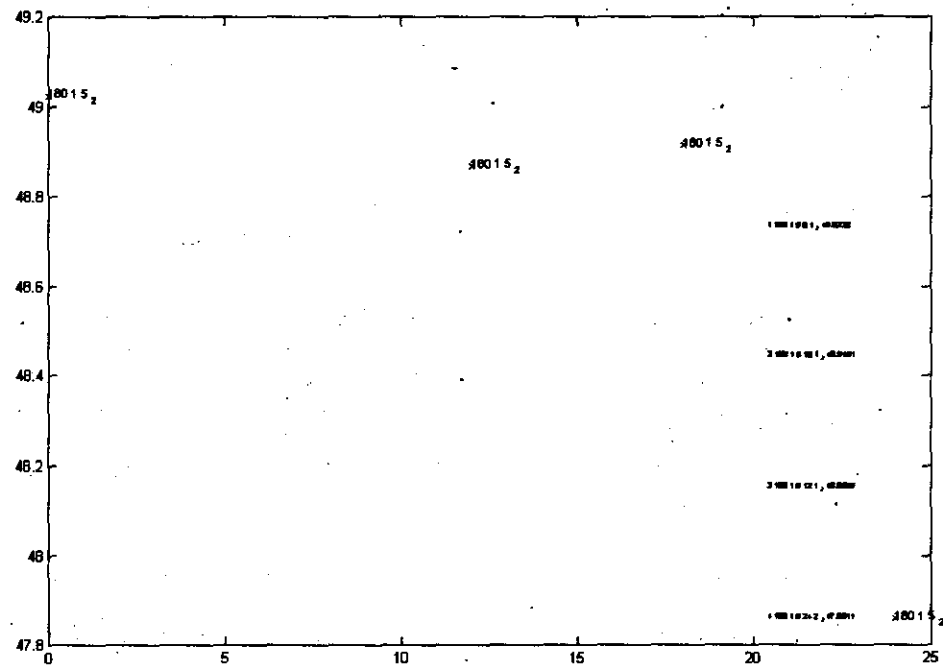


Figure 90: Feature $HFD\_5\_6\_kHz\_X$ (square root of SOS of HFD components in magnitude spectrum – from 5–6kHz – on x axis) for large defect

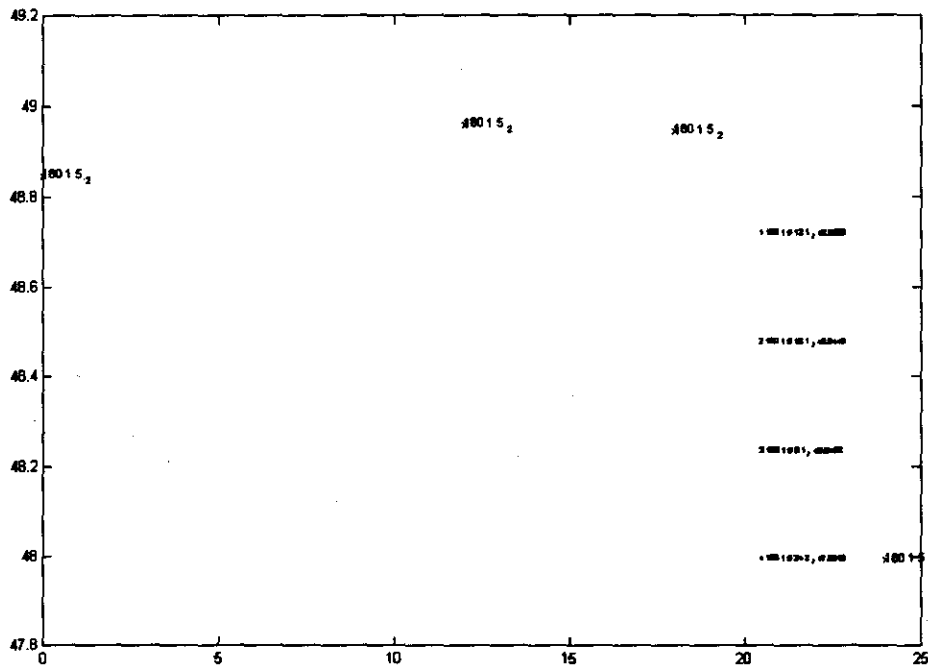


Figure 91: Feature $HFD\_6\_7\_kHz\_X$ (square root of SOS of HFD components in magnitude spectrum – from 6–7kHz – on x axis) for large defect

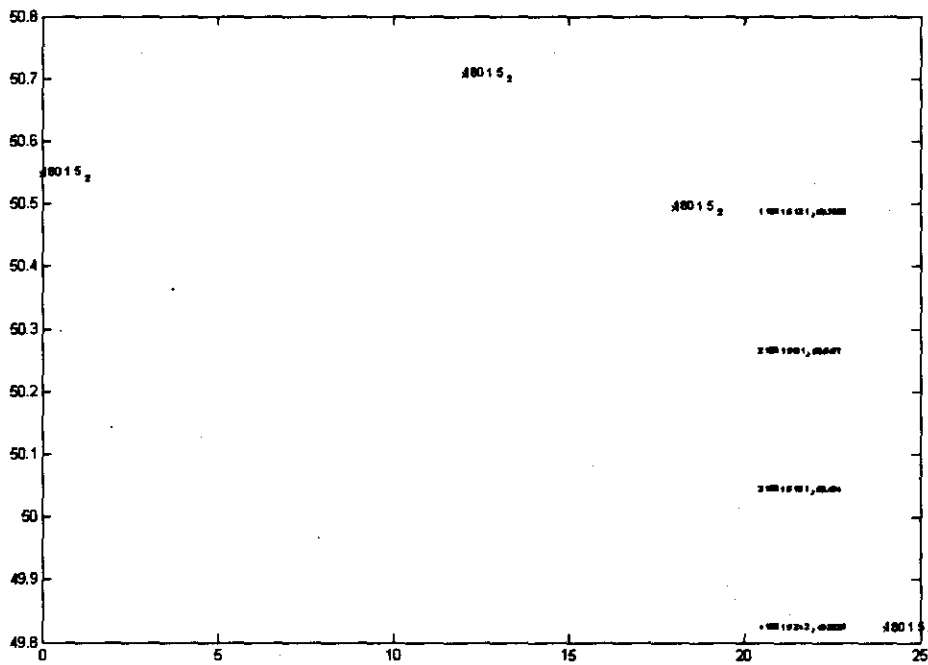


Figure 92: Feature $HFD\_8\_10\_kHz\_X$ (square root of SOS of HFD components in magnitude spectrum – from 8–10kHz – on x axis) for large defect

11.2.2 Classifier

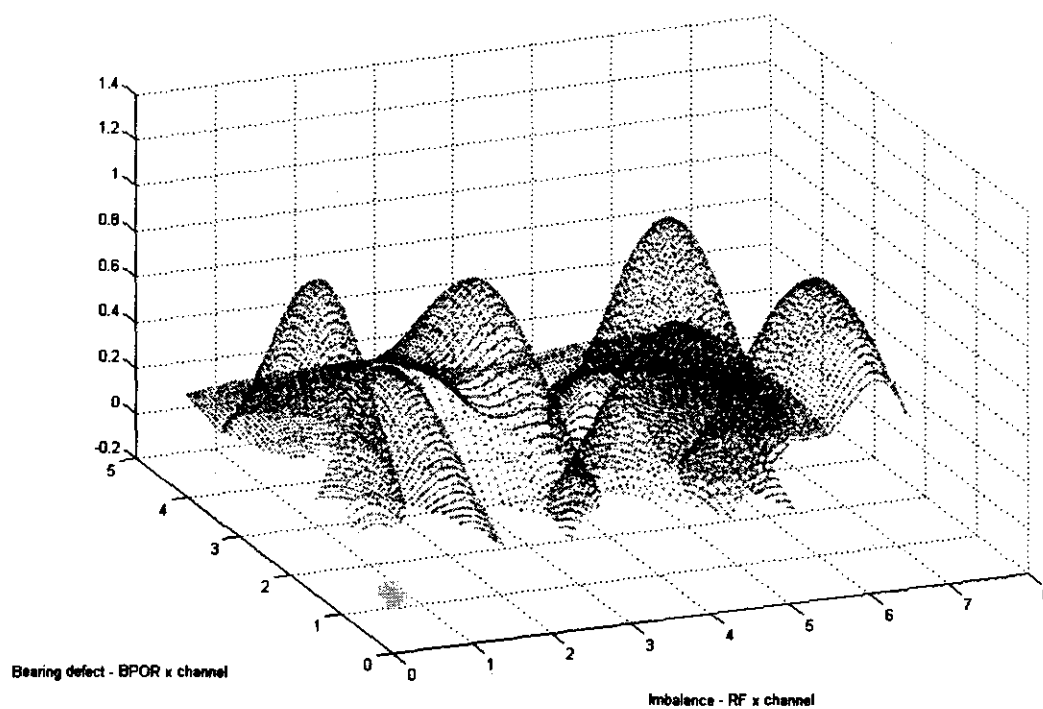


Figure 93: Outputs of RBF for 0g normal training set data – Vib demo data set 2

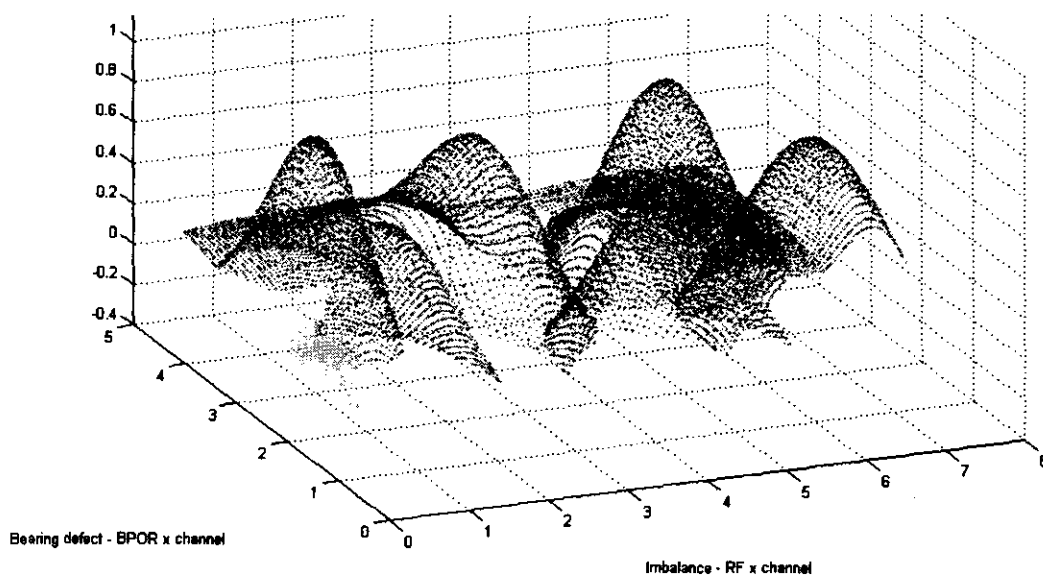


Figure 94: Outputs of RBF for 0g bearing defect training set data – Vib demo data set 2

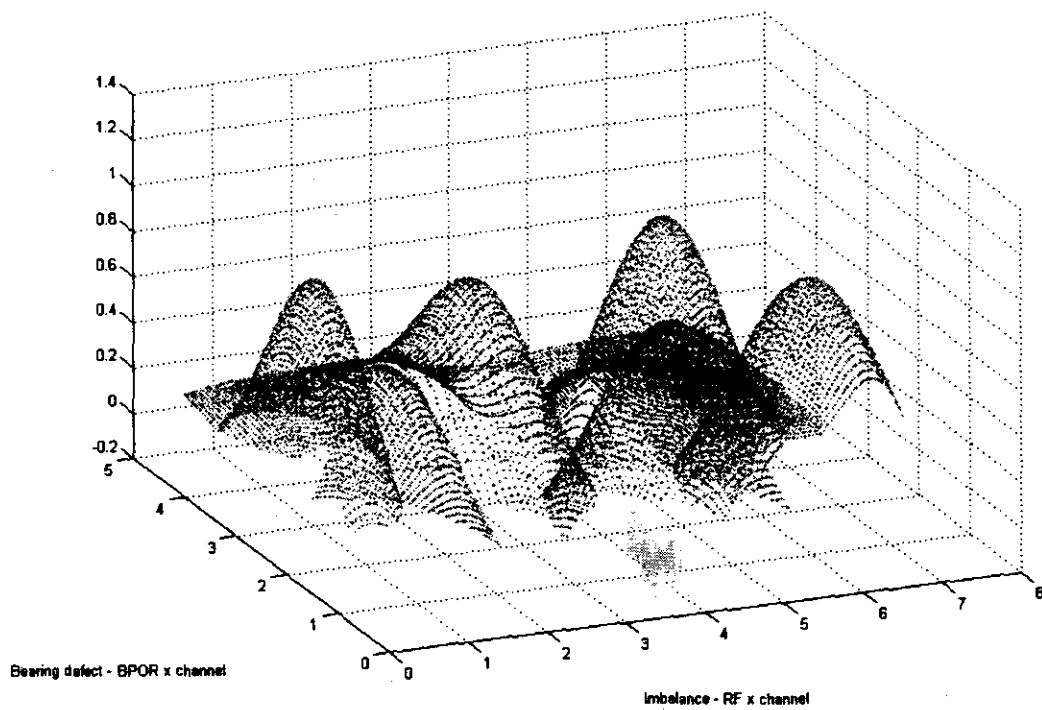


Figure 95: Outputs of RBF for 12g normal training set data – Vib demo data set 2

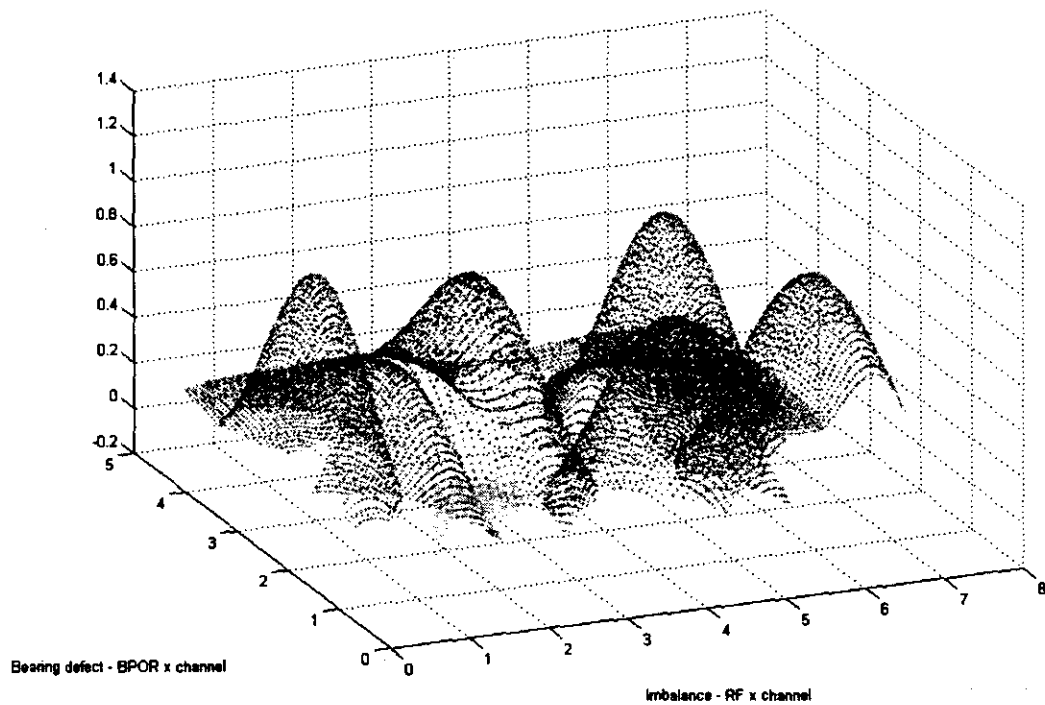


Figure 96: Outputs of RBF for 12g bearing defect training set data – Vib demo data set 2

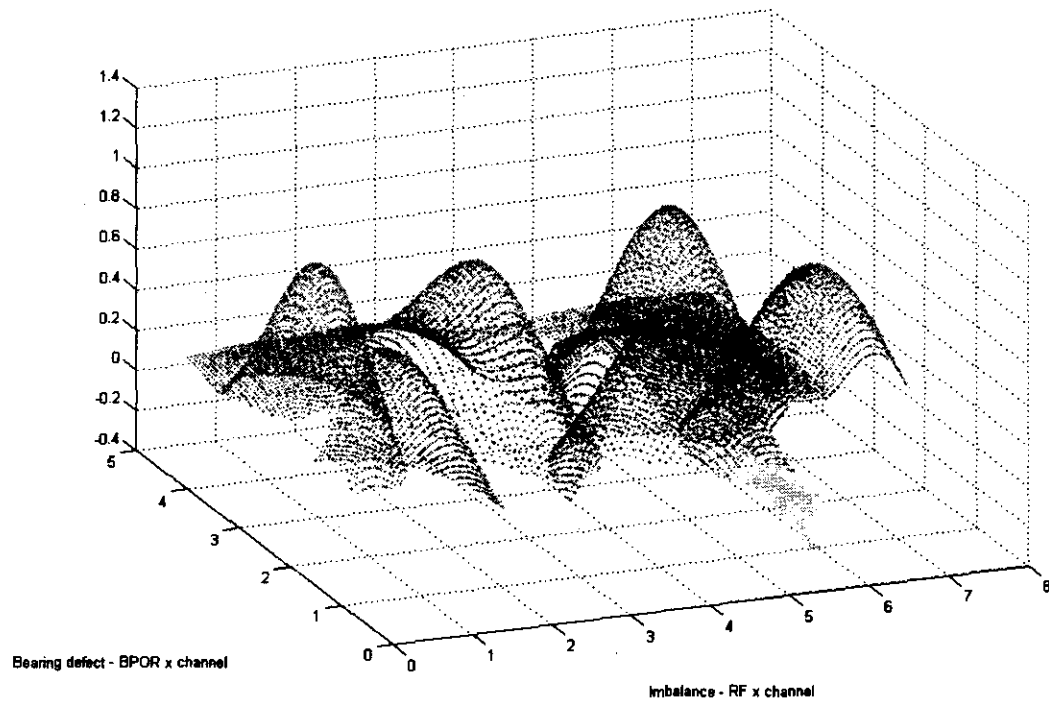


Figure 97: Outputs of RBF for 24g normal training set data – Vib demo data set 2

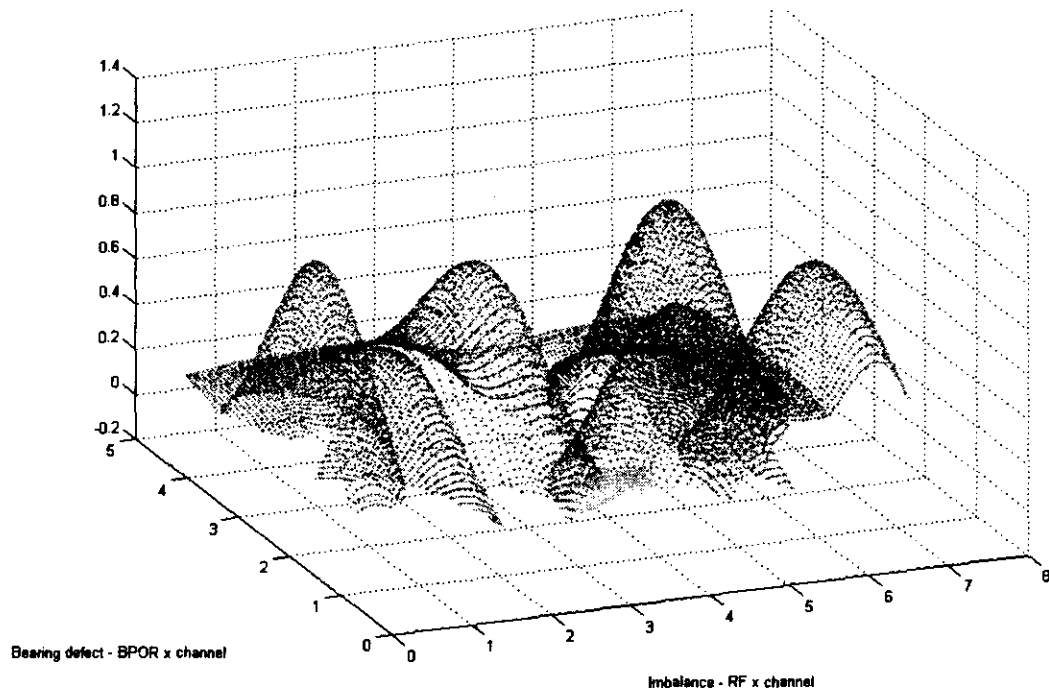


Figure 98: Outputs of RBF for 24g bearing defect training set data – Vib demo data set 2

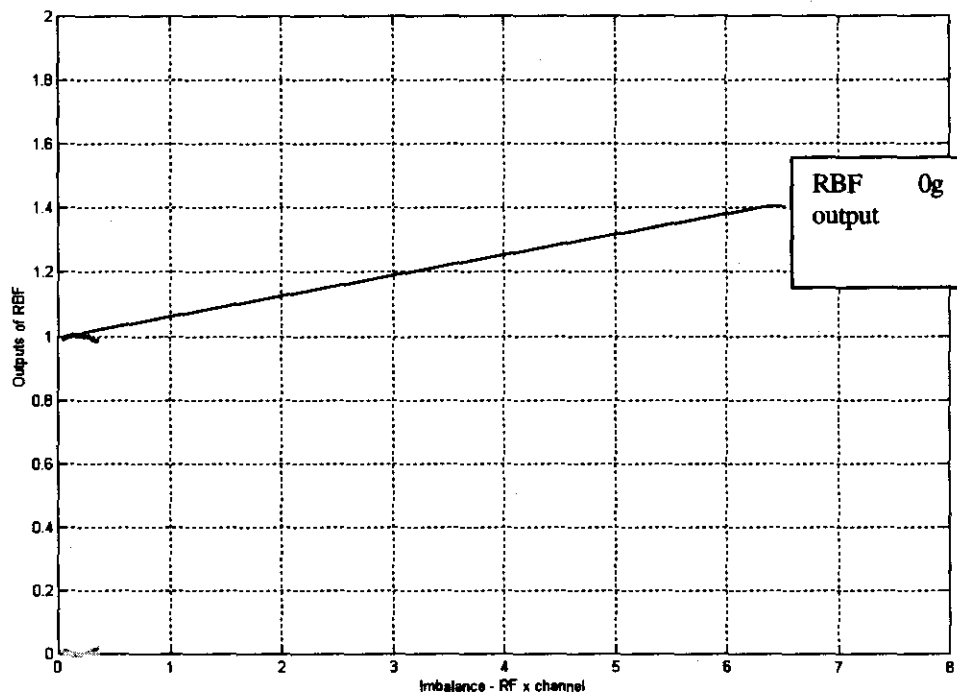


Figure 99: Outputs of RBF for 0g training set data (measured on the normal bearing) – Vib demo data set 2

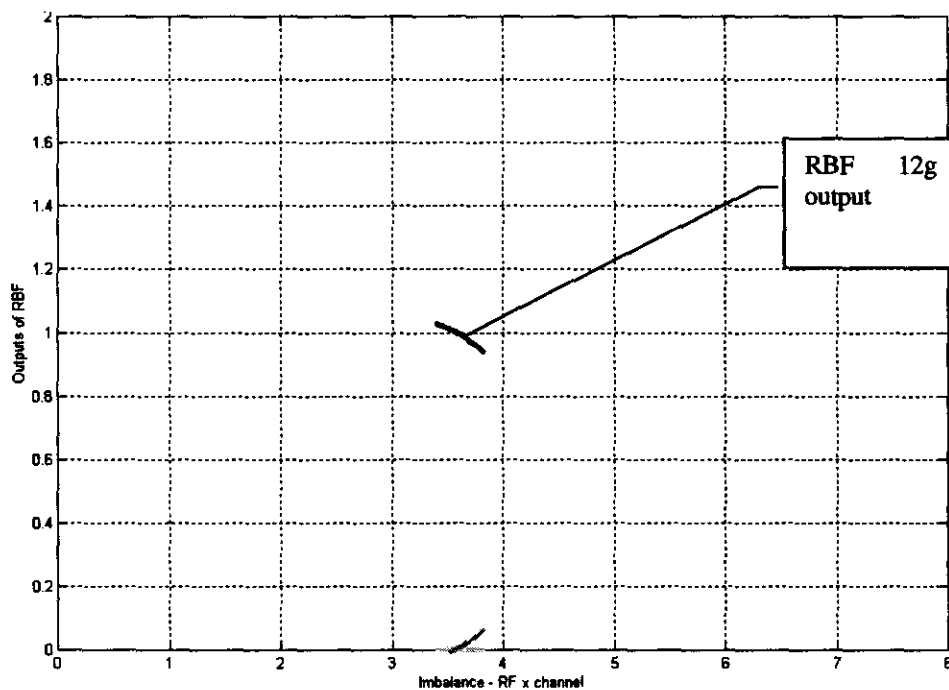


Figure 100: Outputs of RBF for 12g training set data (measured on the normal bearing) – Vib demo data set 2

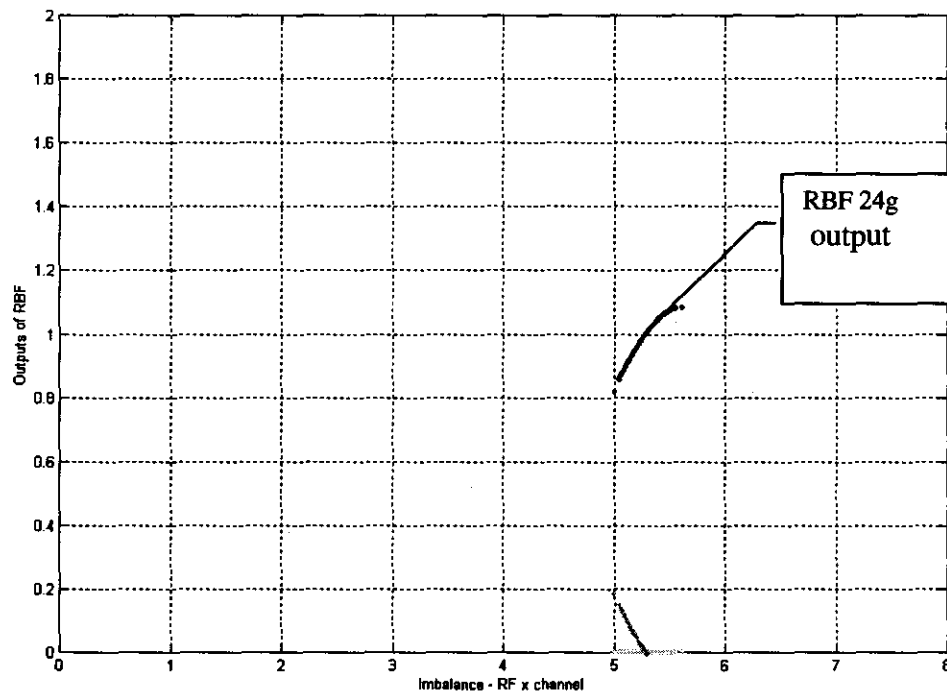


Figure 101: Outputs of RBF for 24g training set data (measured on the normal bearing) – Vib demo data set 2

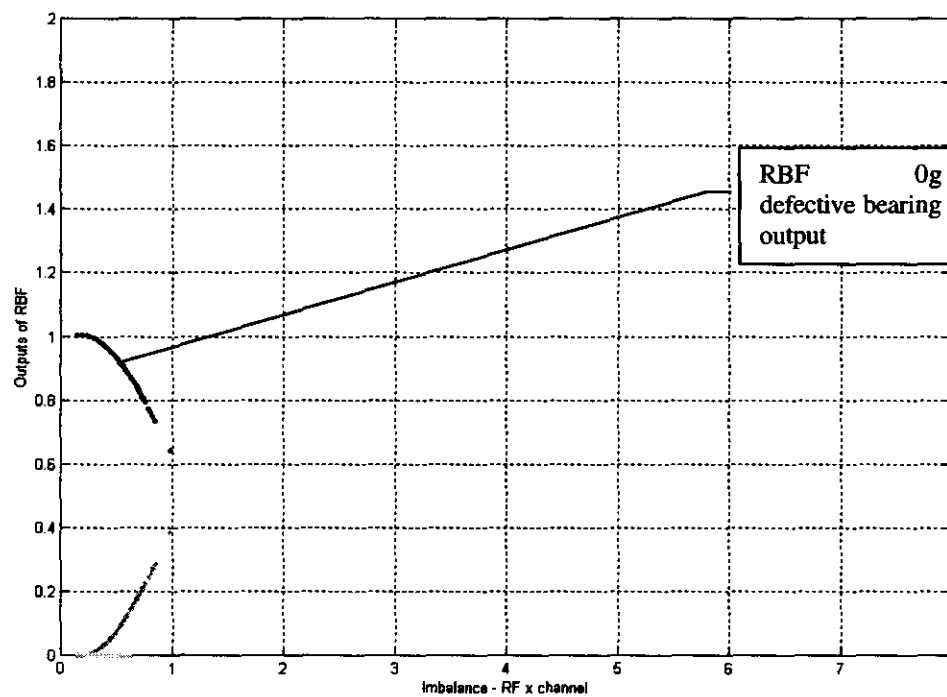


Figure 102: Outputs of RBF for 0g test set data (measured on the defective bearing) – Vib demo data set 2

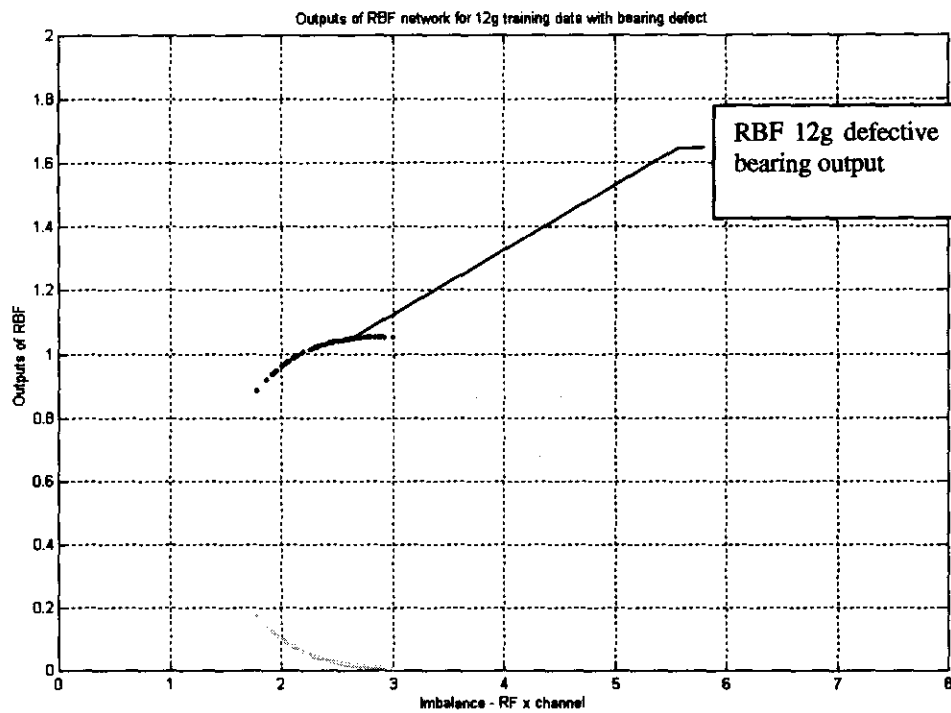


Figure 103: Outputs of RBF for 12g test set data (measured on the defective bearing) – Vib demo data set 2

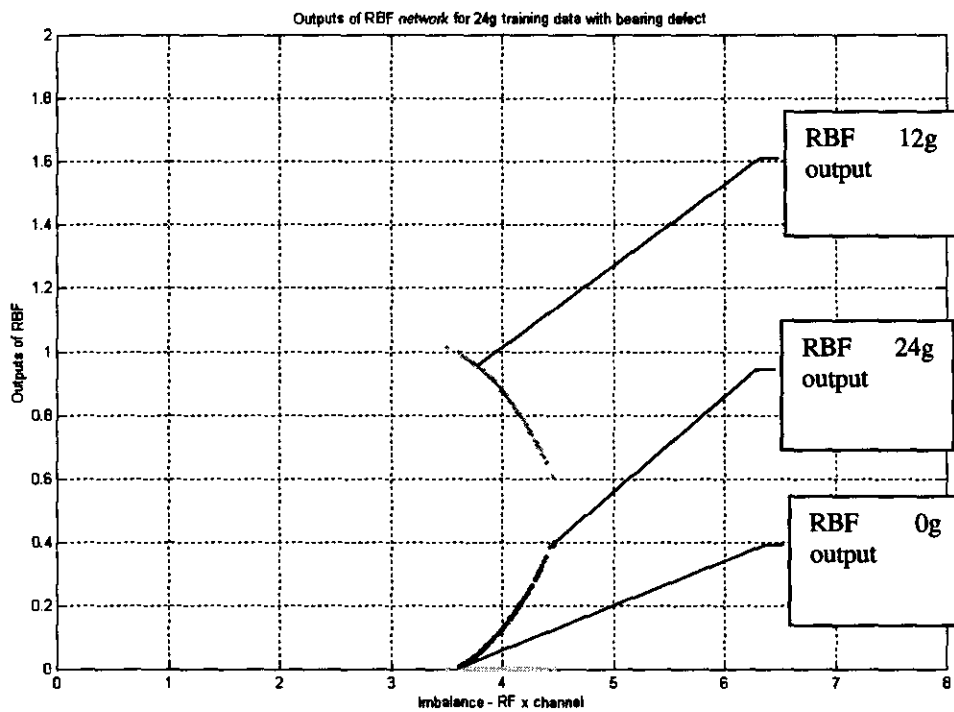


Figure 104: Outputs of RBF for 24g test set data (measured on the defective bearing) – Vib demo data set 2

11.3 Appendix C: Results on Vib demo data set 3 (figures)

11.3.1 Feature extraction

The following features are shown sequentially in the following figures on a log scale (the labels are the same as in appendix B). Note that some of the measurements of the small bearing defect at defect angles of 315°, 360°, 45°, 90°, 135° have been made using a sampling rate of 51 200Hz (20kHz bandwidth) instead of the usual 25 600Hz (10kHz bandwidth).

For each data point indicated by 'x' the defect angle (DA) is shown (180°, 225°, 270°, 315°, 360°, 45°, 90° and 135° for the Vib demo data set 3), the bearing number (1, 4, 5 for the Vib demo data set 3), the bearing defect severity (5, 2, 1 for the Vib demo data set 3 where 5 is the largest bearing defect corresponding to the same bearing used in the Vib demo data set 2) and the distance of the defect from 225° (refer to figure 43 on p. 214 for notational details). The dynamic load (unbalance mass in grams) is shown on the horizontal axis to demonstrate the influence of dynamic loading on bearing defects of decreasing severity and defect angle.

11.3.1.1 Outer race defects (small, medium and large together)¹³¹

The following figures show the effect of applied dynamic load for the following extracted features (the order below is from left to right, then down by row):

<i>Fr_X</i>	<i>BPFO_X</i>	<i>BPFO2_X</i>	<i>Fr_S_X</i>
<i>BPFO2_S_X</i>	<i>BPF12_S_X</i>	<i>BPF12_X</i>	<i>KH_X</i>
<i>CH_X</i>	<i>LH_X</i>	<i>CHp_X</i>	<i>KHp_X</i>
<i>LHp_X</i>	<i>FFT4_BPFO_X</i>	<i>FFT8_BPFO_X</i>	<i>SA_BPFO2_0_X</i>
<i>SA_BPFO2_0_p_X</i>	<i>BPFO2_10_Eng_X</i>	<i>BPFO2_10_Max_X</i>	<i>BPFO2_2_Eng_X</i>
<i>BPFO2_2_Max_X</i>	<i>HFD_5_10kHz_X</i>	<i>HFD_2_5kHz_X</i>	<i>HFD_5_6kHz_X</i>
<i>HFD_6_7kHz_X</i>	<i>HFD_8_10kHz_X</i>	<i>FFT2_MEAN_X</i>	

¹³¹ The measurements at defect angle of 180°, including a large number of handheld measurements, are not shown to save valuable screen space.

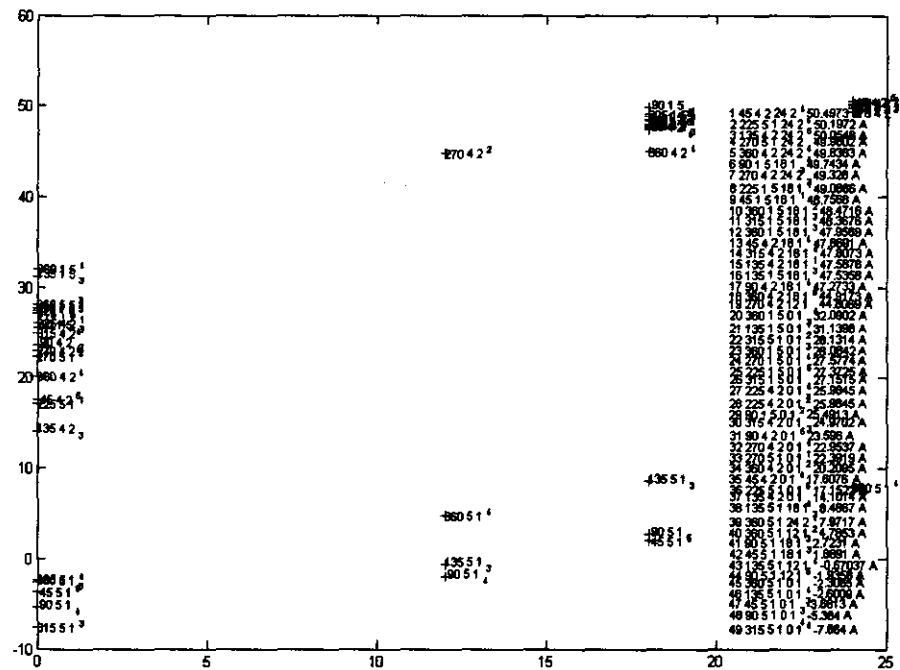


Figure 105: Feature Fr_X (magnitude spectrum value at rotational frequency on x axis) for all defects

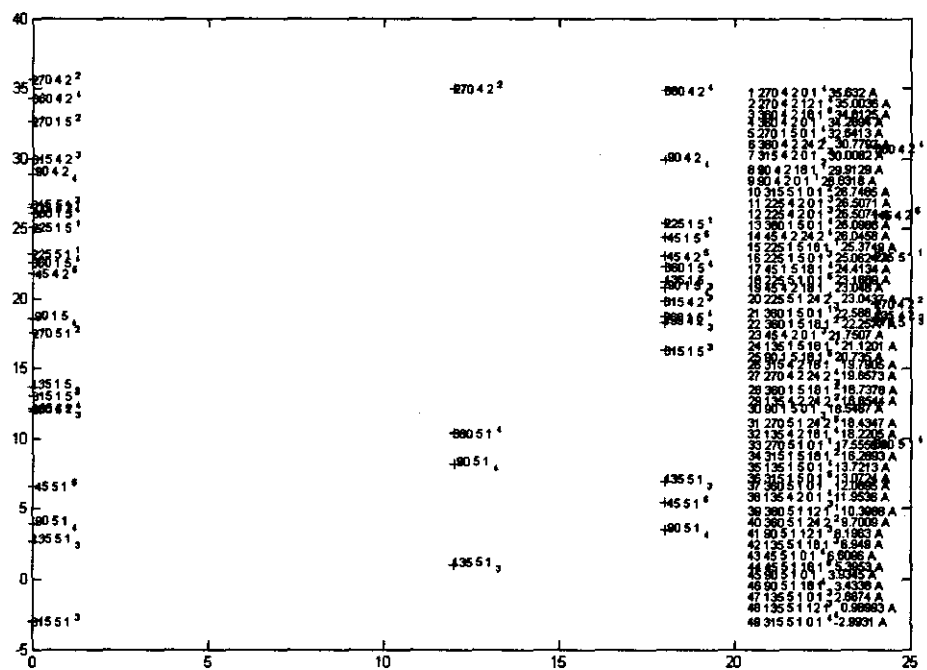
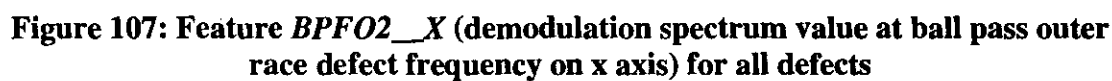


Figure 106: Feature $BPFO\_X$ (magnitude spectrum value at ball pass outer race defect frequency on x axis) for all defects



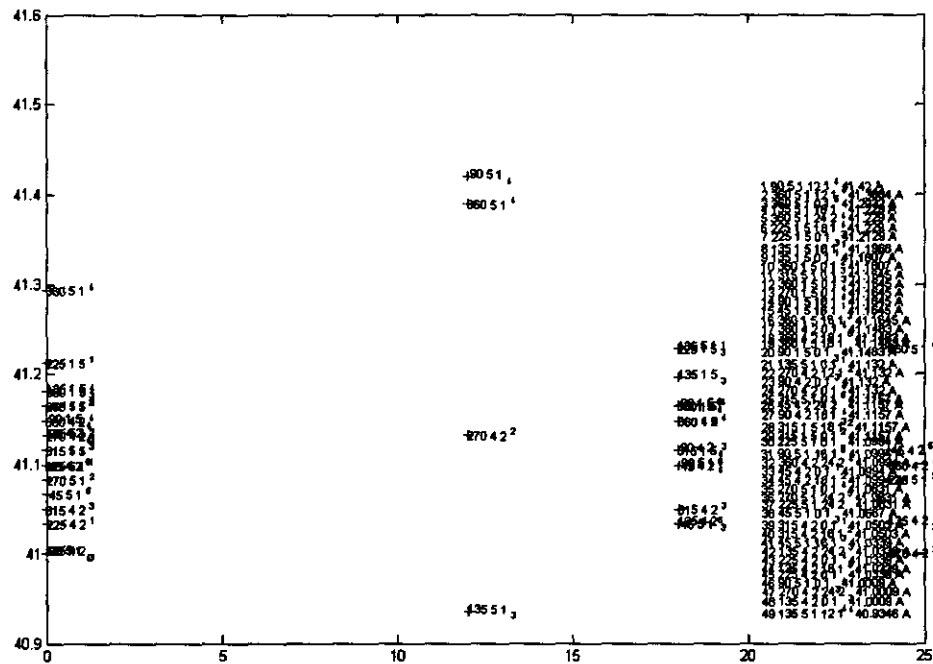


Figure 109: Feature *BPFO2_S_X* (ball pass outer race defect frequency on x axis)
for all defects

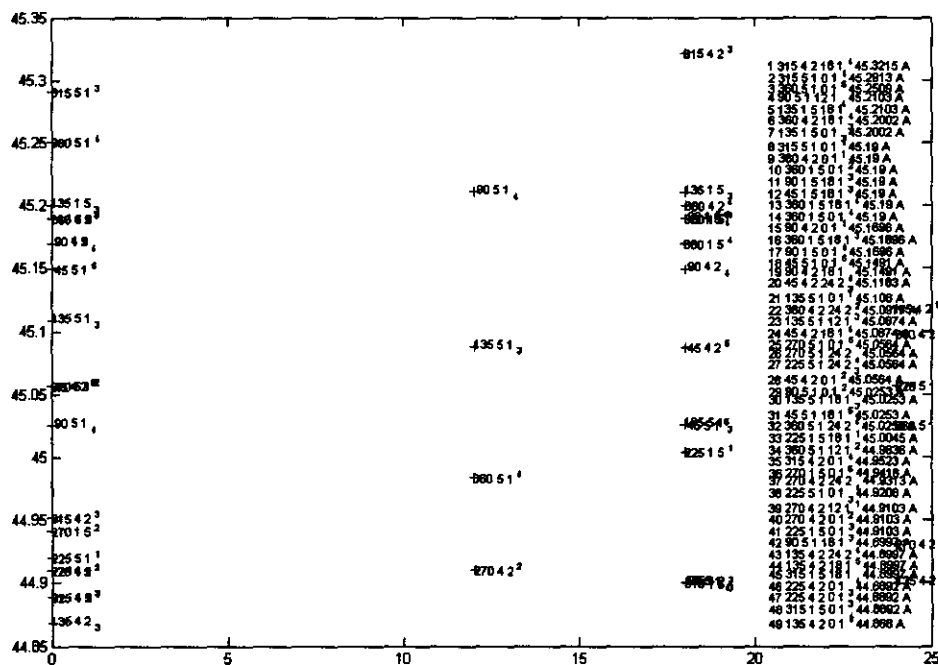


Figure 110: Feature *BPF12_S_X* (ball pass inner race defect frequency on x axis)
for all defects

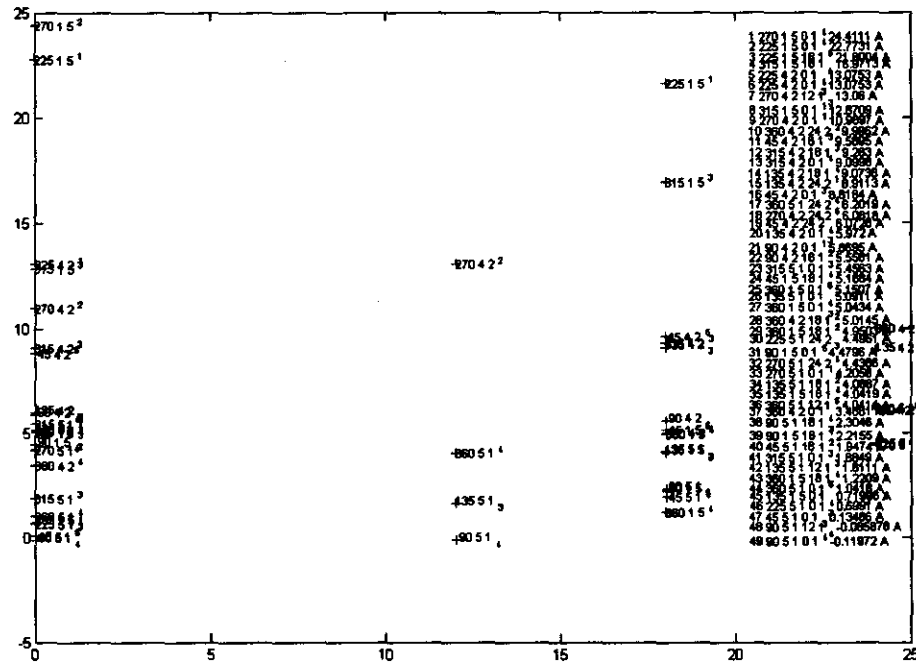


Figure 111: Feature *BPF12_X* (demodulation spectrum value at ball pass inner race defect frequency on x axis) for all defects

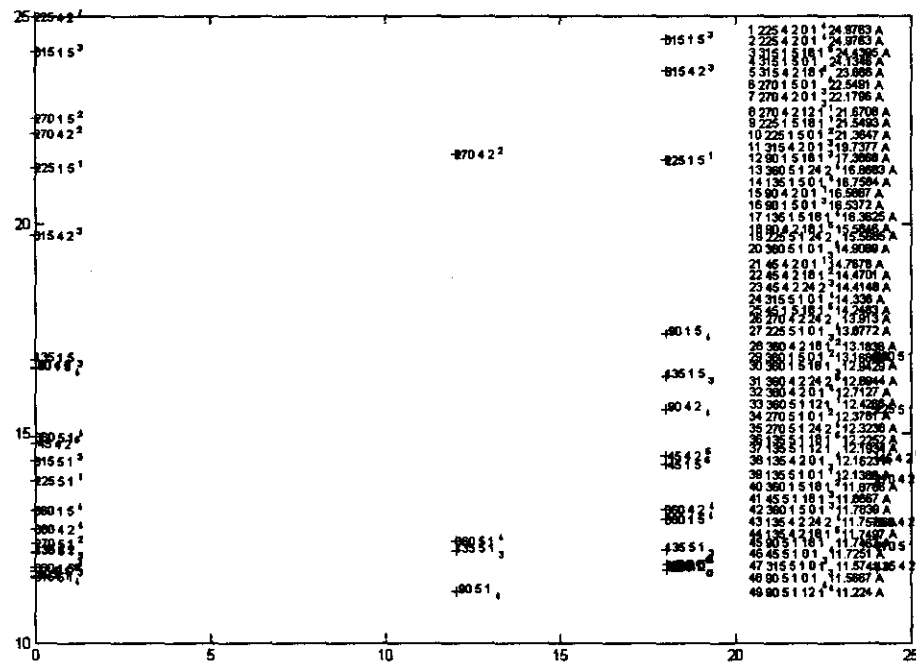


Figure 112: Feature *KH_X* (kurtosis on x axis) for all defects

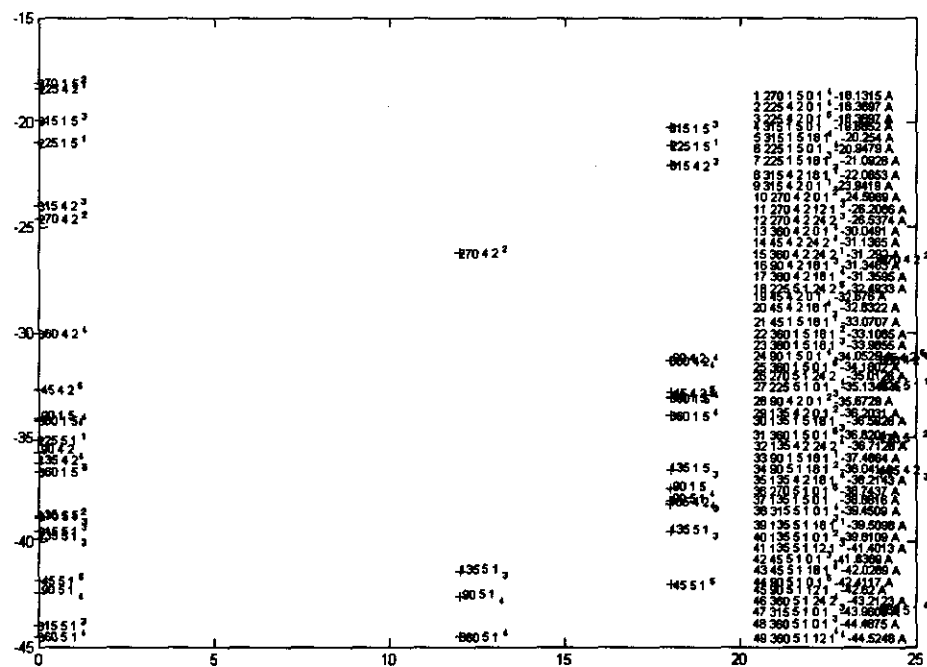


Figure 113: Feature CH_X (real cepstrum on x axis) for all defects

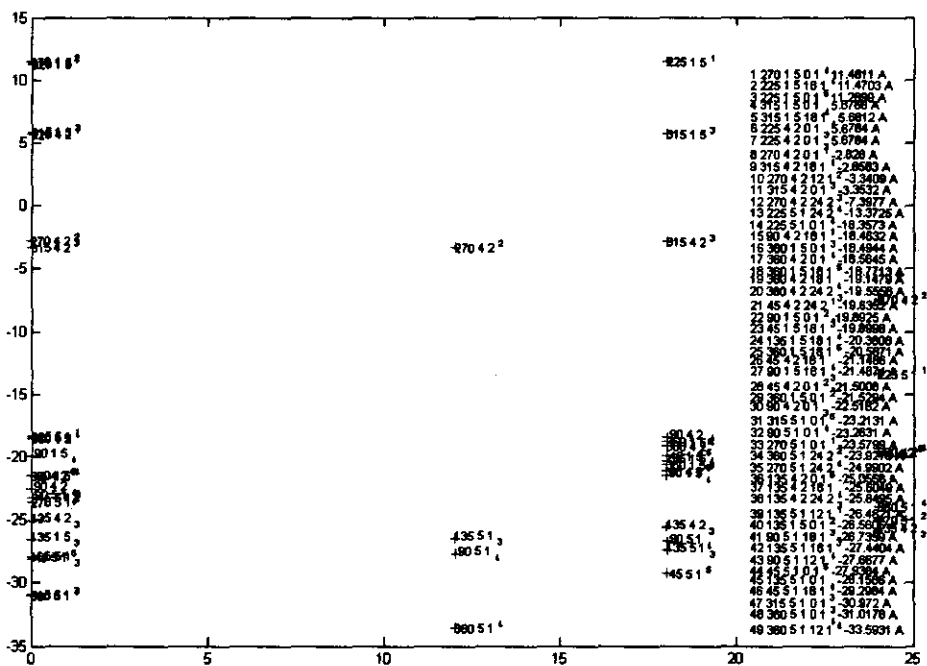


Figure 114: Feature LH_X (linear cepstrum on x axis) for all defects

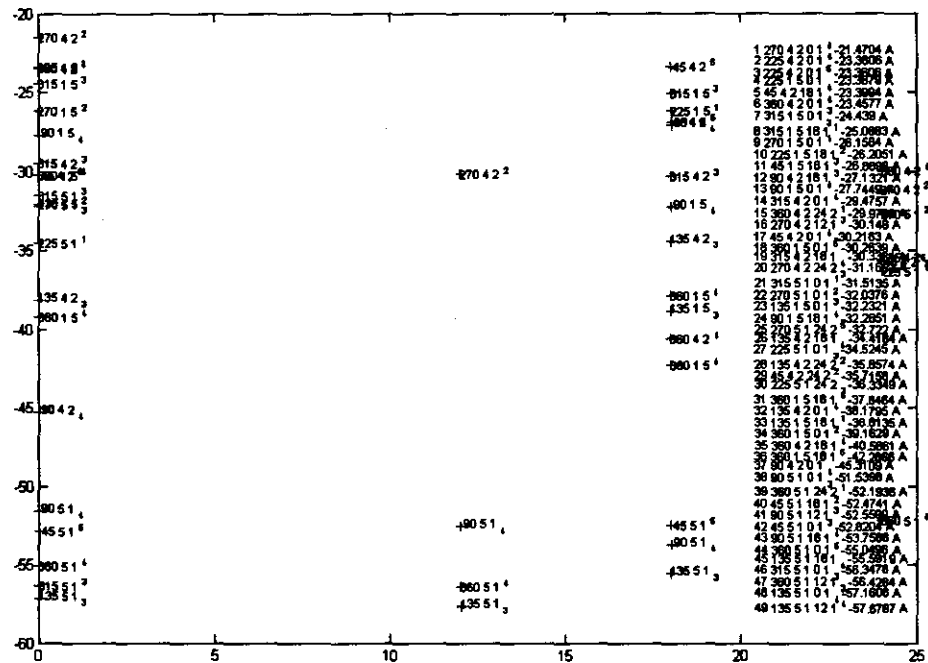


Figure 115: Feature CH_p_X (real cepstrum – after highpass filter – on x axis) for all defects

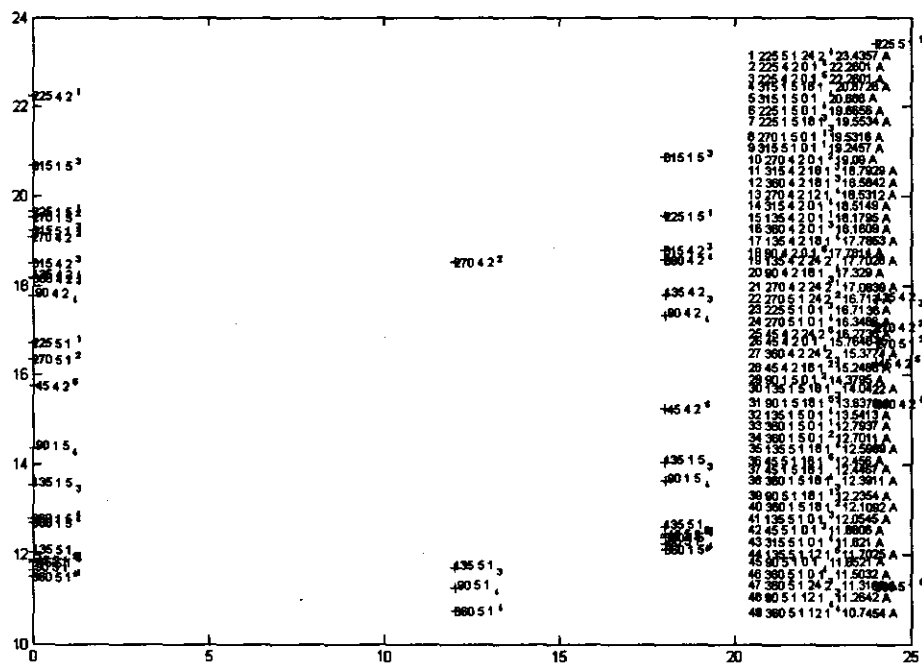


Figure 116: Feature KH_p_X (kurtosis – after highpass filter – on x axis) for all defects

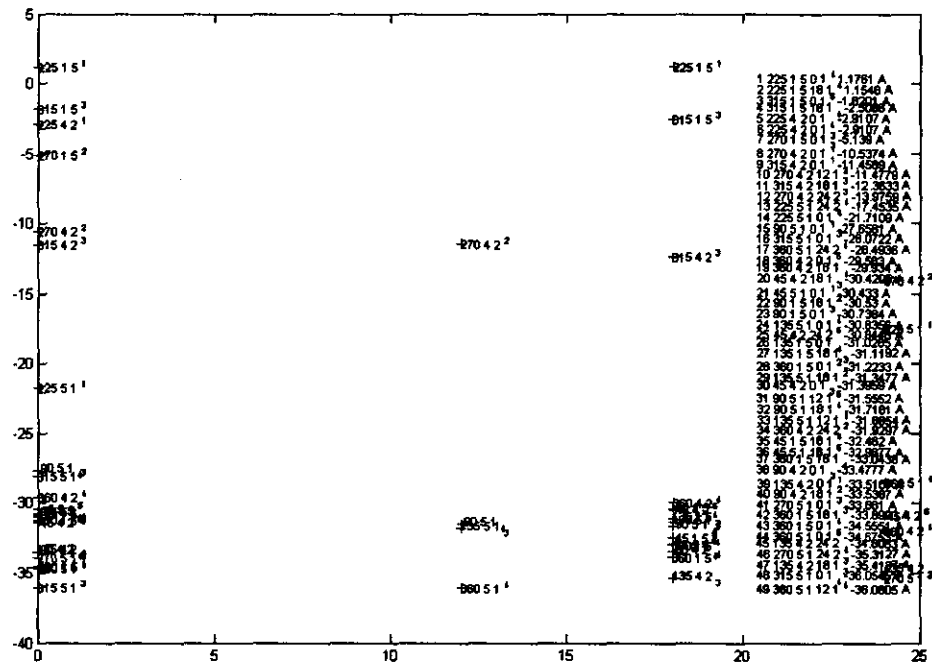


Figure 117: Feature LH_p_X (linear cepstrum – after highpass filter – on x axis) for all defects

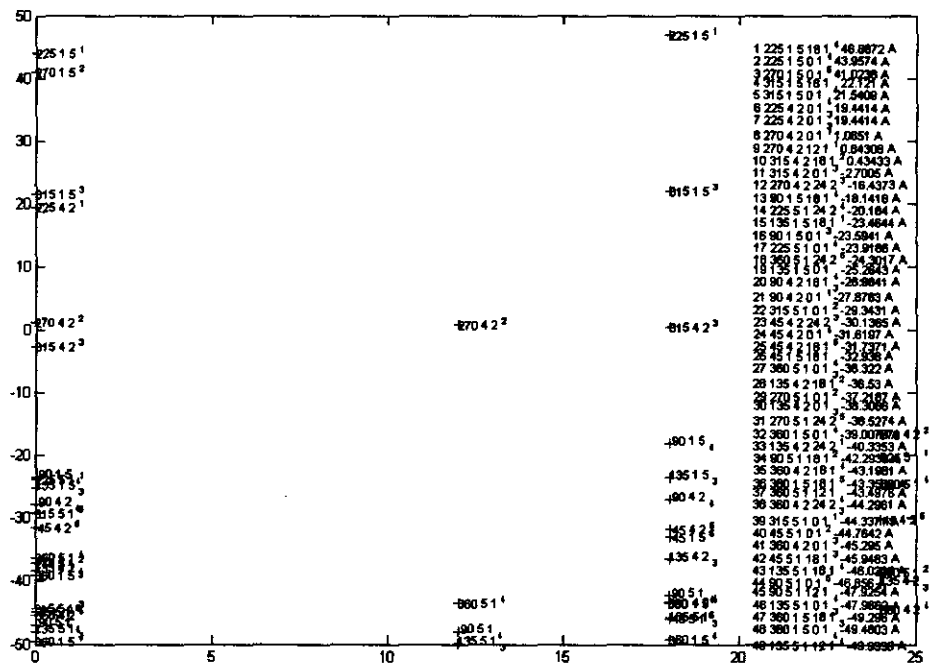


Figure 118: Feature $FFT4\_BPFO\_X$ (magnitude spectrum value of the square of the signal squared at ball pass outer race defect frequency on x axis) for all defects

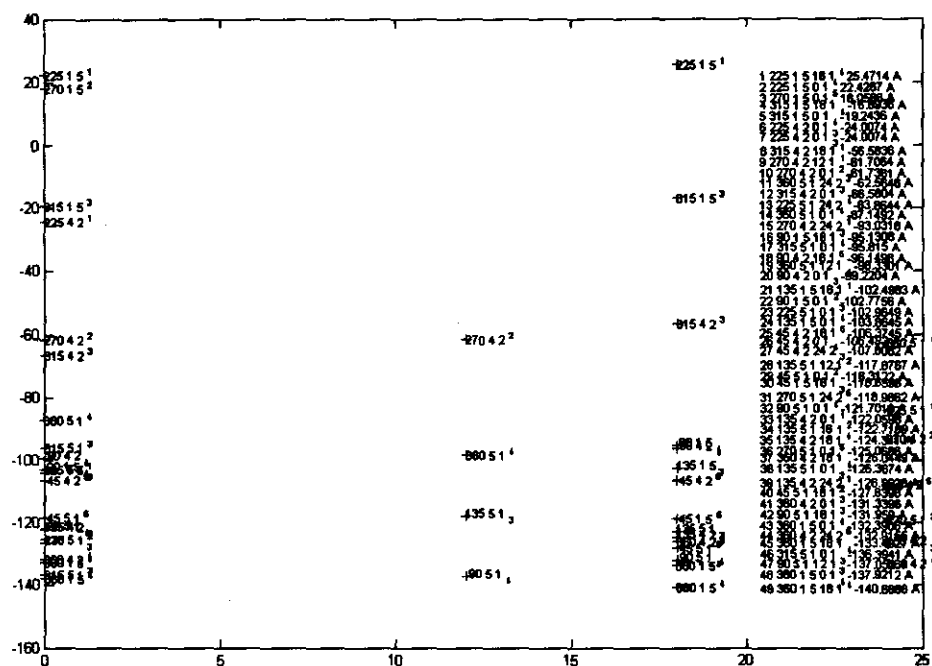


Figure 119: Feature *FFT8_BPFO_X* (magnitude spectrum value of the signal to the 8th power at ball pass outer race defect frequency on x axis) for all defects

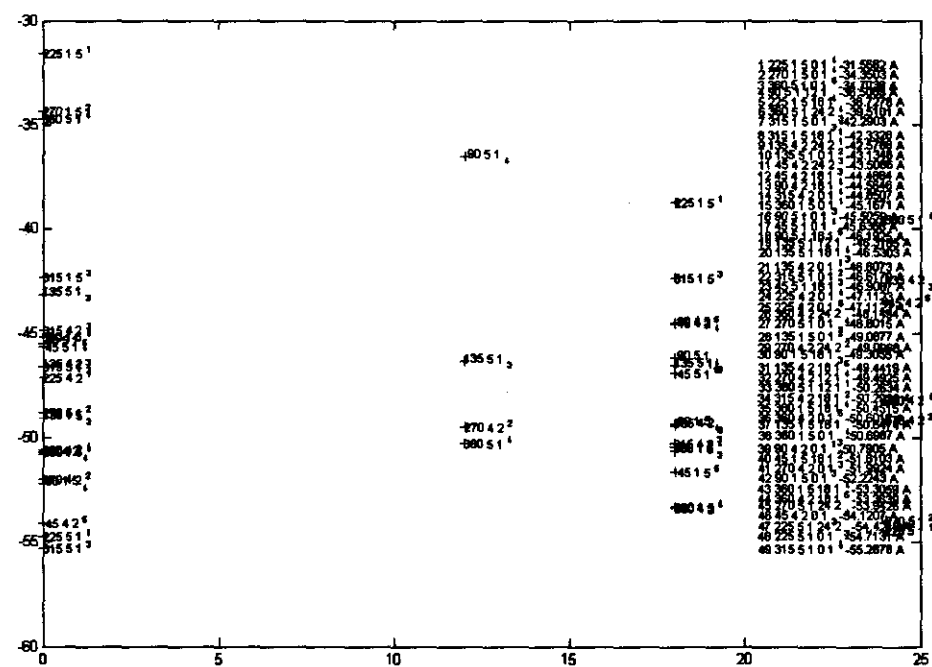


Figure 120: Feature *SA_BPFO2_0_X* (square root of the SOS of synchronous average of the signal at ball pass outer race defect frequency on x axis) for all defects

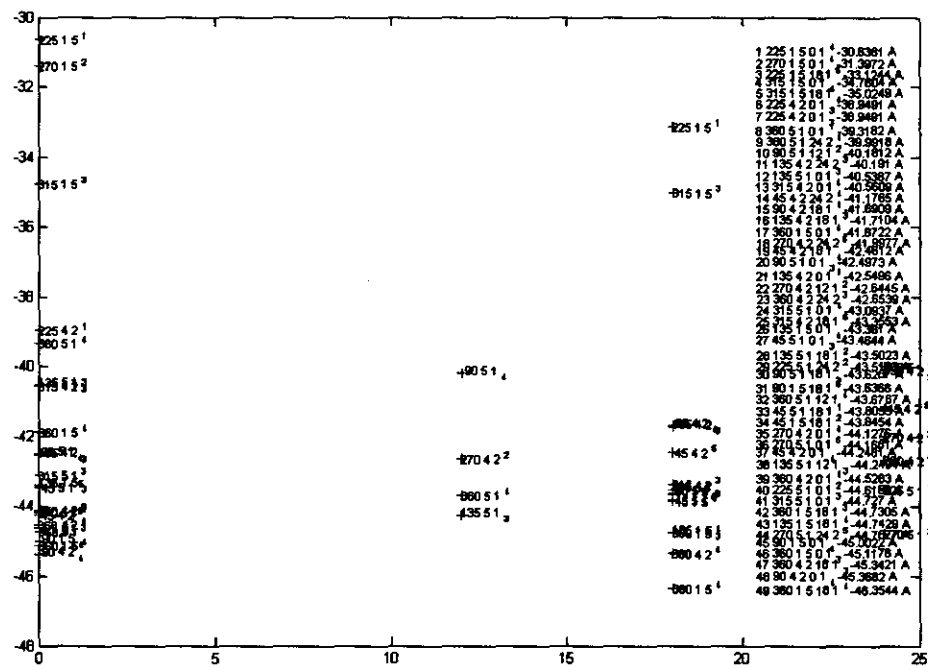


Figure 121: Feature *SA_BPFO2_p_X* (maximum value of synchronous average of the signal at ball pass outer race defect frequency on x axis) for all defects

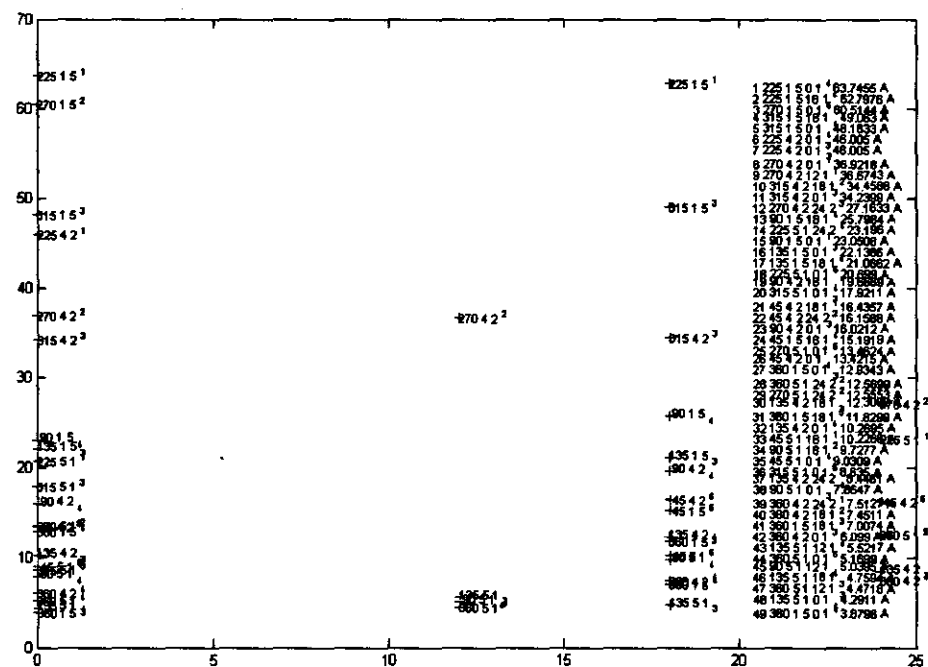


Figure 122: Feature *BPFO2_10_Eng_X* (square root of the SOS of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for all defects

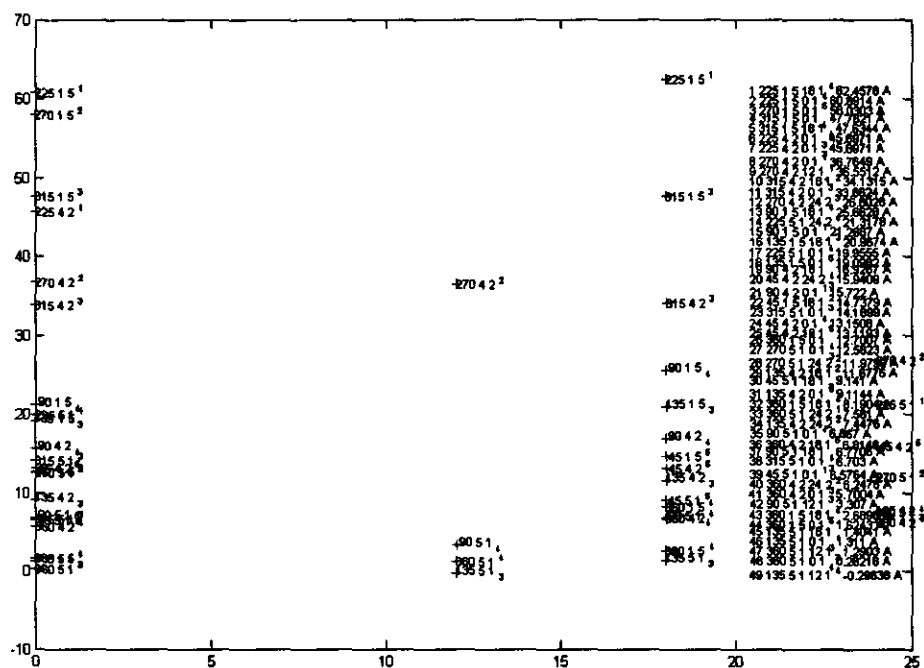


Figure 123: Feature *BPFO2_10_Max_X* (maximum value of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for all defects

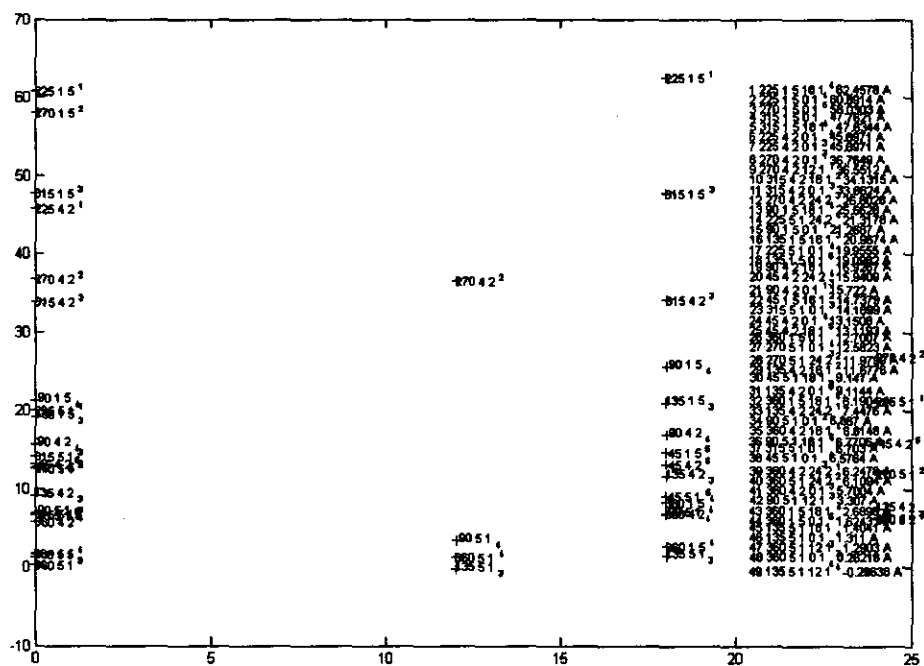


Figure 124: Feature *BPF02_2_Eng_X* (square root of the SOS of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for all defects

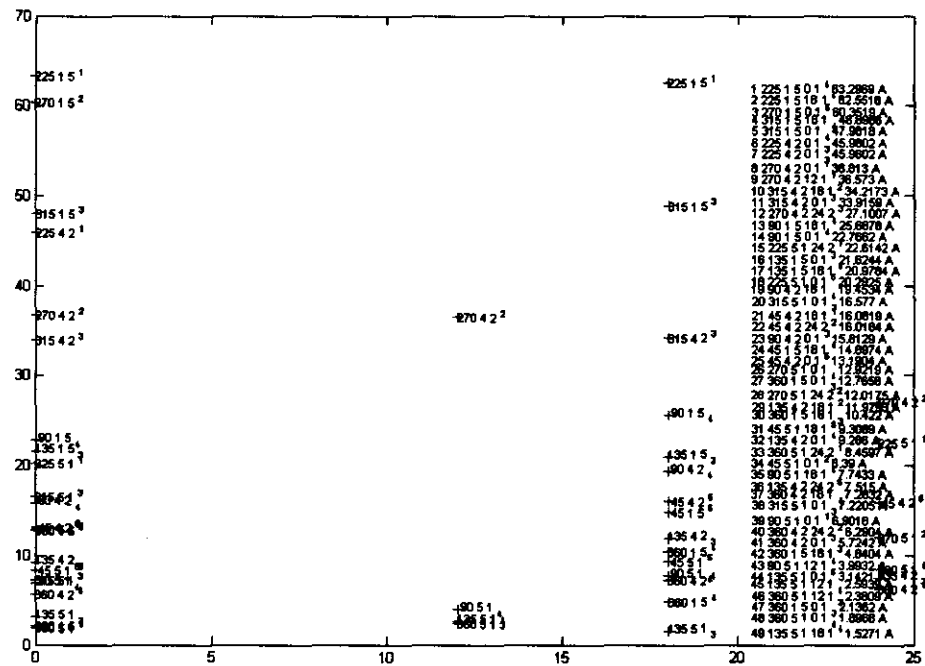


Figure 125: Feature *BPFO2_2_Max_X* (maximum value of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for all defects

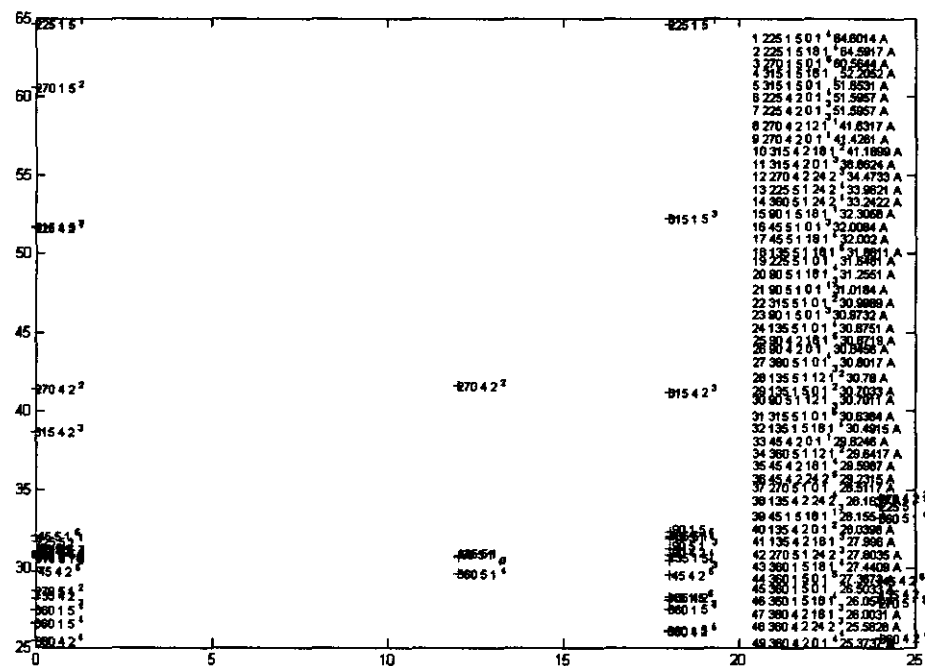


Figure 126: Feature *HFD_5_10_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 5–10kHz – on x axis) for all defects

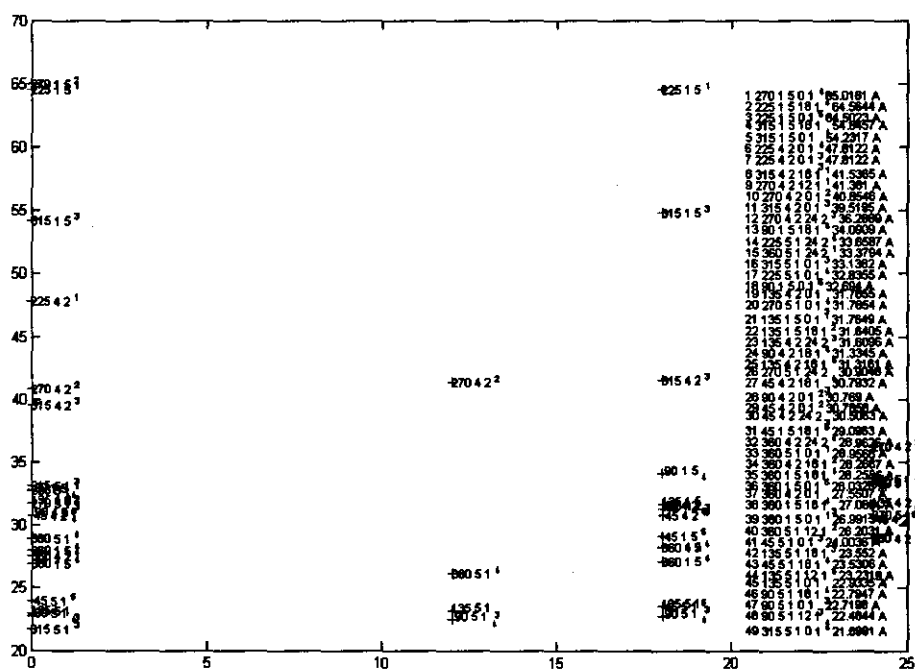


Figure 127: Feature $HFD\_2\_5\_kHz\_X$ (square root of SOS of HFD components in magnitude spectrum – from 2–5kHz – on x axis) for all defects

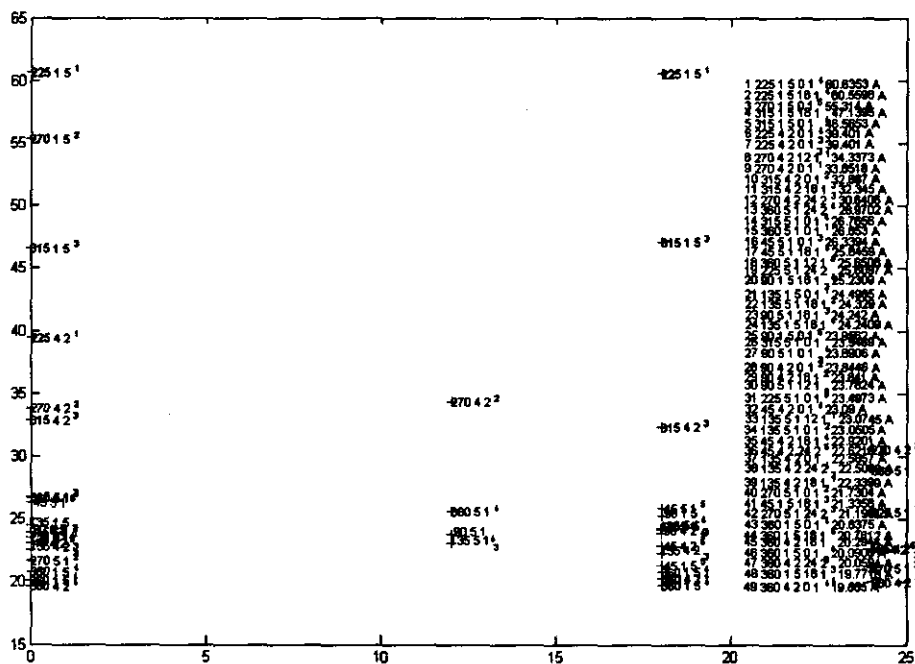


Figure 128: Feature $HFD\_5\_6\_kHz\_X$ (square root of SOS of HFD components in magnitude spectrum – from 5–6kHz – on x axis) for all defects

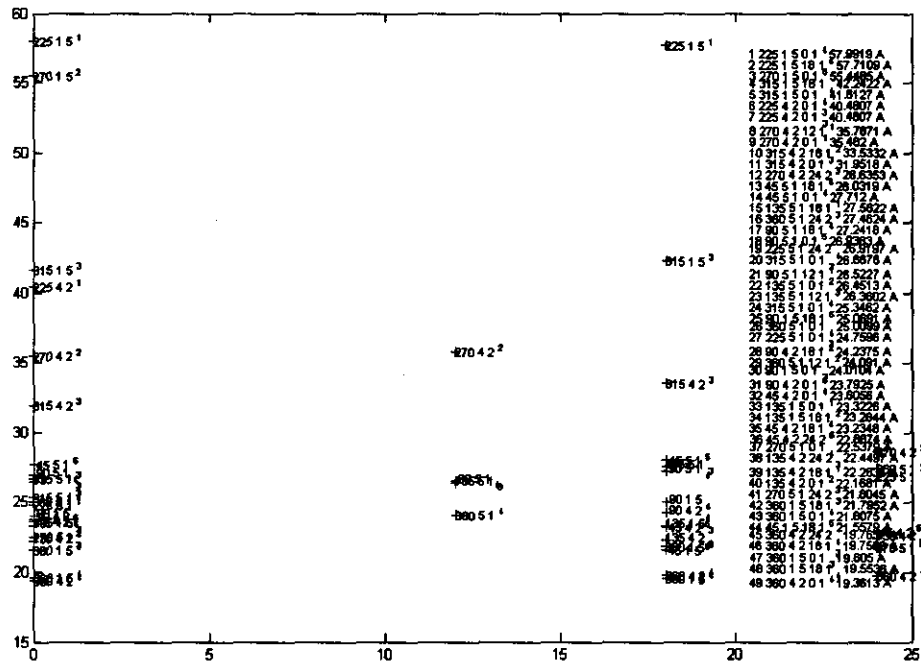


Figure 129: Feature *HFD_6_7_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 6–7kHz – on x axis) for all defects

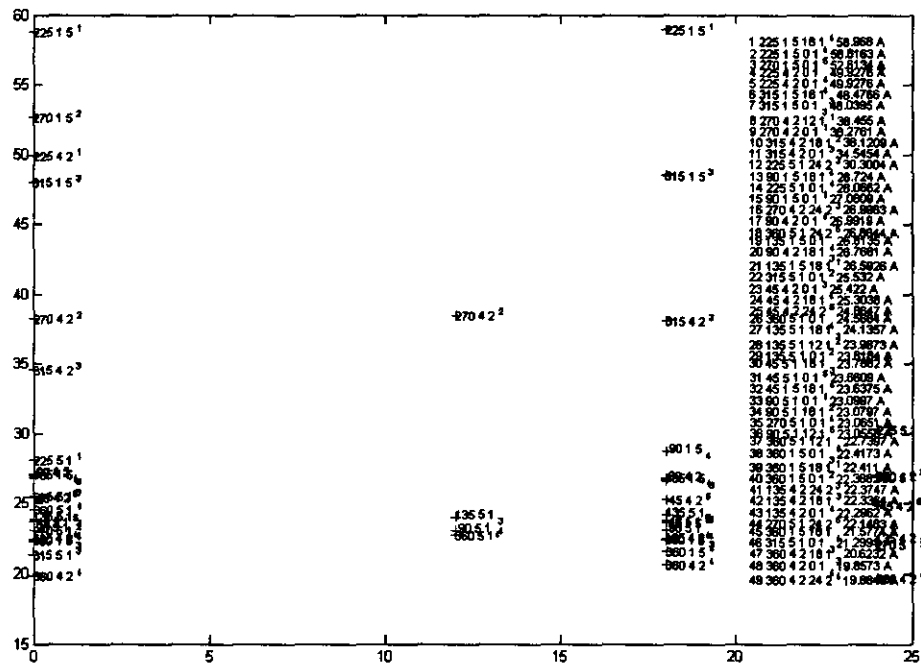


Figure 130: Feature *HFD_8_10_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 8–10kHz – on x axis) for all defects

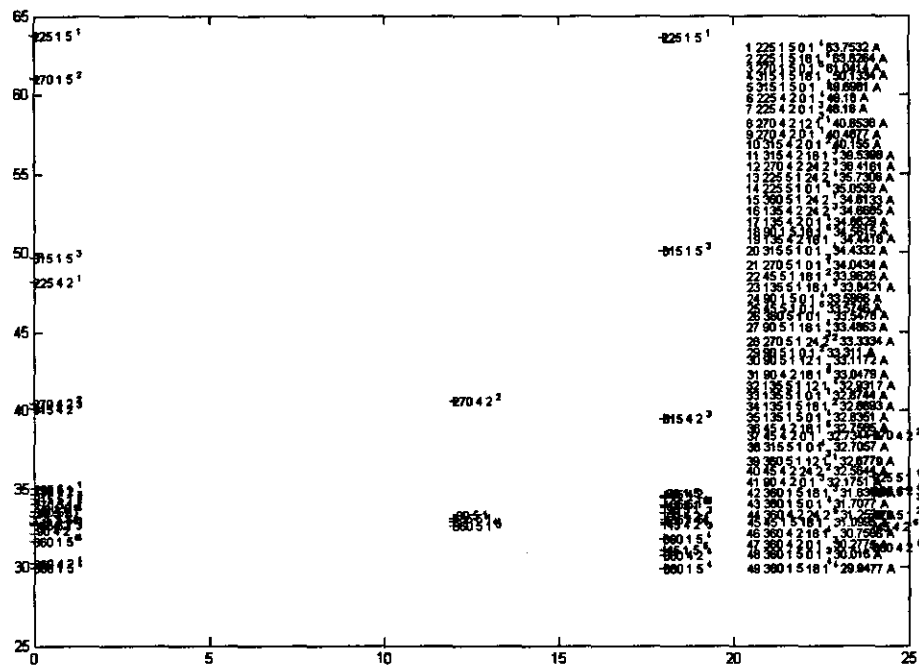


Figure 131: Feature *FFT2_MEAN_X* (first component of demodulation spectrum on x axis) for all defects

11.3.1.2 Large outer race defect

For each data point indicated by 'x' the defect angle (DA) is shown (180°, 225°, 270°, 315°, 360°, 45°, 90° and 135° for the Vib demo data set 3), the bearing number (1, 4, 5 for the Vib demo data set 3), the bearing defect severity (5, 2, 1 for the Vib demo data set 3 where 5 is the largest bearing defect corresponding to the same bearing used in the Vib demo data set 2) and the distance of the defect from 225° (refer to figure 43 on p. 214 for notational details). The dynamic load (unbalance mass in grams) is shown on the horizontal axis to demonstrate the influence of dynamic loading on bearing defects of decreasing severity and defect angle.

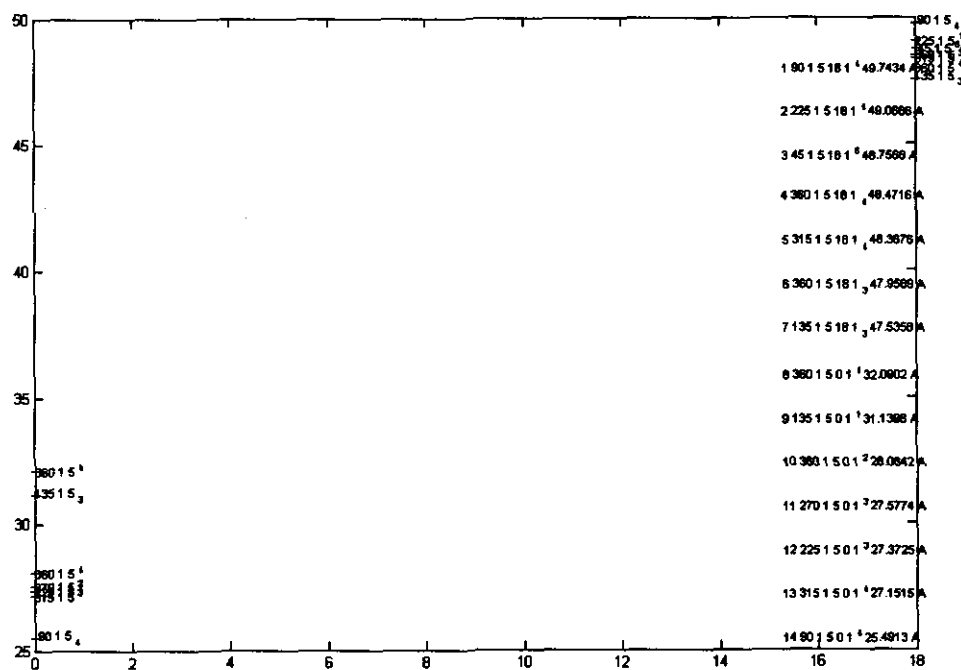


Figure 132: Feature Fr_X (magnitude spectrum value at rotational frequency on x axis) for large defect

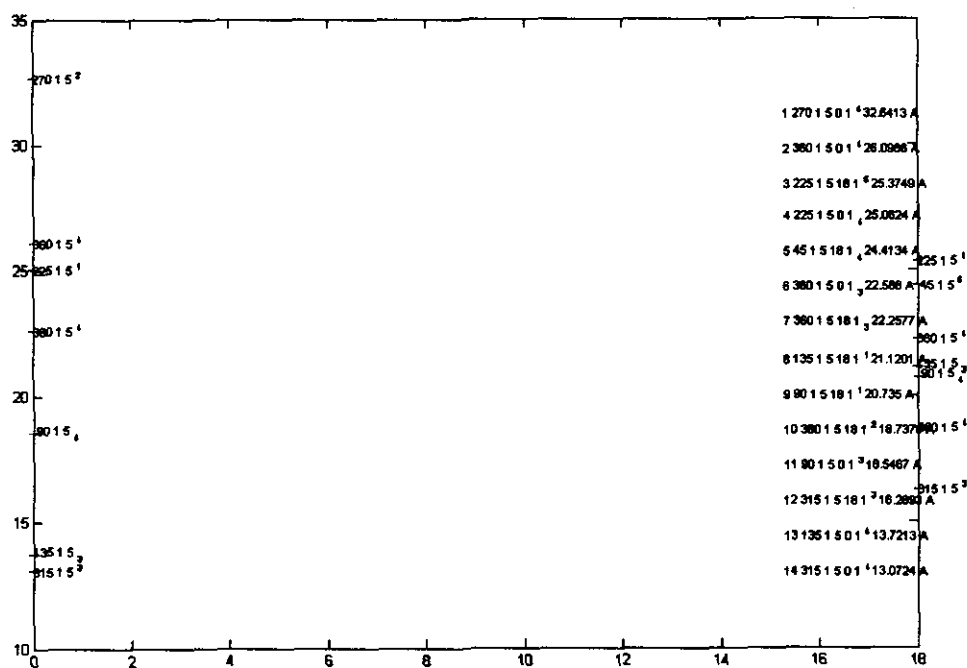


Figure 133: Feature *BPFO__X* (magnitude spectrum value at ball pass outer race defect frequency on x axis) for large defect

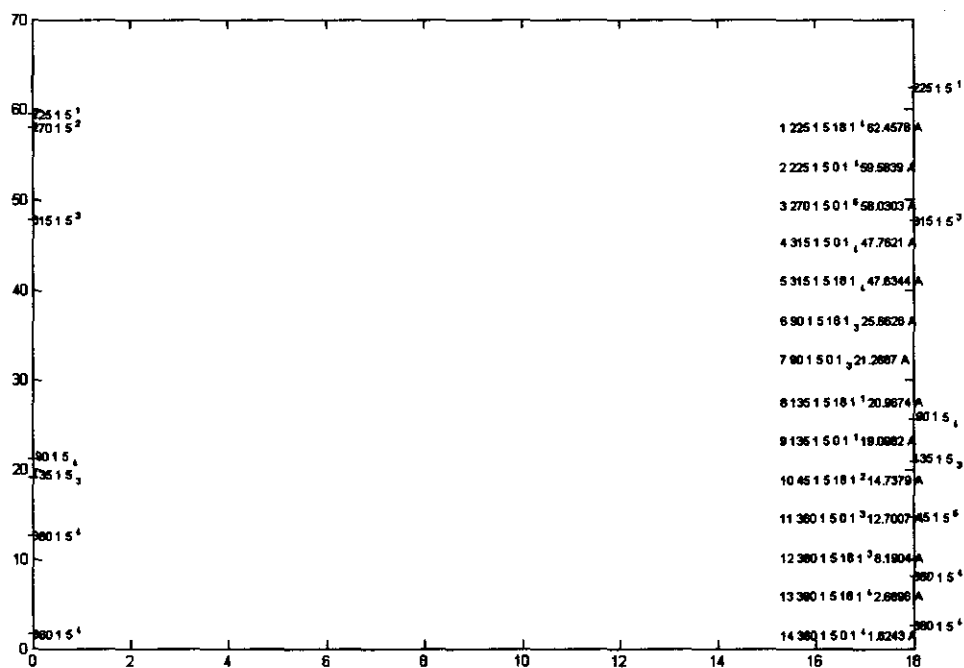


Figure 134: Feature *BPF02_X* (demodulation spectrum value at ball pass outer race defect frequency on x axis) for large defect

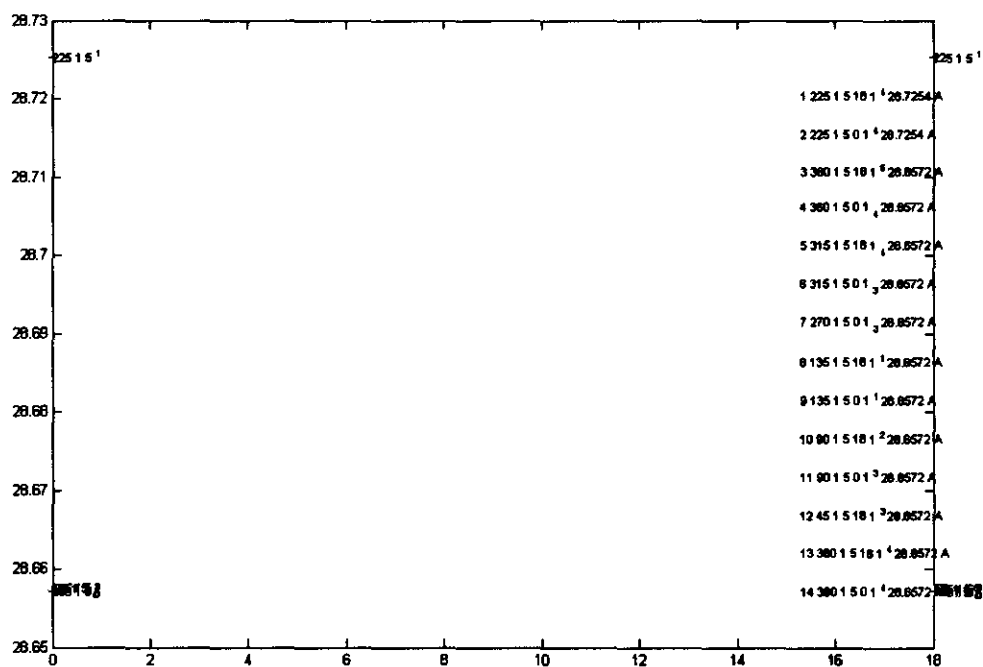


Figure 135: Feature *Fr_S_X* (rotational frequency on x axis) for large defect

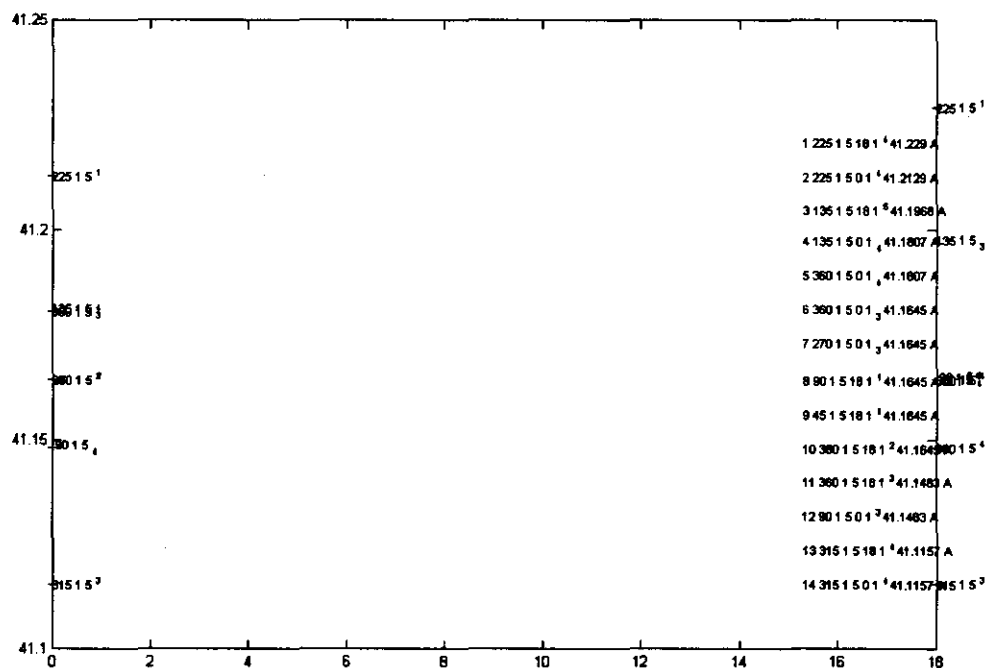


Figure 136: Feature *BPF02_S_X* (ball pass outer race defect frequency on x axis) for large defect

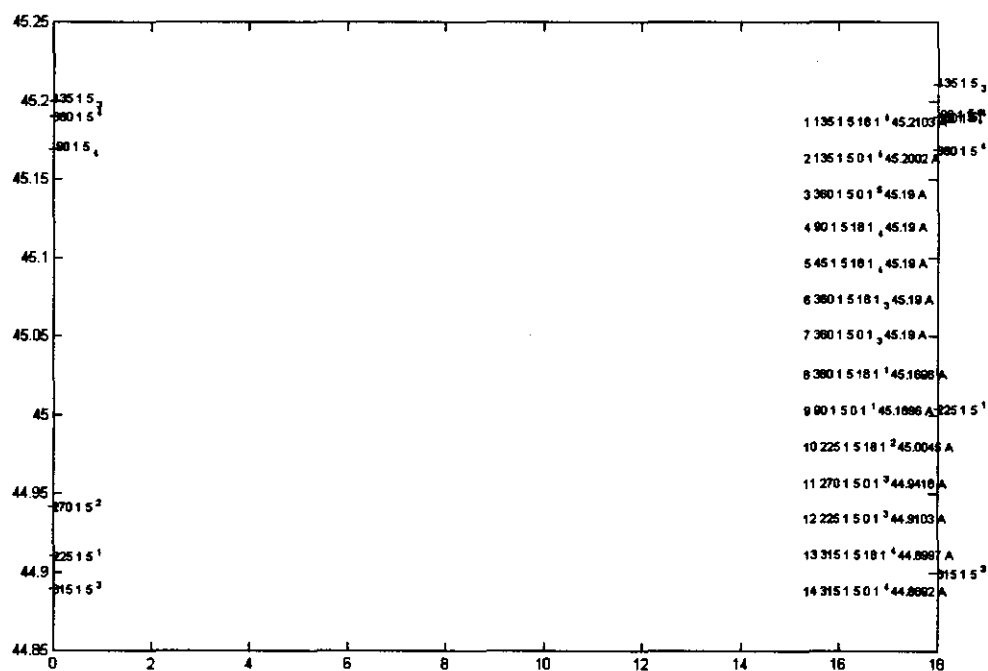


Figure 137: Feature *BPF12_S_X* (ball pass inner race defect frequency on x axis) for large defect

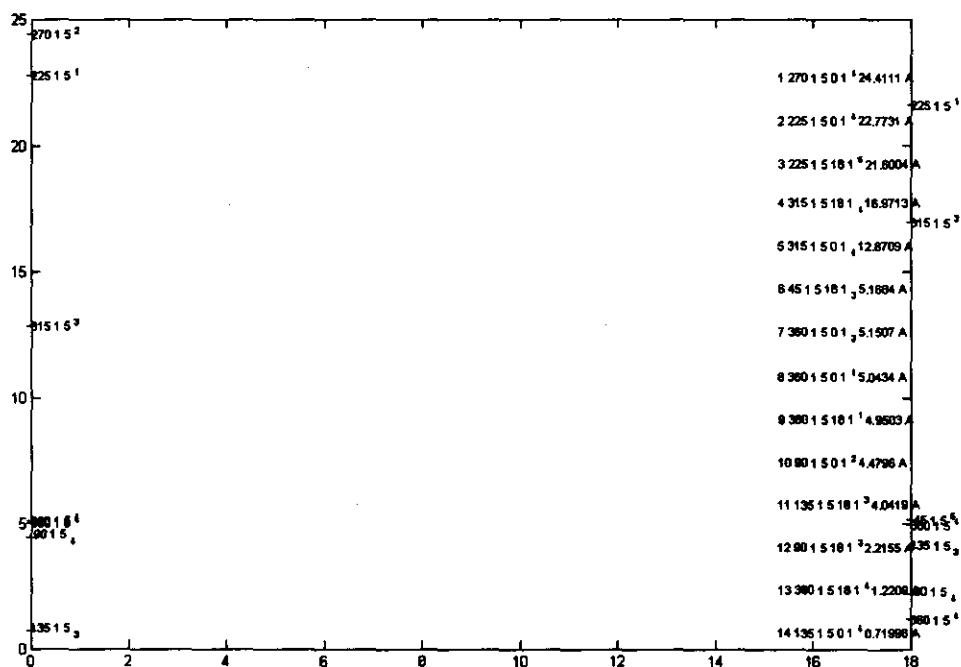


Figure 138: Feature *BPF12_X* (demodulation spectrum value at ball pass inner race defect frequency on x axis) for large defect

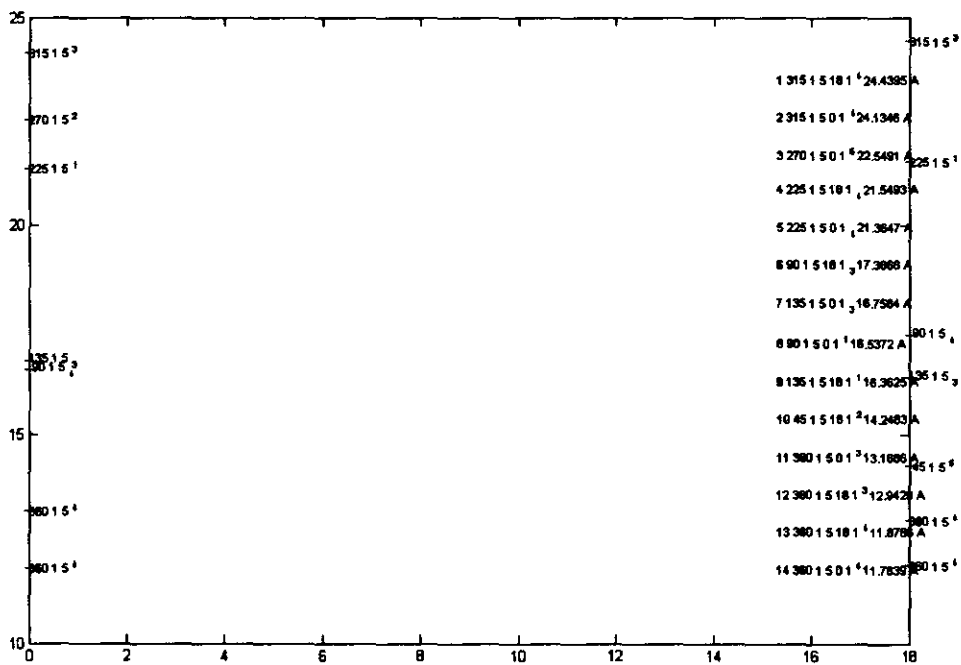


Figure 139: Feature *KH_X* (kurtosis on x axis) for large defect

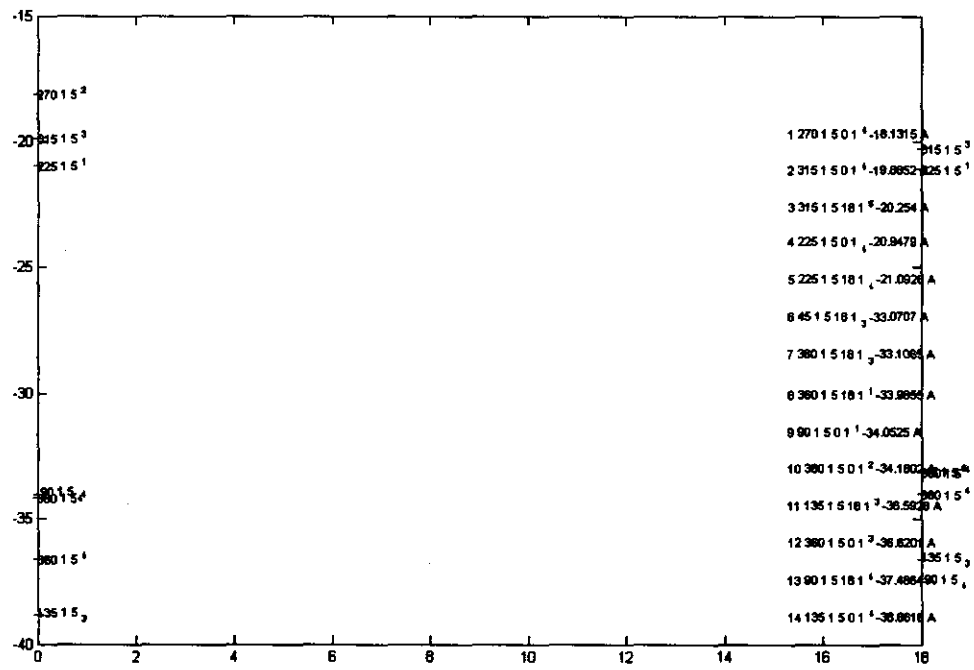


Figure 140: Feature CH_X (real cepstrum on x axis) for large defect

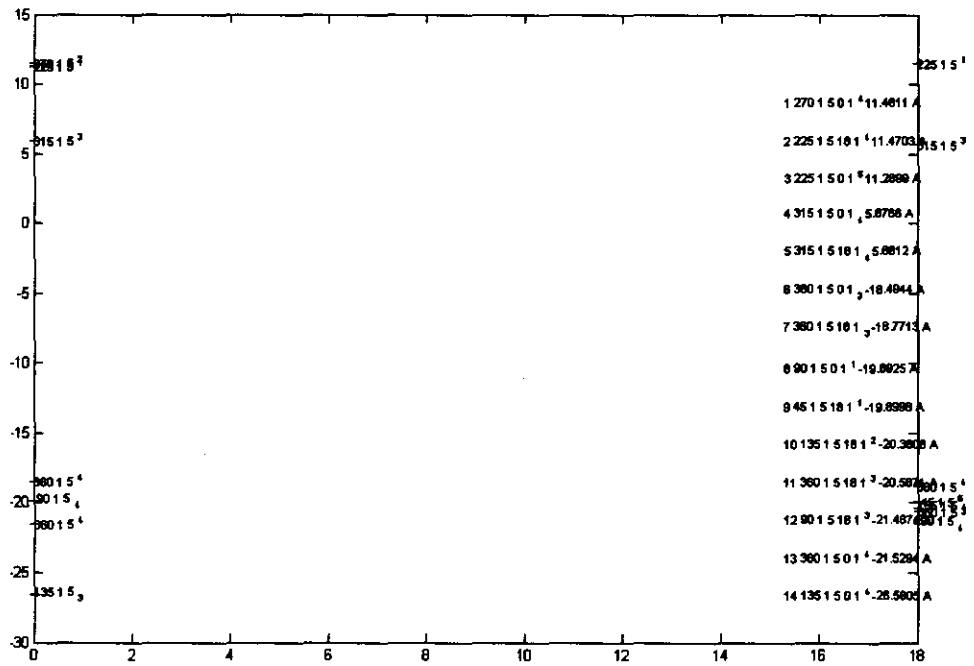


Figure 141: Feature LH_X (linear cepstrum on x axis) for large defect

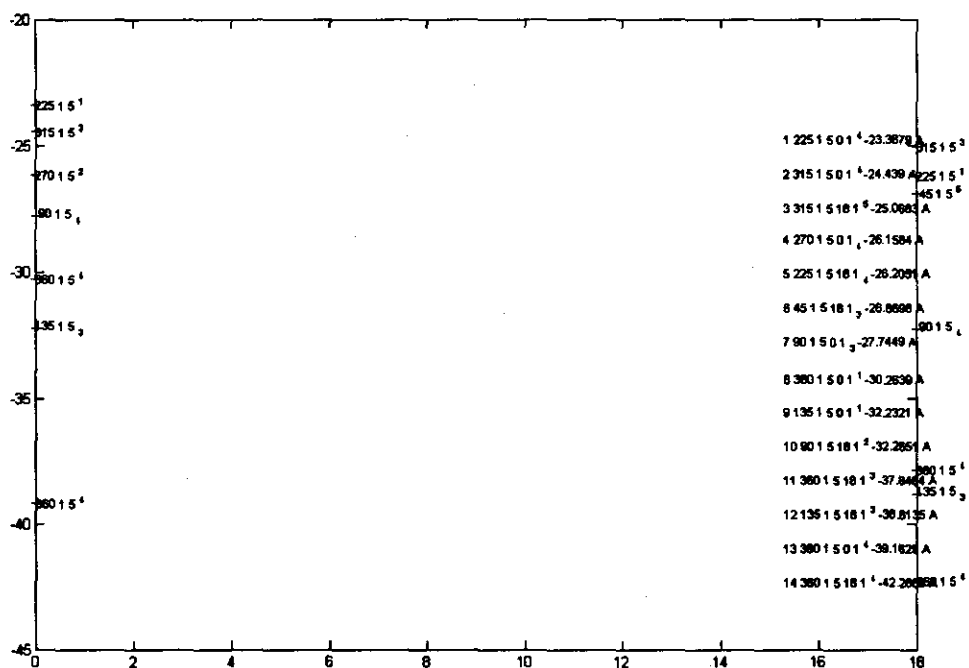


Figure 142: Feature CH_p_X (real cepstrum – after highpass filter – on x axis) for large defect

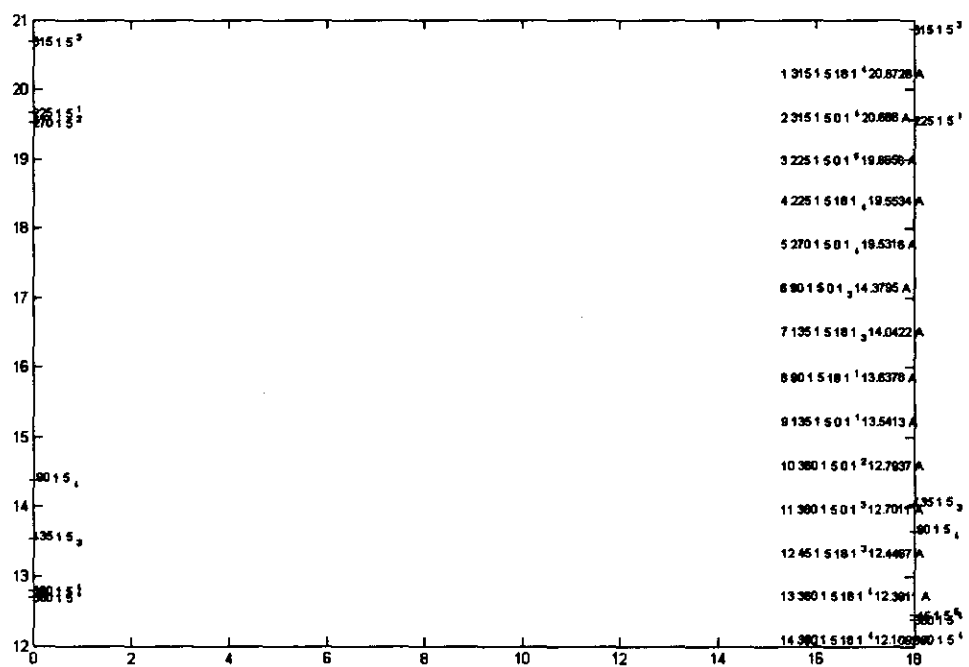


Figure 143: Feature KH_p_X (kurtosis – after highpass filter – on x axis) for large defect

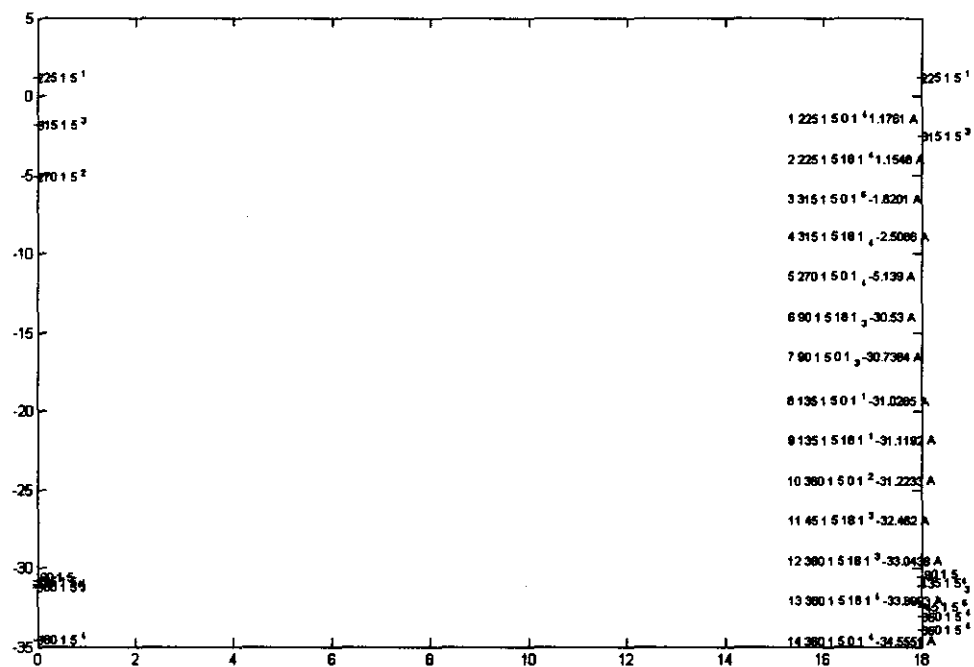


Figure 144: Feature LH_p_X (linear cepstrum – after highpass filter – on x axis) for large defect

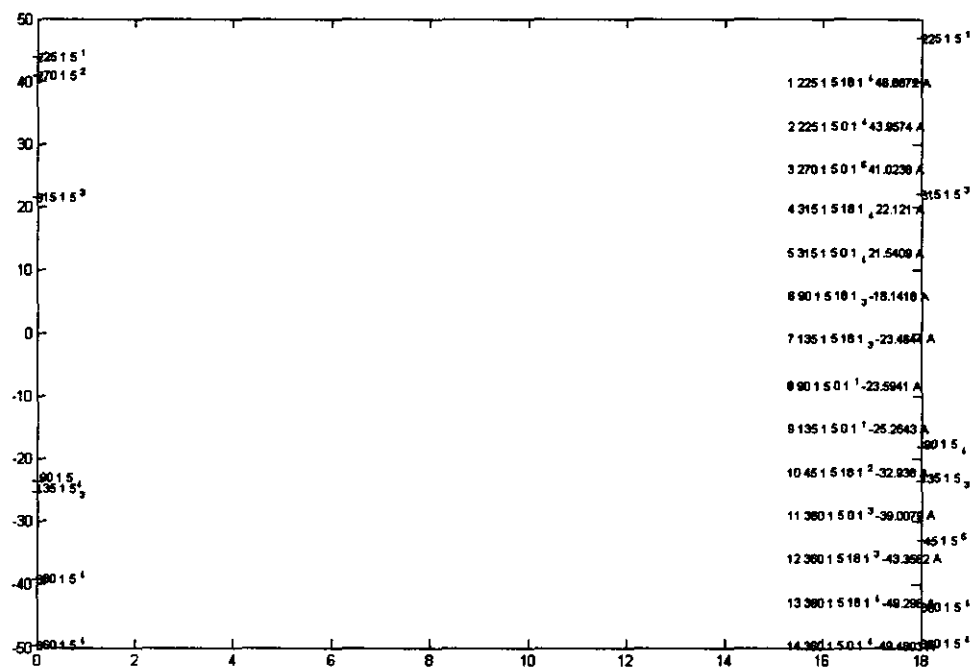


Figure 145: Feature $FFT4\_BPFO\_X$ (magnitude spectrum value of the square of the signal squared at ball pass outer race defect frequency on x axis) for large defect

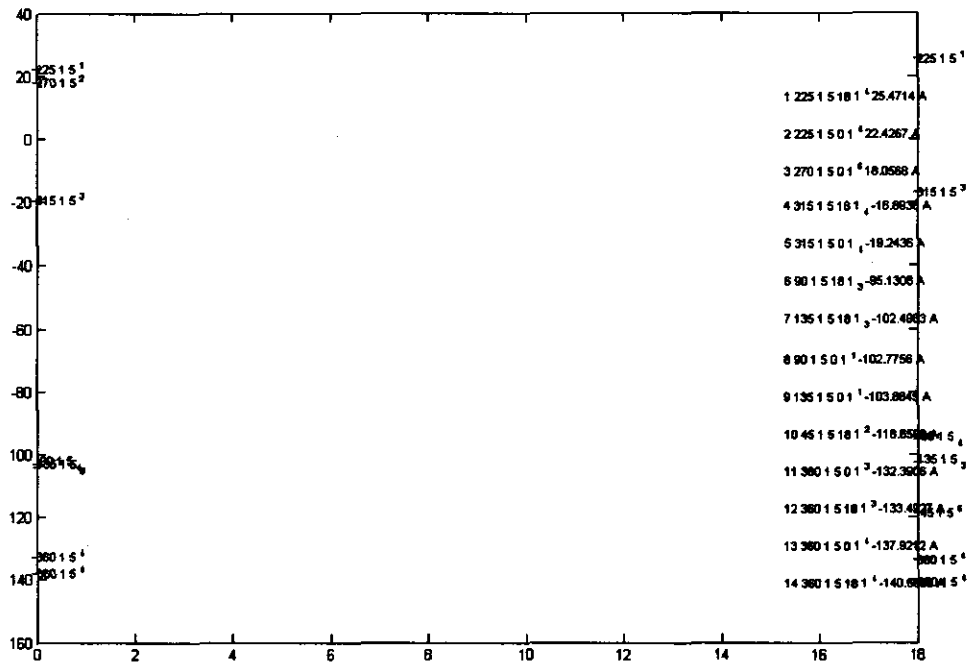


Figure 146: Feature *FFT8_BPFO_X* (magnitude spectrum value of the signal to the 8th power at ball pass outer race defect frequency on x axis) for large defect

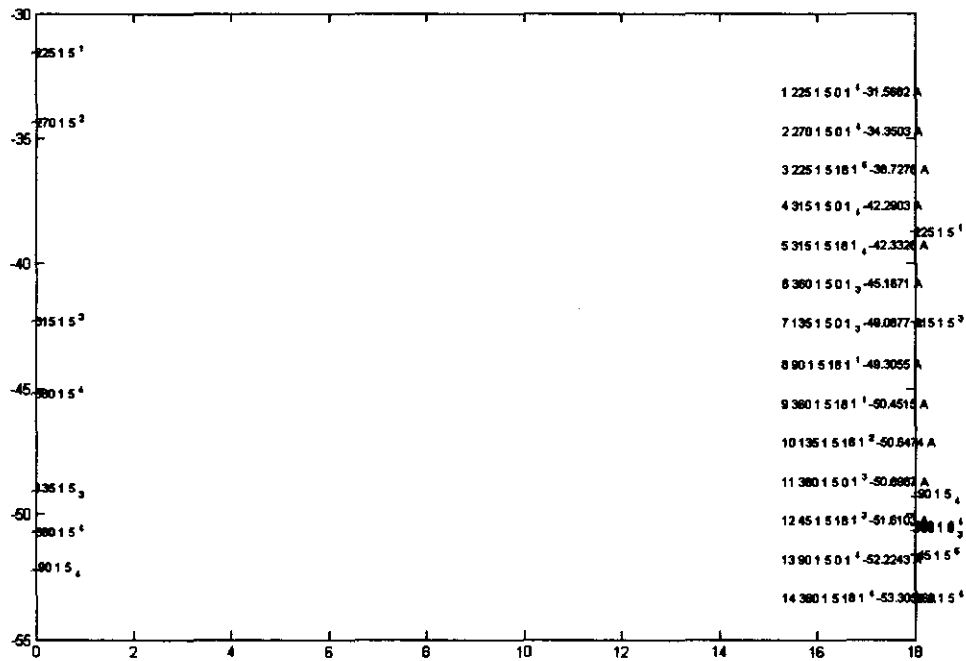


Figure 147: Feature *SA_BPFO2_0_X* (square root of the SOS of synchronous average of the signal at ball pass outer race defect frequency on x axis) for large defect

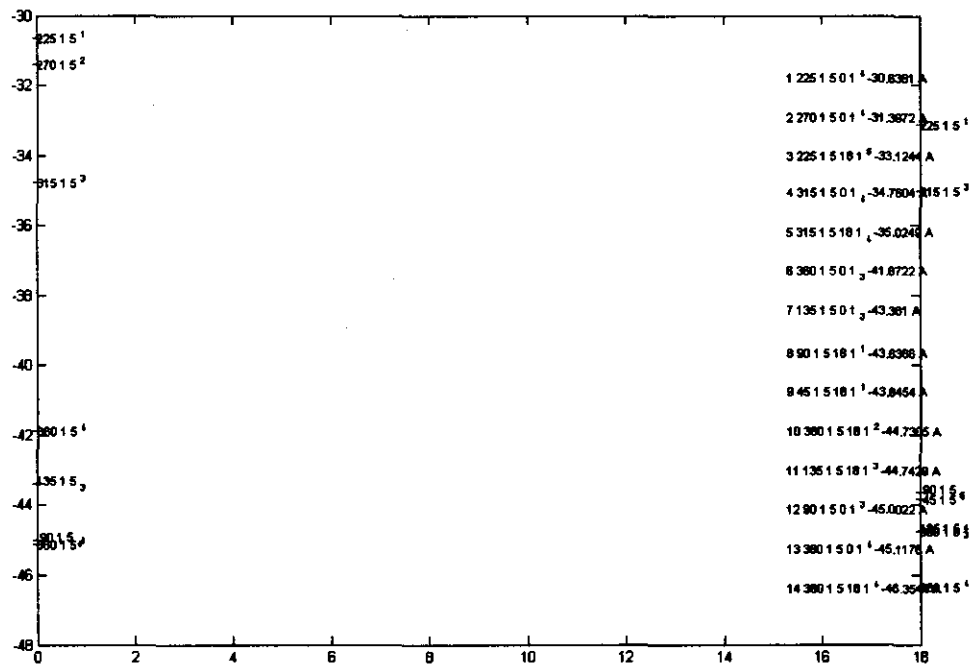


Figure 148: Feature *SA_BPFO2_p_X* (maximum value of synchronous average of the signal at ball pass outer race defect frequency on x axis) for large defect

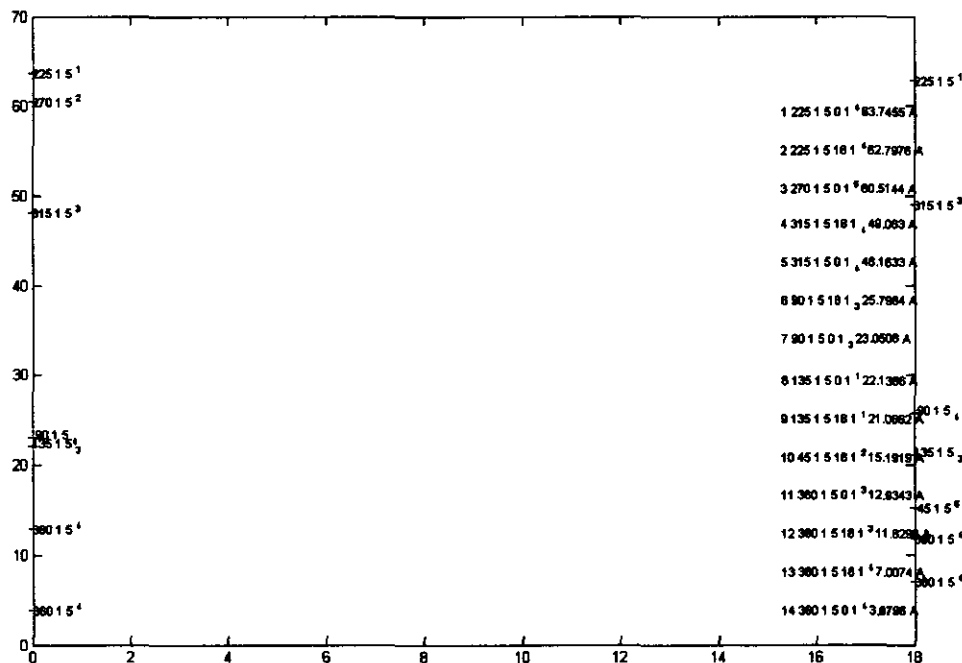


Figure 149: Feature *BPFO2_10_Eng_X* (square root of the SOS of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect

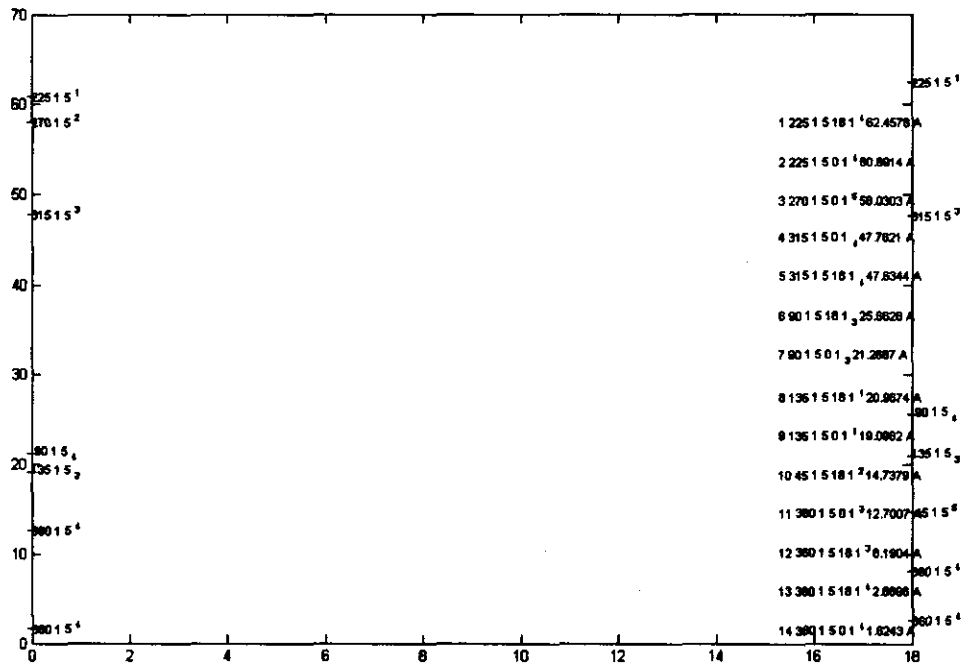


Figure 150: Feature *BPFO2_10_Max_X* (maximum value of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect

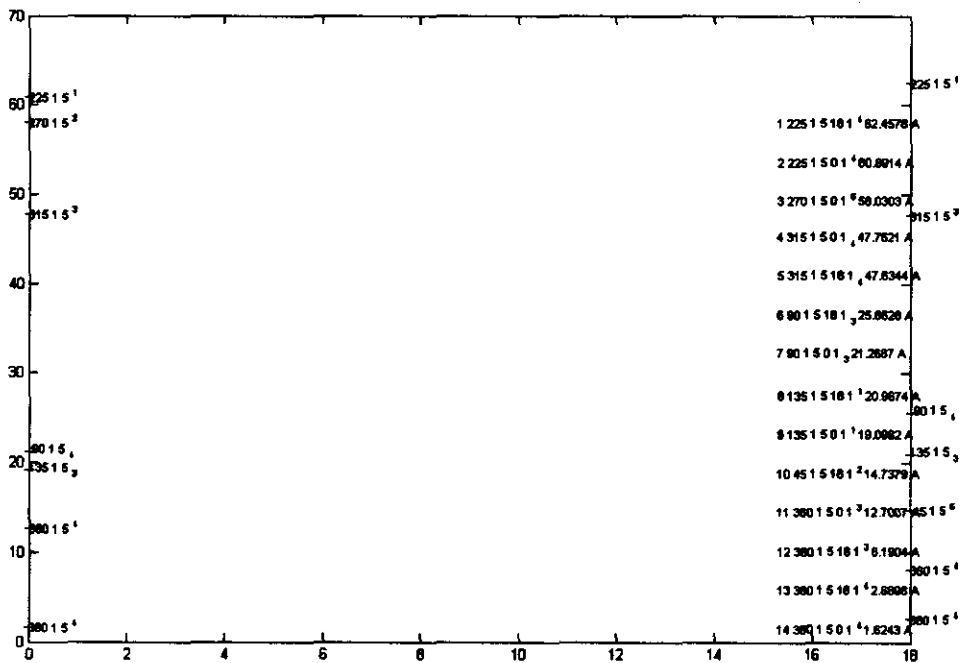


Figure 151: Feature *BPFO2_2_Eng_X* (square root of the SOS of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect

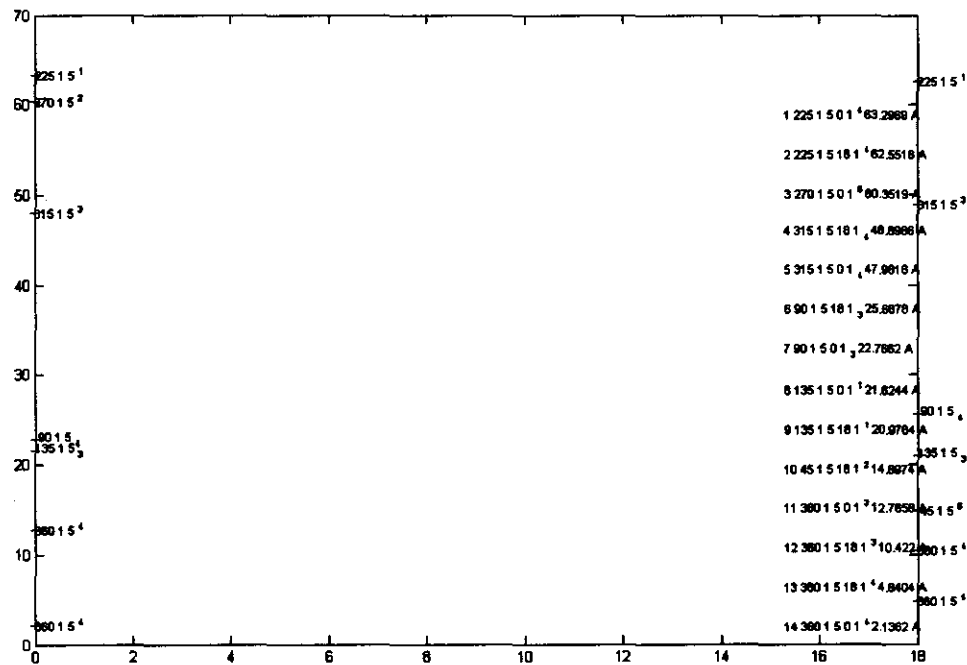


Figure 152: Feature *BPFO2_2_Max_X* (maximum value of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for large defect

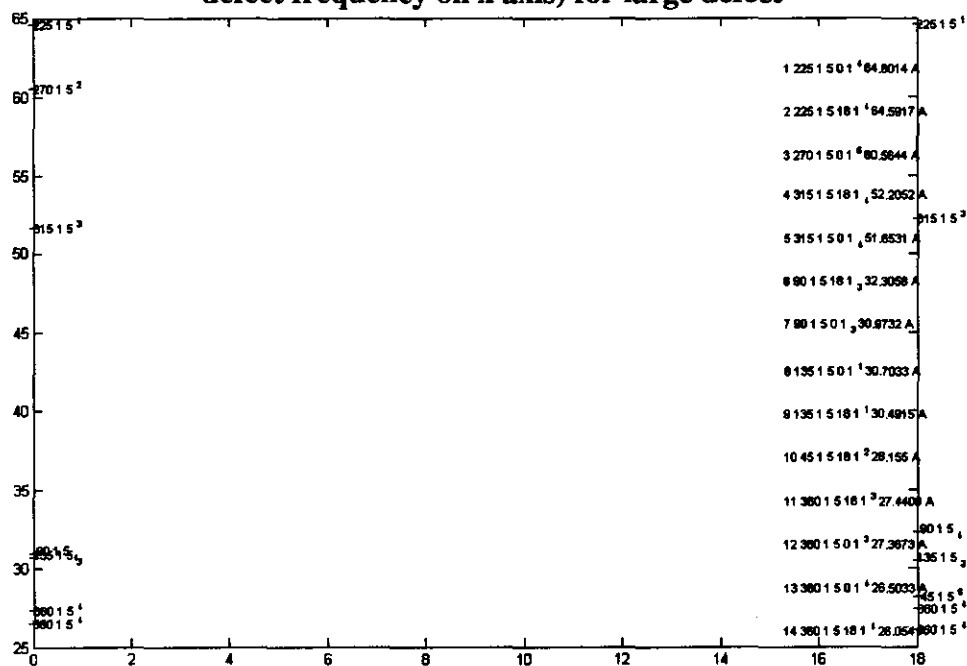


Figure 153: Feature *HFD_5_10_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 5–10kHz – on x axis) for large defect

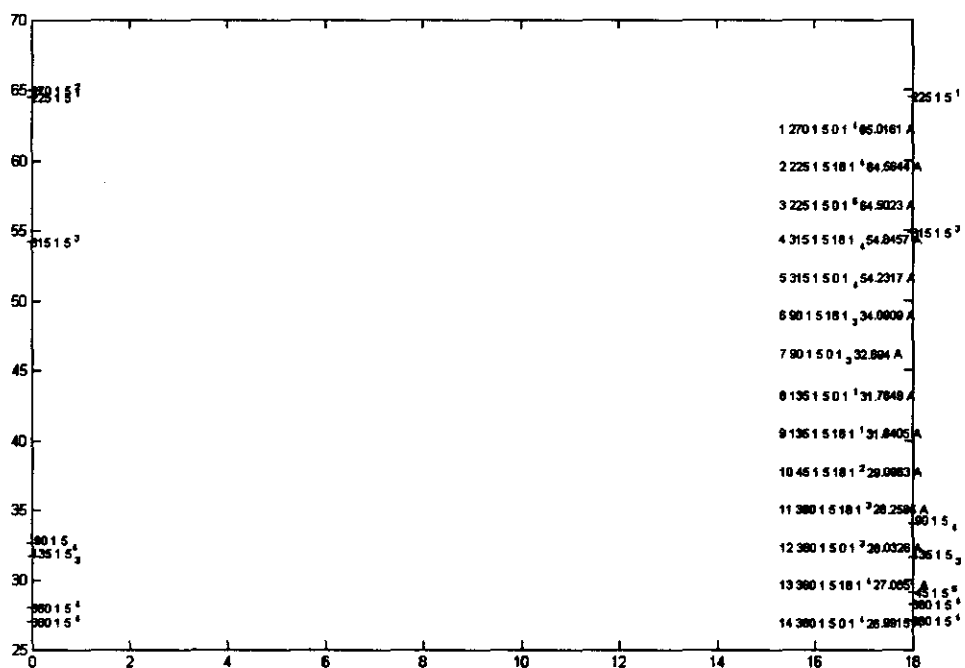


Figure 154: Feature *HFD_2_5_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 2–5kHz – on x axis) for large defect

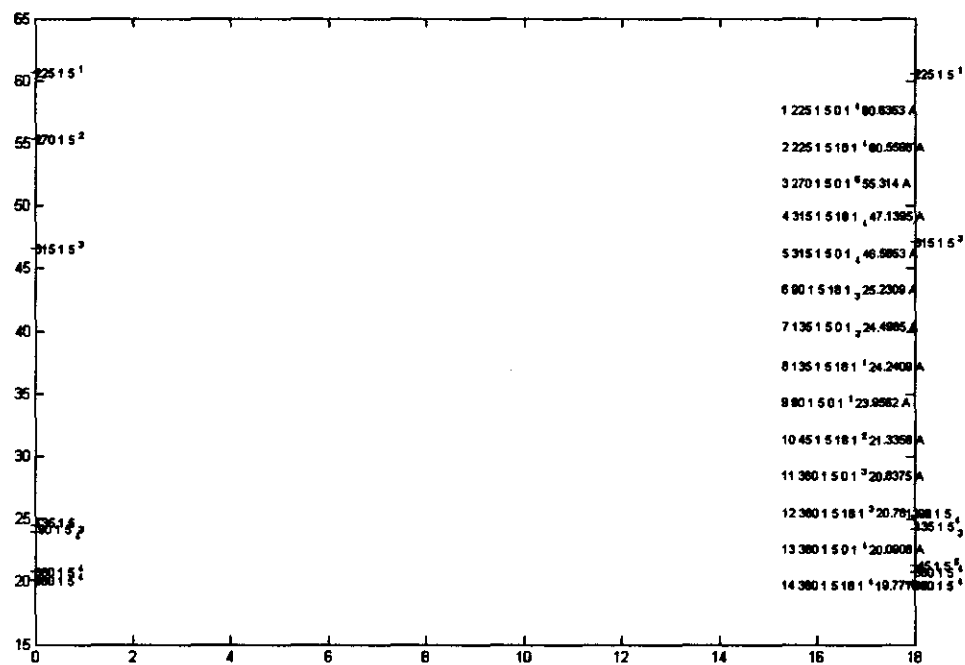


Figure 155: Feature *HFD_5_6_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 5–6kHz – on x axis) for large defect

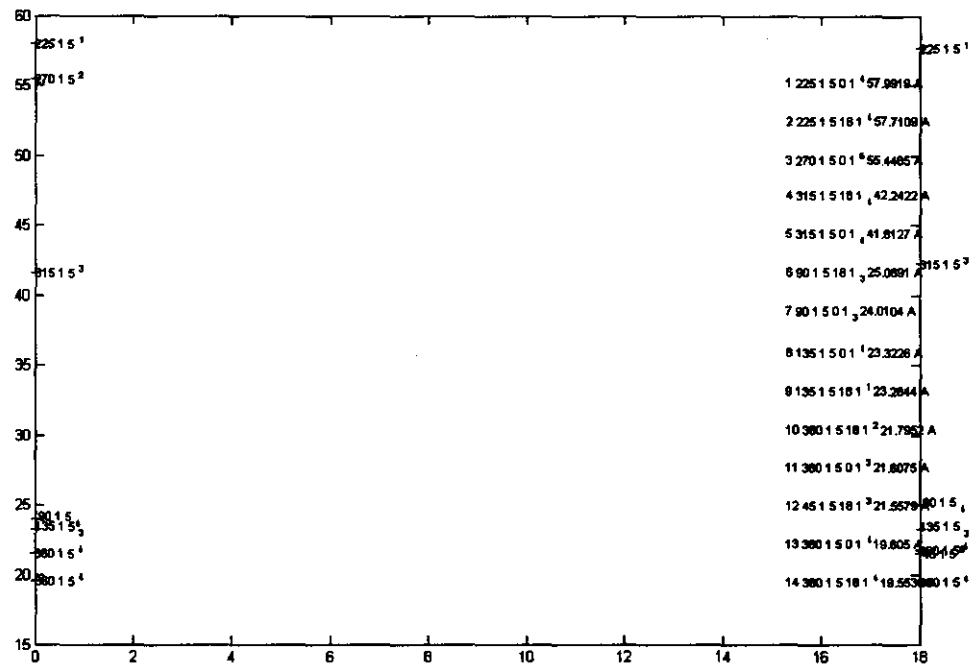


Figure 156: Feature *HFD_6_7_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 6–7kHz – on x axis) for large defect

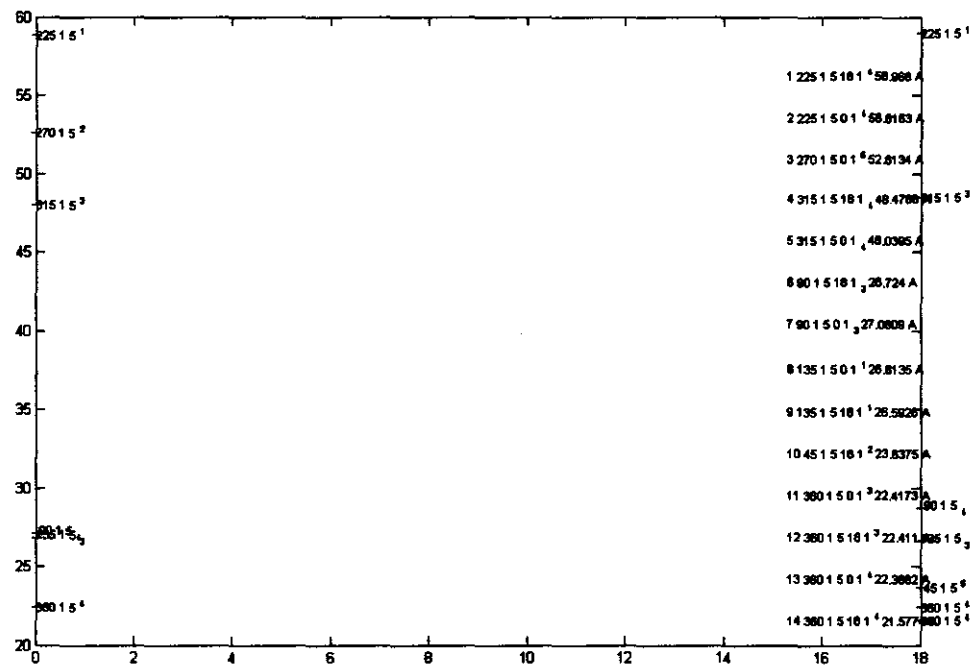


Figure 157: Feature *HFD_8_10_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 8–10kHz – on x axis) for large defect

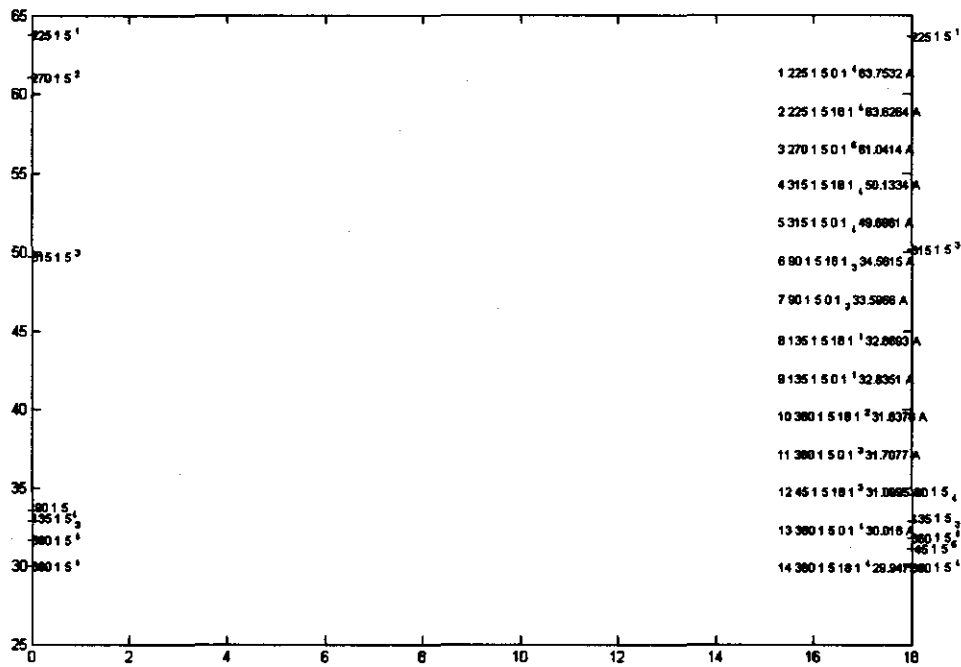


Figure 158: Feature *FFT2_MEAN_X* (first component of demodulation spectrum on x axis) for large defect

11.3.1.3 Medium outer race defect

For each data point indicated by 'x' the defect angle (DA) is shown (180°, 225°, 270°, 315°, 360°, 45°, 90° and 135° for the Vib demo data set 3), the bearing number (1, 4, 5 for the Vib demo data set 3), the bearing defect severity (5, 2, 1 for the Vib demo data set 3 where 5 is the largest bearing defect corresponding to the same bearing used in the Vib demo data set 2) and the distance of the defect from 225° (refer to figure 43 on p. 214 for notational details). The dynamic load (unbalance mass in grams) is shown on the horizontal axis to demonstrate the influence of dynamic loading on bearing defects of decreasing severity and defect angle.

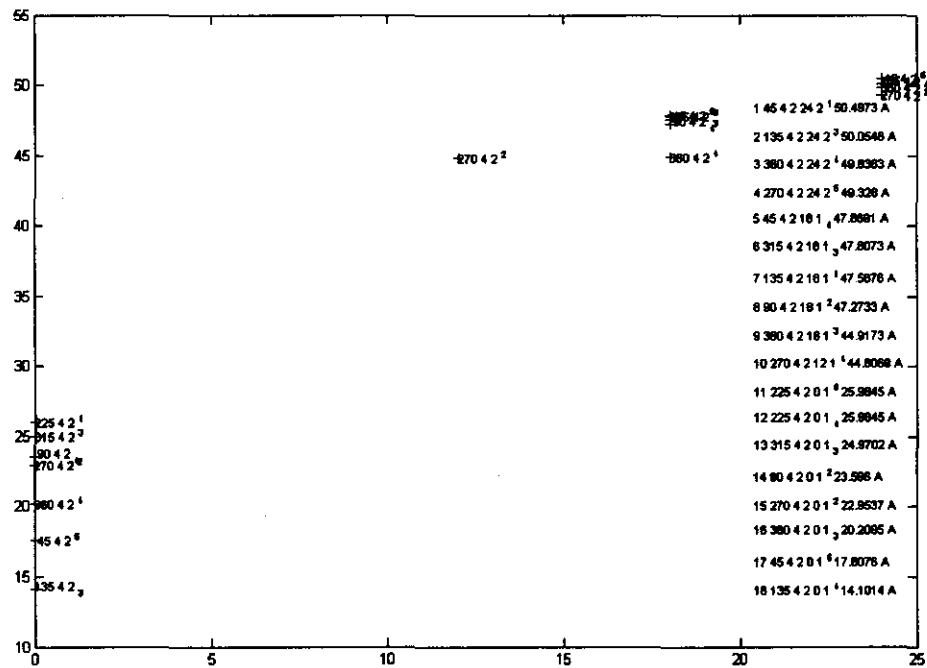


Figure 159: Feature *Fr_X* (magnitude spectrum value at rotational frequency on x axis) for medium defect

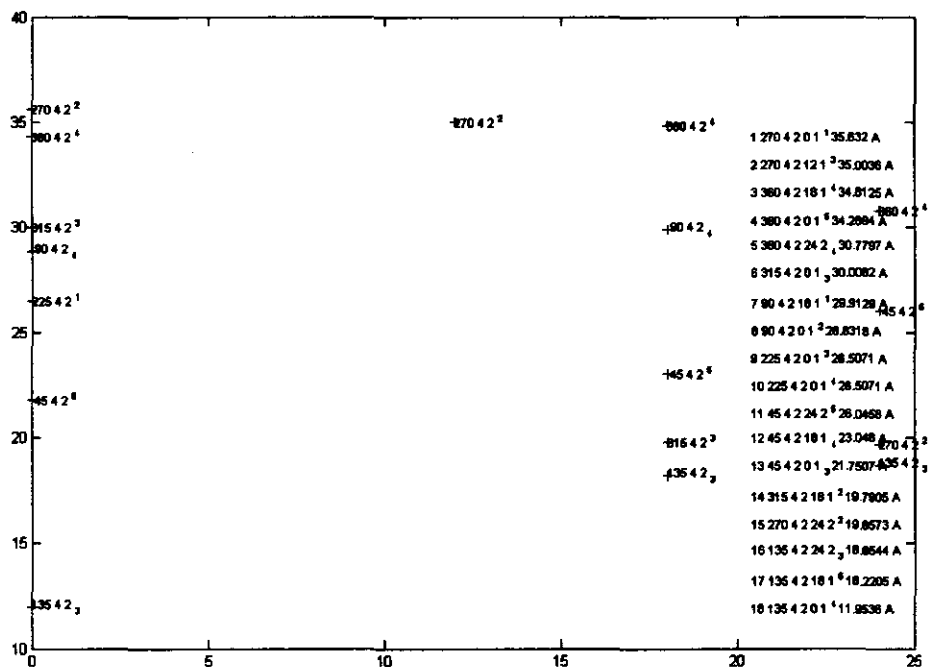


Figure 160: Feature *BPFO_X* (magnitude spectrum value at ball pass outer race defect frequency on x axis) for medium defect

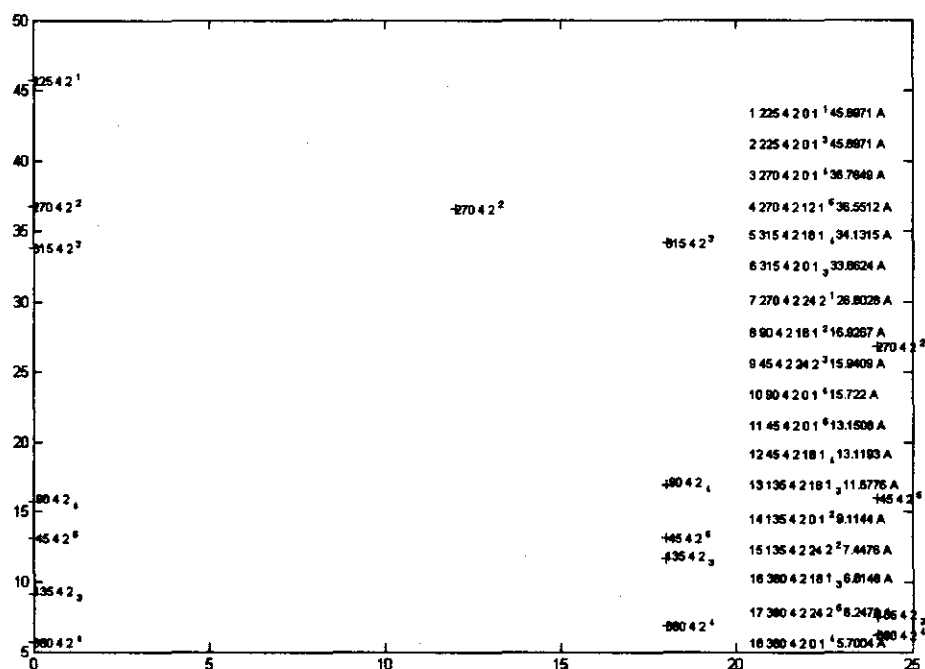


Figure 161: Feature *BPF02_X* (demodulation spectrum value at ball pass outer race defect frequency on x axis) for medium defect

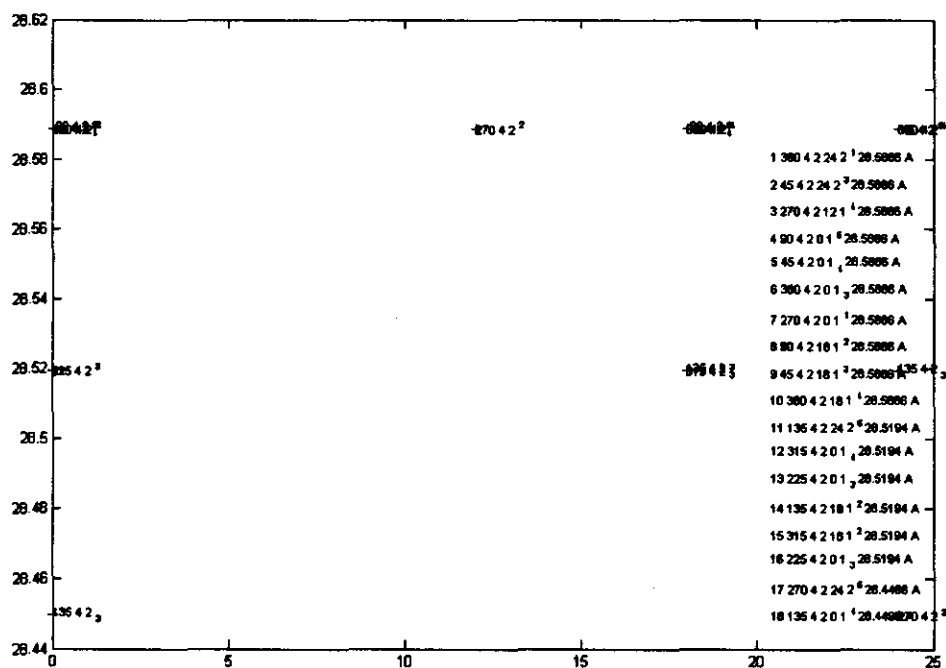


Figure 162: Feature *Fr_S_X* (rotational frequency on x axis) for medium defect

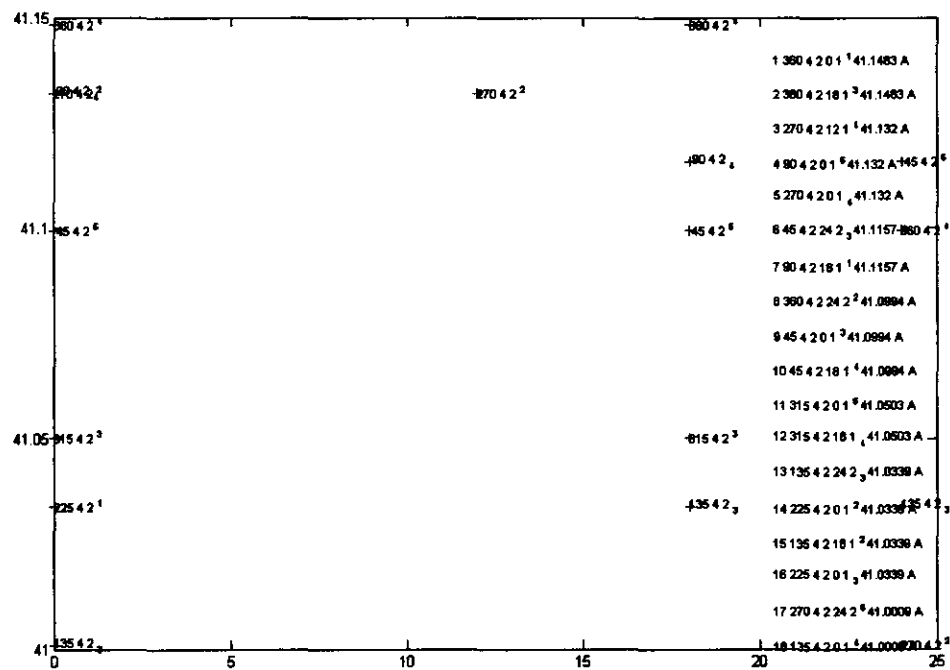


Figure 163: Feature *BPFO2_S_X* (ball pass outer race defect frequency on x axis) for medium defect

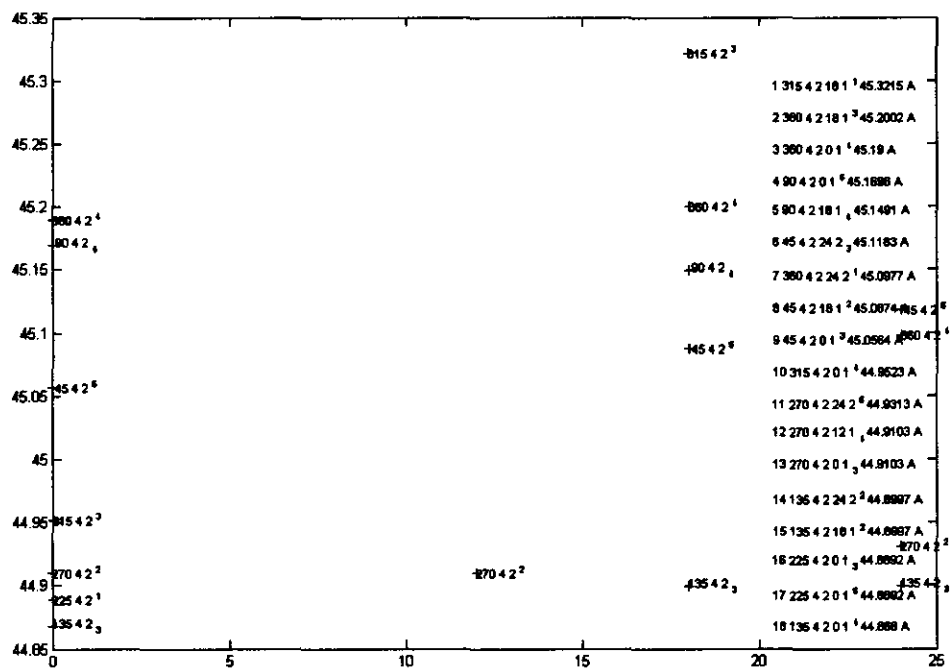


Figure 164: Feature *BPF12_S_X* (ball pass inner race defect frequency on x axis) for medium defect

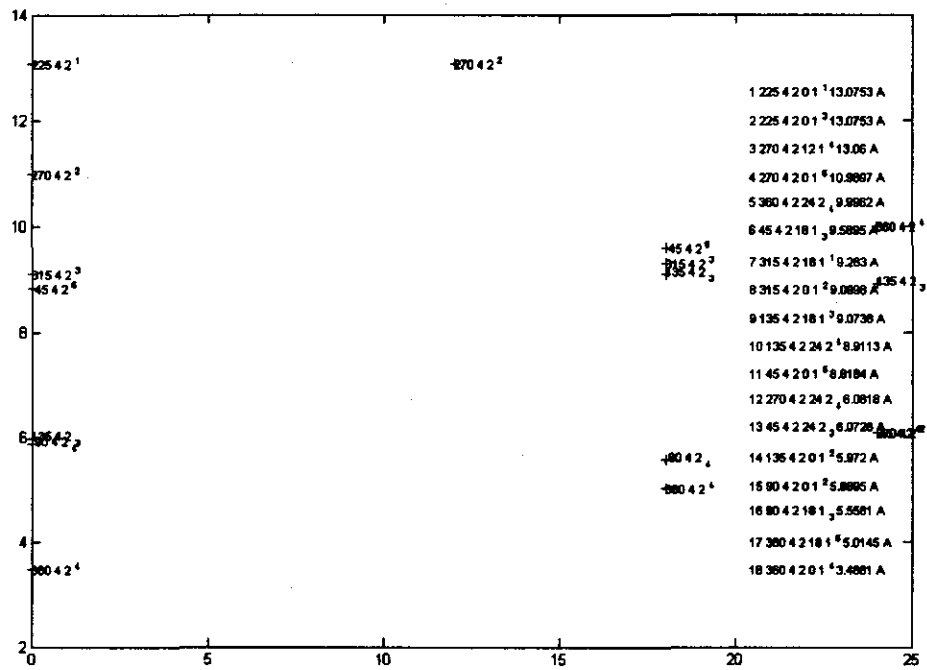


Figure 165: Feature *BPF12_X* (demodulation spectrum value at ball pass inner race defect frequency on x axis) for medium defect

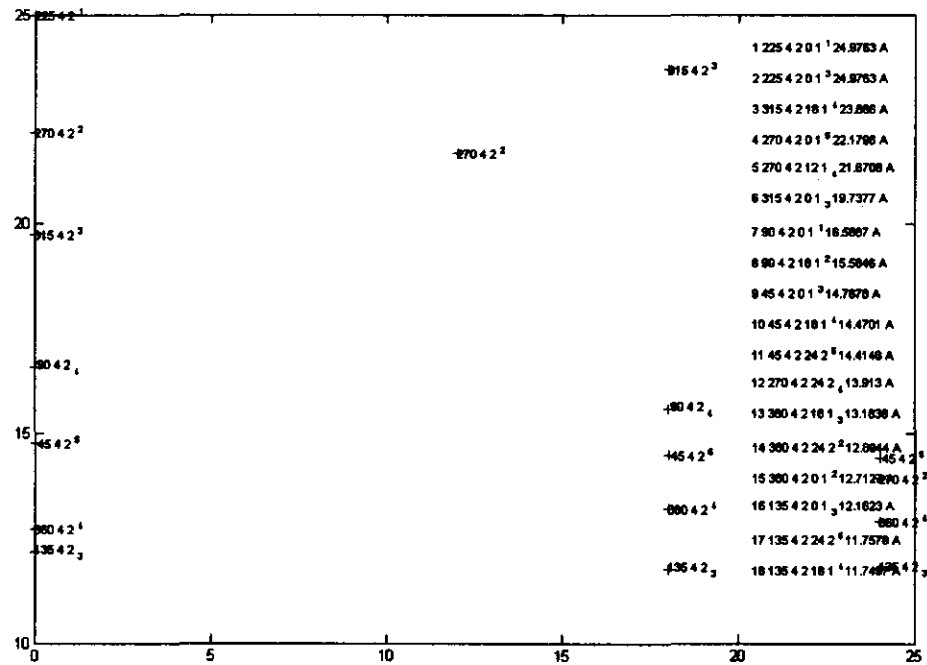


Figure 166: Feature *KH_X* (kurtosis on x axis) for medium defect

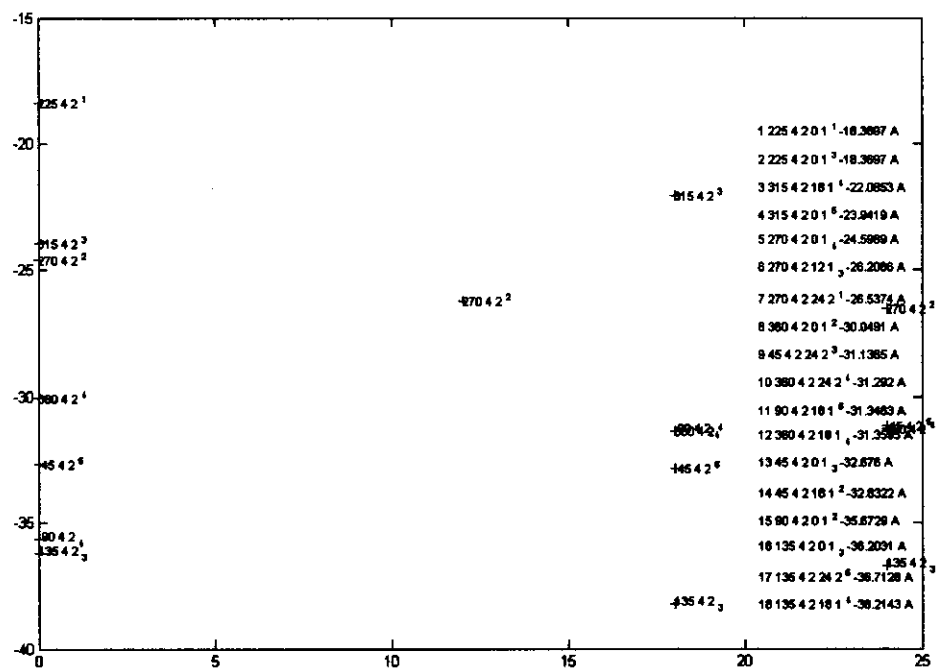


Figure 167: Feature *CH_X* (real cepstrum on x axis) for medium defect

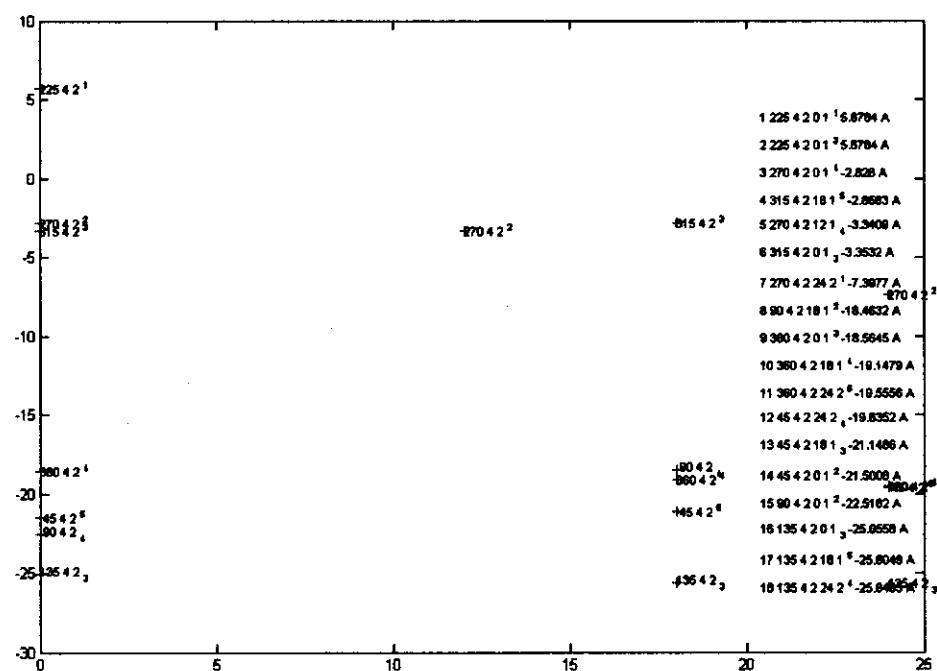


Figure 168: Feature *LH_X* (linear cepstrum on x axis) for medium defect

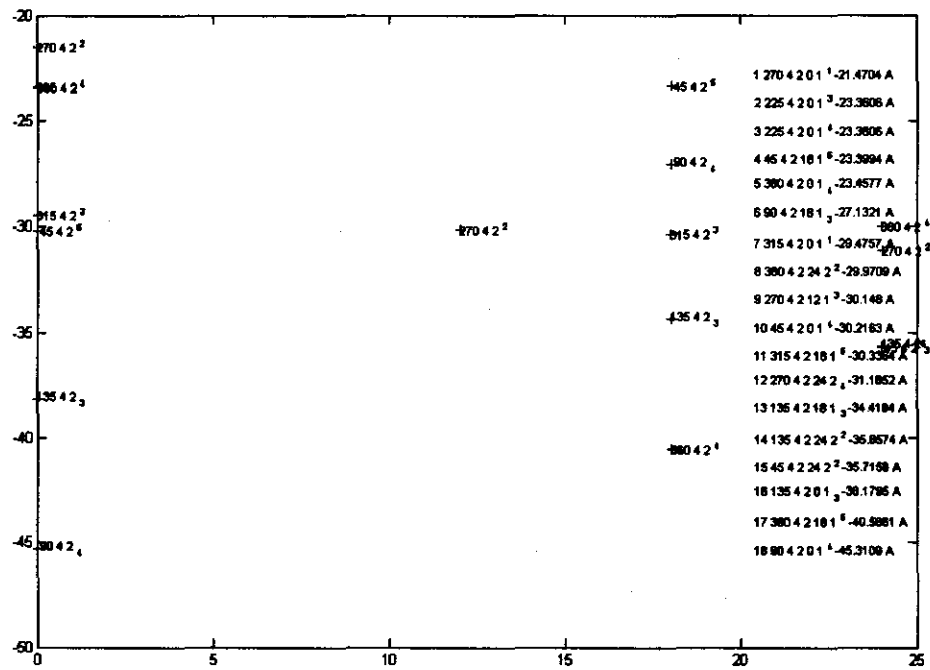


Figure 169: Feature CH_p_X (real cepstrum – after highpass filter – on x axis) for medium defect

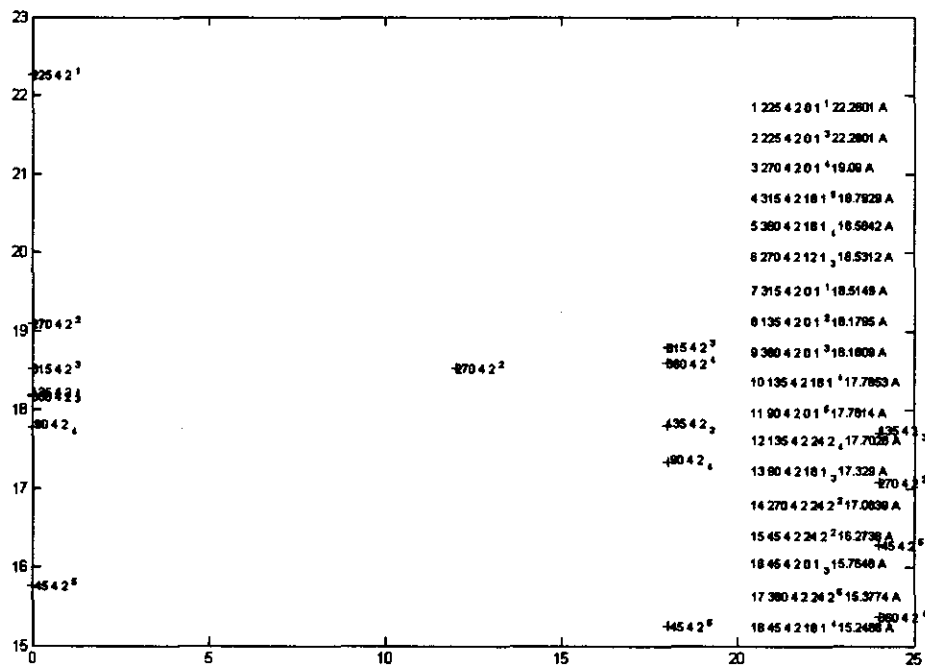
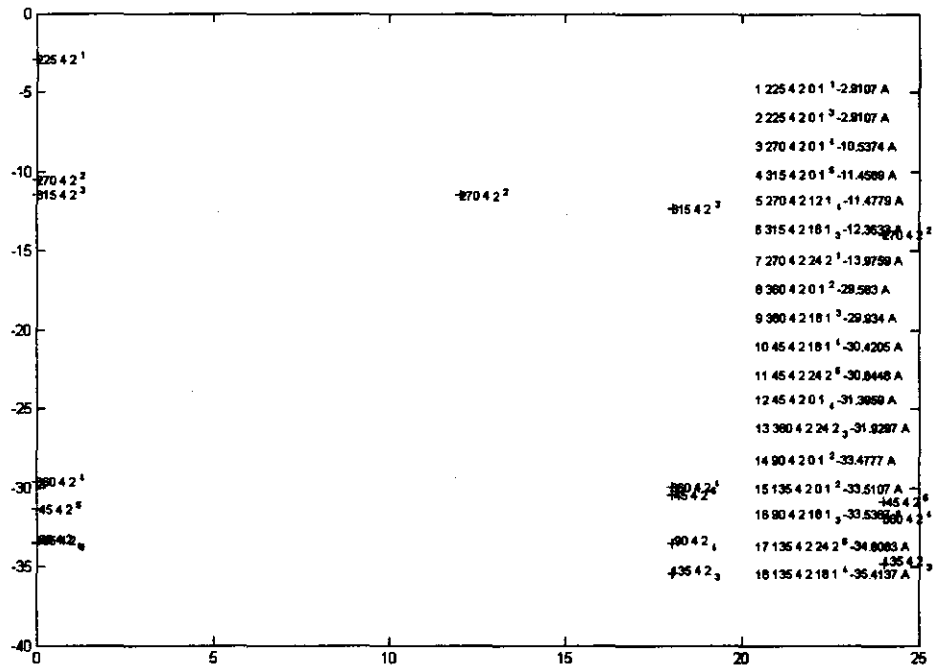


Figure 170: Feature KH_p_X (kurtosis – after highpass filter – on x axis) for medium defect



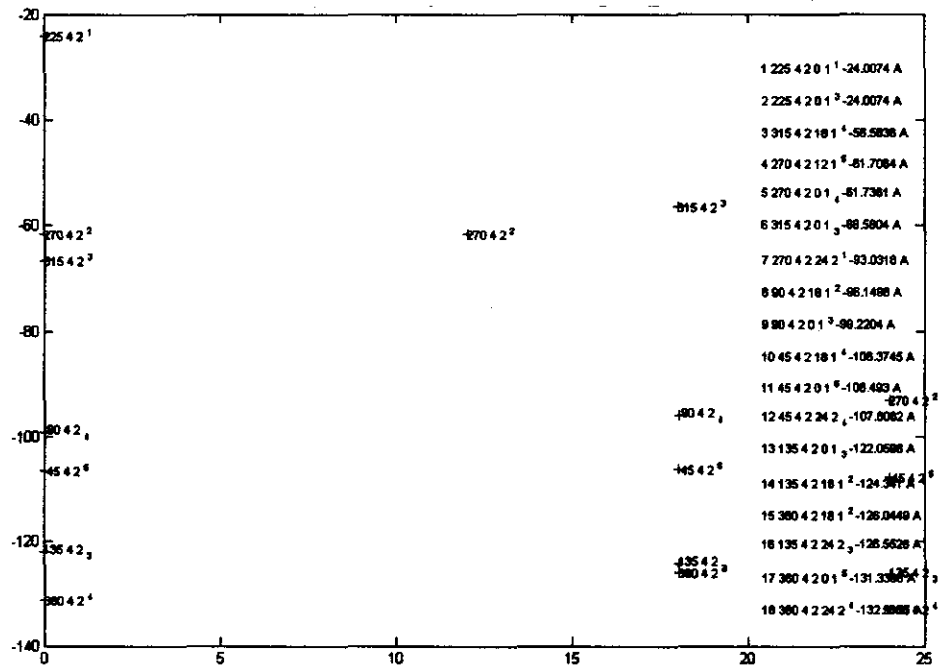


Figure 173: Feature *FFT8_BPFO_X* (magnitude spectrum value of the signal to the 8th power at ball pass outer race defect frequency on x axis) for medium defect

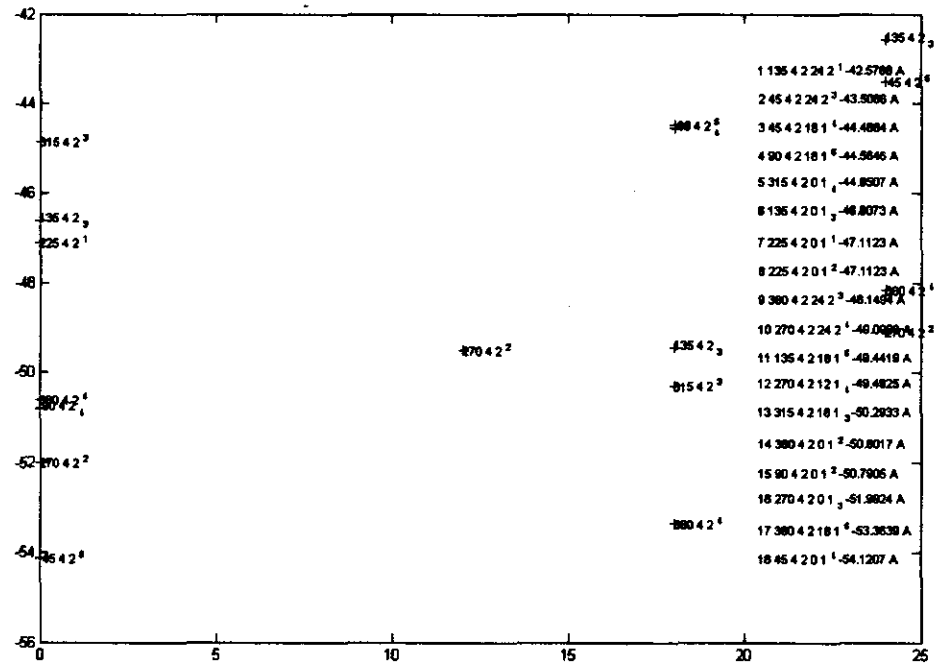


Figure 174: Feature *SA_BPFO2_0_X* (square root of the SOS of synchronous average of the signal at ball pass outer race defect frequency on x axis) for medium defect

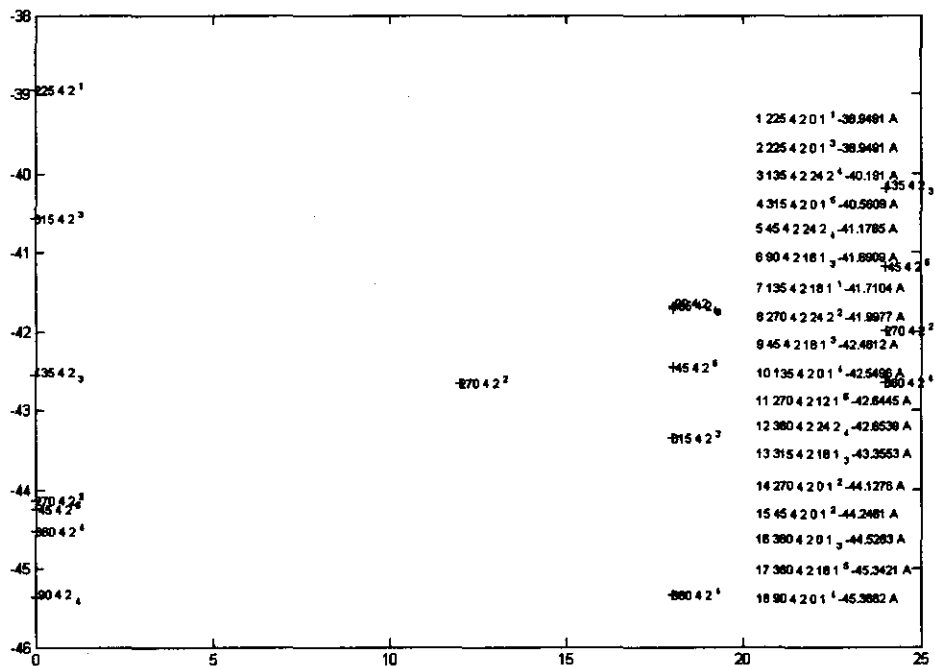


Figure 175: Feature *SA_BPFO2_p_X* (maximum value of synchronous average of the signal at ball pass outer race defect frequency on x axis) for medium defect

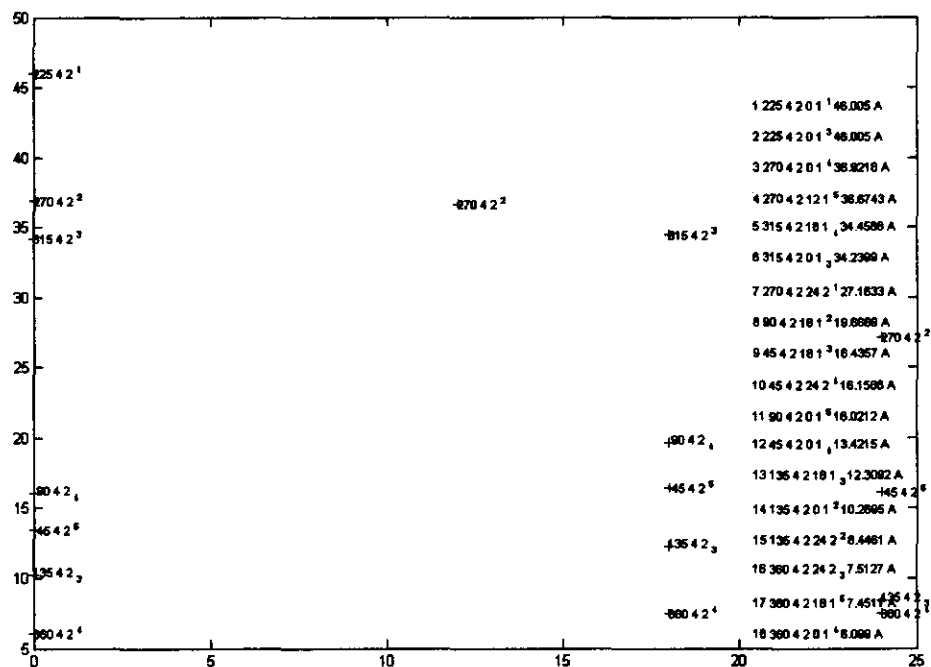


Figure 176: Feature *BPFO2_10_Eng_X* (square root of the SOS of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for medium defect

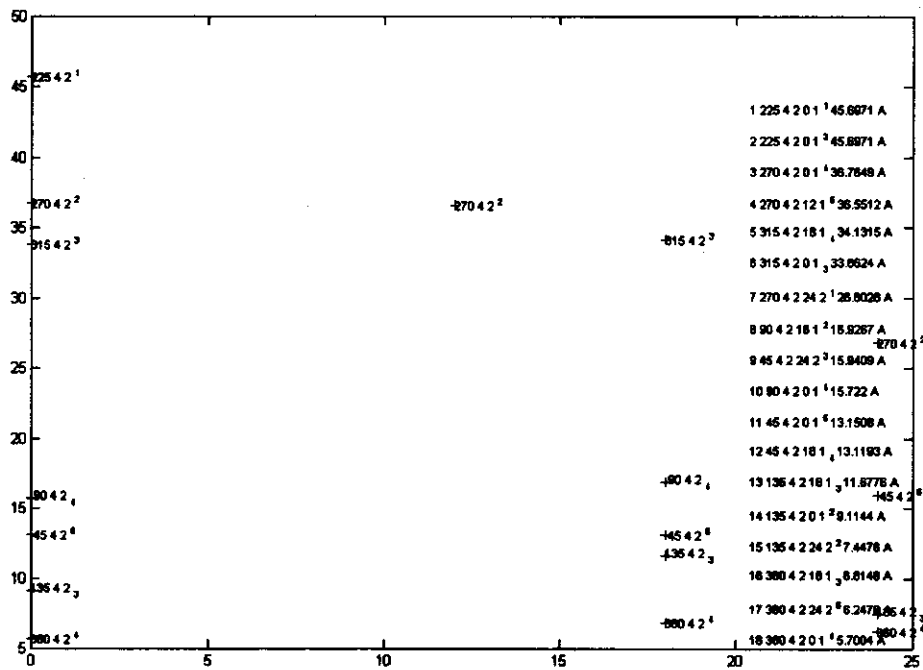


Figure 177: Feature *BPFO2_10_Max_X* (maximum value of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for medium defect

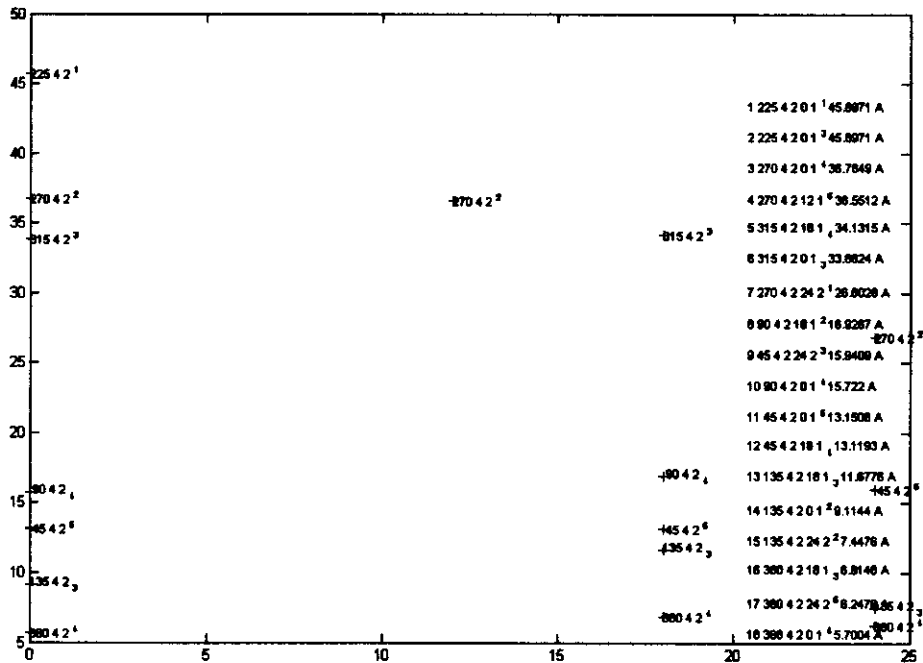
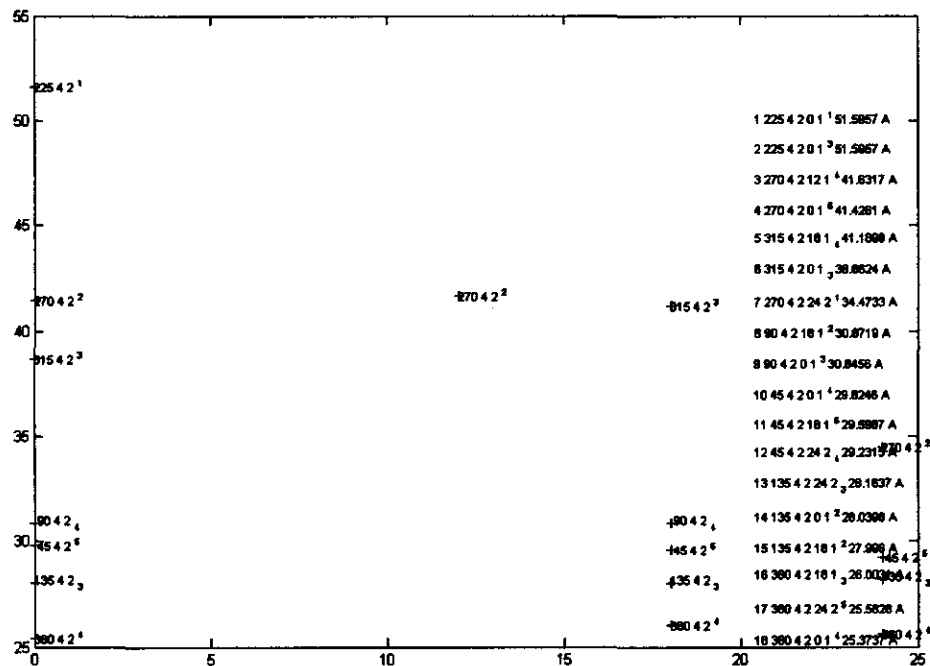
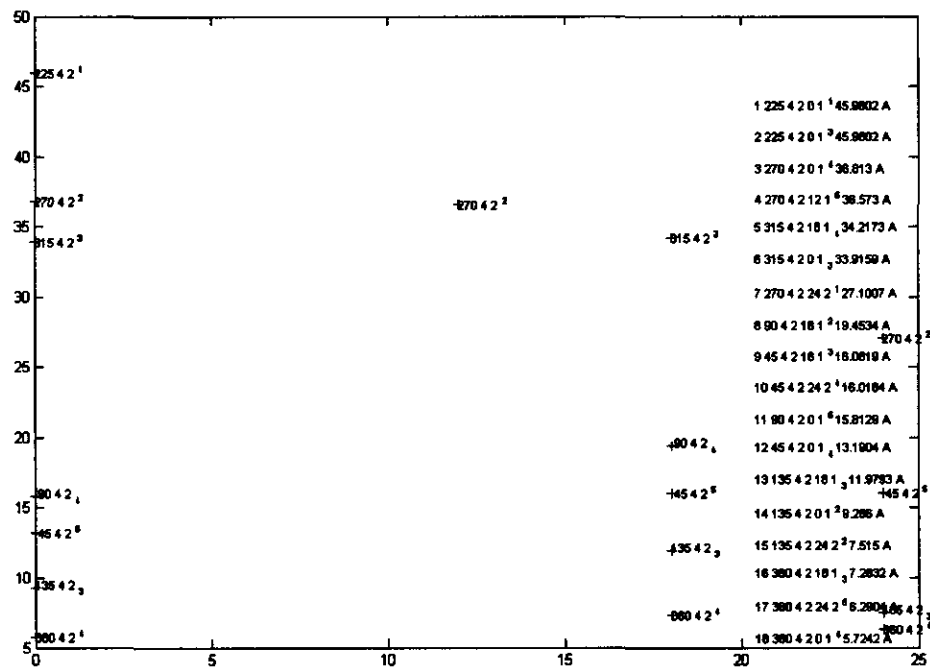


Figure 178: Feature *BPFO2_2_Eng_X* (square root of the SOS of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for medium defect



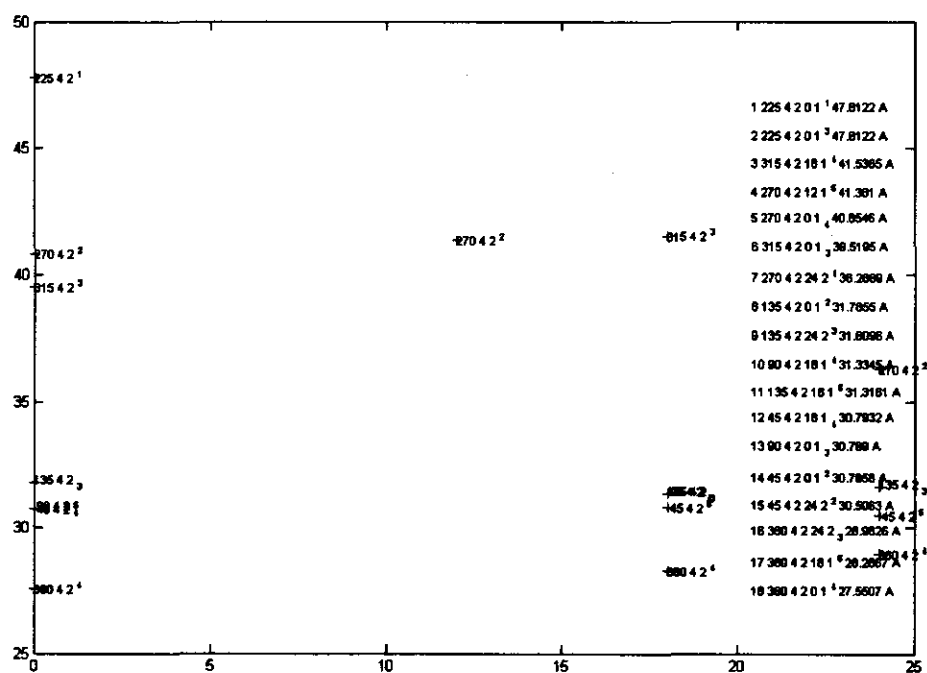


Figure 181: Feature *HFD_2_5_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 2–5kHz – on x axis) for medium defect

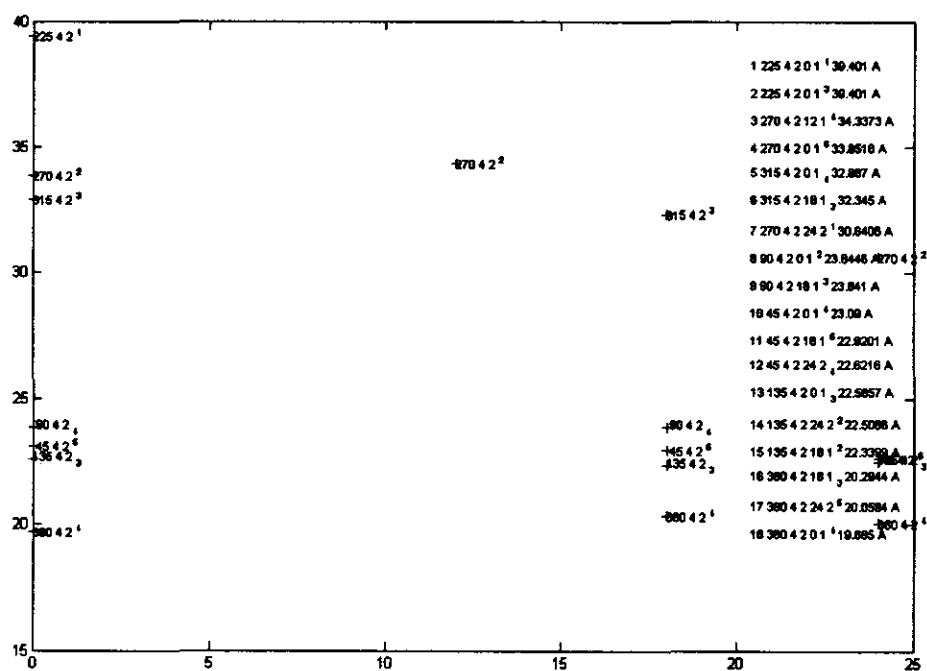


Figure 182: Feature *HFD_5_6_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 5–6kHz – on x axis) for medium defect

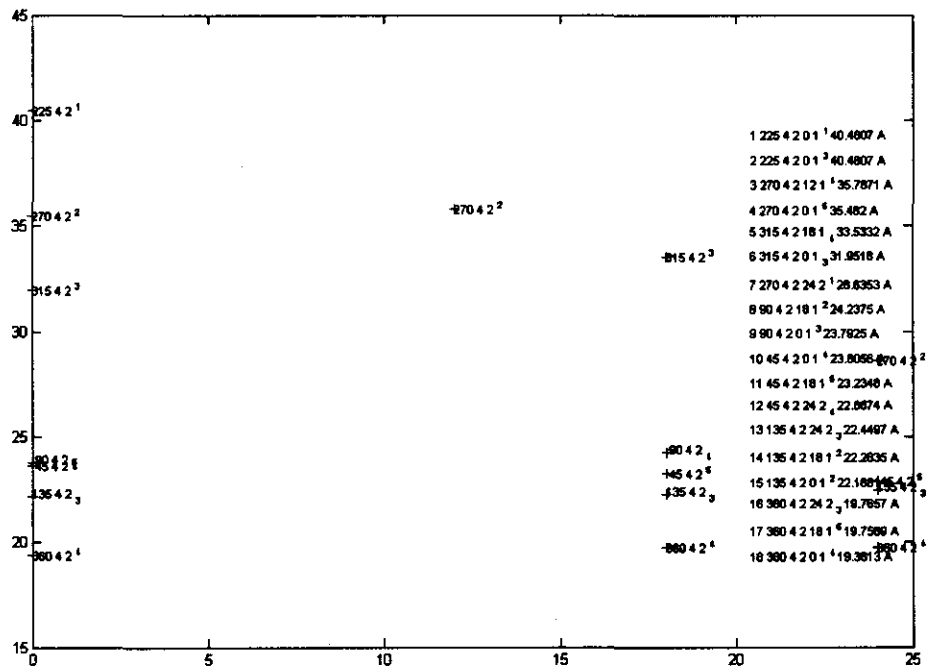


Figure 183: Feature *HFD_6_7_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 6–7kHz – on x axis) for medium defect

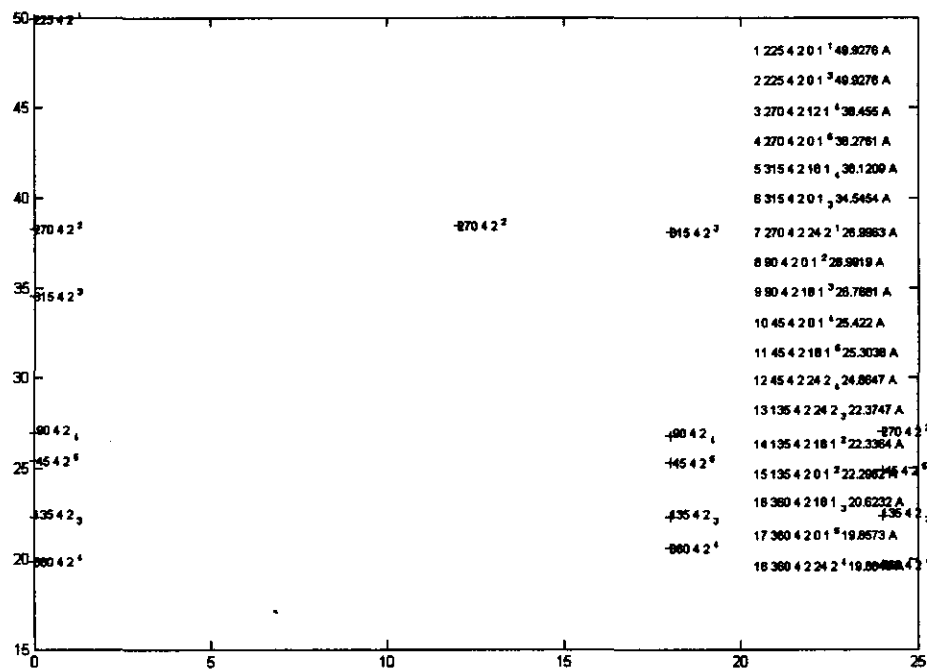


Figure 184: Feature *HFD_8_10_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 8–10kHz – on x axis) for medium defect

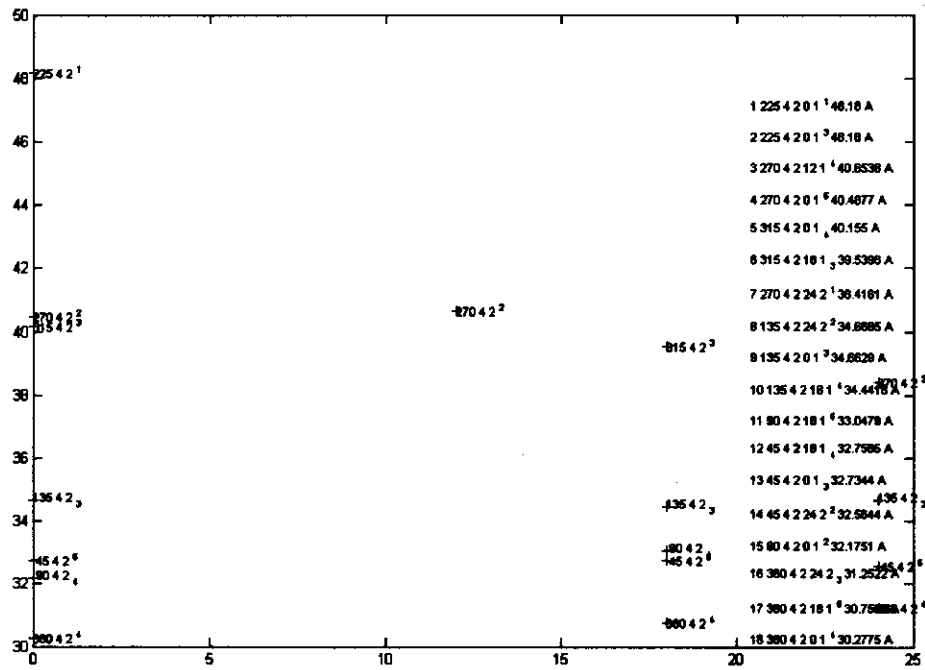


Figure 185: Feature *FFT2_MEAN_X* (first component of demodulation spectrum on x axis) for medium defect

11.3.1.4 Small outer race defect

For each data point indicated by 'x' the defect angle (DA) is shown (180°, 225°, 270°, 315°, 360°, 45°, 90° and 135° for the Vib demo data set 3), the bearing number (1, 4, 5 for the Vib demo data set 3), the bearing defect severity (5, 2, 1 for the Vib demo data set 3 where 5 is the largest bearing defect corresponding to the same bearing used in the Vib demo data set 2) and the distance of the defect from 225° (refer to figure 43 on p. 214 for notational details). The dynamic load (unbalance mass in grams) is shown on the horizontal axis to demonstrate the influence of dynamic loading on bearing defects of decreasing severity and defect angle.

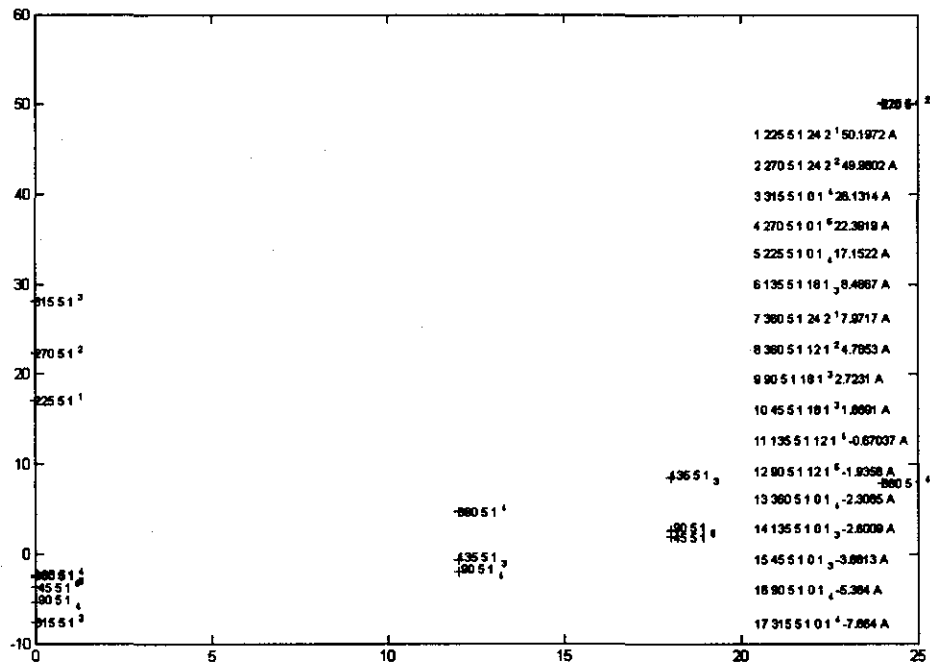


Figure 186: Feature *Fr_X* (magnitude spectrum value at rotational frequency on x axis) for small defect

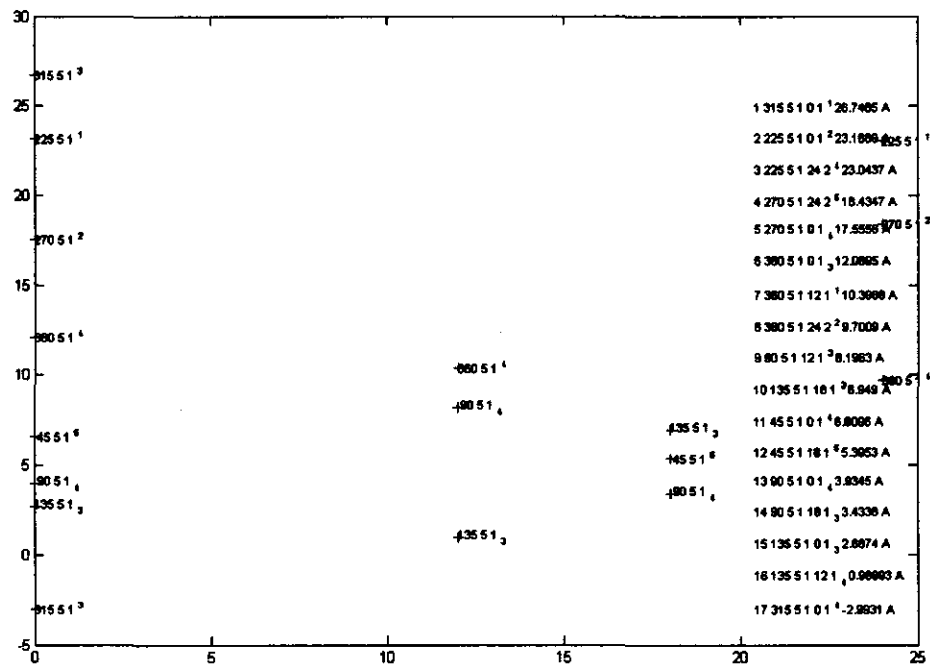


Figure 187: Feature *BPFO_X* (magnitude spectrum value at ball pass outer race defect frequency on x axis) for small defect

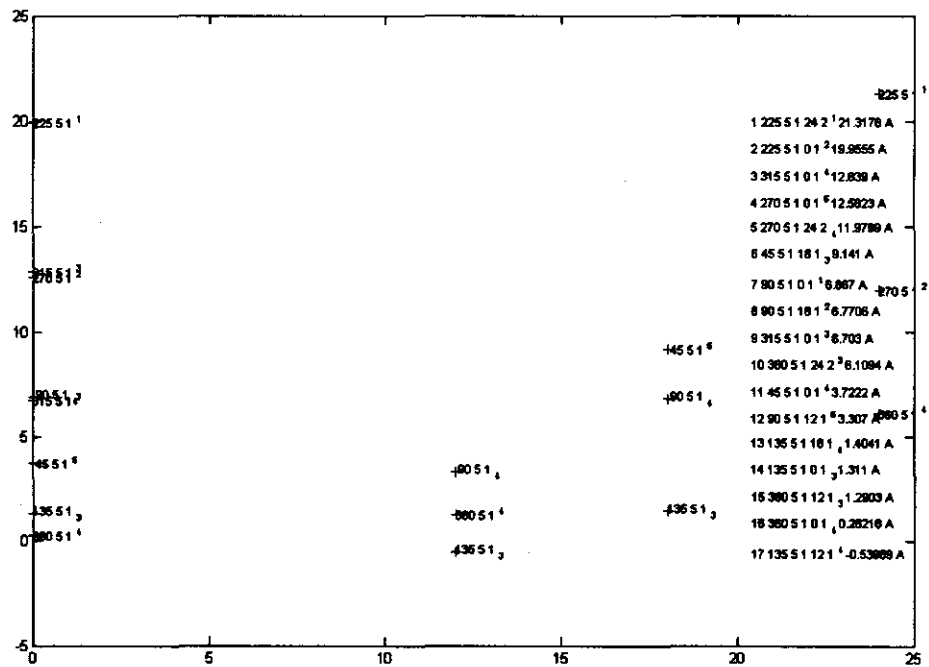


Figure 188: Feature *BPFO2_X* (demodulation spectrum value at ball pass outer race defect frequency on x axis) for small defect

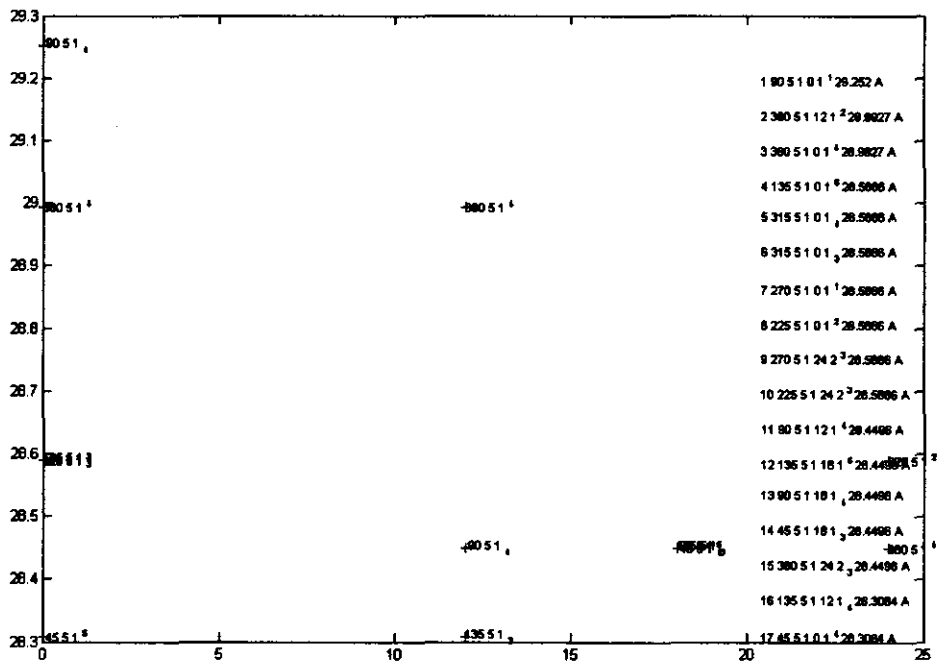


Figure 189: Feature *Fr_S_X* (rotational frequency on x axis) for small defect

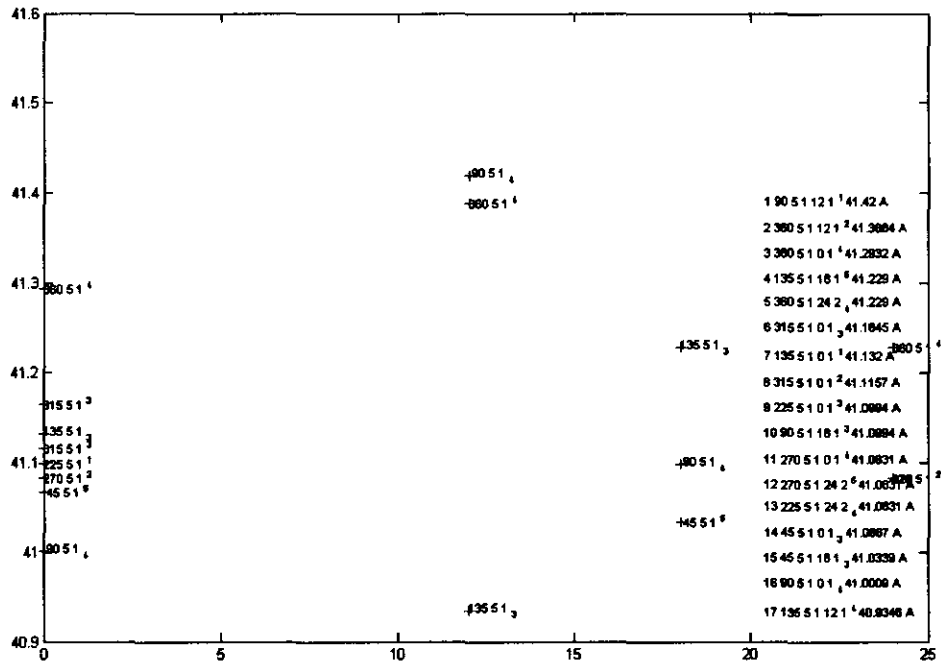


Figure 190: Feature *BPFO2_S_X* (ball pass outer race defect frequency on x axis) for small defect

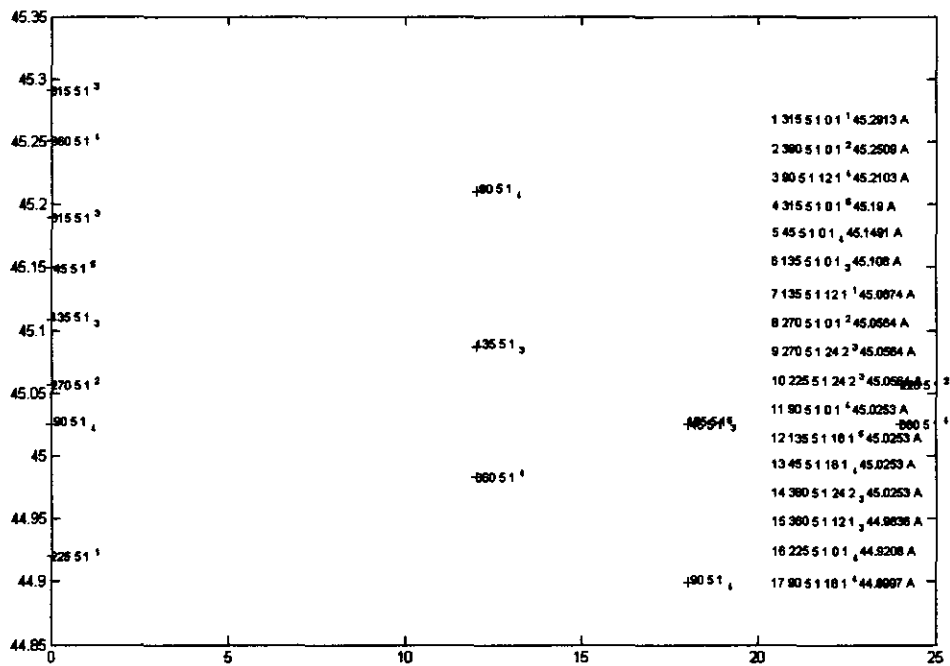


Figure 191: Feature *BPF12_S_X* (ball pass inner race defect frequency on x axis) for small defect

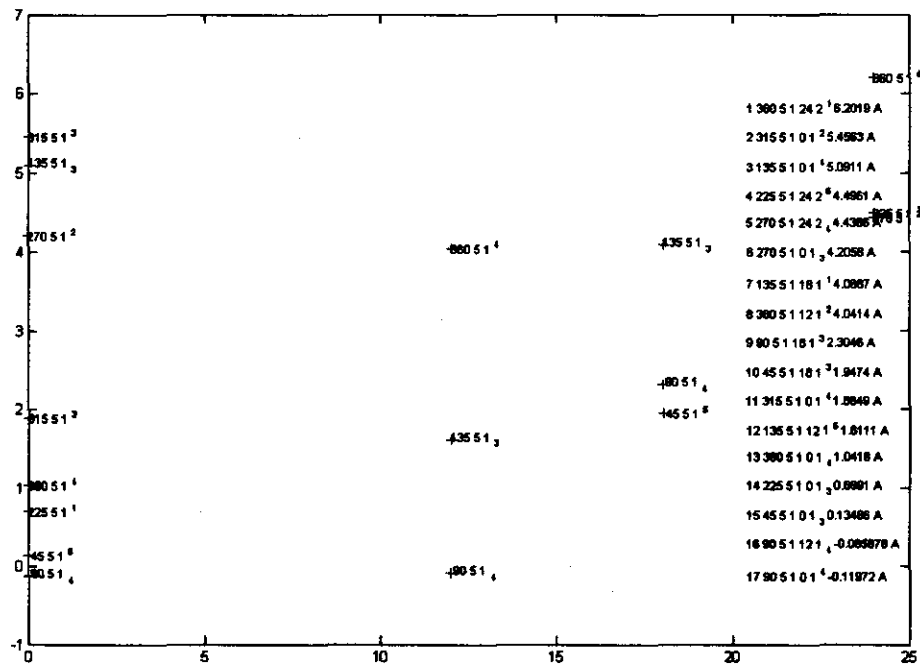


Figure 192: Feature *BPF12_X* (demodulation spectrum value at ball pass inner race defect frequency on x axis) for small defect

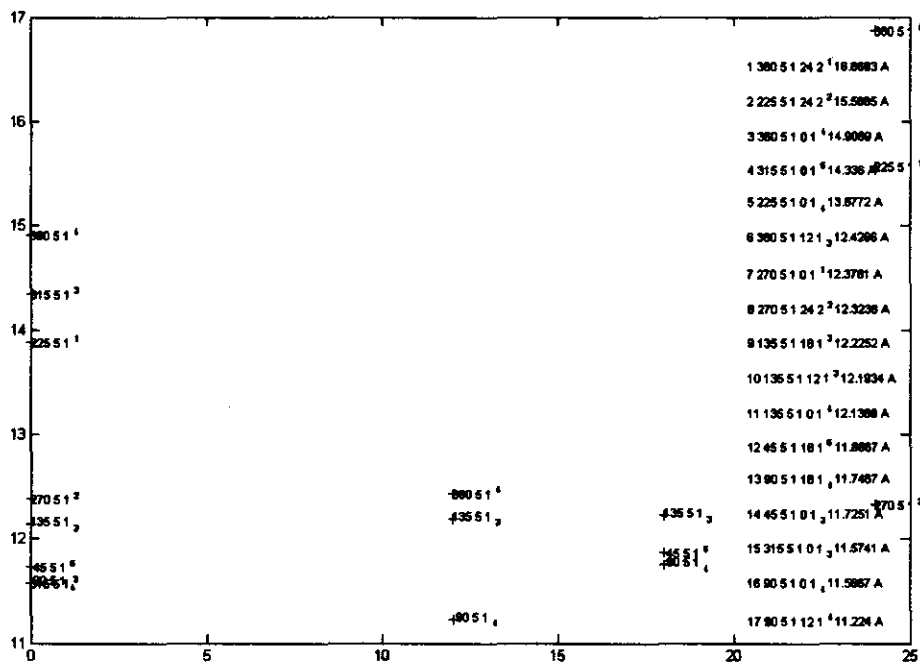


Figure 193: Feature *KH_X* (kurtosis on x axis) for small defect

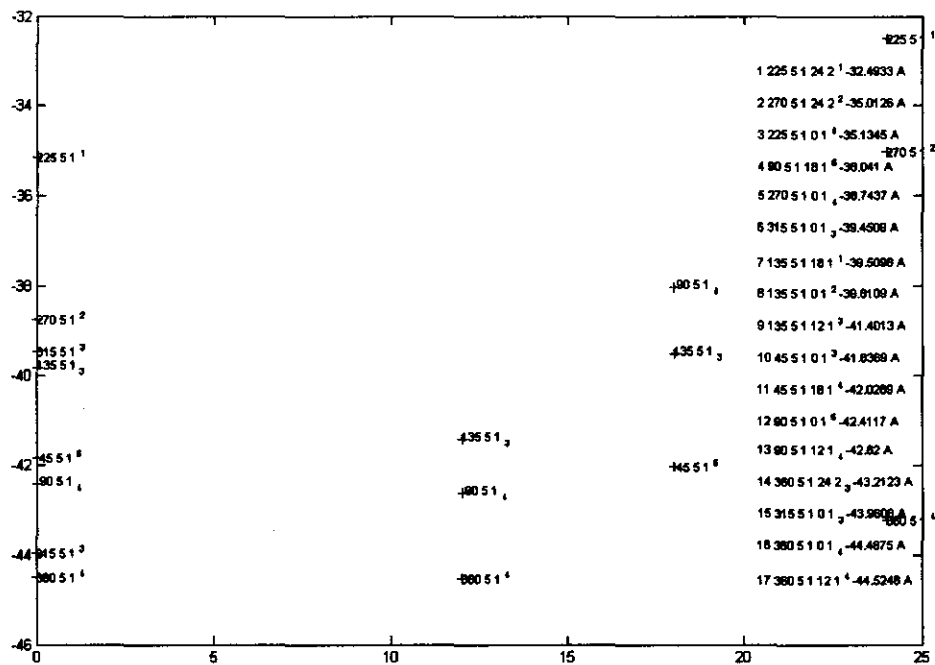


Figure 194: Feature CH_X (real cepstrum on x axis) for small defect

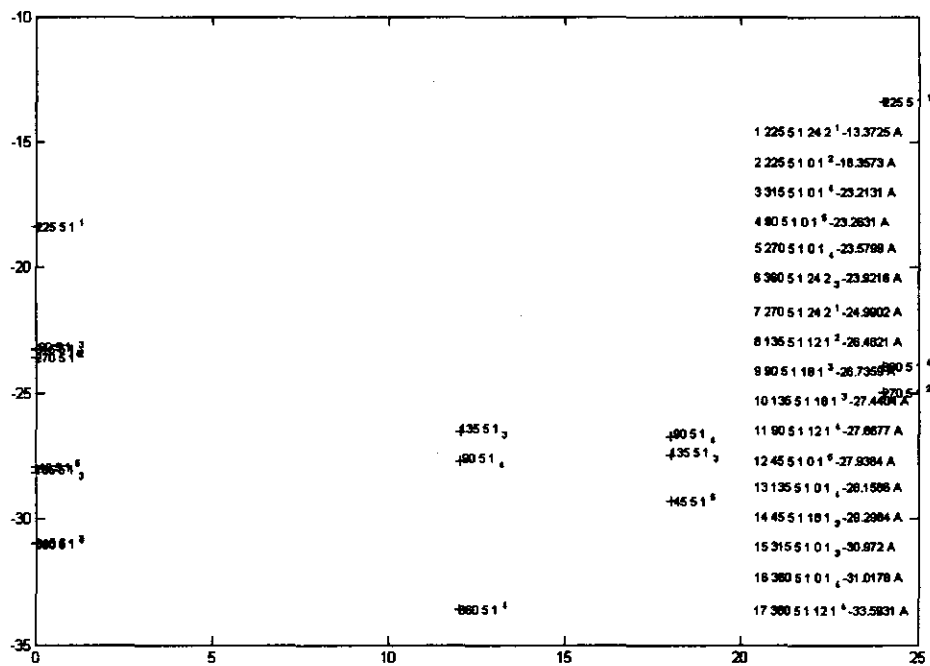


Figure 195: Feature LH_X (linear cepstrum on x axis) for small defect

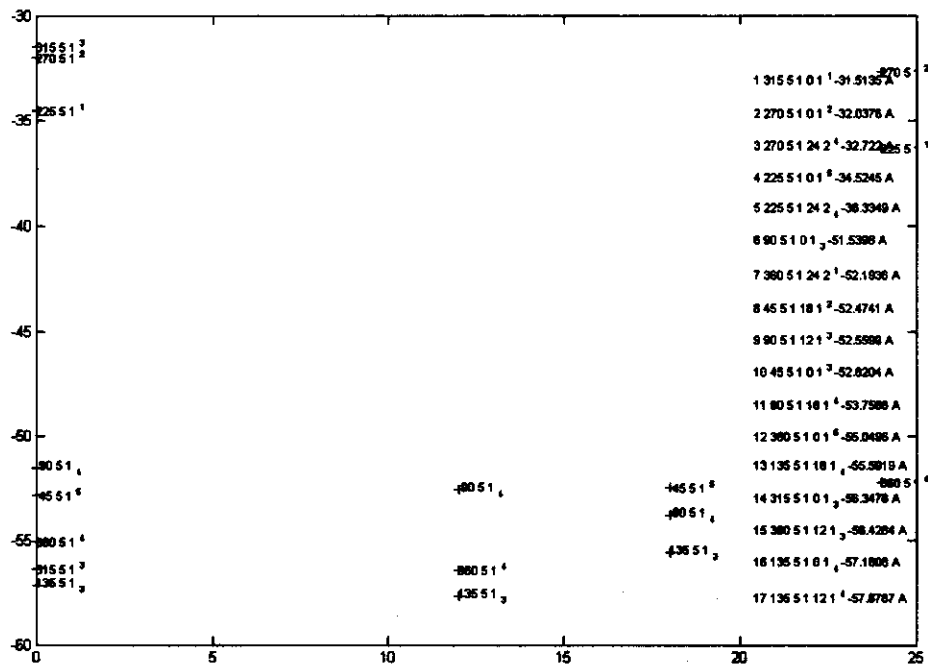


Figure 196: Feature CH_p_X (real cepstrum – after highpass filter – on x axis) for small defect

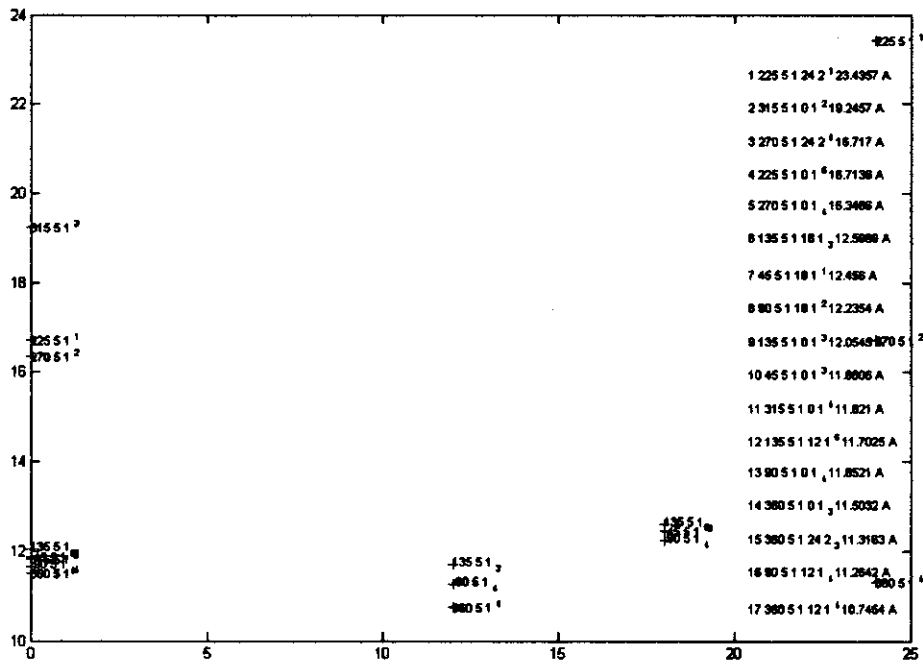


Figure 197: Feature KH_p_X (kurtosis – after highpass filter – on x axis) for small defect

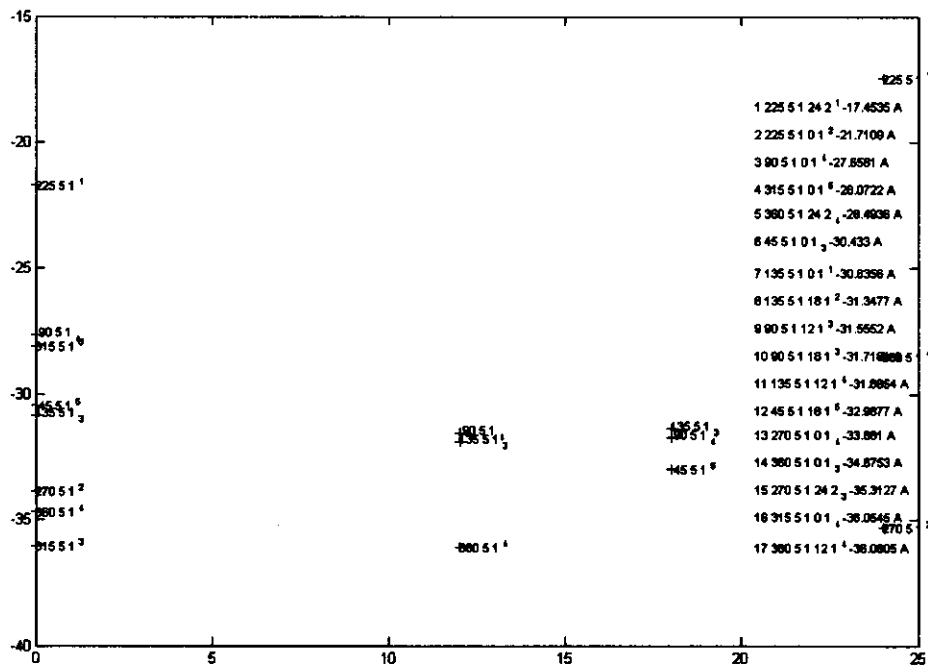


Figure 198: Feature LH_p_X (linear cepstrum – after highpass filter – on x axis) for small defect

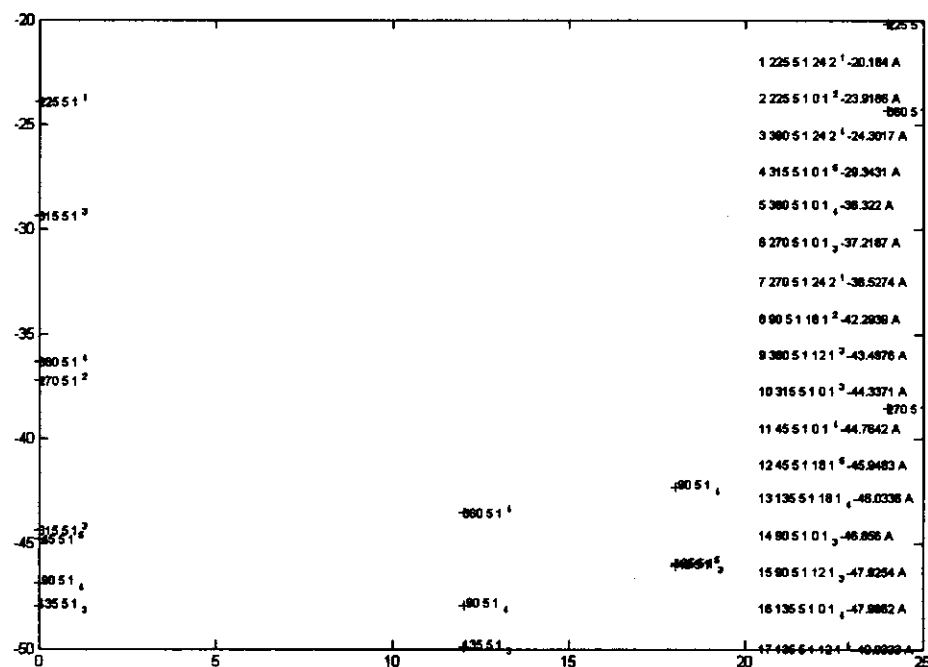


Figure 199: Feature $FFT4\_BPFO\_X$ (magnitude spectrum value of the square of the signal squared at ball pass outer race defect frequency on x axis) for small defect

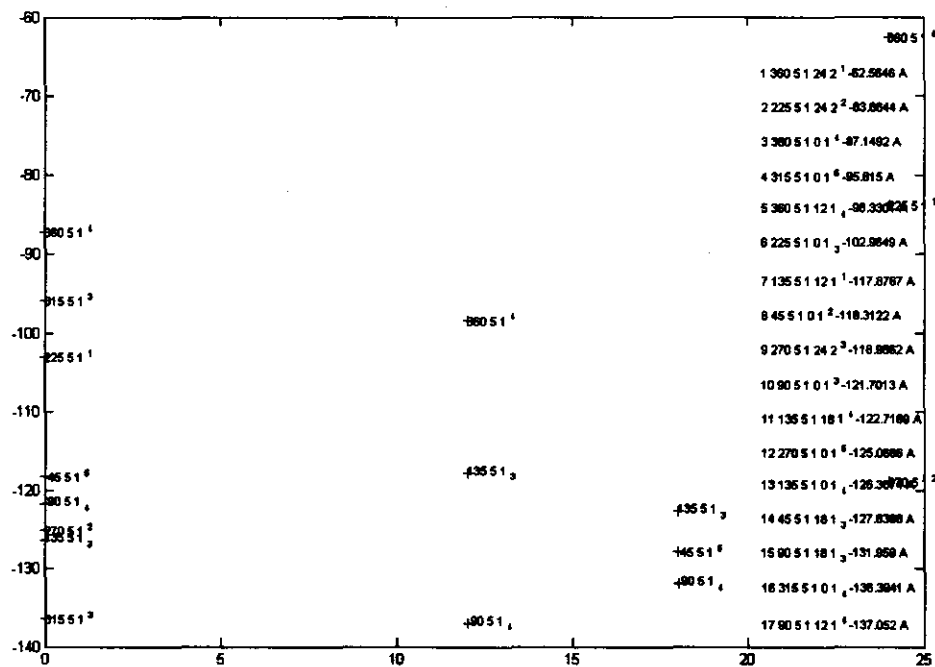


Figure 200: Feature *FFT8_BPFO_X* (magnitude spectrum value of the signal to the 8th power at ball pass outer race defect frequency on x axis) for small defect

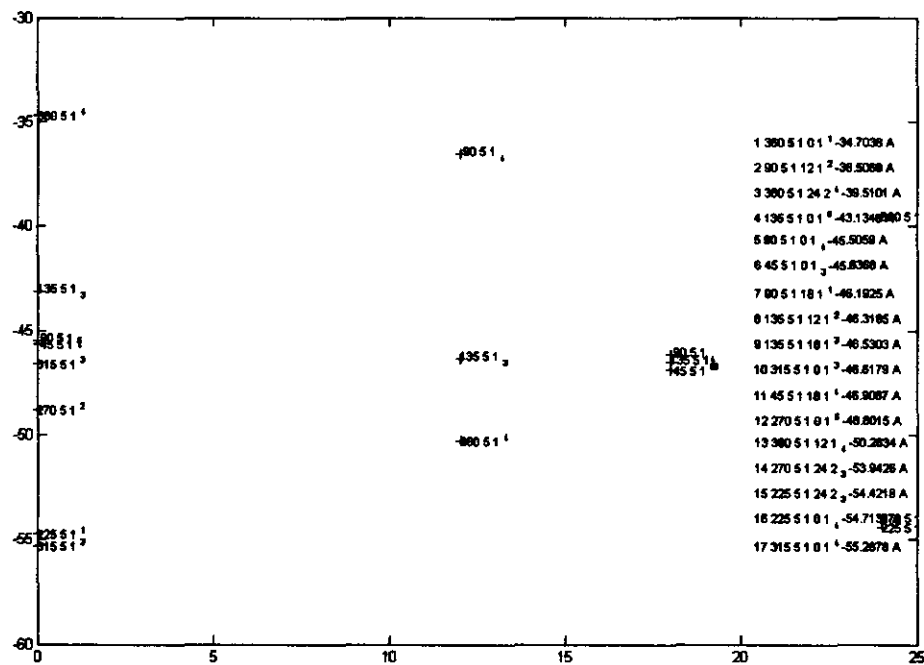


Figure 201: Feature *SA_BPFO2_0_X* (square root of the SOS of synchronous average of the signal at ball pass outer race defect frequency on x axis) for small defect

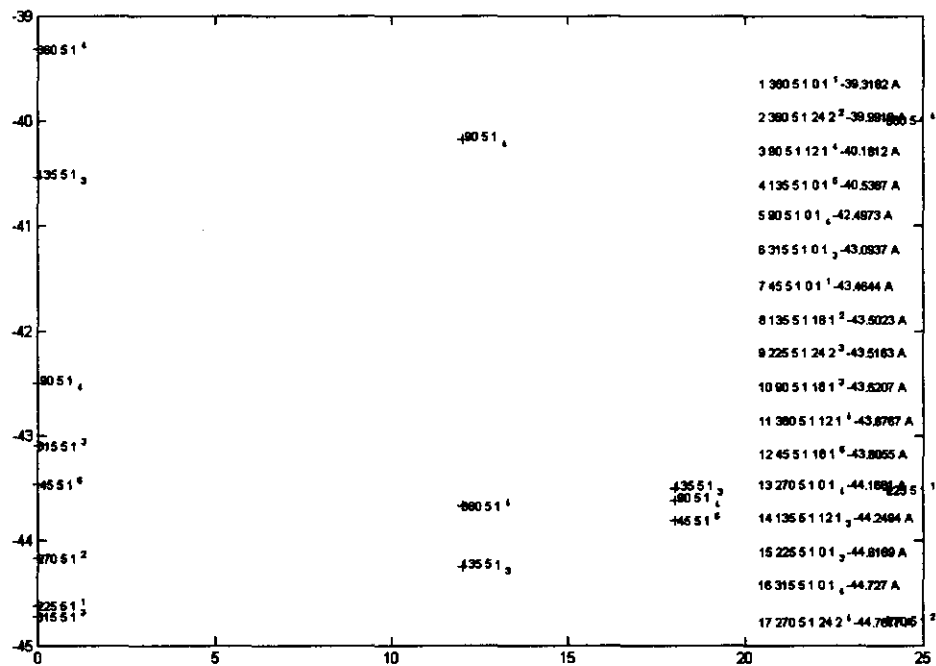


Figure 202: Feature *SA_BPFO2_p_X* (maximum value of synchronous average of the signal at ball pass outer race defect frequency on x axis) for small defect

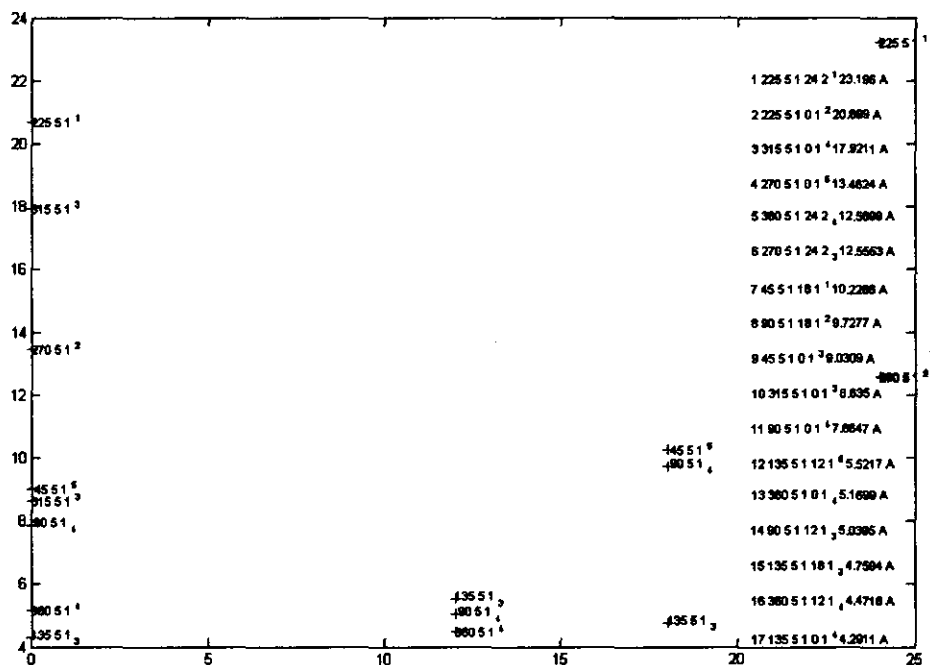


Figure 203: Feature *BPFO2_10_Eng_X* (square root of the SOS of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for small defect

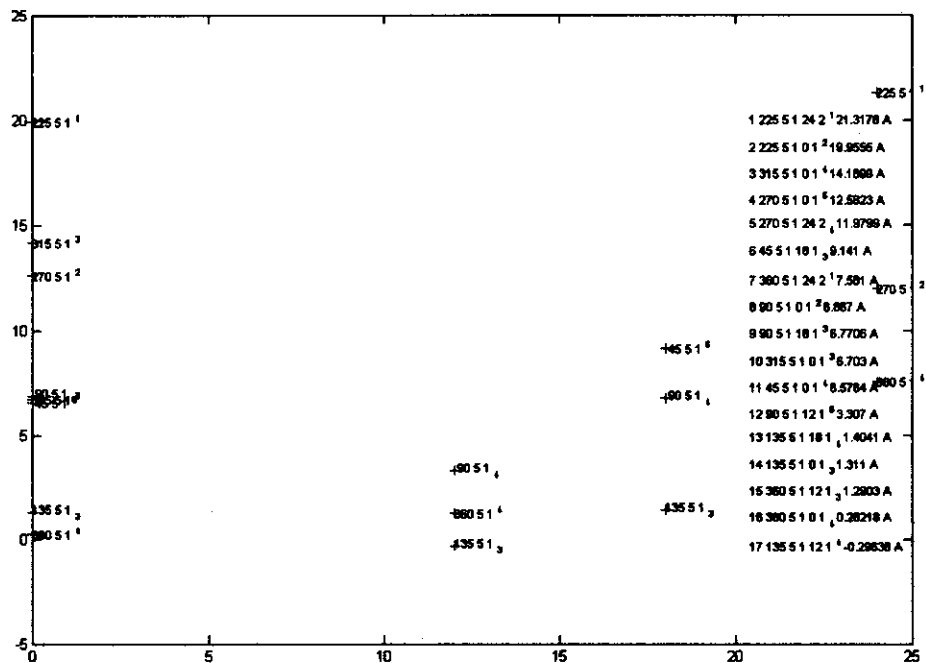


Figure 204: Feature *BPFO2_10_Max_X* (maximum value of the first ten components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for small defect

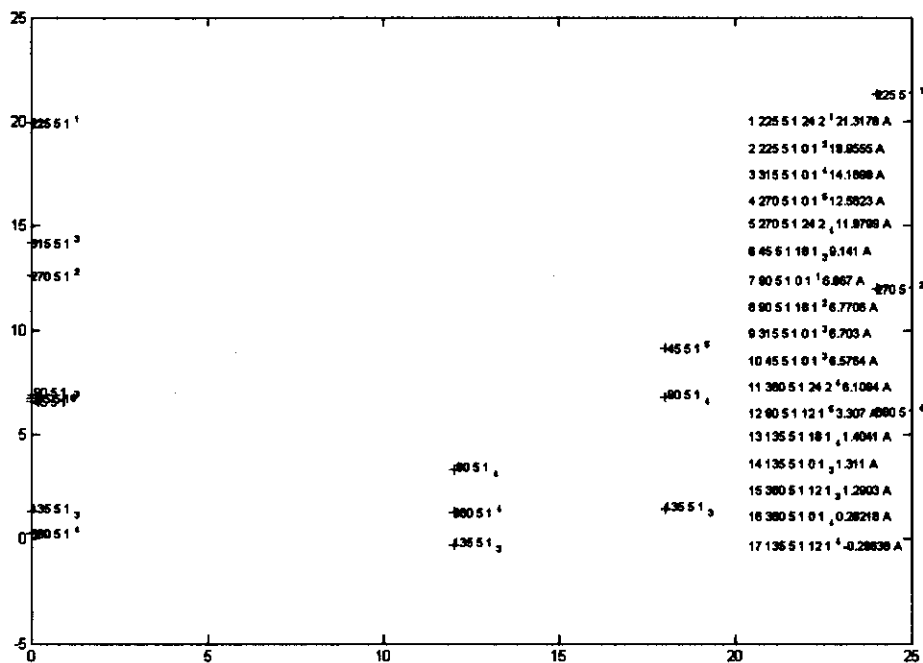


Figure 205: Feature *BPFO2_2_Eng_X* (square root of the SOS of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for small defect

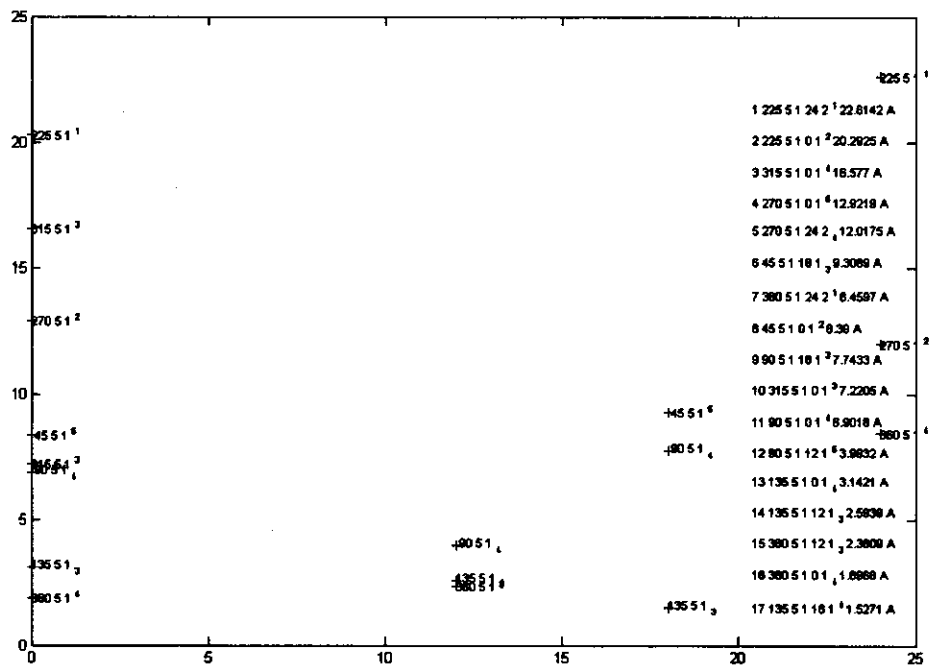


Figure 206: Feature *BPFO2_2_Max_X* (maximum value of the first two components in demodulation spectrum – at multiples of the ball pass outer race defect frequency on x axis) for small defect

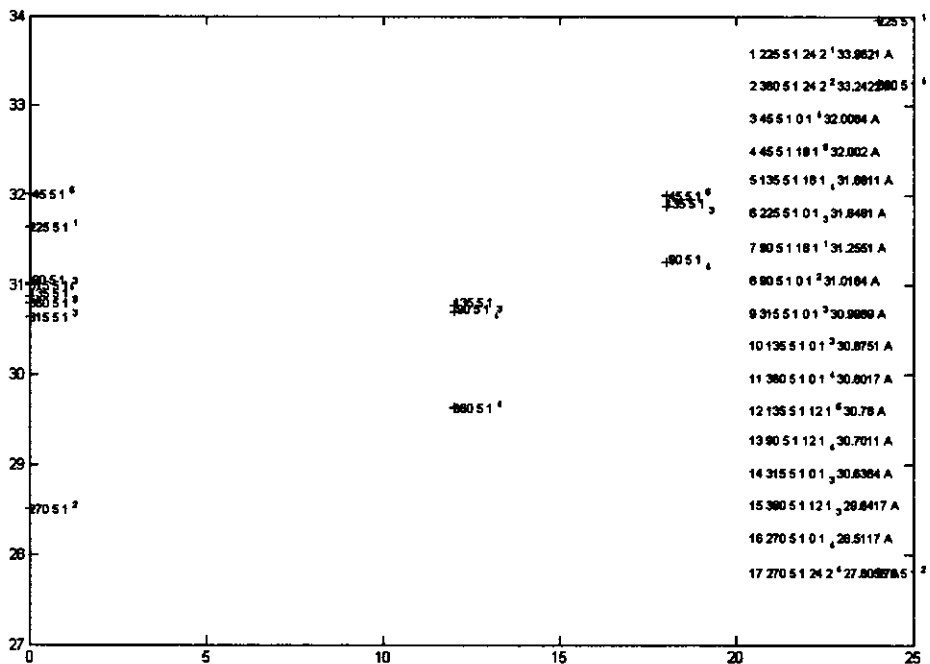


Figure 207: Feature *HFD_5_10_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 5-10kHz – on x axis) for small defect

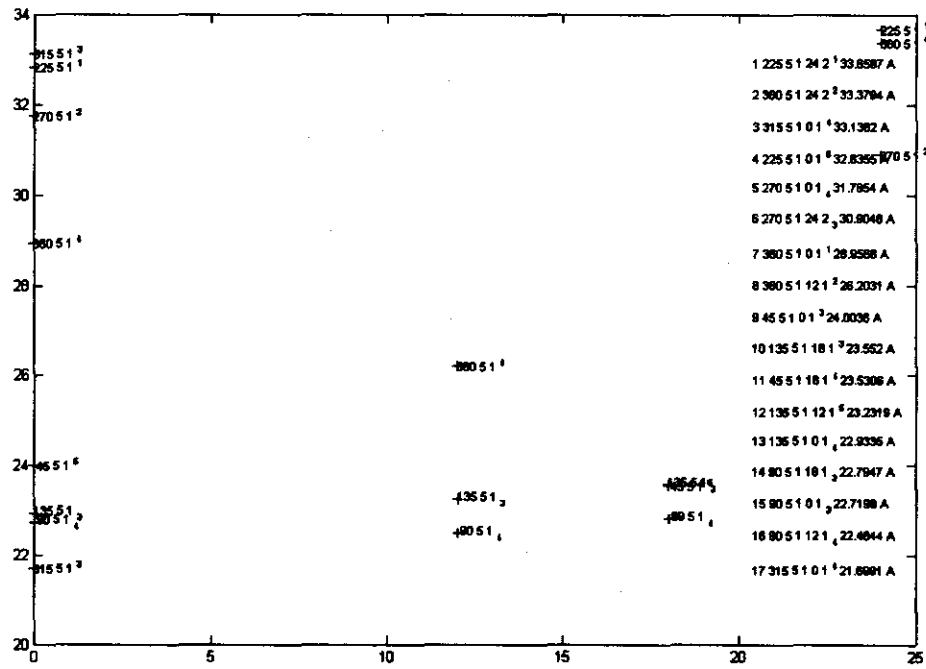


Figure 208: Feature *HFD_2_5_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 2–5kHz – on x axis) for small defect

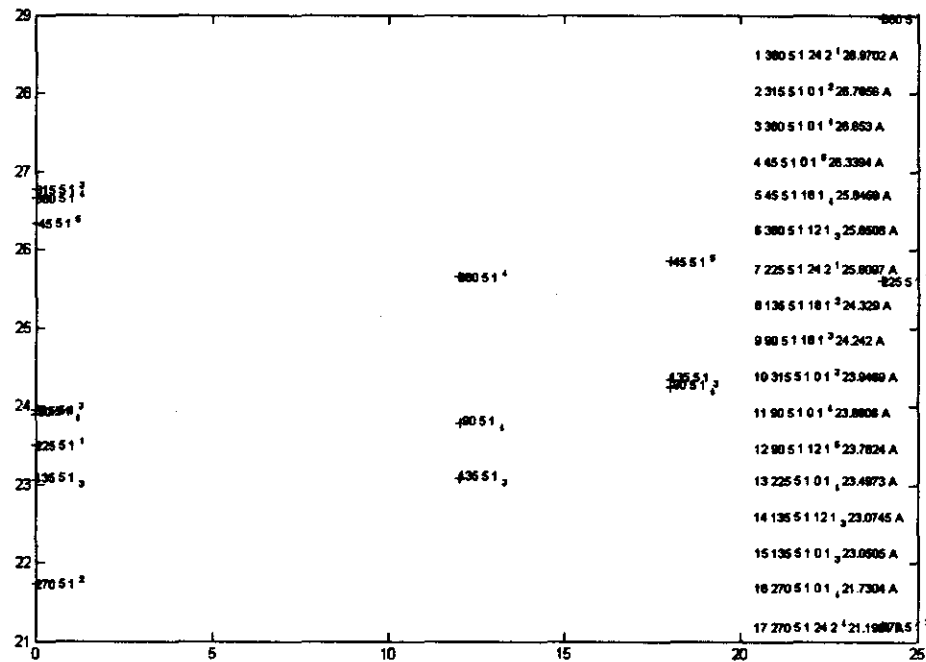


Figure 209: Feature *HFD_5_6_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 5–6kHz – on x axis) for small defect

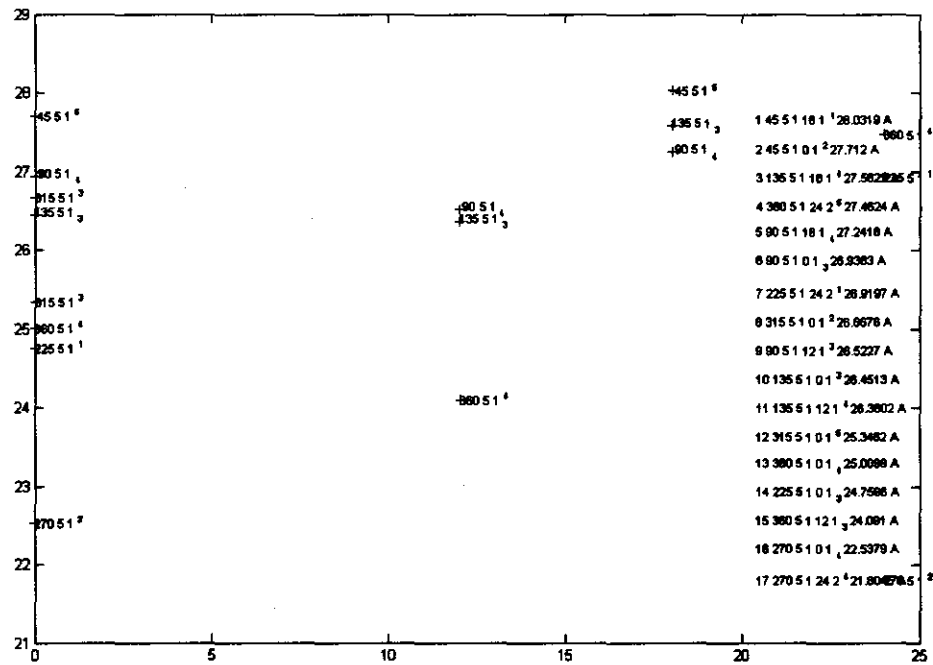


Figure 210: Feature *HFD_6_7_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 6–7kHz – on x axis) for small defect

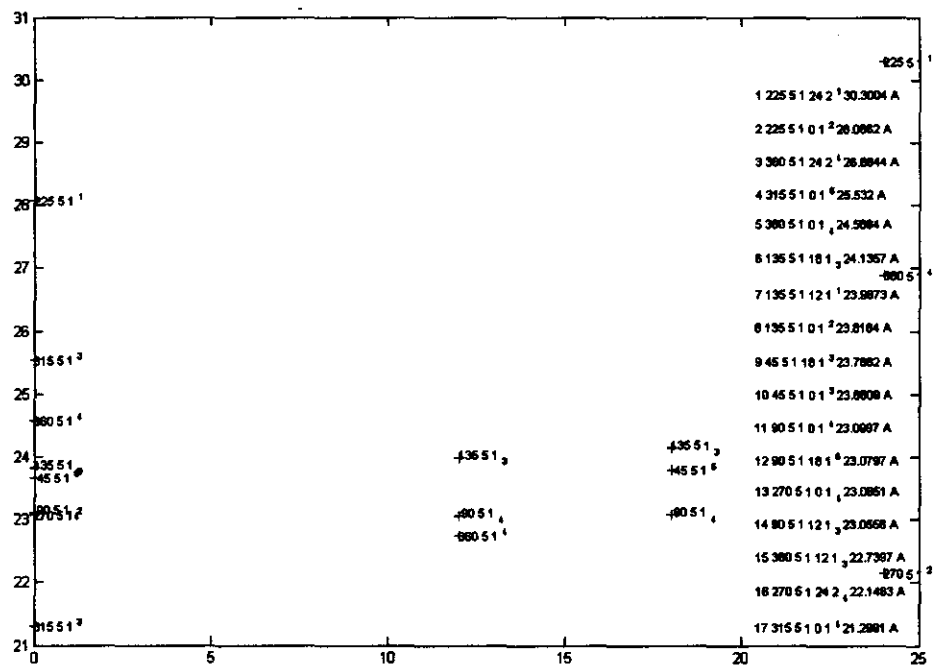


Figure 211: Feature *HFD_8_10_kHz_X* (square root of SOS of HFD components in magnitude spectrum – from 8–10kHz – on x axis) for small defect

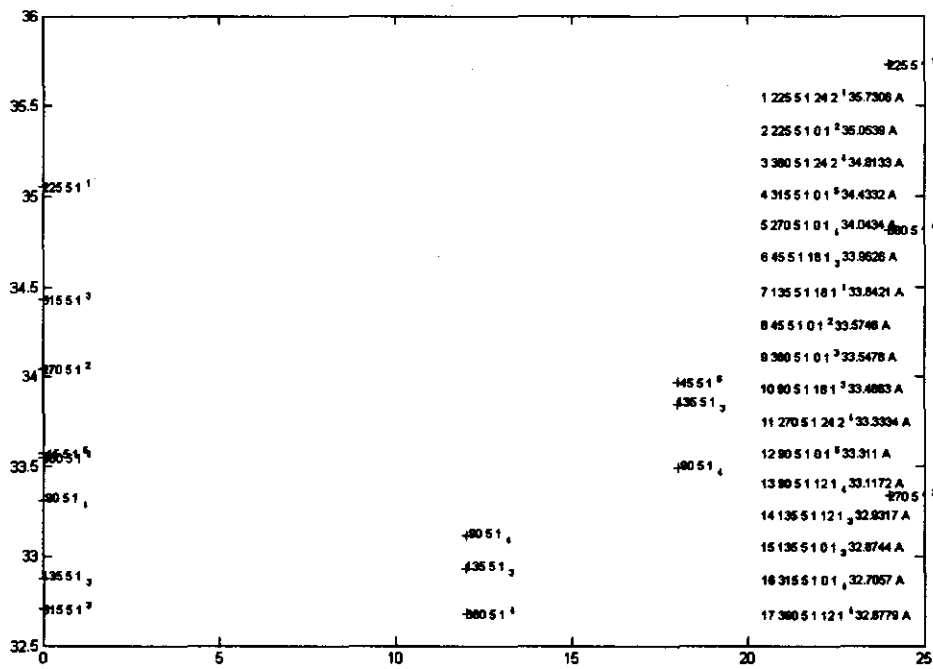


Figure 212: Feature *FFT2_MEAN__X* (first component of demodulation spectrum on x axis) for small defect

11.5 Appendix E: Influence of periodic speed variations

Brie [10] discusses the theoretical influence of speed variations on the spectrum of a signal. Let the sampling signal with variation be given by

$$x = \sum_{k=-\infty}^{\infty} \delta(t - kT_0 - t_k) \quad (129)$$

The autocorrelation is then given by [10]

$$r_{xx}(\tau) = \sum_{k=-\infty}^{\infty} \delta(t) + \sum_{k=-\infty}^{\infty} g(t - kT_0) \quad (130)$$

where $g(\tau)$ represents the convolution of the PDF $p(\tau)$ of the 'timing jitter' with itself

$$g(\tau) = \int_{-\infty}^{\infty} p(t)p(\tau - t)dt \quad (131)$$

If $p(\tau)$ is assumed uniform, the function $g(\tau)$ would be triangular.

The Fourier transform of the autocorrelation function supplies $R_{xx}(\omega)$.

The autocorrelation $r_{xx}(\tau)$ is given by

$$r_{xx}(\tau) = \int_{-\infty}^{\infty} \delta(t) - g(t) \otimes \sum_{k=-\infty}^{\infty} \delta(\tau - kT_0) \quad (132)$$

where the symbol $\otimes$ denotes the convolution operation.

$R_{xx}(\omega)$ is given by

$$R_{xx}(\omega) = 1 - \sum_{k=-\infty}^{\infty} G(\omega = k * 2\pi / T_0) \quad (133)$$

11.6 Appendix F: Some condition monitoring data sets in literature for fault identification

Table 15: Some literature references for fault identification

Authors	Fault mechanisms investigated	Features evaluated
Paya <i>et al.</i> [69]	6 faults (gear and bearing)	20 features (wavelets)
McCormick, A. C. & Nandi, A. K. [46]	NO (Normal), IR, OR, RE	Moments (mean, variance, skewness, kurtosis); preprocessing applied: differentiation, integration, lowpass filtering, highpass filtering
Ypma & Petteri [103]	Bearing failures, imbalance, normal condition	10 NN (nearest neighbour) classifier
Ypma <i>et al.</i> [102]	Inner race bearing defect, imbalance and normal condition	Wavelet decomposition level 5 with Hilbert transform for demodulation; neural network with 5 and 10 hidden units and second- order LM (Levenberg- Marquard) learning algorithm; nearest neighbour algorithm and ASSOM
Hines <i>et al.</i> [34]	Shaft mis-alignment and bearing lifetime	

Authors	Training set size	Accuracy obtained
Paya <i>et al.</i> [69]	48 (8 measurements for a fault)	Guess 90% accuracy on test set; no comparison with other techniques
McCormick, A. C. & Nandi, A. K. [46]	NO (Normal), IR, OR, RE. Train (8) test (16) – 2 seconds @ 12kHz	Refer to article for details
McCormick, A. C. & Nandi, A. K. [31, 32, 43]	Fault conditions: NO, NW, IR, OR, RE, CA. 16 shaft speeds and 10 repetitions for 6 fault conditions = 960 data points.	Refer to [31, 32, 43] for details. Classification rate success from 83% to 97% reported in [31].
McCormick, A. C. & Nandi, A. K. [44]	N-N, N-R, W-N, W-R. 4 Shaft speeds for train set and 7 shaft speeds for test set. Buffer length 24000 samples.	Test set classification rates of 65.6% to 99.43% reported (according to feature set).
Ypma & Petteri [103]	400 samples, split randomly into a training and test set; different sets (of different dimensionality) were constructed by varying the resolution of the spectral estimation method (32, 64 and 128 bins)	Refer to article for details; accuracies above 90%, sometimes 100% reported – i.e. perfect separation
Ypma <i>et al.</i> [102]	300 samples for 3 classes; training set (150); test set (150)	Refer to article for details; 99% reported using nearest-neighbour and neural network algorithms; the SOM technique also performs well

11.7 Appendix G: Derivation of amplitude of sidelobe levels relative to main lobe

Start with the Dirichlet kernel [66]

$$W(e^{j\omega}, M) = \frac{\sin(\omega(M+1)/2)}{\sin(\omega/2)} \quad (134)$$

For the rectangular window the amplitude of the main lobe is $W(0) = M + 1$ [66]. The amplitude of the first sidelobe is obtained when $\omega = 3\pi/(M+1)$

$$\begin{aligned} W(e^{j3\pi/(M+1)}, M) &= \frac{\sin\left(\frac{3\pi}{(M+1)}(M+1)/2\right)}{\sin(\omega/2)} \\ &= \frac{\sin\left(\frac{3\pi}{2}\right)}{\sin\left(\frac{3\pi/2}{(M+1)}\right)} \\ &= \frac{1}{\frac{3\pi/2}{(M+1)}} \quad (\text{for large } M \text{ and small } x, \sin(x) \approx x) \\ &= \frac{M+1}{3\pi/2} \end{aligned} \quad (135)$$

Hence the ratio of the main lobe level relative to the sidelobe level is $1/(3\pi/2)$ or -13dB . In general, the level of the k^{th} sidelobe for the rectangular window would be $1/((2k+1)\pi)$. For the triangular window the level would be $(1/((2k+1)\pi))^2$ (twice the dB value of the rectangular window).

11.8 Appendix H: Derivation of amplitude of demodulation spectrum

In this section, a few examples will be presented to support the earlier derivation of the amplitude of the demodulation spectrum: $|X_2(\omega_m)| \leq |X_2(0)|$.

11.8.1 Example 1

Consider an input signal $x(t)$ consisting of two sinusoidal sequences¹³² with spectral lines $X_0 = |X_0| \angle X_0$ at frequency f_0 and $X_1 = |X_1| \angle X_1$ at frequency f_1 .

$$\text{So } X(f) = X_0 \mathfrak{I}(\cos(\omega_0 t)) + X_1 \mathfrak{I}(\cos(\omega_1 t)) = \frac{1}{2} X_0 \delta(\omega - \omega_0) + \frac{1}{2} X_0 \delta(\omega + \omega_0) \\ + \frac{1}{2} X_1 \delta(\omega - \omega_1) + \frac{1}{2} X_1 \delta(\omega + \omega_1).$$

$$\text{Now } X_2(f) = X(f) \otimes X(f) \tag{136}$$

$$= \frac{1}{2} (X_0 \delta(\omega - \omega_0) + X_0 \delta(\omega + \omega_0) + X_1 \delta(\omega - \omega_1) + X_1 \delta(\omega + \omega_1)) \otimes$$

$$\frac{1}{2} (X_0 \delta(\omega - \omega_0) + X_0 \delta(\omega + \omega_0) + X_1 \delta(\omega - \omega_1) + X_1 \delta(\omega + \omega_1))$$

$$X_2(0) = \frac{1}{2} \frac{1}{2} X_0 X_0 + \frac{1}{2} \frac{1}{2} X_0 X_0 + \frac{1}{2} \frac{1}{2} X_1 X_1 + \frac{1}{2} \frac{1}{2} X_1 X_1$$

$$X_2(\omega_1 - \omega_0) = \frac{1}{2} \frac{1}{2} (X_0 X_1 + X_0 X_1)$$

$$X_2(\omega_0 - \omega_1) = \frac{1}{2} \frac{1}{2} (X_0 X_1 + X_0 X_1)$$

$$X_2(2\omega_0 + \omega_1) = \frac{1}{2} \frac{1}{2} (X_0 X_1)$$

$$X_2(-2\omega_0 - \omega_1) = \frac{1}{2} \frac{1}{2} (X_0 X_1)$$

$$X_2(2\omega_0) = \frac{1}{2} \frac{1}{2} (X_0 X_0 + X_0 X_0)$$

$$X_2(2\omega_1) = \frac{1}{2} \frac{1}{2} (X_1 X_1 + X_1 X_1)$$

$$X_2(\omega_0 + 2\omega_1) = \frac{1}{2} \frac{1}{2} (X_0 X_1)$$

$$X_2(-\omega_0 - 2\omega_1) = \frac{1}{2} \frac{1}{2} (X_0 X_1)$$

$$\text{But } X_2(0) = \frac{1}{2} (X_0^2 + X_1^2) > X_2(\omega_1 - \omega_0) = \frac{1}{2} (X_0 X_1) \text{ because}$$

¹³² It is assumed that the frequencies are all distinct, i.e. $f_0 \neq f_1$, etc.

$$\frac{1}{2}(X_0^2 + X_1^2) = \frac{1}{2}(X_0 + X_1)^2 - \frac{1}{2}2X_0X_1 = \frac{1}{2}(2A_{01})^2 - G_{01}^2$$

where $A_{01} = \frac{1}{2}(X_0 + X_1)$ is the arithmetic mean of X_0 and X_1 and $G_{01} = \sqrt{X_0X_1}$ is the geometric mean of X_0 and X_1 .

$$\text{As } A_{01} \geq G_{01} \text{ then } X_2(0) = 2(A_{01})^2 - G_{01}^2 = (2(A_{01})^2 - 3/2G_{01}^2) + \frac{1}{2}G_{01}^2$$

$$> X_2(\omega_l - \omega_h) = \frac{1}{2}G_{01}^2$$

$$\text{because } A_{01}^2 \geq 3/4G_{01}^2.$$

Similarly, $X_2(0) > X_2(2\omega_h + \omega_l)$, etc.

Hence $X_2(0) > X_2(\omega_m)$ for this example.

11.8.2 Example 2

Consider an input signal $x(t)$ consisting of three sinusoidal sequences with spectral lines $X_0 = |X_0| \angle X_0$ at frequency f_0 , $X_1 = |X_1| \angle X_1$ at frequency f_1 and $X_2 = |X_2| \angle X_2$ at frequency f_2 .

$$\begin{aligned} \text{So } X(f) &= X_0 \mathfrak{I}(\cos(\omega_0 t)) + X_1 \mathfrak{I}(\cos(\omega_1 t)) + X_2 \mathfrak{I}(\cos(\omega_2 t)) = \frac{1}{2}X_0 \delta(\omega - \omega_0) \\ &+ \frac{1}{2}X_0 \delta(\omega + \omega_0) + \frac{1}{2}X_1 \delta(\omega - \omega_1) + \frac{1}{2}X_1 \delta(\omega + \omega_1) + \frac{1}{2}X_2 \delta(\omega - \omega_2) \\ &+ \frac{1}{2}X_2 \delta(\omega + \omega_2). \end{aligned}$$

$$\text{Now } X_2(f) = X(f) \otimes X(f) \tag{137}$$

$$\begin{aligned} &= [\frac{1}{2}(X_0 \delta(\omega - \omega_0) + X_0 \delta(\omega + \omega_0) + X_1 \delta(\omega - \omega_1) + X_1 \delta(\omega + \omega_1) \\ &+ \frac{1}{2}X_2 \delta(\omega - \omega_2) + \frac{1}{2}X_2 \delta(\omega + \omega_2))] \otimes \\ &[\frac{1}{2}(X_0 \delta(\omega - \omega_0) + X_0 \delta(\omega + \omega_0) + X_1 \delta(\omega - \omega_1) + X_1 \delta(\omega + \omega_1) \\ &+ \frac{1}{2}X_2 \delta(\omega - \omega_2) + \frac{1}{2}X_2 \delta(\omega + \omega_2))] \end{aligned}$$

$$X_2(0) = \frac{1}{2}\frac{1}{2}X_0X_0 + \frac{1}{2}\frac{1}{2}X_0X_0 + \frac{1}{2}\frac{1}{2}X_1X_1 + \frac{1}{2}\frac{1}{2}X_1X_1 + \frac{1}{2}\frac{1}{2}X_2X_2 + \frac{1}{2}\frac{1}{2}X_2X_2$$

$$X_2(\omega_l - \omega_h) = \frac{1}{2}\frac{1}{2}(X_0X_1 + X_0X_1)$$

$$X_2(\omega_0 - \omega_2) = \frac{1}{2} \frac{1}{2} (X_0 X_2 + X_0 X_2)$$

$$X_2(\omega_2 - \omega_0) = \frac{1}{2} \frac{1}{2} (X_0 X_2 + X_0 X_2)$$

$$X_2(\omega_1 - \omega_2) = \frac{1}{2} \frac{1}{2} (X_1 X_2 + X_1 X_2)$$

$$X_2(\omega_2 - \omega_1) = \frac{1}{2} \frac{1}{2} (X_1 X_2 + X_1 X_2)$$

$$X_2(\omega_0 - \omega_1) = \frac{1}{2} \frac{1}{2} (X_0 X_1 + X_0 X_1)$$

$$X_2(2\omega_0 + \omega_1) = \frac{1}{2} \frac{1}{2} (X_0 X_1)$$

$$X_2(-2\omega_0 - \omega_1) = \frac{1}{2} \frac{1}{2} (X_0 X_1)$$

$$X_2(2\omega_1 + \omega_2) = \frac{1}{2} \frac{1}{2} (X_1 X_2)$$

$$X_2(-2\omega_1 - \omega_2) = \frac{1}{2} \frac{1}{2} (X_1 X_2)$$

$$X_2(2\omega_0) = \frac{1}{2} \frac{1}{2} (X_0 X_0 + X_0 X_0)$$

$$X_2(2\omega_1) = \frac{1}{2} \frac{1}{2} (X_1 X_1 + X_1 X_1)$$

$$X_2(2\omega_2) = \frac{1}{2} \frac{1}{2} (X_2 X_2 + X_2 X_2)$$

$$X_2(\omega_0 + 2\omega_1) = \frac{1}{2} \frac{1}{2} (X_0 X_1)$$

$$X_2(-\omega_0 - 2\omega_1) = \frac{1}{2} \frac{1}{2} (X_0 X_1)$$

$$\text{But } X_2(0) = \frac{1}{2} (X_0^2 + X_1^2 + X_2^2) > X_2(\omega_1 - \omega_0) = \frac{1}{2} \frac{1}{2} (X_0 X_1 + X_0 X_1) = \frac{1}{2} (X_0 X_1) = \frac{1}{2} G_{01}^2$$

$$\text{as } X_2(0) = \frac{1}{2} (X_0^2 + X_1^2 + X_2^2)$$

$$\geq \frac{1}{2} (X_0^2 + X_1^2) = \frac{1}{2} (2A_{01})^2 - \frac{1}{2} 2G_{01}^2$$

$$2A_{01}^2 - G_{01}^2 = 2A_{01}^2 - \frac{3}{2} G_{01}^2 + \frac{1}{2} G_{01}^2$$

$$\text{and } 2A_{01}^2 \geq \frac{3}{2} G_{01}^2$$

$$\text{and } X_2(0) > X_2(2\omega_0 + \omega_1), \text{ etc.}$$

Hence $X_2(0) > X_2(\omega_m)$ for this example.

11.8.3 Example 3

Consider an input signal $x(t)$ consisting of k sinusoidal sequences with spectral lines $X_0 = |X_0| \angle X_0$ at frequency f_0 , $X_1 = |X_1| \angle X_1$ at frequency $f_1 \dots X_k = |X_k| \angle X_k$ at frequency f_k . So $X(f) = X_0 \mathfrak{I}(\cos(\omega_0 t)) + X_1 \mathfrak{I}(\cos(\omega_1 t)) + \dots X_k \mathfrak{I}(\cos(\omega_k t))$

$$= \frac{1}{2} X_0 \delta(\omega - \omega_0) + \frac{1}{2} X_0 \delta(\omega + \omega_0) + \frac{1}{2} X_1 \delta(\omega - \omega_1) + \frac{1}{2} X_1 \delta(\omega + \omega_1) + \dots \frac{1}{2} X_k \delta(\omega - \omega_k) + \frac{1}{2} X_k \delta(\omega + \omega_k).$$

$$\text{Now } X_2(f) = X(f) \otimes X(f) \quad (138)$$

$$= \frac{1}{2} (X_0 \delta(\omega - \omega_0) + X_0 \delta(\omega + \omega_0) + X_1 \delta(\omega - \omega_1) + X_1 \delta(\omega + \omega_1) + \dots \frac{1}{2} X_k \delta(\omega - \omega_k) + \frac{1}{2} X_k \delta(\omega + \omega_k)) \otimes$$

$$\frac{1}{2} (X_0 \delta(\omega - \omega_0) + X_0 \delta(\omega + \omega_0) + X_1 \delta(\omega - \omega_1) + X_1 \delta(\omega + \omega_1) + \dots \frac{1}{2} X_k \delta(\omega - \omega_k) + \frac{1}{2} X_k \delta(\omega + \omega_k))$$

$$X_2(0) = \frac{1}{2} \frac{1}{2} X_0 X_0 + \frac{1}{2} \frac{1}{2} X_0 X_0 + \frac{1}{2} \frac{1}{2} X_1 X_1 + \frac{1}{2} \frac{1}{2} X_1 X_1 + \dots \frac{1}{2} \frac{1}{2} X_k X_k + \frac{1}{2} \frac{1}{2} X_k X_k$$

$$X_2(\omega_1 - \omega_0) = \frac{1}{2} \frac{1}{2} (X_0 X_1 + X_0 X_1)$$

$$X_2(\omega_0 - \omega_k) = \frac{1}{2} \frac{1}{2} (X_0 X_k + X_0 X_k)$$

$$X_2(\omega_k - \omega_0) = \frac{1}{2} \frac{1}{2} (X_0 X_k + X_0 X_k)$$

$$X_2(\omega_1 - \omega_k) = \frac{1}{2} \frac{1}{2} (X_1 X_k + X_1 X_k)$$

$$X_2(\omega_2 - \omega_1) = \frac{1}{2} \frac{1}{2} (X_1 X_2 + X_1 X_2)$$

$$X_2(\omega_0 - \omega_1) = \frac{1}{2} \frac{1}{2} (X_0 X_1 + X_0 X_1)$$

$$X_2(2\omega_0 + \omega_1) = \frac{1}{2} \frac{1}{2} (X_0 X_1)$$

$$X_2(-2\omega_0 - \omega_1) = \frac{1}{2} \frac{1}{2} (X_0 X_1)$$

$$X_2(2\omega_1 + \omega_k) = \frac{1}{2} \frac{1}{2} (X_1 X_k)$$

$$X_2(-2\omega_1 - \omega_k) = \frac{1}{2} \frac{1}{2} (X_1 X_k)$$

$$X_2(2\omega_0) = \frac{1}{2} \frac{1}{2} (X_0 X_0 + X_0 X_0)$$

$$X_2(2\omega_1) = \frac{1}{2}\frac{1}{2}(X_1X_1 + X_1X_1)$$

$$X_2(2\omega_k) = \frac{1}{2}\frac{1}{2}(X_kX_k + X_kX_k)$$

$$X_2(\omega_0 + 2\omega_1) = \frac{1}{2}\frac{1}{2}(X_0X_1)$$

$$X_2(-\omega_0 - 2\omega_1) = \frac{1}{2}\frac{1}{2}(X_0X_1)$$

$$\text{But } X_2(0) = \frac{1}{2}(X_0^2 + X_1^2 + \dots X_k^2) > X_2(\omega_1 - \omega_0) = \frac{1}{2}\frac{1}{2}(X_0X_1 + X_0X_1) = \frac{1}{2}(X_0X_1) = \frac{1}{2}G_{01}^2$$

$$\text{as } X_2(0) = \frac{1}{2}(X_0^2 + X_1^2 + \dots X_k^2)$$

$$\geq \frac{1}{2}(X_0^2 + X_1^2) = \frac{1}{2}(2A_{01}^2) - \frac{1}{2}2G_{01}^2$$

$$2A_{01}^2 - G_{01}^2 = 2A_{01}^2 - 3/2G_{01}^2 + \frac{1}{2}G_{01}^2$$

$$\text{and } 2A_{01}^2 \geq 3/2G_{01}^2.$$

$$\text{Also } X_2(0) > X_2(2\omega_0 + \omega_1), \text{ etc.}$$

Hence $X_2(0) > X_2(\omega_m)$ for this generalized example with k sinusoids.

11.9 Appendix I: Experimental results on the effect of a time domain window on the SA

The influence of the time domain window on the SA will be illustrated for values of $\beta = 0, 6, 100$ of the Kaiser window ($\beta = 0$ corresponds to the rectangular window and $\beta = 5.44$ is similar to the Hamming window).

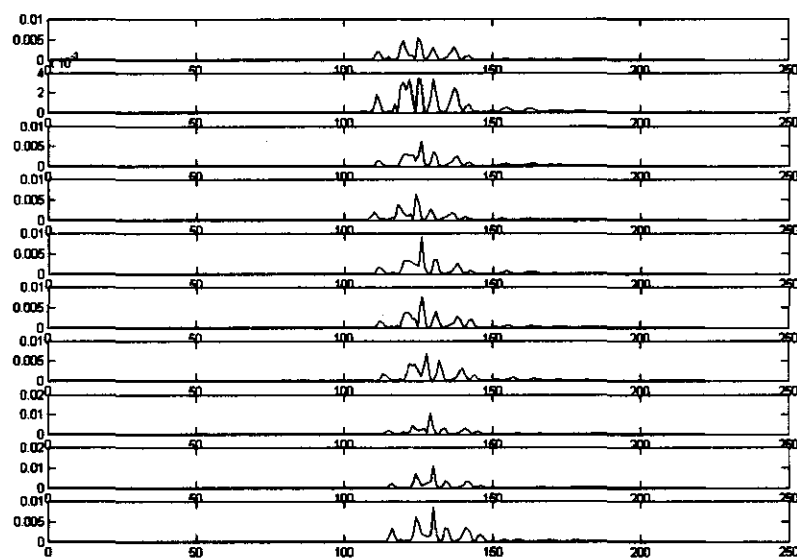


Figure 226: Waterfall plot of signal with Kaiser window ($\beta = 0$)

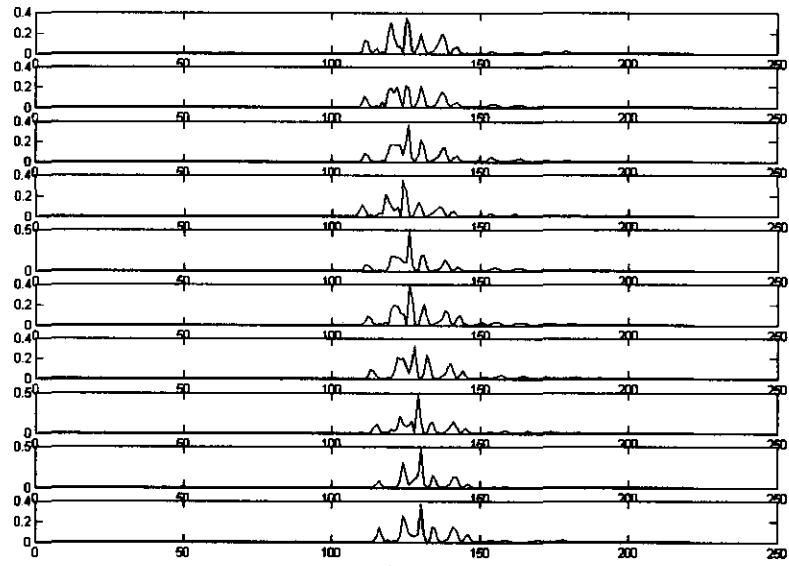


Figure 227: Waterfall plot of signal with Kaiser window ($\beta = 6$)

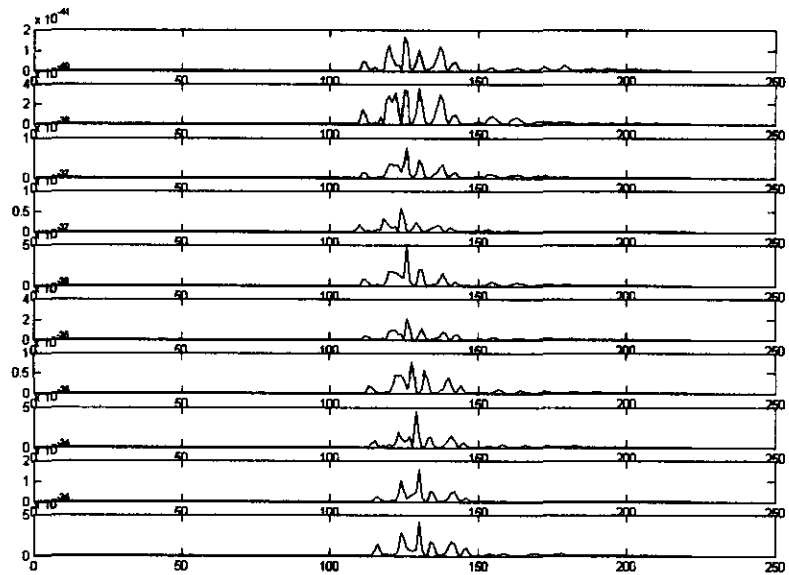


Figure 228: Waterfall plot of signal with Kaiser window ($\beta = 100$)

The SA corresponding to values of β for the Kaiser window is shown in the following figures. Note that the width of the SA decreases as β increases.

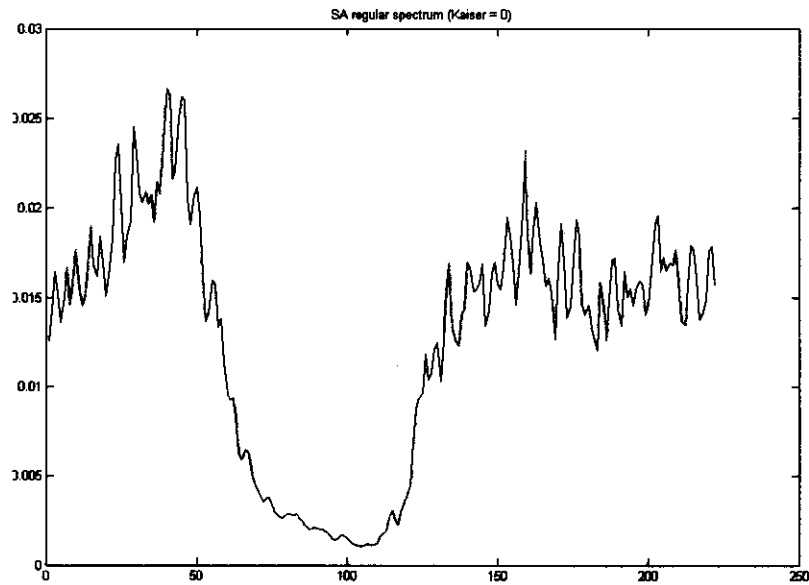


Figure 229: SA of signal with Kaiser window ($\beta = 0$)

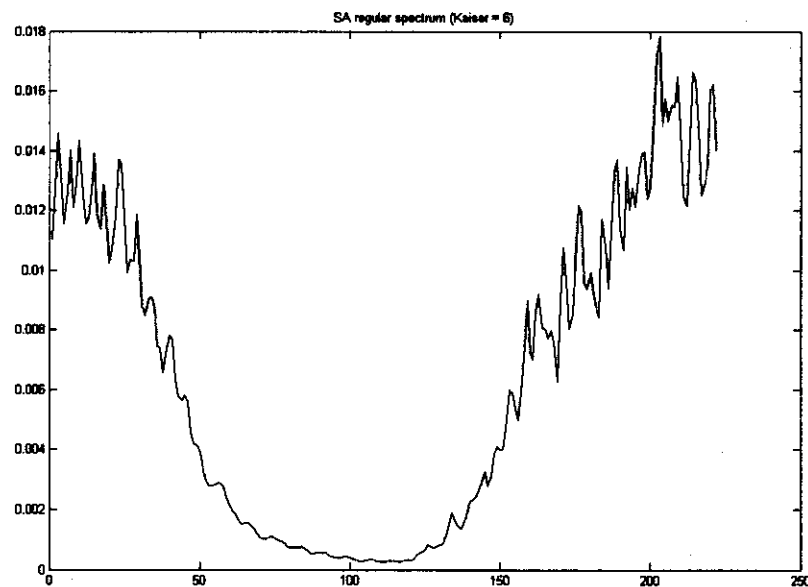


Figure 230: SA of signal with Kaiser window ($\beta = 6$)

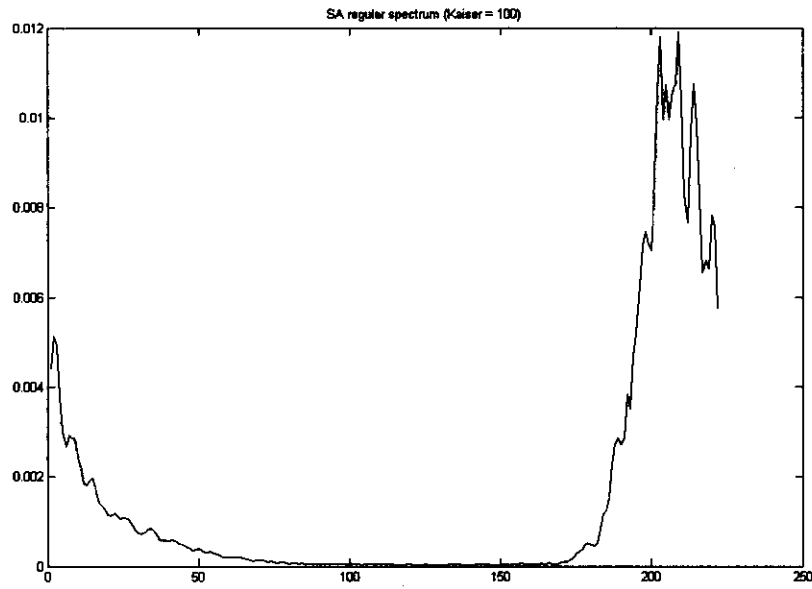


Figure 231: SA of signal with Kaiser window ($\beta = 100$)

The following figures show the repetitive signal corresponding to the SA with increasing values of β . As β increases the resulting pulse train more closely resembles a comb sequence in the time domain (with regard to the width of the pulses that decreases with an increase in β).

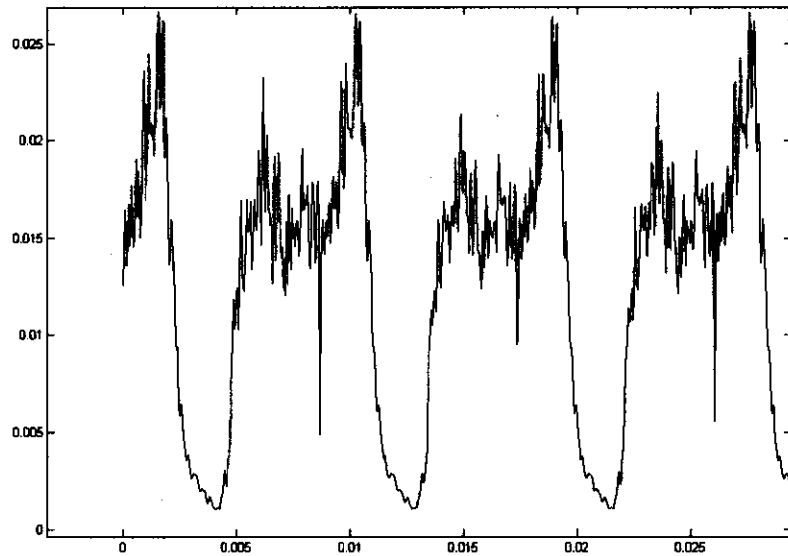


Figure 232: Pulse train corresponding to SA with Kaiser window ($\beta = 0$)

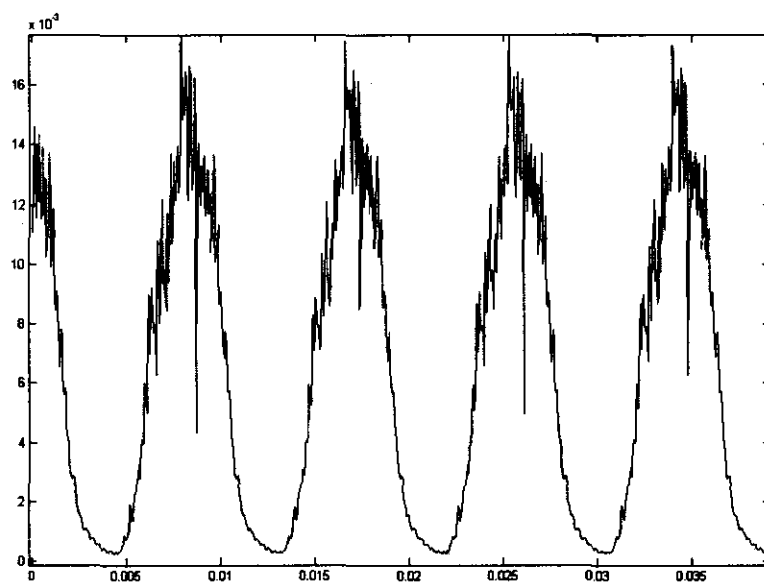


Figure 233: Pulse train corresponding to SA with Kaiser window ($\beta = 6$)

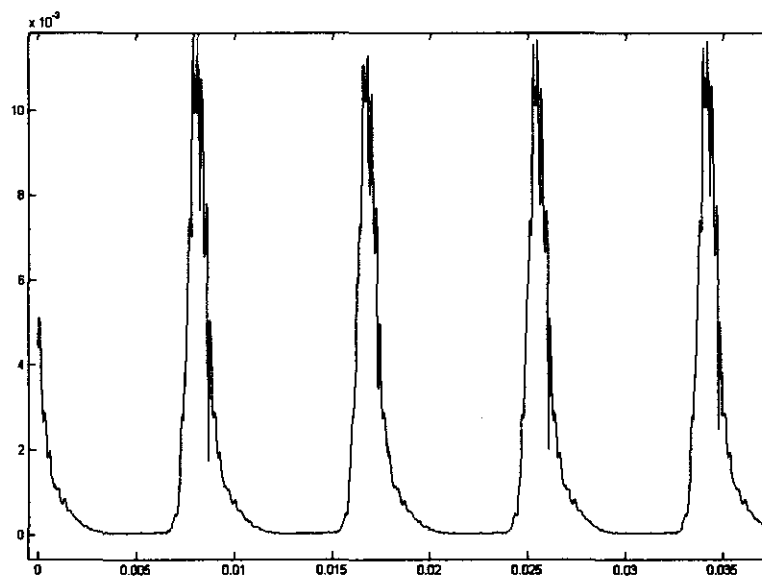


Figure 234: Pulse train corresponding to SA with Kaiser window ($\beta = 100$)

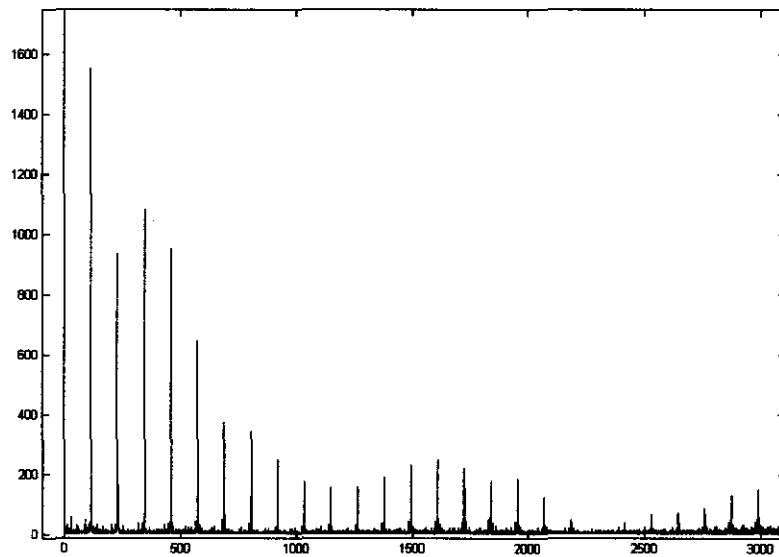


Figure 235: Demodulation spectrum of signal prior to SA for Kaiser window where $\beta = 0$

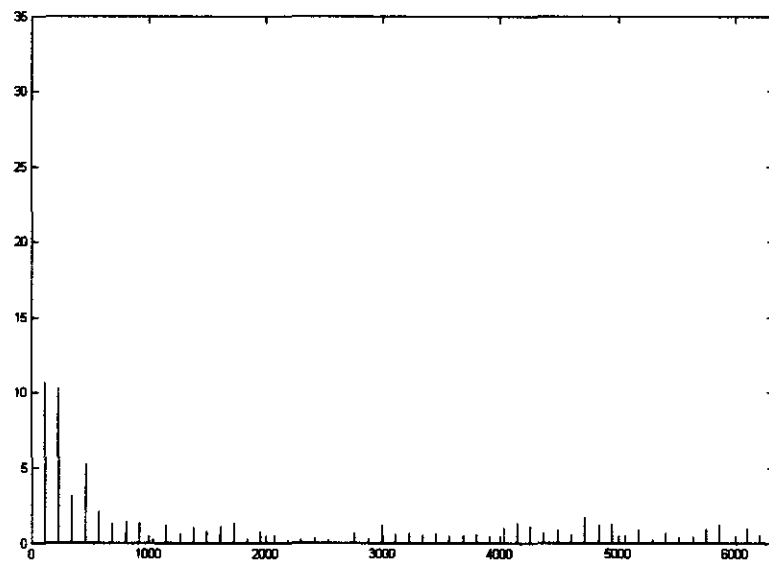


Figure 236: Demodulation spectrum of pulse signal with SA for Kaiser window where $\beta = 0$

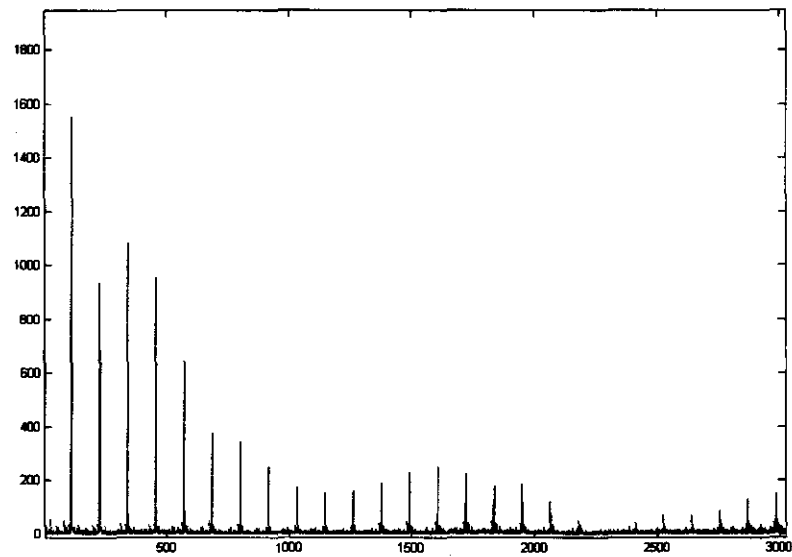


Figure 237: Demodulation spectrum of signal prior to SA for Kaiser window where $\beta = 6$

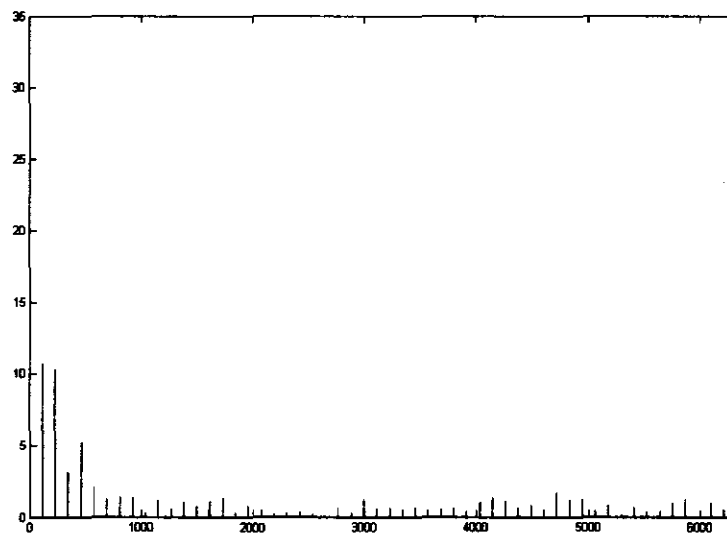


Figure 238: Demodulation spectrum of signal with SA for Kaiser window where $\beta = 6$

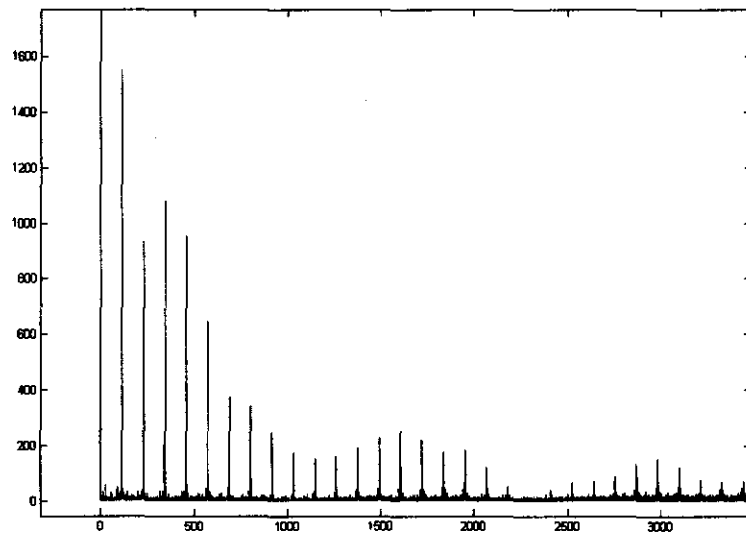


Figure 239: Demodulation spectrum of signal prior to SA for Kaiser window where $\beta = 100$

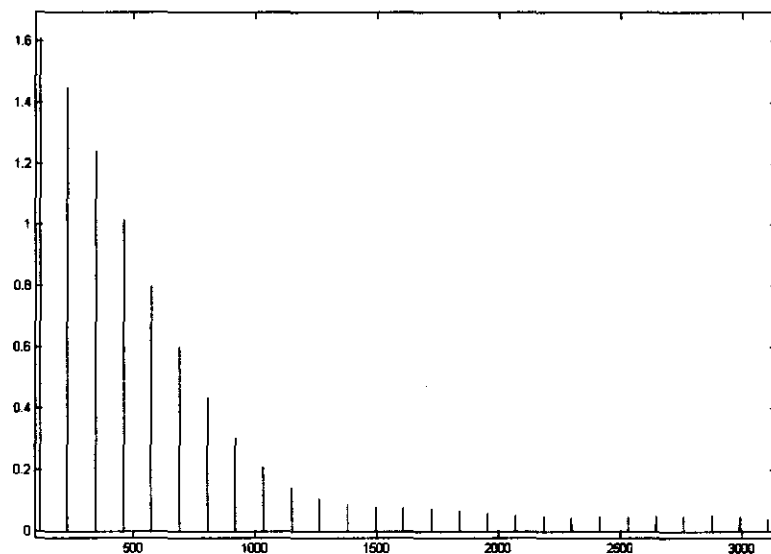


Figure 240: Demodulation spectrum of pulse signal with SA for Kaiser window where $\beta = 100$

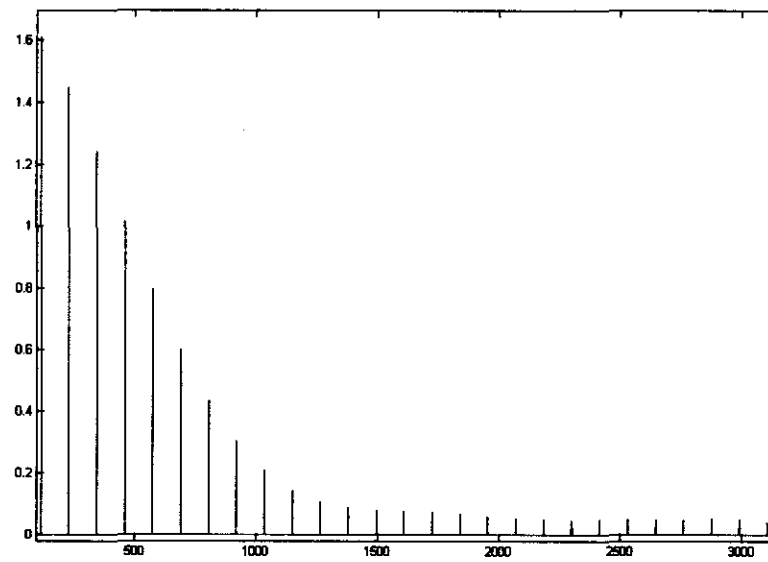


Figure 241: Demodulation spectrum of signal prior to SA for Kaiser window where $\beta = 100$ (zoom plot)

11.10 Appendix J: Recommendations for the application of pattern recognition techniques in condition-monitoring problems

Before applying pattern recognition techniques such as black box neural networks to a problem, do the following checks:

1. Make graphs of each feature versus the output value (class membership or output value) and look for correlations of each individual feature with the output. If sufficient correlation is found, sophisticated modelling techniques, such as neural networks, would be unnecessary because of the simplicity of the problem. If not, make scatterplots of the variables (usually two or three at a time owing to limitations in visualizing higher dimensions). A set of m variables, each scatterplot with two variables at a time, will yield in the order of $(m + 1)(m)/2$ scatterplots (with two variables at a time) owing to symmetry.
2. Has the data set been preprocessed correctly?
3. Determine what features an expert in the field would use for accurate diagnosis. These features refer to the transformation of measurement variables to yield a smaller set of variables while retaining or improving classification – or modelling – performance. The selection of a compact feature set can decrease the size of the database, reduce training times and improve generalization performance.
4. The nearest neighbour (NNR) technique provides a quantitative indication of the complexity of a problem. It requires no lengthy training sessions and parameter optimization. Furthermore, it automatically uses leave one out cross validation (LOO XVAL)¹³³. The NNR technique can also be used in modelling problems if the NNRC technique [94], which is proposed and

¹³³ The NNR has an error rate smaller than twice the optimal Bayes error rate.

described in this thesis, is used. As a first step the NNR technique may be applied to each feature. The resulting classification (or modelling) accuracies quickly ascertain whether additional features (hence modelling) will be required. If the desired performance is not attained, the best individual feature may be selected and compared with all the other features (thus selecting two features). This heuristic strategy of forward selection may be continued until no further improvement can be gained. The Gamma test [36] introduces the number of neighbours in the NNR as an additional parameter. The slope of the graph of the error rate of the NNR with k neighbours may be extrapolated to obtain the y axis intercept, which is called the Gamma estimate and is the best any classifier can (theoretically) hope to achieve¹³⁴.

5. If the complexity of the problem, determined from an initial exploration as indicated in the previous steps, warrants it then neural networks may be considered. However, many topological, parameter optimization and validation issues have to be explored. In particular, the number of free parameters in the neural network should be in proportion to the number of input variables and data points (a heuristic rule of thumb is about ten data points for each weight in the neural network [18]). The generalization performance of the neural network should always be specified (ideally using LOO XVAL¹³⁵, or using separate training, test and validation sets).
6. Determine if *a priori* rules are available about the fault conditions, for example from experts in the field. Even if the rules are not always accurate, they are likely to provide a good set of additional features that can be used in

¹³⁴ It is maintained in this thesis that it is not actually possible to obtain the Gamma estimate in a real situation, as the number of neighbours has fractional values smaller than one as the Gamma estimate is approached (tending to zero, implying an infinite number of data points). Hence the NNR estimate may still be the best estimate of classifier (or model) performance, as it may be dangerous to extrapolate to the Gamma estimate. Hence the Gamma test has not been used in this thesis, as it is held that the NNR estimate is accurate enough.

¹³⁵ As LOO XVAL is very computationally demanding, the CVCP technique [97], proposed in this thesis, may be considered. Alternatively, LOO XVAL may be implemented much more efficiently [56].

a pattern recognition technique (for example fuzzy neural networks that allow the programming of *a priori* rules inside the structure of the neural network. The process of extracting accurate rules from an expert may present practical problems, including the high costs involved, fear that the expert may be replaced by a machine, and semantic difficulty in expressing rules. It may be easier to provide an expert with a number of signals to classify in different fault conditions. One problem is that not all information is stored in a computer database. The diagnostic expert will frequently use additional information and/or tests to provide more reliable diagnostic decisions. The computer can still be used very profitably to handle the routine diagnostic decisions and to assist the operator, leaving the more difficult cases for the expert.

7. Any statistical pattern recognition technique, including neural networks, can make predictions outside of the training set. When the prediction is requested between samples of the training set this is called interpolation. The neural network can be expected to perform linear interpolation, which might happen to perform well if the underlying phenomena are also linearly related. When the prediction is requested outside of the samples of the training set this is called extrapolation. If the neural network transfer function is the sigmoid function, then the output will saturate for very large or very small values (in a variable of single dimension). Hence the neural network would tend to produce the same value (do nothing but keep the value of the input signal), which seems a reasonable strategy in the absence of further knowledge. Thus the extrapolation performance can be expected to saturate if regions far from the training set are explored. This might perform well if the underlying phenomena also tend to saturate.

However, the only reliable predictions of a statistical pattern recognition technique are made on samples of the training set itself (assuming, of course, that the training set itself is accurate and complete). Evaluating the model

resulting from a statistical pattern recognition technique on samples outside of the training set is bound to result in an error.

The garbage in, garbage out (GIGO) principle should be borne in mind: all the knowledge obtained by a statistical pattern recognition technique depends very crucially on the accumulation of a training set. Knowledge about a specific fault condition, for example dynamic loading, can only be obtained by exposing the statistical pattern recognition technique to enough samples under different conditions to obtain a representative database. The statistical pattern recognition technique cannot be expected to learn the desired relationships magically from a brief exposure to the fault condition.

In this respect, the methodology employed for the application of statistical pattern recognition techniques is deemed to be much more important than the contribution of any individual technique. Accumulating enough data to make a meaningful contribution does require time and money. Consequently, it is important to obtain the support of management if the data based approach is to be successful. It also requires discipline to monitor machinery at regular intervals (which is crucial if lessons are to be learnt from previous failures). Employing a disciplined approach is vital if the statistical pattern recognition techniques are to be employed successfully. Judging from the lack of readily available databases, this ideal seems to be the exception rather than the rule today. A disciplined approach can lead to several competitive advantages to the firm, including a systematic database of prior failures, documentation of baseline and actual levels of operation of various machinery, and automation of (robust) failure detection. Furthermore, it is imperative to balance the training set for classification techniques so that each class has approximately the same number of samples. Failure to do so immediately invalidates any model trained on the unbalanced data set. As the failure data are usually scarce relative to the normal data, this step is almost always required.

Standard pattern recognition techniques, such as neural networks, do not provide any indication of 'certainty' in the classification/modelling decision. Intuitively it would be expected that a classifier should not have high confidence when interrogated on samples far removed from training set data. Research is currently in progress, using fuzzy logic neural networks, in this regard [100]. The neural network can only be expected to provide good performance on samples 'close' enough to the training set data provided (although the neural network can interpolate and extrapolate to a certain degree). Hence it can essentially be viewed as a large lookup table (corresponding to the training set data). After the neural network has been trained successfully, it provides good performance on data samples located close to the training data set. Data samples far from the training set data would, in general, give poor performance.

Thus it remains the responsibility of the user to populate the training set data with representative data of the different fault conditions under varying environmental conditions. Failure of the neural network in actual use would imply that the actual operating conditions are far from the training set data¹³⁶, therefore an examination of the latest data (in order to establish deviations from the training set data) and possible retraining of the neural network would be required.

In literature a mechanical analogy is suggested that may be helpful in understanding the complexities associated with neural network modelling. The data set is modelled as a series of spikes on a surface (assuming for the moment a two-dimensional feature space). Each spike corresponds to a single data set output value (associated with the input features). At this stage it would be desirable to divide the spikes into three additional

¹³⁶ Other obvious possibilities, such as erroneous sensor data to the neural network, should of course first be eliminated.

separate sets (arbitrarily called the training, validation and test sets). Uniting the spikes associated with the three data sets provides the complete data set, of course.

The task of the pattern recognition technique may now be viewed as similar to the task of fitting a surface (for example a plate) through all the spikes. This should be done in such a way as to minimize the deviations of the plate from the spikes on the test set (i.e. to minimize the error on the test set). The training algorithm can use a second data set (the validation set) in order to obtain an estimate of the error on samples close to the training set. The validation error can then, for example, be used to terminate training.

The plate could be made of a material that is extremely rigid (we visualize the spikes penetrating through this material). It is clear that it would be impossible to position this rigid plate in such a way that the plate aligns perfectly with all the spikes, unless the spikes happened to be positioned exactly in a plane surface. This situation corresponds to a neural network with just a single neuron in a single hidden layer and effectively performs a linear regression on the input data. As the pattern recognition technique would have too few free parameters (degrees of freedom), the bias would be rather high (unless the surface to be modelled is a very simple surface such a plane surface).

Increasing the number of free parameters in the pattern recognition technique corresponds to choosing a much more flexible material for the plate, a material that can bend locally in order to come closer to the spikes (hence minimizing the error on the training set). As the material becomes extremely flexible it would be possible to match any number of spikes exactly (provided that enough free parameters in the pattern recognition technique are available to 'memorize' the bends in the plate). This is of course the well-known universal approximation theorem. However, the results on the validation and test sets would not be as good. The more flexible the surface, the greater the deviation of the spikes (using the validation and test sets) would be from the plate. Thus the error on the test set would increase as the number of free parameters increases. More formally, the variance would increase with an increase in the number of free parameters (degrees of freedom) of the pattern recognition technique. From the preceding discussion it is clear that training of a pattern recognition technique always involves a trade-off between the bias and variance

achievable for a specific problem (by analogy choosing the 'right' material for the plate). Stating results on the training set only provides no indication of the generalization capabilities of the pattern recognition technique (since the universal approximation theorem states that 100% accuracy can always be achieved on the training set, provided that enough free parameters and training time are available). The more important question is to minimize the number of free parameters (for example by using a neural network with the minimum number of layers, neurons and weights). This can also be seen by analogy to optimizing the material of the plate to achieve a compromise between the rigidity and flexibility. It would seem obvious to start with the most rigid material (with large bias and small variance) and then gradually to add more flexibility as required (small bias and large variance). This corresponds to the well-known growing (heuristic) strategy, starting with the smallest size network and gradually increasing in size. The opposite (heuristic) pruning strategy starts with a large network and gradually prunes away unneeded layers, neurons or weights. For both strategies it is essential to add proper validation (for example by using a test set) when evaluating the success of the pattern recognition technique.

The training of a neural network MLP with sigmoidal transfer function may be visualized as follows (for a single feature). Assume a single neuron in the hidden layer. If the data set consists of two spikes only at some distance from each other, the neural network training algorithm would tend to do linear interpolation when interrogated between the two spikes. Using the previous analogy this may be visualized as fitting a very stiff, rigid plate exactly between the two spikes.

Owing to the saturation of the sigmoidal transfer function to one for large input values, the result would be one for large values (when the spike to be evaluated is very far from the closest spike in the training set). Hence the neural network output would tend to remain the same (saturate) when the spike associated with the neural network output is 'far' from the closest feature in the training set. Using the previous analogy this can be visualized as a very stiff, rigid plate placed (nearly) horizontally at the position of the closest spike.

The NNR can be understood simply as choosing a spike (data sample in the training set) 'close to' the current spike in such a way that the distance to the nearest spike (which can

be quantified in a number of ways) is minimized. Hence the training set is artificially increased in size by choosing the nearest spike(s) (neighbour(s)) to do validation. By interpolating between the current spike and the closest spike (for example using linear interpolation) it would be easy to add additional spikes closer to the current spike for validation purposes (as is implicitly done with the Gamma test). A pattern recognition model that is too flexible would not follow the linear interpolation and would consequently be penalized. Thus the generalization performance of the model may be enhanced by constraining the model to be more rigid during training (to follow a linear interpolation). The above assumes, of course, that the closest spike is really 'close enough' to the current spike to give a meaningful interpretation of the current spike. As mentioned in this thesis, the same idea from the NNR may be extended to any model (such as neural networks) simply by evaluating the output of the neural network (which may be a continuous variable and does not need to be discrete as for a classification problem) on samples 'close' to the current spike.

Hence the process of validation may be understood quite simply as constraining the possible output values of the model (such as neural networks). By analogy the material is chosen to be 'rigid enough' between the spikes (if the material is too flexible the generalization performance would suffer). It stands to reason that this goal can be achieved using a variety of techniques, some of which have been discussed in this thesis.

11.11 Appendix K: Examples of simulated bearing signals and the application of a non-linear operator

11.11.1 Introduction

In the following sections some examples of simulated bearing signals and the application of a non-linear operator (squaring) are discussed. The following figures are shown:

- Time domain signals plotted over each other for one period
- Mean of squares (MOS) $\Sigma(X^2)/N$ of the time domain signal is plotted for one period of the synchronous average
- The mean squared – $(\Sigma X/N)^2$ – of the time domain signal is plotted for one period of the synchronous average
- The ratio – $(\Sigma X/N)^2/(\Sigma X^2/N)$ – of the mean squared to MOS of the time domain signal plotted for one period of the synchronous average
- The synchronous kurtosis (SK) of samples of SA – plotted for one period of the synchronous average
- FFT magnitude spectrum of time domain signal
- FFT demodulation magnitude spectrum of time domain signal

The noise level and speed are varied in subsequent sections.

Please note that the synchronous kurtosis (SK) of, for example, figure 246 is computed using the Matlab©® function kurtosis (4th order moment divided by the square of 2nd order moment). However, no test for division by zero has been performed. The kurtosis definition should be changed in future to give, for example, kurtosis = N (instead of infinity) when the 2nd order moment is close to zero. In the current examples the kurtosis would be large (or undefined, in which case the plot function of Matlab©® would not

show it), which is acceptable for the purposes of this thesis. Note from sections 11.11.2 and 11.11.3 that increasing the speed variation also increases the synchronous kurtosis (with no additive noise), as expected. The synchronous kurtosis (SK) may hence be interpreted as a measure of the spread in time of samples of the synchronous average (SA).

11.11.2 No noise added, speed variation = 0%

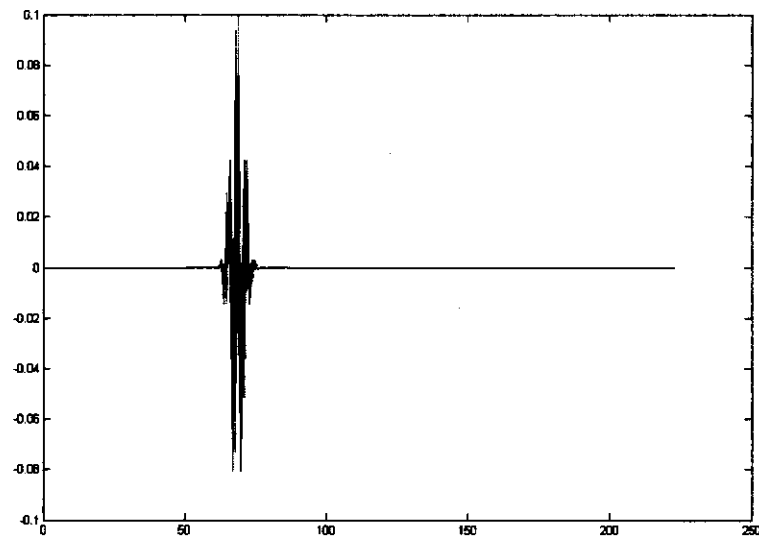


**Figure 242: Time domain signals plotted over each other for one period
(STD of noise = 0, speed variation = 0%)**

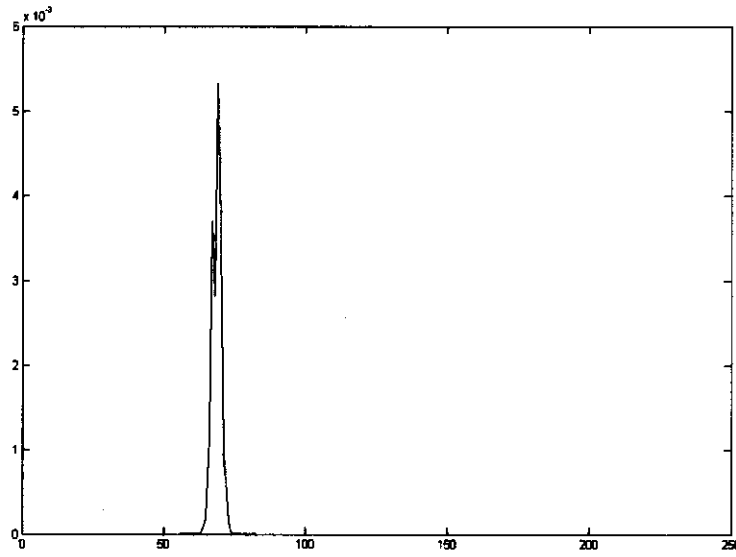


Figure 243: MOS $\Sigma(X^2)/N$ of time domain signal plotted for one period of the synchronous average (STD of noise = 0, speed variation = 0%)

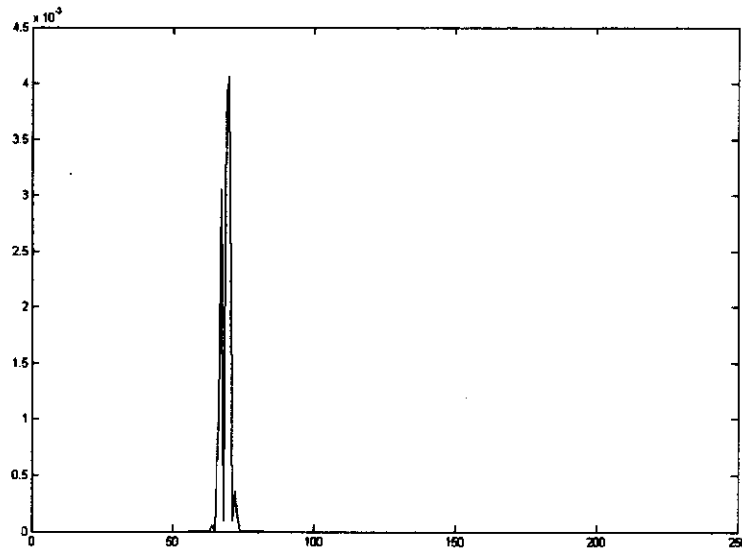


Figure 244: Mean squared – $(\Sigma X/N)^2$ – of time domain signal plotted for one period of the synchronous average (STD of noise = 0, speed variation = 0%)

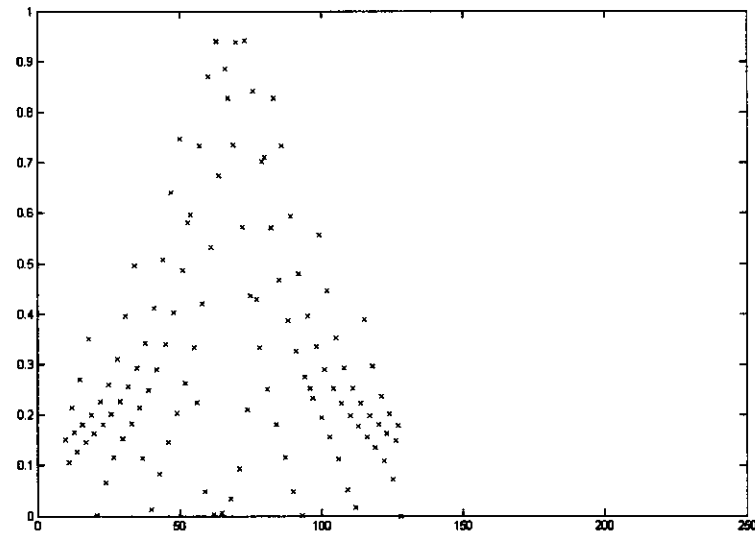


Figure 245: Ratio of mean squared to MOS of time domain signal – $(\sum X/N)^2/(\sum X^2/N)$ – plotted for one period of the synchronous average (STD of noise = 0, speed variation = 0%)

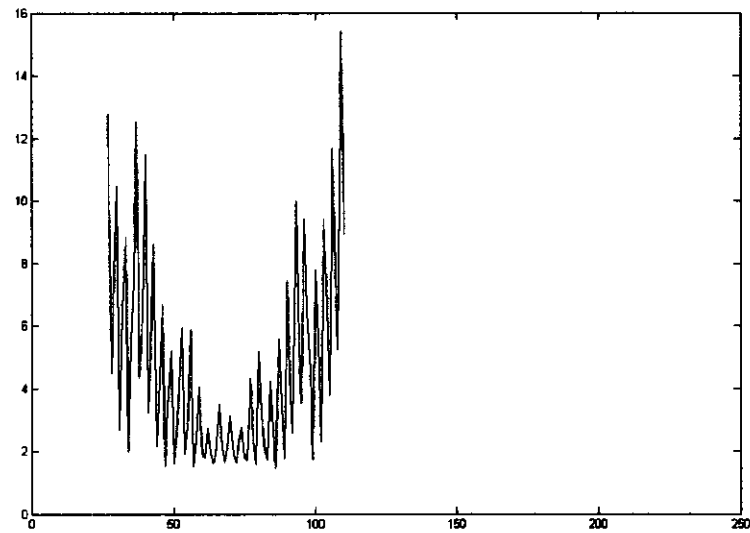


Figure 246: Kurtosis of samples of SA – plotted for one period of the synchronous average (STD of noise = 0, speed variation = 0%)

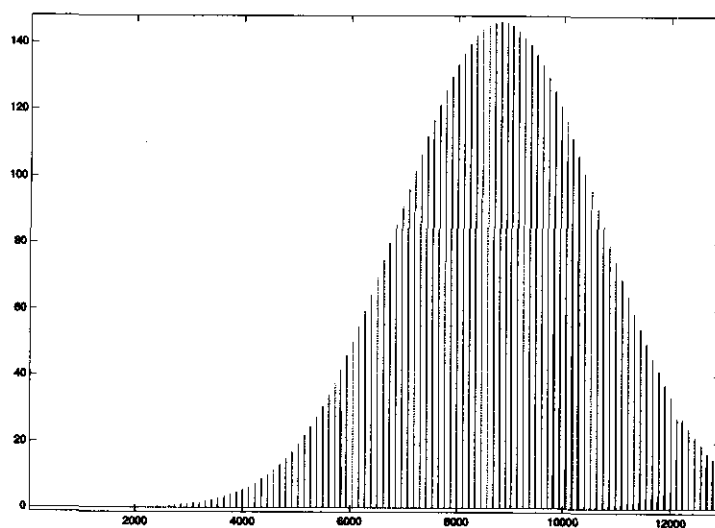


Figure 247: FFT magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0, speed variation = 0%)

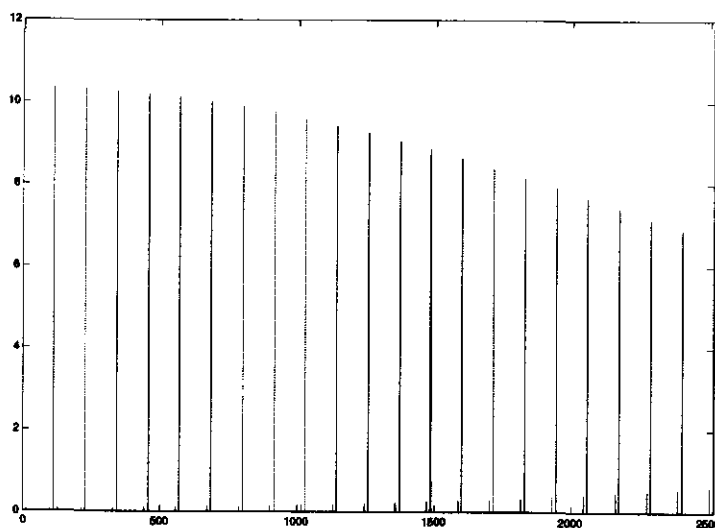


Figure 248: FFT demodulation magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0, speed variation = 0%)

11.11.3 No noise added, speed variation = 20%

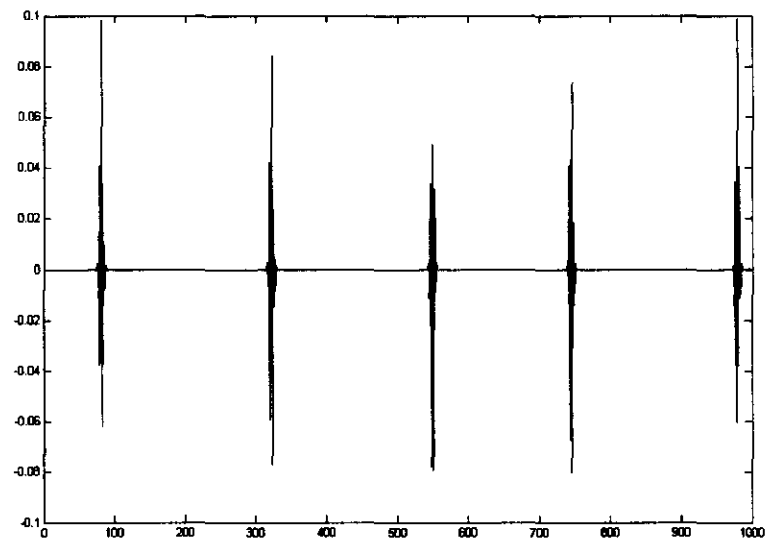


Figure 249: Time domain signal plotted for about five periods (STD of noise = 0, speed variation = 20%)

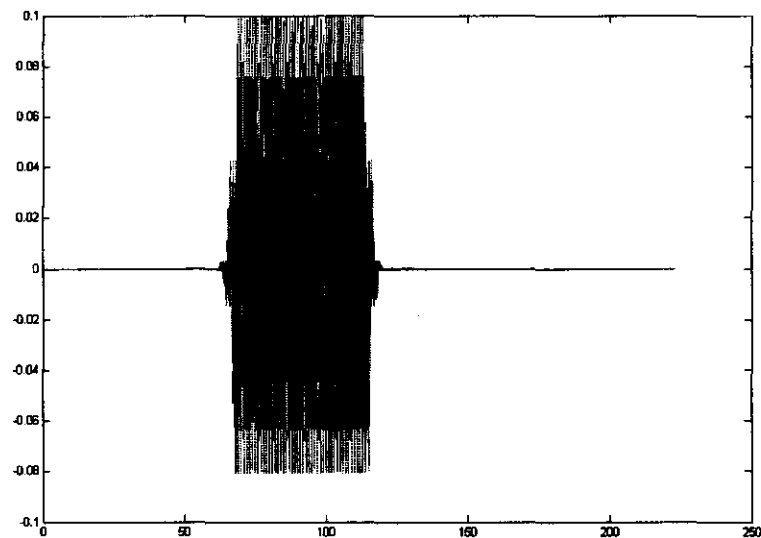


Figure 250: Time domain signals plotted over each other for one period (STD of noise = 0, speed variation = 20%)

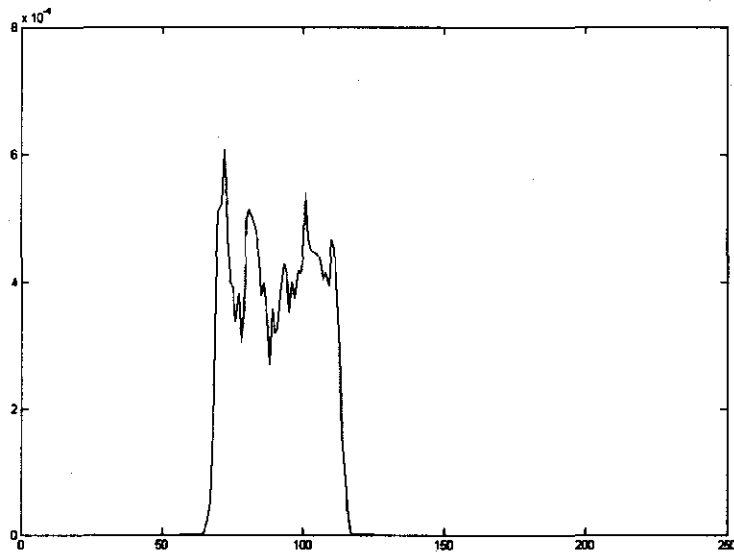


Figure 251: MOS $\Sigma(X^2)/N$ of time domain signal plotted for one period of the synchronous average (STD of noise = 0, speed variation = 20%)

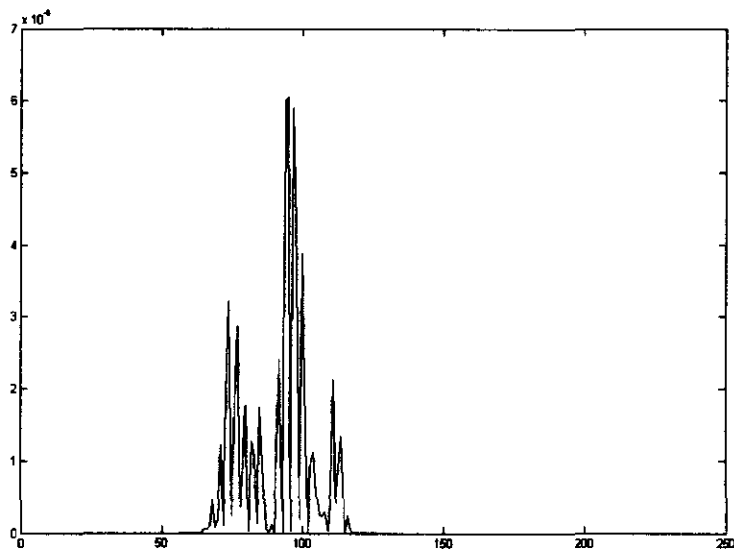


Figure 252: Mean squared – $(\Sigma X/N)^2$ – of time domain signal plotted for one period of the synchronous average (STD of noise = 0, speed variation = 20%)

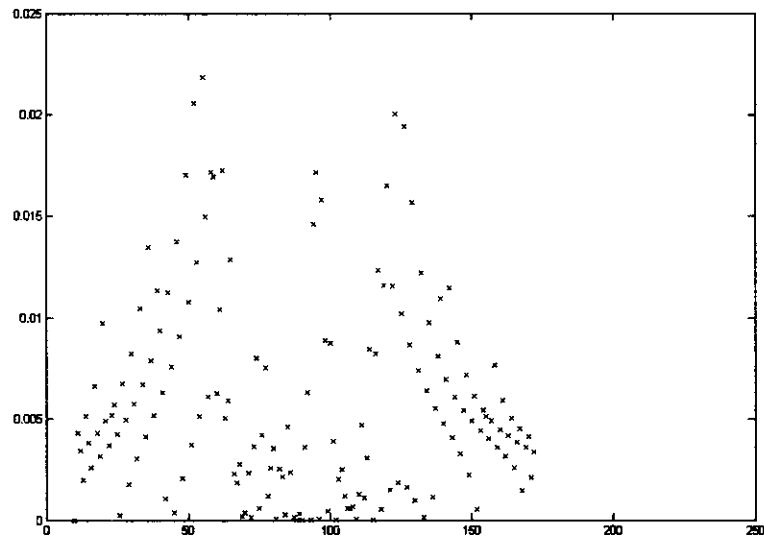


Figure 253: Ratio of mean squared to MOS of time domain signal – $(\sum X/N)^2 / (\sum X^2/N)$ – plotted for a period of the SA (STD of noise = 0, speed variation = 20%)

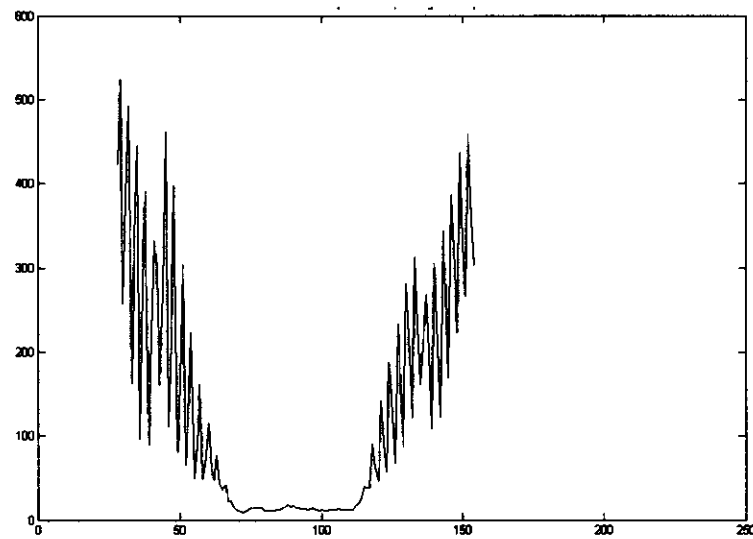


Figure 254: Kurtosis of samples of SA – plotted for one period of the synchronous average (STD of noise = 0, speed variation = 20%)

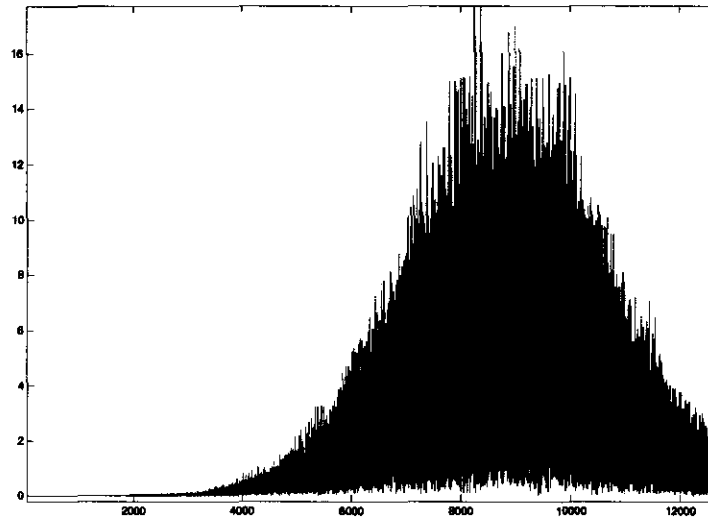


Figure 255: FFT magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0, speed variation = 20 %)

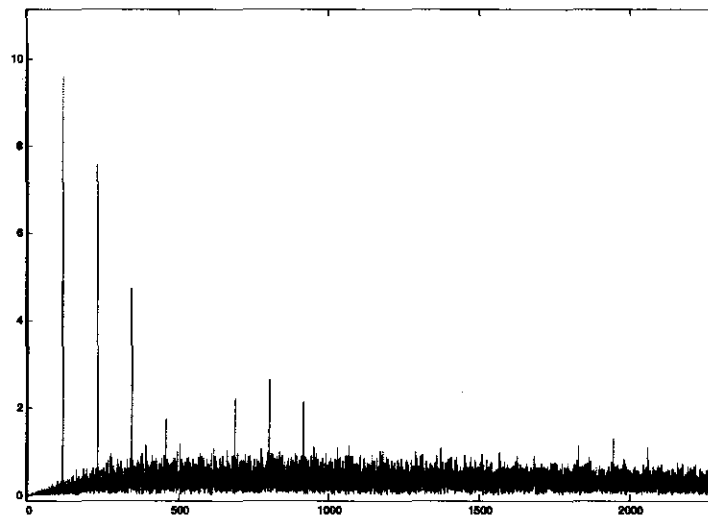


Figure 256: FFT demodulation magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0, speed variation = 20 %)

11.11.4 Noise with normal distribution added, speed variation = 20%

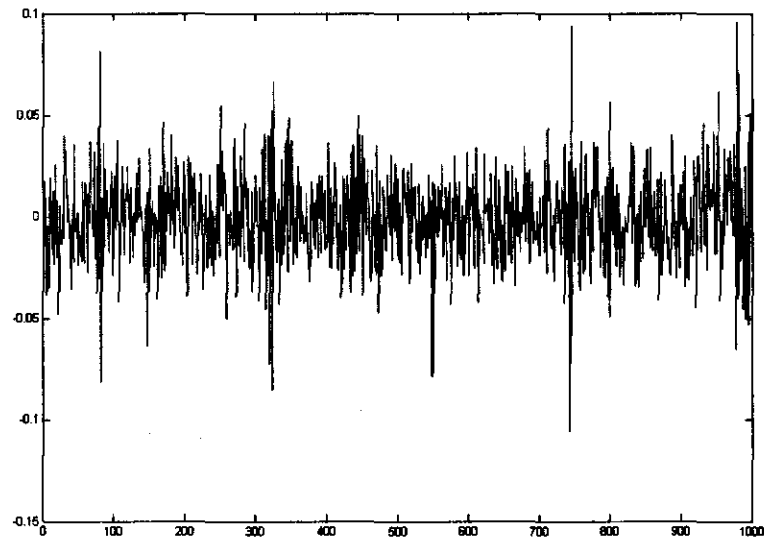


Figure 257: Time domain signal plotted for about five periods (STD of noise = 0.02, speed variation = 20%)

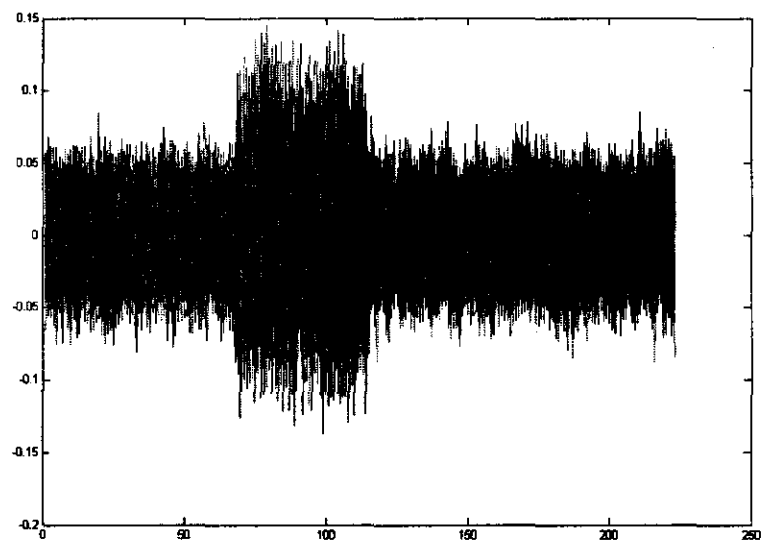


Figure 258: Time domain signals plotted over each other for one period (STD of noise = 0.02, speed variation = 20%)

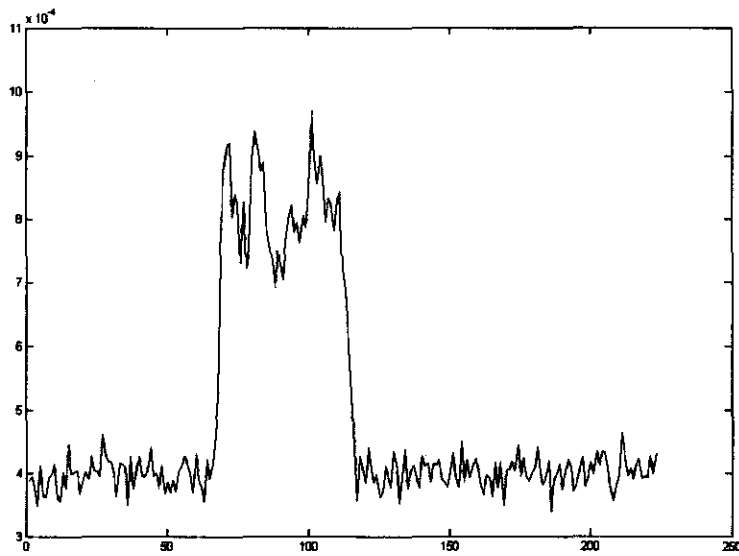


Figure 259: MOS $\Sigma(X^2)/N$ of time domain signal plotted for one period of the synchronous average (STD of noise = 0.02, speed variation = 20%)

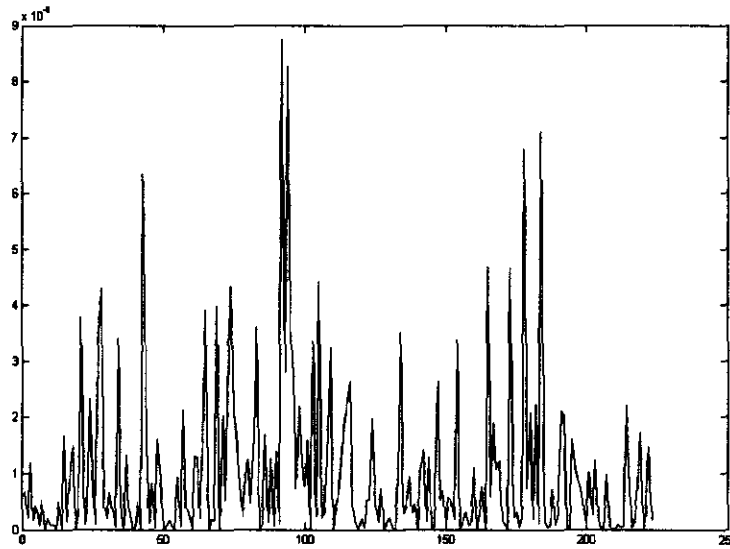


Figure 260: Mean squared – $(\Sigma X/N)^2$ – of time domain signal plotted for one period of the synchronous average (STD of noise = 0.02, speed variation = 20%)

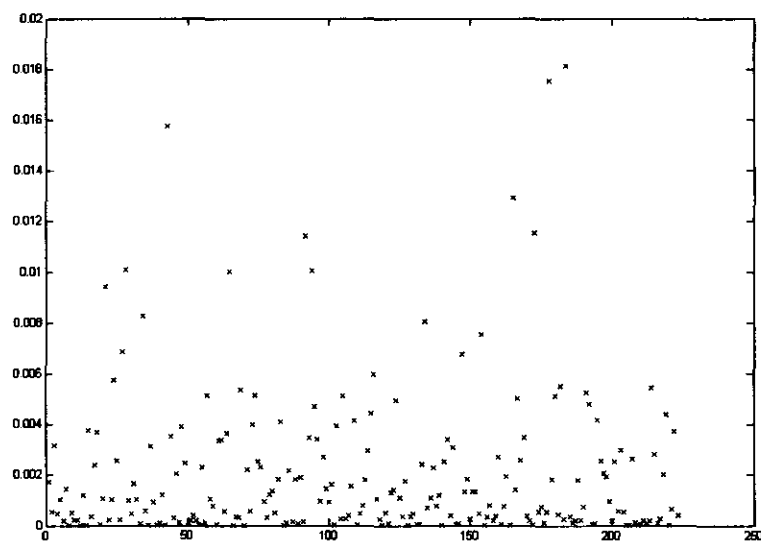


Figure 261: Ratio of mean squared to MOS of time domain signal – $(\sum X/N)^2 / (\sum X^2/N)$ – plotted for one period of the synchronous average (STD of noise = 0.02, speed variation = 20 %)

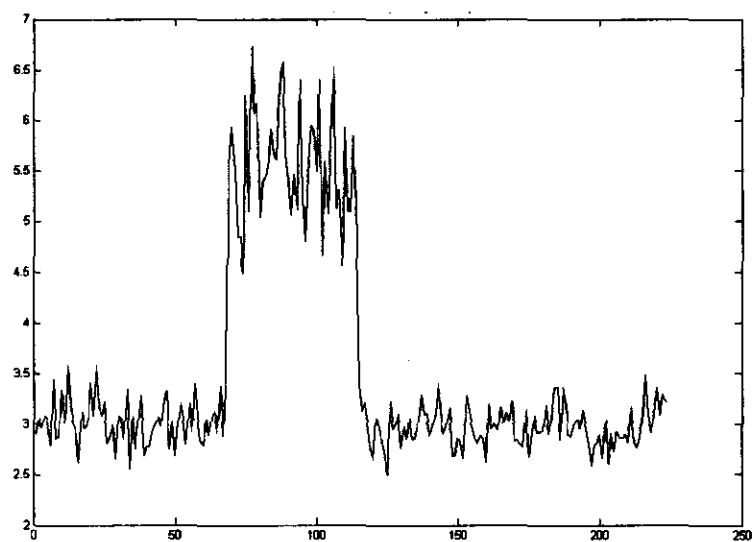


Figure 262: Kurtosis of samples of SA – plotted for one period of the synchronous average (STD of noise = 0.02, speed variation = 20 %)

11.11.5 Noise added, speed variation = 0%

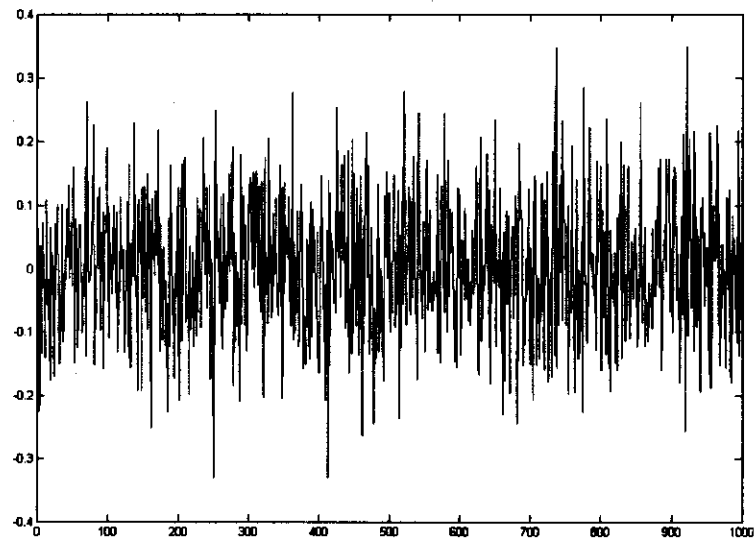


Figure 263: Time domain signal plotted for about five periods (STD of noise = 0.1, speed variation = 0%)

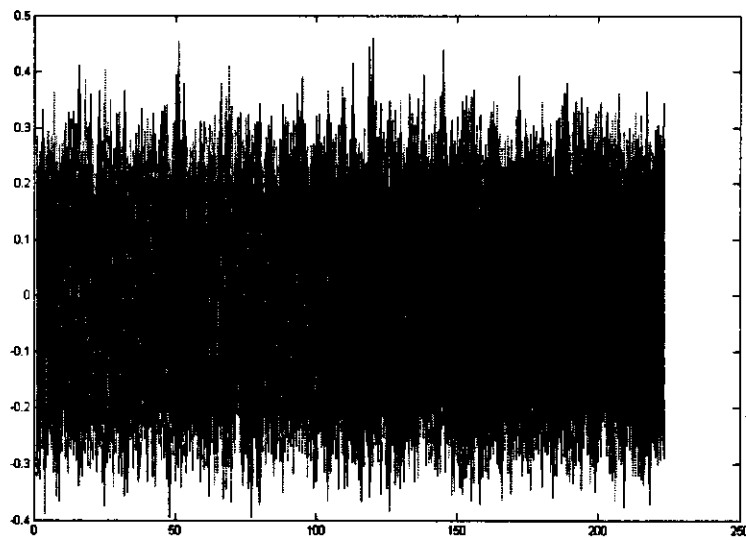


Figure 264: Time domain signals plotted over each other for one period (STD of noise = 0.1, speed variation = 0%)

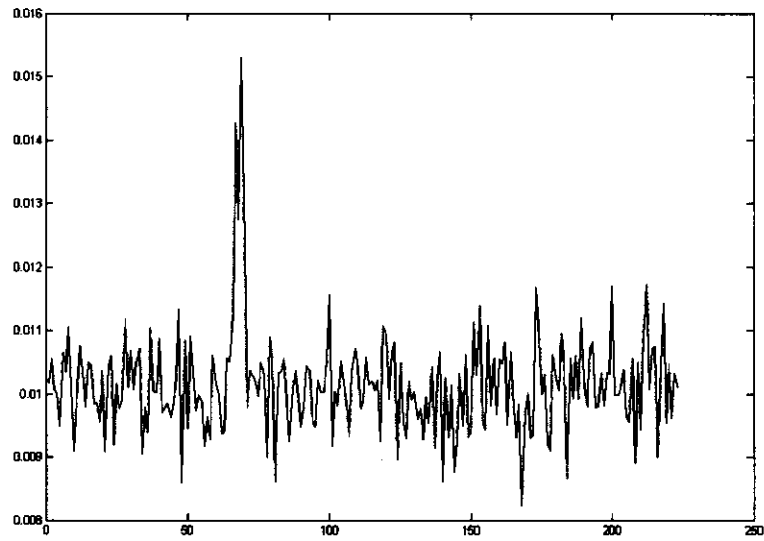


Figure 265: MOS $\Sigma(X^2)/N$ of time domain signal plotted for one period of the synchronous average (STD of noise = 0.1, speed variation = 0%)

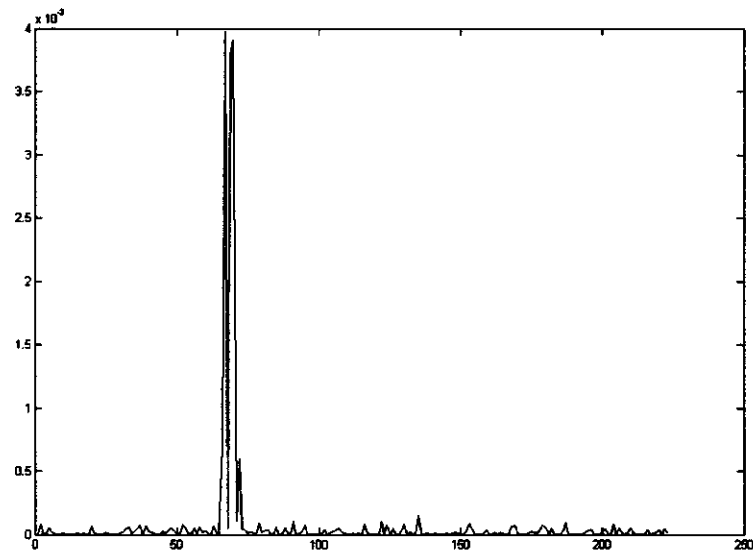


Figure 266: Mean squared – $(\Sigma X/N)^2$ – of time domain signal plotted for one period of the synchronous average (STD of noise = 0.1, speed variation = 0%)

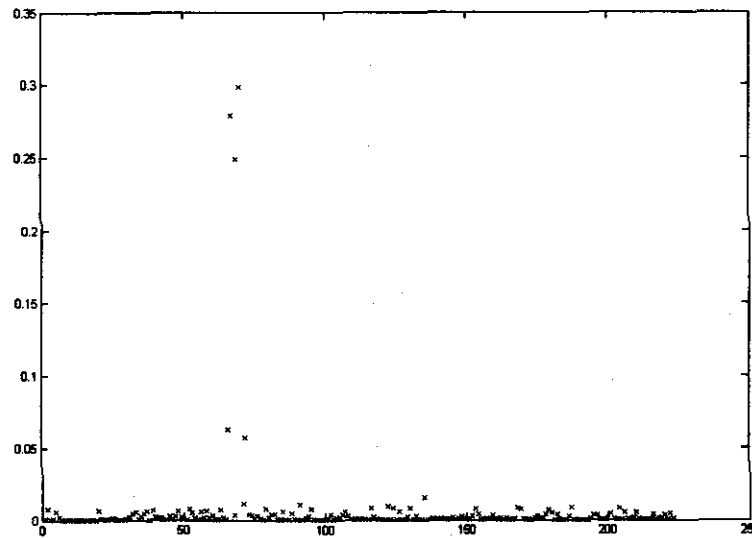


Figure 267: Ratio of mean squared to MOS of time domain signal – $(\sum X/N)^2/(\sum X^2/N)$ – plotted for one period of the synchronous average (STD of noise = 0.1, speed variation = 0%)

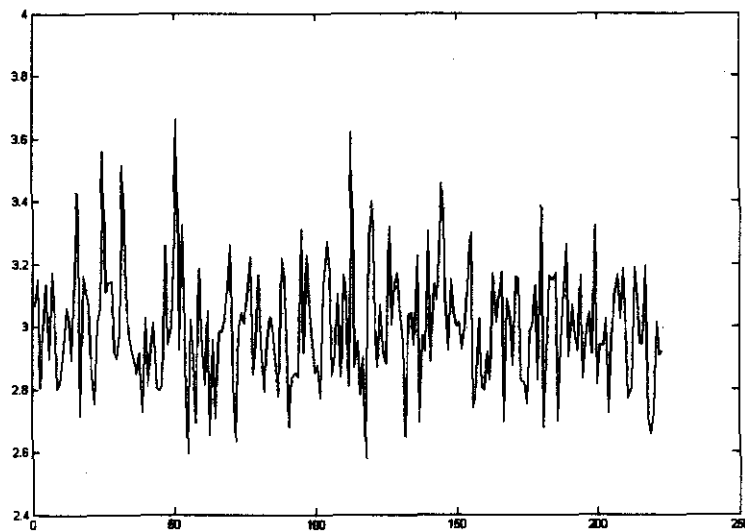


Figure 268: Kurtosis of samples of SA – plotted for one period of the synchronous average (STD of noise = 0.1, speed variation = 0%)

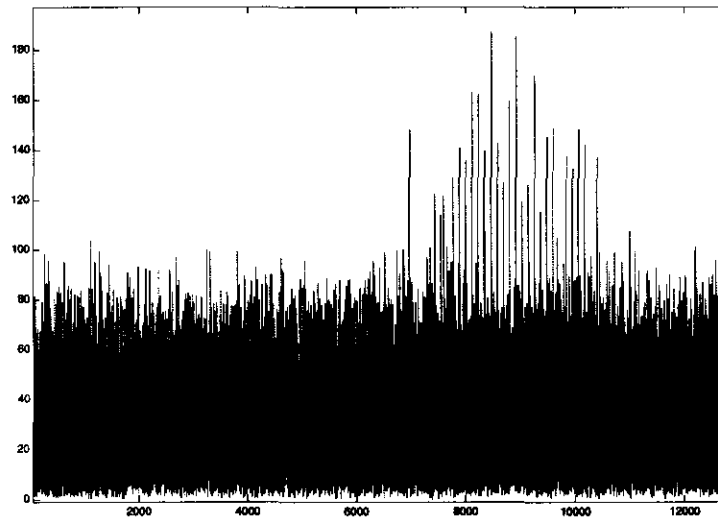


Figure 269: FFT magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0.1, speed variation = 0%)

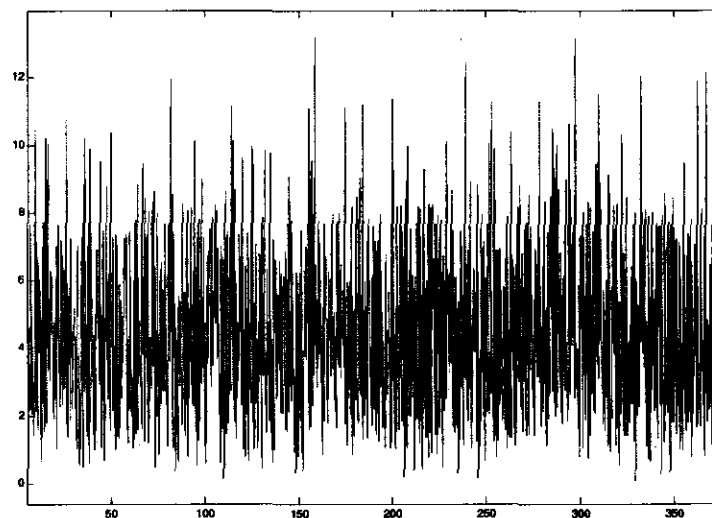


Figure 270: FFT demodulation magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0.1, speed variation = 0%)

Note that the repeating spectral lines equally spaced at multiples of the defect frequency of 114Hz in figure 269 are just visible above the broadband noise. The demodulation peak of figure 270 at 114Hz is also just visible above the background noise.

11.11.6 More noise added, speed variation = 0%

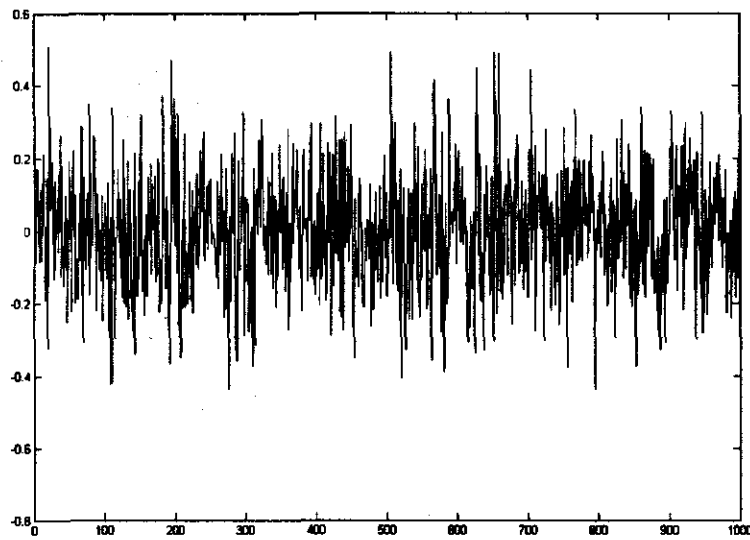


Figure 271: Time domain signal plotted for about five periods (STD of noise = 0.15, speed variation = 0%)

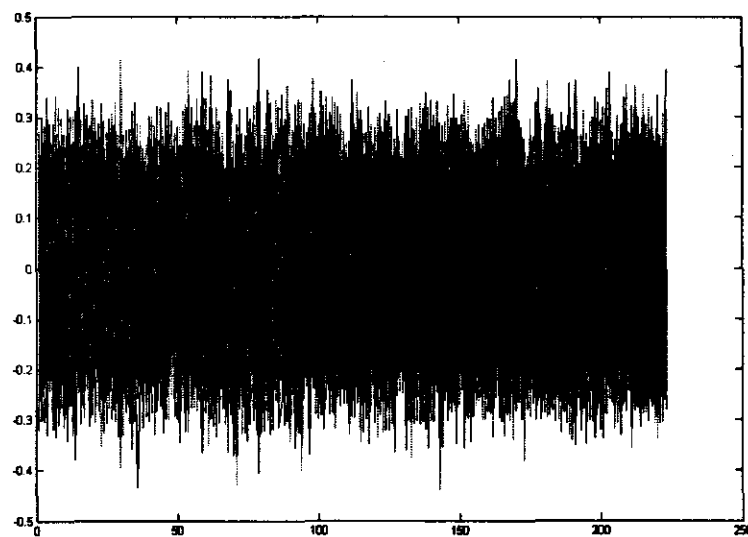


Figure 272: Time domain signals plotted over each other for one period (STD of noise = 0.15, speed variation = 0%)

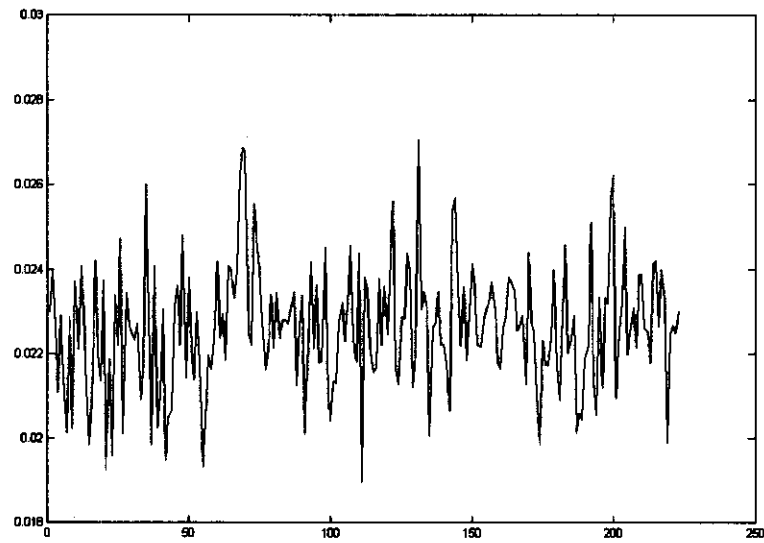


Figure 273: MOS $\Sigma(X^2)/N$ of time domain signal plotted for one period of the synchronous average (STD of noise = 0.15, speed variation = 0%)

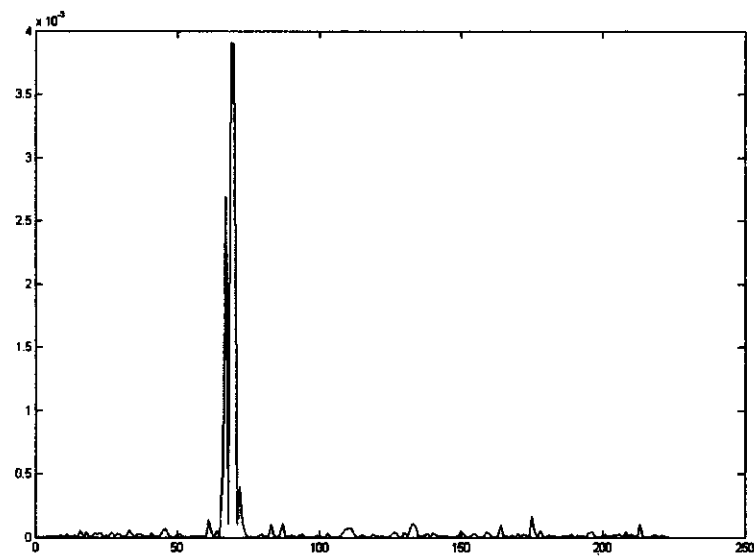


Figure 274: Mean squared – $(\Sigma X/N)^2$ – of time domain signal plotted for one period of the synchronous average (STD of noise = 0.15, speed variation = 0%)

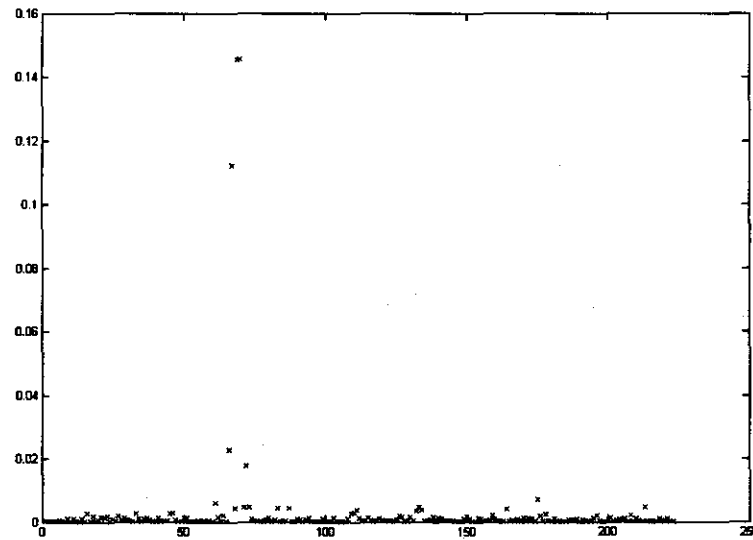


Figure 275: Ratio of mean squared to MOS of time domain signal – $(\Sigma X/N)^2/(\Sigma X^2/N)$ – plotted for one period of the synchronous average (STD of noise = 0.15, speed variation = 0%)

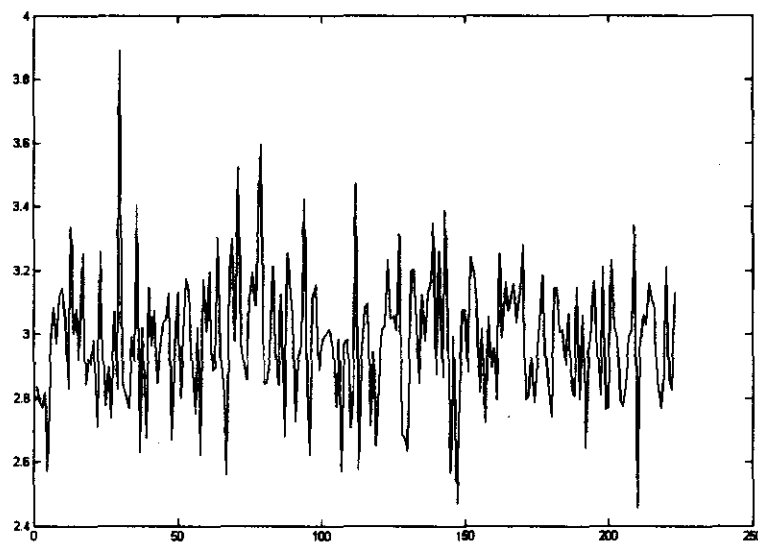


Figure 276: Kurtosis of samples of SA – plotted for one period of the synchronous average (STD of noise = 0.15, speed variation = 0%)

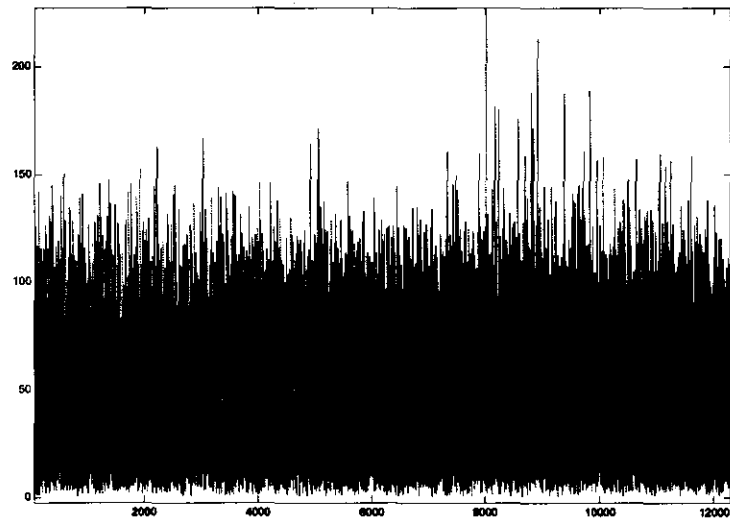


Figure 277: FFT magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0.15, speed variation = 0%)

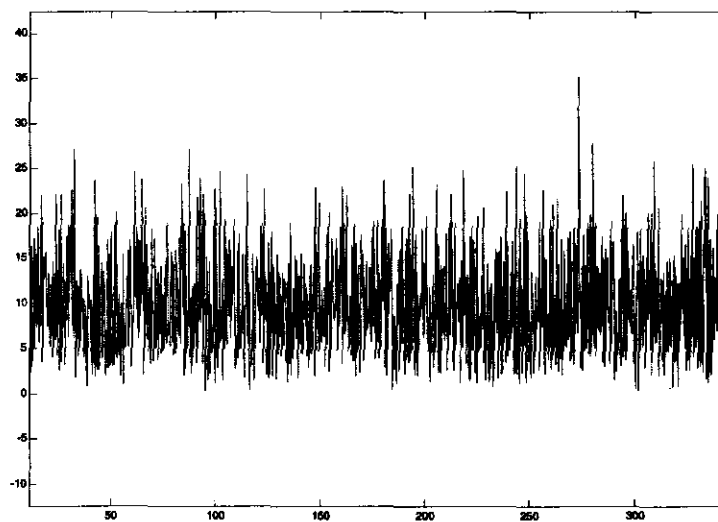


Figure 278: FFT demodulation magnitude spectrum of time domain signal, horizontal axis in Hz (STD of noise = 0.15, speed variation = 0%)

11.11.7 No noise added, linear chirp – speed variation = 0.4084%

Figure 279 shows a linear increase in the position of the simulated vibration signal¹³⁷. Figure 280 shows the usual FFT (the graph of figure 282 on p. 431 showing the log amplitude spectrum with speed variation may be compared with the graph of figure 29 on p. 194 with no speed variation). Note that it is easy to recognize the bearing defect frequency from visual examination of figure 280 and figure 285.

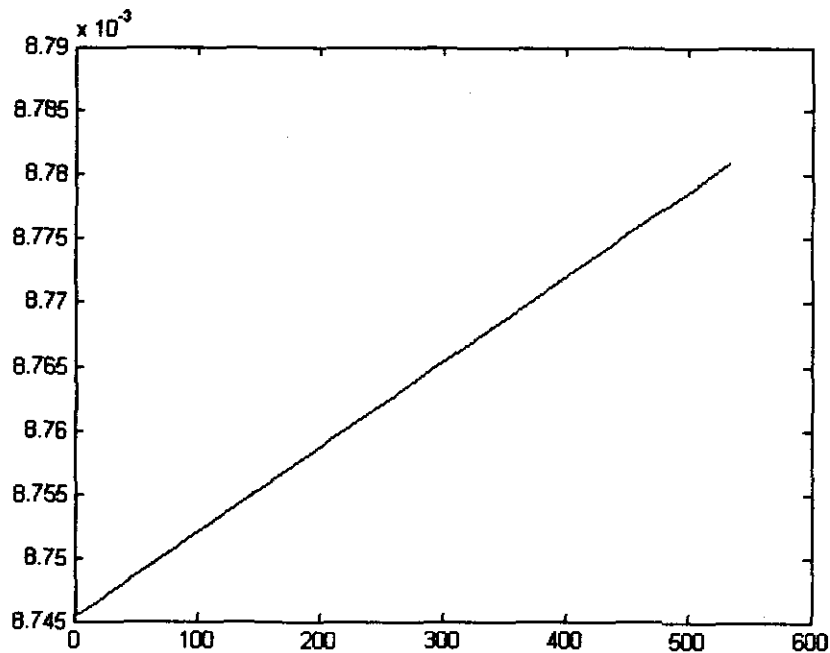


Figure 279: Position of the Gaussian pulse as a function of time (1/defect frequency = 0.008745 seconds)

¹³⁷ The linear chirp is used because the PDF of position against time is uniform. Any real vibration signal with bearing defect signal would also have a PDF that might be approximated by a suitable noise distribution. For modelling purposes the simplest assumption would be a uniform PDF.

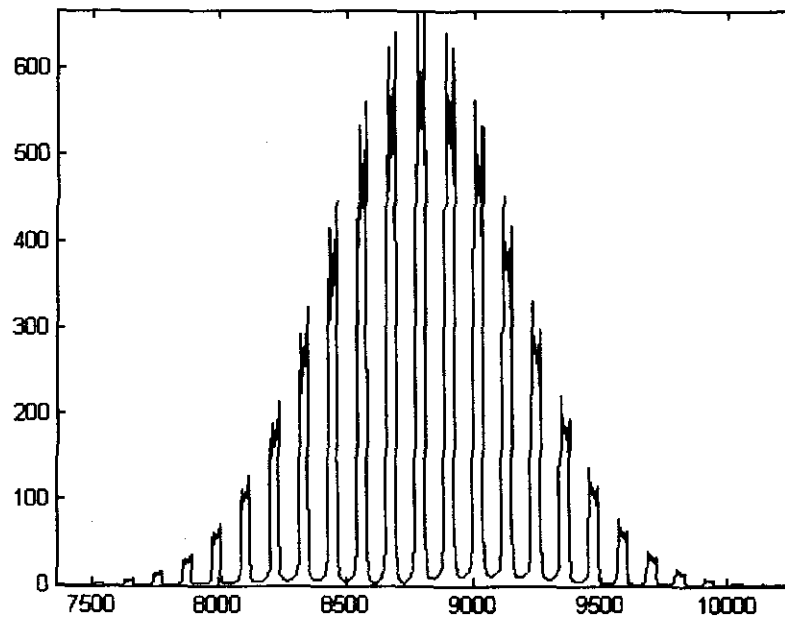


Figure 280: Amplitude spectrum of signal – linear increase of starting position of pulse in time, maximum variation = 0.41 %

The real cepstrum of figure 281 shows a large peak at twice the defect frequency, hence the effect of speed variations has been to confuse the interpretation of the peak at the defect frequency (which is close to zero). The reason for this strange behaviour can be easily seen in figure 282, which shows the log amplitude spectrum of the signal. The log operation amplifies small signals in the amplitude spectrum to such a degree that the resulting log amplitude spectrum resembles a square wave at twice the defect frequency. Thus it may be concluded that the real cepstrum is not suitable for bearing fault identification even with minor speed variations (also see section 6.12 on p. 192 for results on the cepstrum with no speed variations).

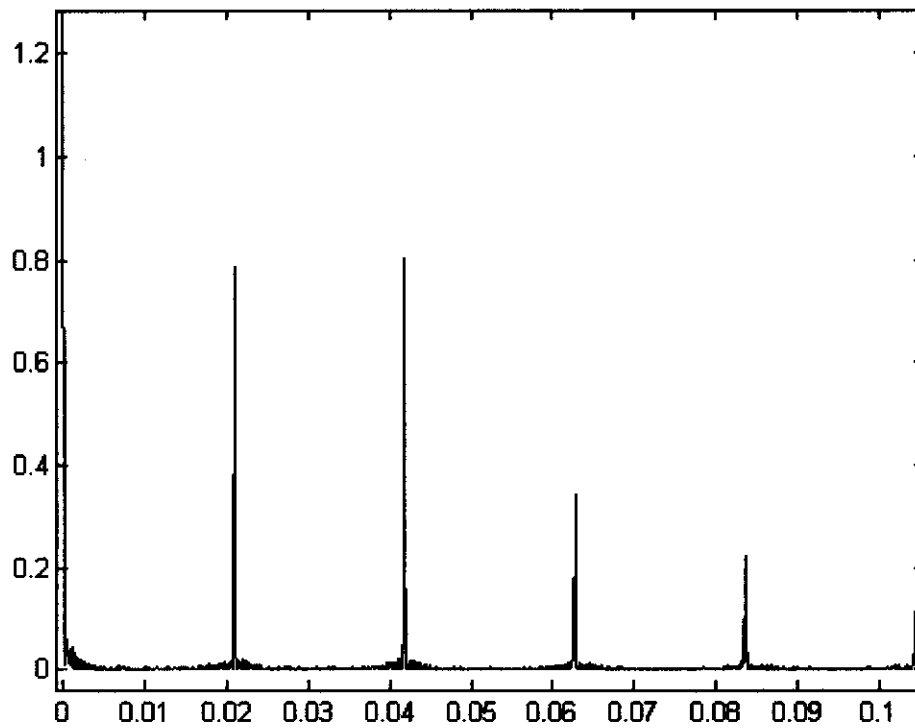


Figure 281: Real cepstrum of signal – linear increase of starting position of pulse in time, maximum variation = 0.41 %

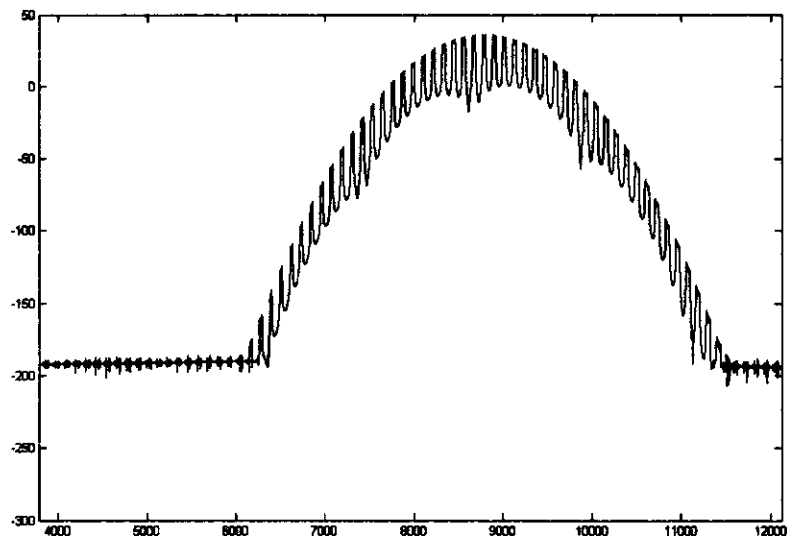


Figure 282: Log amplitude spectrum of signal – linear increase of starting position of pulse in time, maximum variation = 0.41 %

The complex cepstrum of figure 283 is likewise difficult to interpret unambiguously. Although a peak in the cepstral domain at 1/defect frequency is evident, the presence of other peaks complicates analysis.

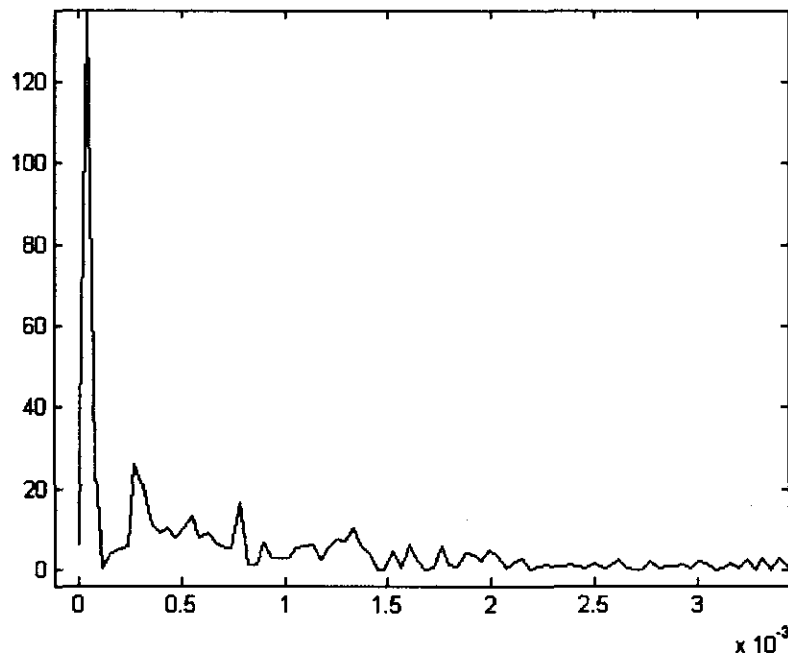


Figure 283: Complex cepstrum of signal – linear increase of starting position of pulse in time, maximum variation = 0.41 %

Figure 284 shows the proposed linear cepstrum (LCEPS) of the signal in which a reasonably clear demodulation peak is evident (no attempt has been made to optimize the results – the LTCEPS algorithm could be used, for instance).

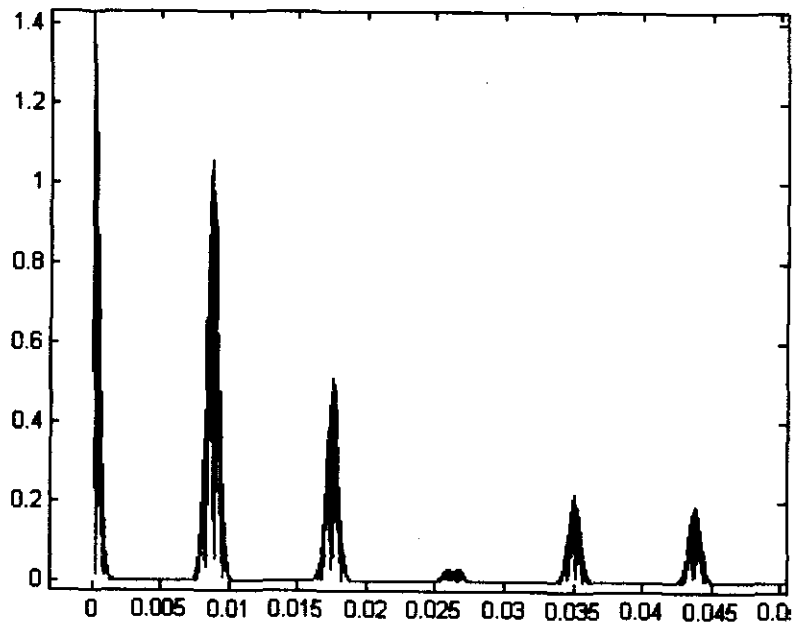


Figure 284: Linear cepstrum (LCEPS) of signal – linear increase of starting position of pulse in time, maximum variation = 0.41 %

Finally, figure 285 shows the amplitude spectrum of the signal squared (demodulation spectrum), in which a very clear demodulation peak is evident at the defect frequency.

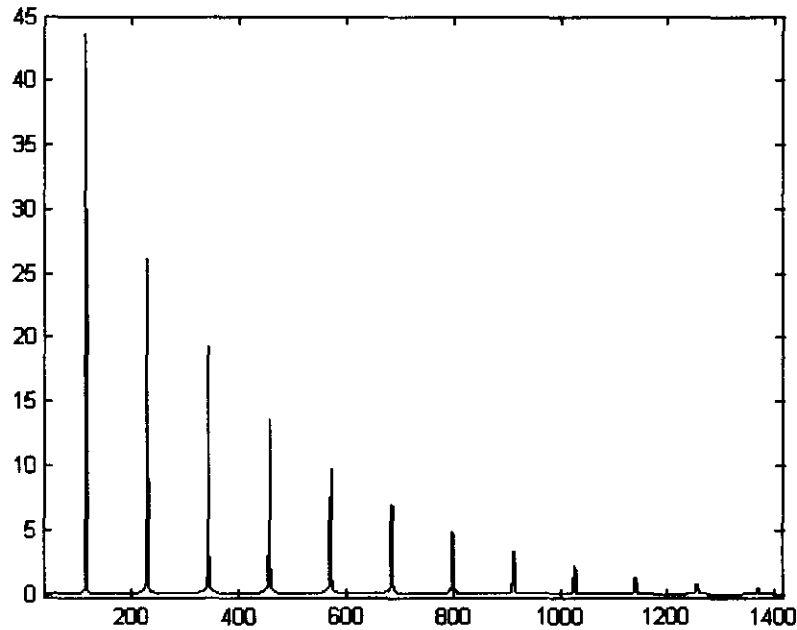


Figure 285: Amplitude spectrum of signal squared – linear increase of starting position of pulse in time, maximum variation = 0.41 %

The spectrum of the sampling sequence, a linear chirp in time, is shown in figure 286. Note that spectral smearing seems to increase towards the higher frequencies¹³⁸. Figure 287 shows a zoom plot – note that the peaks seem to increase in width at multiples of the defect frequency as the frequency increases. Figure 288 shows a zoom plot of the high frequency spectrum – note the amount of spectral smearing compared with figure 287. Also note from figure 287 that the width of the defect peaks increases with increasing frequency. Yet, the formula for spectral smearing of timing jitter in appendix E indicates a constant width of peaks with increasing frequency.

This anomaly may be explained as follows. A frequency difference of ΔF at a fundamental component becomes a difference of $N\Delta F$ at the N^{th} harmonic owing to a different effective

¹³⁸ This can be understood intuitively as follows. Consider the delta pulse sequence (with speed variation) as being composed of the sum of individual delta pulse sequences (with no speed variation) in which each delta pulse sequence D_k (with no speed variation) has a different period T_k . Owing to the linearity of the FFT the resulting spectrum will be the summation of the spectrum of the individual pulse sequences. The increasing period T_k of the pulse sequences results in shifting the delta pulse sequences to the left in the frequency domain. Hence spectral smearing resulting from speed variation becomes evident.

sampling rate. Let us now examine the conditions for spectral smearing to occur as a result of speed variation.

Consider first of all a delta impulse sequence $D_1(t) = \sum \delta(t - kT_1)$, where k is an integer and T_1 is the repetition period. Although the summation should ideally be to infinity, in any practical system with finite time buffers it has to be truncated at some point, for example $k = -M$ to M (where M is a finite integer). The effect is to multiply $D_1(t)$ by a rectangular window. In the frequency domain the spectrum of $D_1(t)$ will be convolved with the Fourier transform of the window function (sinc function). Let us assume that the time buffer is long enough for the finite time buffer to have negligible impact. Then the Fourier transform of $D_1(t)$ is $D_1(f) = 1/T_1 \sum \delta(f - k/T_1)$. In order to simulate the delta impulses of the signal $D_1(t)$ on a computer it is further necessary to substitute a signal of amplitude A , which is as large as possible, and width W , which is as small as possible (so that the area AW stays constant).

Now consider a new delta impulse sequence $D_2(t) = \sum \delta(t - kT_2)$ where $T_2 > T_1$. Now $D_2(f) = 1/T_2 \sum \delta(f - k/T_2)$, so this delta impulse sequence has impulses spaced closer together in the frequency domain.

Let us examine what conditions give rise to spectral smearing. This is frequency dependent. If $f_1 = 1/T_1$ and $f_2 = 1/T_2$ then $dF = f_1 - f_2$. The delta impulse sequence $D_1(t)$ will have impulses spaced at kf_1 , whereas the delta impulse sequence $D_2(t)$ will have impulses spaced at mf_2 . The question arises when $kf_1 = Cmf_2$, where k, m and C are integers. Hence $f_1/f_2 = C(m/k)$. Spectral smearing will result when $mdF = f_1$ or $m = f_1/dF$, where $dF = f_1 - f_2$. Similarly, spectral smearing results when $kdF = f_2$ or $k = f_2/dF$. Furthermore, $C = k/m$, $m = f_1/dF = f_1/(f_1 - f_2) = 1/(1 - f_2/f_1)$ and $k = f_2/dF = f_2/(f_1 - f_2) = 1/(f_1/f_2 - 1)$.

The above derivation indicates that complete spectral smearing occurs at a multiple $1/(f_1/f_2 - 1)$ of f_1 (or a multiple $1/(1 - f_2/f_1)$ of f_2). The closer f_1 and f_2 in frequency, the further away spectral smearing occurs in the frequency domain (in the limit of $f_1 = f_2$ no spectral smearing occurs, as expected). If $f_1 > f_2$ then $SV = (f_1/f_2 - 1) * 100$ is precisely the percentage speed variation. Hence spectral smearing occurs at $f_1/(SV/100) = 100f_1/SV$.

Thus the delta pulses of the sequence $D_2(t)$ will be spaced closer together in the time domain and at a frequency of $100f_I/SV$ the impulse of $D_2(t)$ will touch $D_1(t)$ exactly. It is reasonable to assume that the spectrum of the combined finite length sequences derived from $D_1(t)$ and $D_2(t)$ would be related to the sum of the spectrum of $D_1(t)$ and $D_2(t)$ ¹³⁹. In particular, it is clear that the width of the pulses associated with $D_1(t)$ and $D_2(t)$ increases as the frequency increases. The maximum width of the pulses ($T_{max} = 1/f_I$) associated with $D_1(t)$ and $D_2(t)$ is reached at a frequency of $100f_I/SV$. Consider a thought experiment in which k delta impulse sequences $D_k(t)$ are used (they are concatenated in the time domain one after the other). Assume for now that the minimum frequency is f_I and the maximum frequency is f_2 . If enough delta impulse sequences $D_k(t)$ are used the resulting spectrum will be the sum of the individual spectrum. Hence the (regular) spectrum would be almost flat between frequencies f_I and f_2 . However, the width of the pulses between $2f_I$ and $2f_2$ would almost double. The width of the pulses would reach f_I at a frequency of $100f_I/SV$ (where SV is the percentage speed variation). In figure 287 the width of the pulses can be seen to increase in proportion to frequency. The spectral smearing is also evident in figure 287.

¹³⁹ Assuming that the sequence $D_1(t)$ is placed first and the sequence $D_2(t)$ next, the effect of the rectangular windows multiplying $D_1(t)$ and $D_2(t)$ would have to be considered (also the effect of a shift in the rectangular window associated with $D_2(t)$). It is assumed in this thesis that long enough buffers are used in the time domain for the sinc functions in the frequency domain to be of short enough duration.

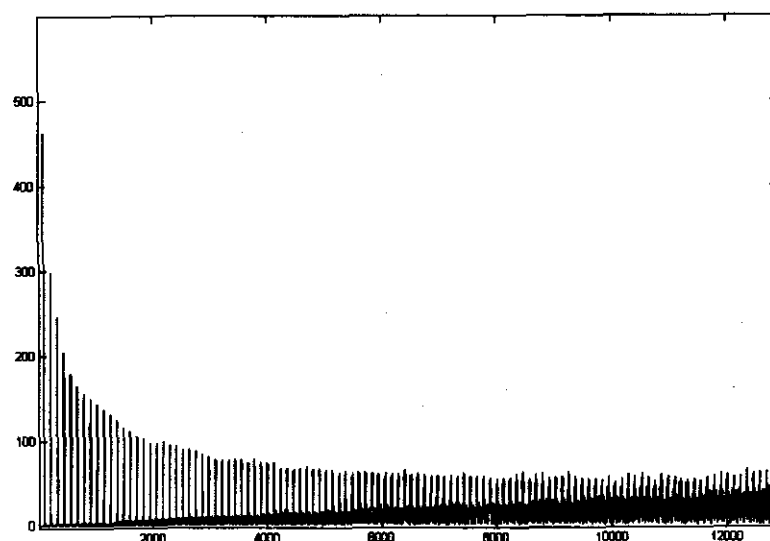


Figure 286: Spectrum of sampling sequence (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41 %

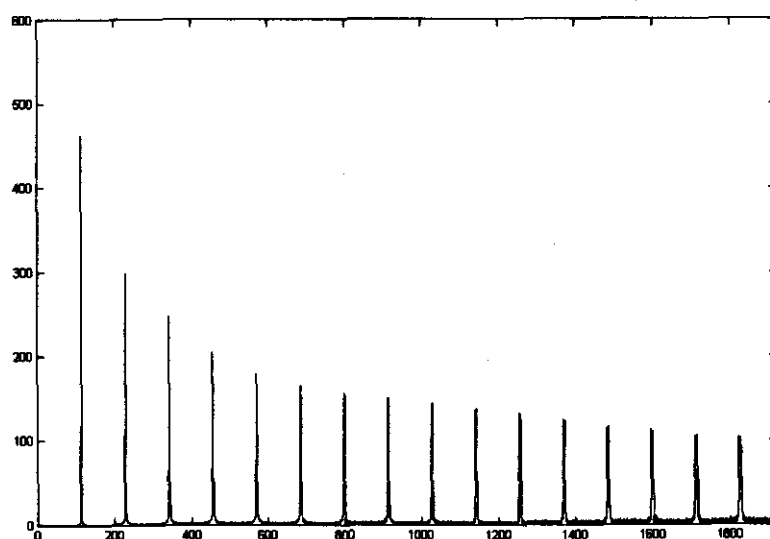


Figure 287: Spectrum of sampling sequence (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41 % (zoom plot)

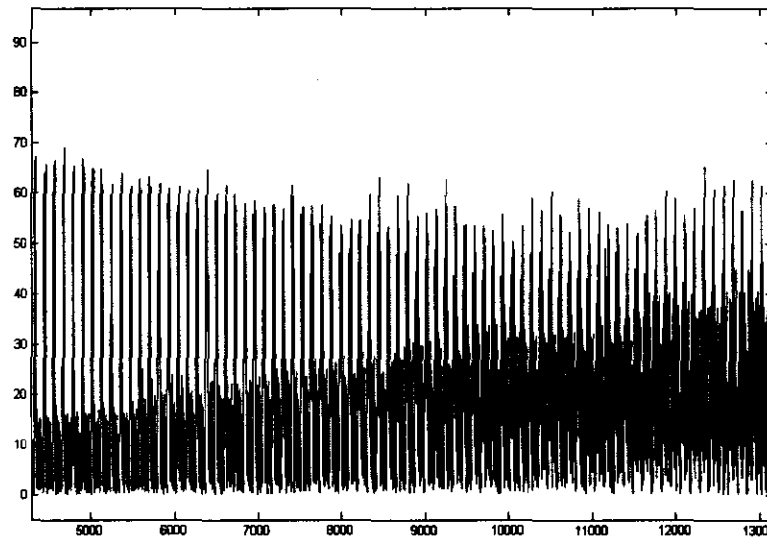


Figure 288: Spectrum of sampling sequence (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41 % (zoom plot)

The effect of the square law modulator will now be illustrated. Figure 289 shows that a narrow bandpass filter around 8000Hz is used to limit the number of convolutional terms that will be required for the computation of the demodulation spectrum. Note that the width of the peaks is about 30Hz (complete spectral smearing would occur around $100f_i/SV = 100 * 114.125/0.41 = 27.835\text{kHz}$, hence at 8kHz about $114.12/3.5\text{Hz} = 33\text{Hz}$ spectral smearing would be expected theoretically). Figure 290 shows the resulting demodulation spectrum of the squared signal. Observe the very narrow peak at the defect frequency of 114.12Hz (approximately 1Hz wide) to be contrasted with the frequency variation of 30Hz at 8kHz in figure 289.

Figure 291 shows the effect of removing the phase information from the amplitude spectrum of $X(f)$. The spectrum of the signal squared is given by $\text{FFT}(x^2) = X(f) \otimes X(f) = |X(f)|\angle X(f) \otimes |X(f)|\angle X(f)$. In figure 291 the phase information $\angle X(f)$ has been removed, which clearly illustrates the rather drastic effect on the demodulation spectrum. The convolution of the amplitude spectrum only, consisting of three shapes close to a rectangle, should provide a triangular function in theory, but a sharp peak is also evident at the defect frequency. Thus the phase information $\angle X(f)$ has a crucial role to play in the

convolution process. An examination of the phase spectrum of figure 293 does indicate random phase variations around 8000Hz and 8100Hz between $-\pi$ and π (in the areas where the magnitude of $X(f)$ is not zero). If the amplitude spectrum is approximated as three rectangular blocks, the effect of the phase variations would be to cancel nearly when all the vectors are summed together. This is illustrated in figure 297. The frequency is shown on the vertical axis, whereas the real and imaginary components of the FFT are illustrated on the two other axes. From 0Hz to before 8000Hz little change is evident in the value of the demodulation peak at 112.213Hz. Then a circle with a radius around $1e-3$ becomes evident and disappears very quickly (from the zoom plot of figure 298 probably around 8050Hz). The process is then repeated for the 'negative' frequencies of the FFT from 12 500Hz to 25 000Hz.

By contrast, figure 295 shows the value of the demodulation peak at 114.347Hz (which correlates well with figure 290), with frequency again depicted on the vertical axis. A clear increase in the value of the peak can be observed in the zoom plot of figure 296 from just before 8000Hz to around 8050Hz. Furthermore, no circles are evident as in figure 298 – instead an almost linear increase is evident in the value of the frequency component at 114.347Hz.

Figure 299 shows the individual convolutional products resulting from $X_2(f) = X(f) \otimes X(f)$ for a frequency of 114.347Hz. Note that the graph of the zoom plot is mostly in the first quadrant with positive values of the real and imaginary convolutional frequency components. Consequently, the summation of all the convolutional products can be expected to supply a large value. By contrast, figure 300 shows both positive and negative values for a frequency of 112.213Hz. Hence it is clear that many of the convolutional products would cancel, leading to a small value. Consequently, the phase information is very important.

Hence the operation of squaring does improve the visual appearance of the amplitude spectrum, but the exact reason is not presently known owing to the complexities of the convolution process. Furthermore, it is clear that the spectral smearing of the amplitude spectrum (without squaring of the signal) increases in proportion to the frequency. Even

though a large peak of almost 30Hz is evident at 8000Hz in the example of 0.41% speed variation, the demodulation peak (after squaring of the signal) is very narrow around the defect frequency (about 1Hz). This may be attributed largely to the phase spectrum, which seems to be important in the (periodic)¹⁴⁰ convolution of $X(f)$ with itself in the frequency domain. Thus the demodulation spectrum is effective in removing the noise associated with speed variations, even though it is perhaps not that easy to visualize in the frequency domain.

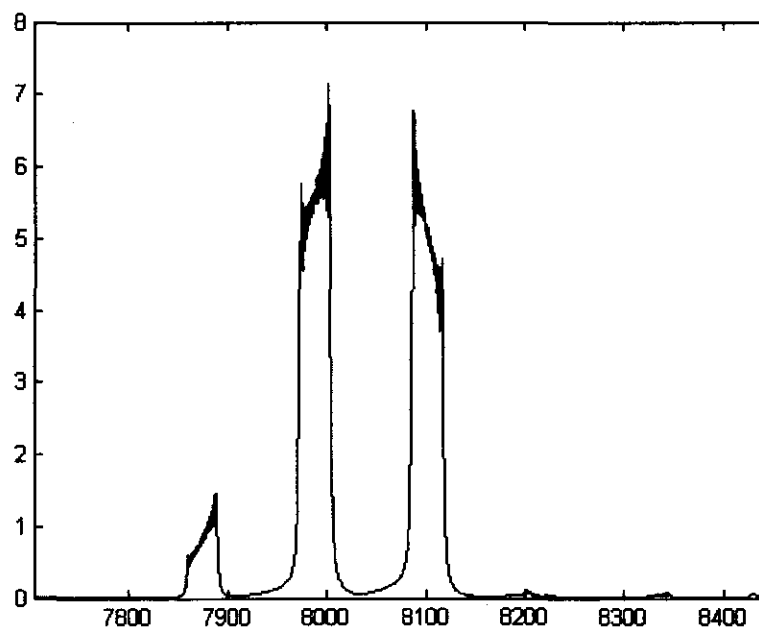


Figure 289: Amplitude spectrum after bandpass filter (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41 % (zoom plot)

¹⁴⁰ The linear convolution will not give the correct result, hence the circular or periodic convolution has to be used.

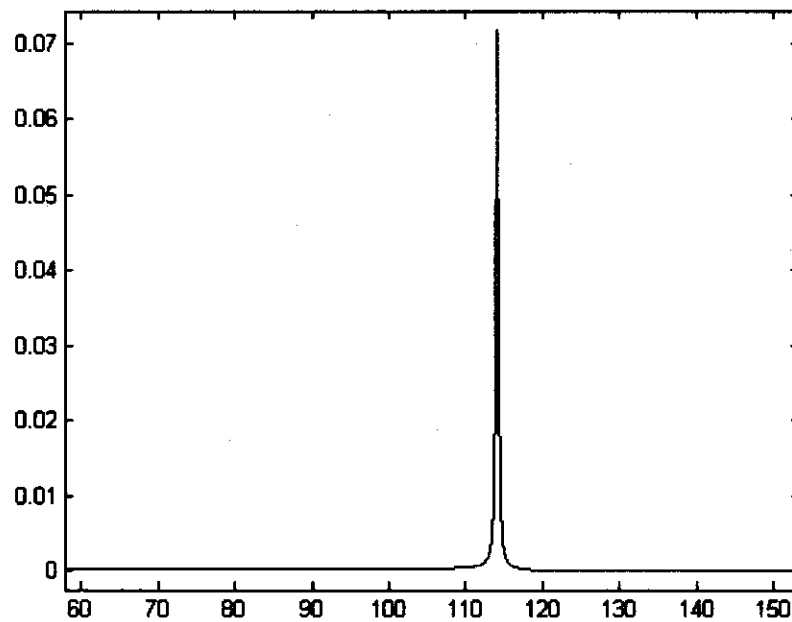


Figure 290: Amplitude spectrum of squared signal after bandpass filter (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41 % (zoom plot)

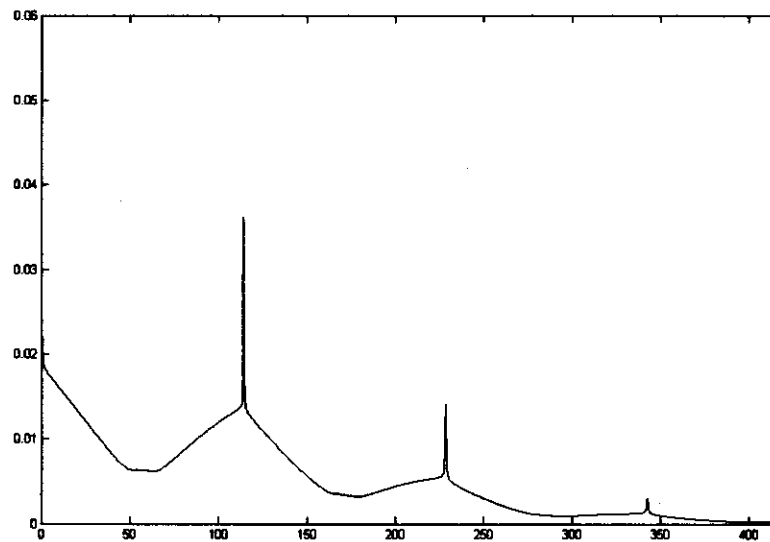


Figure 291: Amplitude spectrum of squared signal after bandpass filter and with removal of all phase information (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41 % (zoom plot)

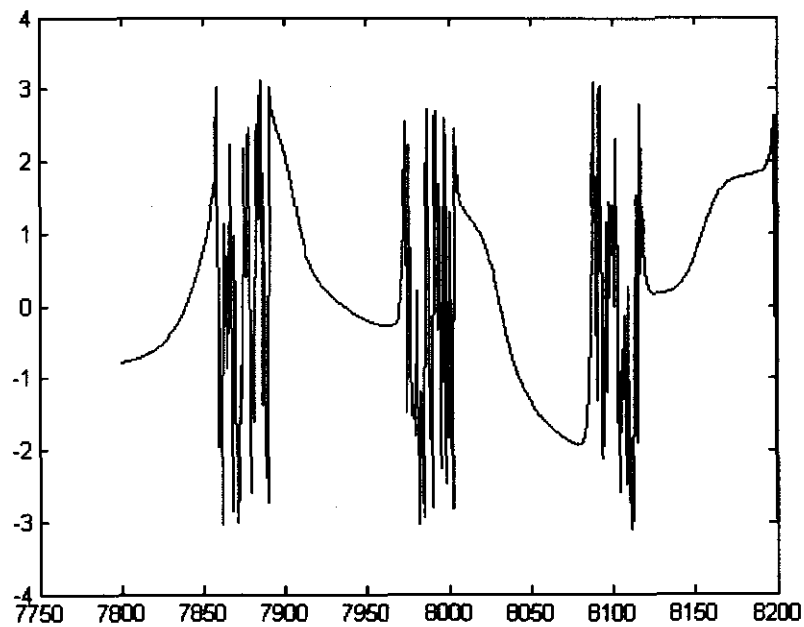


Figure 292: Phase spectrum of squared signal after bandpass filter (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41 % (zoom plot)

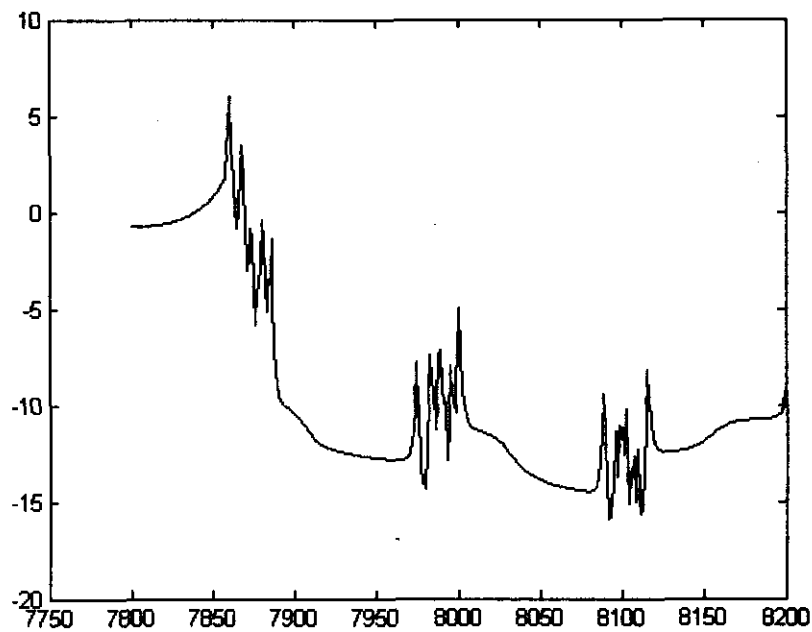


Figure 293: Phase spectrum (unwrapped) of squared signal after bandpass filter (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41 % (zoom plot)

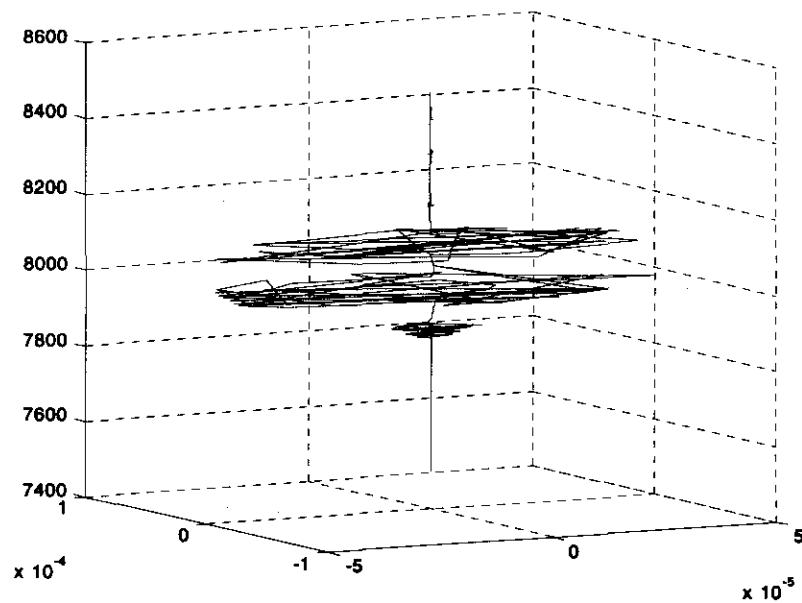


Figure 294: Combined amplitude and phase diagram of signal after bandpass filter (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41 % (zoom plot); the frequency axis is vertical (zoom plot) and the two horizontal axes give the real and imaginary components

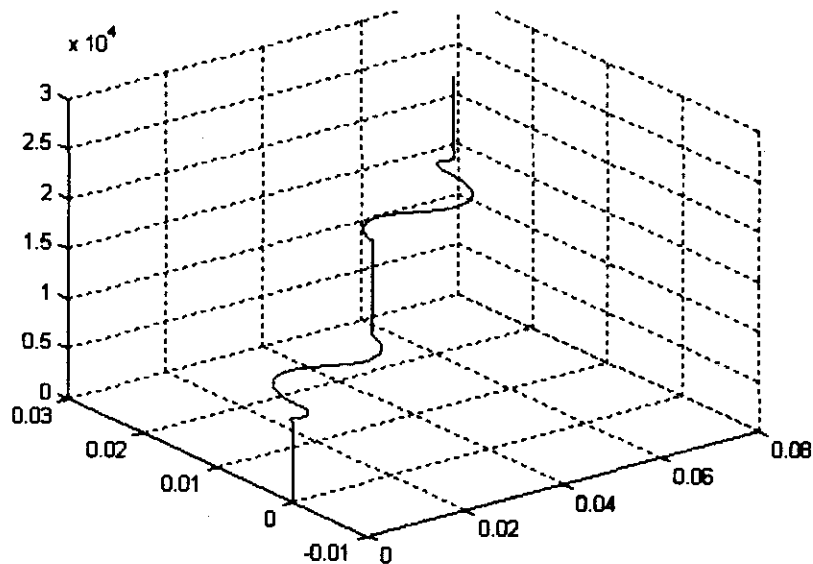


Figure 295: Combined amplitude and phase diagram of squared signal after bandpass filter for a frequency of 114.347Hz (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41 % (zoom plot); the frequency axis is vertical and the two horizontal axes give the real and imaginary components

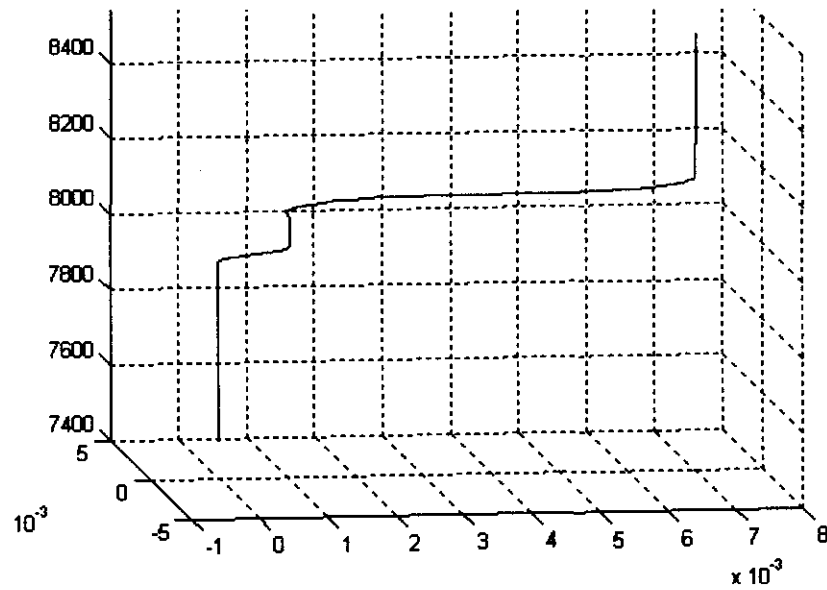


Figure 296: Combined amplitude and phase diagram of squared signal after bandpass filter for a frequency of 114.347Hz (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41 % (zoom plot); the frequency axis is vertical (zoom plot) and the two horizontal axes give the real and imaginary components

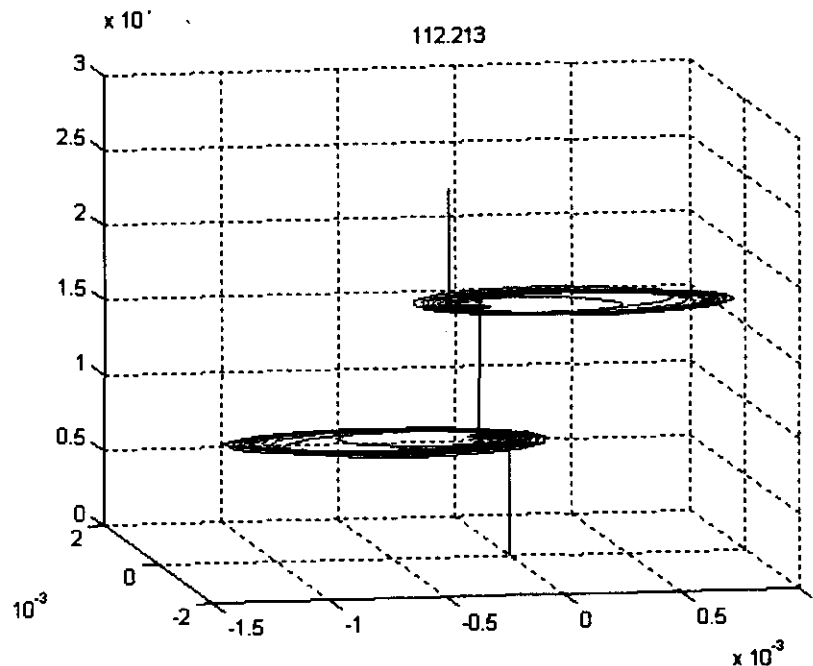


Figure 297: Combined amplitude and phase diagram of squared signal after bandpass filter for a frequency of 112.213Hz (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41 %; the frequency axis is vertical and the two horizontal axes give the real and imaginary components

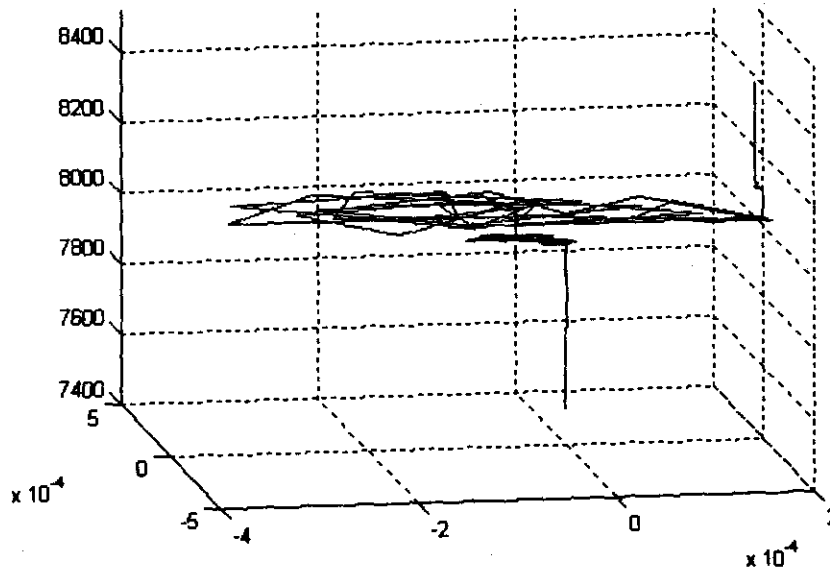


Figure 298: Combined amplitude and phase diagram of squared signal after bandpass filter for a frequency of 112.213Hz (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41 % (zoom plot); the frequency axis is vertical and the two horizontal axes give the real and imaginary components

114.347

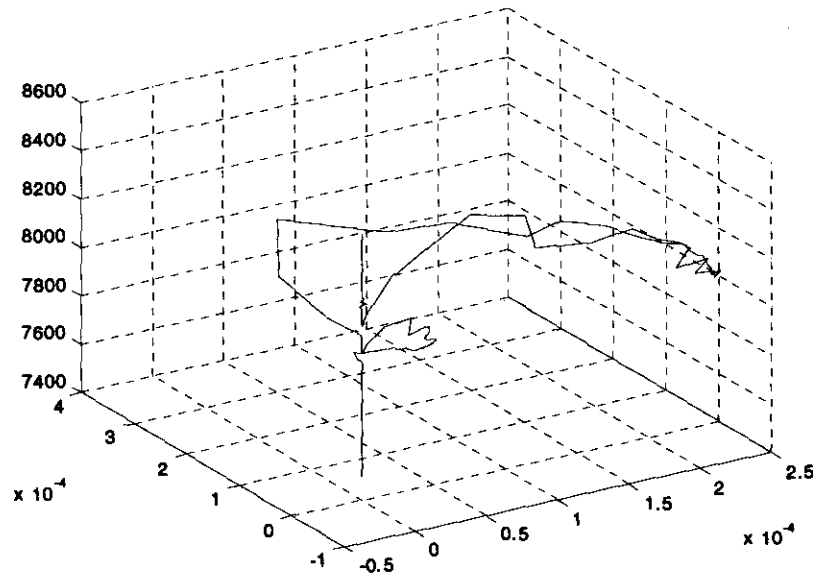


Figure 299: Combined amplitude and phase diagram (convolutional products) of squared signal after bandpass filter for a frequency of 114.347Hz (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41 % (zoom plot); the frequency axis is vertical (zoom plot) and the two horizontal axes give the real and imaginary components

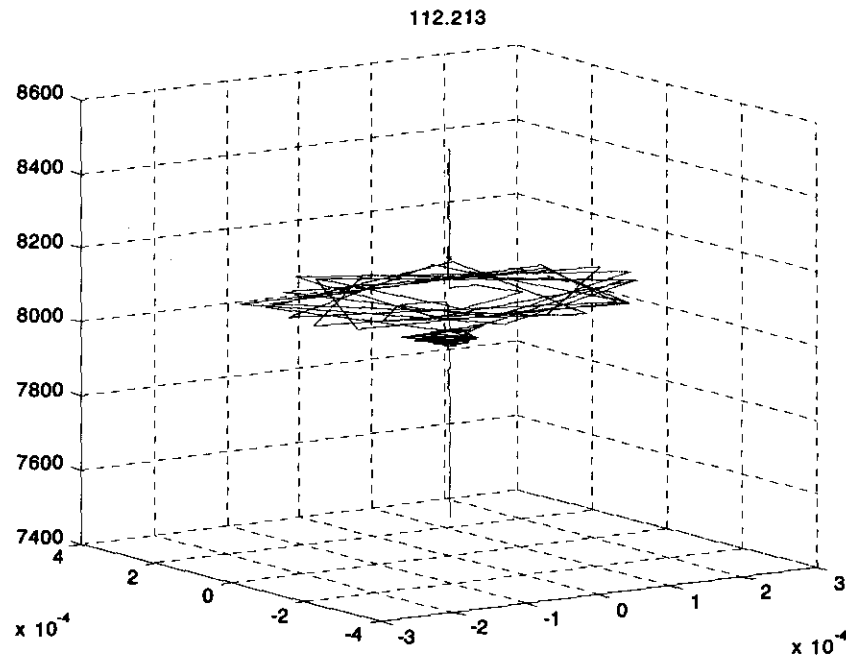


Figure 300: Combined amplitude and phase diagram (convolutional products) of squared signal after bandpass filter for a frequency of 112.213Hz (linear chirp with timing jitter) – linear increase of starting position of pulse in time, maximum variation = 0.41 % (zoom plot); the frequency axis is vertical and the two horizontal axes give the real and imaginary components

11.11.8 No noise added, linear chirp – speed variation = 62%

11.11.8.1 Results

Figure 301 shows a linear increase in the position in time of the simulated vibration signal. Figure 302 shows the usual FFT. Note that it is very difficult to recognize the bearing defect frequency from visual examination of figure 302. The cepstral operation is no longer successful, hence it is not shown.

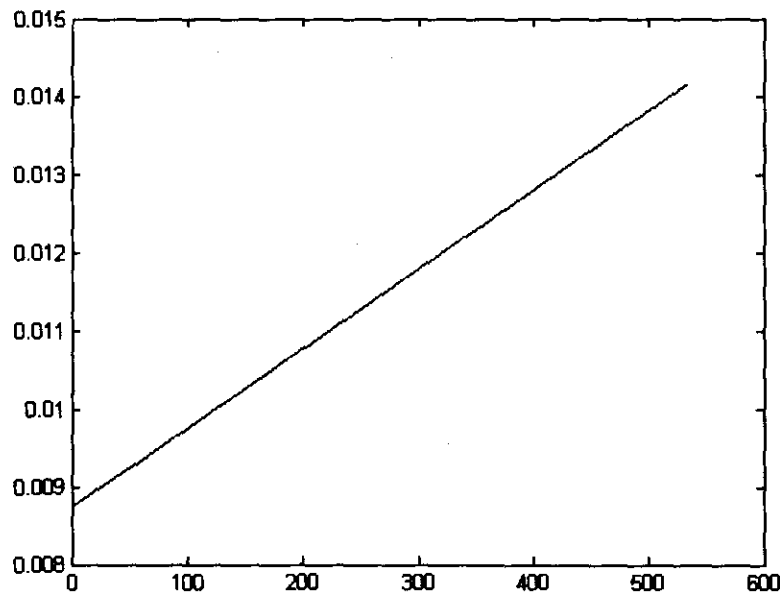


Figure 301: Position of the Gaussian pulse (refer to section 3.3) as a function of time ($1/\text{defect frequency} = 0.008745$ seconds)

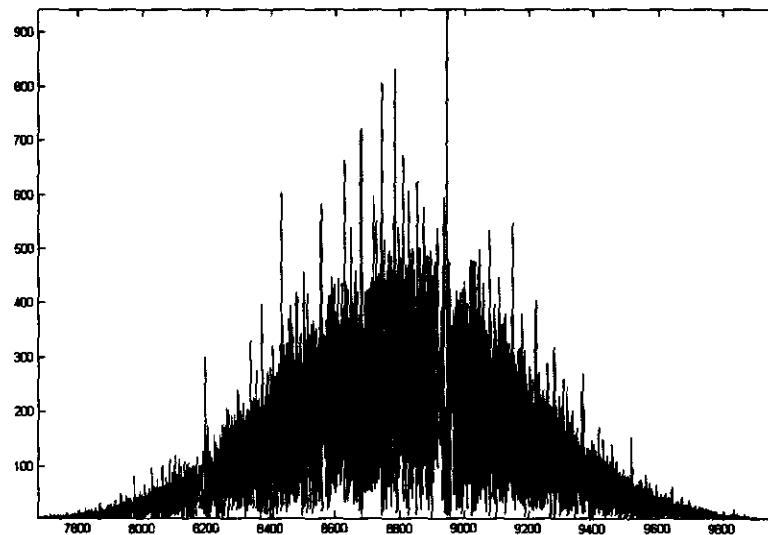


Figure 302: Amplitude spectrum of signal – linear increase of starting position of pulse in time, maximum variation = 62%

By contrast, figure 303 shows the demodulation spectrum (magnitude spectrum of the FFT of the signal squared), in which clear structure is evident around the fundamental bearing defect frequency. The peak around the defect frequency is very broad owing to the large speed variation (compare with section 6.12 on p. 192 with no speed variation).

11.11.8.2 *Theoretical analysis*

The results can be analysed theoretically as follows. The increase in the starting position of the period of the Gaussian pulse, which for easy analysis is a linear chirp in time, may be modelled as a small speed variation on a series of delta impulse sequences spaced at the defect period. Hence the results of appendix E may be applied if the speed variation is regarded as timing jitter with uniform probability density function (PDF). Thus the spectrum of the signal (the Gaussian pulse) will be smeared by the spectrum of the timing jitter. For a rectangular function in the time domain the spectrum of the timing jitter is a sinc function.

Convolution of the Gaussian pulse with the linear chirp in time provides the signal with speed variations. Consequently, the spectrum of the Gaussian pulse sequence (which is itself a Gaussian function) in the frequency domain is multiplied by the spectrum of the delta impulse sequence with timing jitter. The spectrum of the delta impulse sequence with zero timing jitter is a delta impulse sequence (for example [81]). It is postulated that the spectrum of the delta impulse sequence with timing jitter would be given by a convolution of the spectrum of the timing jitter with the delta impulse sequence (refer to appendix E), resulting in a smearing of the delta impulse sequence. This interpretation seems to be supported by figure 303, in which a broad peak with sinusoidal variations (probably owing to the sinc function of the uniform timing jitter) is evident around the defect frequency.

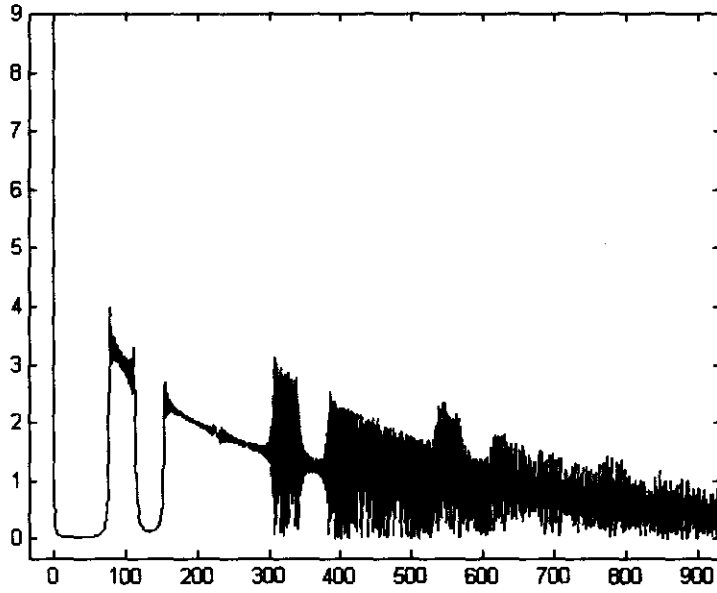


Figure 303: Amplitude spectrum of signal squared – linear increase of starting position of pulse in time, maximum variation = 62%

Convolution of the spectrum $X(f)$ with itself supplies the demodulation spectrum $X_2(f) = X(f) \otimes X(f)$. The spectrum $X(f) = P(f)D(f)$ is the result of multiplying the spectrum of the timing jitter $D(f)$ by the spectrum of the Gaussian pulse $P(f)$ (owing to convolution in the time domain). With no timing jitter the convolution of the periodic delta impulse sequence with the Gaussian pulse $p(t)$ is given by $D(f)P(f) = P(f = k/T)$, where T is the period of the delta impulse sequence, in other words a sampling of the spectrum $P(f)$ at regularly spaced (frequency period = $1/T$) intervals. Thus the demodulation spectrum is given by the (periodic or circular) convolution of $P(f = k/T)$ where $P(f = k/T)$. The resulting spectrum $P_2(f = k/T)$ will also be periodic with period T . Hence if $D(f) = \sum \delta f - n/T$ is the periodic delta impulse sequence (if the period between delta pulses is equally spaced $G(f)$ in appendix E is given by the delta function δf) and $P_2(f) = D(f)P(f) \otimes D(f)P(f) = \sum G(f - n/T)P(f) \otimes \sum G(f - n/T)P(f) = P(f = k/T) \otimes P(f = k/T)$.

Consequently, only specific frequency components (regularly spaced with period T) are convolved together. The spectrum $P_2(f)$ will only have components at frequencies $f = n/T$ (which is a direct result of assuming that the elementary pulse sequence $g(t)$ has no

variation over time). Note that for this special case of zero speed variation no advantage is obtained by using the square of the signal.

However, the regular sampling of $P(f)$ corresponds to the synchronous average of the signal $p(t)$ (refer to section 6.3). Hence in the time domain the synchronous average of the signal $p(t)$ is multiplied by itself.

If the signal has some timing jitter $d(t)$ and $G(f)$ is the spectrum of the timing jitter (refer to appendix E) then $P_X(f) = \sum G(f - n/T)P(f) \otimes \sum G(f - n/T)P(f)$.

Thus the demodulation spectrum $X_2(f) = X(f) \otimes X(f) = P(f)G(f) \otimes P(f)G(f)$.

11.11.8.3 Simplified example of sinusoidal signal with additive noise

To illustrate the characteristics of the non-linear operator in the presence of noise, a simple example will be presented. Instead of a series of impulses spaced at a defect frequency, consider a single sinusoidal waveform (with standard deviation STD = 1) in the presence of additive white noise (the STD of the noise will be varied). The white noise will be generated by using a random number generator to generate a signal $n(t)$. Then $N(f) = \text{FFT}(n(t))$ is computed, but the amplitude of $N(f)$ (which is $|N(f)|$) will not be a constant. Hence $N(f) = N(f)/|N(f)| * N_s$ will normalize the amplitude values to N_s . Thus the resulting time domain signal $n_s(t) = \text{real}(\text{IFFT}(N_s(f)))$ will have a white spectrum $N_w(f) = \text{FFT}(n_s(t)) = N_s$ for all frequencies f .

Let the sinusoidal signal be $x(t) = A \cos(\omega t)$ where $\omega = 2\pi f_c$ and let the noise signal have a white spectrum of magnitude B .

Then $X(f) = \text{FFT}(x(t)) = M * A_x/2 \delta(f - f_c) + M * A_x/2 \delta(f + f_c)$ would give a single peak at frequency f_c and amplitude $A/2 = MA_x/2$ ¹⁴¹. In addition, $N(f) = \text{FFT}(n(t)) = B/2$ would be a constant over frequency (in theory¹⁴²) with magnitude $B/2$.

¹⁴¹ A negative frequency component is also present, of course, but can be ignored. In Matlab© the forward FFT is not scaled, but the IFFT is scaled by $1/N$. Some prefer a symmetrical scaling of $1/\sqrt{N}$ for the forward and inverse FFT operations.

¹⁴² Any computation performed by the computer necessarily has to use a finite time buffer, corresponding to a convolution of the white noise spectrum with a sinc function.

Now $X_2(f) = \text{FFT}(x(t)^2) = X(f) \otimes X(f)$

$$= 1/(2M) * (M^2 A_x^2 / 2 (\delta f) + A_x^2 M^2 / 2 \delta f - 2f_c + A_x^2 M^2 / 2 \delta f + 2f_c))$$

$$= A_x^2 M / 4 (\delta f) + A_x^2 M / 4 \delta f - 2f_c + A_x^2 M / 4 \delta f + 2f_c))$$

Hence this is a peak of $1/4$ the amplitude of $A^2 = A_x^2 M / 4$ at twice the frequency f_c . Furthermore, $N_2(f) = \text{FFT}(n(t)^2) = N(f) \otimes N(f)$. But $N_2(0) = (B/2)^2 / M = \sum n(t)^2$ (mean square value of signal). However, because of the random phase variations in the phase spectrum of $N(f)$, the value of $N_2(2f_c)$ is unfortunately not that easy to compute analytically (because of the autoconvolution of $N(f)$), even though the spectrum of $N(f)$ itself is white (constant over frequency). If $N(f) = |N(f)| \angle N(f)$, then $N_2(f) = |N(f)|^2 \angle (\angle N(f) \otimes \angle N(f)) = B^2 / 4 \angle N_2(f)$ as $|N(f)| = B/2$, where the symbol $\otimes$ means convolution. Thus the random phase variations in $\angle N(f)$ will add to each constant amplitude value $|N(f)|^2$ in a rather complex way, so $N_2(f)$ will no longer have a flat spectrum (as $N(f)$ did).

It is clear that $R(f_c) = X(f_c) / N(f_c) = (A/2) / (B/2) = A/B$. Yet, $R_2(2f_c) = X_2(2f_c) / N_2(2f_c) = A_x^2 M / 4 / (B^2 / 4 \angle N_2(2f_c)) = A_x^2 / B^2 (M/4) / \angle N_2(2f_c)$

So the variance of $n(t) = \sum n(t)^2 / M = \sum N(f)^2 / M^2 = (B/2)^2 * 1/M^2 = (B/(2 * M))^2$ or $B = 2 * M * \sqrt{(\text{var}(n))} = 2 * M * \text{std}(n)$. Hence if the standard deviation of $n(t)$ increases by a factor of 2, $1/B$ would also decrease by a factor of 2 (assume the signal amplitude remains constant at A).

Figure 304 depicts the FFT of the sinusoidal signal $x(t)$. Figure 305 shows that the squaring operator enables easy detection of the sinusoidal signal at twice the frequency with the standard deviation of the noise $\text{STD} = 2$. Figure 305 also shows that the noise spectrum after the squaring operation is no longer white (constant over frequency)¹⁴³. Figure 306 illustrates that the signal plus noise can be detected easily (as expected). The signal value at f_c is $A/2 = A_x/2 * M = \sqrt{2}/2 * (5 * 1024) = 3620.4$, whereas the noise value

¹⁴³ This is as a result of the phase term, which contains random variation and influences the result owing to the convolution of $N(f)$ with itself. Setting the phase term to a constant is not a solution, as the IFFT will be an impulse in the time domain (which is, of course, no longer random).

at f_c is $B/2 = 143.1084$. Hence the detection ratio of signal-to-noise amplitudes for the usual FFT is $R(f_c) = 25.3$. After applying the square law modulator, the signal value in the spectrum of $x(t)^2$ is $A_x^2/4 * M = 2560$, whereas the noise value in the spectrum of $n(t)^2$ is 362.1. Thus the detection ratio after applying the square law modulator is only $R_2(2f_c) = 7.1$. The square law modulator decreased the overall detection ratio by $R_2(2f_c)/R(f_c) = 7.1/25.3 = 1/3.56$. Figure 306 also shows that the noise spectrum is nearly white (constant) over frequency (owing to the proposed normalization technique)¹⁴⁴.

Figure 307 shows an overall FFT of the square of the signal plus noise for a standard deviation of 5. A very large peak may be observed at DC (zero frequency). However, the frequency component is no longer clearly visible at twice the frequency of the sinusoidal signal x . Figure 308 zooms in on the lower frequency range to confirm this. The signal value at f_c is $A/2 = A_x/2 * M = \sqrt{2}/2 * (5 * 1024) = 3620.4$, whereas the noise value at f_c is $B/2 = 357.77$. Hence the detection ratio of signal-to-noise amplitudes for the usual FFT is $R(f_c) = 10.12$. After applying the square law modulator, the signal value in the spectrum of $x(t)^2$ is $A_x^2/4 * M = 2560$, whereas the noise value in the spectrum of $n(t)^2$ is 2262.9. Thus the detection ratio after applying the square law modulator is only $R_2(2f_c) = 1.13$. The square law modulator decreased the overall detection ratio by $R_2(2f_c)/R(f_c) = 1.13/10.12 = 1/8.95$. Note that the detection ratio has decreased by a factor of $8.95/3.56 = 2.5$, which is almost the increase of variance of $5/2 = 2.5$.

This is no coincidence. Increasing the standard deviation of the noise by a factor K will also increase the value of $N_2(2f_c)$ ¹⁴⁵ by K . For example, for a standard deviation of 10, which is twice the previous value of 5, the value of $R(f_c) = 3620.4/715.6 = 5.06$ and $R_2(2f_c)$ ¹⁴⁶ $= 2560/9051.6 = 0.2828$. Thus $R_2(2f_c)/R(f_c) = 1/17.8925$, corresponding to a doubling in the value. Alternatively, it is clear that the noise value after squaring has

¹⁴⁴ It is important also to consider the effect of the phase spectrum when convolving the signals in the frequency domain (in literature [2] the effect of the amplitude spectrum only is often considered).

¹⁴⁵ Strictly speaking, the 'average value' of $N_2(f)$ has been used in the results (excluding the large term $N_2(0)$ in the summation).

¹⁴⁶ The average value of $N_2(f)$ has been used (excluding $N_2(0)$).

increased from 2262.9 to 9051.6 (hence by a factor of 4), which corresponds to the square of the increase in standard deviation by 2. In another previous example the noise value after squaring has increased from 362.1 to 2262.9 (by a factor of 6.25), which corresponds to the square of the increase in standard deviation by 2.5.

In summary, it is not possible in this section to relate the absolute value of $N_2(2f_c)$ to the value of $N(f_c) = B/2$, but it is possible to indicate a relative change in values. An increase of K in the standard deviation of the noise will increase the value of $N_2(2f_c)$ by a factor of K^2 . The value of $R_2(2f_c)$ will decrease by K^2 , but the value of $R(f_c)$ just by K . Consequently, the value of the detection ratio $R_2(2f_c)/R(f_c)$ will change by $\text{std}(n)/(\text{std}(n)^2) = 1/\text{std}(n) = 1/K$, as indicated above. It is clear that $R(f_c) = 1$, in which case the signal and noise can no longer be discriminated using the FFT of the signal plus noise, when: $R(f_c) = X(f_c)/N(f_c) = (A/2)/(B/2) = (MA_x/2)/(\sqrt{M} * \text{std}(n)) = \sqrt{M}A_x/(2\text{std}(n)) = 1$.

Hence $R(f_c) = 1$ when $\text{std}(n) = \sigma_n = (\sqrt{M})A_x/2$, or equivalently when $\text{var}(n) = MA_x^2/4$ (where A_x is the amplitude of the sinusoidal signal and M is the number of samples in the FFT). A similar expression for the condition $R_2(2f_c) = 1$ is unfortunately not available at present and is left for future work. Table 16 summarizes the results (the last column shows the ratio of DR of different σ). Note that the critical variance when $R(f_c) = 1$ occurs when $\text{var}_c(n) = MA_x^2/4 = 5120 * 1/4$ or $\sigma_c = \sqrt{1280} = 35$. From table 16 this theoretical prediction almost matches the simulation results for $\sigma = 30$.

Table 16: Detection ratio (DR) as a function of additive noise σ (the symbol NA means not available)

σ	$R(f_c)$	$R_2(2f_c)$	DR $= R_2(2f_c)/R(f_c)$	$DR(m)/DR(n)$
1	3678.4/63.54 $= 57.9$	2674.9/90.9135 $= 29.42$	29.4/57.9 $= 0.51$	
2	25.3	7.1	7.1/25.3	$DR(1)/DR(2)$ $= 1.816$ $\approx (\sigma = 2)/(\sigma = 1)$

			$= 1/3.56$	$= 2$
5	10.12	1.13	$1.13/10.12$ $= 1/8.95$	$DR(2)/DR(5)$ $= 2.514 \approx$ $(\sigma = 5)/(\sigma = 2)$ $= 2.5$ $DR(1)/DR(5) =$ 4.56 $\approx (\sigma = 5)/(\sigma = 1)$ $= 5$
10	$4076/643$ $= 6.3$	NA		
15	$3500/954.5$ $= 3.7$	NA		
20	$3392/1282$ $= 2.65$	NA		
30	$2500/1923$ $= 1.3$	NA		

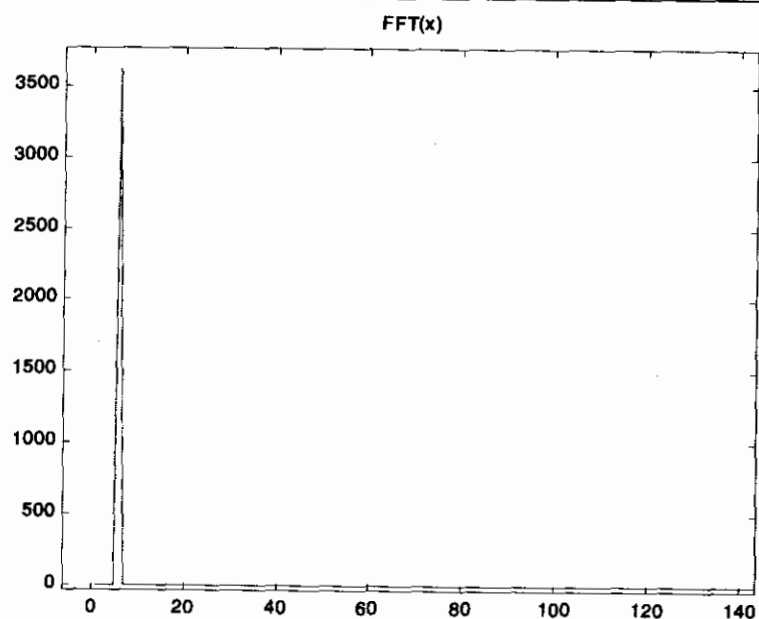


Figure 304: FFT of sinusoidal signal, $STD(x) = \sigma_x = 1$

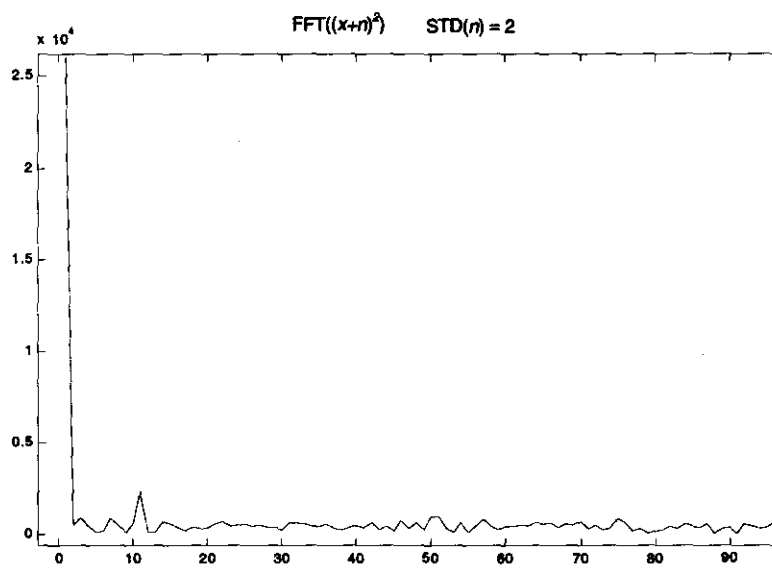


Figure 305: $\text{FFT}((x(t) + n(t))^2)$, $\text{STD}(n) = \sigma_n = 2$

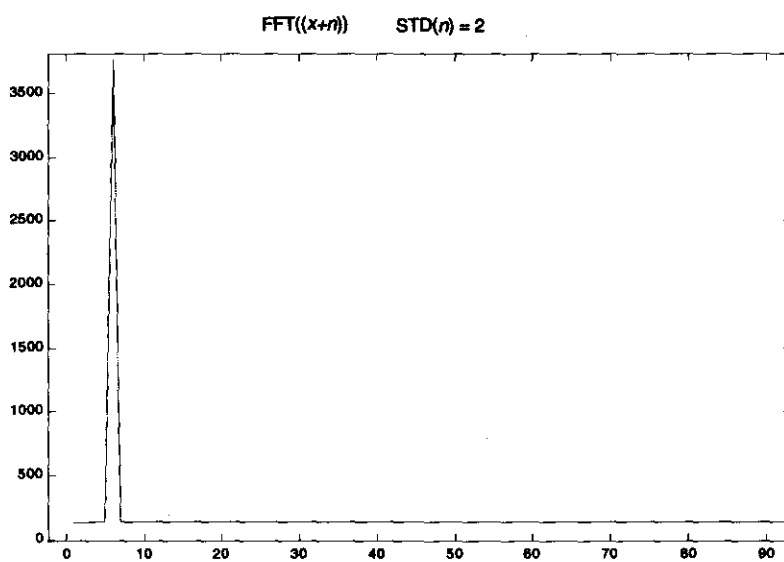


Figure 306: $\text{FFT}(x(t) + n(t))$, $\text{STD}(n) = \sigma_n = 2$

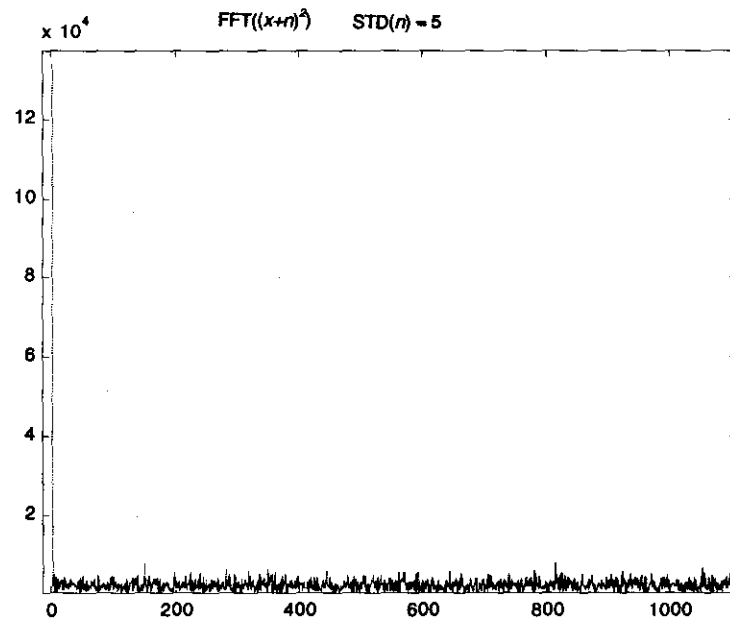


Figure 307: $\text{FFT}((x(t) + n(t))^2)$, $\text{STD}(n) = \sigma_n = 5$

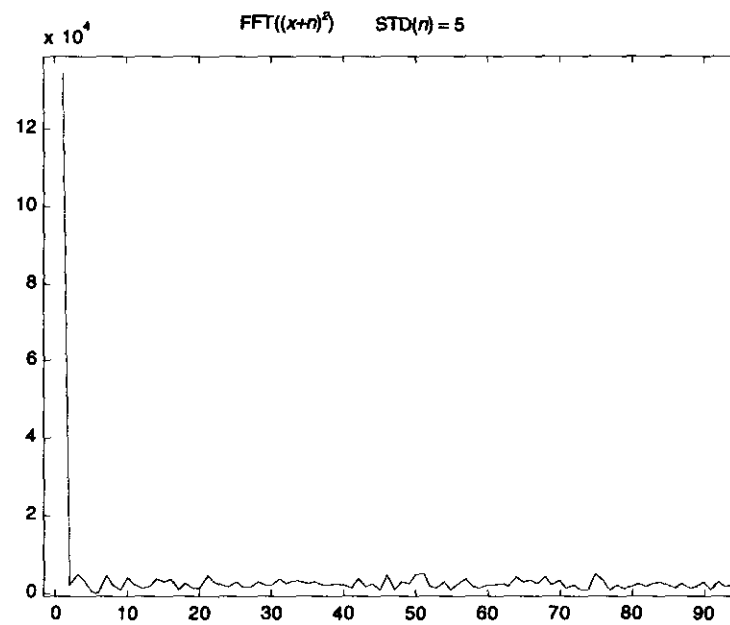


Figure 308: $\text{FFT}((x(t) + n(t))^2)$, $\text{STD}(n) = \sigma_n = 5$ (zoom plot)

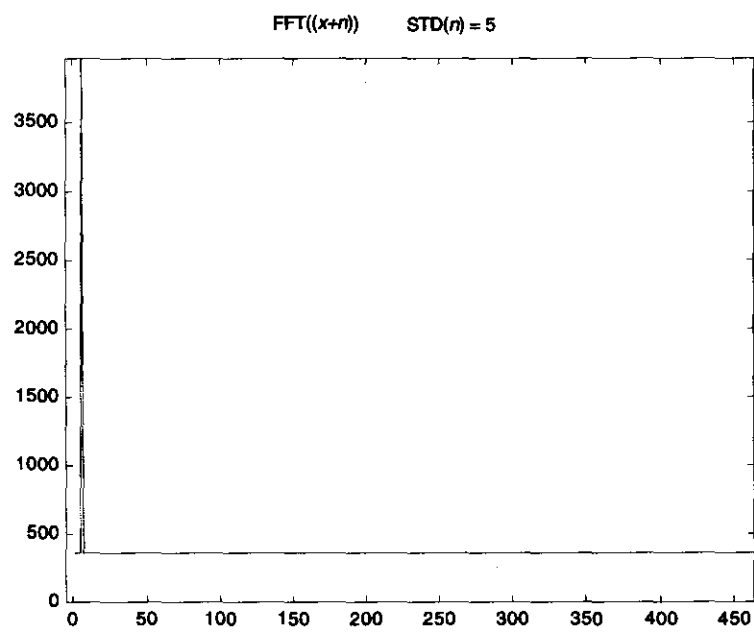


Figure 309: FFT($x(t) + n(t)$), STD(n) = $\sigma_n = 5$

11.11.9 Simplified vibration model

11.11.9.1 Introduction

A simplified vibration model is proposed using a parametric model. The influence of additive and multiplicative noise is illustrated using the regular and demodulated spectrum¹⁴⁷. Furthermore, the influence of speed variation is shown by a simple speed variation function consisting of a step change in frequency as well as a measured vibration signal. A theoretical expression for the detection ratio of the regular and demodulated spectrum is derived for the simplified (parametric) vibration model. It is shown under what conditions the demodulated spectrum outperforms the regular spectrum.

11.11.9.2 Simplified vibration model

To capture some of the behaviour of a real defect vibration signal a simplified vibration model is proposed

$$y(t) = \sin(\omega t)^p \quad (139)$$

The parameter p controls the spread in time of the model signal. As p is increased the signal becomes more localized. Choosing even values of p results in a positive function, eliminating the need for an absolute value function. Although the above equation is always smaller than or equal to one, it is essentially a series of pulses – quite similar to an ideal impulse sequence. In the frequency domain the FFT of y is given by

$$Y(j\omega) = X(j\omega) \otimes X(j\omega) \otimes X(j\omega) \dots X(j\omega) \quad (140)$$

where $X(j\omega) = \mathcal{F}(\sin(\omega t))$ and the number of convolutional terms in the above equation is given by p (assuming that p is an integer). The spectrum of $X(j\omega)$ is, of course, a single spectral component at frequency ω (using a single-sided spectrum for illustration). As the function $X(j\omega)$ is convolved with itself the bandwidth of the resultant increases. A series of harmonic lines spaced at frequency ω_t (where ω_t is the carrier frequency of the sinusoidal

¹⁴⁷ The demodulated spectrum is the signal squared (after applying a bandpass filter to accentuate the high frequency resonances).

signal) results. For large values of p a close approximation to the ideal impulse sequence results. To further increase the width of the time domain signal $y(t)$ a bandpass filter may be applied (convolution in the time domain with the impulse response of the filter and multiplication by a frequency window in the frequency domain). By modulating the amplitude and frequency (phase) of the sine wave the effect of speed variations may be analysed theoretically. The energy of the model signal slowly decreases as a function of p (for values of p larger than 2^{18} the energy essentially stays constant).

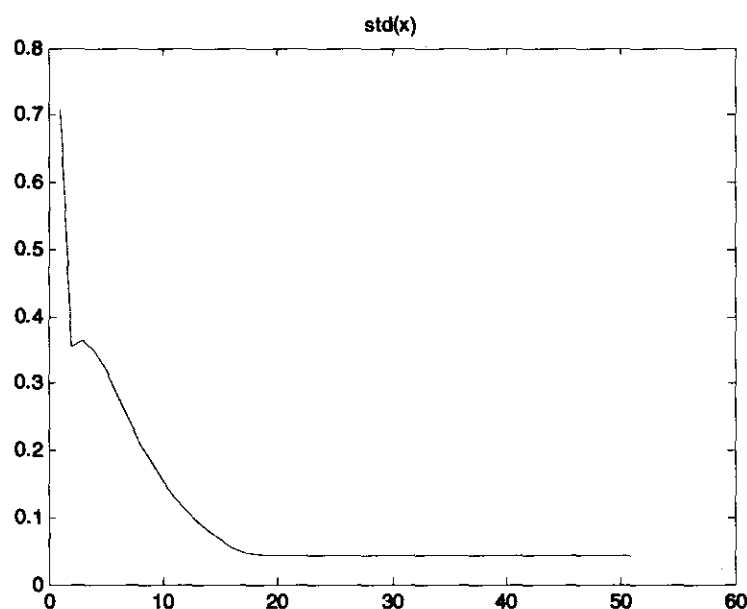


Figure 310: Standard deviation of model signal as function of 2^p

As 2^p is always even no absolute value function is required.

11.11.9.3 *Detection ratio of the simplified vibration model*

In figure 311 the ratio of the noise in the (FFT of the) squared signal relative to the noise in the (FFT of the) signal is shown as a function of the noise standard deviation. For $\sigma_n = 20$ it is evident that the noise in the demodulated spectrum is about 30 times as large as the noise in the regular spectrum.

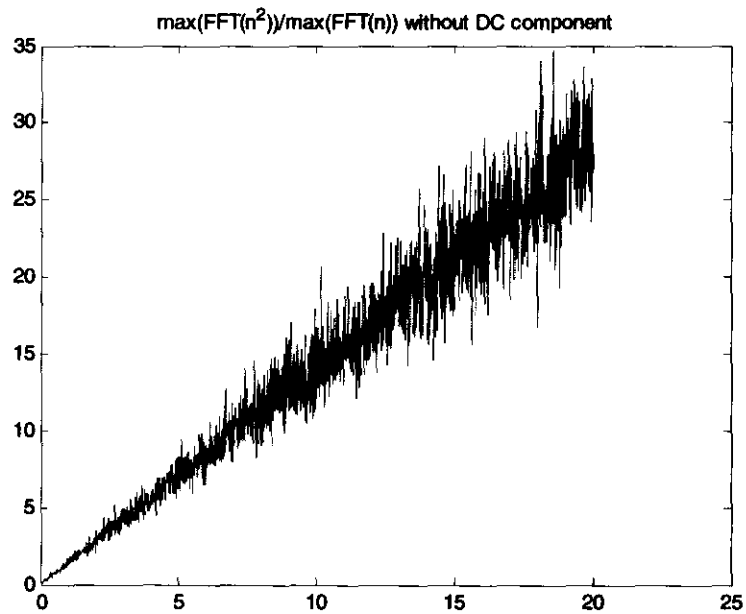


Figure 311: Ratio of $\text{FFT}(n^2)/\text{FFT}(n)$ (ignore DC component in spectrum) using maximum value function

In figure 312 a zoom plot of the ratio of the noise in the (FFT of the) squared signal relative to the noise in the (FFT of the) signal is shown as a function of the noise standard deviation. From values of about $1/\sqrt{2}$ it is evident that the noise in the demodulated spectrum is smaller than the noise in the regular spectrum.

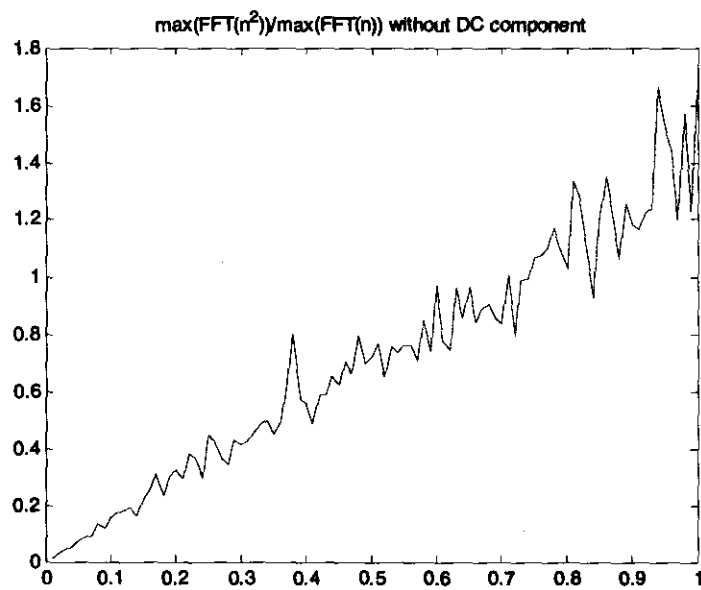


Figure 312: Ratio of $\text{FFT}(n^2)/\text{FFT}(n)$ (ignore DC component in spectrum) using maximum value function – zoom plot

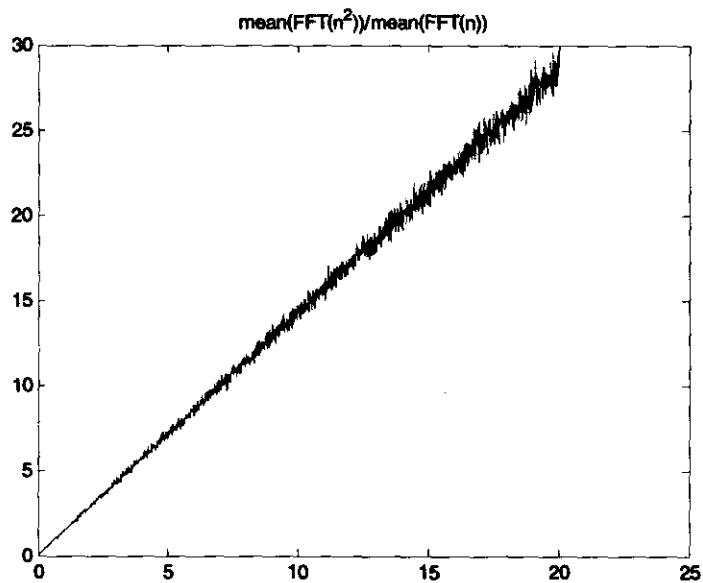


Figure 313: Ratio of $\text{FFT}(n^2)/\text{FFT}(n)$ (ignore DC component in spectrum) using mean value function

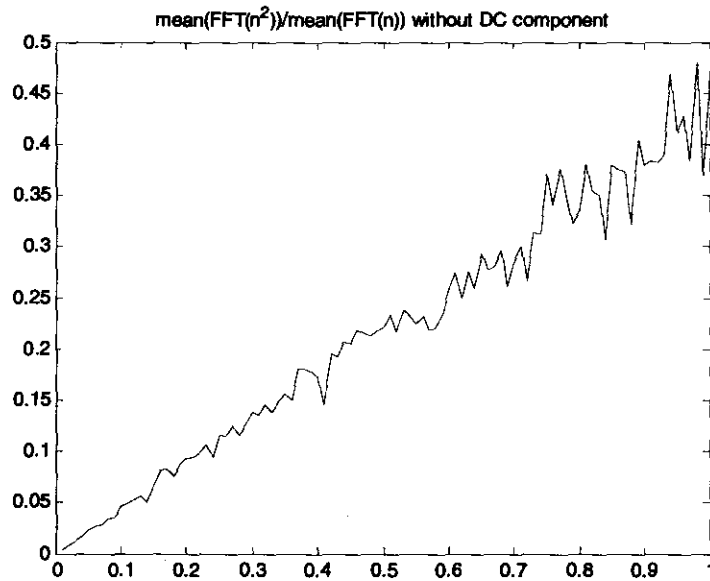


Figure 314: Ratio of $\text{FFT}(n^2)/\text{FFT}(n)$ (ignore DC component in spectrum) using the mean value function – zoom plot

The ratio of the (FFT of the) squared signal relative to the (FFT of the) signal is shown in the following figure as a function of the parameter p (where the signal $y = K_{\text{Sig}}(x/\sqrt{2})^p$) for a signal amplitude of $K_{\text{Sig}} = 10$:

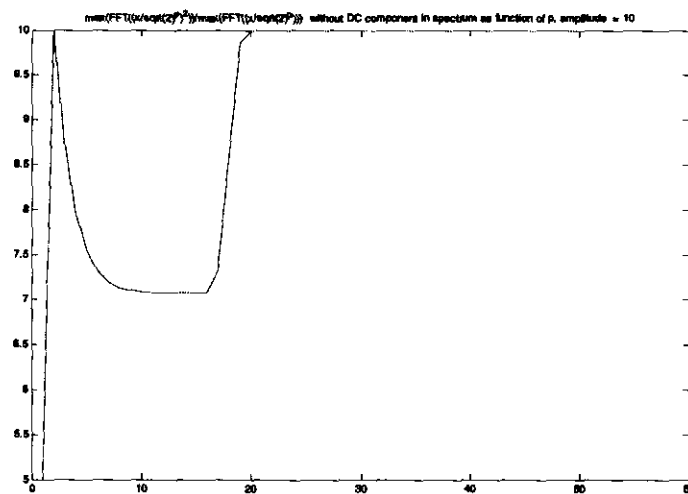


Figure 315: Ratio of $\text{FFT}(n^2)/\text{FFT}(n)$ (ignore DC component in spectrum) as a function of model parameter p

The ratio of the (FFT of the) squared signal relative to the (FFT of the) signal is shown as a function of the parameter p (where the signal $y = K_{Sig}(x/\sqrt{2})^p$) in figure 316 for a signal amplitude of $K_{Sig} = 1$:

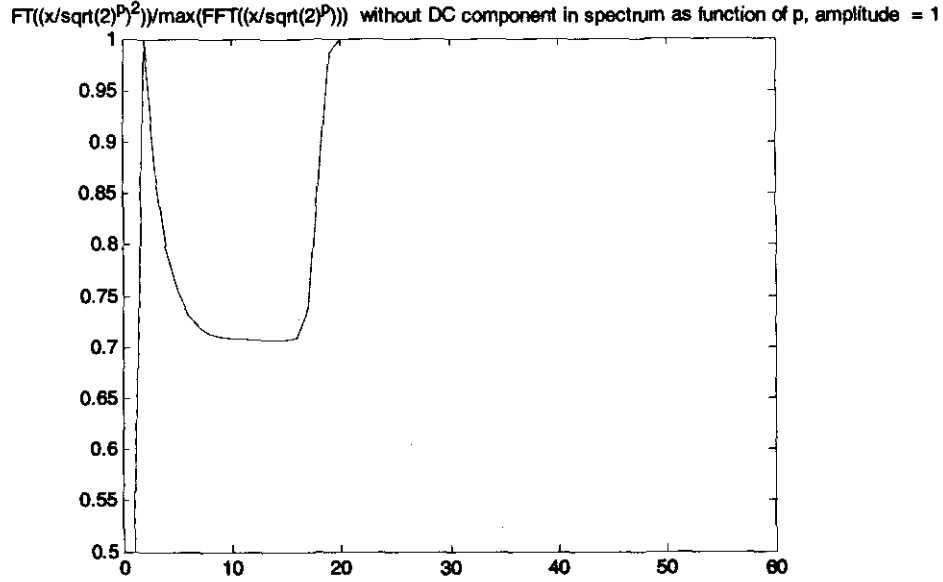


Figure 316: The ratio of the (FFT of the) squared signal relative to the (FFT of the) signal as function of the parameter p (the signal $y = K_{sig}(x/\sqrt{2})^p$, $K_{sig} = 1$)

From comparison of $K_{Sig} = 10$ and $K_{Sig} = 1$ it is evident that the ratio is proportional to the signal amplitude K_{Sig} . As the minimum value in the figure above is 0.7 it is concluded that the ratio of demodulated spectrum to regular spectrum is at most $(1/\sqrt{2})K_{Sig}$. For small signal values (say $K_{Sig} < 1$) the regular spectrum should perform better if the ratio of the noise in the demodulated spectrum to regular spectrum had been the same, whereas the reverse applies for larger signal values.

Define $R_{rs}(p, K_{sig})$ as the ratio of the signal in the demodulated spectrum to regular spectrum and $R_m(K_{noise})$ as the ratio of the noise in the demodulated spectrum to regular spectrum¹⁴⁸. Define $D_{sn}(p, K_{sig}, K_{noise})$ as the signal-to-noise ratio in the regular spectrum and $D_{sn2}(p, K_{sig}, K_{noise})$ as the signal-to-noise ratio in the demodulated spectrum¹⁴⁹. The

¹⁴⁸ The mean FFT value is used in this definition, although the maximum FFT value displays a similar trend.

¹⁴⁹ The mean FFT value is used in this definition, although the maximum FFT value displays a similar trend.

overall performance of the demodulated spectrum relative to the regular spectrum is given by $D_{sn2}(p, K_{sig}, K_{noise})/D_{sn}(p, K_{sig}, K_{noise}) = R_{rs}(p, K_{sig})/R_m(K_{noise})$. The equation for R_{rs} may be approximated by $R_{rs} = K_{sig}$ (compensating for the value of p may be done with the aid of the previous figure if necessary). Hence $D_{sn2}(p, K_{sig}, K_{noise})/D_{sn}(p, K_{sig}, K_{noise}) = R_{rs}(p, K_{sig})/R_m(K_{noise}) \approx K_{sig}/R_m(K_{noise})$. As plots of R_m have already been shown previously, the overall performance of the demodulated spectrum may be readily determined. If the additional simplification of $R_m = 1.4K_{noise}$ is made (from visual examination of the previous plot of R_m), the performance of the demodulated spectrum is $D_{sn2}(p, K_{sig}, K_{noise})/D_{sn}(p, K_{sig}, K_{noise}) \approx K_{sig}/(1.4K_{noise})$.

An additional question of interest is under what conditions the ratio $D_{sn}(p, K_{sig}, K_{noise}) > 1$ (previously defined as the signal-to-noise ratio in the regular spectrum). In figure 317 a normalized plot (in which the FFT values have been divided by the buffer length) of the (maximum value of the) regular spectrum is shown. Note that for large values of p the values become quite small (the effective signal level decreases as p increases). Thus the allowable noise level should be small for large values of p to enable detection of the signal.

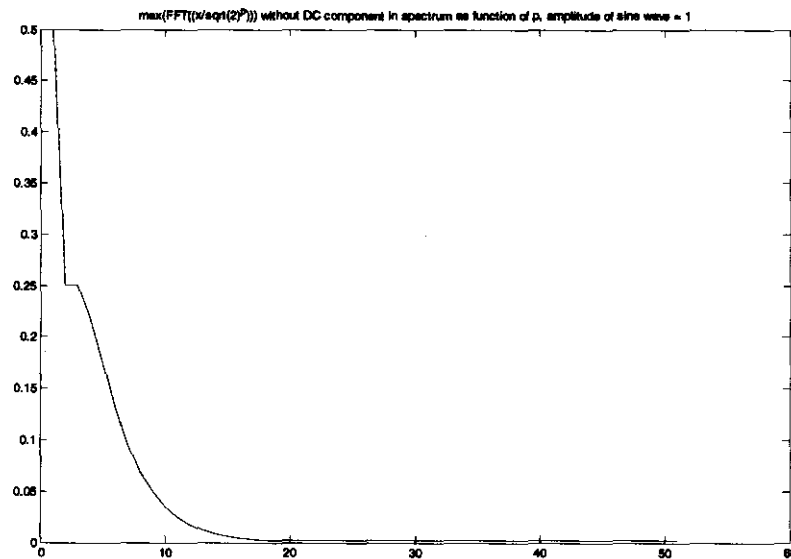


Figure 317: Maximum value of regular magnitude spectrum of model signal $((x/\sqrt{2})^p)$ as function of p

In the figure below a normalized plot (in which the FFT values have been divided by the buffer length) of the (maximum value of the) demodulated spectrum is shown, displaying a similar trend to the previous plot of the regular spectrum.

In the figure below the mean value of the regular spectrum of a noise source is shown as a function of the noise level (normalized by the square root of the buffer length¹⁵⁰).

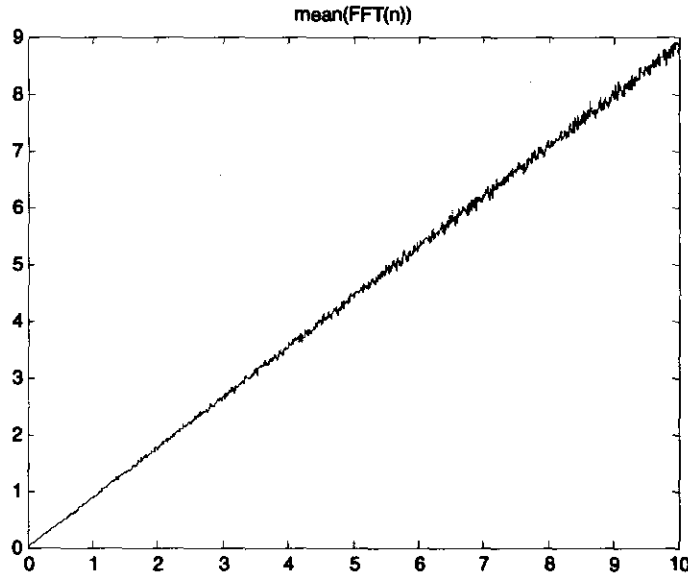


Figure 318: Mean value of spectrum versus noise level

Previous figures have already shown that the amplitude of the magnitude spectrum is proportional to the signal level K_{sig} . The amplitude is also related to the parameter p , controlling the spread in time of the model signal. For small values of p no correction factor is required. For larger values of p the effective signal level decreases (owing to the convolution in the frequency domain) as shown in previous figures and compensation has to be made. The decreased signal level as p increases effectively implies that the allowable noise level for detection of the signal decreases as well. Alternatively, the constant K_{sig} could be increased to offset the loss in signal level.

¹⁵⁰ If the square of the magnitude spectrum is used the normalization factor would be the buffer length.

If the effect of the parameter p is temporarily disregarded, the condition for $D_{sn}(p, K_{sig}, K_{noise}) > 1$ is $K_{sig}/K_{noise} > 1$ or $K_{sig} > K_{noise}$. The effect of the parameter p is to decrease the signal level effectively, which can also be seen as a reduction in the value of K_{sig} . Hence K_{sig} could also be increased (keeping the noise level constant) to combat the decreased signal level for large values of p . Figure 317 shows how much K_{sig} should be adjusted (the inverse of the values in figure 317 should be used for compensation). If the constant K_x is obtained from figure 317 for a specific value of the parameter p , the expression is $K_{sig}/K_{noise} > 0.5/K_x$ ¹⁵¹.

When the FFT is applied the effect of buffer length also has to be compensated. The actual magnitude spectrum value¹⁵² of a sinusoidal sine wave with amplitude K_{sig} after the FFT is $\frac{1}{2}MK_{sig}$, where M is the buffer length of the signal. The actual magnitude spectrum value of a noise source with standard deviation of $K_{noise} = \sigma$ after the FFT is $\sqrt{M}\sigma$, where M is the buffer length of the signal. Furthermore, the effect of the parameter p , which decreases the signal level as p increases, must be included in the constant K_x (refer to figure 317). The actual detection ratio with the FFT is therefore $D_{FFT} = 2K_x \frac{1}{2}MK_{sig} / ((\sqrt{M})K_{noise}) = K_x \sqrt{M}K_{sig}/K_{noise}$.

An additional question of interest is under what conditions the ratio $D_{sn2}(p, K_{sig}, K_{noise}) > 1$ (previously defined as the signal-to-noise ratio in the demodulated spectrum). In figure 319 a normalized plot is shown (in which the FFT values have been divided by the buffer length) of the (maximum value of the) demodulated spectrum. Note the similarity to figure 317 and that for large values of p the values become quite small (the effective signal level decreases as p increases). Hence the allowable noise level should be small for large values of p to enable detection of the signal.

¹⁵¹ The factor $0.5/K_x$ arises because the FFT of a pure sine wave splits into two components (positive and negative frequencies) with half the amplitude.

¹⁵² For example, the amplitude of the positive frequency component only.

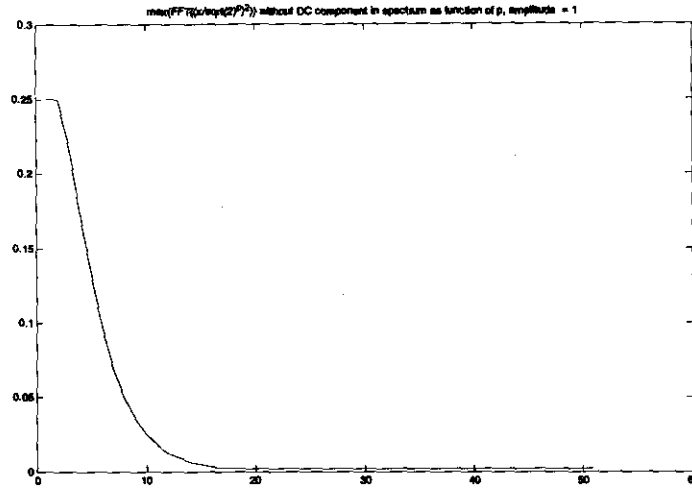


Figure 319: Maximum value of regular magnitude spectrum of model signal $((x/\sqrt{2})^p)$ as function of p – zoom plot

For small values of p (and very large – $p > 2^{20}$) no additional correction factor (owing to the use of the demodulated spectrum instead of the regular spectrum) is required for the demodulated spectrum (refer to figure 316). However, the demodulated noise spectrum differs substantially from the regular spectrum (as shown in figure 311 and figure 312). For small noise levels (standard deviation of noise smaller than about $1/\sqrt{2}$) the noise level of the demodulated spectrum is smaller than that of the regular spectrum with additive noise (figure 311 and figure 312). For large noise levels the opposite applies. Thus for small noise levels (such as with an increase in the parameter p , which decreases the signal level and hence the noise level for detection) the demodulated spectrum would perform better than the regular spectrum. For large noise levels the opposite applies.

To relate the demodulated spectrum to the regular spectrum, an approximation of $R_m \approx 1.4K_{noise}$ seems reasonable (figure 311 and figure 312). Hence the detection ratio of the demodulated spectrum is $D_{sn2}(p, K_{sig}, K_{noise}) \approx R_{rs}(p, K_{sig})/(1.4K_{noise}K_{noise})$. Assuming that the correction factor of the demodulated signal spectrum relative to the regular spectrum is disregarded (as discussed previously) the detection ratio is $D_{sn2}(p, K_{sig}, K_{noise}) \approx K_p(K_{sig})^2/(1.4(K_{noise})^2)$ (where K_p is the correction factor for the parameter p as discussed previously). For small values of p the correction factor K_p is small. Hence an

11.11.9.4 Speed variation function

To keep the theoretical analysis simple the speed variation (SV) function is chosen to be a simple step change in frequency. The signal can therefore be viewed as consisting of the superposition of two sinusoidal signals of different frequency. As mentioned previously, a small change in frequency (i.e. a small frequency difference of the sinusoidal signals) has a large impact on the higher harmonics (spread in frequency domain causes spectral overlap). Note that the amplitude of the signal in the frequency domain effectively decreases as the parameter p increases (each convolution increases the number of harmonic lines, but the amplitude of the pulse in the frequency domain decreases to keep the energy constant).

11.11.9.5 Results

11.11.9.5.1 Small spread of pulse in time domain, no speed variation, $\sigma_n = 2$, $\sigma_x = 0.36$, $p = 4$

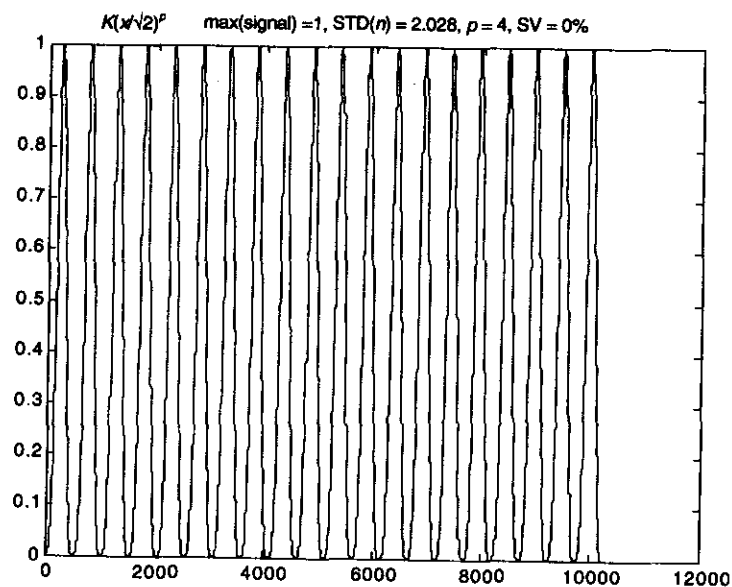


Figure 322: Vibration time domain signal $p = 4$

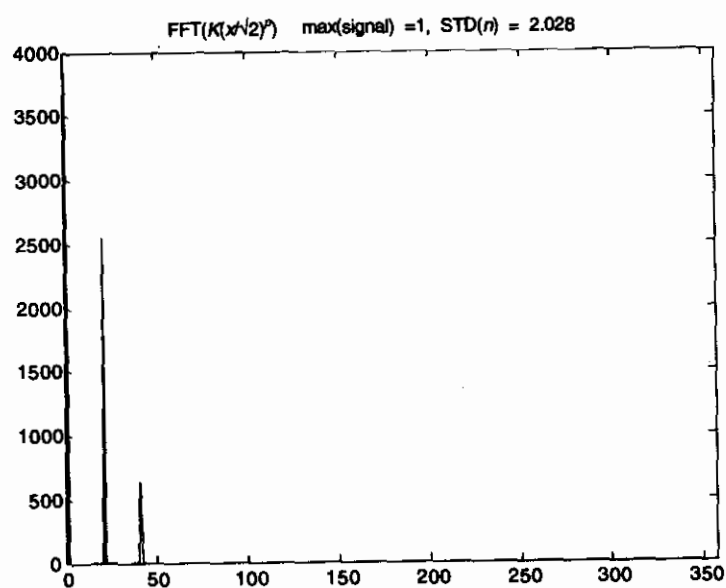


Figure 323: FFT of time domain signal $p = 4$

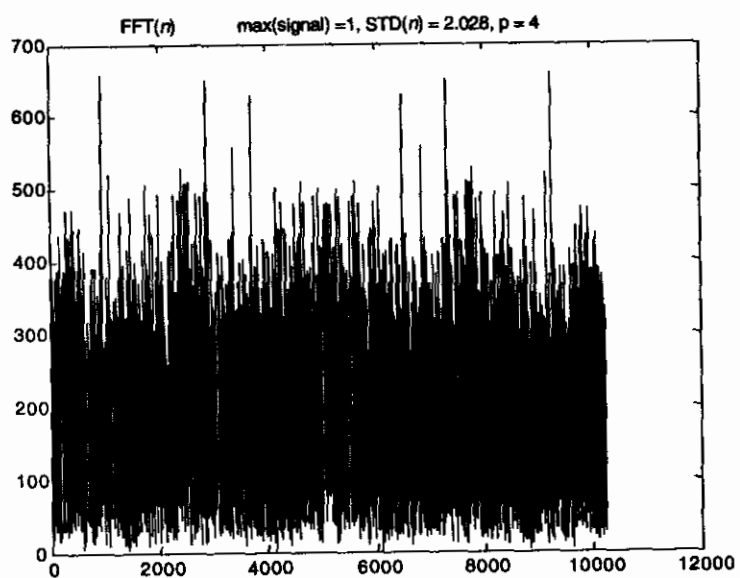


Figure 324: FFT of additive noise with $\text{STD}(n) = \sigma_n = 2, p = 4$

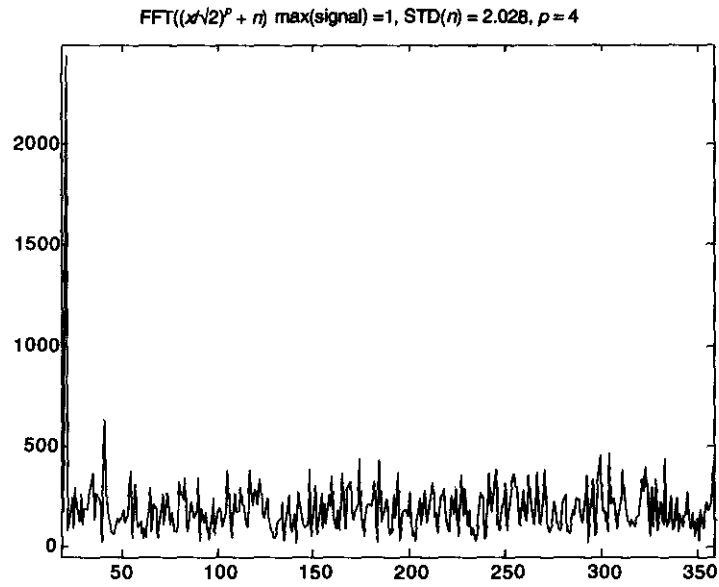


Figure 325: FFT of time domain signal plus additive noise with $\text{STD}(x) = \sigma_x = 0.36$, $\text{STD}(\eta) = \sigma_\eta = 2, p = 4$

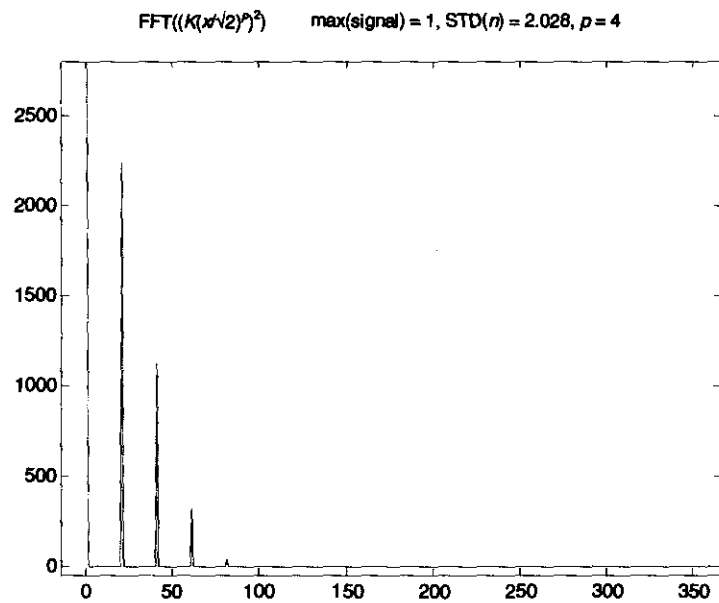


Figure 326: FFT of time domain signal squared using $\text{STD}(x) = \sigma_x = 0.36$, $\text{STD}(\eta) = \sigma_\eta = 2, p = 4$

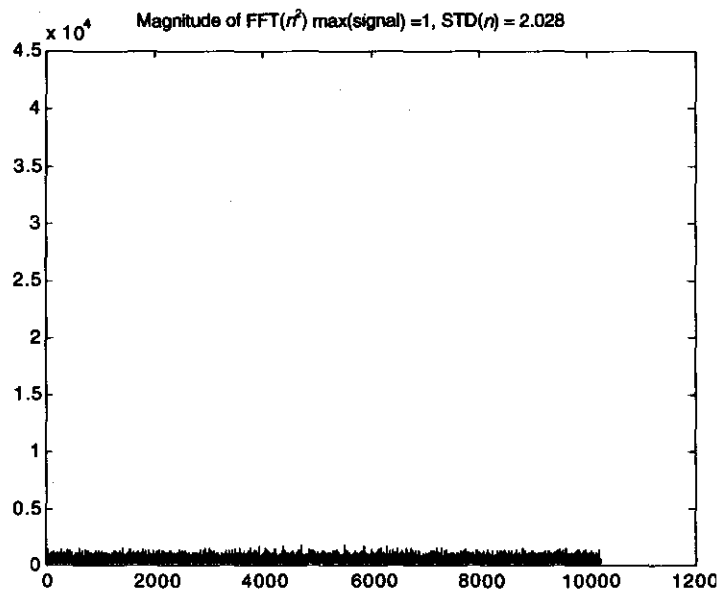


Figure 327: FFT of square of noise signal with $\text{STD}(n) = \sigma_n = 2$, $p = 4$

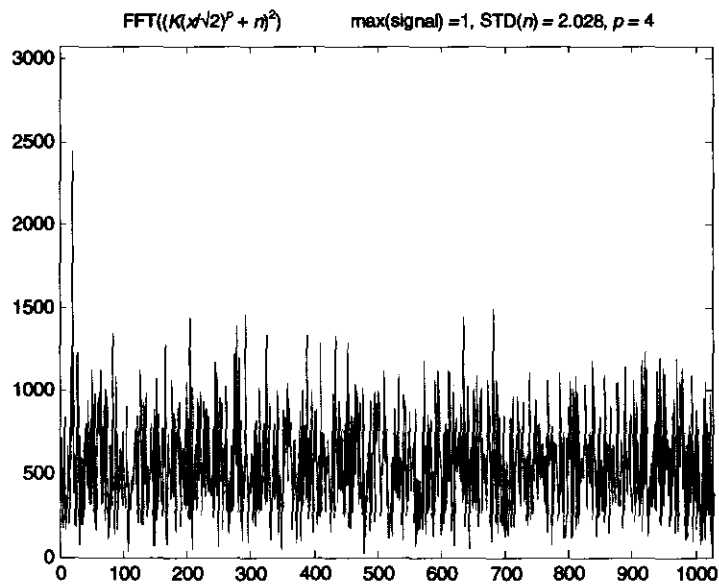


Figure 328: FFT of time domain signal plus additive noise squared using $\text{STD}(x) = \sigma_x = 0.36$, $\text{STD}(n) = \sigma_n = 2$, $p = 4$

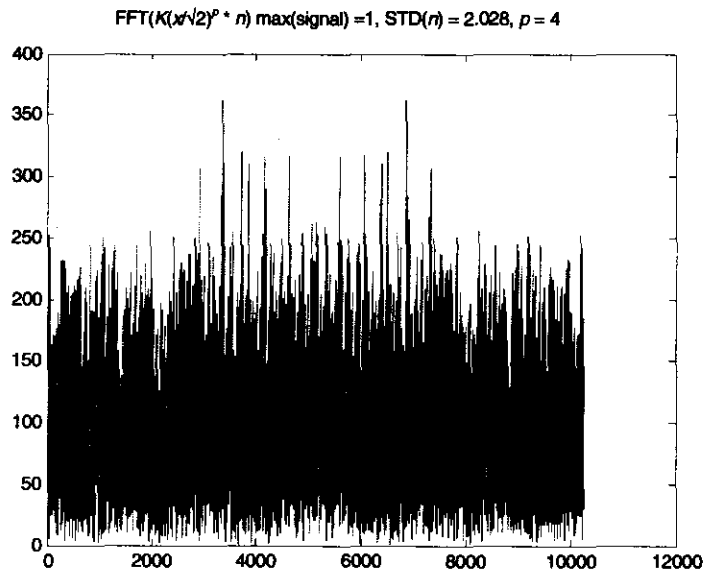


Figure 329: FFT of product of time domain signal and noise signal (multiplicative noise)

11.11.9.5.2 *Small spread of pulse in time domain, no speed variation, $\sigma_n = 0.1$,*

$\sigma_x = 0.18$, $p = 256$

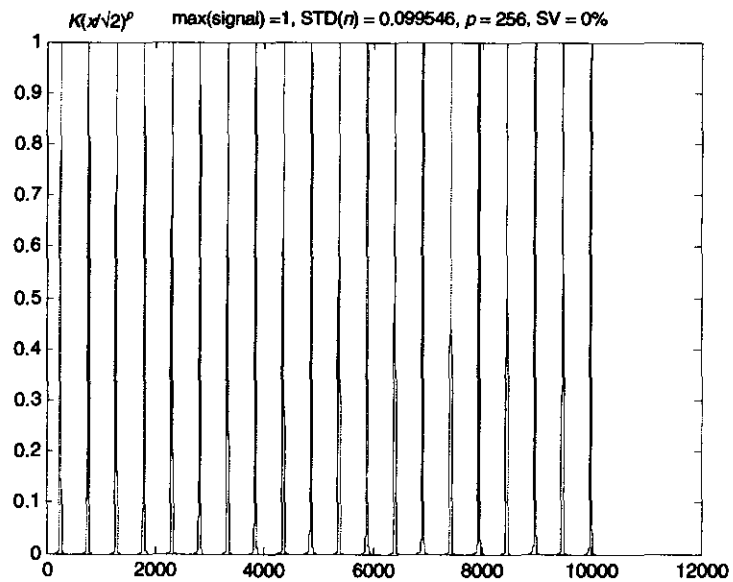


Figure 330: Vibration time domain signal $p = 256$

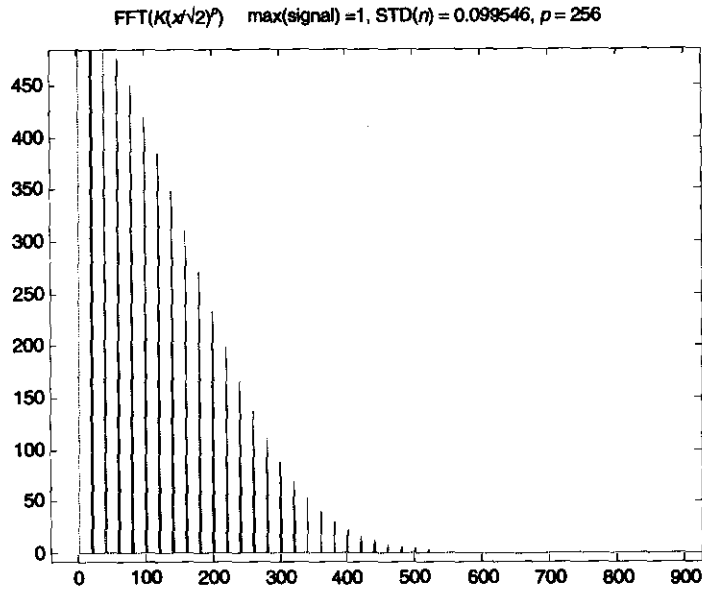


Figure 331: FFT of time domain signal $p = 256$

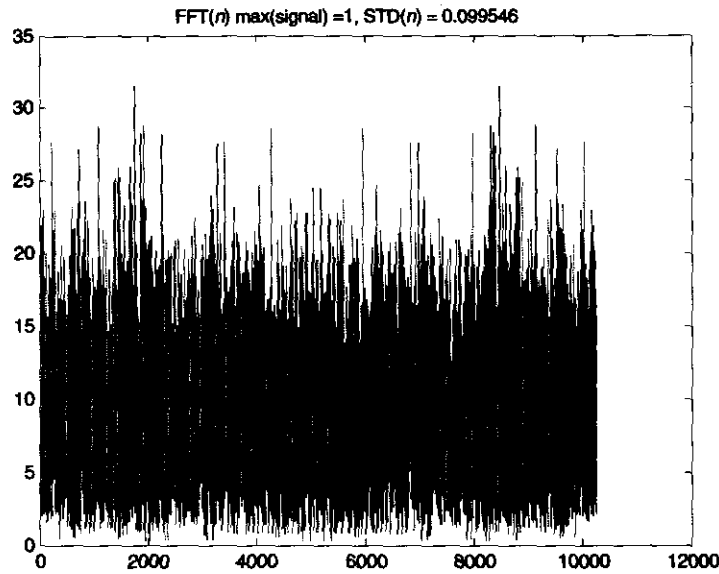


Figure 332: FFT of additive noise with $\text{STD}(n) = \sigma_n = 0.1, p = 256$

When the FFT is applied the effect of buffer length has to be compensated. For $p = 2^8 = 256$ the value of $0.5/K_x = 10.1147$. The predicted detection ratio with the FFT is $D_{FFT} = 2K_x/2MK_{sig}/((\sqrt{M})K_{noise}) = K_x\sqrt{MK_{sig}}/K_{noise} = 0.0494(\sqrt{10240})/0.1 = 50.023$.

The actual mean noise value from the previous figure is about ten (maximum value is about 30). The maximum signal value (excluding the DC component) in the figure below

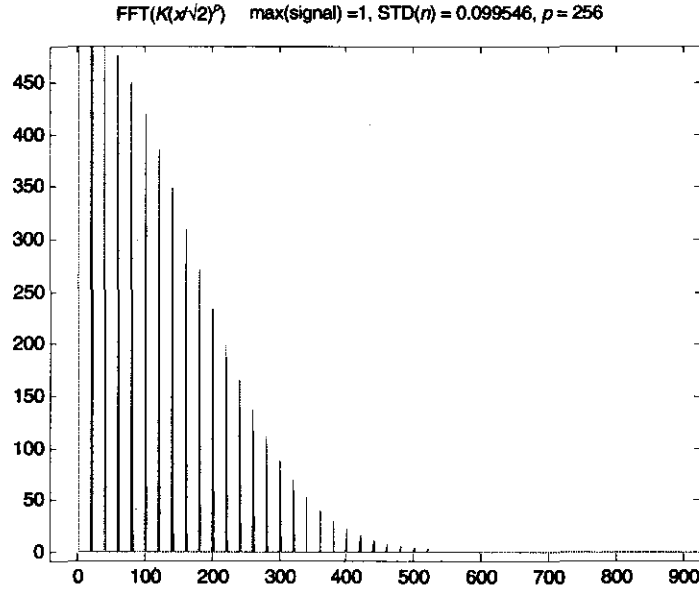


Figure 331: FFT of time domain signal $p = 256$

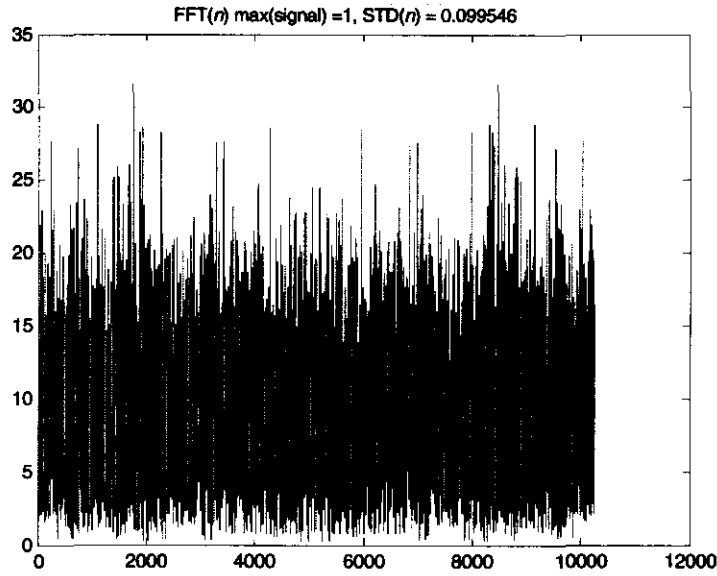


Figure 332: FFT of additive noise with $\text{STD}(n) = \sigma_n = 0.1, p = 256$

When the FFT is applied the effect of buffer length has to be compensated. For $p = 2^8 = 256$ the value of $0.5/K_x = 10.1147$. The predicted detection ratio with the FFT is $D_{FFT} = 2K_x^{1/2}MK_{sig}/((\sqrt{M})K_{noise}) = K_x\sqrt{MK_{sig}/K_{noise}} = 0.0494(\sqrt{10240})/0.1 = 50.023$.

The actual mean noise value from the previous figure is about ten (maximum value is about 30). The maximum signal value (excluding the DC component) in the figure below

is 512 when the FFT is applied. The actual (optimistic) detection ratio is therefore $512/10 = 51.2$. This compares favourably with the theoretical prediction $D_{FFT} = 50.023$ (it can be seen that the signal peak can readily be detected in the presence of the noise source).

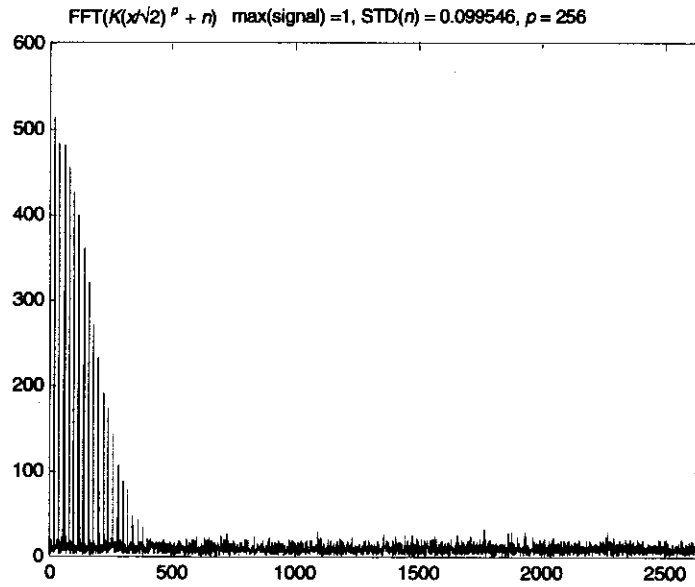


Figure 333: FFT of time domain signal plus additive noise with $\text{STD}(x) = \sigma_x = 0.18$, $\text{STD}(n) = \sigma_n = 0.1$, $p = 256$

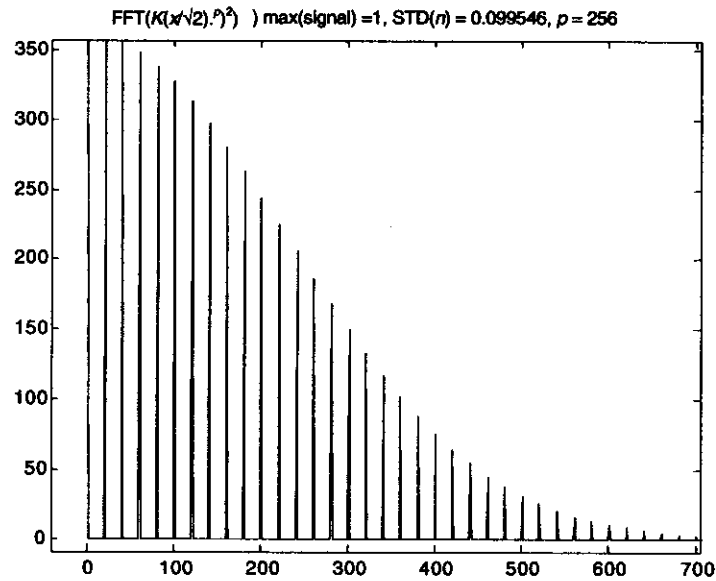


Figure 334: FFT of time domain signal squared with $\text{STD}(x) = \sigma_x = 0.18$, $\text{STD}(n) = \sigma_n = 0.1$, $p = 256$

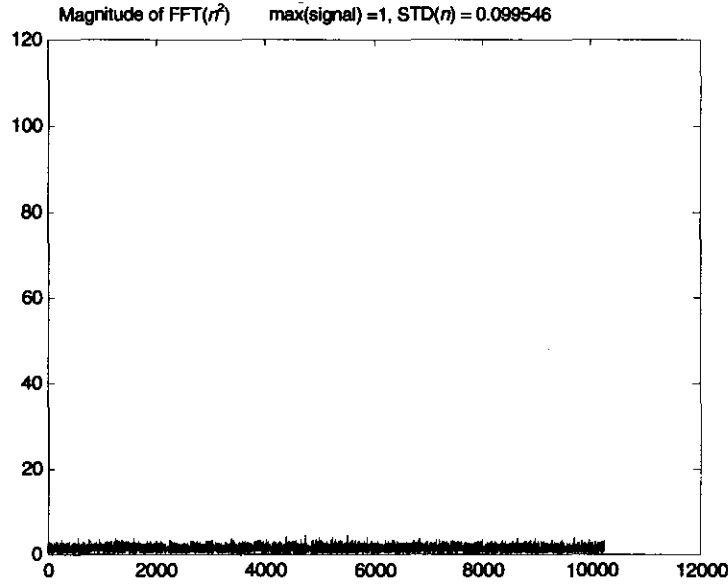


Figure 335: FFT of square of noise signal with $\text{STD}(n) = \sigma_n = 0.1, p = 256$

When the FFT is applied the effect of buffer length has to be compensated. For $p = 2^8 = 256$ the value of $0.5/K_x = 10.1147$. The predicted detection ratio of the demodulated spectrum with the FFT is $D_{FFT2} = 2K_p K_x^{1/2} M K_{sig}^2 / (1.4 K_{noise} (\sqrt{M}) K_{noise}) = 0.71 K_p K_x \sqrt{M} (K_{sig}/K_{noise})^2 = 0.5(0.0494)(\sqrt{10240})(10)^2 = 248.5$. Hence $D_{FFT2}/D_{FFT} = 248.5/50.023 = 4.96$.

The actual mean noise value from the previous figure is about 2 (maximum value is about 5). The maximum signal value (excluding the DC component) in the figure below is 355 when the FFT is applied. The actual (optimistic) detection ratio is therefore about $355/2 = 177.5$. This compares favourably with the theoretical prediction $D_{FFT2} = 248.5$ (it can be seen that the signal peak can readily be detected in the presence of the noise source). As $K_{sig}/K_{noise} = 10 > \sqrt{2K_p}^{-1}$ the demodulated spectrum provides a larger detection ratio than the regular spectrum.

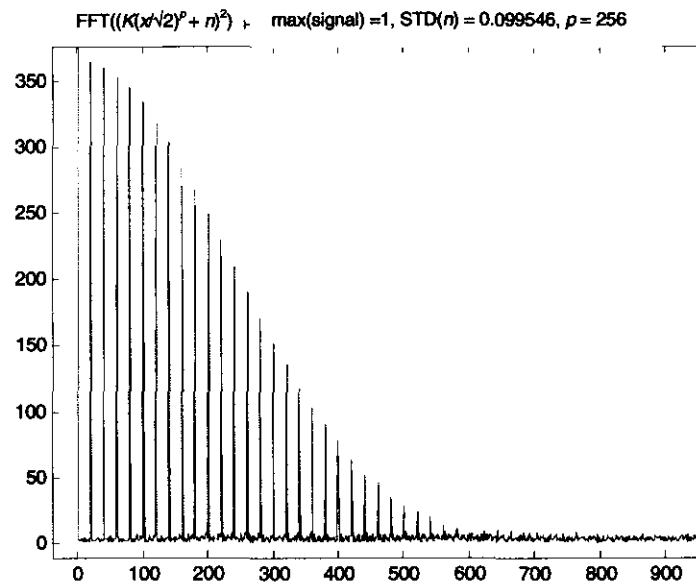


**Figure 336: FFT of time domain signal plus additive noise squared
with $\text{STD}(x) = \sigma_x = 0.18$, $\text{STD}(n) = \sigma_n = 0.1$, $p = 256$**

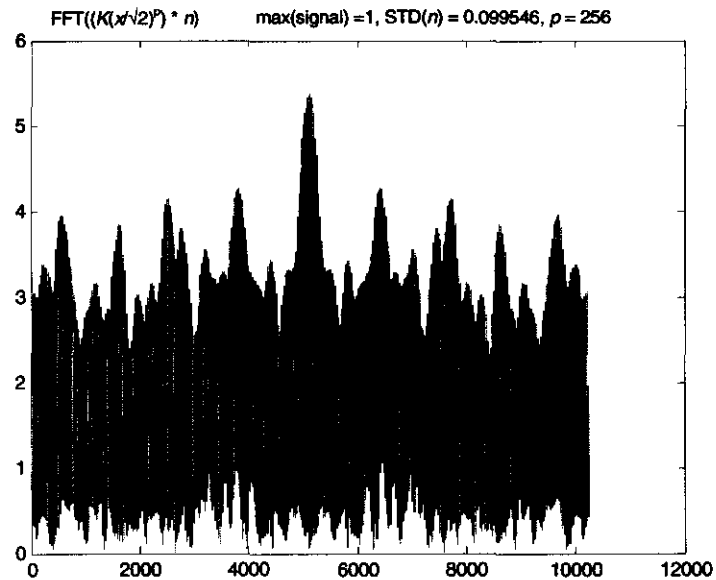


Figure 337: FFT of product of time domain signal and noise signal (multiplicative noise)

11.11.9.5.3 *Small spread of pulse in time domain, no speed variation, $\sigma_n = 0.1$,*

$\sigma_x = 0.0477$, $p = 65\,536$

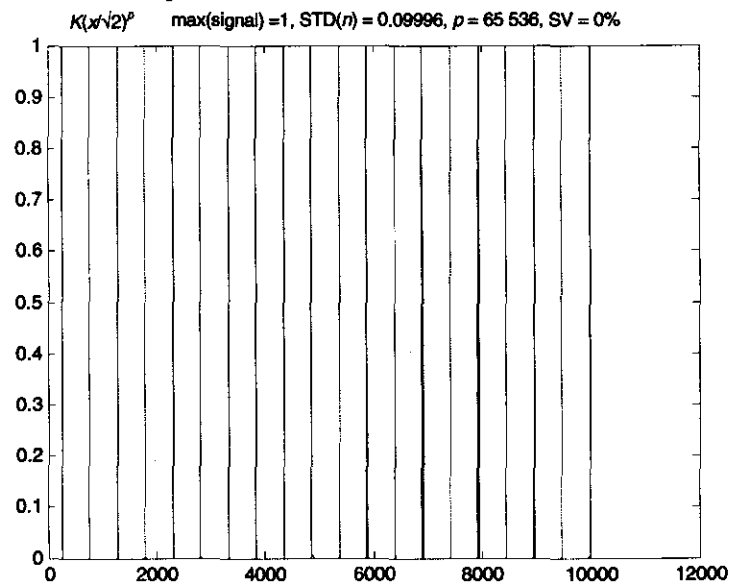


Figure 338: Vibration time domain signal $p = 65\,536$

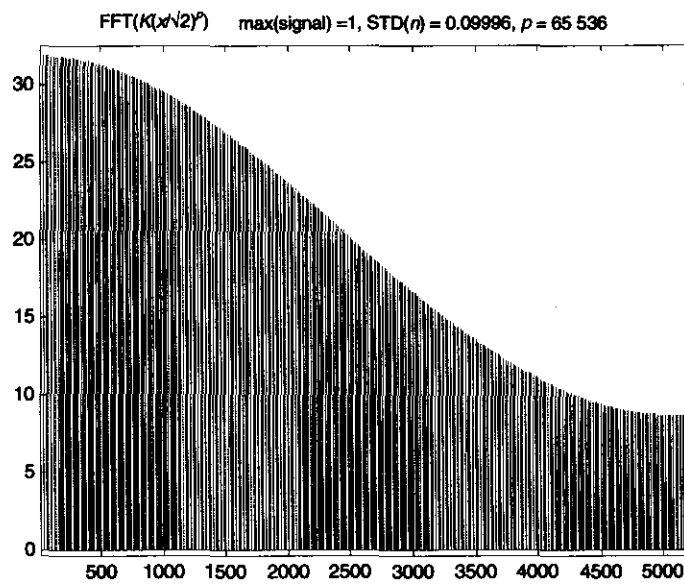


Figure 339: FFT of time domain signal $p = 65\,536$

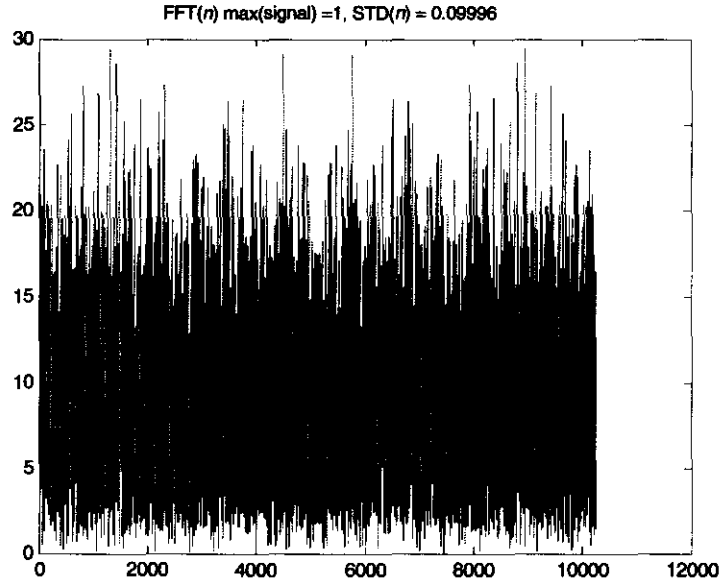


Figure 340: FFT of additive noise with $\text{STD}(n) = \sigma_n = 0.1, p = 65\,536$

When the FFT is applied the effect of buffer length has to be compensated. For $p = 2^{16} = 65\,536$ the value of $0.5/K_x = 160.3$. The predicted detection ratio with the FFT is:

$$D_{FFT} = 2K_x^{1/2}MK_{sig}/((\sqrt{M})K_{noise}) = K_x\sqrt{MK_{sig}/K_{noise}} = 0.0031(\sqrt{10240})1/0.1 = 3.16.$$

The actual mean noise value from the previous figure is about ten (maximum value is about 30). The maximum signal value (excluding the DC component) in the figure below is 30 when the FFT is applied. The actual (optimistic) detection ratio is therefore about $30/10 = 3$. This compares favourably with the theoretical prediction $D_{FFT} = 3.16$ (it can be seen that the signal peak can still be detected in the presence of the noise source).

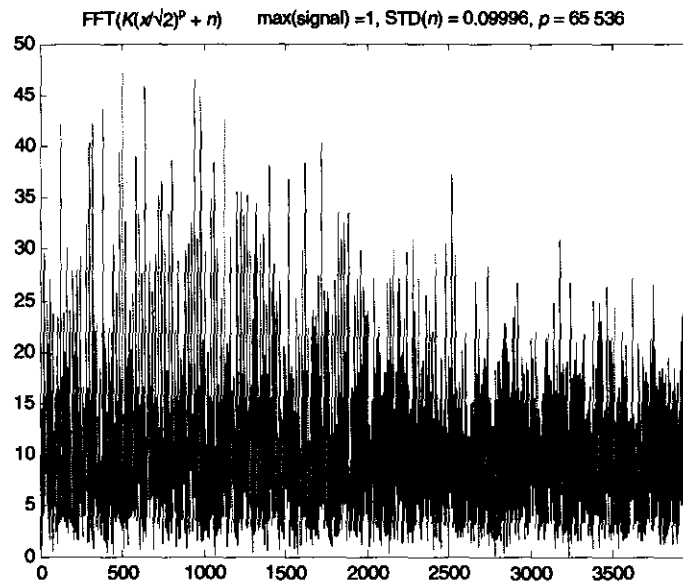


Figure 341: FFT of time domain signal plus additive noise with $\text{STD}(x) = \sigma_x = 0.0477$, $\text{STD}(n) = \sigma_n = 0.1$, $p = 65\ 536$

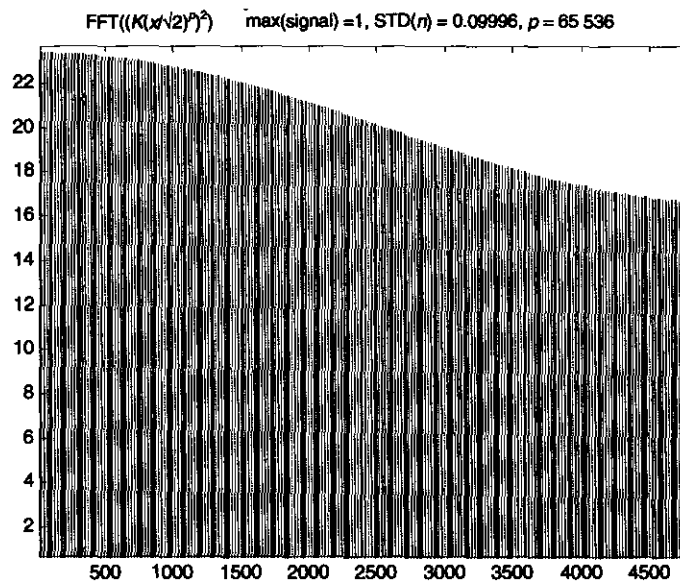


Figure 342: FFT of time domain signal squared with $\text{STD}(x) = \sigma_x = 0.0477$, $\text{STD}(n) = \sigma_n = 0.1$, $p = 65\ 536$

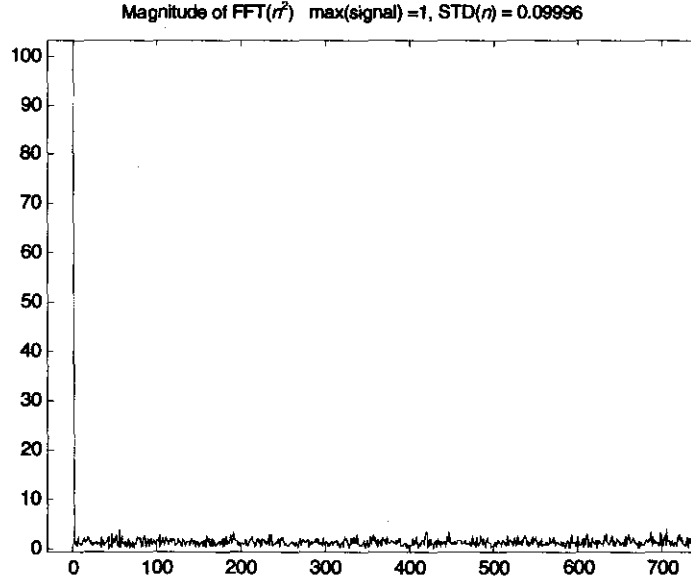


Figure 343: FFT of square of noise signal with $\text{STD}(n) = \sigma_n = 0.1, p = 65\,536$

When the FFT is applied the effect of buffer length has to be compensated. For $p = 2^{16} = 65\,536$ the value of $0.5/K_x = 160.3$. The predicted detection ratio of the demodulated spectrum when applying the FFT is $D_{FFT2} = 2K_p K_x^{1/2} M K_{sig}^2 / (1.4 K_{noise} (\sqrt{M}) K_{noise}) = 0.71 K_p K_x \sqrt{M} (K_{sig}/K_{noise})^2 = 0.5(0.0031)(\sqrt{10240})(10)^2 = 16$. Thus $D_{FFT2}/D_{FFT} = 16/1.69 = 9.47$ (which is about twice the value as for $p = 2^8$, indicating that the demodulated spectrum is helpful as the parameter p increases).

The actual mean noise value from the previous figure is about 2 (maximum value is about 5). The maximum signal value (excluding the DC component) in the figure below is 25 when the FFT is applied. The actual (optimistic) detection ratio is therefore about $25/2 = 12.5$. This compares favourably with the theoretical prediction $D_{FFT2} = 16$ (it can be seen that the signal peak can readily be detected in the presence of the noise source). As $K_{sig}/K_{noise} = 10 > \sqrt{2K_p}^{-1}$ the demodulated spectrum provides a larger detection ratio than the regular spectrum.

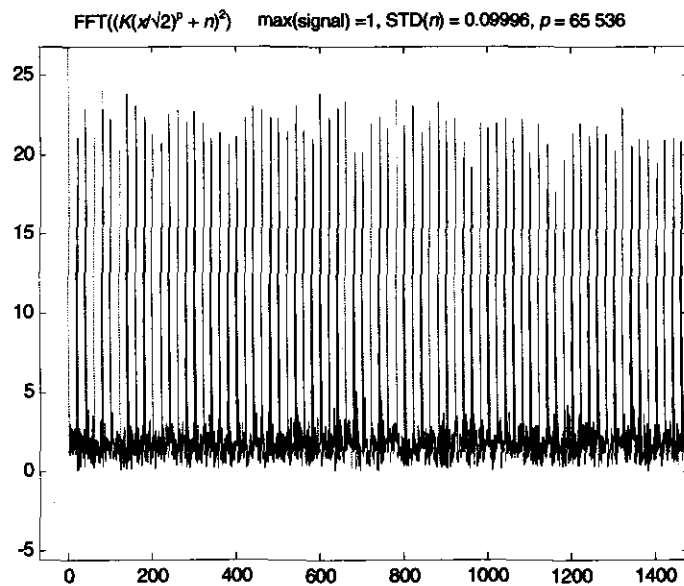


**Figure 344: FFT of time domain signal plus additive noise squared
with $\text{STD}(x) = \sigma_x = 0.0477$, $\text{STD}(n) = \sigma_n = 0.1$, $p = 65\ 536$**

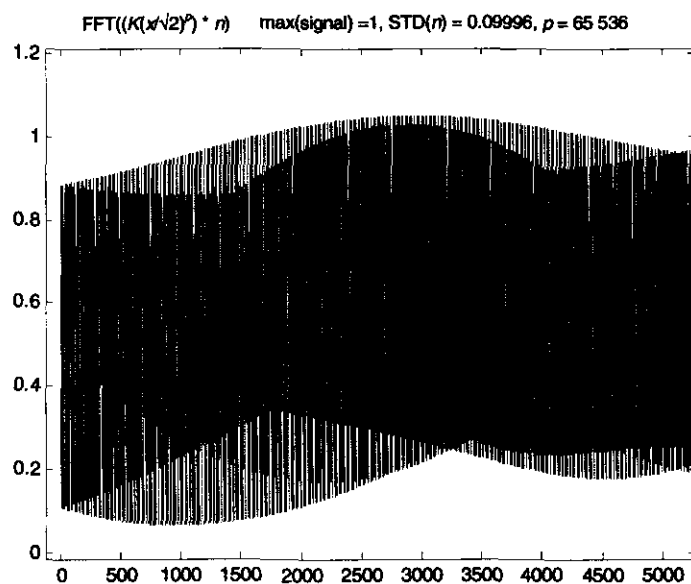


Figure 345: FFT of product of time domain signal and noise signal (multiplicative noise)

11.11.9.5.4 *Small spread of pulse in time domain, no speed variation, $\sigma_n = 20$,*

$\sigma_x = 3.6455$, $p = 4$, amplitude of sine wave = 10

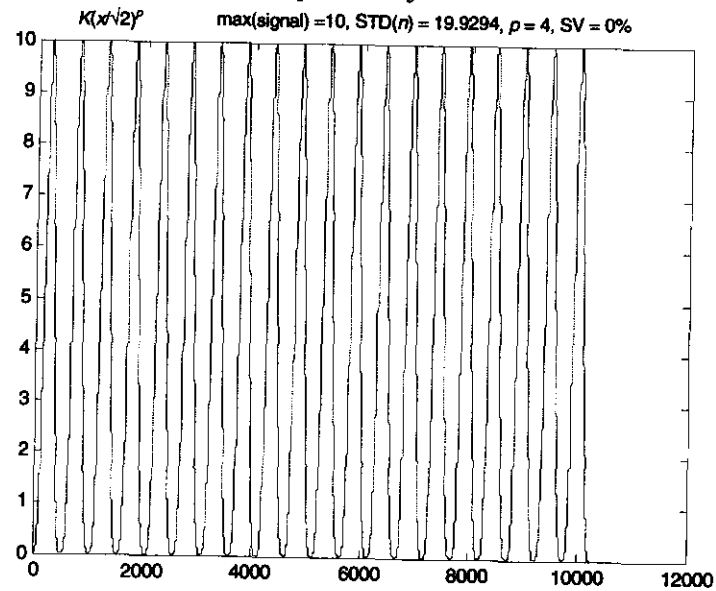


Figure 346: Vibration time domain signal $p = 4$

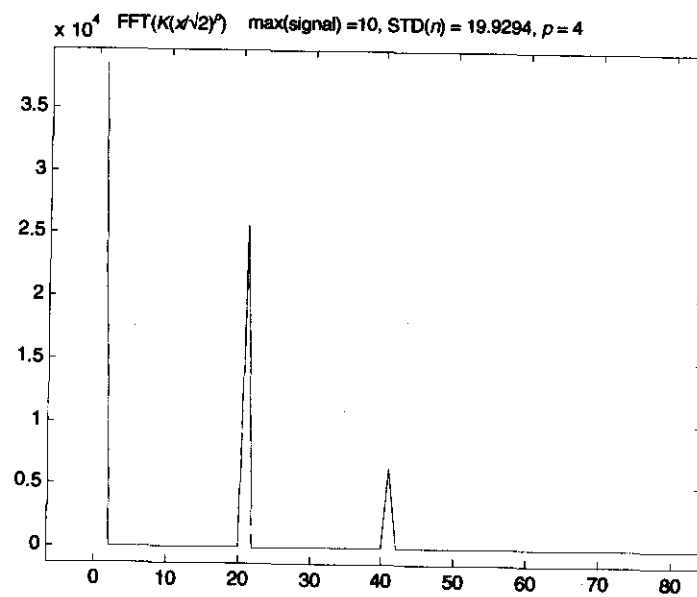


Figure 347: FFT of time domain signal $p = 4$

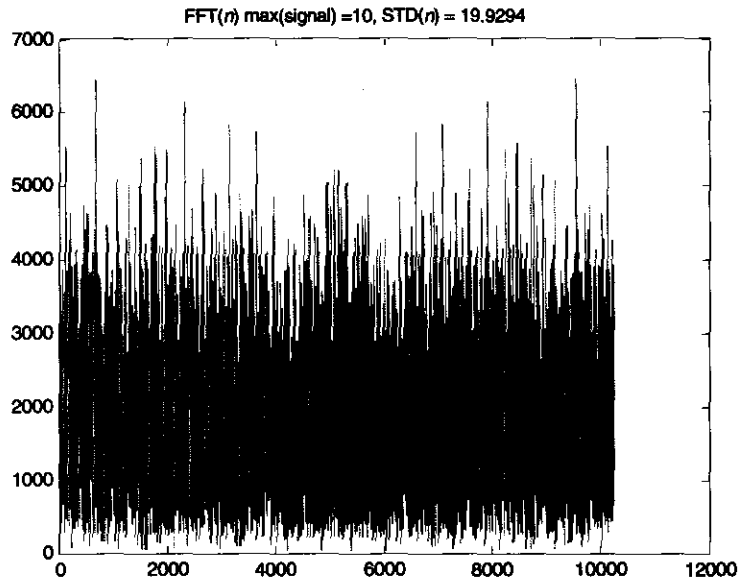


Figure 348: FFT of additive noise with $\text{STD}(n) = \sigma_n = 20, p = 4$

When the FFT is applied the effect of buffer length has to be compensated. For $p = 2^4 = 16$ the value of $0.5/K_x = 2.864$. The predicted detection ratio with the FFT is $D_{FFT} = 2K_x^{1/2}MK_{sig}/((\sqrt{M})K_{noise}) = K_x\sqrt{MK_{sig}}/K_{noise} = 0.1746(\sqrt{10240})10/20 = 8.83$.

The actual mean noise value from the previous figure is about 2000 (maximum value is about 6000). The maximum signal value (excluding the DC component) in figure 349 is 37 000 when the FFT is applied. The actual (optimistic) detection ratio is therefore $37\,000/2000 = 18.5$ This compares reasonably with the theoretical prediction $D_{FFT} = 8.83$ (it is evident that the signal peak can still be detected in the presence of the noise source).

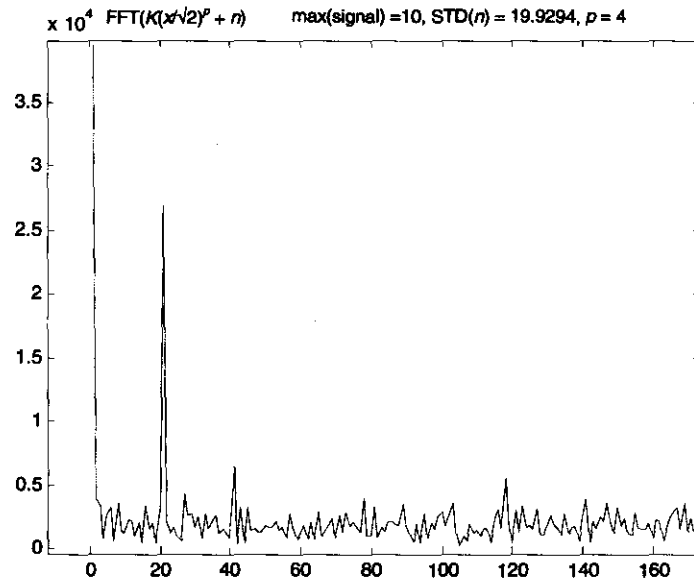


Figure 349: FFT of time domain signal plus additive noise with $\text{STD}(x) = \sigma_x = 3.6455$, $\text{STD}(n) = \sigma_n = 20$, $p = 4$

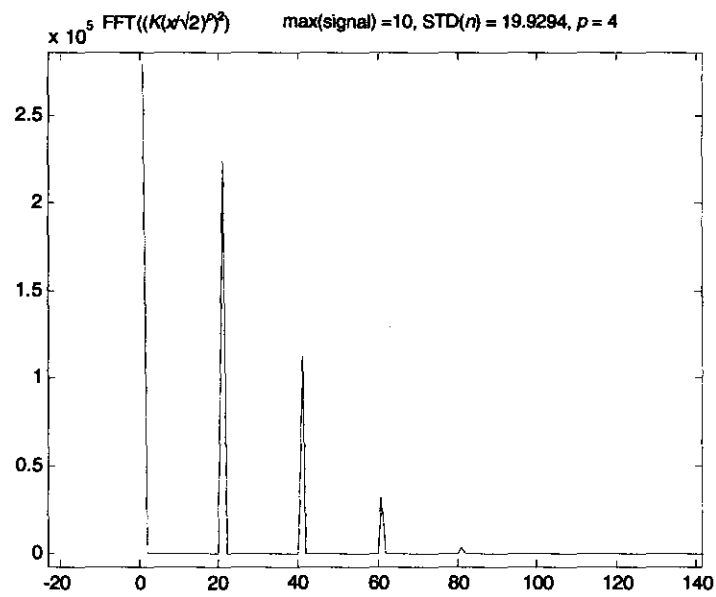


Figure 350: FFT of time domain signal squared with $\text{STD}(x) = \sigma_x = 3.6455$, $\text{STD}(n) = \sigma_n = 20$, $p = 4$

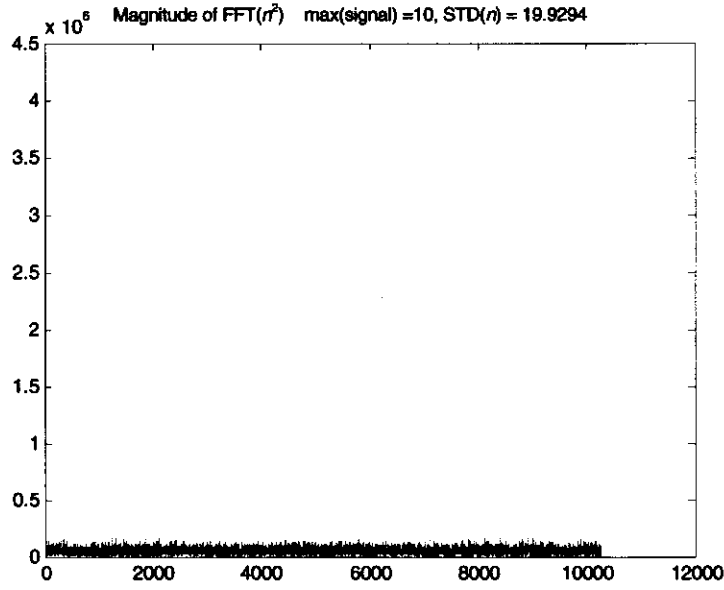


Figure 351: FFT of square of noise signal with $\text{STD}(n) = \sigma_n = 20, p = 4$

When the FFT is applied the effect of buffer length has to be compensated. For $p = 2^4 = 16$ the value of $0.5/K_x = 2.864$. The predicted detection ratio of the demodulated spectrum with the FFT is $D_{FFT2} = 2K_p K_x^{1/2} M K_{sig}^2 / (1.4 K_{noise} (\sqrt{M}) K_{noise}) = 0.71 K_p K_x \sqrt{M} (K_{sig}/K_{noise})^2 = 0.5(0.1746)(\sqrt{10240}) (10/20)^2 = 2.21$. Hence $D_{FFT2}/D_{FFT} = 2.21/8.83 = 0.25$.

The actual mean noise value from the previous figure is about 200 000. The maximum signal value (excluding the DC component) in the figure below is 150 000 when the FFT is applied. The actual (optimistic) detection ratio is therefore about 0.75. This compares reasonably with the theoretical prediction $D_{FFT2} = 0.25$ (it can be seen that the signal peak can hardly be detected in the presence of the noise source). As $K_{sig}/K_{noise} = 0.5 < \sqrt{2} K_p^{-1}$ the demodulated spectrum provides a smaller detection ratio than the regular spectrum.

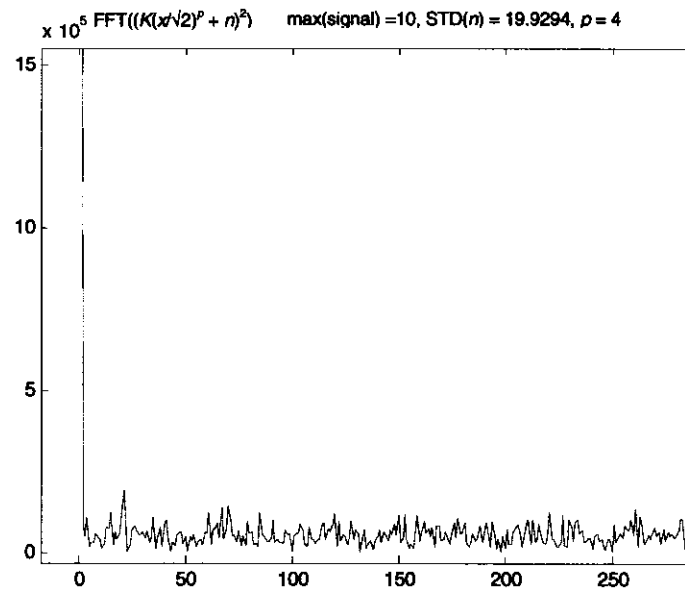


Figure 352: FFT of time domain signal plus additive noise squared with $\text{STD}(x) = \sigma_x = 3.6455$, $\text{STD}(n) = \sigma_n = 20$, $p = 65\ 536$

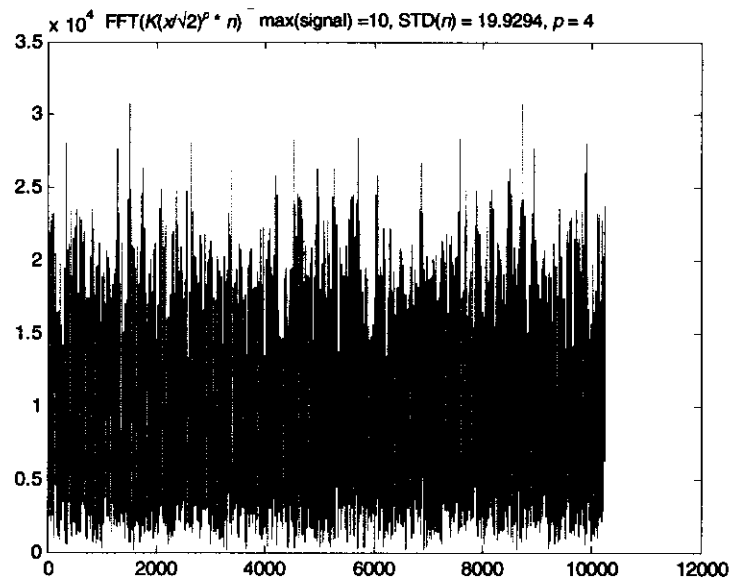


Figure 353: FFT of product of time domain signal and noise signal (multiplicative)

11.11.9.5.5 Multiplicative noise ($\sigma_n = 1$)

In figure 354 a zoom plot of the FFT of the signal with multiplicative noise is shown.

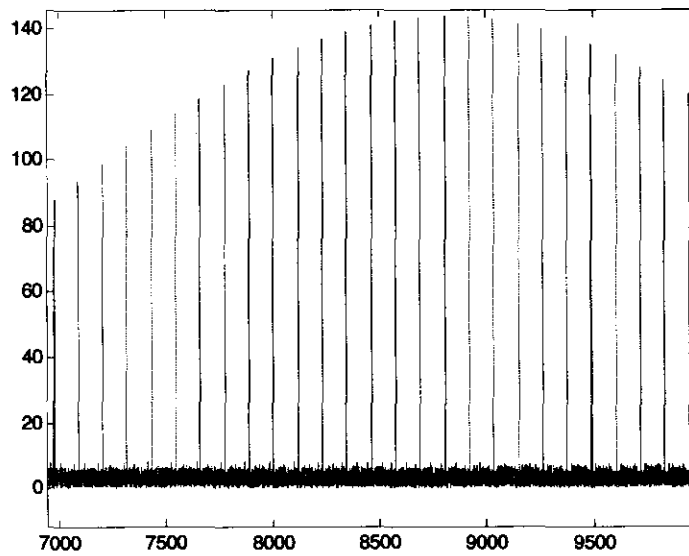


Figure 354: FFT magnitude spectrum of model signal with multiplicative noise ($\sigma_n = 1$)

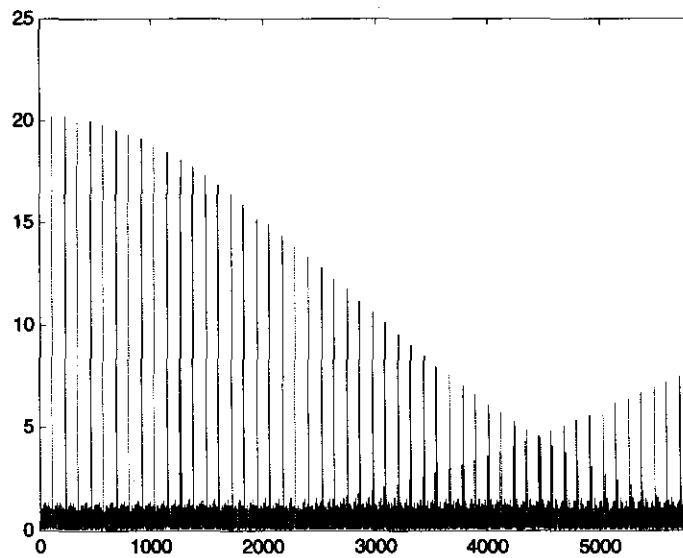


Figure 355: FFT magnitude spectrum of model signal squared with multiplicative noise ($\sigma_n = 1$)

In the figure above the demodulated spectrum is shown.

11.11.9.5.6 Multiplicative noise ($\sigma_n = 5$)

In the figure below a zoom plot of the FFT of the signal plus multiplicative noise is shown.

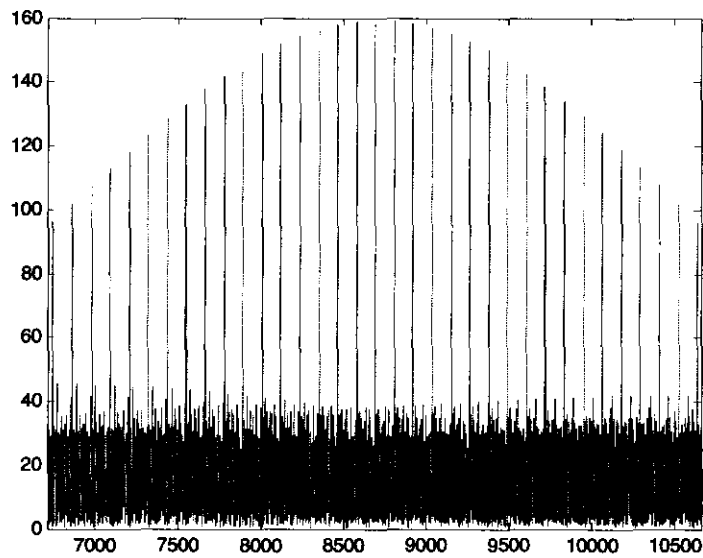


Figure 356: FFT magnitude spectrum of model signal with multiplicative noise ($\sigma_n = 5$)

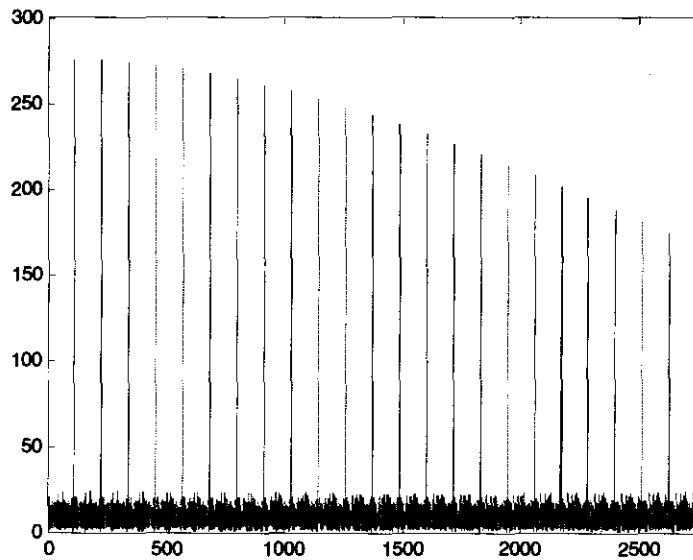


Figure 357: FFT magnitude spectrum of model signal squared with multiplicative noise ($\sigma_n = 5$)

In figure 357 the demodulated spectrum is shown. The noise level with multiplicative noise is much lower than that for the regular spectrum. This is mainly because of the difference between the spectrum of the noise signal and the square of the noise signal. In the former case the spectrum is essentially flat (refer to previous figures), whereas in the latter case the spectrum is also flat but with an impulse at the origin. As multiplication corresponds to convolution in the frequency domain, it can be seen that the demodulated spectrum would perform better with multiplicative noise than the regular spectrum. For additive noise the opposite applies.

11.11.9.5.7 Small spread of pulse in time domain, large speed variation

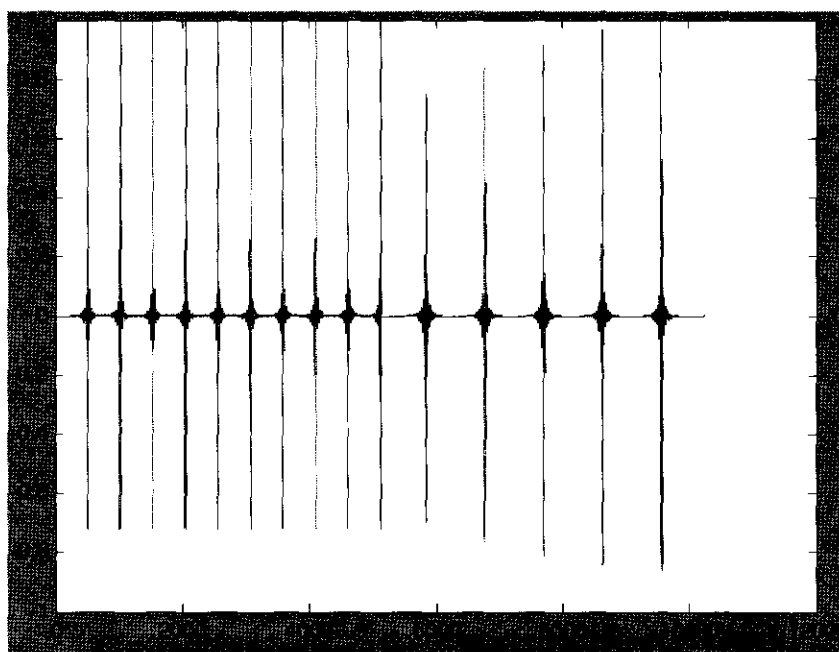


Figure 358: Time domain signal, $p = 65\ 536$, $SV = 45\%$, $\sigma_n = 0.01$

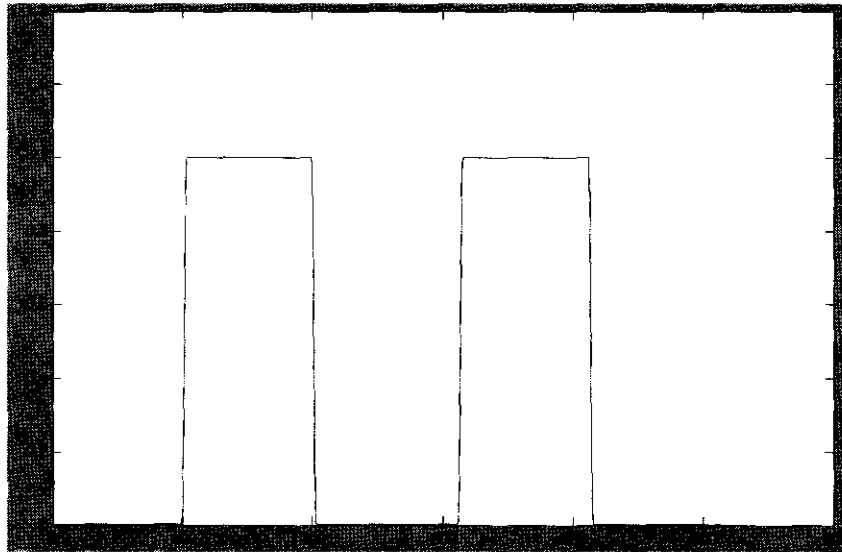


Figure 359: Bandpass filter from 10–20kHz

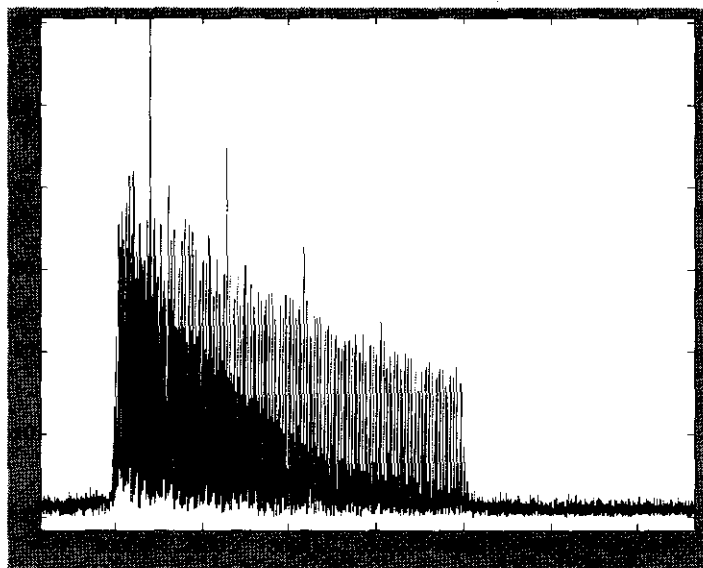


**Figure 360: FFT magnitude spectrum of model signal plus noise ($SV = 45\%$,
 $p = 65\,536$, $\sigma_n = 0.01$)**

In figure 361 a zoom plot of the FFT of the signal plus noise is shown. Note that the spectrum is similar in shape to that of the demodulated spectrum (shown in figure 362).

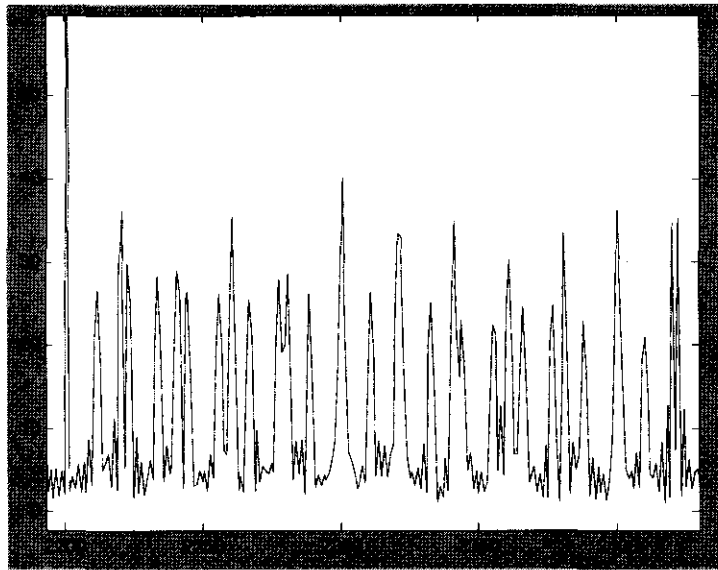


Figure 361: FFT magnitude spectrum of model signal plus noise – zoom plot
($SV = 45\%$, $p = 65\,536$, $\sigma_n = 0.01$)

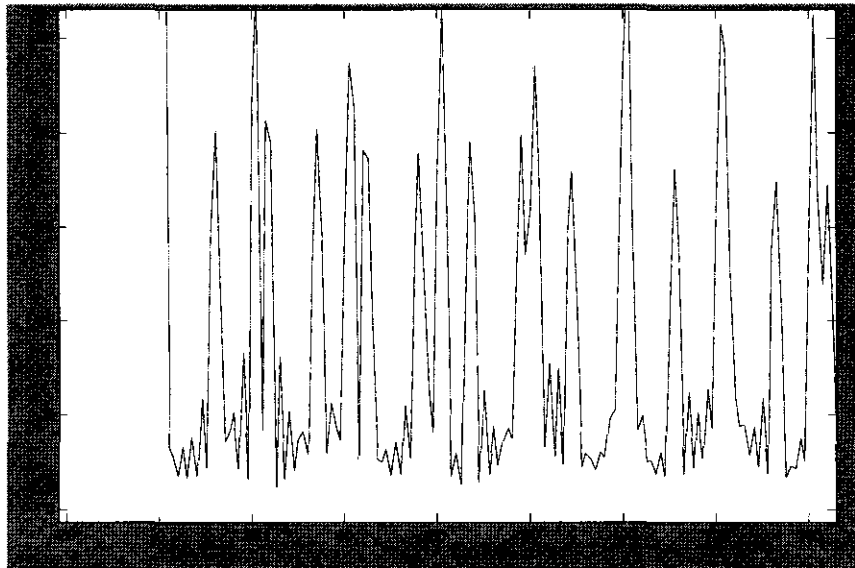


Figure 362: FFT magnitude spectrum of model signal plus noise squared
($SV = 45\%$, $p = 65\,536$, $\sigma_n = 0.01$)

In the figure above the demodulated spectrum is shown. Note the similarity of the demodulated spectrum compared with the regular spectrum.

11.11.9.5.8 *Large spread of pulse in time domain, large speed variation*

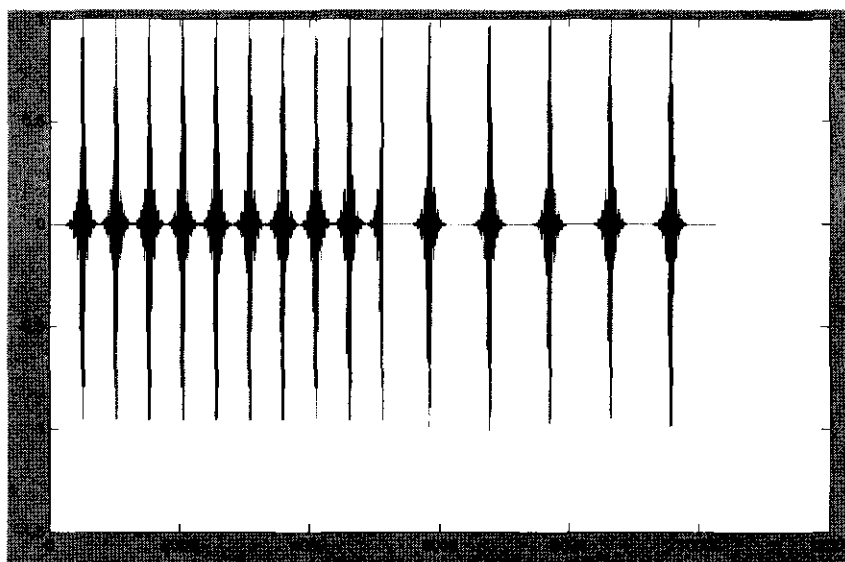


Figure 363: Time domain signal, $p = 65\,536$, $SV = 45\%$, $\sigma_n = 0.01$

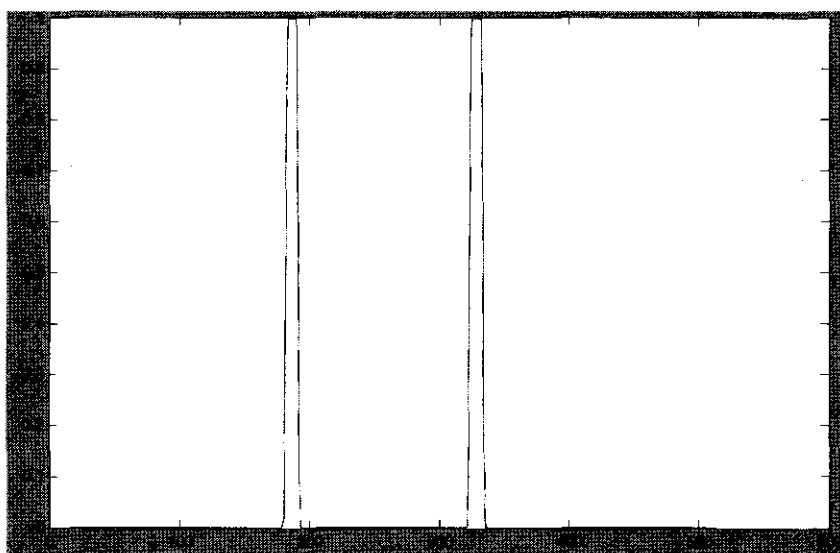


Figure 364: Bandpass filter from 10–20kHz

In figure 365 a zoom plot of the FFT of the signal plus noise is shown. Note that the spectrum is similar in shape to that of the demodulated spectrum (shown in figure 366).

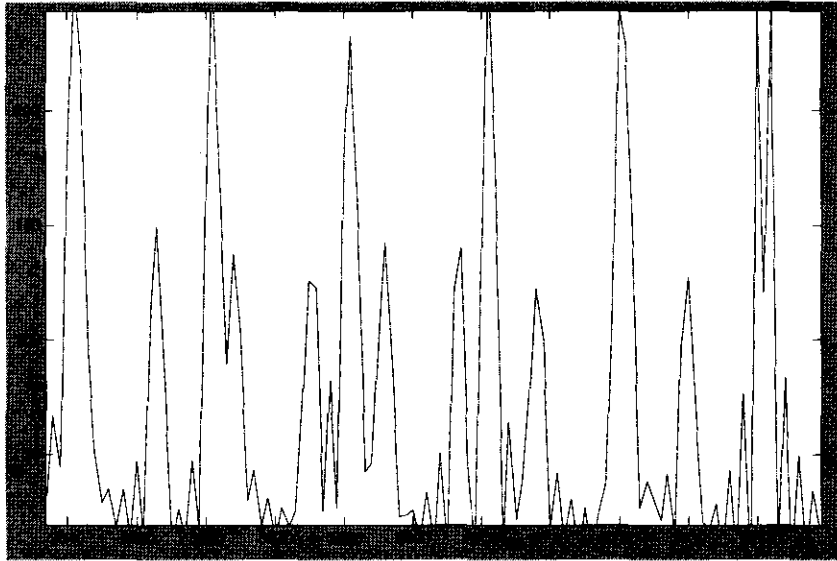


**Figure 365: FFT magnitude spectrum of model signal plus noise ($SV = 45\%$,
 $p = 65\ 536$, $\sigma_n = 0.01$)**

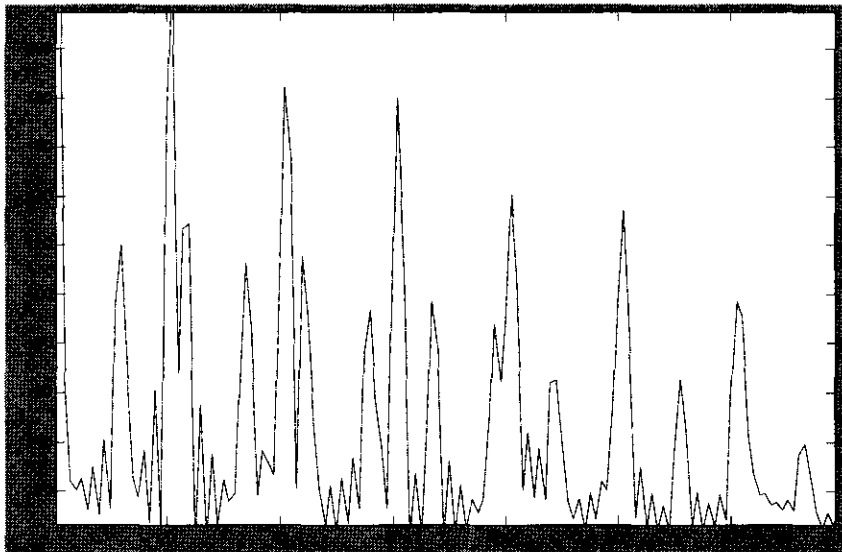


**Figure 366: FFT magnitude spectrum of model signal squared ($SV = 45\%$,
 $p = 65\ 536$, $\sigma_n = 0.01$)**

In the figure above the demodulated spectrum is shown. Note the similarity of the demodulated spectrum compared with the regular spectrum.

11.11.9.5.9 *Small bearing defect signal (DA = 270°, 0g)*

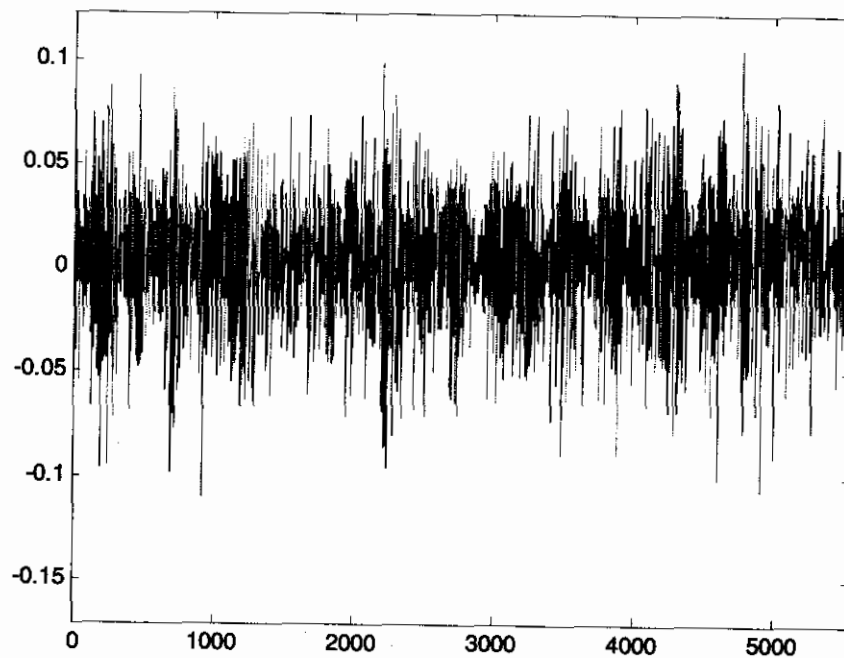


Figure 367: Time domain signal (Vib demo data set 3)

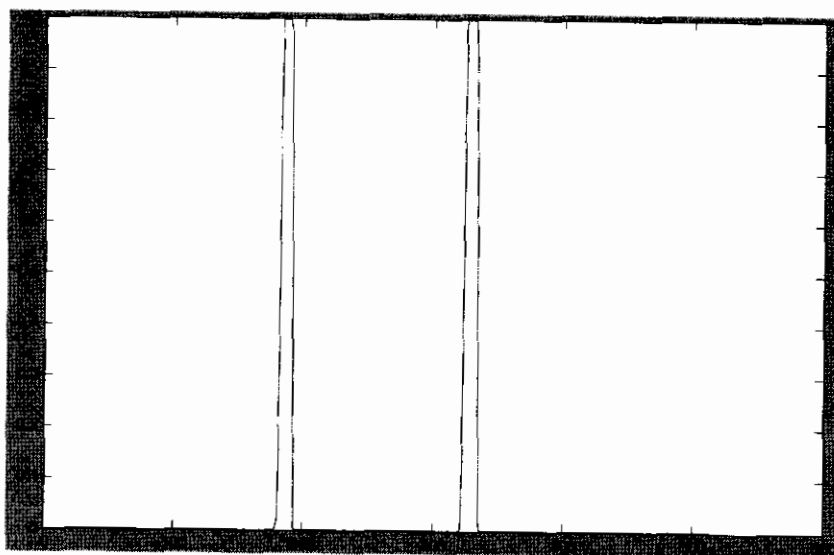


Figure 368: Bandpass filter from 10–20kHz

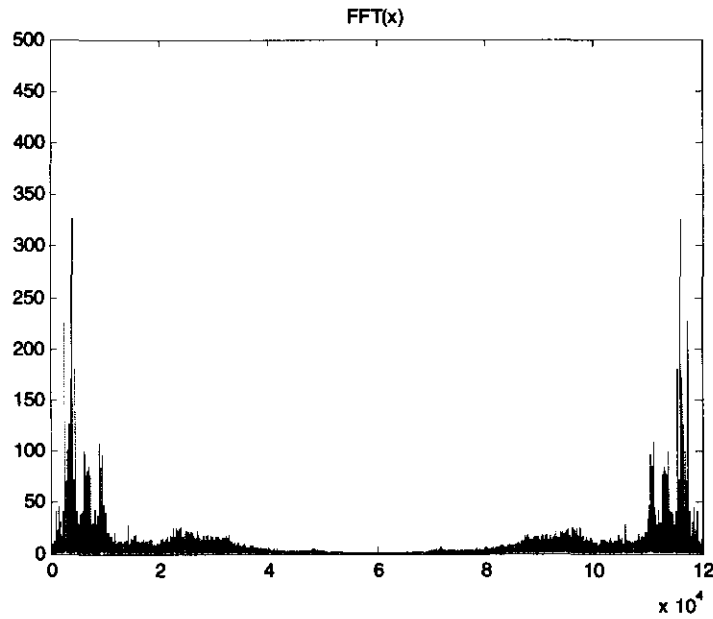


Figure 369: FFT magnitude spectrum of measured vibration signal

In the figure below a zoom plot of the FFT of the signal plus noise is shown. Note that the spectrum is quite similar in shape to that of the demodulated spectrum (shown in the next figure). However, the fundamental defect frequency of 114Hz is not as clear as that of the demodulated spectrum.

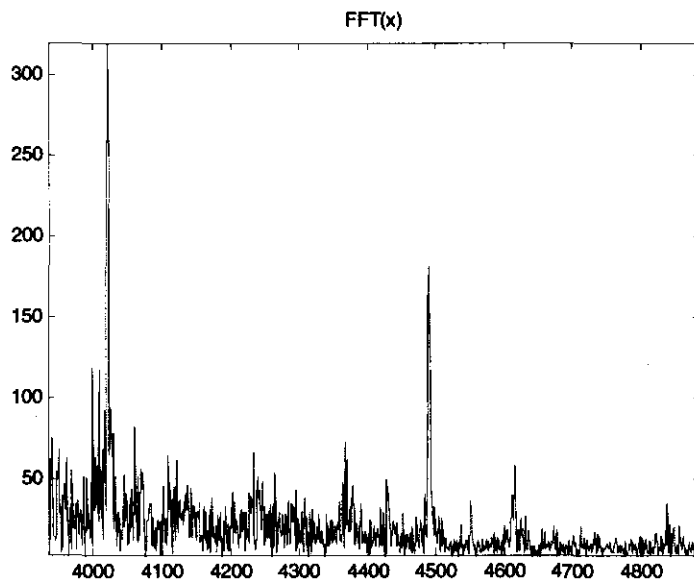


Figure 370: FFT magnitude spectrum of measured vibration signal – zoom plot

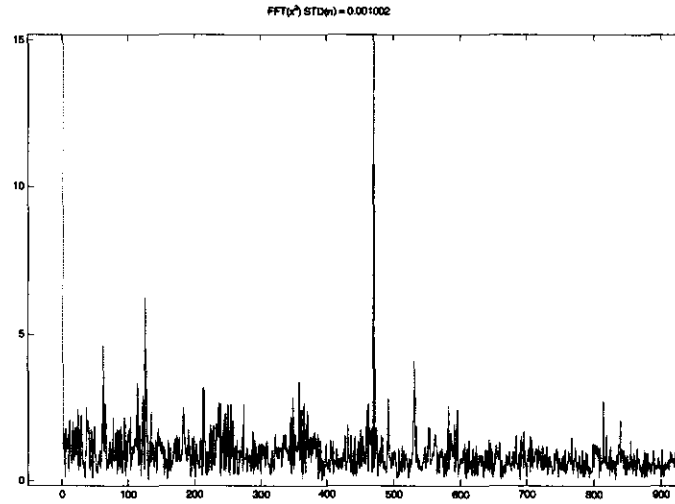


Figure 371: FFT magnitude spectrum of measured vibration signal squared

In the figure above the demodulated spectrum is shown. Note the similarity of the demodulated spectrum compared with that of the regular spectrum.

11.11.9.6 Conclusion

A simplified vibration model, $y = \sin(\omega t)^p$ has been proposed and illustrated on several examples using the FFT. The parameter p controls the spread in time of the simulated defect signal (large values of p correspond to an ideal impulse sequence). It is shown that the demodulated spectrum is better than the regular spectrum when the SNR is greater than $\sqrt{2K_p^{-1}}$. Expressions for the detection ratio of the regular and demodulated spectrum using the FFT and additive noise are derived and illustrated using some examples. The condition for $D_{FFT} > 1$ is $K_{sig}/K_{noise} > 1/(K_x\sqrt{M})$ and is dependent on the buffer length and a constant related to the parameter p (with additive noise). The condition for $D_{FFT2} > 1$ is $K_{sig}/K_{noise} > \sqrt{(1/(0.71K_pK_x\sqrt{M}))}$ and is dependent on the buffer length and constants related to the parameter p (with additive noise). In the simplified model increasing the value of p decreases the signal amplitude of the harmonics (as well as the values of K_x and K_p). For detection of the signal the noise level (for increasing values of p) should hence also be decreased. As long as the SNR is high enough ($> \sqrt{2K_p^{-1}}$) the demodulated spectrum will perform better than the regular spectrum in the detection of the simplified vibration signal.

The effect of speed variations has been considered using a simple step change in frequency. If the low frequency response is filtered out (in a real vibration signal the low frequency components are frequently masked by noise) a zoom plot of the regular spectrum reveals that the spectral smearing near the high frequency resonances is similar in shape to the demodulated spectrum (spectrum of the bandpass-filtered signal squared). However, the demodulated spectrum directly supplies the defect frequency of the high frequency defect pulses. For a large bearing defect in the loading zone the defect frequency could also be determined from the regular spectrum (for example using the real or linear cepstrum). When the SNR is lowered (for example with a small bearing defect), the cepstrum is no longer able to detect the fault (refer to chapter 7 for details).

In summary, the demodulated spectrum is especially useful in the detection of high frequency defect components. The demodulated spectrum is not that useful under very low SNR with additive noise (the SNR should ideally be $> \sqrt{2K_p}^{-1}$ for the demodulated spectrum to perform better than the regular spectrum). However, for multiplicative noise the demodulated spectrum performs better than the regular spectrum¹⁵⁴. Hence the demodulated spectrum should perform better than the regular spectrum with (random) amplitude variations. For speed variations the demodulated spectrum immediately provides the fundamental defect frequency, which is more difficult to obtain with the regular spectrum under conditions of low SNR (for example a small bearing defect).

¹⁵⁴ Squaring of the noise signal provides an additional impulse at DC (0Hz) in the frequency domain (non-zero signal average in the time domain after squaring of the signal), in addition to the (ideally) flat spectrum for other frequency components. For multiplicative noise the signal and noise are multiplied together in the time domain. In the frequency domain this corresponds to a convolution of the signal and noise spectra. As the noise signal (for multiplicative noise) has a non-zero mean the spectrum of the noise signal would also have a component at 0Hz. Hence after multiplication of the signal and noise the resulting convolution in the frequency domain would provide a copy of the signal spectrum (owing to the impulse at 0Hz). The convolution of an impulse at 0Hz with the signal spectrum results in a copy of the signal spectrum. However, a large component at 0Hz would be found for the demodulation spectrum of the noise signal even if the noise signal had a zero mean. Hence the convolution of the demodulation spectrum of the noise signal with the demodulation spectrum of the signal would provide a large contribution from the convolution of the impulse at 0Hz with the demodulation spectrum of the signal. This may help to explain theoretically why the demodulation spectrum performs better with multiplicative noise (squaring of the noise signal provides a copy of the demodulation spectrum of the signal that is added to the other convolutional products, thereby increasing the signal-to-noise level).

