Advanced Vibration Analysis Techniques for Fault Detection and Diagnosis in Geared Transmission Systems

by

B. David Forrester

A thesis submitted for examination for the degree of Doctor of Philosophy

1996

©Copyright 1996 by Barton David Forrester
ACKNOWLEDGMENTS

My thanks goes to the many people at the DSTO Aeronautical and Maritime Research Laboratory who have provided scientific, technical and/or moral support during the course of this research. Particular thanks goes to:

Dr. Bill Schofield for the initial encouragement to start this project and his on-going interest in its progress.

Brian Rebbechi for his scientific and engineering advice throughout the project.

Ken Vaughan for his invaluable technical input in the design and construction of gear rigs and the many hours he spent running the rigs, and Christina Dolan, David Parslow, Peter Stanhope and Madeleine Burchill for their input into the experimental projects.

Dr. Ian Freshwater, of Swinburne University of Technology, for his guidance and advice.
STATEMENT OF ORIGINALITY

This thesis does not contain any material which has been previously submitted for a degree or similar award at any University or other institution. To the best of my knowledge and belief, no material in this thesis has been previously published or written by another person, except where due reference is made.

B. David Forrester
February 1996
ABSTRACT

The primary objective of the research reported in this thesis was the improvement of safety in helicopters by identifying and, where necessary, developing vibration analysis techniques for the detection and diagnosis of safety critical faults in helicopter transmission systems.

A review and, where necessary, expansion of past research is made into

a) the mechanisms involved in the production of vibrations in mechanical systems,

b) the failure modes experienced in geared transmission systems,

c) which failure modes are critical to the safety of helicopters,

d) how the safety critical failure modes affect the vibration signature, and

e) the vibration analysis techniques currently used to detect safety critical failures.

The effectiveness of the currently available vibration analysis techniques is investigated using in-flight vibration data from Royal Australian Navy helicopters and seeded fault data from a purpose built spur gear test rig.

Detailed analysis of techniques for synchronous signal averaging of gear vibration data is undertaken, which includes the development of new methods of modelling and quantifying the effects of synchronous averaging on non-synchronous vibration. A study of digital resampling techniques is also made, including the development of two new methods which provide greater accuracy and/or efficiency (in computation) over previous methods.

A new approach to fault diagnosis is proposed based on time-frequency signal analysis techniques. It is shown that these methods can provide significant improvement in diagnostic capabilities over existing vibration analysis techniques.

Some limitations of general time-frequency analysis techniques are identified and a new technique is developed which overcomes these limitations. It is shown that the new
technique provides a significant improvement in the concentration of energy about the instantaneous frequency of the individual components in the vibration signal, which allows the tracking of small short term amplitude and frequency modulations with a high degree of accuracy. The new technique has the capability of ‘zooming’ in on features which may span only a small frequency range, providing an enhanced visual representation of the underlying structure of the signal.
TABLE OF CONTENTS

ACKNOWLEDGMENTS ................................................................................................. i

STATEMENT OF ORIGINALITY ........................................................................ ii

ABSTRACT ........................................................................................................... iii

TABLE OF CONTENTS ......................................................................................... v

LIST OF ILLUSTRATIONS AND TABLES .............................................................. xv

NOTATION ........................................................................................................... xix

1. INTRODUCTION .............................................................................................. 1

1.1 OBJECTIVE ...................................................................................................... 1

1.2 OUTLINE ......................................................................................................... 1

1.3 SIGNIFICANT ORIGINAL RESEARCH ....................................................... 4

1.4 BACKGROUND .............................................................................................. 5

1.4.1 The need for failure prevention ................................................................. 5

1.4.2 Fault detection, diagnosis and prognosis .................................................... 6

1.4.3 Failure prevention techniques ................................................................... 7

1.4.4 Vibration analysis ..................................................................................... 7

2. A MODEL OF GEARBOX VIBRATION ........................................................... 10

2.1 SOURCES OF GEARBOX VIBRATIONS .................................................... 10

2.1.1 Gear Meshing Vibrations ......................................................................... 10

2.1.1.1 Tooth Profile Deviations ................................................................... 11

2.1.1.2 Amplitude Modulation Effects .......................................................... 13

2.1.1.3 Frequency Modulation Effects ............................................................ 13

2.1.1.4 Additive Impulses ............................................................................. 13
2.1.2 Shaft Vibrations ................................................................. 14
  2.1.2.1 Unbalance ................................................................. 14
  2.1.2.2 Misalignment and Bent Shaft ........................................... 14
  2.1.2.3 Shaft Cracks .............................................................. 15

2.1.3 Rolling Element Bearing Vibrations ................................. 16

2.1.4 Measurement of Gearbox Vibrations ................................. 18
  2.1.4.1 Transmission Path Effects .............................................. 18
  2.1.4.2 Other Measurement Effects ............................................. 19

2.2 MATHEMATICAL MODEL ....................................................... 20
  2.2.1 Gear Vibrations .............................................................. 21
    2.2.1.1 Tooth-meshing vibration .............................................. 21
    2.2.1.2 Additive vibration ...................................................... 23
    2.2.1.3 Tooth-to-tooth meshing .............................................. 23
    2.2.1.4 Multi-mesh gears ...................................................... 24
  2.2.2 Shaft Vibrations .............................................................. 25
  2.2.3 Bearing Vibrations ............................................................ 25
    2.2.3.1 Rotational frequencies ................................................ 26
    2.2.3.2 Angular coordinates .................................................. 29
    2.2.3.3 Bearing vibration signatures ...................................... 30
  2.2.4 Transmission Path Effects ............................................... 33
    2.2.4.1 Static transmission path ............................................. 34
    2.2.4.2 Variable transmission path ........................................ 34
  2.2.5 General Model of Gearbox Vibration ................................ 38

3. FAILURE MODES AND FAULT SIGNATURES ..................... 40
  3.1 CONSEQUENCES OF FAILURE ............................................. 40
    3.1.1 Helicopter transmission systems ................................... 40
    3.1.2 Safety critical failure modes .......................................... 41
      3.1.2.1 Expected safety critical failure modes ............................. 42
      3.1.2.2 Helicopter accident data .......................................... 42
      3.1.2.3 Safety assessment analysis ...................................... 44
3.1.2.4 Summary of safety critical failure modes ........................................... 45

3.1.3 Maintenance and operational considerations ........................................... 46
3.1.3.1 Life limited components ................................................................. 46
3.1.3.2 On-condition maintenance ............................................................. 47
3.1.3.3 The importance of fault diagnosis ................................................... 47

3.2 Failure Prevention ..................................................................................... 48
3.2.1 Usage monitoring .................................................................................. 48
3.2.2 Fault detection and diagnosis ............................................................... 49
3.2.2.1 Temperature monitoring ................................................................. 50
3.2.2.2 Oil debris monitoring ....................................................................... 50
3.2.2.3 Vibration analysis ............................................................................ 50
3.2.3 Prevention of safety critical failures ..................................................... 51
3.2.3.1 Overheating ...................................................................................... 51
3.2.3.2 Incorrect assembly ........................................................................... 51
3.2.3.3 Critical systems ............................................................................... 51
3.2.3.4 Shaft, gear and tooth fracture ......................................................... 52
3.2.4 Priorities for the application of vibration analysis ................................. 52

3.3 Expected Vibration Signatures ................................................................. 54
3.3.1 Shaft fatigue fracture ........................................................................... 54
3.3.2 Gear tooth bending fatigue fracture ..................................................... 54
3.3.3 Gear fatigue fracture ........................................................................... 55
3.3.4 Gear tooth random fracture ................................................................. 55
3.3.5 Shaft unbalance, misalignment and bent shaft ..................................... 56
3.3.6 Gear tooth surface damage ................................................................. 56
3.3.7 Excessive tooth wear and destructive scoring ..................................... 56
3.3.8 Tooth pitting/spalling ......................................................................... 57
3.3.9 Bearing failures .................................................................................... 57

4. REVIEW OF VIBRATION ANALYSIS TECHNIQUES ......................... 58

4.1 TIME DOMAIN ANALYSIS .................................................................... 58
4.1.1 Waveform analysis ................................................................. 58
4.1.2 Time domain signal metrics ............................................... 58
  4.1.2.1 Peak ................................................................................. 59
  4.1.2.2 RMS ................................................................................. 59
  4.1.2.3 Crest Factor ................................................................. 59
  4.1.2.4 Kurtosis ........................................................................... 60
4.1.3 Overall vibration level ....................................................... 60
4.1.4 Waveshape metrics ............................................................ 61
4.1.5 Frequency band analysis .................................................... 62
4.1.6 Advantages ......................................................................... 63
4.1.7 Disadvantages ................................................................... 63
4.1.8 Applicability to safety critical failure modes ................. 63

4.2 SPECTRAL ANALYSIS .............................................................. 63
4.2.1 Conversion to the frequency domain ............................... 64
  4.2.1.1 Bandwidth-time limitation ............................................ 65
  4.2.1.2 FFT Analysers .............................................................. 65
  4.2.1.3 Speed variations .......................................................... 67
4.2.2 Fault detection ................................................................. 68
  4.2.2.1 Spectral comparison ...................................................... 68
  4.2.2.2 Spectral trending .......................................................... 68
  4.2.2.3 Spectral masks .............................................................. 68
4.2.3 Fault diagnosis ................................................................. 69
4.2.4 Advantages ......................................................................... 70
4.2.5 Disadvantages ................................................................... 70
4.2.6 Applicability to safety critical failure modes ................. 70

4.3 SYNCHRONOUS SIGNAL AVERAGING ................................... 71
4.3.1 Fundamental principle ...................................................... 71
4.3.2 Synchronous signal averaging of discrete signals ............ 72
4.3.3 Terminology ......................................................................... 73
4.3.4 Angle domain and shaft orders ....................................... 73
6.1.4.2 Basic signal metrics ................................................................. 149
6.1.4.3 Enhanced signal metrics for pitted Wessex input pinion ............ 149
6.1.4.4 Narrow band demodulation ..................................................... 150
6.1.4.5 Re-analysis of the tapes ......................................................... 152

6.2 SUMMARY OF FINDINGS ............................................................... 155

7. EXPERIMENTAL GEAR RIG DATA .................................................. 157

7.1 SPUR GEAR TEST RIG ................................................................. 157

7.1.1 Description of test rig ............................................................... 157

7.1.1.1 Monitoring equipment ............................................................ 159

7.1.2 Tooth Pitting (Test Gear G3) ..................................................... 161

7.1.2.1 The synchronous signal averages ......................................... 163

7.1.2.2 Trends of signal metrics ....................................................... 165

7.1.2.3 Trended figures of merit ....................................................... 166

7.1.2.4 Narrow band envelope kurtosis ............................................. 167

7.1.2.5 Narrow band demodulation ................................................ 168

7.1.3 Tooth Cracking (Test Gear G6) ................................................ 170

7.1.3.1 Crack initiation and propagation at 45 kW ............................ 172

7.1.3.2 The synchronous signal averages ......................................... 172

7.1.3.3 Trends of signal metrics ...................................................... 173

7.1.3.4 Trended figures of merit ...................................................... 174

7.1.3.5 Narrow band envelope kurtosis ........................................... 175

7.1.3.6 Narrow band demodulation ............................................... 176

7.1.3.7 Crack propagation at 24.5 kW ............................................. 177

7.2 SUMMARY OF FINDINGS .............................................................. 181

8. TIME-FREQUENCY SIGNAL ANALYSIS ....................................... 184

8.1 GENERAL SIGNAL REPRESENTATION ....................................... 185

8.2 TIME-FREQUENCY DOMAIN REPRESENTATIONS ..................... 187

8.2.1 Short-time Fourier transform and spectrogram ........................... 187
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.2.1.1</td>
<td>Multicomponent signals</td>
<td>188</td>
</tr>
<tr>
<td>8.2.1.2</td>
<td>Bandwidth-time limitation</td>
<td>189</td>
</tr>
<tr>
<td>8.2.1.3</td>
<td>Window functions</td>
<td>189</td>
</tr>
<tr>
<td>8.2.1.4</td>
<td>Discrete form</td>
<td>190</td>
</tr>
<tr>
<td>8.2.1.5</td>
<td>Visual representation</td>
<td>191</td>
</tr>
<tr>
<td>8.2.1.6</td>
<td>Advantages and limitations of the spectrogram</td>
<td>193</td>
</tr>
<tr>
<td>8.2.2</td>
<td>The Wigner-Ville distribution</td>
<td>195</td>
</tr>
<tr>
<td>8.2.2.1</td>
<td>The marginal conditions</td>
<td>195</td>
</tr>
<tr>
<td>8.2.2.2</td>
<td>Multicomponent signals and cross-terms</td>
<td>196</td>
</tr>
<tr>
<td>8.2.2.3</td>
<td>‘Negative’ energy</td>
<td>196</td>
</tr>
<tr>
<td>8.2.2.4</td>
<td>Non-stationary signals</td>
<td>197</td>
</tr>
<tr>
<td>8.2.2.5</td>
<td>The Windowed Wigner-Ville Distribution</td>
<td>199</td>
</tr>
<tr>
<td>8.2.2.6</td>
<td>The Discrete Wigner-Ville Distribution</td>
<td>201</td>
</tr>
<tr>
<td>8.2.2.7</td>
<td>Advantages and disadvantages of the WVD</td>
<td>203</td>
</tr>
<tr>
<td>8.2.3</td>
<td>General form of time-frequency distributions</td>
<td>205</td>
</tr>
<tr>
<td>8.2.3.1</td>
<td>Properties of the Cohen class of distributions</td>
<td>205</td>
</tr>
<tr>
<td>8.2.3.2</td>
<td>Reduced Interference Distributions</td>
<td>207</td>
</tr>
<tr>
<td>8.2.3.3</td>
<td>Cross-term attenuation</td>
<td>211</td>
</tr>
<tr>
<td>8.2.3.4</td>
<td>Time-frequency resolution</td>
<td>212</td>
</tr>
<tr>
<td>8.3</td>
<td>SUMMARY</td>
<td>213</td>
</tr>
</tbody>
</table>

9. TIME-FREQUENCY ANALYSIS OF GEAR FAULT VIBRATION DATA | 215 |

9.1 WESSEX INPUT PINION CRACK | 215 |

9.1.1 Comparison with other techniques | 220 |

9.2 WESSEX INPUT PINION TOOTH PITTING | 221 |

9.3 SPUR GEAR TEST RIG - PITTED TEETH | 227 |

9.4 SPUR GEAR TEST RIG - CRACKED TOOTH | 231 |

9.5 SUMMARY OF FINDINGS | 235 |
10. A NEW TIME-FREQUENCY ANALYSIS TECHNIQUE ........ 236

10.1 THEORETICAL DISCUSSION ................................................................. 236

10.1.1 Energy, frequency and bandwidth ..................................................... 237

10.1.2 Desirable properties ......................................................................... 239

10.1.2.1 Global properties ................................................................. 240

10.1.2.2 Time properties ....................................................................... 240

10.1.2.3 Frequency properties ............................................................. 241

10.2 THEORETICAL DEVELOPMENT ......................................................... 242

10.2.1 The spectrogram ............................................................................. 242

10.2.2 The short-time inverse Fourier transform ...................................... 243

10.2.3 The STIFT of the signal and its time derivative ............................... 243

10.2.4 Redistribution of energy ................................................................. 245

10.2.5 Properties of the new distribution .................................................... 246

10.2.5.1 For monocomponent signals ................................................... 246

10.2.5.2 For stationary multicomponent signals ....................................... 246

10.2.5.3 For non-stationary multicomponent signals ................................ 247

10.2.6 Relationship to the Reassignment Method ...................................... 248

10.2.7 The need for bandwidth adjustment ................................................. 249

10.3 EXAMPLES USING THE NEW DISTRIBUTION ............................. 250

10.3.1 Simulated signals ........................................................................... 251

10.3.2 Wessex input pinion cracking .......................................................... 252

10.3.3 Wessex input pinion tooth pitting .................................................... 255

10.3.4 Spur gear test rig tooth pitting ......................................................... 255

10.3.5 Spur gear test rig - cracked tooth ................................................... 257

10.4 SUMMARY .......................................................................................... 259

11. CONCLUSION ....................................................................................... 260

11.1 SUMMARY OF FINDINGS ................................................................. 260

11.2 RECOMMENDATIONS ........................................................................ 263
LIST OF ILLUSTRATIONS AND TABLES

Figure 2.1 Angular contact rolling element bearing geometry........................................ 27
Figure 2.2 Typical epicyclic gear-train........................................................................... 35
Figure 3.1 Schematic of Black Hawk Drivetrain (UTC Sikorsky Aircraft)....................... 41
Table 3.1 Component contribution to transmission related accidents............................ 43
Table 3.2 Safety critical failure modes in helicopter transmissions............................. 45
Table 3.3 Diagnostic evidence of safety critical failure modes................................. 52
Figure 4.1 Synchronous signal averaging................................................................. 72
Figure 5.1 Attenuation of random vibration due to signal averaging....................... 90
Figure 5.2 Attenuation of non-synchronous periodic signal (20.05 order sine wave) .... 93
Figure 5.3 Spectra of reference shaft test signal averages........................................... 102
Figure 5.4 Spectrum of reference shaft after 33 averages.......................................... 103
Figure 5.5 Predicted leakage for test signal (32-288 averages)................................. 104
Figure 5.6 Spectrum of test signal using cubic interpolation (1 average) ....................... 112
Figure 5.7 Spectrum of test signal using cubic interpolation (128 averages)............... 113
Figure 5.8 Spectrum of test signal using low pass filter reconstruction (1 average)..... 114
Figure 5.9 Log spectrum of test signal using LP filter reconstruction (1 average)...... 115
Figure 5.10 Spectrum of test signal using cubic spline interpolation (1 average)........ 119
Figure 5.11 Spectrum of test signal using cubic spline interpolation (128 average).... 120
Figure 5.12 Spectrum of test signal using alternate cubic spline definition (1 average) 121
Figure 5.13 Spectrum of test signal using fifth order spline interpolation (1 average) . 123
Figure 5.14 Log spectrum of signal using fifth order spline interpolation (1 average) . 124
Table 5.1 Performance of digital resampling techniques........................................... 125
Figure 6.1 Spectrum of Wessex MRG WAK143 (263.4 Hrs TSO, 100%)................. 130
Figure 6.2 Recovered 400 hertz speed reference signal.......................................... 131
Table 6.1 Predicted vs. actual RMS values for Wessex input pinion averages............. 132
Figure 6.3 Path of crack in failed Wessex input pinion.............................................. 134
Figure 6.4 Signal averages at 100% rated power (angle domain).............................. 135
Figure 6.5 Signal averages at 100% rated power (frequency domain)....................... 136
Table 6.2 ‘Time domain’ signal metrics for cracked Wessex input pinion............... 137
Table 6.3 Stewart’s Figures of Merit for cracked Wessex input pinion .......... 138
Table 6.4 Narrow Band Kurtosis values for cracked Wessex input pinion ........ 140
Figure 6.6 Demodulated signals at 263.4 hours (103 hours before failure) ..... 141
Figure 6.7 Demodulated signals at 324.3 hours (42 hours before failure) ..... 142
Figure 6.8 Modified demodulated signals at 324.3 hours (42 hours before failure) .... 144
Figure 6.9 Signal averages at 100% rated power (starboard transducer) ....... 148
Figure 6.10 Signal average spectra at 100% rated power (starboard transducer) ... 148
Table 6.5 Basic signal metrics for pitted Wessex input pinion (100% load) .... 149
Table 6.6 Enhanced signal metrics for pitted Wessex input pinion (starboard transducer @ 100% load) ................................................................. 149
Figure 6.11 Demodulated signals at 100% rated power (starboard transducer) .......... 151
Table 6.7 Enhanced signal metrics for pitted Wessex input pinion (5th order spline interpolation, starboard transducer @ 100% load) ................................................................. 153
Table 6.8 Enhanced signal metrics for pitted Wessex input pinion (5th order spline interpolation, starboard transducer @ 75% load) .......................... 153
Figure 6.12 Demodulated signals at 292 hours since overhaul (starboard transducer). 154
Figure 6.13 Band limited spectra at 292 hours since overhaul (8-38 orders) ........ 155
Figure 6.14 Demodulated signals at 339.5 hours since overhaul (100% load) ...... 155
Figure 7.1 Spur Gear Test Rig Schematic .......................................................... 158
Figure 7.2 Spark eroded notch in root of gear tooth ........................................ 159
Figure 7.3 FM4A kurtosis values for test gear G3 ............................................. 162
Figure 7.4 Pitted gear teeth .............................................................................. 163
Figure 7.5 Signal averages of test gear G3 ......................................................... 164
Figure 7.6 Signal average spectra for test gear G3 ............................................. 164
Figure 7.7 Trended angle domain metrics for test gear G3 ............................... 166
Figure 7.8 Trend of ‘Figures of Merit values’ for test gear G3 ......................... 167
Figure 7.9 Narrow band envelope kurtosis values for test gear G3 .................. 168
Figure 7.10 Demodulated signal for test gear G3 (102.9 hours) ........... 169
Figure 7.11 Demodulated signal for test gear G3 (107.9 hours) .............. 169
Figure 7.12 Demodulated signal for test gear G3 (115.5 hours) ............. 169
Figure 7.13 Demodulated signal for test gear G3 (123.1 hours) ........... 169

xvi
Figure 9.8 Zhao-Atlas-Marks distributions of pitted Wessex input pinion
Figure 9.9 Spectrograms of test gear G3 (pitted teeth)
Figure 9.10 Windowed Wigner-Ville distributions of test gear G3 (pitted teeth)
Figure 9.11 Windowed Choi-Williams distributions of test gear G3 (pitted teeth)
Figure 9.12 Zhao-Atlas-Marks distributions of test gear G3 (pitted teeth)
Figure 9.13 Spectrograms of test gear G6 (cracked tooth)
Figure 9.14 Windowed Wigner-Ville distributions of test gear G6 (cracked tooth)
Figure 9.15 Zhao-Atlas-Marks distributions of test gear G6 (cracked tooth)
Figure 10.1 New distribution of frequency modulated signal
Figure 10.2 New distribution of amplitude and frequency modulated signal
Figure 10.3 New distribution of test signals
Figure 10.4 Modified spectrograms of cracked Wessex input pinion
Figure 10.5 Modified spectrograms of pitted Wessex input pinion
Figure 10.6 Modified spectrograms of test gear G3 (pitted teeth)
Figure 10.7 Modified spectrogram of test gear G6 (cracked tooth)
Figure B.1 Sinusoidal (4 per rev) amplitude modulation
Figure B.2 Short term amplitude modulation
Figure B.3 Spectra of sinusoidal (2 per rev) frequency modulated signals
NOTATION

The following is a list of standard notations used in this thesis.

\( t \)  time
\( f \)  frequency
\( x(t) \)  real valued time domain signal
\( s(t) \)  complex valued time domain signal
\( S(f) \)  frequency domain signal
\( \text{Re}[.] \)  real part of a complex value
\( \text{Im}[.] \)  imaginary part of a complex value
\( \bar{x} \)  mean value of vector \( x(t) \)
\( \sigma_x \)  standard deviation (RMS) of vector \( x(t) \)
\( k_x \)  normalised kurtosis of vector \( x(t) \)
\( E \)  total energy of signal
\( a(t) \)  amplitude of a signal at time \( t \)
\( \phi(t) \)  phase of a signal at time \( t \)
\( A(f) \)  amplitude of a signal at frequency \( f \)
\( \Theta(f) \)  phase of a signal at frequency \( f \)
\( \langle f \rangle \)  average frequency of signal
\( \langle t \rangle \)  average time of signal
\( B \)  effective bandwidth of signal
\( T \)  effective duration of signal
\( f_i(t) \)  instantaneous frequency at time \( t \)
\( \tau_g(f) \)  group delay at frequency \( f \)
\( b_i(t) \)  instantaneous bandwidth at time \( t \)
\( t_i(f) \)  instantaneous duration at time \( t \)
\( \mathcal{F}[s(t)] \)  Fourier transform of signal \( s(t) \)
\( \mathcal{H}[x(t)] \)  Hilbert transform of signal \( x(t) \)
\( \rho(t, f) \)  Time-Frequency Distribution (TFD) of signal \( s(t) \)
Chapter 1

INTRODUCTION

1.1 OBJECTIVE

The research reported in this thesis investigates the use of vibration analysis techniques for fault detection and diagnosis in geared transmission systems. The primary objective was the improvement of safety in helicopters by identifying and, where necessary, developing techniques for the detection and diagnosis of safety critical faults in helicopter transmission systems. Reduction of maintenance costs was of secondary consideration. This can be achieved to some extent by improvement in discrimination between critical and non-critical faults; avoiding unnecessary maintenance action.

Although this research concentrated on helicopter transmission systems, many of the techniques described in this thesis are also applicable to other geared transmission systems.

1.2 OUTLINE

The development of failure prevention technology requires the involvement of many engineering disciplines; mechanical, electrical, civil, chemical, and metallurgical. Indeed, Eshleman [29] argued that the lack of clear identity with any one formal academic discipline has hindered the development of machine fault diagnosis and prognosis. In this thesis, it has not been assumed that the reader has a prior knowledge of all disciplines involved. Therefore, the background research is discussed in some detail prior to the presentation of the outcome of the original research components.

Despite the requirement to combine various aspects of different disciplines, an attempt has been made to structure this thesis in an evolutionary fashion. Because of this, the traditional ‘literature survey’ is not presented as a separate chapter, but previous research

1
is discussed and, where necessary, expanded upon as the relevant disciplines are introduced.

A brief introduction to failure prevention and the techniques used (particularly vibration analysis) is given later in this chapter.

In order to identify the areas in which further research on vibration analysis can provide maximum benefit, we need to know

a) how mechanical vibrations are produced,

b) how transmission system components fail,

c) which failure modes are critical to the safety of the helicopter, and

d) how the safety critical failure modes affect the vibration signal.

Chapter 2 addresses the first of these subjects. A review of past research on the mechanisms involved in the production of transmission system vibration is given. This is further developed into a general model of gearbox vibrations which is novel in its approach. Previous models were based on frequency domain representations of vibration and difficulties arose in the description of processes leading to non-stationary signals, such as speed fluctuations and variable transmission path effects. The model developed here is based on the angular position of the various rotating elements which, it is shown, enables complex non-stationary processes to be modelled as simple angular dependencies.

Chapter 3 investigates the consequences of failures on aircraft safety (both in terms of logical expectations and documentary evidence), and discusses the expected vibration characteristics of each of the failure modes. This leads to the identification of the critical failure modes which need to be examined in more detail.

Chapter 4 provides a review of current vibration analysis techniques to determine which methods are best suited to the detection and diagnosis of the safety critical failure modes,
where these methods are deficient and, as a consequence, where further development is needed.

Based on the findings presented in the first four chapters, a number of areas were targeted for further research.

Chapter 5 provides a detailed investigation of synchronous signal averaging techniques which includes; a model of synchronously averaged gear vibration; a theoretical examination of the consequences of synchronous signal averaging, including its effects on non-synchronous vibrations; development of a new method of quantifying and optimising the effects of the synchronous signal averaging process; and detailed examination of coherent resampling techniques including the development of new techniques based on high order spline interpolation.

In Chapter 6, existing vibration analysis techniques are examined and their performance evaluated against actual helicopter fault data.

Chapter 7 describes the development of an experimental spur gear test rig, the generation of seeded faults and the analysis of the seeded fault vibration data using traditional approaches to vibration analysis.

In Chapter 8, a new approach to vibration analysis, based on time-frequency energy distributions, is introduced.

In Chapter 9, the helicopter flight data and seeded fault trial data described in Chapters 6 and 7 are re-examined using the time-frequency analysis techniques discussed in Chapter 8. This shows that, although significant additional diagnostic information can be obtained using time-frequency analysis techniques, existing time-frequency analysis techniques do not provide an adequate representation of small short term frequency modulations which can be important as an early indicator of faults such as cracks in gears and gear teeth.

Chapter 10 describes the development of a new time-frequency analysis technique which has been specifically designed to detect small frequency deviations without sacrificing the
other benefits of time-frequency analysis techniques. The new technique is applied to the helicopter and test rig gear fault vibration data studied in previous chapters and the improvement over the other vibration analysis techniques studied is demonstrated.

1.3 SIGNIFICANT ORIGINAL RESEARCH

The chosen layout of this thesis has resulted in some interspersing of original research with review of research by others. The distinction between the two should be clear when reading the thesis. However, for the benefit of readers who are familiar with the background material, the following is a list of the original research which (in the opinion of the author) provide significant contributions to knowledge in the area of vibration analysis and mechanical failure prevention:

a) A general model of gearbox vibration has been developed which combines gear, shaft and bearing vibrations taking into account transmission path effects (both static and variable) and variable loading (due to torque fluctuations and/or moving load zones). By basing the model in the angle domain, speed variations and the effects of multiple faults can be easily incorporated. The mathematical formulation of the model is described in Chapter 2, Section 2.2.

b) A new method of modelling the effects of synchronous signal averaging has been developed which provides quantification of the attenuation of non-synchronous vibration components. This model is described in Chapter 5, Section 5.2.

c) Based on the model of synchronously averaged vibration data, a method of calculating the ideal number of averages has been developed (Chapter 5, Section 5.3) and, for situations where the ideal number of averages is impractical to implement, a method has been developed for optimising the number of averages (Chapter 5, Section 5.4) which includes methods for the quantification of leakage and the estimation of the signal-to-noise ratio of signal averaged data.

d) An alternate formulation of cubic splines, based on the use of differentiating filters, has been developed (Chapter 5, Section 5.5.2.5) which shows significant improvement
in accuracy over conventional cubic splines when used for digital resampling. A similar approach was used to develop a digital resampling technique using fifth order splines (Chapter 5, Section 5.5.2.6) which showed comparable performance to signal reconstruction using low-pass filters (i.e., close to perfect reconstruction) with a significant reduction in processing time (less than one quarter of the time).

e) An investigation has been made of the application of time-frequency analysis techniques for the study of transmission system vibration. This research is described in Chapters 8 and 9. Time-frequency analysis techniques have been the subject of theoretical study for many years and practical use of some of the techniques has been made in other areas over the last ten years. However, it is believed that the work presented here constitutes the first practical application of these techniques to transmission system fault diagnosis. During the course of this project, a number of papers (including two book chapters) have been published describing portions of this research (see ‘List of Publications’) and, since then, a number of other researchers have started to investigate this area.

f) A new time-frequency analysis technique designed specifically for gear vibration analysis has been developed. The development and application of this new technique is described in Chapter 10.

1.4 BACKGROUND

1.4.1 The need for failure prevention

In many mechanical systems, the cost of component failure can be very high; secondary damage, loss of utility, and human injury or death can far outweigh the value of the failed component. One particularly critical mechanical system is the transmission system in a helicopter. This is required to transmit power from the engines to the rotors, providing lift, thrust and directional control. A mechanical failure in a helicopter transmission system which causes loss of, or significantly reduces, the ability to transmit power can result in catastrophic accident of the aircraft.
The helicopter transmission system needs to efficiently transmit high loads with large speed reduction and, for the sake of aircraft performance, it needs to be of minimal size and weight. This results in complex geared systems in which the components are highly stressed. In addition, duplication of components is impractical, therefore safety margins cannot be increased by redundancy (unlike the propulsion source, for which multiple engines can be used to add a measure of redundancy). Because of this, helicopter transmission systems cannot be made fault tolerant and, in order to increase safety, failure prevention techniques must be used.

1.4.2 Fault detection, diagnosis and prognosis

Failure prevention in any system, be it mechanical, electrical, biological, or whatever, can be viewed at three levels of detail:

a) Fault detection: the essential knowledge that a fault condition exists; without this, no preventative action can be taken to avoid possible system failure.

b) Diagnosis: the determination of the nature and location of the fault; this knowledge can be used to decide the severity of the fault and what preventative or curative action needs to be taken (if any).

c) Prognosis: the forecast or prediction of the probable course and outcome of a fault; based on this, the most efficient and effective method of treatment can be decided upon.

The level of detail required for failure prevention depends very much on the type of system, its perceived value and the consequences of failure. For instance, failure prevention for a relatively inexpensive fuel pump may require only fault detection, with the faulty pump simply being discarded and replaced with a new one. Whereas failure prevention for humans (a mechanism with in-built fault detection capabilities) involves elaborate diagnosis, prognosis and treatment procedures, resulting in huge health care systems.
In the case of helicopter transmissions, and most other geared transmission systems, simple fault detection is not sufficient, as the cost of the system is usually too high to justify total replacement, and some form of diagnosis is required. For simple transmissions which are readily accessible, such as one in an automobile, the diagnosis procedure may involve strip down and visual inspection of the components. In a helicopter, removal and strip down of the transmission is very complex and time consuming, therefore other means of diagnosis must be used. Diagnosis without intervention also allows scope for prognosis; either to predict the possibility of progression from a non-critical fault to a critical fault (indicating the need to closely monitor the fault progression), or to predict the time to failure (allowing scheduling of repairs).

1.4.3 Failure prevention techniques

The most commonly used techniques for failure prevention in geared transmission systems are temperature monitoring, oil debris monitoring and vibration monitoring.

Temperature monitoring is a simple fault detection technique which provides no diagnostic or prognostic capabilities. It is used in a wide range of transmission systems, primarily for the purposes of detecting lubrication and cooling system problems.

Oil debris monitoring is widely used in transmission systems and a large number of detection and diagnosis techniques are available (a review of these is given by Kuhnell [44]). The fundamental limitation in oil debris monitoring is that not all failures generate material debris and, without debris, no detection or diagnosis can take place.

1.4.4 Vibration analysis

Vibration analysis offers the widest coverage of all failure prevention techniques. Virtually any change in the mechanical condition will cause a change in the vibration signature produced by the machine. For a long time, vibration analysis has been used for Rotor Track and Balance (RTB) in helicopters to reduce the vibrations from the rotor
systems. However, vibration analysis techniques are still not widely used for fault
detection and diagnosis in helicopter transmission systems, although this is rapidly
changing.

Past resistance to the use of vibration analysis for fault detection and diagnosis was
probably due to its perceived complexity and lack of rigid procedural guidelines;
although for many years informal (intuitive) vibration analysis has been used for fault
detection in the sense that operators and/or mechanics often detect the presence of a
fault based on a ‘strange sound’. The mechanic who can accurately diagnose a fault by
‘listening’ to the vibrations transmitted via a screwdriver held against the machine casing
is held in some reverence, adding to the perception that vibration analysis is an art rather
than a science.

Early work on the formalisation of vibration diagnostics using spectral analysis
(Blackman and Tukey [4]) progressed slowly through the 1960s, mainly due to the
expense of analysis equipment. The development of the Fast Fourier Transform (FFT) in
1965 (Cooley and Tukey [25]) allowed the development of commercial real-time spectral
analysers and, as the use of these analysers became more widespread, a number of
authors describe the vibration effects of various machine faults and how these could be
diagnosed using spectral analysis (White [82], Minns and Stewart [60], Swansson [77],
Braun [15] and Randall [67]). However, even with the use of spectral analysis, fault
diagnosis using the vibration signature was still relative complex and required specialised
skills.

In the mid 1970s, Stewart [73] made a significant contribution to the use of vibration
analysis as a diagnostic tool for machine faults, especially for gear faults. Based on the
use of synchronous averaging techniques to separate the vibration signatures from
different rotating components, Stewart developed a number of signal metrics (Figures of
Merit) which could be used to indicate (and differentiate) the presence of various
vibration characteristics. The use of these ‘figures of merit’ greatly simplified the
diagnostic task by reducing the complex vibration signals to a handful of parameters
characterising the signal.
Further work by Randall [65] and McFadden [54] in the underlying causes of gear vibration resulted in a better understanding of the correlation between Stewart’s figures of merit and mechanical condition. McFadden [56] showed the importance of phase modulation in the diagnosis of cracks and outlined a signal parameter sensitive to phase modulation.

Up until the late 1980s, machine vibration analysis was based on either the time or frequency domain (spectrum) representation of the vibration signal. Forrester [34] showed that localised machine faults introduce short-term non-stationarities into the vibration signal and that these could be analysed using joint time-frequency signal analysis techniques.

Joint time-frequency analysis techniques were originally developed in the field of quantum mechanics in the 1930s (Wigner [83], Kirkwood [41]) and adapted to signal processing in the 1940s by Gabor [35] and Ville [81]. During the 1950s and 1960s a number of different time-frequency distributions were proposed (Page [62], Margenau and Hill [46], Rihaczek [71]) all of which seemed plausible and showed desirable properties, but produced quite different results. In 1966, Cohen [23] showed that there were an infinite number of ‘plausible’ time-frequency distributions and developed a generalised description from which all of these could be derived. After fifty years of theoretical development, practical use of joint time-frequency distributions has only commenced in earnest over the last 10 years. Cohen [24] and Boashash [8] give comprehensive reviews of time-frequency analysis techniques and applications.
Chapter 2

A MODEL OF GEARBOX VIBRATION

In this chapter, a review is made of the processes involved in the generation of vibrations from various rotating elements in a gearbox. Based on this, a general mathematical model of gearbox vibration is developed which takes into consideration variations in loading and transmission path effects, including variable transmission paths. The model is based on the time-dependent phase of the rotating elements and does not assume a frequency dependency (i.e., it allows for variations in the instantaneous rotational frequencies of the gearbox components).

The model developed here is used in later chapters to describe the effects of various signal processing and vibration analysis techniques, and provides a theoretical basis for the analysis and development of fault detection techniques.

2.1 SOURCES OF GEARBOX VIBRATIONS

The major sources of vibration within a gearbox are the rotating elements, that is the gears, shafts, and bearings. In this section, a review is made of the processes involved in the generation of vibration for each of these elements.

2.1.1 Gear Meshing Vibrations

In a geared transmission system, the main source of vibration is usually the meshing action of the gears. Randall [65] gave a descriptive model of gear vibration in which he divides the vibration into; a periodic signal at the tooth-meshing rate due to deviations from the ideal tooth profile; amplitude modulation effects due to variations in tooth loading; frequency modulation effects due to rotational speed fluctuations and/or non-uniform tooth spacing; and additive impulses which are generally associated with local tooth faults.
2.1.1.1 Tooth Profile Deviations

Deviations from the ideal tooth profile can be due to a number of factors, including tooth deflection under load and geometrical errors caused by machining errors and wear.

2.1.1.1.1 Load Effects

Tooth deflection under load tends to give a signal waveform of a stepped nature due to the periodically varying compliance as the load is shared between different numbers of teeth. The periodic stepped nature of this signal produces vibration components at the tooth-meshing frequency and its harmonics. These vibrations will normally be present for any gear, but the amplitudes are very dependent upon the load. Tooth profile modifications are often used to reduce the level of these vibrations at a particular loading; this compensation will only apply to the design load and it is likely that loadings both below and above the design load will produce higher vibration amplitudes than at the design load. Because of this, it is virtually impossible to predict the expected vibration signature at one load based on measurements at another loading without performing a full dynamic model of the gear based on detailed tooth profile measurements. Therefore, for condition monitoring purposes, it is necessary that vibration measurements are always at the same loading and that this loading be sufficient to ensure that tooth contact is always maintained (i.e., the teeth do not move into backlash).

2.1.1.1.2 Machining Errors

The machining process used in manufacturing the gear often introduces profile errors on the gear teeth. These errors can normally be viewed as a mean error component which will be identical for all teeth and produce vibration at the tooth-meshing frequency and its harmonics, and a variable error component which is not identical for each tooth and, in general, will produce random vibrations from tooth to tooth; note that although these vibrations vary randomly from tooth to tooth they will still be periodic with the gear rotation (i.e., repeated each time the tooth is in contact) which will produce low amplitude vibrations spread over a large number of harmonics of the gear (shaft)
rotational frequency. A special case of machining error gives rise to ‘ghost components’ which will be discussed later.

As the machining errors are geometric variations in tooth profile, the vibration amplitudes are not as load dependent as vibrations due to tooth deflection. As the teeth wear, there is a tendency for these geometric variations to become smaller.

2.1.1.1.3 Ghost Components

The term ‘ghost components’ is used to describe periodic faults which are introduced into the gear tooth profiles during the machining process and correspond to a different number of teeth to those actually being cut. They normally correspond to the number of teeth on the index wheel driving the table on which the gear is mounted during machining, and are due to errors in the teeth on the index wheel (and/or the teeth on its mating gear). The ghost component will produce vibration related to the periodicity of the error as if a ‘ghost’ gear with the corresponding number of teeth existed; that is, at the ghost frequency and its harmonics.

Ghost components, like other machining errors, are fixed geometric errors and therefore should not be very load dependent. This fact may help distinguish ghost components from other periodic vibrations by comparing the relative effects of load on the spectral content of the vibration signal.

Except in the (unlikely) event that the ghost component frequency coincides with a resonance, there is a tendency for the ghost components to get smaller over time as a result of wear.

2.1.1.1.4 Uniform Wear

Systematic wear is caused by the sliding action between the teeth which is present at either side of the pitch circle, but not at the pitch circle itself. Thus, wear will not be uniform over the profile of the tooth and will cause a distortion of the tooth profile. Wear which is uniform for all teeth will cause a distortion of the tooth-meshing frequency which will produce vibration at the tooth-meshing frequency and its harmonics. These may not be apparent until they become larger than the effects due to
tooth deflection. Randall [65] argued that the distortion of the waveform due to heavy wear would generally be greater than that due to tooth deflection and, because the greater distortion leads to more energy in the higher harmonics of the tooth meshing frequency, the effects of wear will often be more pronounced at the higher harmonics of the tooth-meshing frequency than at the tooth-meshing frequency itself.

2.1.1.2 Amplitude Modulation Effects

Randall [65] explained amplitude modulation by the sensitivity of the vibration amplitude to the tooth loading. If the load fluctuates, it is to be expected that the amplitude of the vibration will vary accordingly. A number of faults can give rise to amplitude modulation. These can be generally categorised by the distribution of the fault in the time domain; varying from ‘distributed’ faults such as eccentricity of a gear, which would give a continuous modulation by a frequency corresponding to the rotational speed of the gear, to ‘localised’ faults such as pitchline pitting on a single tooth which would tend to give a modulation by a short pulse spanning approximately the tooth-mesh period, which would be repeated once per revolution of the gear.

2.1.1.3 Frequency Modulation Effects

Variations in the rotational speed of the gears and/or variations in the tooth spacing will give a frequency modulation of the tooth-meshing frequency. In fact, the same fluctuations in tooth contact pressure which give rise to amplitude modulation must at the same time apply a fluctuating torque to the gears, resulting in fluctuations in angular velocity at the same frequency. The ratio of the frequency modulation effects to the amplitude modulation effects is, in general, a function of the inertia of the rotating components; the higher this inertia, the less will be the frequency modulation effects compared to the amplitude modulation effects.

2.1.1.4 Additive Impulses

Most local faults associated with tooth meshing will cause an additive impulse in addition to the amplitude and frequency modulation effects mentioned above. Whereas the modulation effects cause changes in the signal which are symmetrical about the zero line
(in the time domain), additive impulses cause a shift in the local mean position of the signal, that is, the portion affected by the additive impulse is no longer symmetrical about the zero line. Because of the wide frequency spread of a signal with short time duration (Randall [67]), it is quite common for the periodic impulses from a local tooth fault to excite resonances, giving rise to an additive part which is peaked around the frequency of the resonance.

2.1.2 Shaft Vibrations

Vibration related to shafts will generally be periodic with the shaft rotation and appear as components at the shaft rotational frequency and its harmonics.

2.1.2.1 Unbalance

Unbalance occurs when the rotational axis of a shaft and the mass centre of the shaft/gear assembly do not coincide. This causes a vibration component at the shaft rotational frequency, the amplitude of which will vary with shaft speed. Although the force imparted by the mass unbalance will be proportional to the square of the angular frequency of the shaft, the amplitude of the vibration will peak when the rotational speed of the shaft coincides with resonant modes of the shaft (critical speeds).

2.1.2.2 Misalignment and Bent Shaft

Misalignment at a coupling between two shafts usually produces vibration at the shaft rotational frequency plus its harmonics; dependent on the type of coupling and the extent of misalignment. For example, a misaligned universal joint may produce a large twice per revolution component whereas a flexible ‘Thomas’ coupling (consisting of a number of interleaved flexible plates bolted together) may produce a dominant vibration at a frequency equal to the number of bolts in the coupling times the shaft rotational frequency.

Misaligned bearings produce similar symptoms to misaligned couplings. However, they also tend to excite higher harmonics of the shaft rotational frequency.
A bent shaft is just another form of misalignment (with the ‘misalignment’ exhibited at each end of the shaft being in opposite directions) and also produces vibration at the shaft rotational frequency and its lower harmonics.

2.1.2.3 Shaft Cracks

Gasch [36] developed a model of a cracked shaft which showed that in the early stages of cracking, the crack will open and close abruptly once per revolution of the shaft giving a stepped function which is periodic with the shaft rotation. This will give rise to vibration at the shaft rotational frequency and its harmonics. Gasch [36] showed that as the transverse crack depth increases up to half the radius of the shaft, the amplitude of vibration increases with little change in the waveform therefore, the amplitudes at shaft rotational frequency and its harmonics will all tend to increase at the same rate. As the crack depth increases beyond half the shaft radius, the transition of the crack opening and closing occurs over a wider rotational angle of the shaft and the vibration waveform produced tends to become smooth (predominantly a sine wave). Consequently, the higher harmonics of shaft rotational frequency become less significant with larger cracks.

Gasch [36] also looked at the combined case of shaft unbalance and a transverse crack where, dependent upon the size of the unbalance, the shaft rotational speed and the angular relationship of the unbalance force to the crack location, the vibration response at shaft rotational frequency will be either reinforced or suppressed by the response to the mass unbalance. A number of methods can be used to identify shaft cracking in the presence of mass unbalance:

a) In addition to a forward whirl component at shaft frequency, the crack also excites a strong backward whirl component (-ve shaft frequency) which cannot be balanced out. Therefore, an inability to effectively balance a shaft may be indicative of a crack. However, repeated attempts to balance the shaft are not recommended as this can cause additional crack formation and, in the worst case, may lead to a circular crack right around the shaft.
b) Trending of the vibration amplitudes at one, two and three times shaft speed can be used as these will increase to the same extent due to crack growth (up to a crack depth of half shaft radius) and, if the unbalance remains the same, the differences between two successive spectra should give a clear indication of crack growth.

c) During run up or run down, Gasch [36] showed that a crack will display additional resonances at 1/3 and 1/2 critical speed as well as increasing or decreasing the resonance response at the critical speed.

2.1.3 Rolling Element Bearing Vibrations

Howard [38] gives a review of the causes and expected frequencies of vibrations due to rolling element bearings. These will generally be at a much lower amplitude than gear or shaft vibrations and far more complex in nature. From the geometry of the bearing, various theoretical frequencies can be calculated such as the inner and outer race element pass frequencies, cage rotational frequency and rolling element spin frequency. These calculations are based on the assumption that there is no slip in the bearing elements and therefore can only be considered to be an approximation to the true periodicities within the bearing.

McFadden and Smith [50] modelled the vibrations produced by a single point defect in a rolling element bearing as a series of periodic impulses which occur at a frequency related to the location of the fault. The amplitude of each impulse is related to the applied load at the point of element contact with the defect and in the case where the location of the defect moves in relation to the load zone, such as an inner race defect where the inner race is attached to a rotating shaft and the bearing is radially loaded, the amplitude of the impulses will be modulated at the period at which the defect location moves through the load zone. The structural response to each impulse was assumed to take the form of an exponentially decaying sinusoid and consideration was given to the variation in transmission path from the defect location to the measurement point as the shaft rotates.
The model was extended to include the influence of multiple defects [51] which showed that reinforcement or cancellation of various spectral lines could occur based on the phase differences between the vibrations produced by the individual defects.

A further refinement of the McFadden and Smith model was made by Su and Lin [74] which included the influence of shaft unbalance and variations in the diameters of the rolling elements. These produce load variations which result in additional periodic amplitude modulation of the impulses due to the defect. For a rolling element bearing with fixed outer race and inner race rotating at shaft frequency (the most common bearing configuration) they showed that;

a) where there is no unbalance or variation in rolling element diameter, an outer race defect will show no amplitude modulation, an inner race defect will be modulated at the shaft frequency and a rolling element defect will be modulated at the cage rotational frequency;

b) where there is shaft unbalance, the load zone moves with the unbalance whirl of the shaft and an outer race defect is modulated at shaft frequency, an inner race defect will not be modulated and a rolling element defect will be modulated at a frequency equal to the difference between the shaft frequency and the cage frequency; and

c) variations in rolling element diameter will cause a non-uniform load distribution periodic with the cage rotation giving modulation of an outer race defect at the cage rotation frequency, modulation of an inner race defect at the difference between the shaft and cage frequencies and have no affect on a rolling element defect.

Variations in transmission path from the defect location to the transducer will give rise to apparent variations in amplitude and phase of the recorded vibration signal. The transmission path variations will not be significantly changed by the presence of shaft unbalance or variations in rolling element diameter. There will be no variation in transmission path for outer race defects, a variation periodic with shaft rotation for inner race defects and a variation periodic with both cage rotation and rolling element spin (due to alternate contact with inner and outer races) for rolling element defects.
2.1.4 Measurement of Gearbox Vibrations

For an operational gearbox, it is normally impractical to measure the vibrations at their source. Therefore, it is common practice to measure the vibrations at a location remote from the source, typically at a convenient point on the outside of the gearbox casing using a vibration transducer, such as an accelerometer or velocity transducer, which converts a mechanical vibration into an electrical signal. This inevitably leads to some corruption of the vibration signal due to mechanical filtering of the signal from the source to the measurement point (transmission path effects), the interface between the structure and the transducer, inherent limitations of the transducer itself, and inaccuracies in the measurement system.

In developing a model of the vibration recorded from a geared transmission system, these measurement effects need to be taken into consideration.

2.1.4.1 Transmission Path Effects

The transmission path consists of the structure providing a mechanical path from the vibration source to the measurement point. Typically, this will comprise of not only the static structure of the gearbox casing but also the rotating elements (shafts, bearings and gears) intervening between the source and the transducer. The transmission path acts as a filter between the vibration source and the measurement point; that is, it modifies both the amplitude and phase of vibrations dependent on frequency. For instance, sinusoidal vibration at a frequency which corresponds to a resonance within the transmission path may be amplified whereas vibration at a frequency corresponding to a node will be attenuated. Impulses will excite the resonant modes in the transmission path which will normally decay exponentially due to mechanical damping in the system.

Changes in the transmission path can be caused by a number of factors. The obvious one is where the location of the vibration source changes with respect to the measurement location such as with rolling element bearing faults (see above) and with epicyclic gear trains, where the planet gear axes rotate with respect to the ring/sun gear axes. These effects are often taken into account when modelling vibration from the relevant components [49,50,51]. However, other factors which may change the transmission
path such as periodic variations due to the motion of rolling elements in a bearing and/or changes in the number of meshing teeth and non-periodic variations such as flexure of the gearbox casing and variation in operating temperature, are very rarely taken into account. Structural damage, such as cracks in the gearbox casing, will also change the transmission path effects; although it may be possible to detect this type of failure by measurement of the change in transmission path, this is best done by modal analysis techniques using known vibration sources, which is outside the scope of this research program.

The measurement of the frequency response functions associated with the various and varying transmission paths is, at best, very difficult and very often impractical due to the complexity of the structural elements involved and the large number of possible vibration sources. As the transmission path effects are predominantly frequency dependent, it is common practice to perform vibration measurements, for the purpose of machine condition monitoring, at a set constant (or near constant) operating speed. For minor variations due to small speed fluctuations, the variation in the transmission path effects are assumed to be negligible and for variations in the transmission path due to motion of the vibration source, the effects are usually assumed to be linear. As stated above, other factors which may alter the transmission path are generally ignored.

2.1.4.2 Other Measurement Effects

The type of transducer, the transducer/machine interface and the recording mechanism used affect the usable bandwidth and the dynamic range of the measured vibration signal. When use is made of digital recording systems, either digital tape recorders or computer based data acquisition, special care needs to be taken to avoid aliasing. This is usually done by using an anti-aliasing (low pass) pre-filter set at or below half the sampling rate of the digital system. The actual filter cut-off required to avoid aliasing depends upon the type of filter used and/or the type of post processing to be performed (e.g., digital resampling).

These measurement effects are not taken into account in modelling the gearbox vibrations, however it is important to note the inherent limitations they apply to the
analysis of recorded gearbox vibration; frequencies outside the usable (flat) frequency range of the measurement system should not be taken into account (and preferably should be eliminated by pre-filtering) in the analysis and the analysis of signals which are at the lower end of the dynamic range of the system should be avoided.

Some special purpose diagnostic systems make use of features in the monitoring equipment which may otherwise be seen as limitations. For example, a number of systems are available which make use of the resonant frequency of the transducer to perform ‘shock-pulse’ type analysis; a narrow bandpass filter is applied centred at the (known) resonant frequency of the transducer which, it is assumed, will be excited by mechanical impulses. The output of the narrow bandpass filter is amplitude demodulated (typically using a half or full-wave rectifier followed by a low-pass filter) giving an output related to the original impulses. This is a variation of the high frequency resonance technique (or envelope analysis) [38, 52] but using the known transducer resonance in preference to an arbitrarily selected structural resonance. These techniques have proved useful in the diagnosis of bearing faults if the impulses are sufficiently large to adequately excite the required resonances; this may not be the case where the source of the impulse is remote from the transducer.

2.2 MATHEMATICAL MODEL

Based on the above descriptive model of gearbox vibration, a mathematical model will now be developed taking into consideration the functional dependencies and periodicities of the various vibration components.

Conventionally, vibration is expressed as a stationary function of time, having fixed frequency components with the phase of a particular component increasing linearly with time at a slope proportional to its frequency. In order to clearly express the inherent periodicities and angular dependencies in the gearbox vibration, the vibration is expressed in the following derivation as a non-stationary function of time with phase expressed in terms of the angular position of the underlying rotating element. The
(instantaneous) frequency of a particular component is given by the time derivative of its phase $\theta(t)$ (Bendat [3]),

$$f(t) = \frac{1}{2\pi} \frac{d\theta(t)}{dt}.$$ (2.1)

The angular position of a rotating element at time $t$ is given by the integral of its instantaneous frequency since time $t=0$ (Van der Pol [80]):

$$\theta(t) = \theta_0 + 2\pi \int_0^t f(\tau) d\tau,$$ (2.2)

where $\theta_0$ is the angular position at time $t=0$.

In the following, where a function is said to be periodic with an angular variable, it repeats with the modulo $2\pi$ value of the variable (i.e., the value at $\theta + m2\pi$ is the same as that at $\theta$, where $m$ is any integer).

### 2.2.1 Gear Vibrations

The gear vibration is periodic with the angular position of the gear, which is represented in the following by the angular position of the shaft on which the gear is mounted, $\theta_s(t)$.

The vibration due to gear $g$ on shaft $s$, can be described as the sum of the load dependant tooth-meshing vibration and additive vibrations caused by geometric errors:

$$v_{sg}(t) = v_{sg}^{(1)}(t) + v_{sg}^{(2)}(t).$$ (2.3)

#### 2.2.1.1 Tooth-meshing vibration

The gear vibration signal has components at the tooth-meshing frequency and its harmonics with the amplitudes dependent on the mean load and load fluctuations periodic with the shaft rotation (load and amplitude modulation effects) plus a nominally constant amplitude due to mean geometric errors on the tooth profiles (machining errors and wear). The frequency modulation effects due to periodic torque fluctuations and tooth spacing errors are more appropriately described as phase modulation as they only
affect the instantaneous frequency of the vibration, not the mean frequency over the period of rotation.

For the first $M$ harmonics, the vibration at harmonics of tooth-meshing for gear $g$ on shaft $s$, can be expressed as:

\[
v_{sg}^{(1)}(t) = \sum_{m=1}^{M} \left[ A_{sgm}(L, \theta_s) + \bar{E}_{sgm} \right] \cos \left( mN_{sg} (\theta_s + \beta_{sg}(\theta_s)) + \phi_{sgm} \right), \tag{2.4}
\]

where: $\theta_s = \theta_s(t)$ is the shaft angular position,
- $A_{sgm}(L, \theta_s)$ is the amplitude due to tooth deflection (see below)
- $\bar{E}_{sgm}$ is the mean amplitude at harmonic $m$ due to machining errors and wear,
- $N_{sg}$ is the number of teeth on the gear
- $\beta_{sg}(\theta_s)$ is the phase modulation due to torque fluctuations, and
- $\phi_{sgm}$ is the phase of harmonic $m$ at angular position $\theta_s(t) = 0$

The amplitude due to tooth deflection is a function of both the mean load, $\bar{L}$, and fluctuating load periodic with the angular position $\theta_s(t)$

\[
A_{sgm}(L, \theta_s) = \bar{A}_{sgm}(L) \left( 1 + \alpha_{sg}(\theta_s) \right), \tag{2.5}
\]

where: $\theta_s = \theta_s(t)$ is the shaft angular position,
- $\bar{L}$ is the mean load,
- $\bar{A}_{sgm}(L)$ is the amplitude due to the mean load, and
- $\alpha_{sg}(\theta_s)$ is the amplitude modulation due to fluctuating load.

### 2.2.1.1 Phase and amplitude modulations

The load dependant phase modulation, $\beta_{sg}$ in equation (2.4), and amplitude modulation effects, $\alpha_{sg}$ in equation (2.5), are periodic with the shaft rotation $\theta_s(t)$ and therefore they can be expressed as Fourier series.
\[ \beta_{sg}(\theta) = \sum_{k=1}^{K} b_{sgk} \cos(k\theta + \gamma_{sgk}), \]  

and

\[ \alpha_{sg}(\theta) = \sum_{k=1}^{K} a_{sgk} \cos(k\theta + \lambda_{sgk}), \]  

(2.6)

where \( \gamma_{sgk} \) and \( \lambda_{sgk} \) are the phases at \( \theta = 0 \). Note that these modulation effects both have a mean value of zero.

### 2.2.1.2 Additive vibration

In addition to the vibration at harmonics of tooth-meshing and modulations due to load fluctuations, other vibrations related to gear meshing are those caused by geometric profile errors (including ‘ghost components’) which are not identical for each tooth and those due to additive impulses. The amplitude of these additive errors are not significantly affected by load or rotational speed and are expressed as \( K \) harmonics of the shaft rotation:

\[ v_{sg}^{(2)}(t) = \sum_{k=0}^{K} E_{sgk} \cos(k\theta_s(t) + \xi_{sgk}), \]  

(2.7)

where \( \xi_{sgk} \) is the phase of harmonic \( k \) at shaft angular position \( \theta_s=0 \).

Note that a DC component \( (k=0) \) is included in the above, indicating that the additive vibration does not necessarily have a mean value of zero.

### 2.2.1.3 Tooth-to-tooth meshing

Although the tooth-meshing vibration has been expressed above as if it were due solely to, and periodic with, a single gear, it is obvious that tooth-meshing vibrations can only be produced by the interaction between meshing gears. This results in a combined periodicity of the meshing vibration, which is only repeated when the angular position of both gears return to the starting point (i.e., when the same tooth pair returns to mesh).

In general, for meshing gears the vibration will repeat over a period equal to the rotational period of one of the gears times the number of teeth on its mating gear. An
exception to this is when there is a common factor in the number of teeth on the gears, in
which case the normally expected period will be reduced by the value of the common
factor.

The tooth-meshing vibration defined in equation (2.4) represents the mean vibration for
one gear over the tooth-to-tooth meshing period. The unmodulated portion of this
vibration (i.e., the mean tooth-meshing component) will be identical for both gears, with
the modulated portion being contributed by the variations between teeth on the
individual gears. For simplicity, the gears have been treated as separate entities here
(with the mean tooth-meshing vibration being divided between the two gears). The
combination of the vibrations by addition is acceptable where small phase modulations
are involved. Where large phase modulations are involved, the tooth-meshing vibration
can be treated as the mean vibration with both the amplitude and phase modulations
being the mean of those for the two gears.

2.2.1.4 Multi-mesh gears

The meshing vibrations defined above relate to the meshing action of the gear teeth with
those of a mating gear. In the situation where a gear meshes with more than one other
gear (i.e., a multi-mesh gear), a set of meshing vibrations will be produced for each of
the gear meshes. However, it cannot be assumed that the vibration characteristics are
identical for each of the gear meshes; loading conditions, depth of mesh and even the
tooth surface which is in contact can be different for each of the mating gears.

Even though the basic periodicities will be the same for each of the gear meshes, the
vibration waveforms may be quite different and the simplest, and most effective, means
of modelling multi-mesh gears is to treat each gear mesh as if it were a separate gear.
For instance, a gear having 22 teeth and meshing with two other gears would be
modelled as if it were two separate 22 tooth gears.
2.2.2 Shaft Vibrations

Vibrations of shafts and their associated couplings occur at harmonics of the shaft rotational speed. These vibrations are predominantly due to dynamic effects and therefore the amplitudes are assumed to be a function of angular speed. Radial load may affect the shaft unbalance to some extent due to deflection of the geometric centre of the shaft axis. However, this is assumed to be negligible in relation to the dynamic effect of any mass unbalance. The vibration due to shaft $s$ can be expressed as

$$v_s(t) = \sum_{k=0}^{K} A_{sk}(f_s) \cos(k\theta_s + \phi_{sk}).$$  \hspace{1cm} (2.8)

where: $\theta_s = \theta_s(t)$ is the shaft angular position,

$f_s = f_s(t)$ is the shaft rotational frequency,

$A_{sk}(f_s)$ is the amplitude of harmonic $k$ at frequency $f_s$, and

$\phi_{sk}$ is the phase of the vibration at shaft angular position $\theta_s = 0$.

The shaft rotational frequency $f_s(t)$ is the instantaneous shaft frequency (2.1) (i.e., the time derivative of the shaft angular position). For a gearbox operating at constant speed, the variations in the rotational frequency will be due to small fluctuations in torque and it can be assumed that the response to these will be negligible in relation to the response to the mean rotational frequency $\bar{f}_s$. In this case, the shaft vibration (2.8) may be approximated by

$$\bar{v}_s(t) \approx \sum_{k=0}^{K} A_{sk}(\bar{f}_s) \cos(k\theta_s + \phi_{sk}).$$  \hspace{1cm} (2.9)

2.2.3 Bearing Vibrations

Bearing faults are generally modelled with respect to various fault frequencies related to the theoretical rate of contact between a defect surface and other elements within the bearing \cite{38,50,51,74}. These theoretical rates of contact assume no slippage within the
bearing and are, therefore, only approximations to the true rates of contact in an operational bearing.

2.2.3.1 Rotational frequencies

2.2.3.1.1 Inner and outer race frequencies

The inner and outer races of the bearing are almost invariably attached to other components for which the rotational frequency is easily calculated or inferred. Therefore, the inner race frequency \( f_i \) and outer race frequency \( f_o \) are generally known from the configuration of the bearing; typically, the inner race is attached to a shaft and therefore has the same rotational frequency as the shaft and the outer race is static (i.e., it has a constant rotational frequency of zero).

2.2.3.1.2 Cage frequency

As the circumferential velocity of a rolling element due to rotation about its own axis is equal and opposite at the point of contact with the inner and outer races, the axes of the rolling elements (and therefore the cage holding the rolling elements) must move with a velocity equal to the mean of the circumferential velocities of the inner and outer races in order for the elements to maintain contact without sliding.

Where \( V_c \) is the tangential velocity of the cage at the pitch circle and \( V_i \) and \( V_o \) are the circumferential velocities of the inner and outer race respectively,

\[
V_c = \frac{V_i + V_o}{2}. \tag{2.10}
\]

The velocity relationship given in equation (2.10) can be easily converted to a relationship of the rotational frequencies based on the geometry of the bearing shown in Figure 2.1, giving

\[
f_c = \frac{D_i f_i + D_o f_o}{2D}. \tag{2.11}
\]
The relationship between the inner and outer race diameters ($D_i$ and $D_o$ respectively) and the pitch circle diameter ($D$) is based on the diameter of the rolling elements ($d$) and the contact angle ($\alpha$):

$$D_i = D - d \cos(\alpha), \text{ and}$$

$$D_o = D + d \cos(\alpha).$$

(2.12)

The rotational frequency of the cage can be expressed in terms of the pitch circle diameter ($D$), the diameter of the rolling elements ($d$) and the contact angle ($\alpha$) by substituting the relationships given in equation (2.12) into equation (2.11), giving

$$f_c = \frac{(1 - \frac{d}{D} \cos(\alpha))f_i + (1 + \frac{d}{D} \cos(\alpha))f_o}{2}.$$  

(2.13)

### 2.2.3.1.3 Ball pass frequencies

The ball (or roller) pass frequencies are the rate at which rolling elements pass a given point on the inner or outer race. Given the rotational frequencies of the inner race ($f_i$), outer race ($f_o$) and cage ($f_c$) and the number of rolling elements ($N_e$), the theoretical ball
(or rolling element) pass frequencies are easily determined. The inner race ball pass frequency \( f_{bpi} \) is

\[
f_{bpi} = N_e(f_c - f_i) = \frac{N_e}{2}(f_o - f_i)(1 + \frac{d}{D}\cos(\alpha)), \quad (2.14)
\]

and the outer race ball pass frequency \( f_{bpo} \) is

\[
f_{bpo} = N_e(f_o - f_c) = \frac{N_e}{2}(f_o - f_i)(1 - \frac{d}{D}\cos(\alpha)). \quad (2.15)
\]

### 2.2.3.1.4 Ball spin frequency

The ball (or roller) spin frequency is the frequency at which a point on the rolling element contacts with a given race (which is sometimes ambiguously defined as the rate at which the element spins about its own axis [38]; as the angular orientation of the element axis changes with the angular rotation of the cage, this definition may cause confusion). The ball spin frequency \( f_{bs} \) is the reciprocal of the time taken for the element to traverse a distance equal to its diameter \( d \) along either the inner or outer race (both giving the same result):

\[
f_{bs} = (f_c - f_i)\frac{\pi(D - d\cos(\alpha))}{\pi d} = (f_o - f_c)\frac{\pi(D + d\cos(\alpha))}{\pi d}
\]

\[
= \frac{1}{2}(f_o - f_i)\left(\frac{D}{d} - \frac{d}{D}\cos^2(\alpha)\right). \quad (2.16)
\]

### 2.2.3.1.5 Defect frequencies

A defect on the inner or outer race will cause an impulse each time a rolling element contacts the defect. For an inner race defect this occurs at the inner race ball pass frequency, \( f_{bpi} \) (2.14), and the frequency for an outer race defect is the outer race ball pass frequency, \( f_{bpo} \) (2.15).

A defect on one of the rolling elements will cause an impulse each time the defect surface contacts the inner or outer races, which will occur at twice the ball spin frequency, \( 2f_{bs} \) (2.16).
2.2.3.2 Angular coordinates

2.2.3.2.1 Theoretical angular positions

For consistency with the models developed for the other gearbox components, the bearing vibration will be modelled with respect to the angular periodicities within the bearing rather than the stationary theoretical frequencies given above. Given that the inner race angular position ($\theta_i = \theta_i(t)$) and outer race angular position ($\theta_o = \theta_o(t)$) are known, other angular positions can be derived in a similar fashion to the frequencies.

The cage angle ($\theta_c(t)$) is

$$
\theta_c(t) = \frac{(1 - \frac{d}{D}\cos(\alpha))\theta_i + (1 + \frac{d}{D}\cos(\alpha))\theta_o}{2} + \phi_c \quad (2.17)
$$

where $\phi_c$ is the angle at $\theta_i = \theta_o = 0$.

The relative angle of the cage to the inner race ($\theta_{ci}(t)$) is

$$
\theta_{ci}(t) = \theta_c - \theta_i = \frac{1}{2}(\theta_o - \theta_i)\left(1 + \frac{d}{D}\cos(\alpha)\right) + \phi_c. \quad (2.18)
$$

The relative angle of the cage to the outer race ($\theta_{co}(t)$) is

$$
\theta_{co}(t) = \theta_o - \theta_c = \frac{1}{2}(\theta_o - \theta_i)\left(1 - \frac{d}{D}\cos(\alpha)\right) - \phi_c. \quad (2.19)
$$

The angular rotation of the rolling elements ($\theta_b(t)$) (in relation to their angular rotation at $\theta_i = \theta_o = 0$) is

$$
\theta_b(t) = \frac{1}{2}(\theta_o - \theta_i)\left(\frac{D}{d} - \frac{d}{D}\cos^2(\alpha)\right). \quad (2.20)
$$

2.2.3.2.2 Slip and skidding

The theoretical angular positions defined above do not take account of slipping of the cage or skidding of individual rolling elements. Because of the random nature of these events and the difficulties in calculating or detecting their occurrence, it is normal
practice not to take account of them in modelling bearing vibrations. If necessary, an accumulative ‘slip’ angle could be subtracted from the cage angle (2.17) based on a random function with an assumed mean slip rate (this requires a ‘guesstimate’ of the slip rate). A similar function could be subtracted from the rolling element angles (2.20) to account for skidding of the individual elements. As the model presented here is based on angular positions, these two simple adjustments are all that is required to allow for slip and skidding and the remainder of the model (developed in the following sections) requires no modification.

2.2.3.3 Bearing vibration signatures

Because of the natural symmetry in a rolling element bearing, an undamaged bearing under constant load and speed tends toward a state of dynamic equilibrium and generates very little vibration. When a defect, such as pitting, exists in one of the bearing components, a transient force occurs each time another bearing component contacts the defective surface, resulting in rapid acceleration of the bearing components. Although this can cause quite complex reactions within the bearing, for the purpose of modelling, the reaction can be approximated by a short term impulse.

Separate equations are developed for each defect type, and these are subscripted by $b$ for the bearing and $d$ for the defect number to allow for multiple defects on one bearing. The vibration due to multiple defects are combined by addition.

The amplitude of the impulse is affected by both the applied load and the angular velocity at the point of contact. The rate at which the impulses occur are related to the location of the defect. The impulses are modelled using the Dirac operator, $\delta(\theta)$, which has the value $\delta(0) = 1$ and $\delta(\theta \neq 0) = 0$. The mean amplitude of the impulses is assumed to be a function of the mean load ($\bar{L}$) and the mean velocity, which is represented by the mean frequency of the shaft ($\bar{f}_s$) on which the bearing is mounted.

2.2.3.3.1 Motion of the load zone

As the amplitude of the impulses is sensitive to load, the motion of the load zone relative to the defect position needs to be defined. Previous models of bearing vibration have
used different models for different bearing configurations in order to reflect the relative motion of the load zone [38,50,51,74]. To develop a more general model, a variable $\theta_L(t)$ is defined which represents the angular position of the load zone. The relative angular distance between the defect and the load zone is used to define the amplitude modulation due to the motion of the load zone. Typically, $\theta_L(t)$ would be a constant, indicating no motion of the load zone. In the case of shaft unbalance the load zone moves with the shaft rotation and $\theta_L(t)$ would be equal to the shaft angular position $\theta_s(t)$ plus some fixed angle.

### 2.2.3.3.2 Inner race defect

For an inner race defect an impulse will occur each time a rolling element passes the defect location. Allowing for variation in the response to the individual rolling elements and assuming that each rolling element will produce the same response each time it contacts the defect, the vibration produced by the inner race defect can be modelled as a series of impulses repeating periodic with the angular rotation of the cage relative to the inner race $\theta_{ci}(t)$ (2.18). The pulses will also be amplitude modulated periodic with the relative difference between the angular position of the inner race and the load zone ($\theta_i(t) - \theta_L(t)$). For inner race defect $d$ on bearing $b$:

$$I_{bd}(t) = \alpha_{bd}(\theta_{iL}) A_{bd}(\bar{L}, \bar{f}_s) \sum_{n=0}^{N_e-1} \left(1 + a_{bd}(n)\right) \delta\left(\theta_{ci} - n\frac{2\pi}{N_e} - \phi_{bd}\right), \quad (2.21)$$

where:

- $\theta_{ci} = \theta_{ci}(t)$ is the angular position of the cage relative to the inner race,
- $\theta_{iL} = \theta_{iL}(t)$ is the angular position of the inner race relative to the load zone,
- $\alpha_{bd}(\theta)$ is the amplitude modulation due to relative motion of the load zone,
- $A_{bd}(\bar{L}, \bar{f}_s)$ is the impulse amplitude due to mean load and shaft speed,
- $N_e$ is the number of rolling elements,
- $a_{bd}(\theta)$ is the deviation from the mean amplitude due to rolling element $n$, and
- $\phi_{bd}$ is the angular offset to the first element at angular position $\theta_{ci}(t) = 0$. 


2.2.3.3 Outer race defect

The model of vibration produced by an outer race defect is similar to that produced by an inner race defect but with periodic dependencies based on the angular position of the cage relative to the outer race, \( \theta_{co}(t) \) (2.19), and the angular position of the outer race relative to the load zone (\( \theta_{L}(t) - \theta_{o}(t) \)),

\[
O_{bd}(t) = \gamma_{bd}(\theta_{oL}(t))B_{bd}(\tilde{L}, \tilde{f}_s) \sum_{n=0}^{N_e-1} (1 + b_{bd}(n)) \delta(\theta_{co} - \frac{n2\pi}{N_e} - \phi_{bd}),
\]

(2.22)

where:
- \( \theta_{co} = \theta_{co}(t) \) is the angular position of the cage relative to the outer race,
- \( \theta_{oL}(t) \) is the angular position of the outer race relative to the load zone,
- \( \gamma_{bd}(\theta) \) is the amplitude modulation due to relative motion of the load zone,
- \( B_{bd}(\tilde{L}, \tilde{f}_s) \) is the mean impulse amplitude due to mean load and frequency,
- \( N_e \) is the number of rolling elements,
- \( b_{bd}(n) \) is the deviation from the mean amplitude due to rolling element \( n \), and
- \( \phi_{bd} \) is the angular offset to the first element at angular position \( \theta_{co} = 0 \).

2.2.3.4 Rolling element defect

A defect on a rolling element will produce an impulse each time the rolling element contacts either the inner or outer race. Assuming that the response is different for contact on the inner and outer race, but is the same each time the defect contacts a particular race, there will be a periodicity based on the rotation of the rolling element, \( \theta_{b}(t) \) (2.20). The motion of the point of contact with respect to the load zone is given by the relative angle between the cage and the load zone (\( \theta_{c}(t) - \theta_{L}(t) \)):

\[
R^{(i)}_{bd}(t) = \psi_{bd}(\theta_{cL})O^{(i)}_{bd}(\tilde{L}, \tilde{f}_s) \delta(\theta_{b} - \lambda_{bd})
\]

\[
R^{(o)}_{bd}(t) = \psi_{bd}(\theta_{cL})O^{(o)}_{bd}(\tilde{L}, \tilde{f}_s) \delta(\theta_{b} - \pi - \lambda_{bd})
\]

(2.23)

\[
R_{bd}(t) = R^{(i)}_{bd}(t) + R^{(o)}_{bd}(t)
\]
where: \( R_{bd}^{(i)} \) is the vibration due to element contact with the inner race
\( R_{bd}^{(o)} \) is the vibration due to element contact with the outer race
\( \theta_b = \theta_b(t) \) is the angular rotation of the rolling element,
\( \lambda_{bd} \) is the angular offset of the defect location to the inner race at \( \theta_b = 0 \)
\( \Omega_{bd}^{(i)}(L, \tilde{f}_s) \) is the mean inner race impulse amplitude,
\( \Omega_{bd}^{(o)}(L, \tilde{f}_s) \) is the mean outer race impulse amplitude,
\( \theta_{cL} = \theta_{cL}(t) \) is the angular position of the cage relative to the load zone, and
\( \psi_{bd}(\Theta) \) is the amplitude modulation due to relative motion of the load zone.

### 2.2.4 Transmission Path Effects

The above model is of the vibration as seen at the point of contact. The transmission path from the point of contact to the measurement point will act as a filter (i.e., a convolution in time). Variations in the transmission path due to relative motion of the point of contact with respect to the measurement point are modelled as a time varying filter; as the relative motion of the point of contact is usually periodic, a filter with variation periodic with the relative motion of the contact point is used. Other variations in the transmission path, such as those caused by the rotation of intervening bearings, are assumed to be small and have a mean value of zero and are therefore neglected in this model for the sake of simplicity.

In the time domain, the effect of the transmission path on the vibration signal is modelled as a convolution of a vibration signal \( v(t) \) with an impulse response function \( h(t) \) describing the transmission path effects.

\[
x(t) = v(t) * h(t) = \int_{-\infty}^{\infty} v(u) h(t-u) du.
\] (2.24)
2.2.4.1 Static transmission path

Vibration from components having a nominally static transmission path, such as shafts and fixed axis (not epicyclic) gears, can be convolved with a fixed filter response function representing the transmission path effect. This filter will be different for each rotating component.

2.2.4.1.1 Measured vibration for fixed axis gears

The measured vibration for fixed axis gear \( g \) on shaft \( s \), where \( h_{sg}(t) \) is the impulse response of the transmission path, is

\[
x_{sg}(t) = v_{sg}(t)*h_{sg}(t),
\]

(2.25)

where \( * \) represents the convolution integral (2.24).

2.2.4.1.2 Measured vibration for shafts

The measured vibration for shaft \( s \), where \( h_{s}(t) \) is the impulse response of the transmission path, is

\[
x_{s}(t) = v_{s}(t)*h_{s}(t).
\]

(2.26)

2.2.4.2 Variable transmission path

Where variation in the transmission path due to relative motion of the vibration source to the measurement point occurs, such as in bearings and epicyclic gearboxes, the transmission path effect is no longer constant. However, if we assume that the variation in transmission path is small relative to the overall transmission path, the variation in the transmission path effects can be approximated by a linear function. Based on this assumption, the variation in the transmission path effect is modelled as an amplitude modulation of the vibration which would be measure for a fixed (mean) transmission path effect.
2.2.4.2.1 Measured vibration for epicyclic gear-train components

An epicyclic gear-train typically consists of three or more identical planet gears meshing with a sun and a ring gear (see Figure 2.2); they are used where there is a requirement for large speed reductions at high loads in a compact space, such as the final reduction stage in a helicopter main-rotor gearbox. A common configuration for an epicyclic gear-train is with the sun gear rotating about a fixed central axis providing the input to the gear-train, the planet gears orbiting the sun gear, and a stationary ring gear. The axes of the planet gears, which rotate relative to the ring and sun gears, are fixed relative to each other and housed in a ‘planet carrier’ which rotates with the planet axes, providing the output of the gear-train.

![Diagram of an epicyclic gear-train]

In addition to modelling the multiplicity of mesh points (as described in Section 2.2.1.4), the variations in the location of the mesh points due to the orbiting of the planets needs to be taken into account. The variation in the transmission path effect caused by the motion of a mesh point is modelled as an amplitude modulation periodic with the rotation of the planet axis about the sun gear (equivalent to the the planet carrier rotation).
The change in transmission path will affect all of the gear meshes in the same fashion and there will be a constant known phase difference between each of the mesh points, due to the planet axes being at fixed angular positions relative to each other. Therefore, the measured vibration for the complete epicyclic gear-train, including transmission path effects, can be modelled as two equations; one describing the vibration for the ring-planet meshing points and the other for the sun-planet meshing points.

For the ring-planet meshing points:

\[
\chi_r^{(r)}(t) = \left[ \sum_{p=0}^{P_r-1} \Phi_r \left( \theta_r - \frac{p2\pi}{P_r} \right) \left( v_{pr}(t) + v_{rp}(t) \right) \right] * \eta_r(t), \tag{2.27}
\]

where: 
- \( \theta_r = \theta_r(t) \) is the angular position of the planet carrier,
- \( P_r \) is the number of planet gears,
- \( \Phi_r(\theta) \) is the amplitude modulation due to change in transmission path,
- \( v_{pr}(t) \) is the vibration for planet \( p \) at due to meshing with the ring gear,
- \( v_{rp}(t) \) is the vibration for the ring gear due to meshing with planet \( p \), and
- \( \eta_r(t) \) is the filter defining the mean ring-planet transmission path effects.

For the sun-planet meshing points:

\[
\chi_r^{(s)}(t) = \left[ \sum_{p=0}^{P_r-1} \Gamma_r \left( \theta_r - \frac{p2\pi}{P_r} \right) \left( v_{ps}(t) + v_{sp}(t) \right) \right] * \mu_r(t), \tag{2.28}
\]

where: 
- \( \theta_r = \theta_r(t) \) is the angular position of the planet carrier,
- \( P_r \) is the number of planet gears,
- \( \Gamma_r(\theta) \) is the amplitude modulation due to change in transmission path,
- \( v_{ps}(t) \) is the vibration for planet \( p \) at due to meshing with the sun gear,
- \( v_{sp}(t) \) is the vibration for the sun gear due to meshing with planet \( p \), and
- \( \mu_r(t) \) is the filter defining the mean sun-planet transmission path effects.
The total measure vibration for the epicyclic gear-train is simply the sum of the measure vibrations for the ring-planet meshing points and the sun-planet meshing points

\[ \chi_r(t) = \chi_r^{(r)}(t) + \chi_r^{(s)}(t) \]  

(2.29)

2.2.4.2.2 Measured vibration for bearings

The transmission path effects for bearing vibrations can be modelled in a similar fashion to that used for epicyclic gears. A simple amplitude modulation function is used to model the variation in transmission path effect, with a constant impulse response function used to model the mean transmission path effect.

2.2.4.2.2.1 Inner race defects

The transmission path for an inner race defect will move with the location of the defect, that is, it will be periodic with the rotation of the inner race, \( \theta_i(t) \):

\[ I_{bd}(t) = [Q_b(\theta_i - \zeta_{bd})I_{bd}(t)]*q_b(t) \]  

(2.30)

where: \( \theta_i = \theta_i(t) \) is the angular position of the inner race,

\( Q_b(\theta) \) is the amplitude modulation due to variations in the transmission path,

\( \zeta_{bd} \) is the angular offset of the defect at \( \theta_i = 0 \), and

\( q_b(t) \) is the filter defining the mean transmission path effects.

2.2.4.2.2.2 Outer race defects

The transmission path for an outer race defect will be periodic with the rotation of the outer race, \( \theta_o(t) \):

\[ O_{bd}(t) = [C_b(\theta_o - \zeta_{bd})O_{bd}(t)]*c_b(t) \]  

(2.31)

where: \( \theta_o = \theta_o(t) \) is the angular position of the outer race,

\( C_b(\theta) \) is the amplitude modulation due to variations in the transmission path,

\( \zeta_{bd} \) is the angular offset of the defect at \( \theta_o = 0 \), and

\( c_b(t) \) is the filter defining the mean transmission path effects.
2.2.4.2.3 Rolling element defects

The transmission path for a rolling element defect will be periodic with the rotation of the cage, $\theta_c(t)$. Using the separate vibrations for inner and outer race contact given in equation (2.23):

$$\rho_{bd}^{(i)}(t) = \left[ K_b^{(i)}(\theta_c - \nu_{bd}) R_{bd}^{(i)}(t) \right] \ast \kappa_b^{(i)}(t)$$

$$\rho_{bd}^{(o)}(t) = \left[ K_b^{(o)}(\theta_c - \nu_{bd}) R_{bd}^{(o)}(t) \right] \ast \kappa_b^{(o)}(t)$$

$$\rho_{bd}(t) = \rho_{bd}^{(i)}(t) + \rho_{bd}^{(o)}$$

where: $\theta_c = \theta_c(t)$ is the angular position of the cage,

$\nu_{bd}$ is the angular offset of the defect at $\theta_c = 0$,

$K_b^{(i)}(t)$ defines inner race transmission path variations,

$K_b^{(o)}(t)$ defines outer race transmission path variations,

$\kappa_b^{(i)}(t)$ is the filter defining the mean inner race transmission path, and

$\kappa_b^{(o)}(t)$ is the filter defining the mean outer race transmission path.

2.2.5 General Model of Gearbox Vibration

The vibration measured by a transducer mounted on the gearbox casing is the sum of all the vibrating components in the gearbox modified by the transmission path effects. This is the sum of the vibration for all fixed axis shafts plus the meshing points of their mounted gears, plus the vibration from epicyclic gear trains and any bearing defects. Using the definitions of measured vibrations given in Section 2.2.4, the measured vibration can be defined as:
\[
x(t) = \sum_{s=1}^{S} \left( x_s(t) + \sum_{g=1}^{G_s} x_{sg}(t) \right) + \sum_{r=1}^{R} \chi_r(t) \\
+ \sum_{b=1}^{B} \left( \sum_{d=1}^{D_{bi}} t_{bd}(t) + \sum_{d=1}^{D_{bo}} o_{bd}(t) + \sum_{d=1}^{D_{bb}} \rho_{bd}(t) \right)
\]

(2.33)

where:  
- \( S \) is the number of fixed axis shafts in the gearbox,
- \( G_s \) is the number of gears on shaft \( s \),
- \( R \) is the number of epicyclic gear trains in the gearbox,
- \( B \) is the number of bearings in the gearbox,
- \( D_{bi} \) is the number of inner race defects on bearing \( b \),
- \( D_{bo} \) is the number of outer race defects on bearing \( b \),
- \( D_{bb} \) is the number of rolling element defects on bearing \( b \),
- \( x_s(t) \) is the vibration due to shaft \( s \) (2.26),
- \( x_{sg}(t) \) is the vibration due to gear \( g \) on shaft \( s \) (2.24),
- \( \chi_r(t) \) is the vibration due to epicyclic gear train \( r \) (2.29),
- \( t_{bd}(t) \) is the vibration due to inner race defect \( d \) on bearing \( b \) (2.30),
- \( o_{bd}(t) \) is the vibration due to outer race defect \( d \) on bearing \( b \) (2.31), and
- \( \rho_{bd}(t) \) is the vibration due to rolling element defect \( d \) on bearing \( b \) (2.32).
In this chapter, analysis of failure mechanisms is made to identify

a) potential consequences of the failure,

b) methods which can be used to prevent the failure, and

c) where appropriate, the expected vibration signature produced by the failure and how this would be portrayed in the vibration model established in Chapter 2.

Using the results of this analysis, a suite of faults will be selected for further investigation in later chapters. The selection criteria are based on the perceived benefits to be derived from further development of vibration analysis techniques specific to particular faults.

3.1 CONSEQUENCES OF FAILURE

The consequences of a particular failure mode depend on the context in which the ‘failed’ component is operating, therefore, at this point some definition of the operational context needs to be made. The main aim of this research project is to improve fault detection and diagnosis in helicopter transmission systems, with a primary view to improvement in helicopter safety.

3.1.1 Helicopter transmission systems

In a helicopter, the transmission system provides a critical link between the engines and the rotors, which provides lift, thrust and directional control. In addition to having the capability to reliably transmit high loads with large reduction ratios, a helicopter transmission system needs to be of minimal size and weight. This dual requirement results in highly stressed components which need to be designed and manufactured to a high degree of precision and, consequently, at a high cost. For example, a gear set in a helicopter transmission may be required to transmit 20 times the power of a similarly
sized gear set in an automobile transmission and may cost two to three orders of magnitude more (Drago [28]).

As an added complication, duplication of the main helicopter transmission components is impractical therefore little redundancy can be built into the system. Figure 3.1 shows a schematic of a typical transmission system in a modern twin engine helicopter (Sikorsky Black Hawk). Except for the engine input modules and their associated accessory drive systems, there is no duplication (or redundant components) in this transmission; loss of power transmission capability in the main module will result in loss of lift and in the tail rotor drive will result in loss of control, both of which can lead to potentially catastrophic accidents.

Figure 3.1 Schematic of Black Hawk Drivetrain (UTC Sikorsky Aircraft)

3.1.2 Safety critical failure modes

It would be expected that those failures which result in loss of power transmission capabilities to provide lift and/or directional control would have the most serious consequences on helicopter safety.
3.1.2.1 Expected safety critical failure modes

Of the failure modes which may occur in geared transmission systems (these are summarised in Appendix A), those which result directly in loss of power transmission capability are hot flow due to overheating and fractures in gears and shafts (either due to fatigue or overload). Other faults such as excessive wear, destructive scoring, interference, destructive pitting and spalling, and tooth surface damage in gears can eventually lead to fracture.

Bearing failures, in themselves, will not cause loss of power transmission however, secondary damaged such as abrasive wear due to bearing debris and gross misalignment due to collapse of a bearing can eventually lead to failure of shafts or gears which can result in loss of power transmission.

3.1.2.2 Helicopter accident data

The validity of the assumptions made relating to safety critical failure modes can be checked by examination of helicopter accidents statistics.

3.1.2.2.1 Transmission failure related accidents

Astridge [1] reviewed documentation on helicopter accidents in the world-wide civil fleet which showed that 22% of all airworthiness-related accidents (causing death or serious injury, or resulting in loss or substantial damage to the aircraft) between 1956 and 1986 were attributed to transmission system failures. The other major causes were engines (28%) and rotors (27%).

Astridge [1] provided a further breakdown of transmission related accidents by component (reprinted in Table 3.1) which shows that approximately 75% of all transmission related helicopter accidents were caused by shaft and gear failures.

Note that the data given in Table 3.1 relate to component failures causing serious accidents rather than all component failures and, therefore, provides a meaningful statistic for the impact of component failure on aircraft safety but do not provide any information on actual component failure rates. For instance, bearing failures in a
helicopter transmission would occur far more often than tail rotor drive shaft failures however, bearing failures very rarely result in aircraft accidents whereas tail rotor drive shaft failures almost invariably do.

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentage of accidents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tail rotor drive shaft</td>
<td>31.9</td>
</tr>
<tr>
<td>Gears</td>
<td>19.1</td>
</tr>
<tr>
<td>Main rotor drive shaft</td>
<td>14.9</td>
</tr>
<tr>
<td>Lubrication system</td>
<td>8.5</td>
</tr>
<tr>
<td>Main gearbox input shaft</td>
<td>8.5</td>
</tr>
<tr>
<td>Bearings</td>
<td>4.3</td>
</tr>
<tr>
<td>Freewheels</td>
<td>4.3</td>
</tr>
<tr>
<td>Cooling fan drive</td>
<td>4.3</td>
</tr>
<tr>
<td>Unknown</td>
<td>4.3</td>
</tr>
</tbody>
</table>

*Table 3.1 Component contribution to transmission related accidents*

3.1.2.2 Component failure modes

Further investigation of the failure modes involved in the accidents listed in Table 3.1, showed that in the case of shafts and gears, all the failures were fractures (Astridge [1]) either due to overload or fatigue.

In the cases where bearings were listed as the primary failure, subsequent gear fracture was identified as the ultimate cause of accident except for one case, in which the failed bearing was the tail-rotor pitch control bearing. This particular failure highlights the need to look at the operational context of a component when assessing the impact of failure; although no loss of drive occurred, the failure of the bearing caused loss of adequate control of the tail-rotor pitch.

Lubrication system failures (due to oil pipe/connection and filter bowl failures or low oil levels) and cooling fan drive failures (due to drive gear fracture) caused disintegration of the gearbox due to overheating (see hot flow in Appendix A.1.5.2).

Freewheel (clutch) failures resulting in loss of power transmission were identified as being due to incorrect installation.
3.1.2.2.3 Summary of accident data

The accident data provide confirmation of the expected safety critical failure modes identified in Section 3.1.2.1 and, in addition, provide some indication of the statistical relevance of failures in particular components.

3.1.2.3 Safety assessment analysis

In the above, an attempt has been made to identify safety critical failure modes in helicopter transmissions in a very general sense. Failures which may lead to accidents were identified and statistics on accident rates in the world-wide civil fleet were examined to give a statistical assessment of safety critical components in all helicopters.

More specific safety assessment analysis can be done on individual aircraft types based on design data, models and, where available, in-service failure statistics to determine those components and failure modes which are safety critical for a particular aircraft. A number of formalised failure analysis approaches exist (Astridge [1]) including:

a) failure modes and effects analysis (FMEA), where the progression of all the failure modes (as given in Appendix A) for all components in the transmission are analysed to establish the ultimate effect on aircraft safety; and

b) fault tree analysis (FTA), where the final effect is considered and then traced back to the possible primary causes.

Although safety assessment analysis is outside the scope of this research project, it does have a significant impact on how the outcome of this research will be used in the future. The intended aim of safety assessment analysis is to identify potentially catastrophic faults and provide a model of the fault progression from initiation to failure, allowing strategies to be developed to reduce the probability of the failure occurring in service. The probability of the failure may be reduced by component redesign or restriction of operational limits, however, reliable fault detection and diagnosis offers a much more cost effective strategy. In this instance, the fault detection and diagnosis forms an integral part of the transmission system design, and it is imperative that the fault
detection and diagnosis tools are sufficiently sensitive and reliable to perform their intended function.

### 3.1.2.4 Summary of safety critical failure modes

<table>
<thead>
<tr>
<th>Failure</th>
<th>Failure Mode</th>
<th>Cause</th>
<th>Contributing factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shaft fracture</td>
<td>fatigue</td>
<td>unbalance</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>misalignment</td>
<td>coupling</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>bearing failure</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bent shaft</td>
<td></td>
</tr>
<tr>
<td></td>
<td>overload</td>
<td>interference</td>
<td>incorrect assembly</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>bearing failure</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>operational</td>
</tr>
<tr>
<td>Gear fracture</td>
<td>fatigue</td>
<td>life limit exceeded</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>surface damage</td>
<td></td>
</tr>
<tr>
<td></td>
<td>resonance</td>
<td>design</td>
<td></td>
</tr>
<tr>
<td>Tooth fracture</td>
<td>bending fatigue</td>
<td>life limit exceeded</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>surface damage</td>
<td>process related</td>
</tr>
<tr>
<td></td>
<td></td>
<td>thin tooth</td>
<td>excessive wear</td>
</tr>
<tr>
<td></td>
<td>random fracture</td>
<td>surface damage</td>
<td>process related</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>foreign object</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>pitting/spalling</td>
</tr>
<tr>
<td></td>
<td>overload</td>
<td>interference</td>
<td>incorrect assembly</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>bearing failure</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>operational</td>
</tr>
<tr>
<td>Overheating</td>
<td>lubrication</td>
<td>insufficient oil</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>loss of oil</td>
<td>oil line failure</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>filter bowl failure</td>
</tr>
<tr>
<td></td>
<td>insufficient cooling</td>
<td>cooling fan failure</td>
<td>shaft/gear fracture</td>
</tr>
</tbody>
</table>

*Table 3.2 Safety critical failure modes in helicopter transmissions*

From both the expected failure modes and helicopter accident statistics, the ultimate failure modes which can result in aircraft accident are (in order of priority):

a) shaft fracture,

b) gear and gear tooth fracture,

c) overheating,
d) incorrect assembly, and
e) component deterioration (such as bearing failure) in critical control systems.

Table 3.2 summarises a number of different processes (described in Appendix A) which can lead to shaft fracture, gear/tooth fracture, or overheating; including the failure modes, possible causes precipitating the failure, and factors which may contribute to the cause of failure.

3.1.3 Maintenance and operational considerations

In the above, only the failure modes contributing to aircraft accidents have been considered. From the primary point of view of this research project (aircraft safety), this is of particular relevance. However, as original equipment, repair and ongoing maintenance costs of helicopters continue to escalate, there is increasing pressure on helicopter operators (and airframe manufacturers) to find areas in which reductions can be made in the cost of ownership; this is especially true in military forces, because of the world-wide pressure on governments to reduce defence budgets.

3.1.3.1 Life limited components

For most helicopters, maintenance of the transmission system is based on predicted minimum safe lives of the transmission system components; gearboxes are removed and overhauled after a specified number of flight hours, with replacement of life limited components and inspection (and replacement where necessary) of all other components. This procedure is very expensive both in labour and part replacement costs. In addition, because of the safety critical nature of helicopter transmission systems and the difficulty in predicting helicopter operational regimes, the safe life estimates need to be very conservative meaning that, more often than not, components are replaced when they may still have substantial usable life left.
3.1.3.2 On-condition maintenance

Because of the high cost of overhaul (approaching one million Australian dollars for a large military helicopter main rotor transmission) there is a trend to switch to a system of ‘on-condition’ maintenance and modular gearboxes in more modern aircraft (such as the Sikorsky Black Hawk/Seahawk). Rather than performing regular overhauls on large gearboxes, a gearbox module is repaired or replaced only when necessary. This maintenance philosophy can lead to very large savings in maintenance cost over the life of the aircraft, however, it places a much greater burden on fault detection and diagnosis systems.

At present, condition assessment of transmission system components in the Black Hawk and Seahawk relies upon chip detectors in each of the transmission system modules. These can only give warning of failure modes which produce debris; such as wear, scoring, pitting and spalling in gears and bearings. Although a number of these failure modes can contribute to safety critical component failures if left undetected (as summarised in Table 3.2), fractures themselves do not generate any debris therefore any fracture which is not the result of the progression of a debris producing failure mode will go undetected. This places even greater emphasis on the need to develop detection methodologies for these failure modes as, without regular overhaul and component replacement, any undetected critical fault will continue to propagate until catastrophic failure occurs.

3.1.3.3 The importance of fault diagnosis

The ability to detect faults early, although important in terms of aircraft safety, may in fact have a negative effect on operation costs without reliable diagnostic capabilities. Without proper diagnosis, unnecessary costly maintenance action may need to be carried out due to a ‘potentially’ safety critical fault which turns out to be a less serious fault. For instance, pitting on a gear tooth, which progresses at a low rate and may even heal over, may be confused with bending fatigue failure, which can very quickly lead to catastrophic failure.
3.2 FAILURE PREVENTION

In Section 3.1.2.4, the safety critical failure modes which need to be addressed have been identified and the processes leading to those failures have been summarised in Table 3.2. Two main approaches to prevention of failure can be taken;

1. reducing the probability of the failure occurring, either by
   a) redesign or
   b) (for fatigue failures) monitoring the fatigue life usage and replacing the component when the statistical probability of failure exceeds a specified limit; or

2. detecting the initiation of the failure mode and replacing the component before catastrophic failure occurs.

Recently, substantial development of on-board Health and Usage Monitoring Systems (HUMS) [1] has been performed, which combine the approaches of usage monitoring (1(b)) and health monitoring (2) above. This work has mainly been prompted by the efforts of civil aviation bodies, particularly the UK Civil Aviation Authority (CAA), because of the large disparities between accident rates in helicopters and fixed-wing aircraft.

3.2.1 Usage monitoring

Detail discussion on usage monitoring is outside the scope of this thesis, however, it does form an adjunct to health monitoring (fault detection and diagnosis) in HUMS and, in that capacity, it warrants some mention here.

To a certain extent, usage monitoring alleviates some of the problems associated with predicted component life limits (see Section 3.1.3.1) by measuring the actual flight regimes experienced rather than those assumed when establishing component lives. In principle, this should allow the used fatigue life of a component to be predicted with far greater accuracy than by the use of an average life limit based on flight hours under an
assumed mix of flight regimes. However, the calculated fatigue life for a component is still a statistical parameter; only the probability of failure at any time can be stated.

Usage monitoring can also indicate when overload conditions have occurred. Aircrew are currently relied upon to detect overload conditions by observance of cockpit instrumentation. However, this is not a reliable detection method as situations which may lead to overload are often emergency or extreme flight conditions; diligent observance of cockpit instrumentation is usually not a priority for aircrew under these conditions.

Evidence from usage monitoring can be used in conjunction with other evidence in the fault diagnosis process, or indicate an increase in the probability of a fault occurring, which can be used to direct more monitoring effort towards detection of that particular fault.

### 3.2.2 Fault detection and diagnosis

In order to detect (and diagnosis) an impending failure, a good understanding of the evidence relating to the failure mode and methods of collecting and quantifying the evidence is needed. Although many faults may be easily detectable by physical examination of a component, using techniques such as microscopy, X-ray, dye penetrants, magnetic rubber, etc., these methods usually cannot be performed without removal of, and in some cases physical damage to, the component. Whilst physical examination techniques still play a critical role during manufacture, assembly and overhaul, they are impractical in an operational helicopter transmission and other (non-intrusive) fault detection methods need to be employed for routine monitoring purposes.

Almost all failure modes in the rotating elements in geared transmission systems will cause some change in the vibration signature, and many will produce material debris and/or increased friction causing a change in surface temperature. Most helicopters have some form of temperature and oil debris monitoring systems but very few currently have vibration monitoring systems.
3.2.2.1 Temperature monitoring

Often the temperature monitoring system is too crude to pick up individual component faults and is used more as a warning of overheating due to problems such as insufficient cooling and degradation or loss of lubricant. In order to detect temperature changes due to component degradation, the temperature of the individual component or, preferably, the temperature gradient across the component (by measuring the difference between the inlet and outlet oil temperatures) needs to be monitored. Monitoring of this type is impractical for every component in a large helicopter transmission system however, it may be practical on certain critical components.

3.2.2.2 Oil debris monitoring

A large number of oil debris monitoring techniques are available, ranging from simple magnetic plugs and chip detectors through to sophisticated in-line inductive debris monitoring systems. A number of oil debris analysis techniques are also available, such as spectrometric oil analysis, X-ray diffraction, scanning electron microscopy, particle counters, and ferrography, which can provide some diagnostic information based on the quantity, elemental composition, form, size, and size distribution of the particles. Detailed and careful analysis can give an indication of the type of component (e.g., gear or bearing), failure mode (e.g., wear, pitting/spalling, or scoring) and rate of degeneration. However, it is not possible to differentiate between components of the same material composition and sufficient debris must be ‘washed’ to the collection site (or remain suspended in the oil) for the fault to be detected. Kuhnell [44] gives a review of oil debris detection, collection and analysis techniques.

3.2.2.3 Vibration analysis

At the present time, regular vibration analysis is not used widely for fault detection and diagnosis in helicopter transmission systems, however, this is rapidly changing. Part of the resistance to the use of vibration analysis in the past was its perceived complexity; maintenance personnel find it easy to visualise the causal relationship between component degradation and increased temperature or wear debris, but find the correlation with a vibration signature far more difficult to make.
With the advent of cheap, powerful microprocessors and the continuing development of more sophisticated vibration analysis systems which reduce the complexity of the diagnostic information (e.g., from a 1000 line spectrum to a few ‘condition indices’), the potential of vibration analysis as a diagnostic tool is starting to be realised. Vibration analysis techniques will be discussed in detail in the next chapter and expanded further in the remainder of this thesis.

3.2.3 Prevention of safety critical failures

In light of the above, methods for the prevention of the previously identified safety critical failure modes will now be discussed.

3.2.3.1 Overheating

In many older helicopters, transmission system disintegration will occur in a matter of seconds without lubricant; explaining the relatively high number of accidents attributed to lubrication system failures in Table 3.1. In modern helicopters (mainly due to legislative measures), transmission systems will continue to operate for 30 minutes or more after a lubrication system failure, giving the pilot time to land the aircraft before catastrophic failure occurs. Because of this built-in tolerance to lubrication system problems in current helicopters, it is anticipated that overheating will be less of a safety critical issue in future. That is, the problem has already been addressed by the manufacturers and solved by transmission system redesign.

3.2.3.2 Incorrect assembly

Incorrect assembly can only be prevented by due diligence.

3.2.3.3 Critical systems

Where critical systems are identified in which any deterioration of the components can endanger the aircraft, individual monitoring systems (vibration, oil debris and/or temperature) can be used to simplify the fault detection and diagnosis process.
### 3.2.3.4 Shaft, gear and tooth fracture

Table 3.3 gives a summary of the expected diagnostic evidence generated by the various processes which can lead to the identified safety critical failure modes.

<table>
<thead>
<tr>
<th>Failure Mode</th>
<th>Cause</th>
<th>Contributing Factors</th>
<th>Diagnostic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>V</td>
</tr>
<tr>
<td><strong>Shaft fracture</strong></td>
<td>fatigue</td>
<td>unbalance</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>misalignment</td>
<td>coupling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bearing failure</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bent shaft</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>overload</td>
<td>interference</td>
<td>incorrect assembly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bearing failure</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operational</td>
<td>?</td>
</tr>
<tr>
<td><strong>Gear fracture</strong></td>
<td>fatigue</td>
<td>life limit exceeded</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>surface damage</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>resonance</td>
<td>design</td>
<td>?</td>
</tr>
<tr>
<td><strong>Tooth fracture</strong></td>
<td>bending fatigue</td>
<td>life limit exceeded</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>surface damage</td>
<td>process related</td>
</tr>
<tr>
<td></td>
<td></td>
<td>thin tooth</td>
<td>excessive wear</td>
</tr>
<tr>
<td></td>
<td>random fracture</td>
<td>surface damage</td>
<td>process related</td>
</tr>
<tr>
<td></td>
<td></td>
<td>foreign object</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pitting/spalling</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>overload</td>
<td>interference</td>
<td>incorrect assembly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bearing failure</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operational</td>
<td>?</td>
</tr>
</tbody>
</table>

*Table 3.3 Diagnostic evidence of safety critical failure modes*

The column headed V indicates vibratory evidence, O indicates oil debris, T indicates temperature and U indicates information from Usage Monitoring. Tick marks indicate that the specified diagnostic evidence is expected for the failure mode, a cross indicates that the diagnostic evidence is not expected and a question mark indicates possible diagnostic evidence.

### 3.2.4 Priorities for the application of vibration analysis

The areas in which maximum benefit can be derived from the application of vibration analysis are those which are (a) safety critical, (b) not adequately covered by other fault detection methods, and (c) perceived to produce distinguishable vibration signatures.
Based on the information in Table 3.3, the failure modes which are expected to provide good vibratory evidence are fatigue fractures of shafts, gears and gear teeth, and random fractures of gear teeth. Because of the sudden, unexpected nature of overload fractures, these cannot be predicted, however, if total failure does not occur immediately, the initial fracture may progress in a similar fashion to a fatigue fracture.

To provide sufficient warning of impending failure and, where practical, enable preventative maintenance to be carried out, it is desirable that the causal factors leading to the safety critical failures also be detected. These are of secondary importance to the diagnosis of the fractures themselves. Therefore, in order of importance, the failures for which vibration analysis methods are to be examined in more detail are:

a) shaft fatigue fracture,

b) gear tooth bending fatigue fracture,

c) gear fatigue fracture,

d) gear tooth random fracture,

e) shaft unbalance, misalignment and bent shaft, and

f) gear tooth surface damage.

In addition to the priority failure modes listed above, vibration analysis may be useful in providing confirmatory detection and diagnostics of potentially safety critical failure modes for which other failure evidence (e.g., oil debris and/or temperature) is available. These failure modes are:

a) excessive tooth wear and destructive scoring,

b) tooth pitting/spalling, and

c) bearing failures.
3.3 EXPECTED VIBRATION SIGNATURES

In this section, a brief description of the expected vibration signature for each of the selected failure modes is given. This will not only aid in the review of vibration analysis techniques given in the next chapter, but will also give some idea of the complexity of the vibration analysis task for each of the failure modes and, as a consequence, the level of effort required to develop and test reliable diagnostic systems based on vibration analysis.

3.3.1 Shaft fatigue fracture

A description of the vibration expected from a cracked (fractured) shaft was given in Chapter 2 (Section 2.1.2.3). This showed that cracks only have a significant effect on the vibration levels at one, two and three times shaft rotational frequency, with all three increasing equally until the crack reaches half radius, from which point the amplitude at the shaft rotational frequency increases at a faster rate. This makes detection and diagnosis of shaft cracks a relatively simple tasks using vibration analysis if the first three harmonics of the shaft vibration $x_s(t)$ (2.26) can be clearly identified in the measured vibration signal $x(t)$ (2.33).

3.3.2 Gear tooth bending fatigue fracture

A crack in a gear tooth will reduce the bending stiffness of the tooth, leading to greater deflection for a given load (Mcfadden [56]). This will affect both the amplitude and the phase of the tooth-meshing vibration over the period in which the tooth is engaged. As the crack progresses, it would be assumed that the amplitude and/or phase deviations would increase and, in advanced stages of cracking, disturbance of the engagement of subsequent teeth may occur, causing additive impulses due to impacts.

The short term amplitude and phase deviations will be reflected in the vibration amplitude and phase modulation effects, $\alpha_{sg}(\theta)$ and $\beta_{sg}(\theta)$, given in equation (2.6). The expected form of any additive impulse is given in equation (2.7).
For diagnostic purposes, it is important that a bending fatigue crack be distinguished from other localised tooth failures such as pitting, spalling or tooth surface damage, which progress at a far lower rate and will not necessarily lead to catastrophic failure. The major distinguishing characteristic between tooth cracking and other localised tooth damage is the extent of tooth deflection, which will be more noticeable in the phase modulation $\beta_{sg}(\theta)$ (2.6).

### 3.3.3 Gear fatigue fracture

The initiation of cracks in gears and in gear teeth are very similar (Appendix A.1.6.4) with the major difference being the crack progression. Therefore, the vibration signature for gear cracks would be expected to be very similar to that previously described for tooth fracture. In the advanced stages, gear cracking would be expected to affect the meshing of a number of teeth, as the crack remains open longer, and to have a more pronounced phase change than for a single tooth crack; the phase change being due to an increase in tooth spacing between the adjacent teeth as the crack opens rather than the deflection of a single tooth due to bending.

The distinction between gear and gear tooth cracking is not as important as that between cracking and other localised faults. Both fracture modes propagate rapidly to catastrophic failure, therefore similar preventative action is indicated.

### 3.3.4 Gear tooth random fracture

If the initiation site of a random fracture is near the root of the tooth, it would be expected to show similar propagation and vibratory evidence to tooth bending fatigue fracture; as the preventative action required would also be similar, the distinction between the two is not important in this case.

If the crack initiation site is close to the tip of the tooth, very little reduction in tooth stiffness may be evident and crack propagation may continue (at a relatively low rate due to the small bending moment at the crack location) without significant vibratory evidence until a ‘chunk’ of the tooth separates. At this stage, it would be expected that large
impulses will occur as the sharp edges of the broken tooth impact with the mating teeth; this may be accompanied by abrupt phase changes. Significant cutting damage may occur to the mating teeth, leading to a succession of random fractures on both gears. In addition to the vibratory evidence, the separation of relatively large chunks of gear tooth should be readily detectable by the ‘chip detectors’ used in most modern helicopter transmissions.

3.3.5 Shaft unbalance, misalignment and bent shaft

The expected vibration signature for these faults, which may contribute to shaft fatigue fracture, are described in Chapter 2, Section 2.1.2. As with shaft fatigue fracture itself, diagnosis of these faults is relatively straightforward if the vibration amplitudes at the lower harmonics of shaft rotation frequency can be clearly identified.

3.3.6 Gear tooth surface damage

Tooth surface damage, which may be a contributing factor to tooth fracture, would be expected to cause a localised change in the amplitude of the tooth-meshing vibration, which should be evident in the amplitude modulation effects, $\alpha_{sg}(\theta)$ (2.6). If the surface damage is sufficiently large, some additive impulsive vibration may also be evident. However, as the surface damage does not affect the tooth stiffness, little or no phase modulation should be evident; this provides a distinguishing feature between surface damage and tooth cracks.

3.3.7 Excessive tooth wear and destructive scoring

These faults, which may eventually lead to tooth fatigue fracture, both progress at a relatively low rate and produce wear debris which should be detectable by oil debris analysis. Similar vibratory evidence would be expected for both these failure modes; they both produce systematic destruction of the tooth profile, with the extent of material removed being proportional to the distance from the pitch line (Drago [28], Randall [65], and Appendix A.1.1.3 and A.1.2.3). This will result in a change in the vibration
amplitudes at the tooth-meshing frequency and its harmonics, with a more significant change expected in the higher harmonics (see Chapter 2, Section 2.1.1.1.4). Although vibration analysis cannot provide distinction between these two failure modes, oil debris analysis can (Kuhnell [44]).

### 3.3.8 Tooth pitting/spalling

Like excessive tooth wear and destructive scoring, these failures can eventually contribute to tooth fracture (usually random fracture due to the pits/spalls acting as crack initiation sites). The expected vibration signatures for these are the same as for gear tooth surface damage; as the consequences are the same, distinction between these modes is not of any great significance. Oil debris monitoring may be used to distinguish between pitting/spalling and tooth surface damage, as the later does not generate any debris.

### 3.3.9 Bearing failures

A large number of bearing failure modes can occur and the vibration signatures produced are complex and of relatively low energy. It is debatable whether distinction between the various bearing failure modes is of any significant benefit. Wear, scoring and surface fatigue on any of the constituent components all lead to progressive degradation of all the bearing elements, producing material debris and friction generated heat.

Because of the above factors, and the low perceived safety impact of bearing failures (except as a contributory factor), very little emphasis will be placed on vibration analysis of bearing faults in the following chapters.

In the small number of cases where bearings are a safety critical item (such as in the tail rotor pitch control), direct monitoring using debris analysis, temperature monitoring, relative simple vibration analysis or a combination of these, would provide a far more reliable solution than employing sophisticated vibration analysis techniques via remote sensors.
Chapter 4

REVIEW OF VIBRATION ANALYSIS TECHNIQUES

In this chapter, a review is made of some current vibration analysis techniques used for condition monitoring in geared transmission systems. The perceived advantages, disadvantages, and the role each of these techniques may play in the diagnosis of safety critical failure modes is discussed. A summary of the findings is then made to establish which techniques to pursue further, and to identify any deficiencies which need to be addressed.

4.1 TIME DOMAIN ANALYSIS

4.1.1 Waveform analysis

Prior to the commercial availability of spectral analysers, almost all vibration analysis was performed in the time domain. By studying the time domain waveform using equipment such as oscilloscopes, oscillographs, or ‘vibrographs’, it was often possible to detect changes in the vibration signature caused by faults. However, diagnosis of faults was a difficult task; relating a change to a particular component required the manual calculation of the repetition frequency based on the time difference observed between feature points.

4.1.2 Time domain signal metrics

Although detailed study of the time domain waveform is not generally used today, a number of simple signal metrics based on the time domain waveform still have widespread application in mechanical fault detection; the simplest of these being the peak and RMS value of the signal which are used for overall vibration level measurements.
### 4.1.2.1 Peak

The peak level of the signal is defined simply as half the difference between the maximum and minimum vibration levels:

\[
\text{peak} = \frac{1}{2} \left( \max(x(t)) - \min(x(t)) \right)
\]  

(4.1)

### 4.1.2.2 RMS

The RMS (Root Mean Square) value of the signal is the normalised second statistical moment of the signal (standard deviation):

\[
\text{RMS} = \sqrt{\frac{1}{T} \int_0^T (x(t) - \bar{x})^2 \, dt}
\]  

(4.2)

where \( T \) is the length of the time record used for the RMS calculation and \( \bar{x} \) is the mean value of the signal:

\[
\bar{x} = \frac{1}{T} \int_0^T x(t) \, dt
\]  

(4.3)

For discrete (sampled) signals, the RMS of the signal is defined as:

\[
\text{RMS} = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} (x(n) - \bar{x})^2}
\]

\[
\bar{x} = \frac{1}{N} \sum_{n=0}^{N-1} x(n)
\]  

(4.4)

The RMS of the signal is commonly used to describe the ‘steady-state’ or ‘continuous’ amplitude of a time varying signal.

### 4.1.2.3 Crest Factor

The crest factor is defined as the ratio of the peak value to the RMS of the signal:

\[
\text{Crest Factor} = \frac{\text{peak}}{\text{RMS}}
\]  

(4.5)
The crest factor is often used as a measure of the ‘spikiness’ or impulsive nature of a signal. It will increase in the presence of discrete impulses which are larger in amplitude than the background signal but which do not occur frequently enough to significantly increase the RMS level of the signal.

### 4.1.2.4 Kurtosis

Kurtosis is the normalised fourth statistical moment of the signal. For continuous time signals this is defined as:

\[
Kurtosis = \frac{1}{T} \int_0^T \left( x(t) - \bar{x} \right)^4 \, dt \quad \text{(4.6)}
\]

For discrete signals the kurtosis is:

\[
Kurtosis = \frac{1}{N} \sum_{n=0}^{N-1} \left( x(n) - \bar{x} \right)^4 \quad \text{(4.7)}
\]

The kurtosis level of a signal is used in a similar fashion to the crest factor, that is to provide a measure of the impulsive nature of the signal. Raising the signal to the fourth power effectively amplifies isolated peaks in the signal.

### 4.1.3 Overall vibration level

The most basic vibration monitoring technique is to measure the overall vibration level over a broad band of frequencies. The measured vibration level is trended against time as an indicator of deteriorating machine condition and/or compared against published vibration criteria for exceedences. The measurements are typically peak (4.1) or RMS (4.2) velocity recordings which can be easily made using a velocity transducer (or integrating accelerometer) and an RMS meter.

Because the peak level is not a statistical value, it is often not a reliable indicator of damage; spurious data caused by statistically insignificant noise may have a significant
effect on the peak level. Because of this, the RMS level is generally preferred to the peak level in machine condition monitoring applications.

Trending of overall vibration level may indicate deteriorating condition in a simple machine, however it provides no diagnostic information and will not detect faults until they cause a significant increase in the overall vibration level. Localised faults in complex machinery may go undetected until significant secondary damage or catastrophic failure occurs.

4.1.4 Waveshape metrics

The overall vibration level provides no information on the wave form of the vibration signal. With a number of fault types, the shape of the signal is a better indicator of damage than the overall vibration level. For example, faults which produce short term impulses such as bearing faults and localised tooth faults, may not significantly alter the overall vibration level but may cause a statistically significant change in the shape of the signal.

Crest factor (4.5) or kurtosis (4.6) are often used as non-dimensional measures of the shape of the signal waveform. Both signal metrics increase in value as the ‘spikiness’ of the signal increases (i.e., as the signal changes from a regular continuous pattern to one containing isolated peaks). Kurtosis, being a purely statistical parameter, is usually preferable to crest factor in machine condition monitoring applications for the same reasons that RMS is preferable to peak. However, crest factor is in more widespread use because meters which record crest factor are more common (and more affordable) than kurtosis meters.

Because of the non-dimensional nature of the crest factor and kurtosis values, some assessment of the nature of a signal can be made without trend information. Both waveshape metrics will give a value of 0.0 for a DC signal and 1.0 for a square wave. For a pure sine wave, the crest factor will be $\sqrt{2} \approx 1.414$ and the kurtosis will be 1.5. For normally distributed random noise, the kurtosis will be 3.0 and the crest factor will
be approximately 3 (note that because the crest factor is not a statistical measure, its value in the presence of random noise will vary).

Trending of the waveshape metrics can also be used to help identify deteriorating condition. However, the trend of these values may be misleading in some cases; faults which produce a small number of isolated peaks (such as the initial stages of bearing damage) may cause an increase in the crest factor and kurtosis but, as the damage becomes more widely spread, a large number of impulses may occur causing both the crest factor and kurtosis to decrease again. Both the kurtosis and crest factor will decrease if the number of pulses increase (increasing the RMS value of the signal) without an increase in the individual pulse height.

As with the overall vibration level, the waveshape metrics will not detect faults unless the amplitude of the vibration from the faulty component is large enough to cause a significant change in the total vibration signal. This limits their use to components whose vibration signature forms a significant portion of the measured overall vibration.

### 4.1.5 Frequency band analysis

Often, the fault detection capability using overall vibration level and/or waveshape metrics can be significantly improve by dividing the vibration signal into a number of frequency bands prior to analysis. This can be done with a simple analogue band-pass filter between the vibration sensor and the measurement device. The rationale behind the use of band-pass filtering is that, even though a fault may not cause a significant change in overall vibration signal (due to masking by higher energy, non-fault related vibrations), it may produce a significant change in a band of frequencies in which the non-fault related vibrations are sufficiently small. For a simple gearbox, with judicious selection of frequency bands, one frequency band may be dominated by shaft vibrations, another by gear tooth-meshing vibrations, and another by excited structural resonances; providing relatively good coverage of all gearbox components.
4.1.6 Advantages

Meters for recording overall vibration levels, crest factor and/or kurtosis are readily available, relatively cheap and simple to use. Because of this, they can be a very cost effective method of monitoring simple machine components which are relatively cheap and easily replaceable but perform a critical role (for example small pumps and generators). The time domain signal metrics may detect the imminent failure of these components allowing replacement prior to total failure; although the damaged component may be beyond repair by this time, the component replacement cost is generally insignificant compared to the potential cost of catastrophic failure (secondary damage, loss of utility, etc.).

4.1.7 Disadvantages

For more complex or costly machines, it is generally preferable to detect damage at an early stage to allow the machine to be repaired rather than replaced. This requires techniques which are more sensitive to changes in the vibrations of individual components and which can provide at least some diagnostic capabilities.

4.1.8 Applicability to safety critical failure modes

Simple time domain signal metrics, even with the use of band pass filtering, do not provide any diagnostic information and, therefore, cannot be used to distinguish any of the safety critical failure modes from other failure modes.

For very simple safety critical systems, overall vibration level and/or kurtosis level (in combination with oil debris and/or temperature monitoring) may be useful as part of a cost effective failure detection system.

4.2 SPECTRAL ANALYSIS

Spectral (or frequency) analysis is a term used to describe the analysis of the frequency domain representation of a signal. Spectral analysis is the most commonly used vibration
analysis technique for condition monitoring in geared transmission systems and has
proved a valuable tool for detection and basic diagnosis of faults in simple rotating
machinery [65,67]. Whereas the overall vibration level is a measure of the vibration
produced over a broad band of frequencies, the spectrum is a measure of the vibrations
over a large number of discrete contiguous narrow frequency bands.

The fundamental process common to all spectral analysis techniques is the conversion of
a time domain representation of the vibration signal into a frequency domain
representation. This can be achieved by the use of narrow band filters or, more
commonly in recent times, using the discrete Fourier Transform (DFT) of digitised data.
The vibration level at each ‘frequency’ represents the vibration over a narrow frequency
band centred at the designated ‘frequency’, with a bandwidth determined by the
conversion process employed.

For machines operating at a known constant speed, the frequencies of the vibrations
produced by the various machine components can be estimated (as per the model
described in Chapter 2) therefore, a change in vibration level within a particular
frequency band can usually be associated with a particular machine component. Analysis
of the relative vibration levels at different frequency bands can often give an indication of
the nature of a fault, providing some diagnostic capabilities.

4.2.1 Conversion to the frequency domain

The frequency domain representation of a signal can be described by the Fourier
Transform [67] of its time domain representation

\[ X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt . \]  \hspace{1cm} (4.8)

The inverse process (Inverse Fourier Transform [67]) can be used to convert from a
frequency domain representation to the time domain

\[ x(t) = \int_{-\infty}^{\infty} X(f) e^{j2\pi ft} df . \]  \hspace{1cm} (4.9)
There are a number of limitations inherent in the process of converting vibration data from the time domain to the frequency domain.

### 4.2.1.1 Bandwidth-time limitation

All frequency analysis is subject to a bandwidth-time limitation (often called the Uncertainty Principle [27,67] due to the analogous concepts in quantum mechanics, enunciated by Werner Heisenberg in 1927).

Frequency analysis made with bandwidth of $B$ hertz for each measurement and a duration in time of $T$ seconds has a bandwidth-time limitation of:

$$BT \geq 1$$  \hspace{1cm} (4.10)

If an event lasts for $T$ seconds, the best measurement bandwidth (the minimum resolvable frequency) which can be achieved is $1/T$ hertz. If an analysing filter with a bandwidth of $B$ hertz is used, at least $1/B$ seconds will be required for a measurement.

The measurement uncertainty due to the bandwidth-time limitation imposes a resolution restriction on the frequency conversion. To resolve frequencies separated by $B$ hertz at least $1/B$ seconds of data must be taken.

### 4.2.1.2 FFT Analysers

Most modern spectrum analysers use the Fast Fourier Transform (FFT) [25], which is an efficient algorithm for performing a Discrete Fourier Transform (DFT) [61,67] of discrete sampled data.

The Discrete Fourier Transform is defined as [61]

$$X(m) = \frac{1}{N} \sum_{n=0}^{N-1} x(n)e^{-j2\pi \frac{mn}{N}},$$ \hspace{1cm} (4.11)

and the Inverse Discrete Fourier Transform [61] is
\[ x(n) = \sum_{m=0}^{N-1} X(m)e^{j\frac{2\pi mn}{N}}. \]  

(4.12)

The sampling process used to convert the continuous time signal into a discrete signal can cause some undesirable effects.

### 4.2.1.2.1 Aliasing

Frequencies which are greater than half the sampling rate will be aliased to lower frequencies due to the stroboscope effect. To avoid aliasing, an analogue low-pass ‘anti-aliasing’ filter is used prior to sampling to ensure that there are no frequencies above half the sampling rate.

### 4.2.1.2.2 Leakage

When applying the FFT, it is assumed that the sampled data is periodic with the time record. If this is not the case, spurious results can arise from discontinuities between the start and end points of the time record. This ‘leakage’ is normally compensated for by applying a smooth window function which has zero values at the start and end of the time record. This entails a resolution trade-off since it effectively reduces the time duration of the signal. For a simple machine, the time record can be synchronised with the rotation of the machine, ensuring that the major vibration components are periodic within the time record; this is difficult to achieve with complex machines due to the large number of non-harmonically related frequencies.

### 4.2.1.2.3 Picket Fence Effect

The picket fence effect is a result of the discrete frequency nature of the FFT. Where a frequency does not lie on one of the discrete frequency lines, the amplitude will be reduced. If the frequency is well separated from other frequency components, a correction can be made by curve fitting the samples around the peak. Windowing reduces the effect due to the increase in bandwidth caused by the windowing process.
4.2.1.3 Speed variations

The ability to resolve frequency components is not only related to the bandwidth-time limitation but also to the stability of the vibration signal over the analysis period. For FFT analysers, the resolution imposed by the bandwidth-time limitation is constant for all frequencies, however, the frequency of vibration signals due to the mechanically linked rotating components in a geared transmission system will vary proportionally with variations in the rotational speed of the machine, imposing a resolution limitation which is a constant percentage of the frequency.

Even with ‘constant’ speed machines, some drift in operating speed over time is likely to occur and, in some cases, may cause frequency variations (and uncertainty) which are greater than those due to the bandwidth-time limitations. For example, performing an FFT on one second of data would give a spectrum with a resolution of one hertz, however, a one percent speed variation over the analysis period would cause a 5 hertz uncertainty at a frequency of 500 hertz.

4.2.1.3.1 Synchronous sampling

The effects of speed variations can be overcome to a certain extent by the use of ‘synchronous sampling’, in which the sampling rate of the analyser is linked to the speed of the machine. However, this adds further complication to the monitoring process as it requires a speed sensor attached to the machine being monitored, a frequency multiplier to convert the speed sensor signal into a ‘clock pulse’ signal suitable for driving the signal analyser, and often needs an external anti-aliasing filter to avoid aliasing problems (although almost all modern FFT analysers have in-built anti-aliasing filters, when they are driven from an ‘external clock’ these are often bypassed or have inappropriate frequencies due to the unknown external clock frequency).
4.2.2 Fault detection

4.2.2.1 Spectral comparison

The most common spectral analysis technique employed for machine condition monitoring is spectral comparison, where a baseline power (magnitude squared) spectrum is taken under well defined normal operating conditions with the machine in known good condition (preferably soon after commissioning). This ‘baseline’ spectrum is used as a reference for subsequent power spectra taken at regular intervals throughout the machine life under similar operating conditions (Mathew [47]). The comparison is usually done on a logarithmic amplitude scale, with increases of 6-8 dB considered to be significant and changes greater than 20dB from the baseline considered serious (Randall [66]).

4.2.2.2 Spectral trending

In addition to spectral comparison, various forms of spectral trending [47,59,66] can be used to give some indication of the rate of fault progression. In its simplest form, spectral trending involves the trending of the changes in amplitude of all (or a number of selected) spectral lines over time. For complex machines, this can often involve a large amount of data, resulting in information overload due to the large number of significant spectral lines [47]. In an attempt to simplify the detection process, a number of parameters based on the spectrum have been proposed which provide statistical measurements of spectral differences. Mechefske and Mathew [59] give an overview of spectral parameters and a comparison of their detection and diagnostic performance for a number of bearing faults. They found that a number of these parameters performed well in the detection of the faults but that none of the parameters provided significant diagnostic information.

4.2.2.3 Spectral masks

Spectral masks are a method of spectral comparison sometimes employed to identify and evaluate changes in the signature spectrum [47,66], with allowances made for variation in operating condition. A spectral mask is derived from the baseline spectrum by adding
an allowable tolerance limit to the logarithmic amplitude. To allow for variations in speed, the constant bandwidth spectrum is sometimes converted to a constant percentage bandwidth spectrum [66], with the percentage bandwidth being determined by the estimated speed differences which can occur between recordings (note that this is different to, and would be expected to be much larger than, the speed variations during the recording time described in Section 4.2.1.3).

Once a spectral mask is defined, comparison of individual recordings is made with reference to the mask to identify exceedences.

### 4.2.3 Fault diagnosis

Even for relatively simple machines, the vibration spectrum can be quite complex due to the multiple harmonic structures of the vibration from various components in combination with the transmission path effects (as detailed in Chapter 2). This makes detailed diagnostic analysis of an individual spectrum very difficult. The diagnostic process is simplified when performed in conjunction with spectral comparison and/or trending; typically, only the frequencies identified as having significant changes are analysed in detail for diagnostic purposes.

Randall [65] provides details of the expected spectral differences associated with various gear faults, and Su and Lin [74] provide similar information for bearing faults.

Distributed faults which cause significant change in the mean amplitude of the vibration at discrete frequencies, such as heavy wear and unbalance, should be relatively simple to diagnose using spectral analysis, as they would simply translate to changes in a few associated frequency lines in the spectrum.

Faults which cause low frequency sinusoidal modulations, such as an eccentric or misaligned gear, may also be diagnosed as they will translate to increases in the sidebands surrounding the tooth meshing frequency and harmonics.

Very localised faults, such as tooth cracking or spalling, are not easily diagnosed (and may not even be detected) as the short term impulsive vibrations produced translate to a
large number of low amplitude lines in the spectrum (McFadden [56] and Randall [65]). An example where the diagnostic information available using spectral analysis was not sufficient to detect a fatigue crack in a gear (resulting in a fatal helicopter accident) is provided by McFadden [54].

4.2.4 Advantages

A number of companies manufacture and/or supply high quality FFT analysers at a reasonable price. In addition to marketing analysers, several of these companies also provide comprehensive after sales support in the form of literature and training in diagnostic methods using their equipment.

Because of the fairly widespread use of spectral analysis over a number of years, there is a fairly comprehensive collection of literature on its use for machine fault diagnosis (e.g., Randall [67] and Braun [15]).

4.2.5 Disadvantages

The major disadvantage with spectral analysis lies in its complexity. Even with the amount of literature available, specialist skills are still required to exploit the diagnostic capabilities of spectral analysis. When dealing with complex machines or with localised faults such as gear tooth faults, even expert analysts find diagnosis difficult.

4.2.6 Applicability to safety critical failure modes

For relatively simple machines, and those where the first few harmonics of the shaft vibration frequencies can be clearly identified (i.e., can be well separated from other vibration frequencies within the limits of bandwidth and/or speed variations), diagnosis of shaft related faults (fracture, unbalance, misalignment and bent shaft) should be quite simple with spectral analysis, by trending of the amplitudes of the shaft related vibrations or use of spectral masks.
The other safety critical faults identified in the previous chapter all produce impulsive signals and, as was demonstrated by McFadden [54], these faults cannot be reliably diagnosed (or, in some cases, even detected) using spectral analysis.

4.3 SYNCHRONOUS SIGNAL AVERAGING

Stewart [73] showed that with ‘time synchronous averaging’ the complex time-domain vibration signal from a machine such as a helicopter transmission could be reduced to estimates of the vibration for individual shafts and their associated gears. The synchronous average for a shaft is then treated as if it were a time domain vibration signal for one revolution of an individual, isolated shaft with attached gears.

4.3.1 Fundamental principle

The fundamental principle behind synchronous signal averaging is that all vibration related to a shaft, and the gears on that shaft, will repeat periodically with the shaft rotation (see Chapter 2, Sections 2.2.1 and 2.2.2). By dividing the vibration signal into contiguous segments, each being exactly one shaft period in length, and ensemble averaging a sufficiently large number of segments, the vibration which is periodic with shaft rotation will be reinforced and vibrations which are not periodic with the shaft rotation will tend to cancel out; leaving a signal which represents the average vibration for one revolution of the shaft. Figure 4.1 illustrates how this process might be performed on a continuous time signal from a gearbox, using a tacho multiplier to calculate each rotational period of the shaft.
4.3.2 Synchronous signal averaging of discrete signals

The process illustrated in Figure 4.1 assumes the vibration signal being averaged is a continuous time signal. In practice, the signal averaging process usually takes place on a discrete sampled signal (e.g., via an analogue-to-digital converter in a PC) and, in addition to defining the start and end points of the shaft rotation, some mechanism is needed to ensure that the sample points are at equally spaced angular increments of the shaft and that these are at the same angular position for each revolution of the shaft. That is, the sampling must be coherent with the rotation of the shaft. Originally, Stewart [73] used a phase-locked frequency multiplier however, McFadden [58] showed that far greater accuracy and flexibility could be achieved using digital resampling of time sampled vibration based on a reference derived from a simultaneously time sampled tacho signal. This method will be expanded upon in later chapters.
4.3.3 Terminology

It should be noted that the terminology related to this process is somewhat confused. The same process has been referred to as ‘time domain averaging’ (McFadden [53] and Braun [14]), ‘time synchronous averaging’ (Stewart [73]), ‘coherent rotational signal averaging’ (Swansson, et al [78]) and ‘synchronous averaging’ (Succi [75]).

The principle of synchronising the averaging with some other process (in this case the rotational frequency of a shaft) is fundamental to the technique; whether it is performed on continuous or discrete signals. As was seen above, when the process is performed on discrete signals the sampling must be coherent with the rotation of the shaft (hence the term ‘coherent rotational signal averaging’ used by Swansson, et al). Note that the process can be performed in the time or frequency domain (as long as the frequency domain averaging is performed on the complex frequency domain representation). The term ‘time domain averaging’ ([53] and [14]) was used to distinguish the technique from that of averaging of amplitude or power spectra to reduce variance in spectral analysis (Randall [67]).

To properly describe the process when applied to discrete signals, it should probably be referred to as ‘rotationally coherent synchronous signal averaging’. However, this is quite clumsy and therefore the technique will normally be referred to as ‘synchronous signal averaging’ in the remainder of this thesis (the rotational coherency being implied when the technique is applied to discrete data).

4.3.4 Angle domain and shaft orders

Because the synchronous signal average is based on the rotation of the shaft rather than time, it is no longer correct to refer to it as a ‘time’ domain signal (although this is often done). Throughout the remainder of this thesis, a signal resulting from a synchronous signal averaging process (or any other angular based process) will be said to be in the angle domain. The angle domain is expressed in radians (or degrees) of revolution of the shaft (2\(\pi\) radians = 360 degrees = 1 revolution).
In the spectrum (Fourier transform) of an angle domain signal the frequency is expressed in **shaft orders** rather than Hertz (or RPM), where 1 shaft order = 1 cycle per revolution of the shaft.

### 4.3.5 Signal enhancements

In addition to treatment of a synchronous signal average as a simplified vibration signal (i.e., time domain analysis and spectral analysis can be applied), its periodic nature allows far more scope for signal manipulation than does a conventional vibration signal. The Fourier transform of a periodic signal is a pure line spectrum not subject to leakage (Section 4.2.1.2.2) or the picket fence effect (Section 4.2.1.2.3), and for which ideal filtering can be used; that is, one or more frequency lines (here representing shaft orders) can be completely removed from the spectrum without causing discontinuities when the signal is translated back to the angle domain.

This allows various signal enhancement techniques, specifically designed for the treatment of synchronous signal averages, and a number of related signal metrics to be used as an aid to fault detection and diagnosis.

#### 4.3.5.1 Stewart’s Figures of Merit

Stewart developed a number of non-dimensional parameters based on the synchronous signal average, which he termed ‘Figures of Merit’ [73]. These were originally defined as a hierarchical group, with which Stewart described a procedure for the detection and partial diagnosis of faults.

##### 4.3.5.1.1 FM0

The zero order figure of merit, FM0, was proposed as a general purposes fault detector to be applied to all signal averages. It is calculated simply by dividing the peak-to-peak value of the angle domain signal by the sum of the amplitudes of the tooth-meshing frequencies and harmonics. In simple terms, FM0 is a relative measure of the peak deviation of the signal from that defined purely by the mean gear tooth meshing vibration. Stewart claimed this to have good detection capabilities for
a) localised faults (such as tooth breakage, pitting, and spalling), bearing instability and misalignment as these increased the peak-to-peak level with little change in tooth-meshing level, and

b) heavy wear, as this produces little change in the peak-to-peak level with a reduction in tooth-meshing level.

4.3.5.1.2 Mesh-specific figures of merit (FM1, FM2 and FM3)

In the cases where FM0 indicated a significant change in the average, higher order ‘mesh-specific’ figures of merit (FM1, FM2 and FM3) were used to detect various patterns in the signal average indicative of certain types of faults.

FM1 is the relative measure of the low frequency (first and second order) modulation to the tooth-meshing amplitudes. This can be calculated for individual tooth-meshes (by dividing the sum of the two upper and lower sideband amplitudes by the tooth-meshing amplitude) or for all tooth-mesh related vibration (by dividing the sum of the two upper and lower sidebands of all tooth-mesh related frequencies by the sum of the amplitudes of the tooth-mesh related frequencies). FM1 will respond to misalignment, eccentricity, swash or shaft failures.

FM2 was designed specifically to detect single tooth damage, such as fracture or chipping, in multi-mesh gears. This is done by matching a pattern, consisting of spikes spaced at the mesh angular separation, with the envelope of the signal average. The matching is done using a circular matched filter (cross-correlation) of the pattern and the envelope. The ratio of the kurtosis (fourth statistical moment) of the matched filter output to the kurtosis of the envelope is used as the detection parameter. If there are impulses in the signal which are correlated to the angular separation between the mesh points, the matched filter operation emphasises these, forcing the kurtosis of the matched filter output above that of the envelope (giving FM2 > 1). Where no significant correlation of impulses to angular separation between the mesh points exists, the matched filter output tends to be flat with a kurtosis less than that of the envelope (giving FM2 < 1).
FM3 was designed to detect sub-harmonic meshing (i.e., significant sinusoidal vibration at a lower frequency than that of the tooth-meshing) and was claimed to be an indicator of parametric excitation and heavy wear. The signal average is low-pass filtered to include only frequencies up to and including the highest tooth-meshing fundamental frequency (e.g., filter set to 1.1 times the highest tooth-meshing fundamental). FM3 is the ratio of the zero-crossing count of a signal consisting only of the mesh related frequencies to the zero-crossing count of the low-pass filtered signal. Where there is strong sub-harmonic activity, the zero-crossing count of the filtered signal will be lower than that of the mesh related frequencies, causing an increase in the FM3 value.

4.3.5.1.3 FM4

If the change in the signal average could not be attributed to any of the predefined mesh-specific faults, FM4 (or ‘bootstrap reconstruction’) was used to detect other mesh related faults.

Stewart [73] reasoned that if one could define the expected frequency content for the vibration from a particular shaft (the regular signal), then all other vibration represents the deviation (the residual signal) from the expected signal. He proposed that the regular signal would normally include all tooth-meshing frequencies and their harmonics, plus their immediately adjacent sidebands. The residual signal is simply calculated by converting to the frequency domain, eliminating all components defined by the regular signal, and converting back to the angle domain.

Two parameters were proposed based on the residual signal:

a) FM4A: the kurtosis (4.6) of the residual, which will respond to impulsive signals such as those produced by pitting, spalling and tooth cracks.

b) FM4B: the ratio of the RMS of the residual to that of the original signal, which will respond to distributed faults which cause an increase in non-mesh related vibration.

The inclusion of the upper and lower sidebands in the regular signal meant that faults causing once per revolution modulations would not be detected using FM4. However,
these are covered by FM1 and Stewart found that the removal of the sidebands from the residual increased sensitivity to other faults.

Experimental investigation by Stewart [73] showed that FM0 was not a very sensitive detector of localised tooth faults, such as tooth cracking, and he suggested that FM4 be used as a more sensitive general detector.

Because of the sensitivity of FM4 as a general fault detector, and its ease of implementation in comparison to the more specific fault detectors (FM1, FM2 and FM3), a number of experimental investigations based on Stewart’s work have concentrated on enhancement techniques related to FM4 (e.g., McFadden [54] and Zakrajsek [84]).

### 4.3.5.2 Trend analysis

Stewart [73] also proposed a method of trending signal averages by calculating and trending the kurtosis and RMS of the difference between a baseline signal average and subsequent signal averages (from data taken from the same transducer location under the same operating conditions). The signal averages at different times need to be aligned in the angle domain before the difference signal is calculated. In cases were an absolute position reference is not available, the angular alignment is done by using a circular matched filter (cross-correlation) of the two signals; the location of the maximum correlation defining the angular offset between the two signals.

The kurtosis and RMS of the difference signal are interpreted in a similar fashion to FM4A and FM4B respectively.

### 4.3.5.3 Narrow-band Envelope Analysis

In the process of examining an in-service gear fatigue failure, McFadden [54] recognised the importance of phase modulation in the diagnosis of cracks and, based on this, developed a method of including phase information in a signal enhancement technique.

Assuming that the sideband structure surrounding a strong tooth-meshing harmonic was predominantly due to the modulation of the harmonic, McFadden [54] proposed that by narrow bandpass filtering about the selected tooth-meshing harmonic, removing the
tooth-meshing harmonic, and calculating the angle domain envelope, a signal could be obtained which contains contributions from both the amplitude and phase modulations. The kurtosis of this signal was used as a measure of localised damage.

McFadden [54] showed that, in the early stages of cracking, this parameter had a better response than Stewart’s FM4A.

4.3.5.4 Demodulation

McFadden [56] further developed the technique based on narrow bandpass filtering by using demodulation to extract an estimate of both the amplitude and phase modulation signals. He showed that a cracked tooth displays an amplitude drop with simultaneous phase change as the tooth comes in to contact. By displaying the amplitude and phase modulations simultaneously as a polar plot, these characteristic modulations could be seen as loops.

This particular development is of great significance, as it allows a distinction between tooth cracking and other localised tooth faults, such as pitting or spalling, to be made.

4.3.6 Advantages

The main advantages of synchronous signal averaging is that it allows a complex vibration signal to be reduced to a number of much simpler signals, each of which is an estimate of the vibration from a single shaft and its associated gears. The resultant signals are purely periodic and can be enhanced using ideal filtering.

The signal parameters developed by Stewart [73] and McFadden [54,56] further simplify the analysis task.

4.3.7 Disadvantages

Very few pieces of equipment are currently available which accurately implement synchronous signal averaging, and the ones that are available are very expensive. A number of analysers have ‘time synchronous averaging’ capabilities however, these
generally have the capability of synchronising the start of each ensemble, but no method of ensuring coherent rotational sampling.

Most research equipment implementing synchronous signal averaging, such as the one used in this research project, are constructed and programmed by the researchers themselves based on PC’s, analogue-to-digital converters and anti-aliasing filters. Accurate tacho signals and phase-locked frequency multipliers or digital interpolation techniques are also needed.

Further work needs to be done on methodologies for determining the parameters defining the synchronous signal averaging process, such as the number of averages and the sampling accuracy required. These issues will be discussed in subsequent chapters.

**4.3.8 Applicability to safety critical failure modes**

Synchronous signal averaging appears to be applicable in the detection of all the safety critical failure modes except bearing failures, which were considered to be a minor contributory factor in the area of safety.

From the brief descriptions given above, it would appear that McFadden’s demodulation technique is the only one which provides discrimination between safety critical tooth fractures and other localised tooth faults such as pitting and spalling.

**4.4 CEPSTRAL ANALYSIS**

The power cepstrum is the power spectrum of the logarithm of the power spectrum and the complex cepstrum is the spectrum of the logarithm of the complex spectrum [16]. Both the power cepstrum and the complex cepstrum result in a time domain signal, which in the terminology of cepstral analysis is the quefrency domain, and give a measure of periodic structures in the spectrum. The usefulness of the cepstrum is in the fact that a series of harmonically related structures reduce to predominantly one ‘quefrency’ at the reciprocal of the harmonic spacing. This allows faults which produce a number of low-level harmonically related frequencies, such as bearing and localised tooth faults, to be
detected. In this respect, it has advantages over spectral analysis in the detection of safety critical faults such as tooth cracking however, it still does not provide any distinction between these and less serious faults such as pitting and spalling.

Cepstral analysis has proved to be a useful tool in the detection of bearing faults; as bearing faults produce a series of impulses which excite structural resonances, the periodicity of the excitation is commonly evident in the ‘quefrency’ domain but, in the frequency domain, it appears as a number of low-level sidebands (separated by the frequency of the impulses and centred about each of the resonant frequencies) which are often difficult to detect.

Cepstral analysis is not very useful in the analysis of synchronously averaged signals because, even though the signal is periodic in the angle domain, and a pure line spectrum in the frequency domain, it looses its periodicity when translated to the quefrency domain and manipulation will introduce discontinuities.

4.5 ADAPTIVE NOISE CANCELLATION

Adaptive noise cancellation (ANC) has been used to increase the effective signal-to-noise ratio for bearing fault detection [79]. Deterministic methods for attenuating non-synchronous signal components (such as synchronous signal averaging) cannot be used for bearings because of the uncertainty in the rotational periodicities due to slippage and skidding. Bearing faults often go undetected due to the low level of the fault signature in relation to vibration from other components in the gearbox. However, it has been shown in laboratory experiments (e.g., Swansson and Favaloro [76]) that even simple time domain techniques, such as RMS, Crest Factor and Kurtosis, provide good bearing fault detection capabilities if the signal-to-noise ratio is sufficiently high. In an operational gearbox, bearing fault detection capabilities can be increased by increasing the signal-to-noise ratio of the fault signature rather than increasing the complexity of the diagnostic techniques.

ANC is implemented by using two transducers; one in close proximity to the bearing being monitored and the other remote from the bearing (usually closer to the source of
the major interfering vibrations or ‘noise’). It is assumed that where there is a bearing fault, the vibration signal from the monitoring transducer contains the fault signature plus ‘noise’ and the signal from the reference transducer contains only the ‘noise’. The two ‘noise’ signals are assumed to be from the same source and correlated, with the difference being due to transmission path effects. An adaptive filter is used, minimising the difference between the two signals, which simulates the transmission path effects between the two transducers. The adaptive filter is applied to the ‘noise’ signal from the reference transducer, giving an estimate of the ‘noise’ signal at the monitoring transducer location, which is then subtracted from the signal from the monitoring transducer to provide an estimate of the bearing fault vibration signature.

The advantage of using ANC is that it requires no a-priori knowledge of the vibration frequencies in the interfering ‘noise’ signal. However, to give good results it requires a monitoring transducer for each bearing (or group of bearings in close proximity) and a remote reference transducer (although it may be possible to use a monitoring transducer for a remote bearing as the reference transducer).

ANC has no advantage in the monitoring of shafts and gears, for which synchronous signal averaging provide greater attenuation of non-synchronous vibration.

4.6 SUMMARY

Of the techniques discussed above:

a) Overall vibration level, crest factor and kurtosis monitoring of the time domain vibration signal do not provide any diagnostic information but may have limited application in fault detection in simple safety critical accessory components.

b) Spectral analysis may be useful in the detection and diagnosis of shaft faults.

c) Synchronous signal averaging has the potential of greatly simplifying the diagnosis of shaft and gear faults (i.e., the safety critical failures) by providing significant attenuation of non-synchronous vibrations and signals on which ideal filtering can be
used. Further development needs to done on the implementation of synchronous averaging techniques and the analysis of results.

d) Cepstral analysis and adaptive noise cancellation mainly have application in bearing fault detection and diagnosis.

Based on the above, and the priorities placed on the safety critical failure modes in Chapter 3, the remainder of this thesis will concentrate on the synchronous signal averaging technique, firstly on further investigation, development, and testing of the process itself and subsequently on analysis methods to improve diagnostic capabilities for the safety critical failure modes.
Chapter 5

SYNCHRONOUS SIGNAL AVERAGING

The concept of synchronous signal averaging was introduced and briefly discussed in Chapter 4 (Section 4.3). In the remainder of this thesis, synchronous signal averaging will be used as a basis of all vibration analysis techniques studied. Therefore, detailed examination of the technique is required, both in terms of its theoretical consequences and practical implementation.

In this chapter,

a) a model of synchronously averaged gearbox vibration is developed to provide a theoretical basis for the description of vibration analysis techniques used in subsequent chapters,

b) a theoretical examination is made of the consequences of synchronous signal averaging, including its effects on non-synchronous vibrations,

c) a new method of quantifying and optimising the effects of synchronous signal averaging is developed which includes a measure of the leakage from non-synchronous vibrations, and

d) methods for the practical implementation of synchronous signal averaging are examined.

5.1 MODEL OF SYNCHRONOUSLY AVERAGED VIBRATION

In Chapter 2, a general model of gearbox vibration was developed expressed in terms of the time-dependant phase of the rotating elements. This model can easily be reformulated in terms of the angular position of any of the rotating elements by replacing
the time dependant phase functions (e.g., $\theta_s(t)$) with an explicit ratio to the reference phase angle, for example

$$\theta_s(t) = R_s \theta_{ref}, \quad (5.1)$$

where $\theta_{ref}$ is the cumulative phase angle of the reference shaft and $R_s$ is the ratio of the rotational position of the shaft of interest to that of the reference. Note that all angles are expressed as a cumulative angle since time $t=0$, and not the actual modulo $2\pi$ angle. In the manner in which the model in Chapter 2 has been formulated, this simple ratio replacement is valid for all components, including the inner and outer bearing races (with the cage and rolling element angles still being expressed in terms of inner and outer race angles and any accumulated slip angle).

With the reformulation expressed in equation (5.1) in place, the synchronously averaged vibration signal over $N_a$ revolutions of a shaft $s$ can be defined in terms of the general gearbox model (2.33) as

$$\bar{x}_s(\theta) = \frac{1}{N_a} \sum_{n=0}^{N_a-1} x(\theta + n2\pi), \quad (5.2)$$

where $\theta$ is the shaft angle of the averaged vibration signal over the period $[0,2\pi]$.

If we assume that all vibration from components other than those related to the shaft of interest are not synchronous with the shaft rotation, and that all non-synchronous vibrations are completely eliminated, then the averaging process performed synchronously with the rotation of the shaft will reduce the total measured vibration signal to the average of the vibration synchronous with the shaft.

For the various gearbox components, this ‘ideal’ average is as follows.
5.1.1 Fixed axis shafts and gears

For a fixed axis shaft, the average (5.2) will reduce to the mean value of the shaft vibration and that of its attached gears,

\[
\bar{x}_s(\theta) = \frac{1}{N_a} \sum_{n=0}^{N_a-1} \left( x_s(\theta + n2\pi) + \sum_{g=1}^{G_s} x_{sg}(\theta + n2\pi) \right) = \frac{1}{N_a} \sum_{n=0}^{N_a-1} \left( v_s(\theta + n2\pi) \ast h_s(\theta + n2\pi) \right) + \sum_{g=1}^{G_s} v_{sg}(\theta + n2\pi) \ast h_{sg}(\theta + n2\pi). \tag{5.3}
\]

Note that although the shaft vibration, \( v_s(\theta) \) (2.9), and gear vibrations, \( v_{sg}(\theta) \) (2.3), are totally periodic with the rotation of the shaft, the measured shaft vibration, \( x_s(\theta) \) (2.26), and measured gear vibrations, \( x_{sg}(\theta) \) (2.25), are subject to transmission path effects which are time/frequency dependant. The net result of this is simply that the synchronous signal average, \( \bar{x}_s(\theta) \), represents the shaft and gear vibration signatures convolved with the mean transmission path effects over the total averaging period;

\[
\bar{x}_s(\theta) = v_s(\theta) \ast \left( \frac{1}{N_a} \sum_{n=0}^{N_a-1} h_s(\theta + n2\pi) \right) + \sum_{g=1}^{G_s} v_{sg}(\theta) \ast \left( \frac{1}{N_a} \sum_{n=0}^{N_a-1} h_{sg}(\theta + n2\pi) \right) \tag{5.4}
\]

\[
= v_s(\theta) \ast \bar{h}_s(\theta) + \sum_{g=1}^{G_s} v_{sg}(\theta) \ast \bar{h}_{sg}(\theta),
\]

where \( \bar{h}_s(\theta) \) and \( \bar{h}_{sg}(\theta) \) represent the mean transmission path effects for the shaft and gears respectively.
5.1.2 Epicyclic gear trains

The following is a model of the synchronously averaged vibration for epicyclic gear train components using ‘conventional’ signal averaging.

5.1.2.1 Planet-carrier (ring gear) average

When the signal averaging process is performed using the planet-carrier rotation as the reference angle, only the ring-planet mesh vibration is synchronous with the reference. From equation (2.27), with the non-synchronous planet-ring mesh vibration \( v_{pr}(\theta_{ref}) \) removed, the signal averaged ring gear (planet-carrier) vibration is

\[
\overline{\chi}_r^{(r)}(\theta) = \frac{1}{N_a} \sum_{n=0}^{N_a-1} \chi_r^{(r)}(\theta + n2\pi)
\]

\[
= \frac{1}{N_a} \sum_{n=0}^{N_a-1} \left[ \sum_{p=0}^{P_r-1} \Phi_r \left( \theta - \frac{p2\pi}{P_r} + n2\pi \right) v_{rp}(\theta + n2\pi) \right] \ast \eta_r(\theta + n2\pi).
\]

After averaging, the time based transmission path effect \( \eta_r(\theta) \) reduces to a mean transmission path effect \( \overline{\eta}_r(\theta) \) giving

\[
\overline{\chi}_r^{(r)}(\theta) = \left[ \sum_{p=0}^{P_r-1} \Phi_r \left( \theta - \frac{p2\pi}{P_r} \right) v_{rp}(\theta) \right] \ast \overline{\eta}_r(\theta).
\]

That is, the signal average for a planet-carrier is the sum of the vibration due to the ring gear meshing with each planet weighted by the planet pass modulation and convolved with a mean transmission path effect.

5.1.2.2 Planet gear average

When the signal averaging process is performed using an epicyclic planet gear rotation as the reference, we will get a signal which is a combination of the planet-ring and planet-
sun mesh vibrations for all planets in the epicyclic gear train (since all the planets rotate at the same rate). From equations (2.27) and (2.28), with the non-synchronous ring-planet and sun-planet mesh vibrations removed, the signal averaged planet vibration is

\[ \chi_{\theta}^{(p)}(\theta) = \frac{1}{N_a} \sum_{n=0}^{N_a-1} \left( \chi_{\theta}^{(r)}(\theta + n2\pi) + \chi_{\theta}^{(s)}(\theta + n2\pi) \right) \]

where \( R_r \) is the ratio of the planet-carrier rotation to the planet rotation. Note that the planet pass modulations (\( \Phi_r \) and \( \Gamma_r \)) are not synchronous with the planet rotation (but with the planet-carrier rotation) and the averaging process will reduce these to their mean values (\( \overline{\Phi}_r \) and \( \overline{\Gamma}_r \)). Because these functions do not have mean values of zero, they cannot be eliminated by the signal averaging process, but their net effect is to introduce constant scaling factors. The time based transmission path effects \( \eta_r \) and \( \mu_r \) reduce to mean transmission path effects, giving

\[ \chi_{\theta}^{(p)}(\theta) = \left[ \Phi_r \sum_{p=0}^{P_r-1} v_{pr}(\theta) \right] \cdot \overline{\eta}_r(\theta) + \left[ \Gamma_r \sum_{p=0}^{P_r-1} v_{ps}(\theta) \right] \cdot \overline{\mu}_r(\theta). \]

That is, the signal average for the planet gears is a scaled version of the sum of all planet-ring gear mesh vibrations convolved with a mean transmission path effect plus a scaled version of the sum of all planet-sun mesh vibrations convolved with a second mean transmission path effect.
5.1.2.3 Sun gear average

When the signal averaging process is performed using the sun gear rotation as the reference angle, only the sun-planet mesh vibration is synchronous with the reference. From equation (2.28), with the non-synchronous planet-sun mesh vibration $v_{ps}(\theta_{ref})$ removed, the signal averaged sun gear vibration is

$$\chi^{(s)}_{r}(\theta) = \frac{1}{N_a} \sum_{n=0}^{N_a-1} \chi^{(s)}_{r}(\theta + n2\pi)$$

or

$$= \frac{1}{N_a} \sum_{n=0}^{N_a-1} \left[ \sum_{p=0}^{P_{r}-1} \Gamma_{r}\left(R_{rs}(\theta + n2\pi) - \frac{p2\pi}{P_{r}}\right)v_{sp}(\theta + n2\pi) \right] * \mu_{r}(\theta + n2\pi),$$

where $R_{rs}$ is the ratio of the planet-carrier rotation to the sun gear rotation. Note that the planet pass modulation ($\Gamma_{r}$) is not synchronous with the sun gear rotation (but with the planet-carrier rotation) and the averaging process will reduce this to its mean value ($\bar{\Gamma}_{r}$). The time based transmission path effect $\mu_{r}(\theta_{ref})$ reduces to a mean transmission path effect, giving

$$\chi^{(s)}_{r}(\theta) = \left[ \bar{\Gamma}_{r} \sum_{p=0}^{P_{r}-1} v_{sp}(\theta) \right] * \bar{\mu}_{r}(\theta).$$

That is, the signal average for the sun gear is the sum of the vibration due to the sun gear meshing with each planet scaled by the mean planet pass modulation and convolved with a mean transmission path effect.

5.1.2.4 Number of averages required for sun and planet gears

In the models of signal averaged vibration for planet gears, equation (5.8) and the sun gear (5.10), it was assumed that the planet pass modulation would be reduced to its mean value. This is only true if the modulation is exactly periodic over the total
averaging time. That is, the planet-carrier must complete an integer number of revolutions in the time taken for $N_a$ averages.

The number of averages for a sun or planet gear will usually need to be an integer multiple of the number of teeth on the corresponding ring gear to ensure the planet-carrier completes an integer number of revolutions during the averaging period. The exception to this is if there is a common factor between the number of teeth on the sun or planet gear and the number of teeth on the ring gear; in this case, the number of averages can be an integer multiple of the number of teeth on the ring gear divided by the common factor in the teeth numbers.

5.1.3 Bearings

Synchronous signal averaging is not usually applied to bearing vibration as the unpredictable slip inherent in bearings makes the accurate calculation of the rotational angles almost impossible.

5.2 ATTENUATION OF NON-SYNCHRONOUS VIBRATION

In the above, it was assumed that all non-synchronous vibration was completely eliminated by the averaging process. In practice, this will generally not be the case. The averaging process provides an attenuation of non-synchronous vibration which is dependant upon the nature of the vibration and the number of averages.

5.2.1 Random (non-periodic) vibration

In the case of a purely random vibration (i.e., one which has no underlying periodicity), the attenuation of the signal is proportional to the square root of the number of averages (Braun [14]). This is shown in Figure 5.1, where the RMS value of a normally distributed random signal (with initial RMS of 1.0) is plotted against the number of averages (1 to 128).
5.2.2 Non-synchronous periodic vibration

A number of authors have presented models of signal averaging of non-synchronous periodic waveforms (e.g., McFadden [53] and Succi [75]), however, they have not presented a simple procedure for quantifying the effect of the number of averages on an arbitrary periodic signal. Such a model will now be developed.

Consider the synchronous signal average of a single cosine wave of amplitude $A_R$, with initial phase $\phi_R$, and at a ratio $R$ to the reference angle $\theta$;

$$x_R(\theta) = A_R \cos(R\theta + \phi_R).$$  \hspace{1cm} (5.11)

The signal average of $x_\theta(\theta)$ over $N$ periods of $\theta$ is
\[
\bar{x}_{RN}(\theta) = \frac{1}{N} \sum_{n=0}^{N-1} x_R(\theta + n2\pi) = \frac{1}{N} \sum_{n=0}^{N-1} A_R \cos\left(R(\theta + n2\pi) + \phi_R\right)
\]

\[
= \frac{A_R}{N} \sum_{n=0}^{N-1} \left(\cos\left(R\theta + \phi_R\right)\cos\left(Rn2\pi\right) - \sin\left(R\theta + \phi_R\right)\sin\left(Rn2\pi\right)\right) \quad (5.12)
\]

\[
= A_R \left\{ \cos\left(R\theta + \phi_R\right) \frac{1}{N} \sum_{n=0}^{N-1} \cos\left(Rn2\pi\right) \right\} \\
- \sin\left(R\theta + \phi_R\right) \frac{1}{N} \sum_{n=0}^{N-1} \sin\left(Rn2\pi\right)
\]

In the case where \( R \) is an integer, it is clear from equation (5.12) that the signal average reduces to the original signal (i.e., no attenuation); the summation in \( \cos(Rn2\pi) \) is \( N \) and the summation in \( \sin(Rn2\pi) \) is zero.

For the case where \( R \) is not an integer, the following observations are made:

a) for each step \( n \) in the averaging process the effective increment in angle is \( \frac{R}{2}\pi \) (where \( \frac{R}{2}\pi \) is the fractional part of \( R \)),

b) if \( \frac{R}{2}\pi \leq 0.5 \) the summations in \( \cos(Rn2\pi) \) and \( \sin(Rn2\pi) \) are equivalent to integration over the period \([0, \frac{R}{2}\pi]\],

c) if \( \frac{R}{2}\pi > 0.5 \) the summation in \( \cos(Rn2\pi) \) and \( \sin(Rn2\pi) \) are equivalent to integration over the period \([0, (\frac{R}{2}\pi-1)]\) (i.e., integration in negative direction).

Based on the above observations, equation (5.12) can be rewritten,

using: \( \alpha_{RN} = \begin{cases} \frac{R}{2}\pi N, & 0 < \frac{R}{2}\pi \leq 0.5 \\ (\frac{R}{2}\pi-1)N2\pi, & 0.5 < \frac{R}{2}\pi < 1, \end{cases} \) \quad (5.13)
\[
\bar{x}_{RN}(\theta) = \frac{A_R}{\alpha_{RN}} \begin{pmatrix}
\cos(R\theta + \phi_R) \int_0^{\alpha_{RN}} \cos(\alpha) d\alpha \\
-\sin(R\theta + \phi_R) \int_0^{\alpha_{RN}} \sin(\alpha) d\alpha
\end{pmatrix}
\]

\[
= \frac{A_R}{\alpha_{RN}} \left( \cos(R\theta + \phi_R)\sin(\alpha_{RN}) + \sin(R\theta + \phi_R)(\cos(\alpha_{RN}) - 1) \right) \tag{5.14}
\]

\[
= \frac{A_R}{\alpha_{RN}} \left( \sin(R\theta + \phi_R + \alpha_{RN}) - \sin(R\theta + \phi_R) \right).
\]

and the RMS (standard deviation) value of \( \bar{x}_{RN}(\theta) \) is

\[
\sigma_R(N) = \frac{A_R}{\alpha_{RN}} \left| \int_{-\infty}^{\infty} \left( \sin(R\theta + \phi_R + \alpha_{RN}) - \sin(R\theta + \phi_R) \right)^2 d\theta \right|
\]

\[
= \frac{A_R}{\alpha_{RN}} \left| \int_{-\infty}^{\infty} \left( \sin^2(R\theta + \phi_R + \alpha_{RN}) + \sin^2(R\theta + \phi_R) - 2\sin(R\theta + \phi_R + \alpha_{RN})\sin(R\theta + \phi_R) \right) d\theta \right|
\]

\[
= \frac{A_R}{\alpha_{RN}} \left| \int_{-\infty}^{\infty} 1 - 2\cos(\alpha_{RN})\sin^2(R\theta + \phi_R) d\theta \right|
\]

\[
= \frac{A_R}{\alpha_{RN}} \left| \int_{-\infty}^{\infty} -2\sin(\alpha_{RN})\cos(R\theta + \phi_R)\sin(R\theta + \phi_R) d\theta \right|
\]

\[
= \frac{A_R}{\alpha_{RN}} \left| \int_{-\infty}^{\infty} 1 - \cos(\alpha_{RN}) \right|
\]

From equation (5.15) it is clear that there are two factors contributing to the attenuation of non-synchronous periodic signals:

a) Firstly, there is the division by \( \alpha_{RN} \) which is proportional to the number of averages (note that this is factored by the fractional part of the ratio \( R \) and that the closer \( R \) is to 0 or 1, that is, the closer the frequency is to an harmonic of the reference shaft, the lower the rate of attenuation).
b) Secondly, and of more significance, is the cosine term $\cos(\alpha_{RN})$. It is obvious that the non-synchronous periodic vibration will be totally eliminated (RMS of 0) if the value of $\alpha_n$ is an integer multiple of $2\pi$. This can be achieved by setting the number of averages $N$ such that $N \times \text{frac}(R)$ is an integer.

To illustrate this effect, Figure 5.2 shows the results of averaging a simulated signal consisting of a sine wave with a frequency 20.05 times that of the reference shaft. The initial RMS value of the signal was 1 ($A=\sqrt{2}=1.414$). From equations (5.13) and (5.15), it would be expected that the averaged signal will have a value of zero at multiples of 20 averages ($20 \times 0.05 = 1$). This effect can clearly be seen in Figure 5.2.

![Figure 5.2 Attenuation of non-synchronous periodic signal (20.05 order sine wave)](image)

The effect of the division by $\alpha_{RN}$ is best illustrated at the mid-points between the nodes; at $N=10$, $\alpha_{RN} = 10 \times 0.05 \times 2\pi = \pi$, which, from equation (5.15), will give an RMS value of $2/\pi = 0.6366$. The difference between the theoretical value and the value gained by experiment (0.6384) is due to the finite number of points (1024) in the experimental average. The signal, being non-synchronous, is not accurately represented by the finite number of discrete points.

93
At \( N=30 \), \( \alpha_{RN} = 30 \times 0.05 \times 2\pi = 3\pi \) which gives a theoretical RMS value of \( 2/3\pi = 0.2122 \). The value gained by experiment (0.2127) is acceptable given the limitations of the finite number of points.

### 5.3 THE IDEAL NUMBER OF AVERAGES

In Section 5.2.2, only a single sine wave was considered when developing the model of signal averaging for non-synchronous periodic signals. However, if we set the number of averages \( N \) such that a signal at ratio \( R \) is totally eliminated by the averaging process, then all frequencies which are harmonics of \( R \) will also be eliminated. This can easily been seen if you consider that (from equation (5.15));

- \( a) \) for the frequency at ratio \( R \) to be eliminated, \( \frac{R}{N} = k \) must be an integer, and
- \( b) \) for all integer multiples \( m \) of \( R \), \( \frac{mR}{N} = (m \times \frac{R}{N} - p)N = mk - pN \) which will also be an integer (note, here \( p \) represents an integer between 0 and \( m-1 \)).

From the model of gearbox vibration developed in Chapter 2, all vibration from a particular shaft and its attached gears are harmonics of the shaft rotation frequency. Therefore, if we calculate the number of averages required to completely eliminate the rotational frequency of a particular shaft, we will eliminate all vibration pertaining to that shaft and its attached gears.

Generally, the ratio of one shaft to another shaft is represented as the ratio of two integers related to the number of teeth on the intervening meshing gears (e.g., \( R = \frac{N1}{N2} \)). When a ratio is expressed in this fashion, and all common factors in \( N1 \) and \( N2 \) have been removed, the minimum number of averages required to eliminate all vibration synchronous with the shaft at ratio \( \frac{N1}{N2} \) is \( N = N2 \). If we also wish to eliminate all vibration related to a second shaft with ratio \( \frac{N3}{N4} \), then the number of averages \( N \) must be an integer multiple of \( N4 \) as well as of \( N2 \) (i.e., \( N = m1 \times N2 = m2 \times N4 \) where \( m1 \) and \( m2 \) are both integers). The minimum number of averages required to eliminate the vibration from both shafts will be \( N = N2 \times N4 \) unless the divisors \( N2 \) and \( N4 \) are common factors of a lower number.
Although it is theoretically possible to eliminate vibration from all other shafts in the gearbox, the minimum number of averages required to do this can become impractically large. For example, in a Sea King helicopter main rotor gearbox, to eliminate all other vibration from the signal average for the high speed input shaft would require 22,236,000 averages; at the shaft rotation frequency of 316 Hz, this would require more than 19.5 hours of continuous data!

Obviously a compromise solution is required; we wish to reduce vibration from other gearbox components to a minimum whilst keeping the number of averages within feasible limits.

5.4 OPTIMISING THE NUMBER OF AVERAGES

Where there is random noise present in the signal (e.g., measurement noise, external noise such as turbulence etc.) a minimum number of averages can be calculated (based on the estimated signal-to-noise ratio) which will reduce the noise to an acceptable level: to reduce the RMS value of the noise by a factor of \( c \) the number of averages \( N \) must be \( \geq c^2 \).

For periodic ‘noise’, it was shown in Section 5.2.2 that although there is a statistical reduction in the amplitude of the ‘leaked’ vibration with increasing number of averages, local minima exist which may be used to optimise the number of averages for the attenuation of vibration at particular frequencies.

5.4.1 Previous methods

Stewart [73] proposed a method of measuring the stability of the signal average by using the zero-lag cross-coefficient between the signal obtained after \( N \) averages and that obtained after \( N/2 \) averages. This stability measure was referred to as the ‘leakage rate’ (this should not be confused with the actual ‘leakage’ of vibration from other components, but it was assumed to be related). A method of stopping the averaging process was proposed in which, after every power of two averages, the stability measure
was compared against some pre-determined value (typically 0.05); the averaging process was stopped if the stability measure was less than the pre-determined value, otherwise the process continued to the next power of two averages. This method assumes that reduction in leakage of vibration from other sources is directly proportional to the number of averages and that changes caused by an increase in the number of averages are purely due to the attenuation of non-synchronous (both random and periodic) vibrations. No attempt was made to identify optimum number of averages which are not a power of two.

McFadden [53] modelled the signal averaging process as a modified comb filter, taking into account the restrictions due to the use of a finite number of discrete samples. Using this model, he showed that the number of averages could be optimised for the rejection of periodic noise at a discrete frequency. This was done by adjusting the number of averages so that a node in the side lobe structure between the major lobes defining the comb teeth coincided with the frequency of the periodic noise. A formal procedure for the calculation of the optimum number of averages was not defined, however the method is analogous to the identification of the points at which \( (\alpha_n \mod 2\pi) \) equals zero for a particular ratio \( R \) in equation (5.13). McFadden [53] did not attempt to expand the optimisation to multiple frequencies or to quantify the total leakage of non-synchronous vibration in the signal average.

Succi [75] modelled the signal averaging process applied to an harmonic series with a fundamental ratio to the reference shaft of \( N1/N2 \) as the summation of an exponential series. From this he concluded that the value of the summation could be minimised by choosing the number of averages equal to an integer multiple of \( N2 \). This is similar to, and results in the same conclusion as, the derivation of the ‘ideal’ number of averages given in Section 5.3. However, Succi did not correctly quantify the attenuation of periodic vibrations where the number of averages was not an integer multiple of \( N2 \), merely assuming this to be proportional to the number of averages.
5.4.2 A new method for optimising the number of averages

Although the methods proposed by McFadden [53] and Succi [75] allow some optimisation of the number of averages for the rejection of a single frequency (McFadden) or an harmonic series (Succi), neither method provides an optimisation strategy for minimising the total amount of leaked vibration in the signal average.

A general optimisation strategy requires the quantification of the leaked vibration in the signal average and an estimate of the signal-to-noise ratio at any point in the averaging process. Based on these, the number of averages can be selected to maximise the signal-to-noise ratio.

In the following derivation, the mean squared values (= RMS²) have been used to simplify the equations; the mean squared value of the summation of a number of sinusoids is simply the sum of their mean squared values.

5.4.2.1 Quantification of leakage

In the past, no method was available for quantifying the value of the ‘leakage’ of non-synchronous vibration in the signal average. Stewart [73] made some attempt to do this by assuming that the change in a signal as the number of averages increased was directly proportional to the reduction in leakage; this is only valid for non-periodic random noise.

Equation (5.15) gives a formula for the calculation of the RMS value of leakage of periodic vibration at any frequency for a given number of averages \( N \). The extension of this to a measure of the total mean squared value (RMS²) of leaked periodic vibrations is achieved by simply summing the mean squared values of the leakage at all frequencies

\[
\sigma^2_p(N) = \sum_{R=0}^{\infty} \sigma^2_R(N) = \sum_{R=0}^{\infty} \left| \frac{A_R}{\alpha_{RN}} \right|^2 (1 - \cos(\alpha_{RN})).
\]  

(5.16)

Given that the initial RMS value of non-periodic random noise in the signal is \( \sigma_e \), the total mean squared leakage of non-synchronous vibration (noise) after \( N \) averages will be
the summation of the mean squares of the random and non-synchronous periodic vibrations:

\[ \sigma_{\text{noise}}^2(N) = \frac{\sigma_e^2}{N} + \sigma_p^2(N) = \frac{\sigma_e^2}{N} + \sum_{R=0}^{\infty} \left| \frac{A_R}{\alpha_{RN}} \right|^2 \left( 1 - \cos(\alpha_{RN}) \right). \] (5.17)

### 5.4.2.2 Estimating the signal-to-noise ratio

The signal-to-noise ratio (SNR) of a signal average after \( N \) averages is defined as the ratio of the RMS level of synchronous vibration components to the non-synchronous vibration (noise) components. Since the signal average is the sum of the synchronous and the (attenuated) non-synchronous components, a simple estimate of the RMS of the synchronous components can be made by subtracting the mean square of the non-synchronous noise (5.17) from that of the signal average:

\[ \sigma_s^2(N) = \sigma_{\text{total}}^2(N) - \sigma_{\text{noise}}^2(N) \] (5.18)

and the square of the SNR after \( N \) averages is

\[ \text{SNR}^2(N) = \frac{\sigma_s^2(N)}{\sigma_{\text{noise}}^2(N)} = \frac{\sigma_{\text{total}}^2(N) - \sigma_{\text{noise}}^2(N)}{\sigma_{\text{noise}}^2(N)} = \frac{\sigma_{\text{total}}^2(N)}{\sigma_{\text{noise}}^2(N)} - 1 \]

\[ = \frac{\sigma_{\text{total}}^2(N)}{\sigma_e^2/N + \sum_{R=0}^{\infty} \left| \frac{A_R}{\alpha_{RN}} \right|^2 \left( 1 - \cos(\alpha_{RN}) \right)} - 1. \] (5.19)

### 5.4.2.3 Optimisation strategy

Within a range of acceptable values for the number of averages for a given shaft, we wish to maximise the signal-to-noise ratio (5.19), which is equivalent to the minimisation of the noise term \( \sigma_{\text{noise}}(N) \). From equation (5.17), this requires an initial estimate of the RMS of the random vibration \( \sigma_e \) and the amplitudes of the non-synchronous vibrations \( A_R \).
An initial estimate of the amplitude of the major periodic vibration components can be obtained by calculating ‘trial’ signal averages for each shaft. To restrict the amount of processing required, and to avoid problems due to cross-leakage (i.e., identification of a synchronous signal as a non-synchronous signal due to leakage in the ‘trial’ signal averages), only components above a specified amplitude (e.g., 5% of the maximum amplitude) should be considered.

NOTE: coincident vibration components (i.e., those which are periodic with the rotation of two or more shafts) will remain un-attenuated in the signal averages for the shafts whose rotation they are periodic with. To avoid duplication of these components, they should either (a) be removed from all but one of the signal averages or (b) have their amplitudes in all signal averages divided by the number of shafts for which they are coincident.

Once the initial estimates of the amplitude of the major components are made, the leakage after $N$ averages due to vibration at the first $M$ harmonics on shaft $q$ with a ratio $R$ to the shaft for which the signal averaging is being calculated can be expressed as

\[
\sigma_{pq}^2(N) = \sum_{m=1}^{M} \left| \frac{A_q(m)}{\alpha_{mRN}} \right|^2 \left(1 - \cos(\alpha_{mRN}) \right),
\]

where $A_q(m)$ is the estimated (peak) amplitude of the $m$th harmonic of shaft $q$. Note that although the summation is shown over the first $M$ harmonics, only those values for which $A_q(m)$ is greater than a specified minimum amplitude need be calculated in practice.

The estimate of the total leakage of non-synchronous periodic vibrations, equation (5.16), becomes the summation of the leakage for all shafts (other than that for which the signal average is being calculated). Where there are $Q$ other shafts in the gearbox,

\[
\sigma_P^2(N) \approx \sum_{q=1}^{Q} \sigma_{pq}^2(N) \approx \sum_{q=1}^{Q} \sum_{m=1}^{M} \left| \frac{A_q(m)}{\alpha_{mRN}} \right|^2 \left(1 - \cos(\alpha_{mRN}) \right)
\]
and the mean squared SNR estimate (5.19) becomes

\[
SNR^2(N) \approx \frac{\sigma_{total}^2(N)}{\frac{\sigma_e^2}{N} + \frac{\sigma_p^2}{N} + \sum_{q=1}^{Q} \sum_{m=1}^{M} \frac{|A_q(m)|^2}{\alpha_{mRN}} (1 - \cos(\alpha_{mRN}))} - 1.
\] (5.22)

An initial estimate of the random noise component can be made empirically or by theoretical examination of the vibration data. To estimate the random noise component from the vibration data itself, use is made of the estimated leakage from periodic noise, \(\sigma_p(N)\), given in equation (5.21) and the relationship between the synchronous signal, total signal and noise signals given in equation (5.18). If we assume stationarity of the signal over the period of 2N averages then \(\sigma_s(N) = \sigma_s(2N)\), and by calculating the signal averages of a shaft for both N and 2N averages, an estimate of the mean squared value of the random noise \(\sigma_e\) can be made as follows;

\[
\sigma_{total}^2(N) - \sigma_{total}^2(2N) = \sigma_{noise}^2(N) - \sigma_{noise}^2(2N)
\]

\[
= \frac{\sigma_e^2}{N} + \sigma_p^2(N) - \frac{\sigma_e^2}{2N} - \sigma_p^2(2N),
\]

\[
\Rightarrow \sigma_e^2 = 2N\left(\sigma_{total}^2(N) - \sigma_{total}^2(2N) - \sigma_p^2(N) + \sigma_p^2(2N)\right).
\] (5.23)

Using the mechanisms described above for quantifying the leakage due to periodic and random ‘noise’, a simple search strategy can be employed to find the number of averages which will maximise the SNR of the signal average as defined in equation (5.22).

Where \(N_{ref}\) is the number of averages used for the ‘trial’ average of the reference shaft,

\(N_{max}\) is the maximum allowable number of averages,

\(SNR_{min}\) is the minimum allowable signal to noise ratio, and

\(SNR_{req}\) is the required (or desired) signal to noise ratio:
1. Estimate the mean squared value of the synchronous signal from the trial signal average

\[ \tilde{\sigma}_s^2 = \sigma_{total}^2\left( N_{ref} \right) - \frac{\sigma_{e}^2}{N_{ref}} - \sum_{q=1}^{Q} \sum_{m=1}^{M} \left| \frac{A_q(m)}{\alpha_{mRN_{ref}}} \right|^2 \left( 1 - \cos(\alpha_{mRN_{ref}}) \right). \]

2. Establish the minimum number of averages \( N_{min} \) required to reduce the random noise term below that of the maximum allowable noise

\[ N_{min} = \frac{\sigma_{e}^2 \cdot SNR_{min}^2}{\tilde{\sigma}_s^2}. \]

3. Estimate the required noise floor

\[ \tilde{\sigma}_{noise}^2 \approx \frac{\tilde{\sigma}_s^2}{SNR_{req}^2}. \]

4. Establish the optimum number of averages \( N_{avg} \) to perform:

For \( N = N_{min} \) to \( N_{max} \)

estimate the total noise at \( N \) averages

\[ \sigma_{noise}^2\left( N \right) \approx \frac{\sigma_{e}^2}{N} + \sum_{q=1}^{Q} \sum_{m=1}^{M} \left| \frac{A_q(m)}{\alpha_{mRN}} \right|^2 \left( 1 - \cos(\alpha_{mRN}) \right) \]

if \( \sigma_{noise}^2\left( N \right) < \tilde{\sigma}_{noise}^2 \)

set \( N_{avg} = N \) and stop the search.

if \( \sigma_{noise}^2\left( N \right) < \sigma_{noise}^2\left( N_{avg} \right) \)

set \( N_{avg} = N. \)

5.4.2.4 An example of optimisation of the number of averages

To demonstrate this procedure, a simulated signal is used which consists of a single sinusoid at 41 orders of a reference shaft plus a series of harmonics of 43 times a ratio of 41/96 and a second series of harmonics of 25 times a ratio of 1763/10464 (based on ratios of the Sea King main rotor gearbox intermediate shaft and main shaft to the high
speed input shaft). The vibration components in the test signal and their RMS amplitude are as follows:

reference shaft: 41 orders (1.414g RMS)

41/96 x reference: 43 orders (2.828g RMS) + 86 orders (0.353g RMS)

1763/10464 x ref: 25 orders (2.828g RMS) + 50 orders (0.141g RMS) + 75 orders (0.707g RMS) + 100 orders (6.01g RMS)

![Spectrum HSAVG32.VIB](image1)

![Spectrum HSAVG64.VIB](image2)

(a) Test signal after 32 averages  
(b) Test signal after 64 averages

*Figure 5.3 Spectra of reference shaft test signal averages*

Trial synchronous signal averages were calculated using 32 averages for each shaft. Figure 5.3 (a) shows the spectrum of the signal average (after 32 averages) for the reference shaft. The only vibration synchronous with this shaft is the 41 order signal at 1.414 g. Leaked vibration components can be seen around 4 orders (≈25 orders @ ratio of 1763/10464) and 17 orders (≈100 orders @ ratio of 1763/10464). A second signal average was performed for the reference shaft using 64 averages. The spectrum of this is shown in Figure 5.3 (b). This shows a small reduction in the leaked vibration components.

Note that the major vibration component in the (unaveraged) signal is at 100 orders of the shaft at ratio 1763/10464 to the reference. If the method suggested by Succi [75], or the ‘ideal’ number of averages were used, we would need to perform 10464 averages of the reference shaft. Using McFadden’s [53] method for removal of just the 100 order signal would require 2616 averages. If we continue to increase the number of averages
by a power of two, as was suggested by Stewart [73], the leakage vibration continues to
decrease, with the 17 order value reducing (from experimental results) from 0.15g after
64 averages, to 0.09g after 128 averages, to 0.02g after 256 averages and so on.

The optimisation strategy outlined in Section 5.4.2.3 was used with ‘trial’ signal
averages calculated using 32 averages and a random noise estimate of 0. The ‘required’
signal-to-noise ratio was set to 100 (40 dB) and the maximum allowable number of
averages was set to 1000.

The optimisation routine selected 33 averages! This is somewhat surprising given that
the original signal-to-noise ratio is less than 0.3 (-11.2dB), the theoretical ideal number
of averages is 10464, and the signals after 32 and 64 averages (Figure 5.3) both have
signal-to-noise ratios of less than 10 (20dB). The spectrum of the signal obtained using
33 averages is shown in Figure 5.4.

As can be seen from the spectrum in Figure 5.4, the leakage of vibration from the other
shafts is almost totally eliminated after 33 averages (compare this to the spectra obtained
after 32 and 64 averages in Figure 5.3). The component at 17 orders has been reduced
to approximately 0.01g (barely detectable in the spectrum) and the component at 4
orders (which cannot be seen here) has been reduced to less than 0.001g. The SNR is
113 (41 dB), confirming the correct operation of the optimisation procedure.
Figure 5.5 shows the RMS of the predicted periodic signal leakage ($\sigma_p$ from equation (5.21)) versus the number of averages ranging from 32 to 288. This shows the local minimum value at 33 averages, plus a series of local minimum values (all approximately 0.01g) at multiples of 33 averages.

Figure 5.5 Predicted leakage for test signal (32-288 averages)

5.4.3 Practical significance

In addition to the formalisation of a procedure for selecting the optimum number of averages for any shaft in a gearbox, the work presented above provides mechanisms for the quantification of a number of statistical properties related to the signal averaged data and the initial signal.

Equation (5.22) gives an estimate of the signal-to-noise ratio of synchronously averaged vibration data. In the past, only very crude estimates of the SNR were available; related to the statistically reduction of non-synchronous vibration modelled as purely random noise. Using the new derivation of SNR, a judgement on the quality of the synchronously averaged data can be made with some degree of confidence. This can be used to assess the validity of various signal metrics based on the signal average.
A new method of estimating the random noise present in a gearbox vibration signal is given in equation (5.23). This can be used in the absence of (or in conjunction with) an empirical estimate of the random noise in the vibration signal to make a judgement on the overall quality of the (original) recorded vibration signal.

5.5 IMPLEMENTATION METHODS

In the preceding sections, the theoretical consequences of performing synchronous signal averaging have been studied. In this section, the practical application of the process will be examined. Because most modern day vibration analyses take place on discrete signals, only the implementation of synchronous signal averaging of discrete signals will be examined here.

To implement the synchronous signal averaging process, we need to

a) synchronise the averaging with the rotation of a particular component within the gearbox, and

b) control the discrete sampling process to ensure that the sample points are coherent with the rotation of the component.

5.5.1 Synchronisation

Synchronisation of the averaging process requires an angular reference for the rotating component for which the synchronous signal average is to be performed. Ideally, this would be a continuous signal providing the instantaneous angular position (or azimuth) of the component. However, in a complex mechanical system such as a helicopter main rotor gearbox, the installation of sensors to provide such a signal for all rotating components is impractical. Therefore, some compromise measure is required.

Often, complex mechanical systems have some existing form of reference signal which can be adapted for the purposes of synchronising the averaging process. This could be a speed reference signal (as long as it is a ‘pulsed’ type speed reference), ignition pulse (for
internal combustion engines), or AC generator/alternator signal (where this is directly geared to the machine). Note that analogue speed reference signals (i.e., ones which give a voltage proportional to frequency) are not suitable because they do not provide any phase reference.

In the absence of any suitable reference signal, a special purpose sensor needs to be used. It is usually only practical to install the sensor on a shaft which extends beyond the casing (i.e., input and output shafts). A number of different types of sensors may be used, such as:

a) Optical pick-ups, which output a signal proportional to the reflectivity of a surface. By attaching a piece of reflective tape to a small part of the circumference of a shaft and positioning the optical pick-up over the shaft, a once-per-rev pulse is obtained each time the reflective tape passes under the optical pick-up. If necessary the reflectivity of the remainder of the shaft circumference may be reduced by using paint or tape with a low reflectivity. Optical pick-ups can cause problems when used in high ambient light conditions or where they are subjected to contaminants such as dust or oil which can reduce the effectiveness of the optical sensor and/or emitter.

b) Displacement or eddy current probes, which output a signal proportional to the distance between the probe and a surface. A small irregularity, such as a drilled indentation, is placed in an otherwise smooth cylindrical surface (e.g., a flange) attached to a shaft and, with the probe directed at the surface, a once-per-rev pulse (positive or negative) is obtained each time the irregularity passes the probe. In some circumstances it is possible to position an eddy current probe over the teeth of a gear such that a pulse occurs each time a gear tooth passes the probe. In aircraft applications it is often not possible (or difficult) to install these type of sensors; although the physical installation may not be difficult, the modification required to the rotating components needs to be considered from an airworthiness point of view.

c) Shaft encoders, which provide multiple pulses per revolution of a shaft. Typically, these are self-contained devices with an internal slotted disc and an optical or electromagnetic sensor which outputs a pulse each time a slot in the disc passes. The shaft
encoder needs to be attached to the end of a shaft and is therefore limited in application to those machines which have accessible overhung shafts.

5.5.1.1 Tacho multiplication

A synchronisation (tacho) signal is normally not available for all rotating components, therefore some means of adapting a synchronisation signal from one shaft for use with other shafts is required.

In a geared transmission system, the ratio of one shaft to any other shaft is easily calculated from the numbers of teeth on the mating gears between the shafts; this may consist of gears on intermediate shafts as well as the gears on the two shafts themselves. The teeth on a pair of meshing gears must mesh at the same rate, therefore the rate of rotation of the gears (and the shafts on which they are mounted) must be related by the ratio of the numbers of teeth on each gear.

For example, if a gear with \( N_{11} \) teeth on a shaft \( s_1 \) rotating at a frequency of \( f_1 \) hertz is meshing with a second gear with \( N_{21} \) teeth, the rotational frequency, \( f_2 \), of the second gear and the shaft \( s_2 \) on which it is mounted must be

\[
f_2 = f_1 \frac{N_{11}}{N_{21}}. \tag{5.24}
\]

If there is another gear on the shaft \( s_2 \) which has \( N_{22} \) teeth and is meshing with a gear with \( N_{31} \) teeth on shaft \( s_3 \), the rotational frequency \( f_3 \) of shaft \( s_3 \) is

\[
f_3 = f_2 \frac{N_{22}}{N_{31}} = f_1 \frac{N_{11} N_{22}}{N_{21} N_{31}}. \tag{5.25}
\]

Note that a similar expansion to that relating the frequency, \( f_3 \), of the third shaft back to that of the first shaft, \( f_1 \), can be extended to any shaft in the gearbox. That is, the frequency of any shaft can be related to that of any other shaft using only the numbers of
teeth on the intervening gears. Therefore we can generalise equation (5.25) to relate the frequency \( f_n \) of any shaft \( s_n \) to the frequency \( f_r \) of a reference shaft \( s_r \),

\[
f_n = f_r \frac{N_n}{M_n}.
\] (5.26)

The angular position of the rotating components can be related in the same fashion as their frequencies. If \( \theta_n \) is the angular position of shaft \( s_n \) and \( \phi_n \) is its angular position at time 0, then the relationship shown for the frequency of the shaft to that of the reference given in equation (5.26) can be rewritten for the angular positions as

\[
\theta_n = \phi_n + (\theta_r - \phi_r) \frac{N_n}{M_n}.
\] (5.27)

Generally, we are concerned with the relative angular positions of the shafts, not their absolute positions. In this case, the angular positions of all components at time 0 are considered to be 0, and equation (5.27) becomes

\[
\theta_n = \theta_r \frac{N_n}{M_n}.
\] (5.28)

Using the relationship in equation (5.28), synchronisation of the signal averaging to any shaft can be done using a single positional reference signal by multiplying the (cumulative) reference angle by the shaft ratio defined by \( N_n \) divided by \( M_n \).

### 5.5.2 Rotationally Coherent Sampling

When the synchronous signal averaging procedure is applied to discrete signals (as opposed to continuous time signals), the sampling must be performed coherently with the rotation the component for which the average is being calculated. That is, the samples must be at equally spaced angular increments of the rotating component and at the same (modulo \( 2\pi \)) angular position for each rotation of the component.
Note that rotationally coherent sampling automatically implies synchronisation; if there are exactly \( N \) sample points per revolution of the component then, by definition, the component will rotate with a periodicity of exactly \( N \) points.

Two methods of providing rotationally coherent samples are in common use; phase-locked frequency multipliers and digital resampling.

### 5.5.2.1 Phase-locked frequency multiplication

In the early research in the application of synchronous signal averaging to gearbox vibration analysis, such as reported by Stewart [73] and McFadden [54], phase-locked frequency multipliers were used to generate sampling pulses from the shaft positional reference signal. These pulses were used to control the sampling process, with a discrete sample of the vibration signal being taken via an analogue-to-digital converter each time a sampling pulse was generated.

The phase-locked frequency multiplier uses a phase-locked loop to track the incoming reference signal and measure its instantaneous frequency. The frequency is multiplied by the desired ratio and a train of sampling pulses is output at the multiplied frequency.

There are some inherent limitations in phase-locked frequency multipliers:

a) they can be sensitive to noise and drop-outs, which can cause the phase-locked loop to lose lock; if this happens even momentarily the rotational coherency of the sampling pulses will be lost and the signal average corrupted,

b) to optimise the performance of the phase-locked frequency multipliers the internal filter characteristics need to be tuned to suit the expected input frequency and multiplication ratio; unless the system is limited to a small range of applications, this can add greatly to the complexity of the circuitry, and

c) the phase-locked frequency multiplier must inevitably lag behind the incoming reference signal; that is, there is some delay between a change in the frequency of the
reference and the corresponding change in the frequency of the output sampling pulses.

### 5.5.2.2 Digital Resampling Techniques

McFadden [58] showed that the inherent limitations of phase-locked frequency multipliers could be overcome by using digital resampling. In this technique, the vibration signal is sampled (via an analogue-to-digital converter) using conventional time based sampling and then digitally resampled at the required rotationally coherent sample points using interpolation of the discrete time signal.

In the technique used by McFadden [58], the angular position reference signal was sampled simultaneously with, and at the same rate as, the vibration signal. The resample points were calculated by determining the required angular position of the shaft of interest then converting this to an angular position of the reference shaft. For the sample point \( p \) in average number \( q \), where there are \( P \) points per revolution, the required angular position of the shaft of interest \( \theta_n \) and hence the required angular position of the reference shaft \( \theta_r \) (from equation (5.28)) is

\[
\theta_n = 2\pi \left( \frac{qP + p}{P} \right),
\]

\[
\theta_r = \theta_n \frac{M_n}{N_n} = 2\pi \left( \frac{qP + p}{P} \right) \frac{M_n}{N_n},
\]

(5.29)

The digital resampling can be viewed as a classic interpolation problem on a non-equidistant net, for which numerous numerical methods are available (for instance, see Dahlquist and Björck [26] Chapters 4 and 7): where \( x_0, x_1, \ldots, x_m \) are \((m+1)\) sampled angular positions of the reference and \( y_0, y_1, \ldots, y_m \) are the simultaneously sampled vibration data, we wish to define an interpolation function \( y = Q(x) \) giving vibration \((y)\) as a function of angular position \((x)\), with the restriction that
that is, the value of the function at any sampled angular position equals the vibration value sampled at that angular position.

McFadden [58] examined the use of sample and hold, linear and cubic interpolation; these can all be viewed as piecewise application of Lagrange’s interpolation formula [26],

\[ Q(x) = \varphi_k(x), \quad x_{k+n/2} \leq x < x_{k+n/2+1} \]

where \( \varphi_k(x) = \sum_{i=0}^{n} y_{k+i} \prod_{j=0}^{n} \frac{x - x_{k+j}}{x_{k+i} - x_{k+j}} \),

with the degree \( n \) of the interpolation polynomial \( \varphi_k \) being \( n=0 \) for sample and hold, \( n=1 \) for linear and \( n=3 \) for cubic interpolation.

McFadden [58] showed that these techniques behave in a similar fashion to low-pass filters, with the frequency response increasing with the order of the interpolating polynomial. Increasing the order of the interpolating polynomial also decreases the attenuation in the passband and produces smaller side lobes in the stopband. He also showed that some corruption of the signal could occur due to aliasing caused by replication of the sidelobes in the baseband; the effects of this ‘aliasing’ are generally reduced by the signal averaging process.

To demonstrate the effects of digital resampling, McFadden [58] used a synthesized test signal, with ratios based on the Wessex helicopter main rotor gearbox. The ‘tacho’ angular position reference signal was a sine wave at 400 hertz corresponding to the AC generator output. The vibration signal consisting of 128 sinusoids all of unit amplitude at 1 to 128 orders of a ‘shaft’ with a ratio of 399/3721 to the angular reference (corresponding to the ratio of the AC generator output to the input pinion rotational frequency). A ‘sampling rate’ of 20480 hertz was used.
For comparative purposes, a signal with the same characteristics has been used here (McFadden [58] did not indicate what the phase relationship of the sinusoids was; the reproduced signal used here has random phase for all sinusoids). The simulated data were ‘sampled’ using 12-bit resolution to simulate a 12-bit A/D converter (McFadden [58] did not indicate the sampling resolution used). Figure 5.6 shows the test signal after 1 average using cubic interpolation (note this is the same as that obtained by McFadden in reference [58], as would be expected). This shows the effects of attenuation due to roll-off in the frequency response of the interpolation technique and aliasing due to replication of the sidelobes. The amplitude at 127 orders is 0.91g (should be 1g) and the maximum level of the aliased vibration is 0.06g. Note that these effects were far more pronounced when using sample and hold or linear interpolation [58].

The spectrum of the test signal after 128 averages using cubic interpolation is shown in Figure 5.7. Note that the aliasing errors have been greatly reduced by the averaging process but that the roll-off in amplitude due to the frequency response limitations is still apparent.
5.5.2.3 Low Pass Filters

Succi [75] proposed that digital low pass filters be used for interpolation of the sampled vibration signal. The ideal low pass filter, passing all frequencies below the Nyquist frequency and rejecting all frequencies above the Nyquist frequency, can be used to perfectly reconstruct the original (continuous) time signal from the sampled data (Oppenheim and Schafer [61]):

\[
y(t) = \sum_{n=-\infty}^{\infty} y_n h(t-n),
\]

where \( h(t) = \frac{\sin(\pi t)}{\pi t} \).

However, this requires convolution of the sampled data with the filter response over an infinite number of samples.

Succi [75] proposed the use of a truncated version of the ideal filter which is multiplied by a window function to drive its value to zero at \(|t| = m\). The suggested filter was:
\[ h(t) = \left( \frac{\sin(\pi t)}{\pi t} \right) \frac{1}{4} \left( 1 + \cos\left( \frac{\pi t}{m} \right) \right)^2, \quad |t| \leq m. \] (5.33)

Succi [75] claimed that setting the filter length \( m \) to 10 would give a filter which would reliably reconstruct a signal up to one quarter the sampling frequency with an error of less than -82dB.

Figure 5.8 shows the results obtained for one average of the ‘McFadden’ test signal (see above) with digital resampling performed using the low pass filter in equation (5.33), with the filter length \( m \) set to 10. The dramatic improvement over cubic interpolation (Figure 5.6 and Figure 5.7) is easy to see; there is virtually no attenuation of the signal and no ‘aliasing’, even after only one ‘average’.

![Figure 5.8 Spectrum of test signal using low pass filter reconstruction (1 average)](image)

To see the errors in the signal reconstruction we need to look at the spectrum with a logarithmic amplitude scale.
Figure 5.9 Log spectrum of test signal using LP filter reconstruction (1 average)

Figure 5.9 shows the logarithmic amplitude spectrum of the signal reconstructed using the low pass filter after one ‘average’. All amplitudes in the stop band are below -82 dB, which supports the accuracy claimed by Succi [75] for the signal reconstruction using this filter. The amplitudes at some frequencies between 128 orders and the filter cut-off are greater than this (with the maximum ‘error’ being -64 dB) however, these are due to the dynamic range limitations in the original ‘sampled’ signal.

Although the low pass filter reconstruction provides dramatic improvement in the accuracy of the digital resampling (over cubic interpolation), this does come at a cost. Because the filter needs to be centred at the resample point, the coefficients of the filter need to be recalculated for each sample point (in addition to performing the convolution with the time sampled signal).

On the computer used for this study (a ‘slow’ 33 MHz 486), the calculation of 128 averages (with 1024 points per average) using cubic interpolation took 7 seconds to compute (including data input and output) and 112 seconds when interpolating with the low pass filter; 16 times longer. (McFadden [58] included timings for the same number of averages on a DEC LSI 11/73, circa 1986. Performing 128 averages with cubic interpolation on this machine took 431 seconds!)
To reduce the calculation time, the low pass filter coefficients can be pre-calculated and
stored over a fine mesh, with the required filter coefficients being interpolated from the
stored filter shape. This inevitably requires some trade off in accuracy for speed.

The convolution process itself cannot be performed using efficient methods such as the
overlap-add method because of the non-equidistant sample points.

5.5.2.4 Cubic Splines

Spline functions are another method of interpolating which can be used for the digital
resampling.

5.5.2.4.1 Derivation of a cubic spline

A spline function of degree \( n \) with nodes at the points \( x_i, i=0,1,...,m \) is defined [26] as a
function \( Q(x) \) with the properties:

a) on each subinterval \( [x_i, x_{i+1}] \), \( i=0,1,...,m-1 \), \( Q(x) \) is a polynomial of degree \( n \), and

b) \( Q(x) \) and is first (\( n - 1 \)) derivatives are continuous on the interval \( [x_0, x_m] \).

Note that it is the second property (continuity of the derivatives) which distinguish spline
functions from piecewise polynomials such as those used by McFadden [58].

A cubic spline function can be defined as

\[
Q(x) = q_i(x), \quad x_{i-1} < x \leq x_i \]

\[
q_i(x) = a_i + b_i z_i + c_i z_i^2 + d_i z_i^3, \quad \text{where } z_i = \frac{x - x_{i-1}}{x_i - x_{i-1}}, \quad (5.34)
\]

with the restrictions due to property (b) that
\( q_i(x_{i-1}) = y_{i-1}, \quad q_i(x_i) = y_i, \)
\( q'_i(x_{i-1}) = y'_{i-1}, \quad q'_i(x_i) = y'_i, \)  
and \( q''_i(x_i) = q''_{i+1}(x_i). \)  

Noting that the first derivative of the function \( q_i(x) \) (5.34) is

\[
q'_i(x) = b_i + 2c_i z_i + 3d_i z_i^2 \tag{5.36}
\]

and using the function values given in (5.35), we can solve for the terms \( a_i, b_i, c_i, \) and \( d_i: \)

\[
a_i = q_i(x_{i-1}) = y_{i-1}, \quad b_i = q'_i(x_{i-1}) = y'_{i-1},
q_i(x_i) = a_i + b_i + c_i + d_i = y_i, \quad q'_i(x_i) = b_i + 2c_i + 3d_i = y'_i,
\Rightarrow c_i = 3(y_i - y_{i-1}) - 2y'_{i-1} - y'_i, \quad d_i = y'_i + y'_{i-1} - 2(y_i - y_{i-1}),
\]
giving, after substitution into equation (5.34),

\[
q_i(x) = y_{i-1} + y'_{i-1}z_i + (3(y_i - y_{i-1}) - 2y'_{i-1} - y'_i)z_i^2
+ (y'_i + y'_{i-1} - 2(y_i - y_{i-1}))z_i^3. \tag{5.38}
\]

The second derivative of equation (5.38) is

\[
q''_i(x) = 2(3(y_i - y_{i-1}) - 2y'_{i-1} - y'_i) + 6(y'_i + y'_{i-1} - 2(y_i - y_{i-1}))z_i
= 6(y_i - y_{i-1})(1 - 2z_i) + y'_{i-1}(6z_i - 4) + y'_i(6z_i - 2) \tag{5.39}
\]

and this leads to a system of equations based on the continuity requirement for the second derivatives.
\[ q_i''(x_i) = q_{i+1}''(x_i) \]
\[ \Rightarrow 3(y_{i-1} - y_i) + y_{i-1}' + 2y_i' = 3(y_{i+1} - y_i) - 2y_i' - y_{i+1}' \]  \hspace{1cm} (5.40)
\[ \Rightarrow y_{i-1}' + 4y_i' + y_{i+1}' = 3(y_{i+1} - y_{i-1}). \]

The system of equations in (5.40) define a relationship between the first derivatives of the function and the function values which will ensure continuity of the spline and its first and second derivatives. However, this system of equations is only defined for \( i=1,2,\ldots,m-1. \) To define the spline function over the interval \([x_0, x_m]\) we need to define values for the function derivatives at the end points. This is can be done by setting the second derivatives at the end points to zero (i.e., assuming the spline to be straight beyond the end points), giving

\[ q_1''(x_0) = 3(y_1 - y_0) - 2y_0' - y_1' = 0 \]
\[ \Rightarrow 2y_0' + y_1' = 3(y_1 - y_0), \quad \text{and} \]
\[ q_m''(x_m) = 3(y_{m-1} - y_m) + y_{m-1}' + 2y_m' = 0 \]
\[ \Rightarrow y_{m-1}' + 2y_m' = 3(y_m - y_{m-1}). \]  \hspace{1cm} (5.41)

Equations (5.40) and (5.41) form a linear tridiagonal system of equations for determining the first derivatives \( y_i' \):

\[
\begin{bmatrix}
2 & 1 \\
1 & 4 & 1 \\
 & 1 & 4 & 1 \\
 & & \ddots & \ddots \\
 & & & 1 & 4 & 1 \\
 & & & & 1 & 2 \\
\end{bmatrix}
\begin{bmatrix}
y_0' \\
y_1' \\
y_2' \\
\vdots \\
y_{m-1}' \\
y_m' \\
\end{bmatrix}
= \begin{bmatrix}
y_1 - y_0 \\
y_2 - y_0 \\
y_3 - y_1 \\
\vdots \\
y_m - y_{m-2} \\
y_m - y_{m-1} \\
\end{bmatrix}
\]
\hspace{1cm} (5.42)

which can easily be solved using efficient methods for linear tridiagonal systems such as those described by Dahlquist and Björck [26] (Section 5.4.2 - ‘Band Matrices’).
5.5.2.4.2 Digital resampling using cubic splines

Using the sampled vibration data, $y_i$, equation (5.42) is solved giving the first derivatives at each sample point.

For each of the required resample points $x$ on the net $[x_0, x_m]$ defined by the sampled angular reference signal, the resampled vibration signal $y(x)$ is found using the polynomial $q_i(x)$ (5.38), where $x_{i-1} < x \leq x_i$:

$$z_i = \frac{x - x_{i-1}}{x_i - x_{i-1}}$$

$$y(x) = q_i(x) = y_{i-1} + y_{i-1}'z_i + \left(3\left(y_i - y_{i-1}\right) - 2y_{i-1}' - y'\right)z_i^2$$

$$+ \left(y_i' + y_{i-1}' - 2(y_i - y_{i-1})\right)z_i^3.$$ (5.43)

Figure 5.10 shows the results of using cubic spline interpolation on the test signal over one ‘average’. This shows substantial improvement over the result for (piecewise) cubic interpolation (Figure 5.6), however some attenuation and aliasing is still noticeable. The amplitude at 127 orders is 0.98g (should be 1g) and the maximum level of the aliased vibration is 0.02g.

![Spectrum T4371S1.VIB](image)

Figure 5.10 Spectrum of test signal using cubic spline interpolation (1 average)
The spectrum of the signal average after 128 averages using cubic spline interpolation is shown in Figure 5.11. As with cubic interpolation, the averaging process has attenuated the aliased signal, but has not affected the attenuation due to the roll off in frequency response.

Figure 5.11 Spectrum of test signal using cubic spline interpolation (128 average)

The calculation of 128 averages using cubic spline interpolation took 10 seconds as opposed to 7 seconds for cubic interpolation and 112 seconds using the low pass filter.

5.5.2.5 An alternative formulation of a cubic spline

Although the derivation of the cubic spline function given in Section 5.5.2.4.1 ensures continuity of the first and second derivatives, it does not necessarily give the correct values for the derivatives. An alternative method of calculating the function derivatives is using a differentiating filter (Oppenheim and Schafer[61]). In a similar fashion to the construction of the low pass filter in Section 5.5.2.3, it was decided to use a truncated form of the ideal differentiator:

\[ h(t) = \left( \frac{\cos(\pi t)}{t} - \frac{\sin(\pi t)}{\pi t} \right)\frac{1}{2} \left( 1 + \cos\left(\frac{\pi t}{m}\right) \right), \quad |t| \leq m. \tag{5.44} \]
Unlike the low pass interpolating filter, the differentiating filter can be implemented using efficient methods because the filter response is only required at the discrete time samples, not at the resampled points. To take advantage of this, the filter was made longer ($m=64$) and the window less severe (a Hann window was used rather than a squared cosine window). The filtering was performed using the overlap-add method (Oppenheim and Schafer [61]).

Using the function derivatives from the differentiating filter in place of those calculated using equation (5.42), the cubic spline interpolation takes place in exactly the same fashion as before, using equation (5.43).

Figure 5.12 shows the results after one ‘average’ using cubic spline interpolation with the function derivatives calculated using the differentiating filter. This shows a marked improvement over the classic cubic spline implementation (Figure 5.10). The attenuation due to roll off in frequency response and the aliased vibration are now barely discernible, although they still exist; the amplitude at 127 orders is 0.99g (should be 1.0) and the aliased vibration signal is 0.003g.

![Spectrum T4371F1.VIB](image)

*Figure 5.12 Spectrum of test signal using alternate cubic spline definition (1 average)*

For 128 averages using the alternate cubic spline, the processing time (including input, output and calculation of the derivatives) was 17 seconds.
5.5.2.6 Higher order spline interpolation

Although cubic splines are the most common spline functions used, there is no reason why higher order spline functions cannot be used.

A fifth order spline can be defined (in a similar fashion to the definition of the cubic spline in equation (5.34)) as

\[ Q(x) = q_i(x), \quad x_{i-1} < x \leq x_i \]

\[ q_i(x) = a_0 + a_1 z_i + a_2 z_i^2 + a_3 z_i^3 + a_4 z_i^4 + a_5 z_i^5, \quad \text{where } z_i = \frac{x - x_{i-1}}{x_i - x_{i-1}}. \quad \text{(5.45)} \]

Noting that the first and second derivatives of \( q_i(x) \) in (5.45) are,

\[ q_i'(x) = a_1 + 2a_2 z_i + 3a_3 z_i^2 + 4a_4 z_i^3 + 5a_5 z_i^4 \quad \text{and} \]

\[ q_i''(x) = 2a_2 + 6a_3 z_i + 12a_4 z_i^2 + 20a_5 z_i^3, \quad \text{(5.46)} \]

and remembering the continuity requirements,

\[ q_i(x_i) = q_{i+1}(x_i) = y_i, \quad q_i'(x_i) = q_{i+1}'(x_i) = y_i', \quad q_i''(x_i) = q_{i+1}''(x_i) = y_i'', \quad \text{(5.47)} \]

the value of the coefficients \( a_0, a_1, \) etc. become, after some manipulation,

\[ a_0 = y_{i-1}, \quad a_1 = y_{i-1}', \quad a_2 = \frac{1}{2} y_{i-1}'' , \]

\[ a_3 = 10(y_i - y_{i-1}) - 6y_{i-1}' - 4y_{i-1}'' - \frac{3}{2} y_{i-1}'' - \frac{1}{2} y_i'' \]

\[ a_4 = 15(y_{i-1} - y_i) + 8y_{i-1}' + 7y_i' + \frac{3}{2} y_{i-1}'' - y_i'' \]

\[ a_5 = 6(y_i - y_{i-1}) - 3y_{i-1}' - 3y_i' - \frac{1}{2} y_{i-1}'' + \frac{1}{2} y_i'' \quad \text{(5.48)} \]

Rather than solving the continuity equations (i.e., placing continuity restrictions on the third and fourth derivatives) to calculate the first and second derivatives, they can be calculated by direct differentiation of the sampled vibration signal using differentiating filters.
The differentiator defined in equation (5.44) is used to calculate the first derivative, with
a double differentiating filter used to calculate the second derivative. Note that the
second derivative is just the derivative of the first derivative, therefore the double
differentiating filter is simply the derivative of the differentiator defined in equation
(5.44).

Figure 5.13 shows the spectrum of the ‘McFadden’ test signal after one ‘average’ using
fifth order spline interpolation. There is virtually no attenuation of the signal and no
‘aliasing’ even with only one average. This signal, in the linear amplitude scale, is
indistinguishable from that obtained using the low pass filter reconstruction (Figure 5.8).
In order to see the differences between the two signals, we must look at the logarithmic
amplitude spectra.

![Spectrum T4371D1.VIB](image)

*Figure 5.13 Spectrum of test signal using fifth order spline interpolation (1 average)*

Figure 5.14 shows the logarithmic amplitude spectrum of the test signal after one
average using fifth order spline interpolation. When this is compared with the
logarithmic amplitude spectrum for the signal average using low pass filter
reconstruction (Figure 5.9), a number of interesting features are noticeable:
a) The two signals are identical up to approximately 207 orders (=8913 Hz = 0.87 x
filter cut-off). This confirms the earlier assumption that the signal between 128 orders
and half the original sampling frequency was due to ‘sampling’ errors in the original data and not the filter response.

b) Above 207 orders, the signal using the low-pass filter reconstruction begins to roll off whereas the signal using the fifth order spline interpolation remains basically ‘flat’ up to the filter cut-off (Nyquist frequency) and then rolls off steeply. This shows that the fifth order spline has a better frequency response and much sharper roll-off than the low pass filter reconstruction. This is due to the longer filter length used to define the differentiating filters used for the fifth order spline.

c) In the stop band (above the Nyquist frequency) the low pass filter reconstruction provides greater attenuation (-82 dB) than the fifth order spline interpolation (-75 dB).

![Log Spectrum T4371D1.VIB](image)

*Figure 5.14 Log spectrum of signal using fifth order spline interpolation (1 average)*

For 128 averages using the fifth order spline interpolation the calculation time (including input, output and calculation of the derivatives) was 25 seconds, as opposed to 112 seconds for the low pass filter technique.
5.5.2.7 Summary of digital resampling techniques

Table 5.1 gives a comparison of the performance of the digital resampling techniques discussed above. To a certain extent, it shows that digital resampling techniques conform to the old adage of “you get what you pay for”. However, consideration should be given to the limitations of the sampled vibration data when selecting a technique. Where we have data captured using a 12-bit AD converter (as is the case for the research reported in this thesis), which has a theoretical maximum dynamic range of 72 dB, there is little benefit to be gained in using the low pass filter technique as opposed to a fifth order spline, as they both introduce errors which are less than those present in the original data.

<table>
<thead>
<tr>
<th>Interpolation technique</th>
<th>Time for 128 averages (sec)</th>
<th>% error at 1/4 sampling rate</th>
<th>stop band attenuation (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>piecewise cubic</td>
<td>7</td>
<td>9</td>
<td>24</td>
</tr>
<tr>
<td>cubic spline</td>
<td>10</td>
<td>2</td>
<td>34</td>
</tr>
<tr>
<td>cubic spline using differentiating filter</td>
<td>17</td>
<td>0.1</td>
<td>42</td>
</tr>
<tr>
<td>fifth order spline</td>
<td>25</td>
<td>0</td>
<td>75</td>
</tr>
<tr>
<td>low pass filter</td>
<td>112</td>
<td>0</td>
<td>82</td>
</tr>
</tbody>
</table>

Table 5.1 Performance of digital resampling techniques

5.5.3 Tachometer (angular position) signal

Typically the angular position reference used is a cyclic signal (e.g., sine wave, pulse, square wave, etc.) which, for the purpose of digital resampling, needs to be converted into a cumulative angular position such that, at each sample point, it represents the total number of revolutions (including fractions of a revolution) since time t=0. Although not detailed in reference [58], McFadden used a simple zero-crossing detection algorithm to determine the zero phase points of the reference and linear interpolation between the zero phase points to determine the fractional angular positions at the intermediate samples.
The accuracy with which the ‘zero phase’ points in the angular position reference signal can be determined depends upon the wave shape of the signal and the rate at which it is sampled.

If the signal is a sine wave, the zero-crossing points (either positive or negative) can be used as the zero phase estimates. Interpolation of the time sampled values about a zero-crossing point can be used to increase the accuracy of the calculation of the zero-crossing locations. Blunt [5] showed that, using linear interpolation to determine the zero crossing points on a 400 hertz AC generator output of a Sea King helicopter sampled at 40 kHertz with a 12-bit analogue-to-digital convertor, the maximum error in the estimated zero crossing locations was less than 0.5 micro-seconds (with the major limiting factor being the quantization error of the analogue-to-digital convertor).

If the signal is a square wave or a pulse, the zero phase locations can only be determined to within half a sample point. Interpolation of the wave shape does not help as, theoretically, the transition occurs instantaneously. Linear interpolation will place the zero phase point at the centre of the time samples either side of the transition. Interpolation of higher orders should not be used on these types of tacho signals as they become unstable at the transition points and give spurious results. Care should be taken with pulsed tachometer signals to ensure that the sampling rate is high enough to detect the pulses (i.e., the time between samples must be less than the minimum pulse width).

Succi [75] modelled the influence of errors in the detection of the zero phase points and concluded that these could have a large effect on ‘leakage’ from one frequency to another in the resampled vibration signal. However, he incorrectly modelled the effect as a frequency domain process, assuming that the maximum phase error (equal to the phase over one time sample) could be directly translated into an equivalent error in frequency. This is true only over a single cycle of the tachometer signal. Note that the phase is the integral of the frequency and, for the assumed error in frequency to be maintained, the phase error would have to accumulate. This is not the case. The maximum error in phase over any number of cycles is equivalent to the phase over one time sample and the mean error in frequency is proportional to this phase error divided by the number of
cycles; that is, the mean frequency error tends to diminish as the monitoring time increases.

An alternative to time sampling the angular reference signal simultaneously with the vibration signal is to use a separate circuit to detect zero-crossing or pulse arrivals and to record the time at which they occur. This ‘arrival time’ needs to be directly related to the time sampling of the vibration signal to allow the angular position at each vibration sample to be calculated. This would normally require the same clock signal be used to drive the analogue-to-digital conversion and to time-stamp the zero-phase arrival times. Typically, A/D converters use a high speed clock, with each conversion taking place after a certain number of clock pulses, therefore, the zero phase event times could be recorded as a sample number plus a number of clock pulses. This type of zero phase detection method can effectively increase the sampling rate of the tachometer signal, however, it adds to the complexity of the analogue-to-digital conversion system.

Once the location of the zero phase points in the angular reference signal have been determined, the angular position at each vibration sample needs to be calculated. This is done by assigning an incremental angular value to each successive zero phase point (e.g., 0 at the first point, $2\pi$ at the second, $4\pi$ at the third, etc) and then interpolating the angular positions at each time sample using their relationship to the zero phase points. It is important to note that the interpolation performed here is not a signal reconstruction process, as was the case with digital resampling, but a signal approximation process.

We have relatively few points to work with and these points may have errors due to incorrect determination of the zero phase points. It is tempting to use a simple least squares approximation, however, by doing this, we would allow the phase to drift away from the zero phase points; although these may be in error, the maximum error at any one point will be the phase deviation over one time sample and we must constrain the approximation function to within this region.

The choice of approximation method will depend on the accuracy to which the zero phase points are known. If these have been determined to a high degree of accuracy (e.g., using very high initial sampling rates, external event timers, etc.) then a method
such as cubic splines can be used to ‘track’ the speed variations in the gearbox. However, if we use too high an order approximation where there are significant errors inherent in the estimated zero phase locations, overshoot of the approximation function may occur in the region between zero phase points. If the machine being monitored is a nominally constant speed machine, and any speed changes are approximately cyclic over the monitoring period, then using simple linear interpolation between zero phase points will divide the errors in the estimates of the zero phase locations equally over the period of rotation, giving an unbiased estimate of the reference phase at each sample point.
Chapter 6

HEICOPTER FLIGHT DATA

6.1 WESSEX HELICOPTER MAIN ROTOR GEARBOX

6.1.1 Historical background

From 1977 until the aircraft were removed from service in 1990, the Royal Australian Navy (RAN) carried out regular in-flight vibration recording on the main rotor gearboxes in their Wessex helicopters. This was performed as part of the routine condition monitoring procedures, which also included oil debris analysis. For each monitoring flight, an accelerometer was fitted to a bolt on the input housing of the main rotor gearbox, adjacent to input bevel pinion. Monitoring flights were performed at intervals of between 50 and 100 flight hours with the vibration signal from the accelerometer being recorded in-flight at a range of torque settings. The vibration signals were recorded on a 4 channel Brüel and Kjær 7003 FM tape recorder. In addition to the vibration signals, the aircraft communication (voice) channel and an attenuated version of the aircraft AC generator signal (for speed reference) were also recorded.

After each flight, spectral analysis (see Chapter 4) was performed on the vibration signal using a Hewlett-Packard HP3582A analyser.

6.1.2 Current condition of the tapes

Tapes from the Royal Australian Navy Wessex helicopter recorded tape vibration analysis program (RTVAP) were re-analysed using the strategy for quantifying signal components and optimising the number of averages described in Chapter 5 (Section 5.4). It was found that the vibration on the tapes had a high level of random noise. It is likely that these tapes have deteriorated over time (a number of ‘clicks’ and ‘pops’ were audible during playback).
An example of the assessment of the vibration on these tapes is given for a recording of Wessex Serial Number N7-215 main rotor gearbox number WAK143, recorded on 1 July 1983 (tape number 14/83). The gearbox had completed 263.4 hours since overhaul and the recording was made at a torque meter pressure reading of 400 psi (100% rated power).

An added complication in the analysis of this recording was that the speed reference signal (the AC generator output @ approx 400 hertz) had been overloaded, making it unusable. However, electrical interference had caused the 400 hertz signal to be superimposed on the vibration signal. This can be seen in the frequency spectrum of the taped vibration shown in Figure 6.1.

![Figure 6.1 Spectrum of Wessex MRGB WAK143 (263.4 Hrs TSO, 100%)](image)

A speed reference signal was ‘simulated’ by band pass filtering the vibration signal about the 400 hertz electrical interference signal (370 to 430 hertz). The first 40 cycles of this recovered signal is shown in Figure 6.2.
Using the ‘recovered’ speed reference signal, trial signal averages were made for all shafts in the gearbox. The major periodic components were the second harmonic of the input pinion meshing (3.4g), the tail drive meshing fundamental (1.96g), and the fifth harmonic of the epicyclic meshing (1.08g). These can be seen in the spectrum (Figure 6.1) at 1890, 1010 and 2500 hertz respectively. For the purposes of the ‘leakage’ calculation (see Section 5.4.3.2), quantification of periodic components was restricted to those over 0.17g RMS (5% of the second harmonic of input pinion meshing vibration). The 400 hertz component was also included (this corresponds to 3 times the generator shaft speed).

The random noise estimate (from equation (5.23)) was 4.85g RMS (to some extent, this noise can be seen in the ‘noise floor’ in Figure 6.1, particularly at the higher frequencies). The random noise dominates the signal-to-noise ratio calculations and, as a consequence, the predicted ‘leakage’ of non-synchronous vibration basically decreases proportional to the square root of the number of averages (i.e., as would be expected for purely random noise).

For the input pinion, the estimated RMS value of the synchronous signal was 3.55g. As a rough check of the estimated signal metrics, the synchronous signal average for the input pinion was calculated for 1 to 64 averages at power of two increments and the actual RMS of the signal averages check against the values predicted by the methods.
developed in Section 5.4. Note that this is only a ‘rough’ check because the prediction assumes that the synchronous signal amplitudes remain constant; this will not be entirely correct due to minor torque fluctuations. For comparison, the RMS values which would be expected if the non-synchronous vibration were purely random were also calculated. This was done by calculating the mean square difference between the signal average after one average and that after two averages, and assuming this to be half the mean square of the noise component. Classic methods assume reduction in the RMS of the noise by the square root of the number of averages (Braun [14]), that is, the mean square value reduces by the number of averages. This method gives an estimated RMS value for the noise of 5.42g and an RMS value for the synchronous signal of 3.33g.

Table 6.1 gives the predicted versus actual RMS values for the methods developed in Chapter 5 (‘new prediction method’) and for purely random noise (‘classic prediction method’) and the percentage errors in the predictions. The values for the classic method are not given for 1 and 2 averages as these were the values used in the calculations (i.e., they are exact but meaningless for evaluation purposes).

<table>
<thead>
<tr>
<th>Number of averages</th>
<th>Actual RMS</th>
<th>New prediction method</th>
<th>Classic prediction method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RMS</td>
<td>% error</td>
</tr>
<tr>
<td>1</td>
<td>6.3665</td>
<td>6.3612</td>
<td>-0.08</td>
</tr>
<tr>
<td>2</td>
<td>5.0820</td>
<td>5.0452</td>
<td>-0.7</td>
</tr>
<tr>
<td>4</td>
<td>4.3242</td>
<td>4.3172</td>
<td>-0.16</td>
</tr>
<tr>
<td>8</td>
<td>3.9322</td>
<td>3.9550</td>
<td>0.58</td>
</tr>
<tr>
<td>16</td>
<td>3.7135</td>
<td>3.7520</td>
<td>1.04</td>
</tr>
<tr>
<td>32</td>
<td>3.6365</td>
<td>3.6522</td>
<td>0.43</td>
</tr>
<tr>
<td>64</td>
<td>3.6085</td>
<td>3.6015</td>
<td>-0.19</td>
</tr>
</tbody>
</table>

*Table 6.1 Predicted vs. actual RMS values for Wessex input pinion averages.*

The maximum error in prediction using the new method is just over 1%. Note that the percentage error fluctuates, which may be due to a non-stationary signal caused by fluctuations in torque and/or speed. In contrast, the error in prediction assuming purely random noise gets progressively worse.

Because of the significant amount of random noise in the signal, the ‘optimum’ number of averages becomes ‘as many as possible’. Although approximately 30 seconds of data was recorded at each flight condition, it was difficult to find continuous sections of more
than 10 seconds duration in which significant audible ‘clicks’ and ‘pops’ were not present. Despite these limitations, synchronous signal averages with digital resampling using fifth order spline interpolation (see Section 5.5.2.6) were calculated from the tapes. Some existing signal averages, calculated by Dr. Peter McFadden at the Aeronautical Research Laboratory (ARL) in 1985/86 [58], were available for these tapes (at torque meter pressure readings of 400 psi). These were calculated on a DEC LSI 11/73 computer with digitally resampled using cubic interpolation over 400 revolutions of the input pinion (requiring approximately 9.3 seconds of data at the input pinion speed of 43 hertz). Because of the availability of these recordings (which can be used as a ‘standard’ against which to check the validity of the new signal averages) and considering the limited usable flight data on the tapes, it was decided to use 400 averages for the analyses discussed here.

6.1.3 Wessex input pinion crack

In December 1983, a Royal Australian Navy Wessex crashed in Bass Strait with the loss of two lives. Crash investigations attributed the accident to the catastrophic failure of the input spiral bevel pinion in the main rotor gearbox, caused by a fatigue crack which started at a sub-surface inclusion near the root of one of the teeth. After the crash, an investigation was carried out at ARL to determine why the RAN condition monitoring program had failed to detect the cracked gear prior to failure (McFadden [54]). McFadden [54] showed that the spectral analysis techniques used at the time of the accident were not sufficient to detect the crack.

Vibration recordings were made for this gearbox at 47.7, 133.4, 263.4 and 324.3 hours since overhaul, representing approximately 318, 233, 103, and 42 hours before failure respectively. For each test flight, recordings were made at torque meter pressure readings of 100, 200, 300, 400 and 440 psi (representing 25%, 50%, 75%, 100% and 110% of rated power respectively). The analysis for the recordings at 25, 75, 100 and 110% are shown here (except the 25% recording at 47.7 hours which was too badly corrupted by noise to give adequate signal averages).
Because there were only four recordings made for this gearbox from the time of overhaul to the time of failure, adequate parameters for the trending of this data cannot be established, therefore the various signal metrics will be given in tabular form.

6.1.3.1 Description of the failure

The Wessex input pinion is a thin rim spiral bevel gear which transmits approximately 1 MW at full load. The path of the crack in the failed Wessex input pinion studied here is shown in Figure 6.3.

![Figure 6.3 Path of crack in failed Wessex input pinion](image)

The crack started at a sub-surface inclusion near the root of one of the teeth and grew radially into the gear body and axially fore and aft toward both the open end of the gear and the neck simultaneously. When the crack reached the neck of the gear it changed direction and travelled circumferentially around the gear until final overload failure occurred, with the head of the gear totally separating, breaching the gearbox casing and destroying the main rotor control rods. With total loss of control, the aircraft crashed into the sea and sank rapidly; with one crew member and one passenger being drowned.
6.1.3.2 Synchronous Signal Averages

The synchronous signal averages for the input pinion were calculated with 400 averages using fifth order spline interpolation (see Section 5.5.2.6).

Figure 6.4 shows the angle domain representations of the input pinion signal averages at 100% rated power taken from recordings at (a) 318 hours before failure, (b) 233 hours before failure, (c) 103 hours before failure and (d) 42 hours before failure.

A change in the vibration signal, which is likely to be due to the crack, is discernible in the signal averages at 103 and 42 hours before failure at approximately 300 degrees in both (locations arrowed). Note that these signals are not phase aligned (i.e., the starting point of each signal is not necessarily at the same location on the pinion) and the similarities in the rotational offset to the disturbance in the two signals is coincidental.

![Figure 6.4 Signal averages at 100% rated power (angle domain)](image)

The frequency domain spectra of the same signal averages are shown in Figure 6.5.
Figure 6.5 Signal averages at 100% rated power (frequency domain)

The dominant frequency in all of the signal averages is at two times the tooth meshing frequency (44 orders). The fundamental of the tooth meshing (22 orders) is at a very low level and almost indiscernible after the first recording (at 318 hours before failure). This is unusual for a Wessex input pinion vibration signature where typically the fundamental and the third harmonic of the tooth meshing are the dominant frequencies. However, the difference in tooth meshing signature does not in itself constitute a fault, it merely indicates that this gear had different tooth profiles to most of the Wessex input pinions.

There is also a ‘ghost’ component present at 25 orders in all the signal averages. This is caused by a machining error (see Section 2.1.1.1.3) and was also noted by McFadden [54] in his investigation of this gearbox. The 25 order ‘ghost’ component is quite common in Wessex input pinion vibration, and does not constitute a fault.

The unusual meshing behaviour and the presence of a strong ghost component indicate that there may have been machining problems with the tooth profiles on this gear. This could have contributed to the failure (by overloading localised regions of the teeth) but, without the presence of the sub-surface inclusion to provide a crack initiation site, it is doubtful if these errors would have disrupted the correct operation of the gear.
6.1.3.3 Basic signal metrics

The ‘time domain’ (or in this case the angle domain) signal metrics of RMS, Crest Factor, and Kurtosis (described in Section 4.1.2) are shown in Table 6.2.

The results for the signal averages calculated in 1985/86 are shown in italics (for loads of 100% at 133.4, 263.4 and 324.3 hours since overhaul); the difference between these and the ‘new’ signal averages are not large, giving some confidence that, despite the suspected deterioration of the tapes, the signal averages are not significantly affected.

<table>
<thead>
<tr>
<th>Load (%)</th>
<th>Time since overhaul (hrs)</th>
<th>RMS</th>
<th>Crest Factor</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>133.4</td>
<td>4.5914</td>
<td>2.2640</td>
<td>1.9483</td>
</tr>
<tr>
<td>25</td>
<td>263.4</td>
<td>6.5374</td>
<td>2.0581</td>
<td>1.8113</td>
</tr>
<tr>
<td>25</td>
<td>324.3</td>
<td>4.9740</td>
<td>2.2718</td>
<td>2.0990</td>
</tr>
<tr>
<td>75</td>
<td>47.7</td>
<td>4.3421</td>
<td>2.1168</td>
<td>2.0096</td>
</tr>
<tr>
<td>75</td>
<td>133.4</td>
<td>3.9153</td>
<td>2.1552</td>
<td>1.8814</td>
</tr>
<tr>
<td>75</td>
<td>263.4</td>
<td>5.1931</td>
<td>1.9947</td>
<td>1.7928</td>
</tr>
<tr>
<td>75</td>
<td>324.3</td>
<td>4.6410</td>
<td>2.2506</td>
<td>1.9880</td>
</tr>
<tr>
<td>100</td>
<td>47.7</td>
<td>3.1791</td>
<td>2.4434</td>
<td>2.4286</td>
</tr>
<tr>
<td>100</td>
<td>133.4</td>
<td>2.9516</td>
<td>2.2594</td>
<td>2.3516</td>
</tr>
<tr>
<td>100</td>
<td>263.4</td>
<td>2.6920</td>
<td>2.2993</td>
<td>2.1394</td>
</tr>
<tr>
<td>100</td>
<td>324.3</td>
<td>3.6630</td>
<td>2.0130</td>
<td>1.7971</td>
</tr>
<tr>
<td>100</td>
<td>324.3</td>
<td>4.0747</td>
<td>2.3085</td>
<td>2.0141</td>
</tr>
<tr>
<td>100</td>
<td>324.3</td>
<td>4.0829</td>
<td>2.4153</td>
<td>2.0023</td>
</tr>
<tr>
<td>110</td>
<td>47.7</td>
<td>2.2388</td>
<td>2.7960</td>
<td>3.0403</td>
</tr>
<tr>
<td>110</td>
<td>133.4</td>
<td>2.7222</td>
<td>2.3058</td>
<td>2.3799</td>
</tr>
<tr>
<td>110</td>
<td>263.4</td>
<td>2.5389</td>
<td>2.2883</td>
<td>1.9287</td>
</tr>
<tr>
<td>110</td>
<td>324.3</td>
<td>2.9845</td>
<td>2.4342</td>
<td>2.2015</td>
</tr>
</tbody>
</table>

Table 6.2 ‘Time domain’ signal metrics for cracked Wessex input pinion

One concern is the RMS amplitude of the signal averages at 263.4 hours since overhaul. Note also that the ‘crest factor’ and ‘kurtosis’ for the signal averages at 263.4 are lower than those at 133.4 hours or 324.3 hours for all loads; this is also seen in the ‘old’ signal averages. It is known that two transducer locations were available for monitoring the Wessex input pinion (one on the port side of the input housing and one on the starboard side). Because of transmission path effects, the response to the input pinion vibration signal was different at the two locations. In the case of the Wessex input pinion, the
starboard side transducer location tended to give a ‘better’ response (i.e., the vibration levels at the mesh frequencies were generally higher, with a cleaner signal appearance). Later in the RTVAP, the transducer location was noted on the run sheets however, when these recordings were made (1983) this was not done. It was noted from the run sheets that the ‘operator’ who attached the transducers and took the recordings was a different person for the recording at 263.4 hours to the one who made the other recordings; it could be that different operators had different preferences for transducer location. Therefore there is no way of knowing if the changes in the signal are due to changes in mechanical condition, a different transducer location or a combination of the two.

The crest factors and kurtosis of the angle domain signal averages give no indication of the presence of a fault.

### 6.1.3.4 Stewart’s Figures of Merit

<table>
<thead>
<tr>
<th>Hours</th>
<th>Load (%)</th>
<th>FM0</th>
<th>FM1</th>
<th>FM4A</th>
<th>FM4B</th>
</tr>
</thead>
<tbody>
<tr>
<td>133.4</td>
<td>25</td>
<td>3.6740</td>
<td>0.3429</td>
<td>2.6992</td>
<td>0.3080</td>
</tr>
<tr>
<td>263.4</td>
<td>25</td>
<td>3.2911</td>
<td>0.2977</td>
<td>4.3498</td>
<td>0.2706</td>
</tr>
<tr>
<td>324.3</td>
<td>25</td>
<td>3.5877</td>
<td>0.6665</td>
<td>3.2812</td>
<td>0.4062</td>
</tr>
<tr>
<td>47.7</td>
<td>75</td>
<td>2.9369</td>
<td>0.2546</td>
<td>2.6609</td>
<td>0.2112</td>
</tr>
<tr>
<td>133.4</td>
<td>75</td>
<td>2.8735</td>
<td>0.2595</td>
<td>3.2278</td>
<td>0.2236</td>
</tr>
<tr>
<td>263.4</td>
<td>75</td>
<td>2.9529</td>
<td>0.3016</td>
<td>6.4536</td>
<td>0.2330</td>
</tr>
<tr>
<td>324.3</td>
<td>75</td>
<td>3.7023</td>
<td>0.7932</td>
<td>4.8527</td>
<td>0.4819</td>
</tr>
<tr>
<td>47.7</td>
<td>100</td>
<td>3.1674</td>
<td>0.2990</td>
<td>2.8682</td>
<td>0.2353</td>
</tr>
<tr>
<td>133.4</td>
<td>100</td>
<td>2.7924</td>
<td>0.3167</td>
<td>2.7930</td>
<td>0.2703</td>
</tr>
<tr>
<td>263.4</td>
<td>100</td>
<td>3.2536</td>
<td>0.4115</td>
<td>4.9502</td>
<td>0.2841</td>
</tr>
<tr>
<td>324.3</td>
<td>100</td>
<td>3.8087</td>
<td>0.7457</td>
<td>5.8773</td>
<td>0.4959</td>
</tr>
<tr>
<td>47.7</td>
<td>110</td>
<td>3.5932</td>
<td>0.3703</td>
<td>2.7693</td>
<td>0.3145</td>
</tr>
<tr>
<td>133.4</td>
<td>110</td>
<td>2.9285</td>
<td>0.3579</td>
<td>2.9287</td>
<td>0.2848</td>
</tr>
<tr>
<td>263.4</td>
<td>110</td>
<td>3.7023</td>
<td>0.5380</td>
<td>4.6777</td>
<td>0.3574</td>
</tr>
<tr>
<td>324.3</td>
<td>110</td>
<td>4.4213</td>
<td>0.9314</td>
<td>7.6385</td>
<td>0.5746</td>
</tr>
</tbody>
</table>

*Table 6.3 Stewart’s Figures of Merit for cracked Wessex input pinion*

Table 6.3 shows the results for Stewart’s Figures of Merit (described in Chapter 4) obtained from the signal averages for the cracked Wessex input pinion. The ‘regular’ signal for the bootstrap reconstruction figures of merit (FM4A and FM4B) was defined
by the tooth-meshing harmonics (multiples of 22 orders) plus the ‘ghost’ frequency at 25 orders.

In Table 6.3 the figures of merit which provide indication of the fault are highlighted in bold. Here the suggested error levels of FM0 > 5, FM1 > 0.5, FM4A > 4.5 and FM4B > 0.4 have been used to determine what constitutes an ‘indication’ of the fault.

The FM4A (kurtosis) indicates the fault at 263.4 hours (103 hours before failure) at loads of 75, 100 and 110% (but not at 25% although this is above the ‘warning’ level of 3.5) and at 324.3 hours (42 hours before failure) at all loads above 25%.

The FM1 and FM4B metrics both detect the fault at 324.3 hours (42 hours before failure) at all loads with FM1 also detecting the fault at 263.4 hours and 110% load. FM0 fails to detect the presence of the crack.

### 6.1.3.5 Narrow Band Envelope Analysis

The kurtosis of the narrow band envelope was calculated for bands of ±14 orders about the first four harmonics of the tooth-meshing frequency (i.e., 22, 44, 66, and 88 orders). In the band about 22 orders, the 25 order ‘ghost’ component was also eliminated. Table 6.4 shows the results of this analysis. The kurtosis values which are considered to give an indication of a fault (> 4.5) have been highlighted.

Band 2, which is centred around the highest tooth meshing harmonic at 44 orders, gives the best results, with the crack being indicated at 263.4 (103 hours before failure) at all loads and at 324.3 hours (42 hours before failure) at all loads except 25%. The values in the other bands are somewhat erratic (as would be expected as the tooth mesh harmonic at 44 orders dominates the signal). However, ‘band 1’ (about the fundamental tooth meshing frequency at 22 orders) does give a strong indication of ‘damage’ at 324.3 hours for all loads (except 25%) with the value increasing with load. Rather than being due to ‘modulation’ of the tooth meshing frequencies (which is very small in amplitude) this is most likely due to a resonance within the analysis band being excited by impacts;
as we would expect the energy at the impact to increase with load, the increase in the kurtosis value with load is not surprising.

<table>
<thead>
<tr>
<th>Hours</th>
<th>Load (%)</th>
<th>Band 1 Kurtosis</th>
<th>Band 2 Kurtosis</th>
<th>Band 3 Kurtosis</th>
<th>Band 4 Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>133.4</td>
<td>25</td>
<td>2.3231</td>
<td>2.3425</td>
<td>2.3502</td>
<td>2.2173</td>
</tr>
<tr>
<td>263.4</td>
<td>25</td>
<td>2.2545</td>
<td>7.0548</td>
<td>2.5037</td>
<td>2.5370</td>
</tr>
<tr>
<td>324.3</td>
<td>25</td>
<td>4.1921</td>
<td>3.6446</td>
<td>2.8626</td>
<td>5.6932</td>
</tr>
<tr>
<td>47.7</td>
<td>75</td>
<td>3.3623</td>
<td>2.3408</td>
<td>3.2654</td>
<td>2.0689</td>
</tr>
<tr>
<td>133.4</td>
<td>75</td>
<td>3.1712</td>
<td>2.2657</td>
<td>3.0506</td>
<td>5.0967</td>
</tr>
<tr>
<td>263.4</td>
<td>75</td>
<td>2.6935</td>
<td>10.6236</td>
<td>7.5586</td>
<td>3.8727</td>
</tr>
<tr>
<td>324.3</td>
<td>75</td>
<td>7.2212</td>
<td>5.4806</td>
<td>2.2275</td>
<td>3.9201</td>
</tr>
<tr>
<td>47.7</td>
<td>100</td>
<td>3.7390</td>
<td>3.4323</td>
<td>3.5415</td>
<td>4.3619</td>
</tr>
<tr>
<td>133.4</td>
<td>100</td>
<td>2.8149</td>
<td>2.4637</td>
<td>3.8315</td>
<td>3.2511</td>
</tr>
<tr>
<td>263.4</td>
<td>100</td>
<td>3.3178</td>
<td>6.9492</td>
<td>7.1186</td>
<td>3.1982</td>
</tr>
<tr>
<td>324.3</td>
<td>100</td>
<td>9.8941</td>
<td>4.8168</td>
<td>4.8989</td>
<td>4.8583</td>
</tr>
<tr>
<td>47.7</td>
<td>110</td>
<td>3.5234</td>
<td>2.8678</td>
<td>2.3870</td>
<td>2.9599</td>
</tr>
<tr>
<td>133.4</td>
<td>110</td>
<td>2.9933</td>
<td>3.2471</td>
<td>4.3478</td>
<td>3.9172</td>
</tr>
<tr>
<td>263.4</td>
<td>110</td>
<td>2.9874</td>
<td>5.9553</td>
<td>6.6300</td>
<td>3.0004</td>
</tr>
<tr>
<td>324.3</td>
<td>110</td>
<td>10.8238</td>
<td>6.3448</td>
<td>5.5397</td>
<td>2.5499</td>
</tr>
</tbody>
</table>

Table 6.4 Narrow Band Kurtosis values for cracked Wessex input pinion

The reduction in the ‘band 2’ kurtosis levels at all loads except 110% between 263.4 and 324.3 hours could be due to a number of factors, (1) the angle of rotation over which the crack disturbs (modulates) the tooth meshing harmonic has increased, thus broadening the peak in the envelope which will cause a drop in the kurtosis, (2) resonances within the analysis band which are excited by an impact have been superimposed on the modulation signal, or (3) a combination of both.

6.1.3.6 Narrow Band Demodulation

Figure 6.6 shows the demodulated amplitude and phase signals for the recordings at 263.4 hours since overhaul (103 hours before failure). The demodulation was performed using a narrow band about the second mesh harmonic (30 to 58 orders). The points of minimum phase are arrowed in each of the demodulated phase signals, with the corresponding points in the demodulated amplitude signals also marked.
Figure 6.6 Demodulated signals at 263.4 hours (103 hours before failure)

The signals at 75% (b), 100% (c) and 110% (d) are very similar, with the signal at a load of 25% (a) being distinctly different from the others. The low point in the demodulated phase diagrams at 75%, 100% and 110% is thought to be caused by a delay in the tooth adjacent to the crack taking up its full share of the load. This phase ‘dip’ gets progressively larger with load and there is a corresponding amplitude drop, with the minimum amplitude occurring just after the minimum phase point. The demodulated
signal at 25% (a) does not show these same features. It is likely that because of the low load the teeth are not fully ‘engaged’.

Figure 6.7 shows the signals obtained when the demodulation technique is applied to the signal averages for 75%, 100%, and 110% load at 324.3 hours since overhaul (42 hours before failure). In his study of this data McFadden [54] used the vibration signal at 75% load to demonstrate the demodulation technique (although this was not stated, it is obvious when his plots are compared to the ones shown here).

(a) 75% rated power: demodulated amplitude
(b) 100% rated power: demodulated amplitude
(c) 110% rated power: demodulated amplitude

Figure 6.7 Demodulated signals at 324.3 hours (42 hours before failure)

McFadden interpreted the discontinuity in phase seen in the demodulated phase signal in Figure 6.7 (a) as “....indicating that there is a complete reversal of phase at one location representing a loss of 360 degrees.” However, this seems highly unlikely in practice. At
the second harmonic of tooth meshing (44 orders) this would represent a delay of 50% of tooth spacing.

Note that as the demodulated amplitudes in Figure 6.7 approach zero, the corresponding demodulated phase shows a discontinuity. What happens if the amplitude becomes negative? The answer is that, theoretically, it cannot become negative. The demodulation process assumes that all of the signal in the analysis band is due to amplitude and/or phase modulation of a single frequency. To represent a signal whose amplitude becomes ‘negative’ its phase must shift by 180 degrees. This is exactly what is happening here.

6.1.3.7 A modified form of narrow band demodulation

We can very easily overcome the problem of the demodulated amplitude ‘trying’ to become negative by increasing the amplitude of the centre frequency.

Figure 6.8 shows the estimated amplitude and phase modulation signals obtained by doubling the amplitude of the centre frequency prior to demodulation and then subtracting the original centre frequency amplitude from the demodulated amplitude signal and doubling the demodulated phase signal.

Using this technique, the amplitude modulation signal can now become negative and the discontinuities in the phase modulation signals have disappeared. The amplitude modulation signals now appear similar for all loads, as was the case in Figure 6.6 for the signals at 263.4 hours. However, there are still some differences in the phase signals.

If we assume that the major phase dip seen in each of the demodulated phase signals in Figure 6.8 (arrowed) is caused by the same mechanism which produced the phase dips at 263.4 hours (Figure 6.6) the second of the two amplitude drops in each of the demodulated phase signals is also due to the same mechanism. The location of the minimum phase point has been arrowed in the demodulated amplitude signals and, referring back to Figure 6.6, it can be seen that the relationship of the second amplitude drop to the minimum phase point is the same as occurred at 263.4 hours in Figure 6.6.
Figure 6.8 Modified demodulated signals at 324.3 hours (42 hours before failure)

It now needs to be determined what is causing the additional amplitude drop. The distance (in degrees) between the two amplitude drops is marked on the demodulated amplitudes in Figure 6.8. This distance actually increases with load from 39 degrees at 75%, to 41.5 degrees at 100% to 44.3 degrees at 110%. Note also that the ‘peak’ between the two amplitude drops remains in a constant position in relation to the first drop with the location of the second amplitude drop being ‘delayed’ with an increase in load.

Measurement of a Wessex input pinion showed that the facewidth of the teeth is close to three times that of the tooth spacing at the pitch circle. Depending on load, each tooth may be in contact for between 2.5 and 3 times the tooth spacing distance. Converting the angular distances between the amplitude drops in Figure 6.8 to tooth spacings gives
2.4 teeth at 75%, 2.53 at 100% and 2.7 at 110%. Therefore it is possible that the two drops in amplitude occur whilst the same tooth is in contact.

Although it is difficult to establish the time history of the crack, it is possible that at the time of this recording (42 hours before failure and 61 hours since the previous recording in which cracking was evident) the crack has grown a considerable way both into the gear body and axially fore and aft. Considering the thin wall nature of this gear and the high loads transmitted, significant distortion of the gear body itself may be present in the region of the crack, causing a gross misalignment of the tooth adjacent to the crack as it comes into mesh. This would cause a considerable impact.

The impact as the tooth adjacent to the crack comes into mesh will excite structural resonances. The peak amplitude of the excited resonances is not related to, and may well exceed, the amplitude of the tooth meshing vibration. If one, or more, resonant frequencies are included in the analysis band they will be superimposed on the modulated tooth meshing harmonic, either adding to or subtracted from it depending on the instantaneous phase relationship. The frequency at which this additional ‘modulation’ occurs will depend on the frequency difference between the tooth meshing harmonic and the resonance.

We can now explain the observed demodulated amplitudes and phases seen in Figure 6.8.

a) There is a sizeable impact as the tooth adjacent to the crack comes into mesh. This excites structural resonances causing an apparent dip in the amplitude of the tooth meshing harmonic. In this case, the peak amplitude of the structural resonances is larger than the tooth meshing amplitude itself giving the appearance of a ‘negative’ amplitude. The effect of the impact on the phase is apparent in the region between the location of the impact and the main phase dip. This area is substantially different at different loads, with the differences becoming progressively less with distance from the impact point; this is due to damping of the structural resonance. The effect of the impact on the phase is far more complex than on amplitude (where it is predominantly additive), being the arctan of the ratio of the sum of the sines to the sum of the cosines of the individual vibration components.
b) Because of the higher amplitude of the tooth meshing harmonic at 75% load, and considering that the amplitude of the impact is likely to be less at the lower load, it is probable that the demodulated phase at 75% load is the closest to the actual phase modulation experienced by the meshing vibration. This tends to indicate that, in addition to the phase lag which was present at 103 hours before failure (Figure 6.6), there is now a preceding phase lead. One possible explanation for this is that it is due to the same distortion of the gear body which causes the tooth to impact as it enters mesh; as the tooth becomes fully engaged, it is pulled into alignment by the teeth on the mating gear.

c) There is a delay in the tooth adjacent to the crack taking up its full share of the load, causing an amplitude drop and phase lag in the tooth meshing vibration.

6.1.4 Wessex input pinion tooth pitting

The following example is for a Wessex input pinion with tooth pitting (from RAN Wessex N7-226, main rotor gearbox serial number WAK152). This gear was removed during 1987 at 490 hours since overhaul due to high iron wear debris in the oil. At the time of removal the gear showed pitting on three teeth.

Although synchronous signal averaging was being used on RAN Wessex input pinion vibration at this time to identify cracking, this particular pinion had been included in a group of ‘rogue’ gears which had high kurtosis values for the narrow band envelope enhancement but did not show a significant phase change in the narrow band demodulated signal (as would be expected for a crack). Subsequent investigation showed that the high kurtosis on all the other ‘rogue’ gears was due to the inclusion of the 25 order ‘ghost’ frequency in the analysis band; removal of the ‘ghost’ frequency reduced the kurtosis to an acceptable level for all gears except the example shown here. It is not clear why an investigation of the high kurtosis values was not performed at the time of these recordings.

Vibration recordings were made for this gearbox at 1.5, 27.7, 30.6, 38.9, 100.8, 152.3, 201.1, 248.9, 292.0 and 339.5 hours since overhaul. For each test flight recordings were
made at torque meter pressure readings of 100, 300 and 400 psi (representing 25%, 75% and 100% of rated power respectively). After each test flight, synchronous signal averages were performed for the input pinion for the 100% load condition (400 averages with digital resampling using cubic interpolation) and analysed using McFadden’s narrow-band enhancement and demodulation techniques (as described in Chapter 4). Analysis was not performed on the first and last recordings (at 1.5 and 339.5 hours since overhaul respectively) because the synchronising signal was not recorded. The synchronous signal average made at 152.3 hours since overhaul showed evidence of aliasing due to incorrect filter setting. No recordings were made between 339.5 hours since overhaul and removal of the gear at approximately 490 hours since overhaul because the aircraft was on a long term mission remote from the base.

An added complication in the analyses was that some of the recordings were made using the ‘port’ transducer location and others using the ‘starboard’ location; these two locations have different structural responses making comparison between recordings at the different locations invalid.

The recordings at 1.5, 30.6, 38.9, 100.8 and 152.3 hours since overhaul were made using the port side transducer location and the recordings at 27.7, 201.1, 248.9, 292.0 and 339.5 hours since overhaul were made using the starboard transducer location.

### 6.1.4.1 Synchronous Signal Averages

The synchronous signal averages made using cubic spline interpolation at the time of recording (for 100% load condition only) appear to have some inconsistency in scaling. This is thought to be due to incorrect input of the amplifier setting into the signal averaging program. The original signal averages for the recordings taken at 100% load from the starboard transducer at 27.7, 201.1, 248.9 and 292.0 hours since overhaul are shown in Figure 6.9. Note that the scales for these signals are not all the same, however it is thought that the scaling shown here compensates for the errors in scaling of the original signals (i.e., the signals can be viewed as if they were to the same scale).
Figure 6.9 Signal averages at 100% rated power (starboard transducer)

Discounting the differences in scale, very little can be deduced from the signal averages.

Figure 6.10 Signal average spectra at 100% rated power (starboard transducer)

Figure 6.10 shows the frequency domain spectra corresponding to the signal averages in Figure 6.9. Discounting the differences in scaling, it can be seen that there is an increase in the overall noise floor over time, especially between the tooth mesh fundamental (1x)
and its second harmonic (2x). There is also a slight increase in the amplitude of the tooth mesh fundamental frequency (1x) and a reduction in the amplitudes of the higher harmonics (2x, 3x, and 4x) with increased operating hours.

6.1.4.2 Basic signal metrics

Table 6.5 shows the basic angle domain signal metrics for the original signal averages.

<table>
<thead>
<tr>
<th>Time since overhaul (hrs)</th>
<th>RMS</th>
<th>Crest Factor</th>
<th>Kurtosis</th>
<th>Transducer Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.7</td>
<td>8.7683*</td>
<td>2.0297</td>
<td>2.3161</td>
<td>Starboard</td>
</tr>
<tr>
<td>30.6</td>
<td>1.7918</td>
<td>2.5744</td>
<td>2.4328</td>
<td>Port</td>
</tr>
<tr>
<td>38.9</td>
<td>1.6776</td>
<td>2.4494</td>
<td>2.8317</td>
<td>Port</td>
</tr>
<tr>
<td>100.8</td>
<td>2.1126</td>
<td>2.1300</td>
<td>2.0727</td>
<td>Port</td>
</tr>
<tr>
<td>201.1</td>
<td>9.3704</td>
<td>2.0178</td>
<td>2.1109</td>
<td>Starboard</td>
</tr>
<tr>
<td>248.9</td>
<td>8.3346</td>
<td>2.2548</td>
<td>2.0952</td>
<td>Starboard</td>
</tr>
<tr>
<td>292.0</td>
<td>9.3512*</td>
<td>2.2430</td>
<td>2.0394</td>
<td>Starboard</td>
</tr>
</tbody>
</table>

* the RMS values at 27.7 and 292 hours have been multiplied by 1.5 and 2 respectively to compensate for scaling errors (assumed) in the original averages.

Table 6.5 Basic signal metrics for pitted Wessex input pinion (100% load)

Apart from the obvious differences due to transducer location (particularly noticeable in the RMS levels), very little can be deduced from the basic signal metrics.

6.1.4.3 Enhanced signal metrics for pitted Wessex input pinion

Table 6.6 shows the results for Stewart’s Figures of Merit and the narrow band envelope kurtosis for the signal averages from the starboard side transducer at 100% of rated power.

<table>
<thead>
<tr>
<th>TSO (Hours)</th>
<th>Stewart’s Figures of Merit</th>
<th>Narrow band envelope kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FM0</td>
<td>FM1</td>
</tr>
<tr>
<td>27.7</td>
<td>2.5264</td>
<td>0.1755</td>
</tr>
<tr>
<td>201.1</td>
<td>2.5343</td>
<td>0.3372</td>
</tr>
<tr>
<td>248.9</td>
<td>2.9341</td>
<td>0.3716</td>
</tr>
<tr>
<td>292.0</td>
<td>3.1237</td>
<td>0.4208</td>
</tr>
</tbody>
</table>

Table 6.6 Enhanced signal metrics for pitted Wessex input pinion (starboard transducer @ 100% load)
The analysis parameters used are identical to those used for the cracked input pinion (see Sections 6.1.3.4 and 6.1.3.5). The values which exceed the suggested error levels (FM0 > 5, FM1 > 0.5, FM4A > 4.5, FM4B > 0.4 and narrow band kurtosis level > 4.5) have been highlighted. The values for the port side transducer location are not shown; as previously stated, direct comparison between the port and starboard side transducers is invalid and none of the enhanced signal metrics for the port side transducer recordings showed any indication of the fault (either due to the recordings being made prior to initiation of the fault or lack of sensitivity at the port transducer location).

Clear indication of the presence of the fault is given by the enhanced signal kurtosis values (FM4A and the first three narrow bands) at 201.1 hours since overhaul, which is approximately 290 hours prior to removal of the gear due to high iron wear debris. Because of the large time gap in recordings at the starboard transducer location, it is not clear when the initial detection of the fault could have been made. As the fault progresses, the kurtosis levels tend to decrease; this is probably due to a change in the signal from an isolated peak associated with pitting on one tooth to a more uniform signal as more teeth become pitted.

Although the FM0, FM1 and FM4B do not exceed their suggested error values, they all show an increasing trend with time. This suggests that, with sufficient trend data, these condition indices could be used to detect and monitor the progress of the fault. Unlike the kurtosis based signal metrics (FM4A and narrow band kurtosis values), these indices continue to increase as the damage spreads.

### 6.1.4.4 Narrow band demodulation

Figure 6.11 shows the demodulated amplitude and phase obtained using a narrow band of ±14 orders about the tooth meshing fundamental frequency (8 to 36 orders). This corresponds to ‘Band 1’ in Table 6.6. Note that the demodulated amplitude scales for the signals at 27.7 and 292 hours since overhaul are two-thirds and half those of the other signals respectively. This is due to the assumed error in scaling in these signals and should be disregarded; the demodulated phase signals are not affected by the scaling error.
The change in the demodulated amplitude and phase signals as the pitting develops is easy to see in Figure 6.11 (compare (a) at 27.7 and (b) at 201.1 hours since overhaul), however interpretation of these changes is not so simple.

Figure 6.11 Demodulated signals at 100% rated power (starboard transducer)

The demodulated amplitude and phase signals retains the same basic pattern from 201.1 hours (Figure 6.11(b)) onwards, with the magnitude of the signals increasing slightly
with time. These do not have the pronounced isolated change seen in the demodulated signals for the cracked input pinion, but rather a distributed sinusoidal oscillation (in both the demodulated amplitude and phase signals) which initially peaks and then dies away. Interpretation of these demodulated signals as ‘modulations’ of the tooth meshing harmonic (McFadden [55,56]) would result in the conclusion that the tooth meshing period and amplitude varies sinusoidally at a rate of approximately 13 or 14 cycles per revolution; this could possibly be caused by a perturbation of the gear about its axis.

However, referring back to the signal average spectra in Figure 6.10, a group of frequencies (characteristic of an excited structural resonance) is evident below the tooth meshing frequency (1x) and another more pronounced group between the tooth meshing frequency and its second harmonic (2x). A portion of both groups of frequencies is included in the narrow band used for demodulation and their relative amplitudes increase as the damage progresses. These ‘additive’ signals are simply treated as part of the modulation signal by the naive signal processing technique employed, and their effect on the demodulated signal magnitudes has no physical meaning beyond an indication of the relative magnitude of the excited resonance to that of the tooth meshing frequency.

Although it is probable that the major deviation seen in the demodulated amplitude (and phase) coincides (rotationally) with impulses caused by the pitted teeth, the ‘frequencies’ of the sinusoidal variations seen in the demodulated signals correspond not to any physical modulation frequencies but to the frequency difference between the tooth meshing frequency and the frequencies excited by the impulse(s).

### 6.1.4.5 Re-analysis of the tapes

For the purpose of this thesis, re-analysis of some of the tapes was performed using 400 averages with digital resampling by fifth order spline interpolation. The recordings at 1.5, 27.7 and 100.8 hours since overhaul could not be re-analysed due to poor tape quality, and the tapes for recordings at 152.3 and 248.9 hours since overhaul could not be located. Re-analysis of the tape at 339.5 hours since overhaul (which did not have a synchronising signal) was performed using the tooth meshing fundamental frequency as a
synchronising signal. This was done using a digital band pass filter (with a bandwidth of 80 hertz) about the tooth meshing frequency to generate a simulated ‘tacho’ signal.

Table 6.7 shows the results for Stewart’s Figures of Merit and the narrow band envelope kurtosis for the signal averages produced using fifth order spline interpolation from the starboard side transducer at 100% rated load.

<table>
<thead>
<tr>
<th>TSO Hours</th>
<th>Stewart’s Figures of Merit</th>
<th>Narrow band envelope kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FM0</td>
<td>FM1</td>
</tr>
<tr>
<td>201.1</td>
<td>2.3185</td>
<td>0.3256</td>
</tr>
<tr>
<td>292.0</td>
<td>3.1731</td>
<td>0.4097</td>
</tr>
<tr>
<td>339.5</td>
<td>3.0755</td>
<td>0.4467</td>
</tr>
</tbody>
</table>

*Table 6.7 Enhanced signal metrics for pitted Wessex input pinion (5th order spline interpolation, starboard transducer @ 100% load)*

Table 6.8 shows the results at 75% load. The results at 25% load were erratic (probably due to the teeth not being fully engaged) and are not shown here.

<table>
<thead>
<tr>
<th>TSO Hours</th>
<th>Stewart’s Figures of Merit</th>
<th>Narrow band envelope kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FM0</td>
<td>FM1</td>
</tr>
<tr>
<td>201.1</td>
<td>3.1555</td>
<td>0.4303</td>
</tr>
<tr>
<td>292.0</td>
<td>3.5139</td>
<td>0.4586</td>
</tr>
<tr>
<td>339.5</td>
<td>3.5244</td>
<td><strong>0.5128</strong></td>
</tr>
</tbody>
</table>

*Table 6.8 Enhanced signal metrics for pitted Wessex input pinion (5th order spline interpolation, starboard transducer @ 75% load)*

The results using the re-calculated signal averages at 201.1 and 292 hours since overhaul and 100% load (Table 6.7) are slightly different to those of the ‘original’ signal averages (see Table 6.6), particularly for the FM4A and narrow bands 3 and 4 kurtosis. These differences could possibly be due to the roll-off in frequency response caused by the cubic interpolation used in the ‘original’ signal averages.

The results at 75% load (Table 6.8) show that the response to the fault is generally better at the 75% load than it is at 100% load. However, the trend of the results at both loads is similar, with the ‘residual’ energy based measures (FM0, FM1 and FM4B) increasing
with progression of damage and the kurtosis based measures (FM4A and the narrow band envelope kurtosis) decreasing with the progression of damage.

Figure 6.12 shows the demodulated amplitude and phase at (a) 75% load and (b) 100% load for the re-calculated signal averages at 292 hours since overhaul. There is a slight decrease in the overall amplitude of the demodulated phase with an increase in load. This adds weight to the hypothesis that the major cause of the ‘modulation’ is the additive signal caused by excited resonances.

![Demodulated Amplitude and Phase](image)

Figure 6.12 Demodulated signals at 292 hours since overhaul (starboard transducer)

Figure 6.13 shows that the relative amplitudes of the excited resonances to the tooth mesh amplitude decrease with the increase in load from 75% to 100%, which reduces their effect on the demodulated signals.

Note that the amplitude of the re-calculate signal average at 292 hours since overhaul shown here is consistent with the amplitudes of the ‘original’ signal averages at 201.1 and 248.9 hours (see Figure 6.11(b) and (c)) but twice that of the ‘original’ signal average at 292 hours (Figure 6.11(d)). This confirms that there is an error in scaling in the ‘original’ signal average taken at 292 hours since overhaul.
Figure 6.13 Band limited spectra at 292 hours since overhaul (8-38 orders)

Figure 6.14 shows (a) the demodulated amplitude and (b) the demodulated phase for the signal average at 339.5 hours since overhaul and 100% load. These show the same basic pattern as the earlier signal averages (Figure 6.11 and Figure 6.12).

Figure 6.14 Demodulated signals at 339.5 hours since overhaul (100% load)

6.2 SUMMARY OF FINDINGS

In this chapter, two naturally occurring faults in operational helicopters have been examined using some of the existing signal enhancement techniques.

It has been shown that, although detection of both cracking and tooth pitting can be made using condition indices based on residual signal energies (Stewart’s Figures of Merit) and the kurtosis of the narrow band envelope, fault diagnosis using these techniques is difficult. The response of the condition indices to cracking and tooth pitting is similar in the early stages of damage. Detailed trending of the indices (and/or their relative values) over time may provide a clearer diagnosis however the limited number of samples available for the faults studied here cannot be used to confirm this. Also, in an operational situation, the safety of the aircraft is paramount and, because of the current limited knowledge of crack growth rates in gears, it may not be feasible to
monitor the progress of the fault long enough to gain a clear diagnosis. Although continued operation of the aircraft with a pitted tooth is not dangerous, it would be irresponsible to allow the aircraft to continue operation with a suspect cracked gear. In the case of the pitted Wessex input pinion, oil wear debris analysis did not indicate the pitting until approximately 290 hours after the first indication using vibration analysis, therefore oil analysis provided no diagnostic aid in the early stages of damage for these examples.

The narrow band demodulation technique did provide some diagnostic information in the early stages of damage, with the demodulated phase showing a relative large isolated ‘dip’ in the early stages of cracking and a smaller sinusoidal type variation in the early stages of tooth pitting. It was shown that the isolated phase ‘dip’ in the early stages of cracking is due to a modulation in the phase of the tooth meshing harmonic (as stated by McFadden [54]) however, the phase change in the early stages of tooth pitting was not due to modulation of the tooth meshing but to excited resonance(s) in the analysis band.

A modified form of the narrow band demodulation technique was developed and used to show that in the later stages of gear cracking the apparent phase reversal was not due to a loss of 360 degrees in the tooth meshing as suggested by McFadden [54] but by the addition of an excited resonance in the analysis band.
Chapter 7

EXPERIMENTAL GEAR RIG DATA

7.1 SPUR GEAR TEST RIG

In the previous chapter, in-flight vibration data was used to evaluate existing vibration analysis techniques. With the in-flight data we are limited by the time between recordings and the inability to monitor fault propagation over an extended period of time. Because of a need to maintain the airworthiness of the aircraft, we cannot continue operation with a known faulty component to see how the fault progresses.

In order to establish a correlation between actual fault growth and the various vibration analysis techniques, an experimental spur gear test rig was constructed; the aim being to grow realistic faults under controlled conditions and provide detailed measurement of the fault growth.

7.1.1 Description of test rig

During the test program, the rig went through a number of modifications. The final configuration is shown schematically in Figure 7.1.

The test gears were specially constructed ‘aircraft quality’ (AGMA Class 13 standard) spur gears made of case hardened EN36A steel with precision ground teeth; this is the same standard and material as used in the Wessex input pinion. A 35 mm wide 27 tooth input gear was used with a 13 mm wide 49 tooth driven gear.

The aim of the experiment was to initiate and propagate cracking in a gear tooth under normal operating conditions in order to correlate fault condition indices with realistic fault data.
To simulate a sub-surface inclusion (which was the initiation mechanism in the Wessex input pinion crack), a notch 1mm long, 0.1mm wide and 0.5mm radius deep was spark eroded into the root of one of the teeth (at the centre of the tooth face width), as shown in Figure 7.2. The notch was not cut across the full face width. Although this would have made crack initiation far easier to achieve, it would have initiated unrealistic crack
growth because the stiffness of the tooth would have been reduced in both the load
direction and in closure, which does not happen in practice.

Figure 7.2 Spark eroded notch in root of gear tooth

The rig was run continuously at 27 kW (rated load of the gear in accordance with
AGMA Class 13 standard) over eight hour periods, with the load being reduced to 21.6
kW for 20 minutes and then increased to 36 kW for 5 minutes prior to shut down. The
load variation was used in an attempt to place load marker striations on the crack as it
progressed to provide accurate correlation of crack growth versus time. After each eight
hour run, the gears were inspected for any signs of crack growth using a boroscope.

7.1.1.1 Monitoring equipment

Three accelerometers were mounted on the rig

- one on the waist of the gearbox (shown in Figure 7.1),
- one over the input bearing, and
- a high frequency accelerometer on the top of the gearbox.
It was found that the accelerometer location on the gearbox waist gave the best response. This was used for monitoring the gearbox during operation, with the output of the other accelerometers being recorded for future use if necessary.

A dedicated computer was set up to continuously monitor the vibration from the gear rig during operation. An optical shaft encoder, giving 1024 pulses per revolution, was used to provide ‘coherent’ sampling pulses to a Data Translation DT2821 analogue-to-digital converter in the analysis computer. The output of the shaft encoder was passed through a Schmitt trigger to give TTL (0-5Volt) compatible signals suitable for use as the DT2821’s external clock input. Output from the accelerometer on the waist of the gearbox was fed via a Wavetek 752A ‘brickwall’ low-pass filter set to five kHertz (to provide anti-aliasing) to the analogue-to-digital converter. An optical tachometer was used to provide a once per revolution synchronising pulse, which was used to trigger the start of data capture. This ensured that the signal averages produced would be synchronised with a known shaft position, and that the location of the cracked tooth in the signal averages could be determined.

Originally, the shaft encoder and optical tachometer were placed on the output shaft, to provide synchronisation and coherent sampling of the driven (output) gear which had the implanted stress riser. Because of the simplicity of the gearbox, the ‘ideal’ number of averages could be used to eliminate all vibration synchronous with the other shaft. This ‘ideal’ needs to be a multiple of 27 for the output shaft and 49 for the input shaft. Synchronous averages of the output shaft (135 averages = 5 x 27) and the input shaft (245 = 5 x 49 averages, using digital resampling of the vibration sampled coherently with the output shaft) were calculated and displayed every three minutes during the running of the rig, with the results being saved to disk approximately every 15 minutes.

During continuous running, the residual kurtosis value (FM4A), with the ‘regular’ signal for each gear being defined by the tooth meshing harmonics ± two shaft orders, was used as a ‘local’ fault shut down condition with a limit of 4.5, and the FM0 metric was used as a ‘general’ fault shut down condition with a limit of 2.5 (based on the mean value plus five standard deviations of recordings made during initial running). The FM4A was used as a ‘local’ fault indicator in preference to the narrow band envelope kurtosis for two
reasons. Firstly, a priori knowledge of the most appropriate ‘narrow band’ to use was not available and secondly, the regular disassembly and reassembly of the rig meant the balance and alignment of the shafts (which affect the once and twice per revolution modulations of the tooth meshing vibration respectively) could change slightly from run to run. The modulations due to imbalance and misalignment could be removed during the FM4 ‘bootstrap reconstruction’ without adversely affecting the process; theoretically, this could not be done with the narrow band envelope.

In addition to on-line monitoring using the computer, vibration data was recorded every twenty minutes using a Racal VHS format 14 channel FM tape recorder.

### 7.1.2 Tooth Pitting (Test Gear G3)

Crack initiation was not achieved after more than 200 hours of running. A reassessment of the test strategy was made and it was decided that overload conditions be used to initiate cracking, with crack propagation taking place under normal loads after initiation. A new gear set was used, with the output gear containing the spark eroded notch being reduced from 13mm to 10mm in width, and the operating load being increased from 27kW to 34kW, with ‘load markers’ of 27kW and 41kW.

Crack initiation had still not been achieved after 35 hours running with the new gear set (designated G3) running at 34kW. It was decided to increase the load to a constant 41 kW (maximum achievable with destroying the motor) with no load markers. After a further 18.5 hours (total run time of 53.5 hours) the motor was replaced with a 45kW electric motor and the test proceeded at a constant 45kW.

After a total of 107.9 hours running on the gear set (G3), the analysis system indicated a local fault on the input gear (i.e., the ‘undamaged’ gear). The rig was shut down and the gears removed for examination. This showed that pitting was present on one tooth of the input gear. The gearbox was reassembled and run for a further 16.5 hours at 45 kW to monitor the progress of the (naturally occurring) pitting.
Figure 7.3 shows the residual kurtosis values (FM4A) of the pitted gear over the last 23.5 hours of running. The disassembly and inspection points at 107.9, 110.9, and 115.5 hours are indicated on the graph, with the disassembly at 107.9 hours being a result of the FM4A value exceeding the shut down condition value of 4.5.

At 107.9 hours, inspection showed pitting on one tooth on the input pinion. Three hours later (110.9 hours) visual inspection showed that the pitting had increased slightly, with the pitted region being approximately 4 mm long and 1.5 mm wide. At 115.5 hours, the pitting had not progressed greatly, however initial pitting had started on two of the neighbouring teeth. A photo of the pitted teeth at this stage is shown in Figure 7.4. At 123.4 hours, destructive pitting similar to that on the first pitted tooth, was present on the two adjacent teeth and initial pitting was present on most of the teeth on the gear. At this stage the test was terminated.

The FM4A values plotted in Figure 7.3 can be explained in terms of the progress of the tooth surface damage for this gear;

a) as the pitting progresses on a single tooth, the FM4A values increase in proportion to the extent of damage (region between approximately 106 and 110 hours),
b) the value stabilises in the region between 110.9 and 115.5 hours, which corresponds to a reduction in growth rate of the pit on the initial tooth combined with commencement of pitting on neighbouring teeth, and

c) as the general condition of the gear deteriorates (from 115.5 to 123.4 hours), with more teeth developing pitting, the FM4A value decreases; instead of an isolated peak caused by a single tooth, the fault signature becomes more distributed resulting in a reduction in the kurtosis.

![Figure 7.4 Pitted gear teeth](image)

7.1.2.1 The synchronous signal averages

Figure 7.5 shows the (angle domain) signal averages at various stages during the test. The gearbox was disassembled and inspected just after the recording at 102.9 hours and no pitting was noticed on the teeth. The other three recordings are just prior to each inspection as detailed previously.

Note that because this gear was on the input shaft and the optical tacho was on the output shaft, the signal averages do not necessarily start at the same point on the gear (i.e., the 0 degree point on each of the signal averages in Figure 7.5 do not necessarily represent the same physical location on the gear). Very little can be made of these ‘raw’ signal averages.
Figure 7.5 Signal averages of test gear G3

(a) 102.9 hours - no pitting  
(b) 107.9 hours - one tooth pitted  
(c) 115.5 - three teeth pitted  
(d) 123.1 - three teeth badly pitted

Figure 7.6 Signal average spectra for test gear G3

(a) 102.9 hours - no pitting  
(b) 107.9 hours - one tooth pitted  
(c) 115.5 - three teeth pitted  
(d) 123.1 - three teeth badly pitted

Figure 7.6 shows the corresponding spectra for the angle domain signal averages in Figure 7.5. These tell us a little more about the effects of the damage. As a pit develops on a single tooth (b) there is a slight increase in the second harmonic of the tooth meshing frequency and also in the higher harmonics (5 and 6 times tooth mesh
frequency). There is also a slight increase in the overall ‘noise floor’. As the damage progresses to other teeth, the amplitude of the tooth mesh fundamental frequency increases, with the amplitude of the second harmonic reducing. The amplitudes at the 5th and 6th harmonic increase slightly.

Neither the angle domain representation (Figure 7.5) nor frequency domain representation (Figure 7.6) of the signal averages are very useful for diagnostic purposes.

### 7.1.2.2 Trends of signal metrics

Because of the large number of recordings taken during these tests (stored at regular intervals of approximately 15 minutes over a period of more than 120 hours), trending of the signal metrics can be used to provide more gear specific ‘limits’ than with the arbitrary values set on the non-dimensional metrics (such as Stewart’s figures of merit or the narrow band envelope kurtosis values).

For the purpose of the data shown here the ‘trends’ were developed based on the preceding eight hours of operation. All values were normalised by subtracting the mean of the trend and dividing by the standard deviation. This gives a value in ‘standard deviations’ from the mean. Typically, a value of ±5 standard deviations would be used as an indication of a fault condition.

Figure 7.7 shows the trended values for the RMS, Crest Factor (CF) and Kurtosis (K) of the ‘raw’ signal average. In the region of single tooth pitting (between approximately 107 and 111 hours) the RMS level drops initially then returns to its mean value and the Crest Factor peaks at a point roughly corresponding to the maximum point of single tooth damage. As the damage progresses to other teeth toward the end of the test, the RMS is slightly greater than the mean value for the undamaged gear, the Crest Factor is little changed from the mean and the Kurtosis has an increasing downward trend. The reduction in Kurtosis is probably due to the increasing dominance of the tooth mesh fundamental frequency (as seen in Figure 7.6).
None of these ‘raw’ signal metrics exceed five standard deviations, prior to the termination of this test.

7.1.2.3 Trended figures of merit

The results for the FM4A value was shown in Figure 7.3. Figure 7.8 shows the other Figure of Merit values (FM0, FM1, and FM4B) which have been trended in a similar fashion to the angle domain metrics above.

Neither the FM0 nor the FM1 values give an indication of the fault. Approximately midway through the progression of the single tooth damage, the FM4B increases to a maximum of seven standard deviations from the mean for the gear prior to damage. The FM4B value then decreases below five standard deviations as the damage spreads to other teeth.
7.1.2.4 Narrow band envelope kurtosis

Figure 7.9 shows the narrow band envelope kurtosis values for the first three ‘bands’, representing bands of ±14 orders about the first three tooth meshing harmonics.

For the reasons stated in the description of the rig monitoring system, the narrow band envelope kurtosis was not used as a shutdown criteria. However, Figure 7.9 shows that the kurtosis of the second band (40 to 68 orders) clearly indicates the presence of the fault one hour before the initial shutdown of the rig. The first band (13 to 41 orders) indicates the presence of the fault at about the same time as the FM4A value (Figure 7.3) with the third band (67 to 95 orders) not indicating the fault. In a similar fashion to the FM4A value, the first two narrow bands give high values in the region of single tooth damage and reduce as the damage progresses to multiple teeth.
Based on the spectrum of the undamaged gear in Figure 7.6 (a), it is probable that both the first and second ‘band’ would have been selected for use if the shut down criteria had been based on the narrow band envelope technique (due to their similarity in height and dominance over the other frequencies).

7.1.2.5 Narrow band demodulation

Figure 7.10 to Figure 7.13 show (a) the demodulated amplitude and (b) the demodulated phase of the signal averages at 102.9, 107.9, 115.5 and 123.1 hours. These correspond to the disassembly points for which the angle domain signal averages are shown in Figure 7.5 and spectra in Figure 7.6. The demodulation was performed using a pass band of ±14 orders about the second harmonic of the tooth mesh (i.e., 40 to 68 orders), corresponding to ‘band 2’ used for the narrow band envelope kurtosis values shown in Figure 7.9.

There a number of notable features in these plots. Firstly, the demodulated phase signal is consistent enough to allow the signals to be visually aligned (as previously mentioned, these signal averages are not physically aligned due to the synchronisation signal being on another shaft). There is a dip in the phase (arrowed) which can be identified even in
the first plot (Figure 7.10) for the ‘undamaged’ gear at 102.9 hours. It is possible that, rather than being a result of the damage, this dip initially indicates a tooth profile variation causing a delay in tooth engagement.
At the stage when pitting on one tooth was present (107.9 hours), a change in the amplitude modulation signal (Figure 7.11) is noticeable. Interestingly, this amplitude change precedes the dip in phase (location arrowed) suggesting that the phase dip is due to a delay in engagement of the tooth following the pitted tooth. As the pitting spreads to other teeth (Figure 7.12 at 115.5 hours and Figure 7.13 at 123.1 hours), the amplitude change increases (the overall downward shift in position of the vector is due to a decrease in the amplitude at the tooth meshing harmonic). The phase dip also increases, however there now a number of sinusoidal phase changes preceding the original dip and an increasing once per revolution variation in phase. This indicates that there is a larger variation in phase around the entire gear as the profiles on all the teeth become progressively damaged. The fact that the amplitude modulation remains localised and the phase modulation becomes more distributed suggests that it is the phase modulation which is responsible for the observed drop in kurtosis as the damage progresses.

7.1.3 Tooth Cracking (Test Gear G6)

For the next series of tests, it was decided to place the notch in the input gear instead of the output gear; the input gear, being smaller, undergoes more mesh cycles. The rig was redesigned to accept a narrower input gear (reduced from 35mm to 10mm) with the shaft encoder and optical tacho being switched from the output to the input shaft. In addition, the notch size was increased to 2mm in length, 0.1mm wide and 1mm radius deep.

With the third gear set (designated G6) installed the rig was run, after bedding in for 4.5 hours at 10kW, at a constant 37.5 kW for approximately eight hours a day with a reduction to 30 kW for 20 minutes followed by an increase to 45 kW for 5 minutes prior to shut down to add load markers. The rig was disassembled after each run with the gears being microscopically examined for signs of crack initiation.

After 30.5 hours of running with no signs of crack initiation, the rig was increased to a constant 45kW. Visual inspection after 8.5 hours at 45kW (total of 39 hours run time) showed no signs of cracking. After a further 3.5 hours at 45kW, the monitoring system
showed a dramatic jump in the local fault indices and the rig was shut down. Visual inspection showed a 2.75mm crack in the root of the spark eroded tooth. The gearbox was reassembled and run at a reduced load of 24.5 kW (AGMA Class 13 rated load for the narrowed input gear) in order to propagate the crack. After 0.5 hours at 24.5kW, the rig was disassembled and the crack remeasured (2.76mm). This procedure was repeated, with the rig being shut down after a further 12 minutes running at 24.5 kW due to a sudden change in the vibration signature. Visual inspection showed the crack had increased to 3.94 mm and showed a sharp change in direction. Metallurgically examination of the fracture surface failed to give any clear indication of propagation rates.

![Cracked gear tooth (Gear G6)](image)

**Figure 7.14** Cracked gear tooth (Gear G6)

A photo of the crack is shown in Figure 7.14. The crack length at the various disassembly points are marked on this photo. Note that the time of actual crack initiation
and propagation can only be estimated at less than 3.59 hours, which is the time between visual inspections showing no crack and visual inspection showing a crack length as indicated.

### 7.1.3.1 Crack initiation and propagation at 45 kW

Figure 7.15 shows the residual kurtosis values (FM4A) of the cracked gear during the period of running at 45 kW (12 hours). This shows the exponential increase over the last few minutes of the run which precipitated the initial shut down of the rig.

![FM4A Residual Kurtosis (Test Gear G6 @ 45 kW)](image)

*Figure 7.15 FM4A kurtosis values for test gear G6 (45 kW)*

### 7.1.3.2 The synchronous signal averages

Figure 7.16 shows the angle domain representation of signal averages for (a) the gear before cracking was present (at 33.2 hours) and (b) the gear with the cracked tooth (at 42.6 hours - the last recording before shutdown).

Both signal averages exhibit an obvious twice per revolution modulation. This is caused by misalignment of the gears which has been accentuated by the narrowing of the input
gear from 35 mm to 10mm (the same as that of the meshing output gear). Note that the shaft for this gear has an optical tacho and shaft encoder, therefore the two signal averages in Figure 7.16 are phase aligned (the 0 degree point on both represents the same physical location on the gear). The effects of the crack can be seen in the signal average (b) at 42.6 hours (crack location marked with an arrow) as a decrease in amplitude as the cracked tooth comes into mesh.

![Signal Averages](image)

(a) 33.2 hours - no cracking  
(b) 42.6 hours - cracked tooth

**Figure 7.16 Signal averages for test gear G6 (45 kW)**

The spectra of the signal averages seen in Figure 7.17 do not show much change in amplitude at the major tooth meshing harmonics. The one notable change is the appearance of a group of frequencies just above the fourth mesh harmonic. This represents a structural resonance being excited by an impulse.

![Spectra](image)

(a) 33.2 hours - no cracking  
(b) 42.6 hours - cracked tooth

**Figure 7.17 Signal average spectra for test gear G6 (45 kW)**

### 7.1.3.3 Trends of signal metrics

The trended angle domain metrics for the cracked gear running at 45 kW are shown in Figure 7.18. The trends were established over the first 4.5 hours of running at 45 kW.
The trended Crest Factor (CF) shows an increase as the crack develops however, none of the metrics deviate by more than 5 standard deviations from their mean values.

**Trended angle domain metrics (Test Gear G6 @ 45 kW)**

![Trended angle domain metrics](image)

*Figure 7.18 Trended angle domain metrics for test gear G6 (45 kW)*

### 7.1.3.4 Trended figures of merit

Figure 7.19 shows the trended values for the figures of merit FM0, FM1 and FM4B. The FM0 rises just above 5 standard deviations at the end of the run with the FM1 value showing no significant change. There is a very dramatic increasing trend for the FM4B value, which has a maximum value of more than 120 standard deviations at the end of the run. More significantly, the trended FM4B value first exceeds five standard deviations at 41.63 hours, which was more than one hour before shut down of the rig.

However, one must treat this result with a certain degree of caution. The vibration signal we are dealing with here is very ‘clean’, and the mean value of the FM4B ratio (which is the ratio of the ‘residual’ signal to that of the original signal average) is only 0.02 with a standard deviation of 0.0004. That is, only 2% of the vibration signal for the undamaged gear was due to variations in the tooth meshing behaviour and other
unexpected vibration and this did not vary significantly until tooth cracking occurred. The maximum value of this ratio was only 0.067. Typically, a value of 0.4 to 0.5 is used as a ‘warning’ level for the FM4B ratio in the absence of reliable trend data and in ‘real world’ gearboxes, which are subject to much greater variations in speed and load than the test rig, it would be expected that both the mean and standard deviation of the FM4B ratio would be much larger, with a corresponding decrease in sensitivity.

7.1.3.5 Narrow band envelope kurtosis

Figure 7.20 shows the narrow band envelope kurtosis for the first four ‘bands’, being ±14 orders about the first four tooth meshing harmonics.

Here we find that the narrow band about the tooth meshing fundamental does not indicate the presence of the crack. This is probably due to the second order sidebands seen about the tooth meshing frequency (Figure 7.17) ‘swamping’ the signal related to the crack. The detection performance tends to improve as we go to higher frequency bands.
The low amplitudes at the third and fourth harmonics of tooth meshing (see Figure 7.17), suggest that what is being detected with the narrow band envelope in bands three and four are not modulations of the tooth meshing but the structural response of the gearbox to impulses caused by the cracked tooth.

![Narrow Band Kurtosis Values (Test Gear G6 @ 45 kW)](image)

*Figure 7.20 Narrow band kurtosis values for test gear G6 (45 kW)*

### 7.1.3.6 Narrow band demodulation

Figure 7.21 and Figure 7.22 show (a) the demodulated amplitude and (b) the demodulated phase signals for the gear before tooth cracking (at 33.2 hours) and with tooth cracking (at 42.6 hours) respectively.

The changes in the demodulated signals before and after crack initiation are quite dramatic. Before crack initiation (Figure 7.21), both the amplitude and phase signals are dominated by a basically sinusoidal twice per revolution variation (due to misalignment). After crack initiation, a very obvious phase drop occurs (arrowed in Figure 7.22(b)) with a corresponding drop in amplitude (Figure 7.22(a)), with the minimum amplitude occurring just after the maximum phase deviation (as arrowed). These are both due to the reduction of stiffness in the cracked tooth. This is quite different to what was seen for the pitted tooth (Figure 7.11), where the phase lag occurred after the amplitude...
deviation and was probably due to a delay in engagement of the following tooth caused by deterioration of the tooth profile(s).

In additional to the twice per revolution variation due to misalignment and the changes during engagement of the cracked tooth, a sinusoidal variation of approximately 12 times per revolution in both the amplitude and phase signals can be seen in Figure 7.22. This is probably due to a resonance being excited by the impact caused by the cracked tooth. Notice that in both the amplitude and phase signals this is at a maximum just after engagement of the cracked tooth and dies away to a minimum just before the cracked tooth re-engages. This is not necessarily (and is unlikely to be) a physical ‘modulation’ of the tooth meshing at 12 orders but the effects of an additive signal (excited structural resonance) which is centred about a frequency approximately 12 orders away from the tooth meshing harmonic.

7.1.3.7 Crack propagation at 24.5 kW

After the initiation of the crack at 45 kW, with subsequent propagation due to delays in the detection of the crack and shut down of the rig, the crack was further propagated at
a load of 24.5 kW (100% rated load for the gear) over a total period of 0.7 hours. For each period of running at 24.5 kW, the load was ramped up slowly over a period of approximately ten minutes. This was to avoid a sudden application of load which may have snapped the tooth off.

The initial period of running at 24.5 kW produced very little crack growth (0.01 mm in half an hour). This was probably due to initially having to overcome the ‘stress relief’ at the crack tip caused by the propagation at 45 kW. Once the crack started to grow again, propagation proceeded rapidly, with a further 1.18 mm being added to the crack length in the final 12 minutes of the test (giving a total crack length of 3.94 mm). Signal averages were calculated, analysed and stored at approximately 30 second intervals during this period.

The termination of the test was prompted by a dramatic change in visual appearance of the signal average in the final 30 seconds, as shown in Figure 7.23.

![Figure 7.23 Signal averages for test gear G6 (final minute of test)](image)

It is thought that the change in the 26 seconds between Figure 7.23 (a) and (b) corresponds to the change in crack direction seen in Figure 7.14. This suggests that the final 0.3 mm or so of crack growth took approximately 30 seconds and that shutdown occurred within seconds of total loss of the tooth.

The FM4A kurtosis value, with the ‘regular’ signal defined by the tooth meshing harmonics plus their two upper and lower sidebands, showed a slow but steady decrease in value throughout the propagation at 24.5 kW (Figure 7.24 - FM4A+2). In the final 30 seconds, the value dropped suddenly. Also shown on Figure 7.24 is the FM4A kurtosis.
with the sidebands not included in the regular signal (FM4A+0). This behaves quite differently, showing a distinct increase over the last ten minutes of the test.

Figure 7.24 FM4A kurtosis values for test gear G6 (24.5 kW)

Including the two upper and lower sidebands of the tooth meshing harmonics in the regular signal (i.e., removing them from the residual), removes the modulation effects of misalignment and imbalance, which (initially) improves the detection of the additive signal due to the crack. However, this also removes the part of the structural response to the impact which occurs at these frequencies. As the crack progresses the impact energy increases and the contribution at the removed sidebands due to the crack becomes more significant than that due to misalignment and/or imbalance. At this stage, removal of the sidebands is effectively equivalent to ‘adding’ sinusoidal components (with the same amplitude as the original component but with a 180 degree phase shift) at these frequencies, which causes the kurtosis value to decrease.

It is thought that the higher, and more irregular, kurtosis values in the first five minutes or so running at 24.5 kW is due to the ‘stress relief’ caused by the initial crack propagation at 45 kW. In effect, the crack is trying to grow but the altered material properties at the crack tip make its behaviour spring like; the crack opens to a certain
extent and then springs back causing an impulse. As the crack progresses beyond this area, the tooth becomes more compliant, with a corresponding reduction in the impulse energy.

Because of the short period of time over which the rig was run at 24.5 kW, and because of the erratic signal behaviour in the first few minutes, trending of fault indices could not be carried out with any degree of confidence. All indices showed sudden changes in the last recording before shutdown. Apart from this, none of the angle domain metrics or figures of merit showed anything remarkable over the period of running at 24.5 kW, therefore they are not shown here.

The narrow band kurtosis values during the propagation at 24.5 kW are shown in Figure 7.25 (using the same ‘bands’ as used for the crack initiation and propagation at 45 kW). The bands about the third and fourth harmonics of tooth mesh show an increase toward the end of the test. As was the case for the initial tooth crack, these increases are probably not due to modulation of the tooth meshing harmonic but to the presence of excited resonances.

![Narrow band kurtosis values (Test Gear G6 @ 24.5 kW)](image)

*Figure 7.25 Narrow band envelope kurtosis for test gear G6 (24.5 kW)*
Figure 7.26 shows the demodulated amplitude and phase signals at (a) the start of the crack propagation at 25.4 kW and (b) 15 seconds before shutdown of the rig. These show the dominance of the resonance seen in Figure 7.22, but without the obvious amplitude drop and phase lag during the period of engagement of the cracked tooth. This indicates that, throughout this run, the major identifying fault signature is an impact which excites structural resonances. The only significant change, even to the point of imminent loss of the tooth, is in the amplitude of the impulse.

Figure 7.26 Demodulated signals for test gear G6 (24.5 kW)

7.2 SUMMARY OF FINDINGS

During this experimental program, two faults have been generated in case hardened aircraft quality gears and the correlation between the extent of the fault and various fault detection methods has been studied. This has shown that:

a) The analysis of the ‘raw’ signal averages does not provide significant diagnostic information with either tooth pitting or cracking.

b) The ‘Figures of Merit’ FM0 and FM1 did not detect the presence of either pitting or cracking.
c) The methods based on the ‘residual’ signal (or ‘bootstrap reconstruction’), FM4A kurtosis and FM4B energy ratio, provided indications of both tooth pitting and cracking. The FM4A kurtosis value responded to both the tooth crack and single tooth pitting with the value remaining high during crack propagation (although falling slightly as the crack progresses) and, for pitting, the value decreasing toward an ‘acceptable’ level as pitting spreads to multiple teeth. The FM4B (energy ratio) needs to be trended. For the very clean signal we get here using ‘ideal’ averaging of the high precision gears, the absolute value of the FM4B ratio is very low (approximately 0.02) and at no time does it reach the recommended warning level of 0.4. However, trending of this ratio provided the earliest warning of the cracked tooth. The FM4B response to tooth pitting was not as good, with detection only occurring in the region of maximum single tooth damage.

d) The narrow band envelope kurtosis provided detection of both the initial stages of single tooth pitting and tooth cracking. The band about the second harmonic of tooth mesh provided the earliest warning of single tooth pitting with the value decreasing towards an ‘acceptable’ level as pitting spreads to multiple teeth. With tooth cracking, the higher frequency bands tended to perform better than those about the major tooth meshing harmonics. The ‘detection’ of the tooth crack in this instance is thought to be caused by the presence of excited structural resonances in the analysis band and not the amplitude and phase modulation of the tooth meshing harmonics.

e) The narrow band demodulation technique provides the most detailed diagnostic information. The relationship between the phase and amplitude modulation signals showed a clear distinction between tooth pitting and cracking. In the later stages of cracking, the effects of excited resonances caused by impacts could clearly be seen.

At the initial stages of damage, narrow band demodulation gives the only clear distinction between single tooth pitting and tooth cracking. In safety critical systems, such as a helicopter transmission, this distinction can be vital. Note that after the initial stages of pitting, the gear was run for a further 16 hours at high load with no danger of loss of transmission capabilities. In contrast to this, the crack progressed rapidly after the initially detection and even with a reduction of load, total failure would have
occurred in less than one hour. Although the distinction between cracking and pitting can be inferred by monitoring the progress of other techniques over time, in a safety critical component this cannot be done (i.e., we cannot allow the damage to progress much beyond the initial fault indication in order to track the progress of the fault indices).

However, two problems exist in the implementation of the narrow band demodulation technique;

a) the need for expert interpretation of the results, and

b) the uncertainties involved in the selection of the most appropriate band to be used.
Chapter 8

TIME-FREQUENCY SIGNAL ANALYSIS

The vibration analysis techniques studied in the preceding chapters have been based on an assumed model of vibration from geared transmission systems. Fault detection and diagnosis has been performed using various measures of the deviation in the actual vibration signal from the expected vibration signal for a particular gearbox. Although there is some validity in this approach, the assumptions made about the nature of the vibration signal can cause problems in both the detection and diagnosis of faults.

A more general approach to the analysis of signals will now be investigated. The aim being to identify the various signal components without making prior assumptions of what these components should (or should not) be. This approach does not preclude forward modelling however, the matching of signal components to expected vibration characteristics is limited to a post-processing step to avoid constraining the analysis to assumptions made in the model.

In this chapter, the theory of joint time-frequency domain signal analysis will be introduced and a number of time-frequency signal analysis methods will be examined. Note that for consistency with the work of others in signal analysis, the domain variables ‘time’ and ‘frequency’ will be used to describe the behaviour of signals in the following discussions. However, the signal analysis methods discussed are directly applicable to the ‘angle domain’ synchronous signal averages with the rotation ‘angle’ being analogous to ‘time’ and, although the theoretical description of techniques will be based on time, practical examples will be based on angle domain signal representations.

In later chapters, the methods discussed here will be expanded upon and their use in the analysis of vibration signals will be investigated.
8.1 GENERAL SIGNAL REPRESENTATION

Commonly, vibration signals are represented in either the time domain or the frequency domain (see Appendix B). The time domain representation of a signal,

\[ s(t) = a(t)e^{j\phi(t)} = a(t)e^{j\left(\phi_0 + 2\pi \int_0^t f_i(\tau)d\tau\right)} \]  

allows simple characterisation of a signal in terms of its (instantaneous) energy

\[ E_i(t) = a(t)^2 = |s(t)|^2 = s(t)s^*(t) \]  

and instantaneous frequency

\[ f_i(t) = \frac{1}{2\pi} \frac{d(\phi(t))}{dt} \]

that is, at time \( t \) the signal has an energy density of \( E_i(t) \) at a frequency of \( f_i(t) \). However, this only has meaning for monocomponent signals, that is, a signal which only has one frequency at any instance of time.

Figure 8.1 Sinusoidal amplitude (4 per rev) and frequency (2 per rev) modulated signal

Figure 8.1 shows (a) the amplitude and (b) the instantaneous frequency of a (monocomponent) signal with an 88 order ‘carrier’ frequency of amplitude 1g, which has a four times per revolution sinusoidal amplitude modulated of ±0.5 and a twice per revolution sinusoidal frequency modulation with a maximum frequency deviation of ±5 orders. Figure 8.1 shows that the amplitude (and hence the energy density) and
instantaneous frequency can be accurately derived from this monocomponent time (angle) domain signal.

For multicomponent signals (i.e., signals which have more than one frequency at a given instance of time) the ‘instantaneous frequency’ is the average frequency of the signal at that time (Cohen and Lee [20]), however this in itself has very little meaning.

Figure 8.2 Multicomponent modulated signal

Figure 8.2 shows (a) the amplitude and (b) instantaneous frequency of the signal in Figure 8.1 with the additional of an unmodulated 40 order sine wave with an amplitude of 1g. The amplitude and instantaneous frequency of this multicomponent signal have little meaning as descriptors of the signal behaviour.

The frequency domain representation of a signal,

\[ S(f) = A(f)e^{j\theta(f)} = \int s(t)e^{-j2\pi ft} dt \]  \hspace{1cm} (8.4)

gives a perfect representation of a signal which consists of multiple simple harmonic oscillators (i.e., with no amplitude or frequency modulation). However, it is shown in Appendix B that this is not an adequate representation of non-stationary signals (i.e., signals whose frequency content change with time).

Note that the frequency domain representation of a signal can also be used to characterise signals which are continuous and monocomponent in frequency (in a similar fashion to continuous monocomponent signals in the time domain) by their instantaneous energy

\[ E_f(f) = A(f)^2 = |S(f)|^2 = S(f)S^*(f) \]  \hspace{1cm} (8.5)
and group-delay \[11\]
\[
\tau_g(f) = -\frac{1}{2\pi} \frac{d(\theta(f))}{df}.
\] \tag{8.6}

Therefore, we can adequately represent monocomponent non-stationary signals in the time domain or multicomponent stationary signals in the frequency domain but how do we represent multicomponent non-stationary signals?

A multicomponent non-stationary signal can be described as the superposition of a number of monocomponent non-stationary signals (8.1), giving
\[
s(t) = \sum_c s_c(t) = \sum_c a_c(t) e^{j\phi_c(t)} = \sum_c a_c(t) e^{j(\phi_c + 2\pi \int_0^t f_c(\tau) d\tau)}.
\] \tag{8.7}

In order to decompose (and understand) a signal of the form given in (8.7) a joint time-frequency domain representation of the signal is required.

### 8.2 TIME-FREQUENCY DOMAIN REPRESENTATIONS

#### 8.2.1 Short-time Fourier transform and spectrogram

The short-time Fourier transform (STFT) and its energy density spectrum (spectrogram) have been the most widely used time-frequency signal analysis tools \cite{12},\cite{22},\cite{24}. The STFT is the natural consequence of an intuitive approach to time-frequency analysis. To define the frequency behaviour of a signal at a particular time, a small section of the signal centred about the time of interest is ‘extracted’ and Fourier transformed to give an estimate of the frequency content of the signal at that time.

To achieve this, at each fixed time of interest \(t\) the time domain signal is multiplied by a moving (or ‘sliding’) window, \(h(\tau - t)\), which emphasizes the signal centred at time \(t\), and the Fourier transform of the resultant windowed signal is calculated \cite{24}, giving the ‘short-time’ Fourier transform (STFT) at time \(t\):
\[ S_t(f) = \int s(\tau)h(\tau - t)e^{-j2\pi ft} \, d\tau. \]  
\[ \text{(8.8)} \]

The spectrogram is the magnitude squared of the STFT,

\[ \rho_s(t, f) = |S_t(f)|^2. \]
\[ \text{(8.9)} \]

### 8.2.1.1 Multicomponent signals

For the multicomponent non-stationary signal defined by equation (8.7), the STFT is

\[ S_t(f) = \int \left[ \sum_c s_c(\tau) \right] h(\tau - t)e^{-j2\pi ft} \, d\tau \]
\[ = \sum_c \int s_c(\tau) h(\tau - t)e^{-j2\pi ft} \, d\tau \]
\[ = \sum_c S_{c,t}(f) \]
\[ \text{(8.10)} \]

that is, the STFT of the sum of the signal components is equivalent to the sum of the STFTs of the individual components. The STFT of an individual component is

\[ S_{c,t}(f) = \int s_c(\tau) h(\tau - t)e^{-j2\pi ft} \, d\tau \]
\[ = \int a_c(\tau)e^{j\phi_c(\tau)} h(\tau - t)e^{-j2\pi ft} \, d\tau \]
\[ = S_c(f)*[H(f)e^{-j2\pi ft}]. \]
\[ \text{(8.11)} \]

It is assumed that the windowed signal at each fixed time of interest \( t \) approximates a stationary signal with the amplitude and frequency of the individual components being the instantaneous amplitude and frequency of the component at time \( t \), giving

\[ S_t(f) \approx \sum_c \left[ \int a_c e^{j\phi_c + 2\pi f_c \tau} h(\tau - t)e^{-j2\pi ft} \, d\tau \right] \]
\[ \approx H(f)*\left[ \sum_c a_c e^{j\phi_c} \delta(f - f_c) \right]. \]
\[ \text{(8.12)} \]
where the subscripted constants are equivalent to their time-based values (i.e., $a_c = a_c(t)$, etc.).

### 8.2.1.2 Bandwidth-time limitation

As the time duration of the window function $h(\tau)$ decreases, the approximation to the instantaneous amplitude and frequency of the individual components improves however, as a consequence of the bandwidth-time limitation (see Chapter 4), the narrowing of the window function in time causes its spectrum to become wider in frequency which results in loss of frequency resolution in the STFT (due to the frequency domain convolution with the spectrum of the window function $H(f)$ seen in equation (8.12)).

Although it is often stated that the trade off between time and frequency resolution described above makes it impossible to set up a true time-frequency distribution, Cohen [22] argued that this is an artificial limitation introduced by the application of the window function in the short-time Fourier transform. The inherent limitation in resolution is imposed by the duration of the signal itself and, theoretically, if the duration of the signal is infinite in time the signal will be continuous in frequency and vice versa. There is no reason to assume that the resolution in joint time-frequency should be inherently less than those for the single domain cases.

### 8.2.1.3 Window functions

The window function should be real valued and symmetrical about $\tau=0$ to avoid time shifts. Window shapes such as rectangular, Hann (often called Hanning), Hamming, etc. are truncated in the time domain (i.e., their values are zero at the extremities of the window and beyond) which causes ripples and negative amplitudes in their Fourier transforms. The Gaussian (or exponential) window (Randall [67]) is not truncated in the time domain, with its value approaching zero at infinity, and its Fourier transform is a Gaussian window in the frequency domain which has no ripples or negative values. This makes it a convenient choice as the windowing function in the STFT. A Gaussian window is typically defined as (Randall [67])
\[ h_G(\tau) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\tau^2/(2\sigma^2)}, \] (8.13)

which derives from the probability density curve of a normally distributed random signal with a mean value of zero, standard deviation \( \sigma \) and integral value in \( \tau \) of unity. This can be rewritten as

\[ h_G(\tau) = \frac{2^{\ln(\lambda)}}{T_\lambda \sqrt{\pi}} e^{-\tau^2 \ln(\lambda)/(T_\lambda/2)^2} \] (8.14)

where \( T_\lambda \) is the distance in \( \tau \) between the points on the curve at which the value of the window function is \( 1/\lambda \) times the maximum value (which occurs at \( \tau = 0 \)). For example, setting \( \lambda = 100 \) will give a logarithmic ratio of -40dB (20 \( \log_{10}(1/100) \)) at the points \( \tau = \pm T_\lambda/2 \).

The Fourier transform of the Gaussian window function defined in equation (8.14) is (Magnus and Oberhettinger [45])

\[ H_G(f) = \int \frac{2^{\ln(\lambda)}}{T_\lambda \sqrt{\pi}} e^{-\tau^2 \ln(\lambda)/(T_\lambda/2)^2} e^{-j2\pi f \tau} d\tau \]
\[ = e^{-(\pi T_\lambda f)^2/(4 \ln(\lambda))}. \] (8.15)

8.2.1.4 Discrete form

Let \( \hat{s}(t) \) be a sampled version of the continuous signal \( s(t) \) with \( N \) samples at a sample interval of \( T \). The sampled signal can be expressed as

\[ \hat{s}(t) = \sum_{k=0}^{N-1} s(kT) \delta(t - kT) \] (8.16)

and the STFT (8.8) of the sampled signal is
\[ S_f(f) = \int \sum_{k=0}^{N-1} s(kT)\delta(\tau - kT)h(\tau - t)e^{-j2\pi f\tau}d\tau \]

\[ = \frac{1}{NT} \sum_{k=0}^{N-1} s(kT)h(kT - t)e^{-j2\pi kT}. \] 

(8.17)

The restriction of the sampled signal to the period NT causes the STFT to become
discrete in frequency, with the interval in frequency being 1/NT (Randall [67]). The
discrete frequency nature can be expressed by substituting m/NT for f in equation (8.17)
giving

\[ S_f\left(\frac{m}{NT}\right) = \frac{1}{NT} \sum_{k=0}^{N-1} s(kT)h(kT - t)e^{-j2\pi mk/N}. \] 

(8.18)

Changing all functions to their discrete form, with the sample intervals being implied, and
the time of interest \( t \) being replaced by the sample of interest \( n \) gives the discrete STFT,

\[ \hat{S}_n(m) = \frac{1}{N} \sum_{k=0}^{N-1} \hat{s}(k)\hat{h}(k - n)e^{-j2\pi mk/N} \] 

(8.19)

and the discrete spectrogram is the magnitude squared of the discrete STFT:

\[ \hat{\rho}_s(n, m) = \left| \hat{S}_n(m) \right|^2. \] 

(8.20)

### 8.2.1.5 Visual representation

The spectrogram (and other time-frequency distributions discussed in this thesis) is a
function of energy density versus time and frequency and can be visualised as a surface in
three-dimensional space. These can be displayed in a number of ways including contour
maps, waterfall plots, grey-scale images or colour images.

In this thesis, colour images have been used with frequency on the horizontal axis, time
(or angle) on the vertical axis and energy density being represented by colour ranging
over the visible light spectrum from blue to red. The energy is relative to the wavelength
of light at each colour, that is red represents high energy levels and blue represents low
energy. Unless otherwise stated, the energy will be a logarithmic ratio of the maximum energy value in the distribution (i.e., 0dB represents the maximum value in the distribution, -20dB is a tenth of the maximum value, -40dB is one hundredth of the maximum value, etc.). Energy levels below a specified minimum value are not displayed.

This self-relative logarithmic scaling has been used to maximise the visual spectrum for each plot in order to gain a clear visual representation of the signal behaviour; the patterns representing the internal structure of the signal are considered more important in this respect than the absolute value of energy density at a given time-frequency location.

Figure 8.3 shows spectrograms of a frequency modulated sine wave with a ‘carrier frequency’ of 88 orders with a twice per revolution frequency modulation of ±5 orders (the frequency domain spectrum of this signal is shown in Appendix B, Figure B.3(a)). Both plots have ‘energy’ levels ranging from 0dB (maximum energy) to -40dB (0.01 times maximum energy). Figure 8.3(a) shows the spectrogram obtained using a window length of 63.28 degrees at the -40dB point of the window ($T_\lambda=63.28$ and $\lambda=100$ in equation (8.14)) and Figure 8.3(b) shows the same signal with the window length reduced to 31.64 degrees. This shows the increase in frequency spread caused by the reduction in the angle domain length of the window. The plot follows the instantaneous frequency of the signal, with the maximum energy point varying between 83 and 93 orders. The energy remains constant in rotation angle as would be expected for a signal with no amplitude modulation.
Figure 8.4 Spectrograms of modulated monocomponent and multicomponent signals.

Figure 8.4 shows the spectrograms of (a) an 88 order sine wave which has a four per revolution sinusoidal amplitude modulation and a twice per revolution sinusoidal frequency modulation and (b) the same signal with a 40 order unmodulated sine wave added (the angle domain amplitude and instantaneous frequency of these signals are shown in Figure 8.1 and Figure 8.2). Both spectrograms were calculated using a Gaussian window with a length of 63.28 degrees at the -40dB point. The spectrogram gives a relatively good description of the signal behaviour in both cases, with the twice per revolution frequency modulation and the four times per revolution amplitude modulation at the 88 order component being identifiable in both signals. Complete separation of the two signal components in Figure 8.4(b) is achieve in the spectrogram with the window length used.

8.2.1.6 Advantages and limitations of the spectrogram

The advantage of the spectrogram is that it is easily interpretable both in terms of its implementation and the visual representation of the results produced. However, the bandwidth-time limitations imposed by the window function mean that if we want to get better resolution in time we must sacrifice resolution in frequency and vice versa. Note that for a particular signal a particular window may be more appropriate than another window. However, in the case of multicomponent signals, the ‘optimum’ window for one component may not be appropriate for the other components. For example, consider a signal consisting of two closely spaced unmodulated sine waves and a sinusoidally frequency modulated sine wave. To give a clear indication of the time
varying frequency in the modulated sine wave requires a window which is relatively narrow in time (which will have a wide spread in frequency) however, to separate the closely spaced components in frequency, we require a window which is narrow in frequency (broad in time).

![Spectrograms](TEST5.OUT)

(a) Narrow window (45 degrees)  
(b) Broad window (180 degrees)

*Figure 8.5 Spectrograms of signal with conflicting window length requirements.*

Figure 8.5 shows spectrograms of a signal which has two unmodulated sine waves at 40 and 52 orders and a component with a mean frequency of 120 orders which is sinusoidally modulated in frequency at four times per revolution (with variation in frequency of ±10 orders). All components have the same (constant) amplitude. Figure 8.5 (a) shows the spectrogram obtained using a Gaussian window of (angular) length of 45 degrees. This gives a relatively good representation of the frequency modulated component but a very poor representation of the two unmodulated components at 40 and 52 orders due to the poor frequency resolution. Figure 8.5 (b) shows the same signal with the window length increased to 180 degrees. This gives a much better frequency resolution and, consequently, a good representation of the unmodulated components at 40 and 52 orders however, because of the poor resolution in the angle domain, the modulated component is now poorly represented. To give a good representation of all components in this signal requires good resolution in both time and frequency, which cannot be achieved using the spectrogram.
8.2.2 The Wigner-Ville distribution

The Wigner-Ville distribution (WVD) is defined as the Wigner distribution \([83]\) of the analytic signal \([81]\),

\[
W(t, f) = \int s\left(t + \frac{\tau}{2}\right)s^*\left(t - \frac{\tau}{2}\right)e^{-j2\pi ft}\,d\tau,
\]

(8.21)

where \(s^*(t)\) is the complex conjugate of \(s(t)\). The Wigner-Ville distribution can also be expressed in terms of the spectrum \([11]\),

\[
W(t, f) = \int S\left(f + \frac{\nu}{2}\right)S^*\left(f - \frac{\nu}{2}\right)e^{j2\pi ft}\,d\nu.
\]

(8.22)

8.2.2.1 The marginal conditions

The integral of the WVD over frequency at a particular time gives the energy density at that time (Cohen \([24]\))

\[
\int W(t, f)\,df = \int s\left(t + \frac{\tau}{2}\right)s^*\left(t - \frac{\tau}{2}\right)e^{-j2\pi ft}\,d\tau\,df
\]

\[= \left|s(t)\right|^2.\]

(8.23)

Similarly, the integral of the WVD over time at a particular frequency gives the energy density in frequency at that frequency

\[
\int W(t, f)\,dt = \left|S(f)\right|^2.
\]

(8.24)

The properties expressed by equations (8.23) and (8.24) are called the marginal conditions (Cohen \([23]\)) and are often stated as a requirement for a time-frequency energy distribution \([23,24,12]\). Note that meeting the marginal conditions ensures that the total energy in the signal is preserved,

\[
\iint W(t, f)\,dt\,df = \int \left|s(t)\right|^2\,dt = \int \left|S(f)\right|^2\,df = \text{total energy}
\]

(8.25)

however, Cohen \([24]\) noted that it is possible for a distribution to give the correct value for the total energy without satisfying the marginals.
### 8.2.2.2 Multicomponent signals and cross-terms

For the multicomponent non-stationary signal defined by equation (8.7), the WVD is

\[
W(t, f) = \int \left[ \sum_c s_c \left( t + \frac{\tau}{2} \right) \sum_c s_c^* \left( t - \frac{\tau}{2} \right) e^{-j2\pi f\tau} \right] d\tau
\]

\[
= \sum_c \int s_c \left( t + \frac{\tau}{2} \right) s_c^* \left( t - \frac{\tau}{2} \right) e^{-j2\pi f\tau} d\tau + \sum_{c, d \neq c} \sum_{d \neq c} W_{c, d}(t, f)
\]

\[
= \sum_c W_c(t, f) + \sum_{c \neq d} W_{c, d}(t, f),
\]

where \( W_{c, d}(t, f) \) is the Cross Wigner-Ville distribution [12,13]

\[
W_{c, d}(t, f) = \int s_c \left( t + \frac{\tau}{2} \right) s_d^* \left( t - \frac{\tau}{2} \right) e^{-j2\pi f\tau} d\tau.
\]

Therefore, the Wigner-Ville distribution of a multicomponent signal is the sum of the Wigner-Ville distributions of each component (‘auto-terms’, Choi and Williams [18]) plus the sum of the Cross Wigner-Ville distributions of each component with all other components (‘cross-terms’, [18]).

### 8.2.2.3 ‘Negative’ energy

Consider a signal consisting of two unmodulated sine waves,

\[
s(t) = a_1 e^{j2\pi f_1 t} + a_2 e^{j2\pi f_2 t}.
\]

From equation (8.21), the Wigner-Ville distribution of this signal is
\[ W(t, f) = \int \left[ \left( a_1 e^{j2\pi f_1 \left(t + \frac{T}{2}\right)} + a_2 e^{j2\pi f_2 \left(t + \frac{T}{2}\right)} \right) \right. \\
\left. e^{-j2\pi\tau} d\tau \right] \]

\[ = a_1^2 \delta(f - f_1) + a_2^2 \delta(f - f_2) + 2a_1a_2 \cos\left(2\pi t(f_1 - f_2)\right)\delta\left(f - \frac{f_1 + f_2}{2}\right). \]

which, in addition to the two ‘auto-terms’ at frequencies \( f_1 \) and \( f_2 \), has a ‘cross-term’ positioned at frequency \((f_1 + f_2)/2\) which is a cosine wave with frequency \((f_1 - f_2)\). It is clear that this component becomes negative and, since the WVD is an ‘energy’ distribution, represents a ‘negative’ energy in the time-frequency domain. Although the concept of ‘negative’ energy is physically meaningless, note that the summation of the components in equation (8.29) over frequency equals the magnitude squared of equation (8.28) and the ‘cross-term’ energy (including the negative parts) is required to reflect the variable energy in the signal over time whilst maintaining constant energy at the ‘auto-terms’.

It can be easily shown that similar cross-terms (and negative energy) occur between discrete time components with the same frequency, by calculating the WVD using an equation similar to (8.28), but discrete in the time domain rather than frequency, which will give a time-domain equivalent of equation (8.29).

### 8.2.2.4 Non-stationary signals

The WVD can be viewed as the Fourier transform of the inner product in equation (8.21), that is (Boashash [11]),

\[ W(t, f) = \int z_t(\tau)e^{-j2\pi f\tau} d\tau = \mathcal{F}\left[ z_t(\tau) \right], \]

where \( z_t(\tau) = s(t + \frac{T}{2})s^\ast(t - \frac{T}{2}) \).

Note that \( z_t(\tau) \) equals \( z^\ast(-\tau) \) giving (Bendat [3]) a Fourier transform of \( z_t(\tau) \) which is real valued and, therefore, the Wigner-Ville distribution is also a real valued function (Boashash [11]).
The function $z_t(\tau)$ can be expressed as (using Van der Pol’s form of the analytic signal (Appendix B, equation B.15)),

$$z_t(\tau) = a(t + \frac{\tau}{2})a(t - \frac{\tau}{2})e^{j2\pi\left(j^{t+\tau/2}f_i(\xi)d\xi - \int_{t-\tau/2}^{t+\tau/2}f_i(\xi)d\xi\right)} = \alpha_t(\tau)e^{j\beta_t(\tau)} \quad (8.31)$$

and, using the ‘product property’ of the Fourier transform [3], the WVD can be expressed as the convolution of the Fourier transforms of the amplitude and phase related components in (8.31),

$$W(t, f) = \Lambda_t(f) \ast \Psi_t(f),$$

where $\Lambda_t(f) = \mathcal{F}\left[a(t + \frac{\tau}{2})a(t - \frac{\tau}{2})\right]$ \quad (8.32)

and $\Psi_t(f) = \mathcal{F}\left[e^{j2\pi\left(j^{t+\tau/2}f_i(\xi)d\xi - \int_{t-\tau/2}^{t+\tau/2}f_i(\xi)d\xi\right)}\right]$.\n
The Fourier transform of the phase related component is,

$$\Psi_t(f) = \int e^{j2\pi\left(j^{t+\tau/2}f_i(\xi)d\xi - \int_{t-\tau/2}^{t+\tau/2}f_i(\xi)d\xi\right)} e^{-j2\pi\xi\tau} d\tau \quad (8.33)$$

and, for signals having a linear instantaneous frequency law, $f_i(t) = f_0 + bt$ (linear FM chirp, Boashash [11]), equation (8.33) becomes

$$\Psi_t^{(lin)}(f) = \int e^{j2\pi\left(j^{t+\tau/2}f_i(\xi)d\xi - \int_{t-\tau/2}^{t+\tau/2}f_i(\xi)d\xi\right)} e^{-j2\pi\xi\tau - j2\pi\xi\tau} d\tau = \int e^{j2\pi\tau(f_0 + bt)} - j2\pi\xi\tau d\tau \quad (8.34)$$

That is, for a linear instantaneous frequency law the frequency related component of the Wigner-Ville distribution becomes a delta function centred at the instantaneous frequency.

198
If the instantaneous frequency is not linear (or constant) the frequency related component of the Wigner-Ville distribution at time $t$ will have the same form as the Fourier transform of a frequency modulated signal (Appendix B.2.1.2, equation B.33), that is, it can be described by a number of frequency domain convolutions. Boashash [11] showed that the average frequency of the WVD at a given time is the instantaneous frequency at that time $f_i(t)$. Therefore, for a signal with a non-linear instantaneous frequency law the frequency related component of the WVD will have a centre of gravity at the instantaneous frequency however, as with the Fourier transform of a frequency modulated signal, the distribution of energies about the centre frequency may be difficult to interpret.

The amplitude related component in equation (8.31), $a(t+\tau/2)a(t-\tau/2)$, is real valued and symmetrical about $\tau=0$. This will have a Fourier transform which is real valued and symmetrical about $f = 0$ (Bendat [3]). Therefore, the convolution of the Fourier transform of the amplitude related component $\Lambda_t(f)$ with the Fourier transform of the frequency related component $\Psi_t(f)$ (8.32) will preserve the centre of gravity of the distribution (at the instantaneous frequency). As is shown in equation (8.23), the integral of the WVD over frequency at a particular time gives the energy density at that time; it can easily be seen that all the energy in the signal is represented by the amplitude related component, with the ‘energy’ in the phase related component being unity.

Unless the amplitude is constant, there will be a ‘spread’ in frequency (similar to that seen with the Fourier transform) caused by the amplitude modulation.

Therefore, the WVD will give a ‘correct’ description of the signal only in cases where we have a monocomponent signal which has a linear variation in frequency over time and constant amplitude.

### 8.2.2.5 The Windowed Wigner-Ville Distribution

Windowing can be used on the Wigner-Ville distribution in a similar fashion (and for similar purposes) as the windowing applied to the Fourier transform in the short-time Fourier transform (Section 8.2.1). The windowed version of the WVD is often called
the windowed or pseudo Wigner-Ville distribution (PWVD) ([11,22]) and is defined as
(Boashash [11])

\[ W_w(t, f) = \int h(\tau)s(t + \frac{\tau}{2})s^*(t - \frac{\tau}{2})e^{-j2\pi ft} d\tau, \]  

(8.35)

where \( h(\tau) \) is a real symmetrical window function as discussed in section 8.2.1.3. The
effects of the windowing process are best understood by viewing the windowed WVD as
a ‘short-time Fourier transform’ of the inner product \( z(\tau) \) of the WVD as defined in
equation (8.30); the discussion in Section 8.2.1 can be directly applied and will not be
repeated here.

As with the short-time Fourier transform, the windowed Wigner-Ville distribution has a
bandwidth-time limitation however, because of the assumption of linear instantaneous
frequency rather than the constant frequency requirement in the STFT, relative good
results can be obtained using a wider (in time or angle) window than would be required
for the STFT.

Figure 8.6 shows (a) the unwindowed WVD and (b) a windowed WVD (using a
Gaussian window of length (in the angle domain) of 63.28 degrees for a 88 order sine
wave which has a twice per revolution frequency modulation of \( \pm 5 \) orders (STFTs for
this signal are shown in Figure 8.3). The effect of the windowing is quite obvious.
Figure 8.6(a) shows that, without windowing, the WVD of this sinusoidally frequency
modulated signal is difficult to interpret. Although the ‘average’ frequency is following
the instantaneous frequency law of the signal, the pattern produced is quite confusing (this is analogous to the frequency domain spectrum of the frequency modulated signal in Appendix B, Figure B.3). It should be noted that the signal components which appear to be ‘reflections’ of the modulated signal about the 88 order ‘carrier’ frequency are in fact cross-terms (although this signal is ‘monocomponent’ in the angle domain it is in fact ‘multicomponent’ in frequency, with the same frequency occurring at different ‘times’). The cross-terms contain both positive and negative energy regions and the centre of gravity is at the instantaneous frequency, not at the central 88 order frequency as might appear the case. The windowed WVD in Figure 8.6(b) gives a clear representation of the signal behaviour and is very similar to the STFT using the same window length (Figure 8.3(a)). The windowed WVD actually has a higher resolution in frequency (narrower spread) than the STFT, however this requires careful inspection of the two plots to detect.

8.2.2.6 The Discrete Wigner-Ville Distribution

A number of discrete forms of the Wigner distribution and Wigner-Ville distribution have been proposed. The major difficulty in defining a discrete form of the Wigner-Ville distribution is the dual requirement to sample the distribution at half the signal sample interval in both time (equation (8.21)) and frequency (equation (8.22)). Claasen and Mecklenbräuker [19] reviewed a number of discrete-time Wigner distributions (i.e., using the real signal and not the analytic signal), all of which introduced aliasing for signals sampled at the Nyquist rate (i.e., twice the signal bandwidth). They advised that real signals to be analysed using the Wigner distribution should be sampled at twice the Nyquist rate.

Boashash [9] claimed that the use of the analytic signal in the Wigner-Ville distribution eliminated aliasing in the discrete implementation due to the elimination of the negative frequency components. This claim was later revised [11] to exclude analysis of signals with short duration or when using short window lengths, both of which can have negative frequency components. It was suggested [11] that in these circumstances the
sample rate of the signal should be increased prior to analysis either by interpolation or oversampling.

Boashash [9] derived a discrete-time version of the Wigner-Ville distribution by substituting $\theta = \tau/2$ into the definition of the WVD in equation (8.21), giving

$$W(t, f) = 2\int s(t + \theta) s^*(t - \theta) e^{-j4\pi\theta f} d\theta.$$  \hspace{1cm} (8.36)

Converting this to discrete form is straightforward. Using the procedure described in section 8.2.1.4,

$$\hat{W}(n, m) = \frac{2}{N} \sum_{k=0}^{N-1} \hat{s}(n + k) \hat{s}^*(n - k) e^{-j4\pi nk/N}$$

that is,

$$\hat{W}(n, m/2) = \frac{2}{N} \sum_{k=0}^{N-1} \hat{s}(n + k) \hat{s}^*(n - k) e^{-j2\pi nk/N},$$

which can easily be implemented using the discrete Fourier transform.

Peyrin and Prost [63] derived a discrete-time/frequency version of the Wigner(-Ville) distribution by considering the effects of discretization in both time and frequency domains simultaneously (note that the discrete-time Wigner-Ville in equation (8.37) has been derived from the time domain definition of the Wigner-Ville without regard to the frequency domain definition). The discrete-time/frequency Wigner-Ville distribution derived by Peyrin and Prost [63] is (the derivation can be found in [63] or Forrester [33]),

$$\hat{W}(n, m) = \frac{1}{2N} \sum_{k=0}^{N-1} \hat{s}(k) \hat{s}^*(n - k) e^{-j\pi m(2k-n)/N},$$  \hspace{1cm} (8.38)

where $0 \leq n < 2N$ and $0 \leq m < 2N$. For computational purposes, equation (8.38) can be divided into its even and odd numbered time samples [63, 33].
\[
\hat{W}(2n, m) = \frac{1}{2N} \sum_{k=0}^{N-1} \hat{s}(n+k)\hat{s}^*(n-k)e^{-j2\pi mk/N}, \quad \text{and}
\]

\[
\hat{W}(2n+1, m) = \frac{1}{2N} e^{j\pi m/N} \sum_{k=0}^{N-1} \hat{s}(n+k)\hat{s}^*(n-k+1)e^{-j2\pi mk/N},
\]

which can be calculated using the discrete Fourier transform for even time samples and a discrete Fourier transform followed by multiplication by an exponential term for the odd time samples. This implementation of the discrete Wigner-Ville has been used in the work presented in this thesis.

The discrete windowed Wigner-Ville distribution is defined by including the discrete version of the window function \( \hat{h}(k) \) in the inner product of the summations in equation (8.39).

### 8.2.2.7 Advantages and disadvantages of the WVD

The most obvious disadvantage of the Wigner-Ville distribution is the cross-terms and associated negative energy regions. A number of methods have been developed to reduce the level of the cross-terms which will be discussed later.

![Figure 8.7 WVDs of modulated monocomponent and multicomponent signals.](image)

Figure 8.7 shows the effect of ‘cross-terms’ (equation (8.26)) in the WVD. Figure 8.7(a) shows the windowed WVD (63.28 degree window) of a signal with sinusoidal amplitude and frequency modulations (as per Figure 8.4). A 40 order sine wave has been
added to the signal in Figure 8.7(b) and, although there is a good representation of the behaviour of both signals, there is an additional ‘artifact’ or ‘cross-term’ between the two. Note that the ‘energy’ in the cross-term fluctuates (and in fact goes negative in parts) and the sum of the energy in the cross-term over angle at a set frequency is actually zero.

The one obvious advantage that the WVD has over the STFT is its improved time-frequency resolution. Properties of the WVD such as the concentration of energy at the instantaneous frequency, meeting the ‘marginal conditions’, and other theoretically desirable properties [11] also make the WVD preferable to the STFT in many signal processing applications [6,7,40,81].

Figure 8.8 shows the windowed WVDs for the multicomponent signal for which the STFTs (with same windows) are shown in Figure 8.5. As with the STFT, the WVD using the narrow window (Figure 8.8(a)) does not adequately represent the two closely spaced sine waves at 40 and 52 orders but gives a good representation of the frequency modulated component. However, with the broad window (Figure 8.8(b)), an acceptable representation (except for the presence of the cross-terms) of all components is achieved. Note that this could not be achieved using the STFT.
8.2.3 General form of time-frequency distributions

In addition to the spectrogram and Wigner-Ville distribution, many other time-frequency distributions have been proposed, including the Page [62], Margenau-Hill [46], Kirkwood-Rihaczek [41,71], Choi-Williams [18], and Zhao-Atlas-Marks [85] distributions, all of which have their own advantages and disadvantages. In 1966, Cohen [23] developed a generalized form of ‘phase-space’ distributions from which all other time-frequency energy distributions could be derived. The general form, which has since become known as Cohen’s class of distributions, is [23]

\[ \rho(t, f) = \iiint e^{j2\pi(v-u-t)}g(v, \tau)s(u + \frac{\tau}{2})s^*(u - \frac{\tau}{2})e^{-j2\pi f\tau}dv
du d\tau \]

where \( g(v, \tau) \) is referred to as the kernel function and is used to define the properties of the distribution. For example, by setting the kernel \( g(v, \tau) = 1 \), equation (8.40) becomes

\[ \rho(t, f) = \iiint e^{j2\pi(v-u-t)}s(u + \frac{\tau}{2})s^*(u - \frac{\tau}{2})e^{-j2\pi f\tau}dv
du d\tau \]

which is the Wigner-Ville distribution (8.21) and, setting \( g(v, \tau) = h(\tau) \),

\[ \rho(t, f) = \iiint e^{j2\pi(v-u-t)}h(\tau)s(u + \frac{\tau}{2})s^*(u - \frac{\tau}{2})e^{-j2\pi f\tau}dv
du d\tau \]

which gives the windowed Wigner-Ville distribution (8.35).

8.2.3.1 Properties of the Cohen class of distributions

The number of distributions which can be generated from equation (8.40) is infinite. Cohen [23] proposed a number of restrictions on the properties of the distributions to be studied, and showed that certain constraints could be placed on the kernel function \( g(v, \tau) \) in equation (8.40)) to ensure that a distribution meets the restrictions. The constraints on the kernel function required to give certain desirable properties have been
extensively studied [11, 18, 22, 23, 24, 85] and the range of ‘desirable’ properties and the kernels required to meet them is being continuously expanded. An extensive range of properties is given by Boashash [11] and Cohen [24]. Some of these are:

a) The signal energy is preserved

The signal energy is preserved (equation (8.25)) if \( g(0,0) = 1 \).

b) The marginal condition in time

for integration over frequency = energy density in time (8.23), \( g(\nu,0) = 1 \).

c) The marginal condition in frequency

for integration over time = energy density in frequency (8.24), \( g(0,\tau) = 1 \).

d) Real valued distributions

The distribution will be real valued if \( g(\nu,\tau) = g^*(-\nu,-\tau) \).

e) Invariance to time and frequency shifts

If two signals are identical except for a shift in time or frequency then the distributions of the signals should also be identical except for a similar shift in time or frequency. Cohen [24] showed this is true as long as the kernel function is independent of time and frequency.

f) The first moments of the distribution equal the instantaneous frequency and group-delay

Boashash [11] showed that the first moment of a distribution in frequency is equal to the instantaneous frequency (8.3) and the first moment in time equals the group-delay (8.6),

\[
\int \frac{f \rho(t,f) df}{\rho(t,f) df} = f_i(t) \quad \text{and} \quad \int \frac{t \rho(t,f) dt}{\rho(t,f) dt} = \tau_g(f),
\]  
(8.43)
if \[11\]

\[
\frac{\partial g(v, t)}{\partial t} \bigg|_{t=0} = \frac{\partial g(v, t)}{\partial v} \bigg|_{v=0} = 0,
\]

\[\quad g(v, 0) = \text{constant for all } v \quad \text{and} \quad g(0, \tau) = \text{constant for all } \tau.\]

\[g) \quad \text{Recovery of signal}\]

Cohen [22] showed that a signal could be recovered up to a constant phase factor if the kernel function \(g(v, \tau)\) is well defined at every point or has isolated zeros, but not regions where it is zero.

\[h) \quad \text{Finite time support}\]

A distribution has finite time support if it is zero before the signal starts and zero after the signal ends. For a signal to have finite time support the kernel must meet the condition (Cohen [22]):

\[
\int g(v, \tau)e^{-j2\pi vt} dv = 0 \quad \text{for } |\tau| < 2|t|.
\]

The Wigner-Ville distribution, for which the kernel function \(g(v, \tau) = 1\), has all the properties listed.

For the spectrogram, the kernel function is (Cohen [22])

\[g_s(v, \tau) = \int h^*(u - \frac{\tau}{2})h(u + \frac{\tau}{2})e^{-j2\pi vu} du,
\]

which has properties (a) if the total energy of the window = 1, plus properties (d) and (e) (i.e., it is real and invariant to time and frequency shifts).

8.2.3.2 Reduced Interference Distributions

The Wigner-Ville distribution possesses a number of mathematically satisfying properties however, as was seen previously, it produce large cross-terms when applied to
multicomponent signals. This can make interpretation difficult. Recently, a number of investigators have made significant advances in the approach to kernel design, with particular effort concentrated on the reduction of the cross-terms.

### 8.2.3.2.1 The Choi-Williams Distribution

Choi and Williams [18] investigated the differences in behaviour of the auto-terms and cross-terms in the Cohen class of distributions. It was found that the auto-terms have a locus which passes through the origin of the generalised ambiguity function (Cohen and Posch [21]),

\[ M(v, \tau) = g(v, \tau) \int s(u + \frac{v}{2})s^*(u - \frac{v}{2})e^{j2\pi vu} du, \quad (8.44) \]

while the cross-terms stay remote from the origin of (8.44). Therefore, a kernel function \( g(v, \tau) \) which is peaked near the origin (i.e., as \( v \) and \( \tau \) approach zero) and diminishes as \( v \) and \( \tau \) move away from the origin, will attenuate the cross-terms with little effect on the auto-terms. Particular effort was made in [18] to maintain the desirable properties of the Wigner-Ville distribution whilst minimizing the cross-terms, resulting in the choice of an exponential kernel function [18],

\[ g_{cw}(v, \tau) = e^{-v^2\tau^2/\sigma}, \quad (8.45) \]

which meets all properties listed in the previous section except for finite time support, and results in the Choi-Williams distribution (CWD) (from equation (8.40)):

\[
\begin{align*}
\rho_{cw}(t, f) &= \iint e^{j2\pi v(u-t)}e^{-v^2\tau^2/\sigma}s(u + \frac{v}{2})s^*(u - \frac{v}{2})e^{-j2\pi \tau \sigma} dv du d\tau \\
&= \iint \sqrt{\frac{\pi \sigma}{\tau}} e^{-\pi^2 \sigma(u-t)^2/\tau^2} s(t + \frac{v}{2})s^*(t - \frac{v}{2})e^{-j2\pi \tau \sigma} dv du d\tau.
\end{align*}
\quad (8.46)
\]

The parameter \( \sigma \) in equation (8.46) is used to control the size of the cross-terms. If \( \sigma \) is large (e.g., 100) the kernel function in (8.45) approximates 1 everywhere, and the distribution is essentially the Wigner-Ville distribution. For smaller values of \( \sigma \), the kernel becomes peaked near the origin of the generalised ambiguity plane and the remote regions (i.e., the cross-terms) are ‘attenuated’. The energy present at the cross-term is
actually preserved but is spread (symmetrically) over a wider area, giving the appearance of attenuation at the original location of the cross-term. Choi and Williams found that as the value of $\sigma$ became very small (e.g., < 0.1) there was some loss of resolution in the auto-terms.

For the discrete implementation of the Choi-Williams ‘exponential distribution’, the ‘Running Windowed Exponential Distribution’ was defined as

$$\hat{\rho}_{cw}(n, m) = \frac{2}{N} \sum_{k=-N/2}^{N/2} \hat{h}(k) K(n, k) e^{-j2\pi km/N}$$

(8.47)

where

$$K(n, k) = \sqrt{\frac{\sigma}{4\pi k^2}} \sum_{u=-M/2}^{M/2} e^{-\sigma u^2/4k^2} \hat{s}(n + u + k)^* \hat{s}^*(n + u - k).$$

The discrete window function $\hat{h}(k)$ and its length $N$ controls the frequency resolution of the distribution (in a similar fashion to the window used for the windowed WVD (8.35) and STFT (8.8)) and, for consistency with these, the same Gaussian window function (8.14) has been used here. The parameter $M$ determines the range over which the time indexed autocorrelation function $K(n,k)$ is calculated. The larger the value of $M$, the better the approximation given by the discrete Running Windowed Exponential Distribution to a smoothed version of the continuous Choi-Williams distribution.

### 8.2.3.2.2 Zhao-Atlas-Marks Distribution

An alternate approach to the reduction of cross-terms was taken by Zhao, Atlas and Marks [85]. They studied the behaviour of time-frequency distributions by considering the theta integration of the kernel [22], which is defined as the ‘time-lag’ kernel by Boashash [11],

$$G(u, \tau) = \int g(v, \tau) e^{j2\pi vu} dv,$$

(8.48)

and the Cohen class of distributions (8.40) in terms of the time-lag kernel is:

$$\rho(t, f) = \iint G(u - t, \tau) s(u + \frac{\tau}{2}) s^*(u - \frac{\tau}{2}) e^{-j2\pi \tau} du d\tau.$$

(8.49)
Zhao et al. [85] showed that for a time-frequency distribution to have finite time support, the non-zero support region of $G(u, \tau)$ must lie within the ‘cone-shaped’ region $-|\tau|/2 \leq u \leq |\tau|/2$. In development of the Zhao-Atlas-Marks ‘cone kernel’, emphasis was placed on finite time support and reduction of cross-terms, rather than meeting the ‘marginal conditions’ and other properties. It was shown [85] that the spectrogram, which displays the best cross-term reduction of the Cohen class of distributions, does not have finite time support because its time-lag kernel has non-zero support outside the region $-|\tau|/2 \leq u \leq |\tau|/2$. It was also shown that the cross-terms in the spectrogram are placed on top of the auto-terms and, because of the smearing in time and frequency, are normally not noticeable. Hence, it is often incorrectly stated that the spectrogram has no interfering cross-terms. Note that the spectrogram shown in Figure 8.5(a) has visible cross-terms between the two closely spaced sine waves (at 40 and 52 orders) due to the inappropriate analysis window used.

Zhao et al. [85] proposed a time-lag kernel (the ‘cone kernel’) which has a non-zero support region covering the cone-shaped region $-|\tau|/2 \leq u \leq |\tau|/2$, that is

$$G_{zam}(u, \tau) = \begin{cases} \omega(\tau), & |u| \leq |\tau|/2 \\ 0, & |u| > |\tau|/2 \end{cases} \quad (8.50)$$

where $\omega(\tau)$ is a bounded taper function, similar to the sliding window function used in the spectrogram. The Zhao-Atlas-Marks (ZAM) distribution [85] is defined by placing the time-lag kernel of equation (8.50) into equation (8.49) giving

$$\rho_{zam}(t, f) = \int \int G(u-t, \tau)s\left(u + \frac{\tau}{2}\right)s^*\left(u - \frac{\tau}{2}\right)e^{-j2\pi ft} du d\tau$$

$$= \int \omega(\tau)e^{-j2\pi ft} \int_{t-|\tau|/2}^{t+|\tau|/2} s(u + \frac{\tau}{2})s^*(u - \frac{\tau}{2}) du d\tau. \quad (8.51)$$

The kernel function $g(v, \tau)$ can be generated from the time-lag kernel by inverting equation (8.48), that is,

$$g(v, \tau) = \int G(u, \tau)e^{-j2\pi vu} du. \quad (8.52)$$
giving a kernel function for the Zhao-Atlas-Marks distribution (from the time-lag kernel (8.50)) of

\[
g_{\text{zam}}(v, \tau) = \int_{-|\tau|/2}^{|\tau|/2} \omega(\tau)e^{-j2\pi vu}du = \frac{\omega(\tau)\sin(2\pi v|\tau|/2)}{\pi v}.
\]  

(8.53)

If \(\omega(\tau)\) is a real symmetrical tapered function, it can be seen from the kernel requirements for the properties listed previously that the ZAM distribution has properties (d), (e) and (h) that is, it is real valued, shift invariant and has finite time support. In this respect, it may be more appropriate to view the ZAM distribution as a finite time support version of the spectrogram rather than a reduced interference version of the Wigner-Ville distribution.

### 8.2.3.3 Cross-term attenuation

![Choi-Williams Distribution](TEST4_2.OUT)

![Zhao-Atlas-Marks: TEST4_2.OUT](Zhao-Atlas-Marks: TEST4_2.OUT)

(a) Choi-Williams Distribution (\(\sigma=0.1\))

(b) Zhao-Atlas-Marks Distribution

*Figure 8.9 Reduced Interference Distributions of multicomponent signal.*

Figure 8.9 shows (a) the Choi-Williams distribution and (b) the Zhao-Atlas-Marks distribution of the multicomponent signal for which the spectrogram was given in Figure 8.4(b) and the WVD in Figure 8.7(b). Gaussian windows defined with a length of 63.28 degrees at the -40dB points were used in both distributions, with the CWD controlling parameter \(\sigma\) set to 0.1. Note that the restriction of the window to the cone shaped time-lag kernel in the ZAM distribution effectively narrows the window in frequency.
(equation (8.50)), providing a narrower concentration of energy in frequency (but not necessarily a higher frequency resolution) for a given window.

In this example, the ZAM (b) provides the clearest visual representation of the signal, with a narrow spread in frequency and only a low amplitude (< -30dB) cross-term being visible. The CWD (a) shows almost identical ‘auto-terms’ to the WVD (Figure 8.7(b)), with attenuation of the maximum cross-term amplitudes (from 0dB in the WVD to < -20dB) with the cross-term energy being spread over a wide frequency range. Although the cross terms are at a much lower level in the CWD, the visual representation used here (colour plot) makes the CWD perhaps more confusing than the WVD. However, if the range of the display was restricted to 20dB, the cross term energy would not be noticeable. Therefore, it could be argued that the ‘reduced interference’ of the CWD actually causes a reduction in the useable dynamic range of the time-frequency representation.

8.2.3.4 Time-frequency resolution

![Figure 8.10 CWD (σ=0.1) of a signal with conflicting window length requirements.](image)

Figure 8.10 shows the Choi-Williams distribution for the multicomponent signal shown in Figure 8.5 (spectrogram) and Figure 8.8 (WVD) which has the dual requirement of high time (angle) resolution (to describe the rapidly varying frequency of the modulated signal centred about 120 orders) and high frequency resolution (to separate the two closely spaced sine waves at 40 and 52 orders).
As would be expected, the CWD gives essentially the same results as the WVD (Figure 8.8) for the auto-terms with the cross-terms being reduced in amplitude and smeared in frequency for (a) a narrow (angle domain) window and in both frequency and angle for (b) a broad window.

![Image](a) Narrow Window (45 degrees) ![Image](b) Broad Window (180 degrees)

*Figure 8.11 ZAM of a signal with conflicting window length requirements.*

Figure 8.11 shows the Zhao-Atlas-Marks distribution for the signal in Figure 8.10 using the same basic window definitions as used with this signal for the spectrogram (Figure 8.5), WVD (Figure 8.8) and CWD (Figure 8.10). Although the ZAM distribution gives a good representation of the frequency varying sine wave centred at 120 orders using a narrow window (a) (with poor representation of the two sine waves at 40 and 52 orders) the representation of the frequency varying sine wave is very poor with a broad window (b). This result is similar to that for the spectrogram (Figure 8.5(b)), whereas both the WVD and the CWD gave reasonable representations of all three signals with the broad window. This result is suggested by the ZAM kernel function in equation (8.53) (and the subsequent discussion of this kernel) which indicate that the ZAM distribution is more closely related to the spectrogram than the Wigner-Ville distribution.

### 8.3 SUMMARY

In this chapter, it was shown that non-stationary multicomponent signals cannot be adequately represented in the time or frequency domain. The concept of time-frequency domain analysis was discussed and it was shown that these methods can give an adequate
representation of non-stationary multicomponent signals, with certain limitations. It was shown that the spectrogram, although giving a representation of simple multicomponent signals which is relatively easy to interpret, has a bandwidth-time limitation which makes it difficult to represent signals with requirements for high resolution in both time and frequency.

It was seen that the Wigner-Ville distribution has high time-frequency resolution capabilities however, it produces high energy interfering cross-terms which may make signals difficult to interpret. Windowing of the Wigner-Ville distribution improves the visual representation of the signal by ‘smearing’ in the time-frequency plane. This has the effect of blurring cross-terms which lie close to the auto-terms but at the cost of resolution. The windowing does not reduce cross TERMS which are remote from the auto-terms (i.e., those produced by widely separated signal components).

Cohen’s general class of distributions [23] was discussed, with particular reference to the constraints which could be applied to the ‘kernel’ to produce distributions with certain desirable properties. Two of the more recent distributions, the Choi-Williams and Zhao-Atlas-Marks distributions, were briefly examined and the properties related to their kernels discussed. It was shown that the Choi-Williams distribution maintains a number of desirable properties of the Wigner-Ville distribution but reduces the amplitude of the cross-terms by spreading their energy over frequency and/or time. However, it was seen that this arbitrary spreading of the cross-term energy can actually confuse rather than enhance the visual interpretation of the signal (when displayed as a colour map), and restriction of the ‘dynamic range’ of the visual representation may be required to make full use of this distribution.

Examination of the properties related to the kernel function of the Zhao-Atlas-Marks distribution indicated that it is more closely related to the spectrogram than the Wigner-Ville distribution and it was seen from practical examples that, although the ZAM appears to have higher time-frequency resolution than the spectrogram, it suffers from the same basic bandwidth-time limitations.
Chapter 9

TIME-FREQUENCY ANALYSIS OF GEAR FAULT VIBRATION DATA

In this chapter, the time-frequency analysis techniques discussed in the previous chapter (i.e., the spectrogram, Wigner-Ville distribution, Choi-Williams distribution, and Zhao-Atlas-Marks distribution) are applied to the vibration data for the gear faults discussed in Chapters 6 and 7.

9.1 WESSEX INPUT PINION CRACK

The synchronous signal averages generated from the in-flight vibration recordings for the Wessex input pinion crack described in Chapter 6 were analysed using the time-frequency analysis techniques described in Chapter 8. The results of the analysis of the 75 and 100% load conditions at 233, 103 and 42 hours before failure are shown for the spectrogram (Figure 9.1), Wigner-Ville distribution (Figure 9.2), Choi-Williams distribution (Figure 9.3) and Zhao-Atlas-Marks distribution (Figure 9.4). A brief explanation of the relevant features is given at the bottom of each figure.

The initial visible feature identifying the crack is a drop in energy at the 44 order (2 x tooth meshing frequency) line at 103 hours before failure; this can be seen with all four techniques. At 42 hours before failure, the drop in energy is more pronounced (using all techniques) with excited structural resonances also being visible. The Zhao-Atlas-Marks distribution gives the clearest description of the excited resonances. Note that the excitation (starting) point of the structural resonances coincides with a sharp drop in the energy level at the 44 order line and the Zhao-Atlas-Marks distribution also shows a small but distinct change in the instantaneous frequency around the 44 order line in the middle of the low energy region.
(a) 233 hours before failure - 75% load
(b) 233 hours before failure - 100% load
(c) 103 hours before failure - 75% load
(d) 103 hours before failure - 100% load
(e) 42 hours before failure - 75% load
(f) 42 hours before failure - 100% load

All spectrograms have been calculated using a Gaussian window with an angle domain width of 90 degrees at the -40dB points. The initial indication of a crack is the drop in ‘energy’ in the 44 order line (2 x tooth mesh) at 103 hours before failure (feature indicated on plots (c) and (d)). The crack area is obvious in (e) and (f) with the broad frequency spread being due to excited resonances. However, the limited frequency resolution makes individual resonances difficult to identify.

Figure 9.1 Spectrograms of cracked Wessex input pinion
All windowed Wigner-Ville distributions have been calculated using a Gaussian window with an angle domain width of 90 degrees at the -40dB points. The initial indication of a crack is the drop in ‘energy’ in the 44 order line (2 x tooth mesh) at 103 hours before failure (feature indicated on plots (c) and (d)). The crack area is obvious in (e) and (f) at approximately 180 degrees and 300 degrees of rotation (vertical axis) respectively. The interfering cross-terms make the plots visually confusing.

Figure 9.2 Windowed Wigner-Ville distributions of cracked Wessex input pinion
All windowed Choi-Williams distributions (Running Windowed Exponential Distribution) have been calculated using a control parameter $\sigma=0.1$ and a Gaussian window with an angle domain width of 90 degrees at the -40dB points. These show the same features as the windowed Wigner-Ville distributions (see Figure 9.2) with a reduction and spread in cross-term energy. In (e) and (f), individual resonances can be detected (arrowed).

Figure 9.3 Windowed Choi-Williams distribution of cracked Wessex input pinion
All Zhao-Atlas-Marks distributions have been calculated using a Gaussian window with an angle domain width of 90 degrees at the -40dB points. The initial indication of the crack at 103 hours before failure (feature indicated on plots (c) and (d)) is not as clear as with the other techniques however, the higher frequency resolution makes the excited resonances (arrowed) easier to detect. Note the ZAM has a lower angle (time) resolution and higher frequency resolution than the other distributions when using the same window.

Figure 9.4 Zhao-Atlas-Marks distributions of cracked Wessex input pinion
The features seen in these plots become more prominent as the crack progresses and are more noticeable at higher load (100%) than lower load (75%).

9.1.1 Comparison with other techniques

The analysis of the vibration for this gear using other techniques (Chapter 6) showed that:

a) For the kurtosis based condition indices, FM4A and the narrow band envelope kurtosis, the highest value was for the recording at 103 hours before failure and 75% load (plot (c) in the figures shown here). In Chapter 6, it was suggested that this might be due to a broadening (in rotation angle) of the region affected by the crack and/or the presence of excited resonances at the higher loads and with the more advanced crack at 42 hours before failure. The time-frequency distributions shown here, particular the Zhao-Atlas-Marks distribution (Figure 9.4), indicate that both these phenomena are occurring.

b) Of the analysis techniques investigated earlier, the narrow-band demodulation techniques gave the clearest indication of what was occurring in the vibration signal. However, this required a lot of detailed analysis and comparison of the demodulated amplitude and phase signals (plus a bit of intuitive guesswork) to gain insight into the signal behaviour. This also required the development of a modified version of the technique (to allow for ‘negative’ amplitudes) before demodulated signals which made any physical sense could be obtained. This analysis lead to the conclusion that in the early stages of cracking (103 hours before failure) there was a simultaneous drop in amplitude and phase at the crack location and, at the later stages of cracking (42 hours before failure), these features were still present and there was an additional impact exciting structural resonances occurring before (in rotation angle) these features. The amplitude drop and excited structural resonances (and their relationship to one another) can be clearly detect using the time-frequency distributions, particular the Zhao-Atlas-Marks distribution, because of the separation of the signal components in frequency. Note that when using the narrow band demodulation...
technique, contributions from both the modulated tooth meshing harmonic and the structural resonances were combined in the signal, making diagnosis difficult.

The one feature which is not clearly represented in the time-frequency distributions is the phase (frequency) change; this can be seen to some extent in the Zhao-Atlas-Marks distribution at 42 hours however, it is not obvious in the other time-frequency analysis techniques or at 103 hours before failure. It was shown by McFadden [54], and in Chapter 6 of this thesis, that this phase change provides the initial indication of the crack (i.e., the high kurtosis at 103 hours before failure). It would be expected that this phase change would be seen as a frequency change in the time-frequency distributions however, as was explained in Chapter 8, the instantaneous frequency is proportional to the time (angle) derivative of the signal phase and therefore is dependant upon the rate of change of the phase (not the magnitude of the deviation). In the time-frequency distributions shown here, the frequency resolution is such that, unless the reflected change in instantaneous frequency is more than a few frequency orders, it will be difficult to detect phase changes visually.

9.2 WESSEX INPUT PINION TOOTH PITTING

The time-frequency distributions of the synchronous signal averages for the Wessex input pinion with tooth pitting described in Chapter 6 are shown for the spectrogram in Figure 9.5, Wigner-Ville distribution in Figure 9.6, Choi-Williams distribution in Figure 9.7 and Zhao-Atlas-Marks distribution in Figure 9.8. All figures show the distribution for signals taken from the starboard transducer at 100% load at (a) 27.7 hours, (b) 201.1 hours, (c) 248.9 hours, (d) 292 hours and (f) 339.5 hours since overhaul with the 75% load condition also being shown at (e) 339.5 hours since overhaul. Comments related to the visible features of the signals are given at the bottom of each figure.
All spectrograms have been calculated using a Gaussian window with an angle domain width of 90 degrees at the -40dB points. The initial indication of pitting is an excited resonance (circled) between 22 and 44 orders (1 and 2 x tooth mesh) at 201.1 hours (b). As damage progresses (c)-(f) there is a loss of energy at the 22 order line occurring before the resonance (in angle) and a second region of excitation of the resonance appears (smaller circle). The low frequency resolution makes distinction between the resonance frequency and the 2nd mesh harmonic (44 orders) difficult.

Figure 9.5 Spectrograms of pitted Wessex input pinion
All windowed Wigner-Ville distributions have been calculated using a Gaussian window with an angle domain width of 90 degrees at the -40dB points. The interfering cross-terms make the excited resonance (circled) difficult to detect. The associated drop in energy on the 22 order line as the damage progresses (c)-(f) is quite clear. The second area of excitation of the resonance (seen in the spectrograms in Figure 9.1) are almost totally obscured by the interfering cross-terms.

Figure 9.6 Windowed Wigner-Ville distributions of pitted Wessex input pinion
All windowed Choi-Williams distributions (Running Windowed Exponential Distribution) have been calculated using a control parameter $\sigma=0.1$ and a Gaussian window with an angle domain width of 90 degrees at the -40dB points. The reduction in cross-term energy (from that of the Wigner-Villes in Figure 9.2) allow the excited resonances (circled) easier to detect. The improvement in frequency resolution over that of the spectrogram (Figure 9.1) makes it easier to distinguish the resonance frequency from the second mesh harmonic (44 orders). The second area of excitation of the resonance (seen in the spectrograms in Figure 9.1) are difficult to detect; these are still obscured by the interfering cross-terms.

Figure 9.7 Windowed Choi-Williams distributions of pitted Wessex input pinion
(a) 27.7 hours since overhaul - 100% load  
(b) 201.1 hours since overhaul - 100% load  

(c) 248.9 hours since overhaul - 100% load  
(d) 292 hours since overhaul - 100% load  

(e) 339.5 hours since overhaul - 75% load  
(f) 339.5 hours since overhaul - 100% load  

All Zhao-Atlas-Marks distributions have been calculated using a Gaussian window with an angle domain width of 90 degrees at the -40dB points. The major region of excitation of the resonance between 22 and 44 orders (circled) can be seen as the damage progresses (c)-(f) with the smaller excitation region (smaller circles) seen in Figure 9.1 being less obvious. The reduction in energy at the 22 order line just prior to the excited resonance can be seen to progress as damage progresses (d)-(f).  

Figure 9.8 Zhao-Atlas-Marks distributions of pitted Wessex input pinion
The initial identifying feature for this fault is the excited resonance between the tooth meshing frequency (22 orders) and two time tooth meshing frequency (44 orders), which is first seen in the plots at 201.1 hours since overhaul (b). This feature is easily detectable in the spectrogram (although not necessarily as a distinct resonance), Choi-Williams distribution and Zhao-Atlas-Marks distribution. The interfering cross-terms in the Wigner-Ville distribution make the excited resonance difficult to detect.

As the damage progresses (c)-(f), a second region of excitation of the resonance is observed in the spectrogram and, to a limited extent, the Zhao-Atlas-Marks distribution. This second region is probably due to pitting developing on teeth on the other side of the gear (separation of the two resonances is a little less than 180 degrees). This second excitation of the resonance is difficult to detect in the Wigner-Ville and Choi-Williams distribution due to the interfering cross-terms (although these are reduced in magnitude in the Choi-Williams distribution, they are still high enough to obscure this feature).

A drop in energy at the 22 order line is also seen as damage progress in all distributions (c)-(f). This occurs before (in rotation angle) the excitation of the resonance, as opposed to the drop in energy seen with the cracked tooth which occurred after the impact exciting the resonances (and was visible in the early stages of cracking, prior to any excitation of resonance being observed). As with the cracked input pinion, the features identified as being associated with the fault become more prominent as the fault progresses.

The other techniques used to analyse these signals in Chapter 6 showed high kurtosis values (for FM4A and narrow band envelope kurtosis) in the initial stages of pitting (at 201.1 hours since overhaul), with a reduction in value as the damage progressed. Interpretation of the narrow band demodulated amplitude and phase led to the conclusion that the major feature was excited resonances (as is observed in the time-frequency distributions shown here).
9.3 SPUR GEAR TEST RIG - PITTED TEETH

All spectrograms have been calculated using a Gaussian window with an angle domain width of 90 degrees at the -40dB points. Before 107.9 hours the gear was assumed to have no pitting, one tooth is pitted at 107.9 hours and three teeth are pitted at 115.5 hours. There is some indication of an excited resonance between 135 and 150 orders (circled) on all plots except 101.3 hours (a); this may be at the fifth harmonic of tooth mesh (135 orders) in (d)-(f). There is also a very slight energy drop on the 27 and 54 order lines (circled) in plots (c)-(f) and the relative energies of the tooth mesh harmonics at 54 and 81 orders diminish as damage progresses (d)-(f).

Figure 9.9 Spectrograms of test gear G3 (pitted teeth)
All windowed Wigner-Ville distributions have been calculated using a Gaussian window with an angle domain width of 90 degrees at the -40dB points. Little evidence of damage can be seen in these plots. The only place the 'excited resonance' (see Figure 9.9) between 135 and 150 orders can be seen is in (c) at 107.9 hours.

Figure 9.10 Windowed Wigner-Ville distributions of test gear G3 (pitted teeth)
All windowed Choi-Williams distributions have been calculated using control parameter $\sigma=0.1$ and a Gaussian window with an angle domain width of 90 degrees at the -40dB points. None of these plots show any useful information for these signals. The cross-term reduction strategy employed in the Choi-Williams distribution has spread the large energy cross-terms between 27 and 54 orders (1 and 2 x tooth mesh) over the entire frequency range. Although this has resulted in a sizeable reduction at the peak cross-term energy, it has obscured the low energy components at the higher frequencies.

*Figure 9.11 Windowed Choi-Williams distributions of test gear G3 (pitted teeth)*
All Zhao-Atlas-Marks distributions have been calculated using a Gaussian window with an angle domain width of 90 degrees at the -40dB points. These plots show little evidence of damage. The excited resonance between 135 and 150 orders (circled) at 107.9 hours (c) was also seen in the Wigner-Ville distribution and spectrogram.

*Figure 9.12 Zhao-Atlas-Marks distributions of test gear G3 (pitted teeth)*
The results of time-frequency analysis of the signal averages for the pitted spur gear from the experimental gear rig (see Chapter 7) are shown for the spectrogram in Figure 9.9, Wigner-Ville distribution in Figure 9.10, Choi-Williams distribution in Figure 9.11 and Zhao-Atlas-Marks distribution in 9.12. Although there are some anomalies present in the spectrograms, none of the time-frequency analysis techniques used here give a clear indication of the fault in this gear.

The narrow-band demodulation technique used on the vibration from this gear in Chapter 7 showed that there was a small relative amplitude modulation and a small phase change associated with this fault. When the phase is converted to an instantaneous frequency estimate (by differentiation), it is found to be a maximum of approximately ±1 order. On the plots shown, and with the frequency resolution of the techniques used, this is not visually detectable.

9.4 SPUR GEAR TEST RIG - CRACKED TOOTH

For the cracked tooth in the spur gear test rig (discussed in Chapter 7), only the spectrogram (Figure 9.13), Wigner-Ville distribution (Figure 9.14) and Zhao-Atlas-Marks distribution (Figure 9.15) are shown. The loss of ‘dynamic-range’ caused by the spreading of the cross-terms in the Choi-Williams distribution (discussed in Chapter 8) obscured the low level fault features for these vibration recordings (as in Figure 9.11) therefore, the Choi-Williams distributions are not shown for these signals.

The three techniques shown all indicate that there are excited resonances from 42.4 hours onwards, and that these get relatively larger with time (even after the reduction of load from 45 to 24.5 kW at 42.8 hours). These features are most easily identified in the spectrogram however, the Zhao-Atlas-Marks distributions provides better localisation of the resonances in frequency. The spectrogram shows a energy drop at the 54 order line (2 x tooth mesh frequency) in the final recording (f) at 43.75 hours. However, this is only seconds before shutdown of the rig.
All spectrograms have been calculated using a Gaussian window with an angle domain width of 90 degrees at the -40dB points. From 42.4 hours (b) onwards a number of excited resonances can be detected (circled). In (f) a distinct drop in the energy at the 54 order (2 x tooth meshing frequency) can also be seen. However, at this stage the total failure of the tooth was imminent, and the run was terminated 15 seconds after this recording was taken.

*Figure 9.13 Spectrograms of test gear G6 (cracked tooth)*
All windowed Wigner-Ville distributions have been calculated using a Gaussian window with an angle domain width of 90 degrees at the -40dB points. The excited resonances are clearly visible in the high frequency regions from 42.4 hours (b) onwards. The resonances at the low frequencies and the energy drop at the 54 order line at 43.75 hours seen in the spectrogram are obscured by the cross-terms.

*Figure 9.14 Windowed Wigner-Ville distributions of test gear G6 (cracked tooth)*
All Zhao-Atlas-Marks distributions have been calculated using a Gaussian window with an angle domain width of 90 degrees at the -40dB points. Excited resonances can be seen from 42.4 hours (b) onwards (circled), with the level of the resonances getting higher as damage progresses. From 42.6 hours (d) onwards (running at 24.5 kW), some modulation (smaller circle) of the tooth mesh frequency (27 orders) can be seen; there is also a larger modulation of the 4th harmonic of tooth mesh (108 orders) in these plots.

Figure 9.15 Zhao-Atlas-Marks distributions of test gear G6 (cracked tooth)
A slight frequency modulation can be detected in the ZAM (Figure 9.15) at the tooth meshing frequency (27 orders) and its 4th harmonic (108 orders) during the run at 24.5 kW (d)-(f), with the maximum deviation in frequency occurring after (in rotation angle) the initial excitation of the resonances; this is consistent with the findings for the Wessex input pinion crack and the analysis of this data using the narrow band demodulation technique in Chapter 7.

9.5 SUMMARY OF FINDINGS

Time-frequency analysis of the Wessex input pinion crack and tooth pitting using the techniques described in Chapter 8 provided far more information about the signal behaviour than any of the techniques studied in Chapter 6, with the Zhao-Atlas-Marks distribution being particularly effective in separating structural resonances and modulations of tooth meshing harmonics. The features associated with cracking and tooth pitting were different and the identifying features became more prominent as damage progressed in both cases. This is in contrast to the kurtosis based techniques FM4A and the narrow band envelope kurtosis, which had high values in the initial stages of damage for both cracking and pitting, with the values going down as damage progressed.

However, none of the time-frequency analysis techniques studied gave any clearly definable features for the tooth pitting from the spur gear test rig. For the cracked tooth from the spur gear test rig, the spectrogram, Wigner-Ville and Zhao-Atlas-Marks distributions identified excited structural resonances but did not provide any clear indication of the expected frequency/amplitude modulations of the tooth meshing harmonics. This is thought to be due to the relatively small levels of modulation for this gear.

Methods of overcoming these limitations will be investigated in the next chapter.
Chapter 10

A NEW TIME-FREQUENCY ANALYSIS TECHNIQUE

In the previous chapter it was seen that, although existing time-frequency analysis techniques could provide useful diagnostic information, they have limitations when it comes to the representation of short term frequency modulations which are relatively small in magnitude. It was seen in early chapters that short term phase modulations (and hence frequency changes) have an important role in the diagnosis of gear faults, particularly cracks.

In this chapter, a time-frequency analysis technique is developed which is able to detect and track small frequency deviations without sacrificing time (angle) resolution.

10.1 THEORETICAL DISCUSSION

It was found in the previous chapter that the spectrogram and Zhao-Atlas-Marks distribution provided the clearest description of the underlying structure of the vibration signals analysed. Although the Wigner-Ville distribution and the Choi-Williams distribution provide a more ‘mathematically correct’ description of a signal by meeting the time and frequency marginal conditions (Cohen [23] and Choi and Williams [18]), from the point of view of vibration diagnostics, this is not particular important. We are more concerned with having an ‘intuitively correct’ signal representation, that is, one which gives a sensible description of the behaviour of the individual signal components. The interfering cross-terms in the Wigner-Ville distribution limits its use as a visual diagnostic tool. Although the Choi-Williams distribution reduces the maximum amplitudes of the cross-terms, the spread of the cross-term energy over a wide frequency and/or time can reduce the ability to detect low energy features which may play an important role in the diagnostic process.
It was shown in Chapter 8 that the Zhao-Atlas-Marks distribution is basically a modified spectrogram with higher time-frequency resolution and finite time support. Rather than attempting to improve upon the Zhao-Atlas-Marks distribution, an alternative method will be derived based on the spectrogram.

### 10.1.1 Energy, frequency and bandwidth

Before entering into a discussion of the desirable properties of a time-frequency distribution for gear fault vibration analysis, a brief review and definitions of the physical properties of a signal will be made.

It was seen in Chapter 8 that for the signal

\[ s(t) = a(t)e^{j\varphi(t)} = a(t)e^{\int_{0}^{t}(\phi + 2\pi f_i(\tau))d\tau}, \]  

we can define the following signal properties:

\[ E_i(t) = |s(t)|^2 = \text{energy density per unit time}, \]

\[ E = \int |s(t)|^2 dt = \text{total energy}, \text{ and} \]

\[ f_i(t) = \frac{1}{2\pi} \varphi'(t) = \text{instantaneous frequency}. \]

Cohen and Lee [20] made the following observations relating to the average frequency and bandwidth of a signal.

The average frequency of a signal over its duration is given by its first moment in frequency

\[ \langle f \rangle = \frac{1}{E} \int f |S(f)|^2 df = \text{average frequency} \]  

(10.2)
and the bandwidth is the standard (root mean square) deviation of the frequency about the average

\[ B^2 = \frac{1}{E} \left\{ \left( f - \langle f \rangle \right)^2 \right\} S(f)^2 \, df = \text{effective bandwidth squared.} \quad (10.3) \]

It was shown [20] that the average frequency and bandwidth could be expressed as functions of the time domain signal; with the average frequency being

\[ \langle f \rangle = \frac{1}{E} \int f_i(t) \left| s(t) \right|^2 \, dt \quad (10.4) \]

and the bandwidth being expressed as

\[ B^2 = \frac{1}{E} \int \left[ \left( \frac{a'(t)}{2\pi a(t)} \right)^2 + (f_i(t) - \langle f \rangle)^2 \right] \left| s(t) \right|^2 \, dt. \quad (10.5) \]

The remarkable notion expressed in this equation is that, although the average frequency over the duration of the signal is the average value of the instantaneous frequency weighted by the absolute square of the signal (10.4), the bandwidth of the signal is not just the mean square deviation of the instantaneous frequency about the average frequency. This suggest that, even for monocomponent signals, the instantaneous frequency is itself an average, with the spread of frequencies about the instantaneous frequency being related to the derivative of the amplitude. Cohen and Lee [20] suggested that the definition of a monocomponent signal be restricted to exclude amplitude modulated signals (i.e., a monocomponent signal is one which has only a single frequency at a particular time) however, as was seen in Chapter 8, a signal which is both amplitude and frequency modulated can be expressed meaningfully in the time domain and, in the opinion of the author, the separation of such a signal into its constant amplitude components is not only unnecessary, it is not meaningful. For example, consider the amplitude modulated signal

\[ A(1 + \beta \cos(2\pi f_0 t))e^{j2\pi f_1 t} \]
which can easily be described as a signal at frequency $f_1$ with a mean amplitude of $A$ and a sinusoidal amplitude modulation at a frequency of $f_0$ and a magnitude of $\beta$. However, if we describe the same signal as the sum of its constant amplitude components,

$$A e^{j2\pi f_1 t} + \frac{\beta}{2} e^{j2\pi (f_1 + f_0) t} + \frac{\beta}{2} e^{j2\pi (f_1 - f_0) t},$$

the physical meaning is obscured.

The bandwidth expressed in equation (10.5) led Cohen and Lee [20] to introduce the concept of an *instantaneous bandwidth*,

$$b_i^2(t) = \frac{1}{4\pi^2} \left( \frac{a'(t)}{a(t)} \right)^2 = \text{instantaneous bandwidth squared.} \quad (10.6)$$

### 10.1.2 Desirable properties

What properties do we want this distribution to have? It was shown in Chapter 8 that *non-stationary monocomponent* signal can be meaningfully represented in the time domain but not the frequency domain and a *stationary multicomponent* signal can be meaningfully represented in the frequency domain but not the time domain. Based on this, we can state a set of meaningful ‘marginal’ conditions in time for *monocomponent* signals and a set of ‘marginal’ conditions in frequency for *stationary* signals to ensure the time-frequency distribution meaningfully represents both these known signal types. However, requiring that the distribution meet all conditions for all signal types is unnecessarily restrictive and may, in fact, be counter productive. Three separate sets of properties will be stated; a set of global properties reflecting what we know about the total signal, a set of properties in time reflecting what is known about a monocomponent signal, and a set of properties in frequency reflecting what is known about a stationary signal.

For the time-frequency energy distribution $\rho(t,f)$ the desirable properties are as follows.
10.1.2.1 Global properties

The global properties relate to what is known about any signal type, and should be met by the distribution for all signals.

**Property G1.** The total energy in the signal is reflected in the distribution. That is,

\[
\int \int \rho(t, f) \, dt \, df = \int |s(t)|^2 \, dt = \int |S(f)|^2 \, df = E = \text{total energy.} \quad (10.7)
\]

**Property G2.** The distribution is positive for all time and frequency.

**Property G3.** The distribution is real valued.

10.1.2.2 Time properties

For the monocomponent signal

\[
s(t) = a(t) e^{j\phi(t)} = a(t) e^{\int_{0}^{t} \{ \phi + 2\pi \int_{0}^{t} f_i(\tau) \, d\tau \}}, \quad (10.8)
\]

the distribution should have the following properties.

**Property T1.** The integral of the distribution over frequency at time \( t \) should equal the energy density at time \( t \) (the time marginal, Cohen [23]):

\[
\int \rho(t, f) \, df = |s(t)|^2. \quad (10.9)
\]

**Property T2.** The first moment of the distribution in frequency equals the instantaneous frequency:

\[
\int \frac{f \rho(t, f) \, df}{\int \rho(t, f) \, df} = f_i(t) = \frac{1}{2\pi} \varphi'(t). \quad (10.10)
\]

The first two time properties are well known and commonly stated as desirable properties (or requirements) of a time-frequency distribution. From the discussions
related to instantaneous frequency and bandwidth in the previous section, it was concluded that the instantaneous frequency is in itself an average frequency therefore, in order to describe the spread in frequency about the instantaneous frequency, the instantaneous bandwidth (10.6) is used, leading to

**Property T3.** The second moment of the distribution in frequency equals the instantaneous bandwidth:

$$
\frac{\int \left( f - f_i(t) \right)^2 \rho(t, f) \, df}{\int \rho(t, f) \, df} = b_i^2(t) = \frac{1}{4\pi^2} \left( \frac{a'(t)}{a(t)} \right)^2.
$$

(10.11)

### 10.1.2.3 Frequency properties

For a multicomponent stationary signal

$$
s(t) = \int S(f) e^{j2\pi ft} \, df,
$$

(10.12)

the distribution should have the following property.

**Property F1.** The integral of the distribution over time at frequency $f$ should equal the energy density at frequency $f$ (the frequency marginal, Cohen [23]):

$$
\int \rho(t, f) \, dt = |S(f)|^2.
$$

(10.13)

Although we could specify properties relating to the group delay and instantaneous duration [20] of the signal at frequency $f$, these would only be meaningful for a signal which has continuity in frequency. Because this property is not possessed by gear vibration data (which consists mainly of isolated frequency components), restrictions relating to the group delay and instantaneous duration will not be placed on the distribution discussed here.
10.2 THEORETICAL DEVELOPMENT

10.2.1 The spectrogram

The spectrogram will be used as a starting point for the development of a time-frequency distribution with the required properties. As a first step, the spectrogram will be studied to see which of the desired properties it does or does not possess.

The spectrogram is defined as the square of the magnitude of the short-time Fourier transform (STFT) ([22] and Chapter 8),

$$\rho_{st}(t, f) = |S_{st}(f)|^2, \quad \text{where}$$

$$S_{st}(f) = \int s(\tau)h(\tau - t)e^{-j2\pi f\tau} d\tau.$$  \hfill (10.14)

It was shown in Chapter 8 that the spectrogram has all the properties we have listed as global properties above, that is, it preserves the total energy in the signal (if the total energy in the window is 1), is positive and real valued.

It was shown in Chapter 8 that the spectrogram meets none of the time or frequency related properties we have specified for our distribution.

Note that the spectrogram is the square of the magnitude of the STFT (10.14) which, taking the inverse Fourier transform at time $t$, gives

$$\int S_t(f)e^{j2\pi ft} df = s(t)h(0)$$  \hfill (10.15)

and the integral over time at a particular frequency is

$$\int S_t(f) dt = \int s(\tau)e^{-j2\pi f\tau} \int h(\tau - t) dt d\tau$$

$$= S(f)H(0).$$  \hfill (10.16)
Therefore, although the spectrogram does not meet the marginal conditions for an energy distribution, the underlying short-time Fourier transform can easily be related back to the original signal (i.e., it has ‘marginal’ values related to the original signal).

### 10.2.2 The short-time inverse Fourier transform

We can define the short-time inverse Fourier transform (STIFT) of a signal as

\[
S_f(t) = S_t(f)e^{j2\pi ft} = \int s(\tau + t)h(\tau)e^{-j2\pi \tau}d\tau
\]

which has an integral over frequency at a particular time of

\[
\int S_f(t)df = s(t)h(0)
\]

and its Fourier transform is

\[
\int S_f(t)e^{-j2\pi ft}dt = S(f)H(0).
\]

Note that the square of the magnitude of the STIFT is identical to that of the STFT therefore the spectrogram can be defined from either,

\[
\rho_s(t, f) = |S_t(f)|^2 = |S_f(t)|^2.
\]

However, the STIFT provides a more convenient basis from which to develop the new distribution, as will be seen later.

### 10.2.3 The STIFT of the signal and its time derivative

If we make the same assumptions for the STIFT as is made for the STFT of a monocomponent signal (see Chapter 8), that is, at each fixed time of interest \( t \) the windowed signal approximates a stationary signal with the amplitude and frequency
being the instantaneous amplitude and frequency of the signal at time \( t \), then the STIFT (equation (10.17)) of the signal can be approximated as

\[
S_f(t) \approx \int a e^{j\phi} h(\tau) e^{-j2\pi(f-f_i)\tau} d\tau
\]

\[
\approx H(f-f_i)ae^{j\phi},
\]

(10.21)

where the constants \( a, \phi \) and \( f_i \) are equivalent to the time based variables \( a(t), \phi(t) \) and \( f_i(t) \) respectively in equation (10.8).

The time derivative of \( s(t) \) (10.8) is

\[
s'(t) = (a'(t) + ja(t)\phi'(t))e^{j\phi(t)}
\]

(10.22)

which has a STIFT (using the same assumptions as in (10.21)) of

\[
S'_f(t) = \int s'(\tau + t)h(\tau)e^{-j2\pi\tau} d\tau
\]

\[
\approx \int (a' + ja\phi')h(\tau)e^{j\phi} e^{-j2\pi(f-f_i)\tau} d\tau
\]

\[
\approx H(f-f_i)(a' + ja\phi')e^{j\phi},
\]

(10.23)

where the constants represent the values of the equivalent time based variables at the fixed time of interest \( t \).

Dividing (10.23) by (10.21) gives

\[
D_f(t) = \frac{S'_f(t)}{S_f(t)} = \frac{a'(t)}{a(t)} + j\phi'(t)
\]

(10.24)

which is the Poletti ‘dynamical’ signal [20] (originally defined as the derivative of the log of the signal). Note that, for a monocomponent signal, the value of (10.24) will be the same at all frequencies and is independent of the window.
10.2.4 Redistribution of energy

From equation (10.24) we can easily obtain an estimate of the instantaneous frequency at each time-frequency location,

\[ f_i(t) = \frac{1}{2\pi} \varphi'(t) \approx \frac{1}{2\pi} \text{Im}[D_f(t)]. \tag{10.25} \]

and an instantaneous bandwidth estimate of

\[ b_i(t) = \sqrt{\frac{1}{4\pi^2} \left( \frac{a'(t)}{a(t)} \right)^2} = \frac{1}{2\pi} \left( \frac{a'(t)}{a(t)} \right) = \frac{1}{2\pi} \text{Re}[D_f(t)]. \tag{10.26} \]

Using the estimates of instantaneous frequency and bandwidth, the signal energy estimates from the spectrogram (10.20) can be redistributed in the time-frequency plane by using a function centred about the instantaneous frequency estimate (10.25) which has a standard deviation of \( b_i(t) \) (10.26). That is, we define a new distribution

\[ \rho_r(t, f) = \rho_s(t,u) g_i(u, f) du, \tag{10.27} \]

where \( g_i(u,f) \) is a function with the properties,

\[ \int g_i(u,f) df = 1, \]
\[ \int f g_i(u,f) df = \frac{1}{2\pi} \text{Im}[D_u(t)] = f_u(t), \quad \text{and} \]
\[ \int \left( f - f_u(t) \right)^2 g_i(u,f) df = \frac{1}{4\pi^2} \left| \text{Re}[D_u(t)] \right|^2. \tag{10.28} \]

A function which meets these properties is the normal (Gaussian) distribution,
\[ g_t(u, f) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(f-f_i)^2/(2\sigma^2)} \),

where

\[ f_i = \frac{1}{2\pi} \text{Im}[D_u(t)], \quad \text{and} \]

\[ \sigma = \frac{1}{2\pi} \left| \text{Re}[D_u(t)] \right|. \]

10.2.5 Properties of the new distribution

The distribution defined by equation (10.27) has all the properties of the spectrogram (i.e., it is real, positive and preserves the energy in the signal) plus:

10.2.5.1 For monocomponent signals

\[ \int \rho_r(t, f) df = \int \rho_s(t, u) \int g_t(u, f) df \, du = \int \rho_s(t, u) du \]

\[ = \int |H(u-f_i)|^2 a^2 \, du = |s(t)|^2 \int |H(u-f_i)|^2 du, \]

which, if the energy in the window is 1, meets the first time property T1 (10.9) within the limits of the assumption made on the stationarity of the signal over the duration of the window. Note, that for monocomponent signals, we can make the window infinitely short (i.e., a delta function), and equation (10.30) will be exact.

The other time properties follow from the property of the energy redistribution function \( g_t(u, f) \) given in (10.28), that is, the energy will be centred about the instantaneous frequency (property T2) with the spread being equal to the instantaneous bandwidth (property T3).

10.2.5.2 For stationary multicomponent signals

For a stationary multicomponent signal,

\[ s(t) = \sum_c A_c e^{j(\theta_c + 2\pi f_c t)} \]
the STIFT is,

\[
S_f(t) = \sum_c A_c e^{j(\theta_c + 2\pi f_c \tau)} h(\tau) e^{-j2\pi(f-f_c)\tau} d\tau
\]

\[= H(f) \left( \sum_c A_c e^{j(\theta_c + 2\pi f_c \tau)} \delta(f - f_c) \right). \tag{10.31}\]

and, similarly, the STIFT of the signal derivative is,

\[
S'_f(t) = H(f) \left( \sum_c j2\pi f_c A_c e^{j(\theta_c + 2\pi f'_c \tau)} \delta(f - f_c) \right). \tag{10.32}\]

If we make the window infinitely long in time, such that it becomes a delta function in frequency (at f=0) (i.e., h(t) = 1), then the STIFT of the signal and its derivative reduce to,

\[
S_{f_c}(t) = A_c e^{j(\theta_c + 2\pi f_c \tau)}, \quad \text{and} \quad S'_{f_c}(t) = j2\pi f_c A_c e^{j(\theta_c + 2\pi f'_c \tau)}. \tag{10.33}\]

From equation (10.33), it can be seen that the bandwidth (10.26) at \(f_c\) will be zero and the instantaneous frequency (10.25) will be \(f_c\) for all values of \(t\). That is, no redistribution of the signal energy will be performed and the integral over time at frequency \(f_c\) will be equal to the energy density in frequency at that frequency (property F1).

**10.2.5.3 For non-stationary multicomponent signals**

We have seen that the limiting factor in the distribution is the window. If we take the window equal to a delta function in time, the distribution will perfectly represent a non-stationary monocomponent signal, and if we take the window equal to a delta function in frequency, we will have a perfect representation of a stationary multicomponent signal.
However, what happens if we have a non-stationary multicomponent signal? Here we are limited to the same compromises inherent in the spectrogram. That is, we need to trade off time and frequency resolution. If the window is sufficiently short in time to assume that the signal is stationary over the window duration, and sufficiently narrow in frequency to separate the components (i.e., the bandwidth of the window is less than the frequency separation of components at a particular time), then we will get a near perfect representation of the individual components; that is, the time properties will be met for each component but not necessarily the entire signal.

It is important to note that, although the concentration of energy about the instantaneous frequency is dramatically improved over the spectrogram, the resolution in frequency is not changed.

### 10.2.6 Relationship to the Reassignment Method

The new time-frequency analysis method proposed here bears some resemblance to the ‘Modified Moving Window Method’ proposed by Kodera et al. [42,43] and recently re-examined and extended to other distributions (‘The Reassignment Method’) by Auger and Flandrin [2]. The major premise behind the reassignment method is that the applied window causes a smearing of the distribution in time and frequency which can be partially ‘undone’ by changing the attribution point of the calculated energy to the centre of gravity of the energy contributions at each time-frequency ‘bin’.

Using this method on the spectrogram gives the same attribution point in frequency as the ‘instantaneous frequency’ of the method proposed here with a similar relocation in time also being applied. The new method proposed here does not relocate the energy in time. However, the time relocation is a simple extension of the development for the frequency relocation given previously, with the instantaneous frequency ‘marginal’ being replaced by the group delay and the instantaneous bandwidth being replaced by the instantaneous duration. The reason this is not done here is that gear vibration signals are predominantly discrete in the frequency domain (i.e., composed mainly of discrete sinusoidal components, being the tooth meshing harmonics) and, because the signal is
not continuous in frequency, the frequency derivative of the signal will be ill defined (e.g., for a single stationary sine wave we would have an infinite duration with an undefined group delay).

### 10.2.7 The need for bandwidth adjustment

The other major difference between the method proposed here and the Reassignment Method is the use of the instantaneous bandwidth to spread the signal about the instantaneous frequency. This is not done in the reassignment method. Apart from the mathematically pleasing result it gives, the bandwidth adjustment has a significant practical value. Note that even if the signal is discrete (i.e., sampled), the definition of the new distribution including the bandwidth modification function (10.30), is continuous in frequency. This allows us to ‘zoom’ in on a small frequency range and see changes in the signal which may well be below the frequency ‘resolution’. Without bandwidth adjustment, although we have reassignment in frequency we will still have a number of discrete frequency ‘lines’. Although we would expect these individual lines to be physically at the same location (the instantaneous frequency) this is not necessarily the case due to the limitations imposed by the window and the effects of amplitude modulation (causing a spread in frequency).

Figure 10.1 shows the new distribution applied to signal with a ‘carrier’ frequency of 88 orders which has a twice per revolution sinusoidal frequency modulation of ±5 orders (the spectrogram of this signal is shown in Figure 8.6). The plots have been calculated using a 31.64 degree window and zoomed to show the frequency range 78 to 98 orders (note that the spectrogram using the same window shown in Figure 8.6(b) has energy present in the range 60-120 orders). Figure 10.1 (a) shows the signal obtained by frequency reassignment only (no bandwidth adjustment) and (b) shows the signal with bandwidth adjustment applied. Note that there is little difference in these two plots apart from a slightly smoother appearance of the bandwidth adjusted signal. The differences here are caused solely by the windowing of the signal (i.e., the instantaneous bandwidth of the signal is zero for all rotation angles, therefore, apart from the influence of the window there should be no difference in these signals).
Figure 10.1 New distribution of frequency modulated signal (31.64 degree window) (Zoomed to 78-98 orders)

Figure 10.2 shows a signal that has both amplitude and frequency modulation. This signal has a carrier frequency of 88 orders with a four times per revolution sinusoidal amplitude modulation (with peak amplitude occurring at 62 degrees, 152 degrees etc.) and a twice per revolution sinusoidal frequency modulation of ±5 orders (with minimum frequency of 83 orders occurring at 45 and 225 degrees).

Figure 10.2 New distribution of amplitude and frequency modulated signal. (31.64 degree window. Zoomed to 78-98 orders)

Figure 10.2 (a) shows the signal without bandwidth adjustment and (b) shows the new distribution with bandwidth adjustment (both plots use a window of 31.64 degrees and are zoomed to 78-98 orders). The effect of the bandwidth adjustment is quite visible here. Apart from the smoother appearance of the bandwidth adjusted signal in (b), the maximum energy points are correctly located at 62 degrees, 152 degrees etc., whereas
the distribution without bandwidth adjustment in (a) has maximum energy points at the minimum and maximum frequencies. The time domain demodulation of this signal is shown in Figure 8.4 and its spectrogram is shown in Figure 8.7 (a).

10.3 EXAMPLES USING THE NEW DISTRIBUTION

10.3.1 Simulated signals

We have already seen the (zoomed) results of the new distribution applied to non-stationary monocomponent signals in Figure 10.1 and Figure 10.2 (with and without the bandwidth adjustment). The bandwidth adjustment will be used in all remaining examples.

Figure 10.3 shows the new distribution applied to the example signals used in Chapter 8 for (a) a multicomponent signal with a stationary signal at 40 orders and a frequency and amplitude modulated signal at 88 orders (as per Figure 10.2) and (b) the test signal which was used to demonstrate conflicting window requirements. Here we have used a medium length window of 90 degrees as opposed to the 45 and 180 degree windows used for the examples in Chapter 8. This ‘compromise’ window was not used for any of the distributions in Chapter 8 as it gave a poor representation of all of the signal components in all cases. Figure 10.3 shows a more pleasing signal representation than any of the other distributions for these signals (see Chapter 8), with a higher concentration of the signal components about their (individual) instantaneous frequencies and very little cross-term interference. The distribution is able to give a relatively good representation of the two closely spaced sine waves in (b) using the ‘compromise’ window because the bandwidth adjustment tends to spread the energy away from the midpoint cross-terms and concentrate it at the ‘auto-terms’.
the distribution without bandwidth adjustment in (a) has maximum energy points at the minimum and maximum frequencies. The time domain demodulation of this signal is shown in Figure 8.4 and its spectrogram is shown in Figure 8.7 (a).

10.3 EXAMPLES USING THE NEW DISTRIBUTION

10.3.1 Simulated signals

We have already seen the (zoomed) results of the new distribution applied to non-stationary monocomponent signals in Figure 10.1 and Figure 10.2 (with and without the bandwidth adjustment). The bandwidth adjustment will be used in all remaining examples.

Figure 10.3 shows the new distribution applied to the example signals used in Chapter 8 for (a) a multicomponent signal with a stationary signal at 40 orders and a frequency and amplitude modulated signal at 88 orders (as per Figure 10.2) and (b) the test signal which was used to demonstrate conflicting window requirements. Here we have used a medium length window of 90 degrees as opposed to the 45 and 180 degree windows used for the examples in Chapter 8. This ‘compromise’ window was not used for any of the distributions in Chapter 8 as it gave a poor representation of all of the signal components in all cases. Figure 10.3 shows a more pleasing signal representation than any of the other distributions for these signals (see Chapter 8), with a higher concentration of the signal components about their (individual) instantaneous frequencies and very little cross-term interference. The distribution is able to give a relatively good representation of the two closely spaced sine waves in (b) using the ‘compromise’ window because the bandwidth adjustment tends to spread the energy away from the midpoint cross-terms and concentrate it at the ‘auto-terms’.
10.3.2 Wessex input pinion cracking

Figure 10.4 shows the new distribution (simple labelled as the ‘modified spectrogram’) of the signal averages for the cracked Wessex input pinion (see Chapter 6 and Chapter 9) for 75 and 100% load conditions at 233, 103 and 42 hours before failure.

The initial identifying feature for the crack is the simultaneous amplitude drop and frequency change at the 44 order line (2 x tooth meshing frequency) which can be seen quite clearly at 103 hours before failure for both the 75% load case (c) and the 100% load case (d). The other time-frequency distributions applied to these signals in Chapter 9, showed the amplitude drop but none of them clearly indicated the frequency change which is so apparent using the modified spectrogram.

At 43 hours before failure, the excited resonances seen using the other time-frequency analysis techniques are also apparent using the modified spectrogram, Figure 10.4 (e) and (f). In addition, the amplitude and phase modulation of the 44 order line is easily detectible. It is clear that the maximum amplitude drop and frequency deviation at the 44 order line occur after the impact exciting the resonances, which is consistent with the findings in Chapters 6 and 9 for these signals.
All modified spectrograms have been calculated using a Gaussian window with an angle domain width of 90 degrees at the −40dB points. The initial indication of the crack at 103 hours before failure (feature indicated on plots (c) and (d)) is clearer than other techniques with a distinct frequency modulation and drop in amplitude being visible because of the higher concentration and accuracy in frequency. The frequency location of the excited resonances (arrowed) is quite clear.

Figure 10.4 Modified spectrograms of cracked Wessex input pinion
(a) 27.7 hours since overhaul - 100% load  
(b) 201.1 hours since overhaul - 100% load

(c) 248.9 hours since overhaul - 100% load  
(d) 292 hours since overhaul - 100% load

(e) 339.5 hours since overhaul - 75% load  
(f) 339.5 hours since overhaul - 100% load

All modified spectrograms have been calculated using a Gaussian window with an angle domain width of 90 degrees at the -40dB points. The initial identifying feature is the excited resonance at 201.1 hours (b) (circled). As damage progresses (c)-(f), there is a noticeable drop in amplitude of the 22 order line (tooth meshing frequency) which occurs before (in rotation angle) the excited resonance.

*Figure 10.5 Modified spectrograms of pitted Wessex input pinion*
10.3.3 Wessex input pinion tooth pitting

Figure 10.5 shows the modified spectrograms of signal averages for the Wessex input pinion with tooth pitting, which was analysed using conventional vibration analysis methods in Chapter 6 and other time-frequency methods in Chapter 9. The excited resonance in the initial stages of pitting (b) at 201.1 hours since overhaul, is quite clear (as was the case for all other distribution except the Wigner-Ville distribution). However, as damage progresses, the modified spectrogram gives a slightly different picture to the other distributions. We can see, just prior to the resonance in rotation, a region of low energy on the 22 order (tooth meshing frequency) line. Although this was identified using the other time-frequency analysis techniques in Chapter 9, in Figure 10.5 we can see that this low energy region becomes progressively longer (in rotation) as the damage progresses and there is also a distinct frequency modulation associated with it. At this stage, the physical interpretation of this is uncertain.

10.3.4 Spur gear test rig tooth pitting

Figure 10.6 shows the results of applying the modified spectrogram, including the zoomed version, to some of the signal averages from the pitting on the teeth of a gear from the spur test rig described in Chapter 7. The results of applying the other time-frequency analysis techniques (spectrogram, Wigner-Ville, Choi-Williams and Zhao-Atlas-Marks) were inconclusive for these signals (see Chapter 9).

The modified spectrogram was applied to the signal averages taken at 102.9, 107.7 hours and 115.5 hours with the results being displayed over the frequency range 0 to 150 orders ((a),(c) and (e)) and a zoomed region of plus and minus 10 orders about the second harmonic of tooth mesh (44-64 orders) being shown in (b), (d) and (f). It was assumed that the gear teeth were not pitted at 102.9 hours as the gear rig was dismantled just after this run and pitting was not noticed. However, this does not guarantee that pitting was absent, just that it wasn’t detected.
All modified spectrograms have been calculated using a Gaussian window with an angle domain width of 90 degrees at the -40dB points. The plots on the left hand side are for 0 to 150 orders with the plot on the right being the zoomed version between 44 and 64 orders. (Note that these signals are not phase aligned and the 0 degree point is not necessarily the same for each plot). This signal shows a similar progression to that of the Wessex input pinion pitting.

*Figure 10.6 Modified spectrograms of test gear G3 (pitted teeth)*
The modified spectrogram at 102.9 hours over 0 to 150 orders, Figure 10.6 (a), suggests that there may have been pitting, or some other surface defect at this time. There is a excited resonance between 135 and 150 orders (circled). The zoomed modified spectrogram (b), shows a distinct frequency change which occurs before (in rotation) the excited resonance; there is no noticeable amplitude change at this stage. At 107.9 hours, we know that one tooth is pitted. Here the 0 to 150 order signal (c) shows the same excited resonance between 135 and 150 orders plus a distinct frequency modulation of the tooth meshing harmonics. The zoomed version of the signal at 107.9 hours in (d) shows a distinct increase (in relation to (b)) in the amount of frequency modulation and there is a slight drop in amplitude.

As damage progresses, the amplitude and frequency modulations increase, with the duration (in rotation angle) of the modulated region also increasing.

10.3.5 Spur gear test rig - cracked tooth.

The modified spectrograms of the signal averaged vibration from the cracked tooth in the spur gear test rig (see Chapter 7) are shown in Figure 10.7, with the frequency ranges 0 to 150 orders and 44 to 64 orders shown for the signal at (a) and (b) 33.2 hours (no cracking), (c) and (d) with initial crack and (e) and (f) just prior to (estimated) total failure. Although the other time-frequency distributions applied to this data in Chapter 9 detected the excited resonance between 105 and 120 orders, none of them showed the modulation of the teeth meshing harmonics which can be seen clearly in the modified spectrogram as the crack develops. The zoomed version of the modified spectrogram around the second tooth mesh harmonic (44 to 64 orders) shows this quite dramatically.

Note that the zoomed distribution of the uncracked gear (b) shows a slight yet distinct twice per revolution frequency modulation. This is due to misalignment.
All modified spectrograms have been calculated using a Gaussian window with an angle domain width of 90 degrees at the -40dB points. The plots on the left hand side are for 0 to 150 orders with the plot on the right being the zoomed version between 44 and 64 orders.

*Figure 10.7 Modified spectrogram of test gear G6 (cracked tooth)*
10.4 SUMMARY

A new time-frequency distribution has been developed based on a set of properties appropriate to the task of analysing gear fault vibration. These requirements were based on the observed deficiencies of existing distributions when applied to actual fault data.

The new distribution has high accuracy and is continuous in frequency, which allows the application of a zoomed version of the distribution. It has been shown that this technique is able to represent small short-term variations in frequency and amplitude, which are the features which provide the early indications of a fatigue crack.

The technique was shown to be more suited to gear vibration analysis than other time-frequency analysis techniques, and has some distinct advantages over ‘conventional’ methods. The ability to identify and separate structural resonances and modulations provides far greater insight into the behaviour of the vibration signal.
11.1 SUMMARY OF FINDINGS

Methods for the early detection and diagnosis of geared transmission system faults have been investigated, with particular attention being paid to safety critical faults in helicopter transmission systems. During the course of this research:

a) The mechanisms involved in the production of vibration from geared transmission systems were reviewed and a general model of gearbox vibration was developed. This model is based on the angular relationships of the transmission system components which enables processes leading to non-stationary signals, such as speed/load fluctuations and variable transmission path effects, to be modelled as simple angular dependencies. This is in contrast to previous models of transmission system vibration, which were based on the frequency domain representation of vibration and had difficulties in describing non-stationary processes.

b) A review was made of component failure modes, the potential consequences of failure, and the diagnostic evidence produced by particular faults. This resulted in the identification of safety critical faults in helicopter transmission systems and the procedures which could be used to detect these faults.

c) A review of existing vibration analysis techniques was made, with particular attention being paid to the expected value of each technique in the detection and diagnosis of safety critical faults in helicopter transmission systems. From this, it was clear that methods based on synchronous signal averaging of the vibration data were the most appropriate diagnostic techniques for the safety critical failure modes identified.
d) A detailed investigation of synchronous signal averaging techniques was undertaken. This resulted in the development of

i) a method of quantifying the attenuation of non-synchronous vibration signals,

ii) a formula for identification of the ‘ideal’ number of averages for any signal,

iii) a definition of the signal-to-noise ratio for synchronously averaged vibration data,

iv) a method for the optimisation of the number of averages, and

v) two new methods for the digital resampling of discrete vibration signals using spline functions based on signal derivatives obtained via differentiating filters; these methods proved to provide higher accuracy and/or greater efficiency over previous methods used for digital resampling.

e) An investigation of the use of existing vibration analysis techniques based on synchronous signal averaging was made using in-flight fault data from Wessex helicopter main rotor gearboxes. This showed that, although some of the existing techniques provided good detection capabilities for gear cracking and tooth pitting, diagnosis of these faults was difficult. The technique which provided the most detailed diagnostic information was the narrow band demodulation technique (McFadden [55]). A new method of performing the demodulation was developed which allows for ‘negative’ amplitudes. This was required to provide physically meaningful results in the presence of high amplitude additive vibration (e.g., excited resonances). Although the demodulated amplitude and phase provided useful diagnostic information, the interpretation of the results was not simple and incorrect selection of the analysis band could give misleading results.

f) In order to provide gear fault data under controlled conditions, a spur gear test rig was developed. Aircraft quality gears were specially manufactured for this rig and, rather than using artificial ‘faults’, a small implanted stress riser was used in an attempt to initiate a realistic tooth crack. Considerable difficulty was encountered (and time spent) initiating a crack. The gear rig and test specimens underwent a
number of modifications before crack initiation and propagation was eventually achieved. Analysis of the cracked tooth and a gear with pitted teeth (which occurred ‘naturally’ during the course of the experiment) using existing vibration analysis techniques showed similar results to those obtained for the helicopter gear faults.

g) An alternative approach to gear fault diagnosis was proposed, based on the identification and interpretation of features actually present in the vibration signal rather than making prior assumptions about what features should, or should not, be present (as is done with existing techniques). A review of general signal theory was made which showed that multicomponent non-stationary signals, similar to those produced by tooth faults in geared transmission signals, could not be adequately portrayed in either the time or frequency domain. Because of this, a study of joint time-frequency domain signal analysis was made. Although these techniques have been studied theoretically for over fifty years, and been used in a number of practical applications over the last few years, they had not previously been applied to the analysis of gear fault vibration.

h) Application of various time-frequency analysis techniques to the in-flight gear faults showed that these techniques provided far greater diagnostic information than existing vibration analysis techniques, particularly in the identification of excited resonances and the separation of these from short term amplitude and frequency modulations. This was achieved without the need to apply any other ‘enhancement’ to the signal averaged data (e.g., removal of tooth meshing harmonics or narrow band filtering of the signal). It was also shown that the features identifying the fault become more prominent as damage progresses; this is not the case with existing techniques, where the indication of damage is often more pronounced in the early stages and reduces as damage becomes more severe. However, these techniques did not perform well on the vibration data from the spur gear test rig faults. It was shown that this was because the existing time-frequency analysis techniques did not clearly portray the small short term frequency modulations which provide the major diagnostic information present in these signals.
i) A new time-frequency energy distribution was developed based on the perceived requirements for a time-frequency representation of gear vibration data. This distribution is real and positive for all time and frequency and preserves the energy in the signal. Instead of requiring that certain ‘marginal’ conditions are meet for all signals, the distribution was based on the ‘intuitive’ notion that different types of signals have different time and frequency properties. A set of time ‘marginals’ were defined which the distribution was required to meet for non-stationary monocOMPONENT signals. These were that the integration over frequency at a particular time equals the energy at that time, the first moment in frequency equals the instantaneous frequency and the second moment equals the ‘instantaneous bandwidth’. For stationary multicomponent signals the integration over time at a certain frequency was required to equal the energy at that frequency. For non-stationary multicomponent signals, the time and frequency ‘marginals’ are partially met. That is, an approximation is made which, it is assumed, gives a reasonable representation of the individual signal components but does not meet the ‘marginals’ for the total signal.

The new technique has high accuracy and continuity in frequency, which allows small short-term frequency modulations to be clearly seen. Because of the continuity in frequency, zooming can be used to examine small frequency ranges in fine detail. This technique clearly showed the small frequency deviations present in the spur gear test rig faults which could not be seen using other time-frequency representations.

### 11.2 RECOMMENDATIONS

For safety critical faults in helicopter transmission systems, it was shown that vibration analysis techniques based on synchronous signal averaging provided the best detection and diagnostic capabilities. Detailed study of the synchronous signal averaging process was undertaken and a method of determining the optimum number of averages for any shaft in a gearbox was developed. In practice, this optimisation process need only be
performed once for each gearbox, with the ‘optimum’ number of averages defined for each shaft being used for all subsequent analyses of that gearbox.

It was shown that digital resampling of the vibration data using higher order spline functions, with the signal derivatives being determined using differentiating filters, provides an efficient method of signal reconstruction which introduces errors which are less than the dynamic range of the analogue-to-digital conversion (i.e., effectively error free). Using this method, in combination with the optimised number of averages, can provide accurate (and stable) synchronous signal averages on which a fault detection and diagnosis system can be based.

Because of the uncertainties in crack propagation rates, particularly in highly loaded systems such as helicopter transmissions, it would be preferable to perform continuous (or near continuous) vibration analysis for critical components. This requires a permanent on-board monitoring system.

Although it was shown that time-frequency analysis techniques (particularly the new technique developed here) provided far more diagnostic information than existing vibration analysis techniques, the existing techniques performed well as initial detectors of gear faults. Because of the relative simplicity and efficiency of techniques such as Stewart’s Figures of Merit [73] and McFadden’s narrow band envelope kurtosis [54], they are preferable to time-frequency analysis techniques as initial fault detectors.

Therefore, the most efficient and flexible system would be to perform synchronous signal averaging and initial fault detection using existing vibration analysis techniques on-board the aircraft, with detailed diagnosis using time-frequency analysis of the synchronous signal averages being performed post-flight and only when necessary (i.e., after a ‘fault’ detection). Because of the uncertainties involved in the cause of the ‘fault’ detection with existing techniques, it is not recommended that in-flight warnings be given based on these (except in extreme instances).
11.3 FUTURE WORK

It has been shown that time-frequency analysis techniques have advantages over current techniques for fault diagnosis in that

a) they require no prior assumptions about the nature of signal,

b) they are able to separate signal features related to particular fault(s) (e.g., excited resonances and modulations can be seen as separate features), and

c) the identifying feature(s) become more pronounced as damage progresses.

However, visual interpretation of these features is still required and further work is needed to make full use of the diagnostic information available with time-frequency analysis techniques. It is anticipated that the areas of research which would provide the most benefit in this respect are:

a) Investigation of pattern recognition techniques to automate the identification of various signal features.

b) Dynamic modelling and/or further generation of faults in experimental rigs to provide data for ‘forward modelling’ of identifying fault features.

c) Time-frequency filtering techniques may also be of benefit by allowing the separation of individual signal components in the time-frequency domain, with the analysis of each component taking place in the time and/or frequency domain using more traditional methods.
Appendix A

FAILURE MECHANISMS

In this appendix, details are given of the mechanisms leading to various types of failures in gears, bearings and shafts.

The term ‘failure’, although readily understood, cannot be strictly defined for components in a mechanical system without reference to the context in which the component is operating. A failed or faulty component is one which causes interruption or degradation of service [28]. It is quite possible that a condition which constitutes a failure in one instance, such as moderate tooth wear on a gear in a precision instrument, may not constitute a failure in another instance, such as a gear in a large mill drive where moderate wear is quite acceptable.

The following sections detail the mechanisms involved in the development of various conditions which may constitute faults in a geared transmission system.

A.1 GEAR FAILURE MODES

Drago [28] identifies a number of failure modes in gears.

A.1.1 Wear

The basic mechanism causing wear is insufficient lubricant film thickness allowing surface-to-surface contact between the mating surfaces of the teeth. Other factors may cause or aggravate wear, such as abrasive particles in the lubricant, corrosion of the tooth surfaces, or tooth surface irregularities which penetrate the lubricant film.
A.1.1.1 Polishing wear

Polishing wear is most frequently observed when relatively low-speed gears are operating with substantial surface-to-surface contact, causing a polishing of the tooth surfaces to an almost mirror-like finish. Further wear generally continues at a very low rate and, therefore, polishing wear is not often regarded as a failure.

A.1.1.2 Moderate wear

Moderate wear is usually caused by insufficient lubricant film thickness. Since wear is proportional to the sliding velocity and the sliding velocity varies from zero at the pitch line to a maximum at the extremes of contact, the tooth shows greatest wear at the tip and root and practically none at the pitch line. Moderate wear in itself is not usually seen as a failure, however, it is a prelude to excessive wear and ultimately to complete gear tooth failure. The rate of wear can often be reduced by increasing oil viscosity, using an extreme pressure oil, improving the tooth surface finish or changing the gear geometry to reduce the sliding velocity.

A.1.1.3 Excessive wear

If not corrected, moderated wear will progress to excessive wear, where the original tooth profile is destroyed, and ultimately catastrophic failure by tooth fracture due to

a) the tooth wearing so thin that its bending strength is exceeded, and/or

b) progression of cracks originating at points of tooth surface damage, and/or

c) high dynamic loads induced by tooth profile damage.

A.1.1.4 Abrasive wear

Abrasive wear is caused by particles in the lubricant with a hardness near or above that of the tooth surface and a diameter equal to or greater than the lubricant film thickness. To avoid abrasive wear, it is necessary to ensure that the lubrication is clean at all times;
using filters where possible and changing the lubricant frequently. The abrasive particles may be the result of another failure (such as a bearing failure).

A.1.1.5 Corrosive wear

Corrosion can cause destructive wear by damaging the finish of the tooth surface and reducing the area of contact which will increase the unit loading on the tooth surface. Both tooth surface damage and increased surface loading will lead to accelerated wear. The corrosion can be due to the breakdown of extreme pressure oil, outside contamination, or contaminants on the gears or other components at assembly.

A.1.2 Frosting, scoring and scuffing

Frosting, scoring or scuffing are caused by an instantaneous welding of the asperities of the tooth surfaces, followed by a breaking of the weld. This occurs when the combination of load, sliding velocity and oil temperature reaches a critical value causing a break down of the oil film separating the tooth surfaces. This allows metal-to-metal contact and, if the surface pressure and sliding velocity are high enough, welding will occur. The difference between frosting and scoring is the extent of the welding and the effect of breaking the welds. Scoring is generally observed only on high-speed, high-load gears operating with low-viscosity synthetic lubricants.

The terms frostning, scoring and scuffing are frequently used interchangeably, with no universal agreement on when each term should apply. Drago [28] uses only the terms frosting and scoring, providing arbitrary delineation on the degree of scoring.

A.1.2.1 Frosting

Frosting occurs when the extent of welding is such that only the extreme tips of the surface asperities are welded and subsequently broken off with little or no further damage. This gives the tooth surface the appearance of a frosted crystal which is caused by micropitting of the surface with no tear marks in the direction of sliding. The initial frosting (removal of the extreme tips of the surface asperities) can increase the contact
surface area, lowering the surface pressure, and the gears can often run for long periods of time with no further damage. If necessary, the frosting may be removed by polishing the affected area with a very fine grit paper and the gears returned to service without a recurrence of the problem.

A.1.2.2 Light to moderate scoring

When the combination of sliding, load and temperature are sufficiently above the critical value, gross instantaneous welding of the surface asperities occur and the subsequent breaking of the weld results in scratching of the tooth profiles as the tooth surfaces slide on one another. If not corrected, this condition will usually be progressive and lead to destruction of the tooth profile. Polishing of the tooth surface may correct the problem. In some circumstances, light or moderate scoring (like frosting) may cease or heal over with continued operation as the tooth surfaces asperities are reduced.

A.1.2.3 Destructive scoring

If the operating conditions are far beyond the critical point or scoring progresses beyond the moderate stage, destructive scoring of the tooth profile occurs. Since the amount of scoring is proportional to the sliding velocity, those areas farthest removed from the pitch line score to the greatest degree. The removal of material from the extremities of the tooth profile leave an area in the vicinity of the pitch line which is full relative to the remainder of the tooth profile (often called a proud pitch line) and the concentration of load at this point can cause pitch line pitting or spalling. The long term consequences of destructive scoring are metal particle generation, destruction of the profile and ultimately tooth breakage.

A.1.2.4 Localised scoring

Conditions which produce non-uniform loading on the tooth surface, such as misalignment and local tooth profile errors, can cause localised scoring. Gears with minimal amounts of localised scoring may continue to operate without further damage if the scoring removes the cause of the non-uniform loading (such as a high spot on a
profile) and the remaining contact surface is capable of supporting the full load. In some circumstances, initial localised scoring can be indicative of underlying problems, such as misalignment, which could lead to failures of a more catastrophic nature if left uncorrected; in this case, localised scoring can be of diagnostic benefit.

**A.1.3 Interference**

Physical interference of one tooth with another generally causes progressive damage; in some circumstances the damage progression may cease after the interference is alleviated. Interference can be caused by a number of conditions, such as operating on tight centres, insufficient involute, thermal expansion, misalignment, insufficient or incorrect profile modification, etc. Tip and root interference is particularly detrimental as it can cause stress concentration near the tooth fillet which can lead to tooth fracture. Interference is normally indicative of poor design, manufacture and/or assembly.

**A.1.4 Surface fatigue**

Surface fatigue is produced by the repeated application and removal of load on the tooth surface which leads to failure when the fatigue capacity of the material is exceeded. The failure modes associated with surface fatigue are pitting and spalling. The fatigue life of a gear is dependant upon the load and the number of load cycles the material is subjected to; a gear with a short designed life can be subjected to much higher surface loads than a similar gear designed for a long life.

Most surface fatigue failures originate at varying depths below the tooth surface, but they are termed ‘surface fatigue’ failures because the surface of the tooth is damaged by the progression of the failure.

**A.1.4.1 Initial pitting**

Localised load concentrations can cause very small pits on the tooth surface either uniformly across the face at the pitch line or locally at one end of the tooth. This ‘initial’
pitting will often only progress until the localised overload condition is relieved, at which time the edges of the pit deform plastically and appear to smooth out; a phenomenon referred to as ‘healing over’ or ‘corrective pitting’.

Initial pitting in itself is not normally viewed as a failure, however if it does not cease it will progress to destructive pitting. Unfortunately, there is no sure way of determining if initial pitting will cease or progress.

**A.1.4.2 Destructive pitting**

When initial pitting does not cease and heal over, it will progress to destructive pitting which will eventually destroy the tooth profile. Destructive pitting occurs when the basic fatigue load capacity of the material has been exceeded, due to application of too much torque or poor load distribution along the tooth or between several pairs of teeth. Carburised, nitrided or induction-hardened gears are often used in surface fatigue critical applications, however these processes tend to distort the gear and grinding or lapping is frequently required to correct the distortion. If full contact is not obtained due to distortion, the effective load capacity can be severely reduced.

Generally, destructive pitting will progress over a long period of time and generate a substantial amount of debris before it progresses to a more catastrophic failure, such as tooth fracture due to stress concentration.

**A.1.4.3 Spalling**

Destructive pitting is often called spalling and vice versa, due to similar appearances in later stages of damage, however they are caused by different failure mechanisms. Spalling is produced by a combination of high surface stresses and relatively high sliding velocities which causes a change in the origin of the surface fatigue failure. In the initial stages of spalling, cracks appear in the tooth surface and spread from the failure origin in a fan like manner in the direction of the sliding. Eventually a piece of the material is removed from the surface, giving the appearance of a great deal of destructive pitting in which the pits have run together forming a ‘spalled’ area.
Spalling usually occurs only on case-hardened gears and because it is related both to the surface stress and sliding, it may be possible to prevent its occurrence by reducing the amount of sliding or the coefficient of friction by shifting the tooth profile, improving the surface finish or changing the oil type.

A.1.4.4 Case crushing

Case crushing, which only occurs in case-hardened gears, is caused by cracks deep below the tooth surface in or near the relatively soft core of the tooth. These cracks are due to the tooth hardness (and hence the shear strength) dropping off faster than the shear stress because the case is too thin. The cracks propagate either within the core or at the case/core junction, with little or no outward evidence of distress, until a large portion of the case is undermined, collapses and breaks away. The visual surface damage is often confused with destructive pitting or spalling, however case crushing usually occurs suddenly on one or two teeth, whereas pitting or spalling progress gradually and are usually evident on many teeth.

A.1.5 Plastic flow

Plastic flow is a yielding and permanent deformation of the tooth surface which can occur under conditions of very heavy loads, usually combined with relatively low rotational speeds.

A.1.5.1 Cold flow

Cold flow is most often observed on medium-hard gears operating with adequate lubrication but with high load and high sliding velocity. The high load can cause the tooth surface to yield with the high frictional component of the load causing the material to cold flow. If allowed to progress, the tooth profile will be destroyed and spalling or tooth fracture is likely to result.

Rippling and ridging are types of cold flow usually observed only on very low speed gears operating with a poor lubricant film. Rippling occurs when the combined effects of
load and sliding cause ripples to form perpendicular to the direction of sliding. In ridging, the plastic flow has a pattern of peaks and valleys forming ‘ridges’ which run parallel to the general direction of sliding. Ridging is most often found on low-speed worm and hypoid gears which have a combination of low entraining and high sliding velocities.

A.1.5.2 Hot flow

Hot flow occurs when the temperature of the gear material increases (and hence the hardness of the gear decreases) sufficiently to allow the material to flow plastically under the applied load. The overheating of the gear generally results in discolouring of the gear material (often a blue or straw colour) which distinguishes hot flow from cold flow. Hot flow is most often the result of a lubrication problem such as decreased flow, interruption of flow or complete depletion of lubricant. Hot flow is a self-sustaining failure in that continued operation, without the aid of lubricant to remove the friction generated heat, generally drives the temperature higher and catastrophic failure occurs in a short period of time.

A.1.6 Fracture

Fracture is probably the most serious failure mode associated with gears. Unlike the failure modes discussed above, which are usually progressive with a long time between initiation and catastrophic failure, fracture can cause almost immediate loss of serviceability or greatly reduced power transmission capability. This can have catastrophic consequences, including human injury or loss of life, in equipment such as helicopters, elevators, cranes, winches, etc., in which the ability to transmit or restrain rotation is critical.

Fracture may occur in a number of ways from a variety of causes.
A.1.6.1 Bending fatigue

The classic bending fatigue failure occurs when the applied load causes a stress level at the root fillet of the tooth, particularly near the tangency point between the fillet and the profile, which exceeds the allowable stress level for a given life expectancy. Typically, the crack initiates near the root of the tooth and grows exponentially; as the crack grows, the bending strength of the tooth is reduced resulting in acceleration of the crack growth rate.

In gears with low axial overlap, such as low contact ratio spur gears, bending fatigue failure usually results in loss of the tooth causing immediate loss of power transmission. In gears with high axial overlap characteristics, such as helical, spiral bevel, worm etc., often only part of the tooth will fracture and the gear set will initially continue to transmit power; successive failures generally occur soon after leading to catastrophic failure of the gear within a short period of time.

Irregularities in the fillet area such as inclusions, tool nicks, steps, grinding and heat treating cracks can act to increase the stress level above the theoretical value, resulting in bending fatigue failure, even if the load is not above that for which the gear was safely designed.

A.1.6.2 Overload

An overload failure occurs when the applied load causes stress in excess of the ultimate strength of the material. Unlike bending fatigue failure where there is some crack progression, the overload failure results in sudden fracture and loss of the tooth. Generally, several adjacent teeth will experience overload failure almost simultaneously resulting in an immediate loss of power transmission. Overload failures are usually caused by sudden, unexpected load application such as mechanical jamming or locking of rotating components.
A.1.6.3 Random fracture

Random fractures, both fatigue and overload, are usually initiated by problems such as a surface or subsurface inclusion, heat treatment cracks, physical damage, excessive pitting and/or spalling, grinding cracks and other stress risers. The random nature of these fractures means that they can appear anywhere on a tooth and generally result in loss of part of the tooth which, although loss of power transmission may not result, will eventually lead to extensive damage.

A.1.6.4 Root/rim/web cracking

In the case of classic bending fatigue fracture described in Section 1.1.6.1, it was assumed that the tooth is mounted on a rigid, massive structure such that tooth bending effects dominate. However, if the rim that supports the gear tooth is thin, the bending of the rim becomes significant and a fatigue crack, rather than progressing through the base of the tooth, may initiate at the tooth root and progress through the rim. This type of damage can lead to total separation of the gear, rather than a single tooth, causing catastrophic damage to the gearbox. A failure of this type occurred with a spiral bevel pinion on a RAN Wessex helicopter in 1983 [54]; gear fragments breached the gearbox casing and destroyed the main rotor control rods, causing a crash in which two lives were lost.

Factors other than rim thickness, such as residual stresses due to faulty surface hardening, inclusions, discontinuities in the root land area, excessive stock removal in the root, grinding and quench cracks, can also lead to crack progression through the rim rather than across the base of the tooth.

A.1.6.5 Resonance induced fracture

Gear resonance can cause relatively rapid and catastrophic failure of a gear, often with separation of large fragments of the gear blank which can cause extensive damage, especially when high rotational speeds are involved. The resonance phenomenon is especially prevalent in highly stressed, lightweight, fatigue-critical gears operating at high
speed, such as gears in helicopter transmissions, gas-turbines drives, and turboprops. Once resonance has been identified as the causal factor, prevention of this type of failure requires a redesign of the gear so that its natural frequencies do not coincide with any operational excitation condition, or the gear can be damped so that, even if a natural frequency is excited, the response of the gear is small enough not to cause failure.

**A.1.7 Process related failures**

In addition to the failures discussed above, which occur during the service life of a gear, there are many types of failures which can occur during the manufacture process, but which may not become apparent until after installation.

Tooth surface damage such as cracks due to improper quenching or grinding, nicks and scratches due to improper handling, and tooling marks can normally be detected by careful inspection prior to installation. If placed into service, tooth surface damage may lead to spalling or fracture.

Other process related damage, such as grinding ‘burns’ (which soften part of the tooth surfaces) and case/core separation, are not easily detectable prior to installation. Grinding burns can lead to early fatigue spall initiation. Case/core separation normally only occurs at sharp corners and tooth tips and, once the small part of the case involved separates, further progression does not usually occur.

**A.2 ROLLING ELEMENT BEARING FAILURE MODES**

The types of material failures which occur in rolling element bearings are similar to those detailed for gears above, however the mechanisms leading to those failures can be somewhat different. Howard [38] provides a summary of bearing failure mechanisms, categorised under the headings of fatigue, wear, plastic deformation, corrosion, brinelling, poor lubrication, faulty installation and incorrect design.

The bearing failure mechanisms here will be summarised under headings similar to those used for gear failure modes in the previous section.
A.2.1 Wear

Unlike gears, there is little or no repetitive sliding action between the mating surfaces in rolling element bearings, and the main causes of systematic wear is not surface-to-surface contact but abrasive and/or corrosive wear which occur in similar fashion to those described in Sections 1.1.1.4 and 1.1.1.5 respectively.

A.2.2 Scoring

Conditions similar to those causing scoring in gears can arise in bearings under conditions of inadequate lubrication. Insufficient film thickness can result in metal-to-metal contact between the rolling elements and the raceways which, in combination with intermittent sliding of the rolling elements, generates heat through increased friction. The combination of sliding, load and heat can lead to the welding and tearing apart of the contact surfaces, resulting in scoring of the contact surface which will cause noisy operation and reduce the life of the bearing.

A.2.3 Surface fatigue

The mechanism causing surface fatigue in bearings is similar to that for gears (Section 1.1.4) in that it is caused by repeated application and removal of load. As there is normally very little sliding at the contact surface, it would be expected that surface fatigue in bearings would normally lead to pitting however, Howard [38] uses the terms pitting, spalling and flaking to describe surface fatigue failures. Components which rotate in relation to the load zone, such as rotating races and rolling elements, will normally experience fairly even loading resulting in a fairly uniform distribution of subsurface fatigue cracks which can ultimately lead to part of the surface lifting away in flakes (flaking). Static races, which do not move in relation to the load zone, will have non-uniform surface fatigue which can ultimately lead to destructive pitting similar to that for pitting at the pitch line for gear teeth (Section 1.1.4.2).
A.2.4 Brinelling

Brinelling is a phenomenon which can occur on bearings, but not on gears, and appears as one or more sets of regularly spaced indentations (with spacing equal to the spacing of the rolling elements) round the circumference of a race. It can be the result of:

a) static overloading causing plastic deformation of the race at the race/element contacts,

b) vibration or shock loading of a stationary bearing, or

c) electric arcing between the rolling elements and race of a stationary bearing.

More than one set of indentations can result when a process causing brinelling is repeated with the bearing in a slightly different position. Brinelling results in noisy operation as the indentations cause impacts as each of the rolling elements pass over them. In itself, brinelling is not necessarily seen as a failure however, because of the increased dynamic loads caused when the rolling elements strike the indentations, it often leads to a reduction in fatigue life of the bearing.

A.3 SHAFT FAILURE MODES

Unlike gears and bearings, shafts normally experience no surface-to-surface contact and therefore are not subject to wear, surface fatigue or scoring. Unbalance, misalignment and/or bent shafts are not in themselves normally considered to be failures however, if they continue uncorrected, they can cause distress to other rotating components, and fatigue cracking of the shaft itself, due to excessive, non-uniform dynamic loads. The principle failure mode which occurs in shafts is cracking.

A.3.1 Fatigue cracking

Cyclic loading of the shaft as it rotates can lead to the formation of fatigue cracks in a similar fashion to bending fatigue in gear teeth. Unbalance, misalignment and bent shafts can accelerate the onset of fatigue cracking due to increased dynamic loads. The
location of the crack and the direction of crack growth depends on the stress distribution within the shaft which is governed by the geometry of the shaft, its attached bearings and gears and the static and dynamic loads. Like bending fatigue cracks in gear teeth, shaft cracks would normally be expected to grow in an exponential fashion as the increasing crack results in increased flexibility. However, this may not always be the case as the stress distribution may change during the progression of the crack, causing a change in crack growth direction.

A.3.2 Overload

Sudden, unexpected application of loads in excess of the ultimate strength of the material can cause rupture of a shaft and catastrophic failure in a similar fashion to overload fracture of gear teeth.
Appendix B

BASIC SIGNAL THEORY

B.1 TIME SIGNALS

Time varying signals appear in almost every branch of physics and the fundamental physical laws governing the behaviour of these signals are similar. For instance, the equations of motion in a mechanical mass-spring system and electromagnetic oscillations in an LC circuit are analogous.

B.1.1 Energy

One of the most important notions in signal analysis is how much ‘energy’ is required to produce the signal, or more specifically how much energy it takes to produce the signal at time \( t \). For any time signal \( s(t) \), the energy density is given by the square of the amplitude of the signal (Cohen [24])

\[
E_t(t) = |s(t)|^2 = \text{energy density per unit time at time } t. \tag{B.1}
\]

This is a fundamental relationship which can be derived from Maxwell’s equations for electromagnetism, Newton’s laws of motion, etc.

The total energy in the system is (Cohen [24], Bendat [3])

\[
E = \int |s(t)|^2 \, dt = \text{total energy} \tag{B.2}
\]

or, where the signal is periodic over time \( T \) (Randall [67]),

\[
E_T = \frac{1}{T} \int_{-T/2}^{T/2} |s(t)|^2 \, dt. \tag{B.3}
\]
(Note: unless otherwise stated all integrals and summations will be over $-\infty$ to $\infty$).

**B.1.2 Simple harmonic motion**

Many signals in nature exhibit oscillatory (harmonic) motion. These can be observed in mechanical systems (such as the vibration of a guitar string, the motion of a mass attached to a spring, etc.) and electromagnetic fields (such as radio waves, microwaves, visible light etc). The same basic mathematical equations are used to described mechanical and electromagnetic oscillations. For example, the equation of motion for an undamped mass-spring pair is derived from Newton’s second law, $F = ma = m(d^2x/dt^2)$ and Hooke’s law, $F = -kx$ (Halliday and Resnick [37]) giving

$$\frac{d^2x}{dt^2} + \frac{k}{m}x = 0 \tag{B.4}$$

and, from Maxwell’s equations, the differential equation describing the oscillations of a resistanceless LC circuit is

$$\frac{d^2q}{dt^2} + \frac{1}{LC}q = 0. \tag{B.5}$$

Mathematically, equations (B.4) and (B.5) are identical. The solution to the equations require that the displacement ($x$) for the mass-spring system and the charge ($q$) for the LC circuit be functions of time whose second derivatives are the negative of the function itself, except for a constant factor ($k/m$ and $1/LC$ respectively). The sine or cosine functions (or combinations of both) have this property, and we can write as a general solution for a simple harmonic oscillator

$$x(t) = a \cos(2\pi ft + \phi_0), \tag{B.6}$$

where $f$ is the frequency of the oscillation and $\phi_0$ is a phase constant defining the phase of the function at time $t=0$. Putting the second derivative of the function,
\[ \frac{d^2(x(t))}{dt^2} = -4\pi^2 f^2 a \cos(2\pi ft + \phi_0) \]  \hspace{1cm} (B.7)

into equations (B.4) and (B.5), it is easy to see that frequency of the oscillation \( f \) is defined for the mass-spring system by \( f^2 = k/(4\pi^2 m) \) and for the LC circuit by \( f^2 = 1/(4\pi^2 LC) \).

In a system in which no nonconservative forces (such as friction in a mechanical system or resistance in an electromagnetic system) are acting, the total energy in the system remains constant. For the simple harmonic oscillator \( x(t) \) in equation (B.6) the potential energy \( U \) at any instant is given by (Halliday and Resnick [37])

\[ U = \frac{1}{2} k x^2 = \frac{1}{2} k a^2 \cos^2(2\pi ft + \phi_0) \]  \hspace{1cm} (B.8)

and the kinetic energy \( K \) is

\[ K = \frac{1}{2} mv^2 = \frac{1}{2} m \left( \frac{dx}{dt} \right)^2 = \frac{1}{2} m 4\pi^2 f^2 a^2 \sin^2(2\pi ft + \phi_0) \]

\[ = \frac{1}{2} k a^2 \sin^2(2\pi ft + \phi_0). \]  \hspace{1cm} (B.9)

The total energy is the sum of the potential energy and the kinetic energy

\[ E = U + K = \frac{1}{2} k a^2 \cos^2(2\pi ft + \phi_0) + \frac{1}{2} k a^2 \sin^2(2\pi ft + \phi_0) = \frac{1}{2} k a^2. \]  \hspace{1cm} (B.10)

For convenience, the displacement vector \( x(t) \) can be combined with its quadrature signal using Euler’s formula to form a complex vector \( s(t) \),

\[ s(t) = a \cos(2\pi ft + \phi_0) + ja \sin(2\pi ft + \phi_0) \]

\[ = a e^{j(2\pi ft + \phi_0)} \]  \hspace{1cm} (B.11)

for which the energy relations in equations (B.1) to (B.3) give values which are proportional to the actual energy in the system.

282
B.1.3 General oscillations

The signal describing simple harmonic motion (B.11) can be easily extended to describe more complex signal behaviour.

B.1.3.1 Nonconservative forces

Nonconservative forces such as friction and resistance change the instantaneous amplitude of the signal (and hence the instantaneous energy), and can be described by the introduction of a time varying amplitude

\[ s(t) = a(t) e^{j(2\pi ft + \phi_0)}. \]  

(B.12)

B.1.3.2 Forced oscillations

When an oscillating external force is applied to the system, the response will be an oscillation at the frequency of the external force, not at the natural frequency of the system. However, the system response will be of the same form as equation (B.11) if the external force is of constant amplitude and frequency. If the external force has a time varying amplitude and constant frequency the response will be of the same form of equation (B.12). When the external force is varying in frequency but has constant amplitude, the response will be of the form (Van der Pol [80])

\[ s(t) = a e^{j \left( \phi_0 + 2\pi \int_0^t f_i(\tau) d\tau \right)} = a e^{j\varphi(t)} \]  

(B.13)

where \( f_i(t) \) is the instantaneous frequency of the external force at time \( t \). This leads to the common definition of the instantaneous frequency of a time signal [80]

\[ f_i(t) = \frac{1}{2\pi} \frac{d}{dt} \left( \varphi(t) \right) \]  

(B.14)

If the external force is varying in frequency and amplitude, the response will be of the form
\[ s(t) = a(t) e^{i \left( \phi_0 + 2\pi \int_0^t f_i(\tau) d\tau \right)}. \] (B.15)

### B.1.3.3 Multicomponent signals

In many instances, two or more waves can traverse the same space independently from one another, producing a signal which is simply the sum of the individual waves,

\[ s(t) = \sum_c s_c(t). \] (B.16)

The *superposition principle* expressed in equation (B.16) holds when the ordinary linear laws of mechanical and electromagnetic action apply. However, if the wave disturbances become very large the governing equations become non-linear and the superposition principle no longer applies. For instance, beyond the elastic limit of deformable media Hooke’s law no longer holds and the linear relation \( F = -kx \) can no longer be used.

### B.2 FREQUENCY DOMAIN REPRESENTATION

Any periodic signal can be represented as the sum of simple harmonic motions (*Fourier series*)

\[ s_T(t) = \sum_{n=0}^{\infty} A_n e^{i(2\pi f_0 t + \phi_n)}, \] (B.17)

where the subscript \( T \) is used here to indicate that the signal is periodic in time \( T \). The fundamental frequency, \( f_0 \), of the *Fourier series* is the reciprocal of \( T \) \( (f_0 = 1/T) \). Equation (B.17) can be extended to the more general case for non-periodic signals by letting the period \( T \to \infty \), resulting in the *Fourier integral*:

---

1 Named after the French mathematician Jean Baptiste Joseph Fourier (1768-1830) whose work on heat transfer lead to the development of this principle.
\[ s(t) = \int A(f) e^{j(2\pi f t + \phi(f))} df \]

\[ = \int S(f) e^{j2\pi f t} df, \quad \text{where } S(f) = A(f) e^{j\phi(f)}. \]  

(B.18)

Inverting equation (B.18) gives

\[ S(f) = \int s(t) e^{-j2\pi f t} dt \]  

(B.19)

which is commonly referred to as the Fourier transform of \( s(t) \) and equation (B.18) is the inverse Fourier transform.

It is clear from equation (B.18) that at frequency \( f \), \( S(f) \) is equivalent to the amplitude of a simple harmonic oscillator (B.11) rotated (in the real-imaginary plane) by a phase constant \( \phi(f) \). Therefore, the energy contribution at frequency \( f \) can be defined as the total energy of a simple harmonic oscillator of amplitude \( |S(f)| \) which, from equation (B.2), is simply the square of the absolute value of \( S(f) \)

\[ E_f(f) = |S(f)|^2 = \text{energy density per unit frequency at frequency } f \]  

(B.20)

and the total energy in the system is

\[ E = \int |S(f)|^2 df = \int |s(t)|^2 dt = \text{total energy} \]  

(B.21)

**B.2.1 Physical meaning of the frequency domain representation**

It is clear that if \( s(t) \) is the superposition of a number of simple harmonic oscillators, then the frequency domain representation \( S(f) \) given by the Fourier transform of \( s(t) \) will provide an exact decomposition of the signal into its individual components. That is, for each oscillator in the system there will be an associated delta function in \( S(f) \) centred at the frequency of the oscillation and with amplitude and phase equivalent to the (constant) amplitude and initial phase (at time \( t=0 \)) of the oscillator respectively.
However, if the signal is not composed entirely of simple harmonic oscillators, its Fourier transform will not be a concise description of the internal structure of the signal and interpretation of the relative values of a number of frequency ‘components’ needs to be made to deduce the underlying signal behaviour.

From the linear property of the Fourier transform (Bendat [3]),

\[
\mathcal{F} \left[ \sum_c s_c(t) \right] = \sum_c \mathcal{F}[s_c(t)] = \sum_c S_c(f) \tag{B.22}
\]

the superposition of signals in the time domain (B.16) will result in a superposition of signals in the frequency domain and, for the moment, we can treat individual signal components separately.

**B.2.1.1 Amplitude modulated signals**

Where the amplitude of the signal is varying in time but the frequency is constant, as in equation (B.12), the product property of the Fourier transform (Bendat [3]) gives

\[
S_c(f) = \mathcal{F}[a_c(t)e^{j(2\pi f_c t + \phi_c)}] = \mathcal{F}[a_c(t)] \ast \mathcal{F}[e^{j2\pi f_c t + \phi_c}] \\
= \int A_c(f - \gamma)e^{j\phi_c} \delta(\gamma - f_c) d\gamma \tag{B.23}
\]

\[
= A_c(f - f_c)e^{j\phi_c}
\]

which is simply the Fourier transform of the amplitude signal \(a_c(t)\) shifted in frequency by the frequency \(f_c\) and shifted in phase by the initial phase \(\phi_c\). The amplitude signal is real valued and, as such, the negative frequency components of its Fourier transform are the complex conjugate of the corresponding positive frequency components (i.e., they have the same amplitude and negative phase). Therefore, for amplitude modulated signals we would expected to see symmetry about the ‘centre’ frequency (the frequency of oscillation \(f_c\):
\[ S_c(f_c - f) = A_c(-f)e^{j\phi_c} = A^*_c(f)e^{j\phi_c} = S^*_c(f_c + f)e^{j2\phi_c}. \tag{B.24} \]

If the frequency content of the amplitude signal is concentrated near zero (e.g., low frequency sinusoidal modulation or slowly decaying exponential damping) the amplitude modulation can usually be identified in the amplitude or power spectrum by a symmetrical grouping of ‘sidebands’ about the frequency of oscillation.

Figure B.1 shows an 88 order sine wave (with mean amplitude of 1 g) which has a sinusoidal 4 per revolution amplitude modulation of ±0.5. The amplitude modulation is clearly seen in the angle domain (a) and the frequency domain (b) representations. In the frequency domain, the modulation is identified by the sidebands at ±4 orders about the centre frequency (88 orders). Both sidebands have the same amplitude (0.1768 = 0.25/sqrt(2)).

(a) angle domain representation  
(b) frequency domain representation

*Figure B.1 Sinusoidal (4 per rev) amplitude modulation*

Where the signal has a change in amplitude over a short period of time, it is often difficult to identify the modulation due to the spread of energy over a wide frequency range.

Figure B.2 shows an 88 order sine wave with a nominal amplitude of 1g which is amplitude modulated over a short time period. The maximum value of the modulation is 0.5 (i.e., the signal has a maximum amplitude of 1.5g). The amplitude modulation can be clearly seen in the angle domain representation (a) but, because of the wide frequency spread, it is not easily identifiable in the frequency domain representation (b).
B.2.1.2 Frequency modulated signals

Randall [65] addressed the case of a simple sinusoidal frequency modulation expressed as

\[ s(t) = a e^{j(2\pi f_0 t + \beta \sin(2\pi f_1 t))}, \]  

(B.25)

where \( \beta = \Delta f f_1 \) is the ‘modulation index’ defining the maximum deviation in the instantaneous frequency of the signal. The signal defined in equation (B.25) is equivalent to that of equation (B.13) with phase constant \( \phi_0 = 0 \) and instantaneous frequency (B.14) of

\[ f_i(t) = \frac{1}{2\pi} \frac{d(2\pi f_0 t + \beta \sin(2\pi f_1 t))}{dt} \quad \text{(B.26)} \]

\[ = f_0 + f_0 \beta \cos(2\pi f_1 t) = f_0 + \Delta f \cos(2\pi f_1 t). \]

Randall [65] showed that the signal expressed in (B.25) could be decomposed into the exponential series,

\[ s(t) = a \left( J_0(\beta) e^{j2\pi f_0 t} + \sum_{n=1}^{\infty} J_n(\beta) \left( e^{j2\pi(f_0 + nf_1)t} + (-1)^n e^{j2\pi(f_0 - nf_1)t} \right) \right) \quad \text{(B.27)} \]
where \( J_0(\beta) \) and \( J_n(\beta) \) are the relative amplitudes of the carrier frequency component and the \( n^{th} \) order sidebands respectively. These functions are dependant on the value of the modulation index \( \beta \). The Fourier transform of (B.27) is

\[
S(f) = a \left\{ J_0(\beta)\delta(f - f_0) + \sum_{n=1}^{\infty} J_n(\beta)\left(\delta(f - f_0 - nf_1) + e^{jn\pi} \delta(f - f_0 + nf_1)\right) \right\}.
\] (B.28)

It can be seen from the above that, even for the case of simple sinusoidal frequency modulation, the resultant spectra will be quite complex.

For the general case, we can view the instantaneous frequency as the series

\[
f_i(t) = f_0 + \sum_{m=1}^{M} f_m \beta_m \cos(2\pi f_m t + \phi_m)
\] (B.29)

which, when substituted into equation (B.13), gives the time signal

\[
s(t) = a e^{jf(\phi_0 + 2\pi f_0 t + \sum_{m=1}^{M} \beta_m \sin(2\pi f_m t + \phi_m))}
\] (B.30)

\[
= ae^{f(\phi_0 + 2\pi f_0 t)} \prod_{m=1}^{M} e^{\beta_m \sin(2\pi f_m t + \phi_m)}.
\]

Using the expansion

\[
s_m(t) = e^{j\beta_m \sin(2\pi f_m t + \phi_m)}
\]

\[
= J_0(\beta_m) + \sum_{n=1}^{\infty} J_n(\beta_m) \left( e^{jn(2\pi f_m t + \phi_m)} + (-1)^n e^{-jn(2\pi f_m t + \phi_m)} \right)
\] (B.31)

and putting
\[ s(t) = ae^{j(\varphi_0 + 2\pi f_0 t)} \prod_{m=1}^{M} s_m(t) \]  

(B.32)

gives the Fourier transform of the signal as the convolution of the Fourier transforms of the individual components,

\[ S(f) = \left( ae^{j\varphi_0} \delta(f - f_0) \right) \ast S_1(f) \ast S_2(f) \ast \cdots \ast S_{M-1}(f) \ast S_M(f), \]  

(B.33)

where the Fourier transforms of the individual components are

\[ S_m(f) = \left\{ J_0(\beta_m) \delta(f) + \sum_{n=1}^{\infty} J_n(\beta_m) \left( e^{jn\varphi_m} \delta(f - nf_m) + e^{jn(\pi - \varphi_m)} \delta(f + nf_m) \right) \right\}. \]  

(B.34)

Note that the Fourier transforms of the individual components (B.34) all have symmetry about \( f = 0 \), therefore the convolution in equation (B.33) will have symmetry about the mean frequency \( f_0 \).

In the case where the signal is modulated in both amplitude and frequency, its Fourier transform can be expressed as the convolution of equation (B.33) with the Fourier transform of the amplitude signal.

In the above, the relative amplitudes of the various components due to frequency modulation, \( J_n(\beta) \), have not been quantified. As indicated, these values are a function of the modulation index \( \beta \). Randall [65] showed that for low values of the modulation index \( \beta < 1 \) the signal energy is concentrated at the carrier frequency and its lower order sidebands. As the value of the modulation index increases, the energy is spread into the higher order sidebands.

Figure B.3 shows the effect of the modulation index in the frequency domain. Both signals shown have a centre frequency of 88 orders, amplitude of 1g and twice per revolution sinusoidal frequency modulations. The spectrum in Figure B.3(a) is for a signal with a modulation index of \( \beta = 0.5 \) (giving a maximum frequency deviation of 1
order). This shows the signal energy concentrated at the carrier frequency (88 orders) and the low order sidebands (-4, -2, +2 and +4). Figure B.3(b) shows the effect of increasing the modulation index to 2.5 (maximum frequency deviation of 5 orders). Here the signal energy has been spread into the higher order sidebands.

![Spectrum TEST3_1.OUT](image1)

![Spectrum TEST3_2.OUT](image2)

(a) $\beta=0.5$ ($\Delta f = 1$)  
(b) $\beta=2.5$ ($\Delta f = 5$)

Figure B.3 Spectra of sinusoidal (2 per rev) frequency modulated signals

The complexity of the spectrum resulting from frequency modulation (as expressed in equations (B.33) and (B.34)) makes this type of signal difficult to interpret in the frequency domain even when only a single (mono-component) signal is analysed.

### B.3 THE ANALYTIC SIGNAL

In general, a recorded time signal will be real valued. As discussed in Section B.1, this represents only one half of the ‘natural’ energy in the signal and, for signal analysis purposes, should be converted into the complex vector equivalent in a similar fashion to equation (B.11). That is, the real part of the complex vector is the recorded time signal and the imaginary part is the quadrature of the recorded signal. Gabor [35] and Ville [81] proposed that the Hilbert transform could be used to form a unique complex analytic signal from a real value signal, with the imaginary part in quadrature to the real part:

$$s(t) = x(t) + j \mathcal{H}[x(t)]$$  \hspace{1cm} (B.35)

where $\mathcal{H}[x(t)]$ is the Hilbert transform of the real valued signal $x(t)$, defined as (Bendat [3])
\[ \mathcal{F}[x(t)] = \int \frac{x(t-\tau)}{\pi \tau} d\tau = x(t)^* \left( \frac{1}{\pi t} \right). \]  \hspace{1cm} (B.36)

The Fourier transform of \( \mathcal{F}[x(t)] \) is [3]

\[
\mathcal{F}[\mathcal{F}[x(t)]] = \mathcal{F}[x(t)] \mathcal{F}\left[ \frac{1}{\pi} \right] = \begin{cases} -jX(f), & f > 0 \\ 0, & f = 0 \\ jX(f), & f < 0 \end{cases} \quad \text{(B.37)}
\]

where \( X(f) \) is the Fourier transform of \( x(t) \). From equation (B.37), the Fourier transform of the analytic signal (B.35) is

\[
S(f) = \mathcal{F}[s(t)] = \mathcal{F}[x(t)] + j\mathcal{F}[\mathcal{F}[x(t)]] = \begin{cases} 2X(f), & f > 0 \\ X(f), & f = 0 \\ 0, & f < 0 \end{cases} \quad \text{(B.38)}
\]

which gives a simple method of calculating the analytic signal by taking the Fourier transform of the real signal \( x(t) \), multiplying the positive frequency components by 2 and setting the negative frequency components to zero, and performing the inverse Fourier transform to give the complex time domain analytic signal \( s(t) \). It should be noted that unless the signal is periodic over the analysis time (as is the case with our synchronously averaged signals), the discrete implementation of the process defined in (B.38) will cause ripples in the computed analytic signal. If this is the case, a Hilbert transform filter [61] should be used to produce the imaginary part of the analytic signal (B.35).
REFERENCES


[9] Boashash, B., “Note on the Use of the Wigner Distribution for Time-


1965.


[34] Forrester, B.D., “Use of the Wigner-Ville Distribution in Helicopter Transmission Fault Detection”, *Proceedings of Australian Symposium on*


[59] Mechefske, C.K. and Mathew, J., “A Comparison of Frequency Domain Trending and Classification Parameters when used to Detect and Diagnose Faults in Low Speed Rolling Element Bearings”, *Machine Condition*
Minns, H. and Stewart, R.M., “An Introduction to Condition Monitoring with Special Reference to Rotating Machinery”, Workshop in On-Condition Maintenance, Section 8, Institute of Sound and Vibration Research, University of Southampton, 1972.


[77] Swansson, N.S., “Application of Vibration Signal Analysis Techniques to Signal Monitoring”, Conference on Friction and Wear in Engineering,
Institution of Engineers, Australia, pp. 262-267, 1980.


LIST OF PUBLICATIONS

The following is a list of publications resulting from the research reported in this thesis:


