

VIBRATION-BASED CONDITION MONITORING

VIBRATION-BASED CONDITION MONITORING

INDUSTRIAL, AUTOMOTIVE AND AEROSPACE APPLICATIONS

Second Edition

Robert Bond Randall

*University of New South Wales
Australia*

WILEY

This edition first published 2021
© 2021 John Wiley & Sons Ltd

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of Robert Bond Randall to be identified as the author of this work has been asserted in accordance with law.

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

This book's use or discussion of MATLAB® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of MATLAB® software.

Library of Congress Cataloging-in-Publication Data

Names: Randall, Robert Bond, author.
Title: Vibration-based condition monitoring : industrial, automotive and aerospace applications / Robert Bond Randall, University of New South Wales.
Description: Second edition. | Hoboken, NJ : Wiley, 2021. | Includes bibliographical references and index.
Identifiers: LCCN 2020048411 (print) | LCCN 2020048412 (ebook) | ISBN 9781119477556 (cloth) | ISBN 9781119477693 (adobe pdf) | ISBN 9781119477655 (epub)
Subjects: LCSH: Vibration--Testing. | Nondestructive testing. | Vibration--Measurement.
Classification: LCC TA355 .R34 2021 (print) | LCC TA355 (ebook) | DDC 621.8/11--dc23
LC record available at <https://lccn.loc.gov/2020048411>
LC ebook record available at <https://lccn.loc.gov/2020048412>

Cover Design: Wiley
Cover Images: © Suriya Desatit/Shutterstock, Borya Galperin/Shutterstock, M101Studio/Shutterstock, PARETO/E+/Getty Images, Alexsey/E+/Getty Images

Set in 10/12pt TimesLTStd by SPi Global, Chennai, India

Dedicated to the memory of Professor Simon Braun, 1933–2020.

Contents

Foreword	xii
About the Author	xv
Preface to The Second Edition	xvi
About the Companion Website	xix
1 Introduction and Background	1
1.1 Introduction	1
1.2 Maintenance Strategies	2
1.3 Condition Monitoring Methods	3
1.3.1 <i>Vibration Analysis</i>	3
1.3.2 <i>Oil Analysis</i>	5
1.3.3 <i>Performance Analysis</i>	6
1.3.4 <i>Thermography</i>	6
1.4 Types and Benefits of Vibration Analysis	6
1.4.1 <i>Benefits Compared with Other Methods</i>	6
1.4.2 <i>Permanent vs Intermittent Monitoring</i>	7
1.5 Vibration Transducers	9
1.5.1 <i>Absolute vs Relative Vibration Measurement</i>	10
1.5.2 <i>Proximity Probes</i>	11
1.5.3 <i>Velocity Transducers</i>	13
1.5.4 <i>Accelerometers</i>	14
1.5.5 <i>Dual Vibration Probes</i>	18
1.5.6 <i>Laser Vibrometers</i>	18
1.6 Torsional Vibration Transducers	19
1.6.1 <i>Shaft Encoders</i>	19
1.6.2 <i>Torsional Laser Vibrometers</i>	20
1.7 Condition Monitoring – The Basic Problem	20
References	23
2 Vibration Signals from Rotating and Reciprocating Machines	25
2.1 Signal Classification	25
2.1.1 <i>Stationary Deterministic Signals</i>	28
2.1.2 <i>Stationary Random Signals</i>	28

2.1.3	<i>Cyclostationary Signals</i>	29
2.1.4	<i>Cyclo-non-stationary Signals</i>	30
2.2	Signals Generated by Rotating Machines	30
2.2.1	<i>Low Shaft Orders and Subharmonics</i>	30
2.2.2	<i>Vibrations from Gears</i>	39
2.2.3	<i>Rolling Element Bearings</i>	46
2.2.4	<i>Bladed Machines</i>	50
2.2.5	<i>Electrical Machines</i>	51
2.3	Signals Generated by Reciprocating Machines	55
2.3.1	<i>Time-Frequency Diagrams</i>	55
2.3.2	<i>Torsional Vibrations</i>	59
	References	60
3	Basic Signal Processing Techniques	63
3.1	Statistical Measures	63
3.1.1	<i>Probability and Probability Density</i>	63
3.1.2	<i>Moments and Cumulants</i>	64
3.2	Fourier Analysis	67
3.2.1	<i>Fourier Series</i>	67
3.2.2	<i>Fourier Integral Transform</i>	68
3.2.3	<i>Sampled Time Signals</i>	70
3.2.4	<i>The Discrete Fourier Transform (DFT)</i>	70
3.2.5	<i>The Fast Fourier Transform (FFT)</i>	72
3.2.6	<i>Convolution and the Convolution Theorem</i>	73
3.2.7	<i>Zoom FFT</i>	83
3.2.8	<i>Practical FFT Analysis</i>	84
3.3	Hilbert Transform and Demodulation	93
3.3.1	<i>Hilbert Transform</i>	93
3.3.2	<i>Demodulation</i>	94
3.4	Digital Filtering	101
3.4.1	<i>Realisation of Digital Filters</i>	102
3.4.2	<i>Comparison of Digital Filtering with FFT Processing</i>	103
3.5	Time/Frequency Analysis	104
3.5.1	<i>The Short Time Fourier Transform (STFT)</i>	104
3.5.2	<i>The Wigner-Ville Distribution</i>	104
3.5.3	<i>Wavelet Analysis</i>	105
3.5.4	<i>Empirical Mode Decomposition</i>	108
3.6	Cyclostationary Analysis and Spectral Correlation	111
3.6.1	<i>Spectral Correlation</i>	112
3.6.2	<i>Spectral Correlation and Envelope Spectrum</i>	114
3.6.3	<i>Wigner-Ville Spectrum</i>	114
3.6.4	<i>Cyclo-non-stationary Analysis</i>	116
	References	119
4	Fault Detection	123
4.1	Introduction	123
4.2	Rotating Machines	123
4.2.1	<i>Vibration Criteria</i>	123

4.2.2	<i>Use of Frequency Spectra</i>	127
4.2.3	<i>CPB Spectrum Comparison</i>	128
4.3	Reciprocating Machines	135
4.3.1	<i>Vibration Criteria for Reciprocating Machines</i>	135
4.3.2	<i>Time/Frequency Diagrams</i>	136
4.3.3	<i>Torsional Vibration</i>	139
	References	146
5	Some Special Signal Processing Techniques	147
5.1	Order Tracking	147
5.1.1	<i>Comparison of Methods</i>	147
5.1.2	<i>Computed Order Tracking (COT)</i>	148
5.1.3	<i>Phase Demodulation Based COT</i>	151
5.1.4	<i>COT Over a Wide Speed Range</i>	156
5.2	Determination of Instantaneous Machine Speed	163
5.2.1	<i>Derivative of Instantaneous Phase</i>	163
5.2.2	<i>Teager Kaiser and Other Energy Operators</i>	168
5.2.3	<i>Comparison of Time and Frequency Domain Approaches</i>	170
5.2.4	<i>Other Methods</i>	174
5.3	Deterministic/Random Signal Separation	177
5.3.1	<i>Time Synchronous Averaging</i>	178
5.3.2	<i>Linear Prediction</i>	180
5.3.3	<i>Adaptive Noise Cancellation</i>	183
5.3.4	<i>Self Adaptive Noise Cancellation</i>	183
5.3.5	<i>Discrete/Random Separation (DRS)</i>	185
5.4	Minimum Entropy Deconvolution	187
5.5	Spectral Kurtosis and the Kurtogram	189
5.5.1	<i>Spectral Kurtosis – Definition and Calculation</i>	190
5.5.2	<i>Use of SK as a Filter</i>	192
5.5.3	<i>The Kurtogram</i>	193
	References	197
6	Cepstrum Analysis Applied to Machine Diagnostics	199
6.1	Cepstrum Terminology and Definitions	199
6.1.1	<i>Brief History of the Cepstrum and Terminology</i>	199
6.1.2	<i>Cepstrum Types and Definitions</i>	202
6.2	Typical Applications of the Real Cepstrum	205
6.2.1	<i>Practical Considerations with the Cepstrum</i>	205
6.2.2	<i>Detecting and Quantifying Harmonic/Sideband Families</i>	208
6.2.3	<i>Separation of Forcing and Transfer Functions</i>	214
6.3	Modifying Time Signals Using the Real Cepstrum	216
6.3.1	<i>Removing Harmonic/Sideband Families</i>	217
6.3.2	<i>Enhancing/Removing Modal Properties</i>	222
6.3.3	<i>Cepstrum Pre-whitening</i>	225
	References	228
7	Diagnostic Techniques for Particular Applications	231
7.1	Harmonic and Sideband Cursors	231

7.1.1	<i>Basic Principles</i>	231
7.1.2	<i>Examples of Cursor Application</i>	232
7.1.3	<i>Combination with Order Tracking</i>	232
7.2	Gear Diagnostics	236
7.2.1	<i>Techniques Based on the TSA</i>	236
7.2.2	<i>Transmission Error as a Diagnostic Tool</i>	238
7.2.3	<i>Cepstrum Analysis for Gear Diagnostics</i>	254
7.2.4	<i>Separation of Spalls and Cracks</i>	263
7.2.5	<i>Diagnostics of Gears with Varying Speed and Load</i>	267
7.3	Rolling Element Bearing Diagnostics	270
7.3.1	<i>Signal Models for Bearing Faults</i>	273
7.3.2	<i>A Semi-Automated Bearing Diagnostic Procedure</i>	277
7.3.3	<i>Alternative Diagnostic Methods for Special Conditions</i>	283
7.3.4	<i>Diagnostics of Bearings with Varying Speed and Load</i>	285
7.4	Reciprocating Machine and IC Engine Diagnostics	295
7.4.1	<i>Time/Frequency Methods</i>	295
7.4.2	<i>Cylinder Pressure Identification</i>	297
7.4.3	<i>Mechanical Fault Identification</i>	303
	References	304
8	Fault Simulation	309
8.1	Background and Justification	309
8.2	Simulation of Faults in Gears	310
8.2.1	<i>Lumped Parameter Models of Parallel Gears</i>	310
8.2.2	<i>Separation of Spalls and Cracks</i>	316
8.2.3	<i>Lumped Parameter Models of Planetary Gears</i>	320
8.2.4	<i>Interaction of Faults with Ring and Sun Gears</i>	322
8.3	Simulation of Faults in Bearings	324
8.3.1	<i>Local Faults in LPM Gearbox Model</i>	325
8.3.2	<i>Extended Faults in LPM Gearbox Model</i>	327
8.3.3	<i>Reduced FE Casing Model Combined with LPM Gear Model</i>	329
8.4	Simulation of Faults in Engines	338
8.4.1	<i>Misfire</i>	338
8.4.2	<i>Piston Slap</i>	347
8.4.3	<i>Bearing Knock</i>	351
	References	354
9	Fault Trending and Prognostics	355
9.1	Introduction	355
9.2	Trend Analysis	355
9.2.1	<i>Trending of Simple Parameters</i>	356
9.2.2	<i>Trending of 'Impulsiveness'</i>	361
9.2.3	<i>Trending of Spall Size in Bearings</i>	364
9.3	Advanced Prognostics	372
9.3.1	<i>Physics-Based Models</i>	372
9.3.2	<i>Data-Driven Models</i>	375
9.3.3	<i>Hybrid Models</i>	377
9.3.4	<i>Simulation-Based Prognostics</i>	380

9.4	Future Developments	387
9.4.1	<i>Advanced Modelling</i>	387
9.4.2	<i>Advances in Data Analytics</i>	389
	References	390
Appendix: Exercises and Tutorial Questions		393
	Introduction	393
A.1	Introduction and Background	393
A.1.1	<i>Exam Questions</i>	393
A.2	Vibration Signals from Machines	394
A.2.1	<i>Exam Questions</i>	394
A.3	Basic Signal Processing	396
A.3.1	<i>Tutorial and Exam Questions</i>	396
A.4	Fault Detection	408
A.4.1	<i>Tutorial and Exam Questions</i>	408
A.4.2	<i>Assignment</i>	413
A.6	Cepstrum Analysis Applied to Machine Diagnostics	414
A.6.1	<i>Tutorial and Exam Questions</i>	414
A.7	Diagnostic Techniques for Particular Applications	415
A.7.1	<i>Tutorial and Exam Questions</i>	415
A.7.2	<i>Assignments</i>	418
A.9	Prognostics	422
A.9.1	<i>Tutorial and Exam questions</i>	422
Index		423

Foreword

About 10 years ago, when the first edition of Prof Randall's *Vibration-based Condition Monitoring* was published, I enthusiastically welcomed it, because a book was finally published that presented a systematic approach to a subject, 'Condition-Based Monitoring', which had grown in complexity over the years, with contributions from many important scholars, but in a non-systematic way.

There were already books that dealt with partial aspects (e.g. *Vibration and Acoustic Measurement Handbook* (1972) by Michael P. Blake and William S. Mitchell, *A Practical Vibration Primer* (1979) by Charles Jackson and *Fundamentals of Noise and Vibration Analysis for Engineers* (1989) by Michael P. Norton), or they dealt with specific fields (including, among the many, Maurice L. Adams Jr.'s *Rotating Machinery Vibration from Analysis to Troubleshooting* in 2001). There was already a reference journal (i.e. *Mechanical Systems and Signal Processing*, founded in 1987 with foresight by the late and recently deceased Simon Braun), but it lacked an organic work, a book that presented a systematic approach and introduced both methods and techniques.

Certainly, in twenty-first century society, one could reflect and debate for a long time whether *a book* still represents 'the instrument' for the transmission of knowledge. Today we have different *media*, but I remain personally convinced that, in the scientific field, *the book* is still a fundamental tool; what has certainly changed is the way it is used: probably many readers of *this book* are not doing it, right now, on paper media, but on a digital medium.

The scientific community and engineers were lucky because this book was written by Prof Randall. I do not think there is any need to present him, because he has a long career of research, of development of signal processing techniques, of case study analysis and of teaching. I was lucky enough to meet him in person 20 years ago, and my esteem for him has always grown, as a monotonic function, not only for its scientific aspects, but also human.

Now, this second edition fills in some inevitable gaps (when writing for the first time a wide-ranging work like this, it is impossible to delve into all the topics or not neglect some, which appeared secondary at that time), but above all introduces and deepens new methods and techniques that have been fully developed in the last ten years, such as tacho-less techniques.

Why is Condition Monitoring so important in engineering and, more generally, in today's world? Prof Randall explains very well the reasons in the introduction of this book: it is a fundamental component for some of the so-called pillars of the technology paradigm of Industry 4.0, at least for IoT ('Internet of Things'), but also for Big Data analytics. Condition Monitoring allows the full implementation of Condition-Based Maintenance (CBM), with remarkable economic advantages, from a single machine to entire plants and industrial facilities, from manufacturing, to services and utilities. Finally, Condition Monitoring is the basis of a predictive – i.e. prognostic – approach to determining the residual useful life (RUL) of a component or a system.

It is certainly ambitious to define what the purpose of science is, and illustrious minds have applied themselves to this: from Greek philosophers to Galileo, from Descartes to Gödel, from Cantor to Popper. If we limit ourselves to the narrow sphere of Engineering and to its purpose, it is not possible to fail to recognise that it must explain '*how*' one does something and '*why*' it is done. In this case, in his book, Prof Randall explains very clearly the '*how*', that is, the most well-established methods and techniques for condition monitoring are analysed in detail and implemented. To do this, he uses one of the most natural signals generated by mechanical systems: vibrations, inextricably linked to the dynamic behaviour of the mechanical systems themselves. Prof Randall also explains in detail '*why*' applying Condition Monitoring is so important.

There is also, however, another interesting '*why*' to analyse, by limiting the scope to the foreword to a book: it concerns Condition Monitoring's rapid development in recent years and its pervasiveness in the modern world. Condition Monitoring is certainly a technological innovation, and as such, its genesis and evolution can be analysed by means of the mechanisms of generating innovation, starting from the more traditional ones, such as the 'Technology-Push', theorised by Joseph A. Schumpeter way back in 1911, and the 'Demand-Pull' most recently introduced by Jacob Schmookler in 1966.

Certainly, the 'epic' and 'primordial' phase of Condition Monitoring (we could call it the 'Proto-Condition Monitoring') was governed by technology-push: without the microprocessors (introduced in the Cold War, not so much for the space race as it is commonly believed, but for the guidance and control of intercontinental missiles), without the invention of miniaturised and reliable sensors (the switch from the strain-gage to the piezoelectric accelerometer happened between the 1940s and the 1950s, with the starting up of manufacturers such as Brüel & Kjær, Columbia Research Laboratories, Endevco, Gulton Manufacturing and Kistler Instruments – some of which are still firmly on the market – or the introduction of the eddy-current proximity probe in 1961 for rotary machines by Bently Nevada), without the personal computers and without the low-cost storage systems, Condition Monitoring would have remained confined to laboratories. Very often, the hardware manufacturer also produced the necessary software and supplied the brainware: for example, minicomputers to collect data and run signal processing methods and rule-based systems for the implementation of condition monitoring. Think, for example, to Hewlett-Packard and Sohre's tables of 1968 or to Bently Nevada and their ADRE systems, and the signal processing methods developed by Donald Bently himself and Agnes Muszynska.

This phase was followed by a 'maturity' phase, governed mainly by the demand-pull, which we could call the 'Meso-Condition Monitoring', during which some large players immediately realised the benefits of Condition Monitoring and implemented it within a CBM approach, as an economic driver for cost reductions. At this stage, the leading roles were big companies and operators of 'big fleets', both in a physical and figurative sense, in various sectors: from the military (think the US Navy) to the transport and aerospace (as in the case of NASA), from the energy (first of all GE and Siemens) to the manufacturing.

However, the two traditional technology-push and demand-pull models do not explain, as it is often the case in technology, why what we might call the 'Neo-Condition Monitoring' is growing so rapidly, in more recent years. The explanation, from a technological innovation point of view, is given by an interactive vision: on the one hand, technological evolution introduces new tools (hardware in the broadest sense: sensors, wireless systems, computers and memory) and new signal processing techniques are proposed, with a frequency if not weekly, at least monthly. On the other hand, as we said, the condition monitoring market, thanks in part to the IoT, has become immense.

In light of these considerations, it is clear how important it is that we have a reference and authoritative text for ‘Vibration-based Condition Monitoring’ and all of us who work in science and technology should be grateful to Bob Randall (I now allow myself to move to a more confidential tone) for writing this second updated and expanded edition, which will certainly become a new milestone.

Paolo Pennacchi
Dept. of Mechanical Engineering
Politecnico di Milano – Milan, Italy
October 2020

About the Author

Bob Randall is a visiting Emeritus Professor in the School of Mechanical and Manufacturing Engineering at the University of New South Wales (UNSW), Sydney, Australia, which he joined as a Senior Lecturer in 1988, and where he is still active in research. Prior to that, he worked for the Danish company Brüel and Kjær for 17 years, after 10 years' experience in the chemical and rubber industries in Australia, Canada and Sweden. His research and publication record while with Brüel and Kjær was treated as PhD equivalent when he joined UNSW. He was promoted to Associate Professor in 1996 and to Professor in 2001, retiring in 2008. He has degrees in Mechanical Engineering and Arts (Mathematics, Swedish) from the Universities of Adelaide and Melbourne, respectively. He is the invited author of chapters on vibration measurement and analysis in a number of handbooks and encyclopedias, including *Shock and Vibration Handbook* (McGraw-Hill) and *Handbook of Noise and Vibration Control* (Wiley). He is a member of the Advisory Board of *Mechanical Systems and Signal Processing* and of the Editorial Board of *Trans. IMechE Part C*. He is the author of more than 350 cited papers in the fields of vibration analysis and machine diagnostics, and has supervised seventeen PhD and three Master's projects to completion in those and related areas. From 1996 to 2012, he was Director of the Defence Science and Technology Organisation (DSTO) Centre of Expertise in Helicopter Structures and Diagnostics at UNSW. He has an interest in languages and has lectured in English, Danish, Swedish, French, German, and Norwegian.

Preface to The Second Edition

In the 10 years since the first edition was published, there have been a very large number of developments in vibration-based machine condition monitoring, not only in the techniques applied, but also in its much wider acceptance in industry, as the most efficient basis for maintenance. This has, for example, changed the relative importance of intermittent manual monitoring and permanent online monitoring. The former approach covered, and still does cover, a much larger number of (simpler) machines, but the latter now represents a much larger investment in monitoring equipment and systems.

Where online monitoring was previously confined largely to expensive and critical machines in say a chemical or power generation plant, which often ran at constant speed and almost constant base load, the huge growth in development of renewable energy sources in the last decade, in particular wind turbines, first onshore, and then offshore, has prompted the development of techniques that can cope with widely dispersed smaller machines, often multiple units in large wind farms, with widely varying speed and load. Such techniques could then also be applied to other machines with difficult access, and variable speed and load, such as automated machines in manufacturing plants, and mobile equipment in mines, etc, the latter often being driverless and remotely controlled. Industry 4.0 and the associated Internet of Things (IoT) recognise the importance of including more transducers in autonomous machines, not only for automatic control, but also for optimised condition-based maintenance (CBM), and the economic benefits of CBM.

The new edition contains nine chapters instead of the original six. Chapters 1, 2, and 4 only have minor updates. Because of the need to process signals with varying speed, considerable advances have been made in order tracking, involving resampling time signals at uniform spacings in rotation angle of a machine. The accuracy of doing this has been greatly improved, allowing multiple transformations back and forth between the time and angle domains. This was found very useful to cope with the fact that shaft speed related signals, such as gearmesh frequencies, are made more periodic in the angle domain, while structural response properties, such as impulse responses, retain constant time scales in the time domain, independent of speed. It is now often possible to extract the speed information from the response signal itself, avoiding the need for a tachometer signal. A related topic is the accurate determination of the machine speed, often from the vibration signals as well. A new Chapter 5, called *Some special signal processing techniques* has thus been added to address this, but it also includes carry-over of some updated topics from the original Chapter 3, *Basic signal processing techniques*. Significant new material has been added to the remaining topics in the updated Chapter 3.

Shortly after the publication of the first edition there was a breakthrough in cepstrum analysis, which meant that time signals could be obtained by editing the cepstrum of continuous signals. This was not previously thought to be possible, since the complex cepstrum, which can be reversed back

to time signals, requires continuous phase spectra, which are a property of transients only. However, there are many applications where the real cepstrum can be edited, to obtain an edited log amplitude spectrum, which can then be combined with the original non-continuous phase spectrum, to give edited time signals with very little error. Just one example of this is where the editing removes families of harmonics from the spectrum (and time signals) by notching in the cepstrum. The phase will be in error at the frequencies of the removed components, but these have been greatly reduced (to the same level as adjacent noise components) and are widely spaced, so the effect of the residual phase errors is usually negligible. This new possibility gives rise to so many new applications, including extraction of separated gear and bearing signals under varying speed conditions, and pre-processing of machine vibration signals as a precursor to operational modal analysis (OMA), that a new Chapter 6, devoted to *Cepstrum analysis applied to machine diagnostics*, has been added. It should be mentioned that OMA is now recognised as being an important part of advanced condition monitoring, to allow the extraction of force signals from measured vibration responses.

The new Chapter 7, *Diagnostic Techniques for particular applications*, is a substantially updated version of the old Chapter 5, *Diagnostic techniques*, even though many of the headings are the same. Quite recently, it has been discovered that the analysis of gear transmission error (TE) is much more powerful as a diagnostic tool than originally thought, and though introduced in the first edition, it was then thought that the required mounting of shaft encoders would restrict it to being a laboratory tool. However, with the rapid progress of Industry 4.0 and the IoT, it is becoming more common to build such transducers into machines at the design phase, also because of their benefits in control, in particular of variable speed machines. This topic has thus been considerably expanded, and important new applications developed. There have also been many other improvements in the diagnostics of gears and bearings, not least under variable speed and load conditions. While still being a very important tool, the main approach to bearing diagnostics in the first edition, based on spectral kurtosis and the kurtogram, has been shown to sometimes fail, for example when impulsive signals from other sources, such as EMI (electromagnetic interference) dominate the kurtosis, so a number of alternative approaches can now be chosen, which have benefits in many situations. Diagnostics of internal combustion (IC) engines has also been greatly improved, for example to include mechanical faults such as piston slap and bearing knock.

Another area which has become much more practicable is the ability to make realistic detailed simulation models of machine components, and even complete complex machines, using CAD, FEM, and multi-body dynamics. This has become a standard part of the machine development process, greatly reducing the number of intermediate prototypes. It has led to the adaptation of such models to individual machines (known as ‘digital twins’), for example to compensate for tolerance variations, and even changes in performance with time, in particular in the now much more common situation where manufacturers are responsible for the whole-life performance and maintenance of machines. For the latter application, it would be advantageous to simulate faults in components, using substructuring techniques to incorporate them in the overall models, in place of the original healthy components. A new Chapter 8 has thus been added, entitled *Fault simulation*, giving typical approaches for the cases of gears, bearings and IC engines, largely developed since the publication of the first edition. It is particularly with dynamic machine models that OMA is required to update the simulation models, not only because they vary more with operating conditions than static structures, but also because the forcing functions are much more complex in general. Extraction of these forces, using the modal models, is also much more important in machine condition monitoring than in structures, where the condition is indicated primarily by modal properties.

Simulation is also now a very important part of prognostics, because of the impossibility of experiencing fully documented failures in sufficient quantities to use purely data driven methods, so the new Chapter 9, replacing the old Chapter 6 on *Fault trending and prognostics*, has been considerably

updated, even though the ideal solution to this problem has still not been found. The sometimes blind faith in ‘Big data’ as a future solution is misplaced, although such techniques will play a large part in compensating for the effects of wide variations in operating conditions, for machines in normal, or near normal, condition, for which big data does exist. This will greatly increase the reliability of detection of departures from normal condition, and at least aid prediction of developments into the near future. The first advances will most likely be made in fleets of similar machines, e.g. wind turbines and helicopter gearboxes, where increasing numbers of documented cases will be incorporated into predictive systems, based initially on physics-based fault development models, including simulation of various degrees of sophistication, right up to digital twins. Watch this space.

Because of the large amount of new material, a large number of additional acknowledgements are required, not only for new work, but also because some earlier work has now increased in importance for machine condition monitoring. In that category must now be added Dr Yujin Gao, whose pioneering work on cepstral methods of modal analysis, forms the basis of much of the added material of Chapter 6, and Dr Yuejin Li, whose work, still not widely published, on the weighting of responses at a fixed point on the casing to faults in rotating planet bearings led to the first application of the log envelope of a bearing signal, and will almost certainly assist in the diagnostics of planet gears and their bearings in the future.

I would like to acknowledge the contributions of all the many co-authors of the large number of new papers by our group at UNSW, from 2011 to the present, but in particular would like to thank my colleagues in that period including Dr Wade Smith, Dr Pietro Borghesani, and Prof. Zhongxiao Peng, as well as my former PhD students Dr Jian Chen, Dr Lav Deshpande, and Dr Michael Coats, whose doctoral works make substantial contributions to the new edition. I am very sad to report that Dr Chen tragically passed away in July 2019, at a very young age, after a short illness. It is also with great sadness that I have to report the passing in June 2015, of Dr Peter McFadden, many of whose contributions are reported in both editions.

One of the greatest losses to the machine diagnostic community was the passing, in March 2020, of Professor Simon Braun, founder of the journal *Mechanical Systems and Signal Processing*. I would like to reiterate my acknowledgement, from the first edition, of his continued support and mentorship over many years. It is perhaps worth mentioning that over 50% of the references in the new material of the new edition were published in MSSP.

A special thanks is again due to Professor Jérôme Antoni, who moved to INSA Lyon, France, several years ago, and who has continued to be a steady inspiration and collaborator in many of the new developments.

To the list of international academic institutions, with whom collaboration has enhanced much of the new material, must be added Vrije Universiteit Brussel (VUB), Belgium, on OMA and wind turbine monitoring, and Professor Paolo Pennacchi’s group at Politecnico di Milano, on bearing diagnostics, initiated by the visit to UNSW in 2011 of Pietro Borghesani, during his PhD studies. SpectraQuest Inc., Richmond USA, supported work on bearing diagnostics and system modelling, while the former Leuven Measurement Systems (LMS, now Siemens), Belgium, supported the groundbreaking PhD project of Dr Jian Chen.

Once again, I would like to acknowledge the support of my wife, Helen, who encouraged me to keep writing at all times, even when it threw a greater load on her.

About the Companion Website

This book is accompanied by a companion website.

www.wiley.com/go/randall/vibration



This website includes:

Exercises and tutorial questions

1

Introduction and Background

1.1 Introduction

Machine Condition Monitoring is an important part of Condition-based Maintenance (CBM), which is becoming recognised as the most efficient strategy for carrying out maintenance in a wide variety of industries. Machines were originally ‘run to break’, which ensured maximum operating time between shutdowns, but meant that breakdowns were occasionally catastrophic, with serious consequences for safety, production loss, and repair cost. The first response was ‘Preventive Maintenance’, where maintenance is carried out at intervals such that there is a very small likelihood of failure between repairs. However, this results in much greater use of spare parts, as well as more maintenance work than necessary.

Even at the time of the first edition of this book, there was a considerable body of evidence that CBM gave economic advantages in most industries. An excellent survey of the development of maintenance strategies was given by Rao in a Keynote paper at a Comadem (Condition Monitoring and Diagnostic Engineering Management) conference in 2009 [1]. Maintenance is often regarded as a Cost Centre in many companies, but Al-Najjar [2–4] has long promoted the idea that CBM can convert maintenance to a Profit Centre. Jardine et al. [5, 6] from the University of Toronto documented a number of cases of savings given by the use of CBM. The case presented in [6], from the Canadian pulp and paper industry, is discussed further in Chapter 9, in connection with their approach to prognostics.

Since the first edition of this book was written, the value of CBM has been even more accepted, and for example a recent report [7] indicates that the global value of the machine condition monitoring market will rise from USD 2.6 to 3.9 billion from 2019 to 2025. This is in part due to the increasing general acceptance of Industry 4.0, or the Fourth Industrial Revolution, which according to Wikipedia ‘is the trend towards automation and data exchange in manufacturing technologies and processes which include cyber-physical systems (CPSs), the internet of things (IoT), industrial internet of things (IIoT), cloud computing, cognitive computing and artificial intelligence’. CBM is expected to be part of the ‘Smart factory’, and this is evidence that it is starting to be more extensively applied in manufacturing plants, for example on complex automated machine tools, whose operations have rapid changes in speed and load. This was just not possible in the early days of condition monitoring, where the monitored machines tended to run for long periods at constant speed and loads, for example in power generation and chemical plants. Some of the newer techniques to

cope with variable speed and loads have been developed in the intervening 10 years and are described in this new edition.

To base maintenance on the perceived condition of operating machines (many of which are required to run continuously for 12 months or more) requires that methods are available to determine their internal condition whilst they are in operation. The two main ways of getting information from the inside to the outside of operating machines are vibration analysis and lubricant analysis, although a few other techniques are also useful.

This chapter includes a description of the background for, and methods used in, condition monitoring, while most of the rest of the book is devoted primarily to the methods based on vibration analysis, which are the most important. This chapter describes the various types of vibration measurement used in condition monitoring, and the transducers used to provide the corresponding vibration signals. It also describes the basic problem in interpretation of vibration signals, in that they are always a compound of forcing function effects (the source) and transfer function effects (the structural transmission path), and how the two effects may be separated.

1.2 Maintenance Strategies

As briefly mentioned above, the available maintenance strategies are broadly:

1. Run-to-break

This is the traditional method where machines were simply run until they broke down. This in principle gives the longest time between shutdowns, but failure when it does occur can be catastrophic, and can result in severe consequential damage, for example of components other than the ones that failed, and also of connected machines. As a result, the time to repair can be greatly increased, including the time required to obtain replacement parts, some of which might be major items and take some time to produce. In such a case, the major cost in many industries would be production loss, this often being much greater than the cost of individual machines. There is still a place for run-to-break maintenance, in industries where there are large numbers of small machines, e.g. sewing machines, where the loss of one machine for a short time is not critical to production, and where failure is unlikely to be catastrophic.

2. (Time-based) Preventive Maintenance

Maintenance is done at regular intervals which are shorter than the expected ‘time-between-failures’. It is common to choose the intervals to be such that no more than 1–2% of machines will experience failure in that time. It does mean that the vast majority could have run longer by a factor of two or three [8]. The advantage of this method is that most maintenance can be planned well in advance, and that catastrophic failure is greatly reduced. The disadvantages, in addition to the fact that a small number of unforeseen failures can still occur, are that too much maintenance is carried out, and an excessive number of replacement components consumed. It has been known to cause reduced morale in maintenance workers (who are aware that most of the time they are replacing perfectly good parts) so that their work suffers and this can give rise to increased ‘infant mortality’ of the machines, by introducing faults which otherwise never would have happened. Time-based preventive maintenance is appropriate where the time to failure can be reasonably accurately predicted, such as where it is based on well-defined ‘lifing’ procedures, which can predict the fatigue life of crucial components on the basis of a given operational regime. Some components do tend to wear or fatigue at a reasonably predictable rate, but with others, such as rolling element bearings, there is a large statistical spread around the mean, leading to estimates such as the one given above, where the mean time to failure is two to three times the minimum [8].

3. Condition-based Maintenance (CBM)

This is also called ‘predictive maintenance’ since the potential breakdown of a machine is predicted through regular condition monitoring, and maintenance is carried out at the optimum time. It has obvious advantages compared with either run-to-break or preventive maintenance, but does require having access to reliable condition monitoring techniques, which are not only able to determine current condition, but also give reasonable predictions of remaining useful life. It has been used with some success for 35–45 years, and for example the abovementioned report [8] by Neale and Woodley predicted back in 1978 that maintenance costs in British industry could be reduced by approximately 65% by appropriate implementation in a number of industries that they identified. However, the range of monitoring techniques was initially quite limited, and not always correctly applied, so it is perhaps only the last 20 years or so that it has become recognised as the best maintenance strategy in most cases. Initially the greatest successes were attained in industries where machines were required to run for long periods of time without shutting down, such as the power generation and (petro-) chemical industries. The machines in such industries typically run at near constant speed, and with stable load, so the technical problems associated with the condition monitoring were considerably reduced. As more powerful diagnostic techniques have become available, it has been possible to extend condition monitoring to other industries in which the machines have widely varying speed and load, and are perhaps even mobile (such as ore trucks in the mining industry). Refs. [9, 10] discuss the potential benefits given by CBM applied to hydroelectric power plants and wind turbines, respectively.

This book aims at explaining a wide range of techniques, based on vibration analysis, for all three phases of machine condition monitoring, namely, fault detection, fault diagnosis, and fault prognosis (prediction of remaining useful life).

1.3 Condition Monitoring Methods

Condition monitoring is based on being able to monitor the current condition and predict the future condition of machines while in operation. Thus, it means that information must be obtained externally about internal effects while the machines are in operation.

The two main techniques for obtaining information about internal condition are:

1. Vibration Analysis

A machine in standard condition has a certain vibration signature. Fault development changes that signature in a way that can be related to the fault. This has given rise to the term ‘Mechanical Signature Analysis’ [11].

2. Lubricant Analysis

The lubricant also carries information from the inside to the outside of operating machines in the form of wear particles, chemical contaminants etc. Its use is mainly confined to circulating oil lubricating systems, although some analysis can be carried out on grease lubricants.

Each of these is discussed in a little more detail along with a couple of other methods, performance analysis and thermography, that have more specialised applications.

1.3.1 *Vibration Analysis*

Even in good condition, machines generate vibrations. Many such vibrations are directly linked to periodic events in the machine’s operation, such as rotating shafts, meshing gear teeth, rotating

electric fields, etc. The frequency with which such events repeat often gives a direct indication of the source, and thus many powerful diagnostic techniques are based on frequency analysis. Some vibrations are due to events that are not completely phase-locked to shaft rotations, such as combustion in IC (internal combustion) engines, but where a fixed number of combustion events occur each engine cycle, even though not completely repeatable. As will be seen, this can even be an advantage, as it allows such phenomena to be separated from perfectly periodic ones. Other vibrations are linked to fluid flow, as in pumps and gas turbines, and these also have particular, quite often unique, characteristics. The term ‘vibration’ can be interpreted in different ways, however, and one of the purposes of this chapter is to clarify the differences between them, and the various transducers used to convert the vibration into electrical signals that can be recorded and analysed.

One immediate difference is between the absolute vibration of a machine housing, and the relative vibration between a shaft and the housing, in particular where the bearing separating the two is a fluid film or journal bearing. Both types of vibration measurement are used extensively in machine condition monitoring, so it is important to understand the different information they provide.

Another type of vibration which carries diagnostic information is torsional vibration, i.e. angular velocity fluctuations of the shafts and components such as gears and rotor discs.

All three types of vibration are discussed in this chapter, and the rest of the book is devoted to analysing the resulting vibration signals, though mainly from accelerometers (acceleration transducers) mounted on the machine casing.

It should perhaps be mentioned that a related technique, based on the measurement of ‘acoustic emission’ (AE), has received some attention and is still being studied. The name derives from high frequency solid-borne rather than airborne acoustic signals from developing cracks and other permanent deformation, bursts of stress-waves being emitted as the crack grows, but not necessarily otherwise. The frequency range for metallic components is typically 100 kHz–1 MHz, this being detected by piezoelectric transducers attached to the surface.

One of the first applications to machine diagnostics was to detection of cracks in rotor components (shafts and blades) in steam turbines, initiated by the Electric Power Research Institute (EPRI) in the USA [12]. Even though they claimed some success in detecting such faults on the external housing of fluid film bearings, the application does not appear to have been developed further. AE monitoring of gear fault development was reported in [13], where it was compared with vibration monitoring. The conclusion was that indications of crack initiation were occasionally detected a day earlier (in a 14 day test) than symptoms in the vibration signals, but the latter persisted because they were due to the presence of actual spalls, while the AE was only present during crack growth. Because of the extremely high sampling rate required for AE, huge amounts of data would have to be collected to capture the rare burst events, unless recordings were based on event triggering. In [14], AE signals are compared with vibration signals (and oil analysis) for gear fault diagnostics and prognostics, but the AE sensors had to be mounted on the rotating components and signals extracted via slip rings.

There is some evidence that AE measurements on rolling element bearings can detect very small incipient faults a little earlier than vibration measurements (e.g. [15]), but it is almost certain that the frequency range of the excitation will fall below 100 kHz as the faults become physically larger, after which the AE transducers could not be relied upon to follow the fault development. In the author’s opinion, other methods which can detect resonant responses in the frequency range 40–100 kHz, are likely to detect bearing faults almost as early as AE transducers, and would still give sufficient advance warning to allow prognostics of the fault development as the frequency range of the excitation reduces with increase in fault size.

One such method is the proprietary ‘PeakVue’ method incorporated in Emerson CSI analysers [16], where the signal used to generate the envelope is down-sampled from the original rate of 102.4 kHz in such a way that envelope information is not lost. If the sample rate is to be reduced by

a factor of 50, for example, to give an envelope spectrum range of about 1 kHz, instead of simply retaining every 50th sample (which might completely miss short high frequency pulses) the absolute peak value in every 50 samples is retained. Because the signal is not lowpass filtered before down-sampling, this of course gives aliasing (see Chapter 3), but only of the high frequency carrier, which does not contain diagnostic information. The important information about the repetition frequency of the response pulses (from a bearing fault, for example), is contained in the envelope signal, which is retained, as explained in [17]. The down-sampling approach in [15] is based on the same principle. The high frequency signals (up to the original lowpass filter frequency of about 40 kHz) containing the bearing fault pulses, are almost certainly dominated by the accelerometer mounting resonance, which as pointed out in [16] varies considerably based on the mounting method. It is therefore unlikely that the PeakVue method will give repeatably scaled measurements, suitable for trending, but it will often give a very early detection and diagnosis of a bearing fault. The same accelerometer as used for the PeakVue method is also used for conventional vibration analysis in the lower frequency range where the accelerometer response is linear and more repeatable.

Another somewhat similar approach, the Shock Pulse Method (SPM), specifically uses the accelerometer resonance to carry information about high frequency fault pulses, such as from bearings, and uses the demodulated signal for a range of diagnostic purposes, including envelope spectrum analysis. In this case the accelerometer mounted resonance is specified as 32 kHz, and is ensured by tight control of the mounting method, including the use of steel rod wave guides to carry the signal from the machine casing to where it can more easily be measured. The conventional SPM techniques use only this resonant response, but from 2015, a new ‘DuoTech’ transducer has been made available which can cover the conventional lower frequency vibration range as well ([18]).

Because of the difficulty of application of AE monitoring to machine condition monitoring there are only limited further discussions in this book, although new developments may change the situation.

1.3.2 Oil Analysis

This can once again be divided into a number of different categories:

1. **Chip detectors.** Filters and magnetic plugs are designed to retain chips and other debris in circulating lubricant systems, and these are analysed for quantity, type, shape, size, etc. Alternatively, suspended particles can be detected in flow past a window.
2. **Spectrographic Oil Analysis Procedures (SOAPs).** Here, the lubricant is sampled at regular intervals and subjected to spectrographic chemical analysis. Detection of trace elements can tell of wear of special materials such as alloying elements in special steels, white metal, or bronze bearings, etc. Another case applies to oil from engine crankcases, where the presence of water leaks can be indicated by a growth in NaCl or other chemicals coming from the cooling water. Oil analysis also includes analysis of wear debris, contaminants and additives, and measurement of viscosity and degradation. Simpler devices measure total iron content.
3. **Ferrography.** This represents the microscopic investigation and analysis of debris retained magnetically (hence the name), but which can contain non-magnetic particles caught up with the magnetic ones. Quantity, shape, and size of the wear particles are all important factors in pointing to the type and location of failure.

Successful use of oil analysis requires that oil sampling, changing, and top-up procedures are all well-defined and documented. It is much more difficult to apply lubricant analysis to grease lubricated machines, but grease sampling kits are now available to make the process more reliable.

Since the first edition of this book was published, more advanced online techniques have been developed ([19, 20]). Ref. [19] describes online measurement of oil viscosity and other parameters such as different particle concentration levels, while Ref. [20] uses online image analysis to obtain particle quantity, size, shape, and colour. This does help to remove some of the problems involved in sending out oil samples to external organisations for analysis, including much more extensive sampling, and greatly reduces the time involved.

1.3.3 Performance Analysis

With certain types of machines, performance analysis (e.g. stage efficiency) is an effective way of determining whether a machine is functioning correctly.

One example is given by reciprocating compressors, where changes in suction pressure can point to filter blockage, valve leakage could cause reductions in volumetric efficiency, etc. Another is in gas turbine engines, where there are many permanently mounted transducers for process parameters such as temperatures, pressures, and flowrates, and it is possible to calculate various efficiencies and compare them with the normal condition, so-called ‘flow path analysis’.

With modern IC engine control systems, e.g. for diesel locos, electronic injection control means that the fuel supply to a particular cylinder can be cut off, and the resulting drop in power compared with the theoretical.

1.3.4 Thermography

Sensitive instruments are now available for remotely measuring even small temperature changes, in particular in comparison with a standard condition. At this point in time, thermography is still used principally in quasi-static situations, such as with electrical switchboards, to detect local hot spots, and to detect faulty refractory linings in containers for hot fluids such as molten metal.

So-called ‘hot box detectors’ have been used to detect faulty bearings in rail vehicles, by measuring the temperature of bearings on trains passing the wayside monitoring point. These are not very efficient, as they must not be separated by more than 50 km or so, because a substantial rise in temperature of a bearing only occurs in the last stages of life, essentially when ‘rolling’ elements are sliding. Monitoring based on vibration and/or acoustic measurements appears to give much more advance warning of impending failure.

1.4 Types and Benefits of Vibration Analysis

1.4.1 Benefits Compared with Other Methods

Vibration analysis is by far the most prevalent method for machine condition monitoring because it has a number of advantages compared with the other methods. It reacts immediately to change, and can therefore be used for permanent as well as intermittent monitoring. With oil analysis for example, several days often elapse between the collection of samples and their analysis, although some online systems do exist. Also in comparison with oil analysis, vibration analysis is more likely to point to the actual faulty component, as many bearings, for example, will contain metals with the same chemical composition, whereas only the faulty one will exhibit increased vibration. There is some development towards the combined use of wear debris analysis and vibration analysis, the first

indicating the type and total amount of wear, and the second the detailed distribution of the wear, but this book concentrates on the vibration analysis part.

Most importantly, many powerful signal processing techniques can be applied to vibration signals to extract even very weak fault indications from noise and other masking signals. Most of this book is concerned with these issues.

1.4.2 Permanent vs Intermittent Monitoring

Critical machines often have permanently mounted vibration transducers, and are continuously monitored so that they can be shut down very rapidly in the case of sudden changes, which might be a precursor to catastrophic failure. Even though automatic shutdown will almost certainly disrupt production, the consequential damage that could occur from catastrophic failure would usually result in much longer shutdowns and more costly damage to the machines themselves. Critical machines are often ‘spared’, so that the reserve machines can be started up immediately to continue production with a minimum of disruption. Most critical high speed turbo-machines in, for example, power generation plants and petro-chemical plants, have built-in proximity probes (Section 1.5.2) which continuously monitor relative shaft vibration, and the associated monitoring systems often have automatic shutdown capability. Where the machines have gears and rolling element bearings, or to detect blade faults, the permanently mounted transducers should also include accelerometers, as explained in Section 1.5.4.

Note that ‘permanent’ or ‘online’ monitoring is not synonymous with ‘real-time’ monitoring, which is rarely required in machine monitoring, as compared, for example, with automatic control. Real-time processing implies causal signal processing, which has severe disadvantages compared with non-causal processing, as typified by the use of the Fast Fourier Transform (FFT), which is perhaps the dominant signal processing tool in machine monitoring. As shown in Chapter 3, use of the FFT involves batch processing of time records selected out of a continuous record, but individually treated as though repeated periodically. This means that the second half of each time record is implicitly treated as negative time (thus non-causal), but gives huge advantages in that it allows for almost ideal filters, with no phase distortion, which can completely exclude adjacent strong frequency components, not possible with causal filters, which have a much more gentle ‘roll-off’, and introduce phase shifts over relatively wide frequency ranges. The delay involved in the non-causal processing is just the processing time of one such time record, and would rarely be more than a second or so, this normally being negligible compared with the time constants associated even with real-time processing. Even if a large turbomachine were automatically tripped (that would be rare because a human would normally have to make that decision) the speed (which is normally reduced over a period of many hours) would not change significantly in the few seconds difference between causal and non-causal processing. Another example where real-time processing is of no advantage is if the machine being monitored were, say, an aircraft engine, what would be the difference between the warning ‘this engine will self-destruct in two seconds’, compared with two milliseconds? Neither would be of much use.

The **advantages** of permanent, or online, monitoring are:

- It reacts very quickly to sudden change, and gives the best potential for protecting critical and expensive equipment.
- It is the best form of protection for sudden faults that cannot be predicted. An example is the sudden unbalance that can occur on fans handling dirty gas, where there is generally a build-up of deposits on the blades over time. This is normally uniformly distributed, but can result in sudden massive unbalance when sections of the deposits are dislodged.

- It is sometimes more economical to have permanently mounted transducers on widely distributed and difficult-to-access machines, such as wind turbines, and automated manufacturing machines, and then the additional cost of transmitting the collected signals back to a centralised monitoring system is economically justified. This approach is now being applied also for mobile equipment, such as mining trucks and machines, many of which are autonomous.

The **disadvantages** of permanent monitoring are:

- The cost of having permanently mounted transducers is very high, so previously could only be justified for the most critical machines in a plant, or where it is difficult for operators to access the machines. However, the cost of transducers such as accelerometers is continuously being reduced, and with the increased development of autonomous machines, more in-built transducers are required anyway, in conformity with the ‘Internet of Things’. Because of the increasing realisation of the benefits of CBM, online monitoring is being extended to more and more machines.
- Where the transducers are proximity probes, they virtually have to be built in to the machine at the design stage, as modification of existing machines would often be prohibitive.
- Since the reaction has to be very quick, permanent monitoring is normally based on relatively simple parameters, such as overall RMS or peak vibration level, and the phase of low harmonics of shaft speed relative to a ‘key phasor’, a once-per-rev pulse at a known rotation angle of the shaft. In general, such simple parameters do not give much advance warning of impending failure; it is likely to be hours or days, as opposed to the weeks or months lead time that can be given by the advanced diagnostic techniques detailed in later chapters in this book.

Of course, if transducers are mounted permanently, it is still possible to analyse the signals in more detail, just not continuously. It gives the advantage that intermittent monitoring can be carried out in parallel with the permanent monitoring, and updated at much more frequent intervals, typically once per day instead of once per week or once per month, to give the best of both worlds. In order to take advantage of the powerful diagnostic techniques, the permanently mounted transducers would have to include accelerometers, for the reasons discussed below in Section 1.5.4.

For the vast majority of machines in many plants, it is not economically justified to have them equipped with permanently attached transducers or permanent monitoring systems. On the other hand, since the major economic benefit from condition monitoring is the potential to predict incipient failure weeks or months in advance, so as to be able to plan maintenance to give the minimum disruption of production and acquire replacement parts etc., it is not always important to do the monitoring continuously. The intervals must just be sufficiently shorter than the minimum required lead times for maintenance and production planning purposes. A procedure for determining the optimum intervals is described in [21]. A very large number of machines can then be monitored intermittently with a single transducer and data logger, and the data downloaded to a monitoring system capable of carrying out detailed analysis.

The **advantages** of intermittent monitoring are:

- Much lower cost of monitoring equipment.
- The potential (through detailed analysis) to get much more advance warning of impending failure, and thus plan maintenance work and production to maximise availability of equipment.
- It is thus applied primarily where the cost of lost production from failure of the machine completely outweighs the cost of the machine itself.

The **disadvantages** of intermittent monitoring are:

- Sudden rapid breakdown may be missed, and in fact where failure is completely unpredictable this technique should not be used. On the other hand, the reliability of detection and diagnostic techniques for predictable faults is increasing all the time, and can now be said to be very good, in that considerable economic benefit is given statistically by correct application of the most up-to-date condition monitoring techniques [2–6].
- The lead time to failure may not be as long as possible if the monitoring intervals are too long for economic reasons. This is in fact an economic question, balancing the benefits of increased lead time against the extra cost of monitoring more frequently [18].

To summarise, **permanent monitoring** is used to shut machines down in response to sudden change, and is thus primarily used on critical and expensive machines to avoid catastrophic failure. It is based on monitoring relatively simple parameters that react quickly to change, and typically uses proximity probes and/or accelerometers. **Intermittent monitoring** is used to give long-term advance warning of developing faults, and is used on much greater numbers of machines and where production loss is the prime economic factor rather than the cost of the machines themselves. It is usually based on analysis of acceleration signals from accelerometers, which can be moved from one measurement point to another.

1.5 Vibration Transducers

Transducers exist for measuring all three of the parameters in which lateral vibration can be expressed, viz. displacement, velocity and acceleration. However, the only practical (condition monitoring) transducers for measuring displacement, proximity probes, measure relative displacement rather than absolute displacement, whereas the most common velocity and acceleration transducers measure absolute motion. This is illustrated in Figure 1.1, which shows a bearing pedestal equipped with one horizontal accelerometer and two proximity probes at 90° to each other. The latter, even though termed vertical and horizontal, would normally be located at $\pm 45^\circ$ to the vertical so as not to interfere with the usual bolted flange in the horizontal diametral plane of the bearing (Figure 1.2).

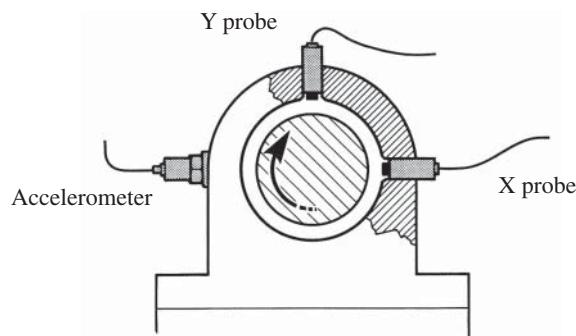


Figure 1.1 Illustration of absolute vs relative vibration. Source: Courtesy Brüel & Kjær.

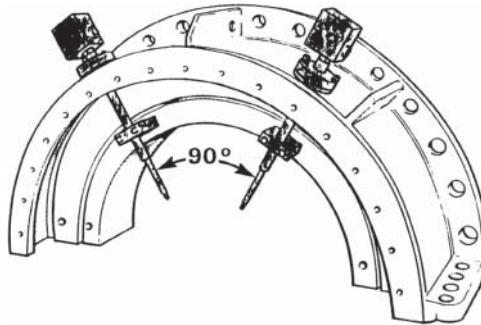


Figure 1.2 Proximity probes installed in a turbine bearing cap.

1.5.1 Absolute vs Relative Vibration Measurement

Proximity probes measure the relative motion between a shaft and casing or bearing housing (as illustrated in Figure 1.1). It is important to realise that this gives very different information from the absolute motion of the bearing housing, as measured by a so-called ‘seismic transducer’ as exemplified by an accelerometer. These two parameters are probably as different as the temperature and pressure of steam, even though sometimes related.

The relative motion, in particular for fluid film bearings, is most closely related to oil film thickness, and thus to oil film pressure distribution, as calculated using Reynolds equation [22]. It is thus also very important in rotor dynamics calculations, as these are greatly influenced by the bearing properties. These questions are discussed in more detail in Chapter 2, which gives further references on fluid film bearings and rotor dynamics. However, a fluid film bearing is a very nonlinear spring, and therefore the amplitude of relative vibration does not give a direct measure of the forces between the shaft and its bearing. An increase in static load, for example, causes the oil film to become thinner, and the bearing stiffer, with reduced vibration amplitude, even though the higher load might be more likely to cause failure.

The absolute motion of the bearing housing, on the other hand, responds directly to the force applied by the shaft on the bearing (these being the same since the inertia of the oil film is negligible), and since the machine structure tends to have linear elastic properties, the vibration amplitude will be directly proportional to the force variation, independent of the static load.

In other words, the journal bearing stiffness and damping properties, and thus the dynamic bearing forces, are most directly related to the relative position and motion of the shaft in the bearing, but the response to these forces is most directly indicated by the absolute motion of the housing. An advantage of proximity probes is that they can measure both the absolute position of the shaft in the bearing and also the vibrations around the mean position. DC accelerometers do exist, but are rarely used in machine monitoring, since it is still not possible to integrate the signals directly to total velocity and displacement because of the lack of constants of integration. Accelerometers are thus used to measure fluctuations in acceleration around a mean value of zero. This can be integrated to absolute velocity and displacement (fluctuations), but excluding zero frequency. It should be kept in mind that if the zero frequency expected values of acceleration or velocity were different from zero, the machine would not be staying in the same place.

Other comparisons between the different types of transducers depend on the technical specifications for dynamic range, frequency range, etc., so each type will be discussed in turn.

1.5.2 Proximity Probes

Proximity probes give a measure of the relative distance between the probe tip and another surface. They can be based on the capacitive or magnetic properties of the circuit including the gap to be measured, but by far the most ubiquitous proximity probes are those based on the changes in electrical inductance of a circuit brought about by changes in the gap. Such probes were pioneered by the company Bently Nevada, now owned by GE, and are very widely used for machine monitoring, becoming a standard for the American Petroleum Institute (API) [23]. Figure 1.2 shows typical proximity probes installed in the bearing cap of a turbine.

The medium in the gap must have a high dielectric value, but can be air or another gas, or for example the oil in fluid film bearings. The surface whose distance from the probe tip is being measured must be electrically conducting, so as to allow the generation of eddy currents by induction. A signal is generated by a ‘proximitor’ (oscillator/demodulator) at a high frequency, and its amplitude is directly dependent on the size of the gap between the probe and the measurement surface. Amplitude demodulation techniques are used to retrieve the signal. A typical probe can measure reasonably linearly in the gap range from 0.25–2.3 mm with a maximum deviation from linearity of 0.025 mm (1.1% of full scale) with a sensitivity of 200 mV mil^{-1} (7.87 V mm^{-1}). Thus, in the sense of the ratio of maximum to minimum value, the dynamic range is <20 dB, but in the sense of the ratio of the maximum to minimum component in a spectrum, this would be limited by the nonlinearity to at best 40 dB.

Linearity is not the only factor limiting the dynamic range of valid measurement. By far the biggest limitation is given by runout, called ‘glitch’ by Bently Nevada [24]. Runout is the signal measured in the absence of actual vibration, and is composed of ‘mechanical runout’ and ‘electrical runout’. Mechanical runout is due to mechanical deviations of the shaft surface from a true circle, concentric with the rotation axis, and these include low frequency components such as eccentricity, shaft bow and out-of-roundness, and shorter components from scratches, burrs and other local damage. Electrical runout is due to variations in the local surface electrical and magnetic properties, and can be affected by residual magnetism, and even residual stresses, as well as subsurface imperfections. Much can be done to minimise runout before a shaft goes into service [24], but in general it is unlikely that the dynamic range from the highest measured component to the highest runout component would be more than 30 dB. It is possible to use ‘runout subtraction’ to compensate to some extent for runout, but the benefits are very limited. In principle, the runout, both mechanical and electrical, can be measured under ‘slow roll’ conditions (<10% of normal operating speed), when it can be assumed that the vibration is negligible, and then subtracted from measurements at higher speed. This is most valid for the first harmonic (fundamental frequency) of rotation, and can often be done by the monitoring system by vector subtraction. It is unlikely to be valid above critical speed for measurements made below critical speed, at least where the runout is due to shaft bow. Another reason why the runout subtraction might not be valid is that on large machines, thermal expansion from low to high load/speed means that the section of shaft on which the slow roll measurements are made is different from that aligned with the probes under normal operating conditions. Some machines are required to run without shutdown for one or more years, and the monitoring position on the shaft is also subject to change through wear of thrust bearings. Proximity probes are in fact used in axial position monitoring of rotors.

Interestingly, the dominant standard for shaft vibration monitoring, the American Petroleum Institute’s API 670 [23], states that no correction is to be made for runout in indicated vibration levels. It also states that the total runout should not exceed 25% of the maximum allowed peak-to-peak vibration amplitude. This corresponds to just –12 dB, and is the best indicator of the valid dynamic range of a proximity probe measurement. Even where runout subtraction can be carried out successfully,

it is unlikely the improvement in dynamic range would be more than 10 dB, say from 30 to 40 dB, and that primarily at low harmonics. The higher the harmonic, the shorter the wavelength, and thus the greater the likelihood that measured runout would be affected by small axial displacement due to thermal expansion or wear.

The valid frequency range of proximity probes is typically 10 kHz, but this is misleading, as the actual limit is likely to be given by a certain number of harmonics of the shaft speed, because of the dynamic range limitation. As explained in Section 4.2.1, mechanical vibrations tend to have roughly uniform spectra in terms of velocity, and thus reducing in terms of displacement as $1/f$, where f is the frequency. It is unlikely that more than 10 or so harmonics would be within the valid dynamic range as restricted by runout. This severely restricts the diagnostic capabilities of proximity probes, in particular for long-term advance warning of incipient failure, and is the main reason why the major part of this book is devoted to analysis of accelerometer signals, which have much larger dynamic and frequency ranges, as explained in Section 1.5.4.

A typical example of the restricted frequency range of proximity probe measurements is given in Figure 1.3, which compares spectra of signals from a proximity probe and an accelerometer on the same machine at the same time. The signals were recorded by the author from the monitoring system of a centrifugal compressor in a Canadian chemical plant.

The spectrum of the proximity probe signal (Figure 1.3a) is completely dominated by harmonics of the shaft speed (133 Hz). However, only the first two or three are presumably valid, as the higher harmonics are quite uniform. In the spectrum of the accelerometer signal (Figure 1.3b), which has been integrated to (absolute) displacement for easier comparison with the (relative) displacement of the prox. probe signal, the first two or three harmonics protrude above the noise. However, at higher frequencies there are four harmonics of the vane pass frequency (11 vanes) visible, which could

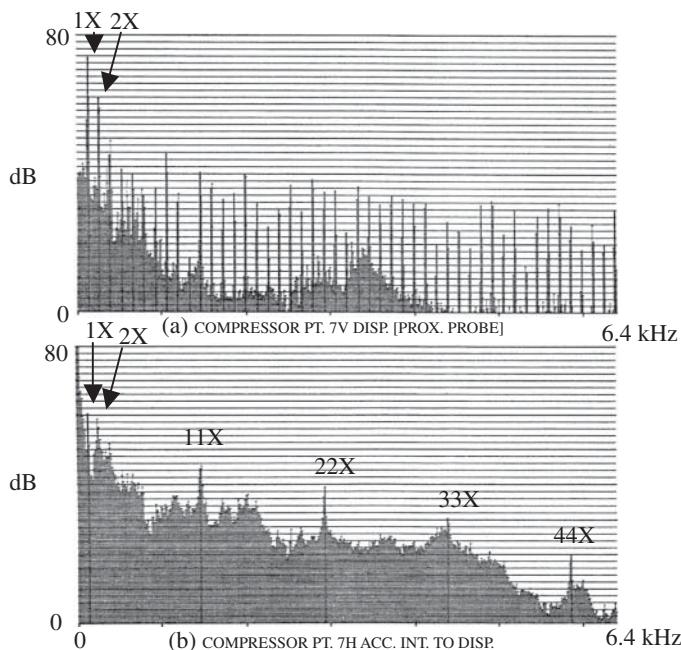


Figure 1.3 Comparison of spectra measured on a centrifugal compressor (a) Proximity probe (b) Accelerometer signal (integrated to displacement).

be useful for diagnostic purposes (see Section 2.2.4). There is nothing remarkable about the same harmonics in the prox. probe spectrum. Note that even though the accelerometer signal contains much more noise from gas flow, the latter can be removed by synchronous averaging (see Section 5.3.1) exposing the harmonics of shaft speed. On the other hand, this cannot be used to remove the runout effects from prox. probe signals, since they are perfectly periodic with the rotation speed of the shaft.

1.5.3 Velocity Transducers

Transducers do exist which give a signal proportional to absolute velocity. They are effectively a loudspeaker coil in reverse, and typically have a seismically suspended coil in the magnetic field of a permanent magnet attached to the housing of the transducer (as in Figure 1.4) or the inverse, where the coil is rigidly attached to the housing and the magnet seismically suspended. A body is said to be seismically suspended when it is attached to another by a spring such that when the second body is vibrated, the first will move with it at low frequencies, but when the excitation frequency exceeds the natural frequency of the suspended mass on its spring, it will remain fixed in space, and the first body will move around it. When the housing of the transducer (or pickup) is attached to a vibrating object, the relative motion between it and the seismically mounted component (for frequencies above the suspension resonance) is equal to the absolute motion of the object in space. To avoid problems with excessive response to excitation in the vicinity of the resonance frequency, the damping of the suspension is usually quite high, typically of the order of 70% of critical damping, and this also means that the amplitude response of the transducer is reasonably uniform almost down to the resonance frequency.

Figure 1.5 (from [25]) shows the frequency response of a generalised vibration transducer of the type described, for different values of damping, against frequency ratio with respect to the natural frequency of the suspension. It is seen that for critical damping ratio $\zeta = 0.7$, the frequency range for amplitude ratio close to 1 (i.e. output equals input) is as wide as possible. On the other hand, for this value of damping, the phase deviation from 180° (the ideal asymptotic value) extends to quite a high frequency. This means that the amplitude spectrum of signals captured with such a transducer will be quite accurate, but waveforms will not necessarily follow the original. As will be seen from the Fourier analysis theory of Chapter 3, to reproduce repetitive impulses for example, all harmonics must be in phase at the time of occurrence, and this would not be the case if they were measured with such a transducer if the low harmonics were in the range of phase distortion.

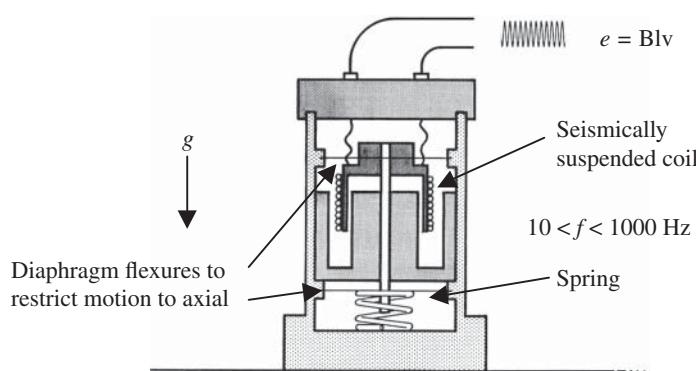


Figure 1.4 Schematic diagram of one realisation of a velocity pickup. Source: Courtesy Brüel & Kjær.

In the case of a velocity pickup, the relative motion of the magnet in the coil gives a voltage signal proportional to velocity (and thus the absolute velocity of the housing). The dynamic range (ratio of largest to smallest measurable signal) of such a transducer is about 60 dB. The lower frequency limit is typically set (by adjustment of the suspension resonance frequency) to 10 Hz, while the highest measurable frequency is limited by the resonances of internal components to about 1–2 kHz. Much of the data for the VDI 2056 and ISO 2372 standards (Section 4.2.1) was gained with velocity transducers of this kind, and that is the main reason why the frequency limits in those and later standards are 10 Hz–1 kHz.

Relative to accelerometers, velocity pickups are much heavier and bulkier. It will be shown in the next section that an accelerometer plus integrator is a much better velocity transducer.

1.5.4 Accelerometers

Accelerometers are transducers which produce a signal proportional to acceleration. By far the most common type for use in machine condition monitoring are piezoelectric accelerometers, which make use of the piezoelectric properties of certain crystals and ceramics. Such piezoelectric elements generate an electric charge proportional to strain. In a typical design as shown in Figure 1.6a, a so-called ‘compression’ type, the piezoelectric elements are sandwiched between a mass and the base, the whole assembly being clamped in compression via a spring. This arrangement can also be considered as one representation of the general vibration transducer whose frequency characteristics are shown in Figure 1.5, except that it is designed to operate below the natural frequency of the suspended element (in the ‘Range for accelerometer’ in Figure 1.5). When the base of the accelerometer is connected to a vibrating object, the mass is forced to follow the motion of the base by the piezoelectric elements, which act as a very stiff spring. The varying inertial force of the mass causes the piezoelectric elements to deform slightly, giving a strain proportional to the variation in acceleration. They then produce an electric charge proportional to this acceleration, and so their sensitivity is quoted in pico-Coulombs per metre per sec², (pC/ms⁻²). As discussed below, this must be converted to a voltage by a charge amplifier. The clamping spring, while very stiff, is much less stiff than

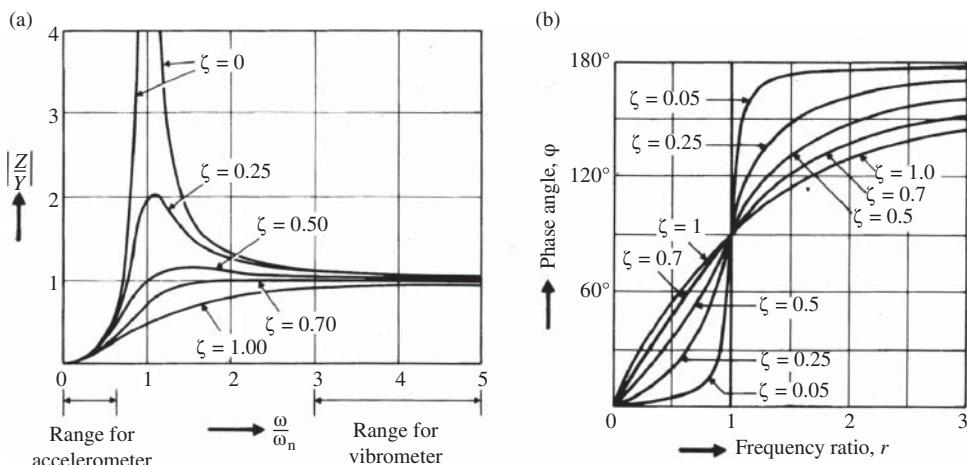


Figure 1.5 Frequency response of a seismically suspended vibrometer [25]. (a) Amplitude characteristic (b) Phase characteristic. ζ = critical damping ratio.

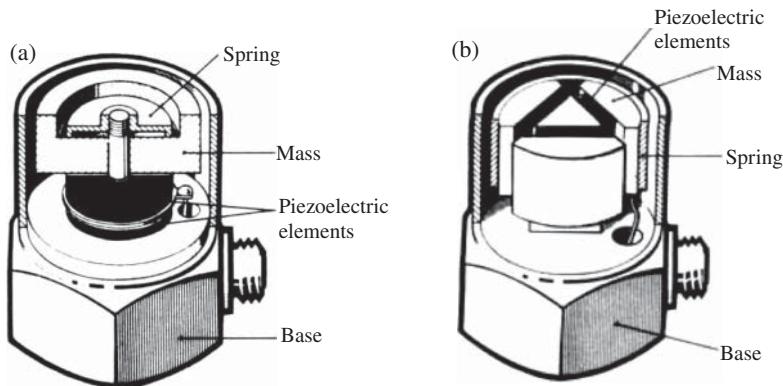


Figure 1.6 Typical accelerometer designs (a) Compression type (b) Shear type. Source: Courtesy Brüel & Kjær.

the piezoelectric elements, so its force remains effectively constant, but it is required to maintain a positive compression force on the assembly.

Figure 1.6b shows an alternative design where the piezoelectric elements deform in shear (they must be polarised so as to produce a charge proportional to shear rather than compressive strain). This particular design is the patented ‘delta shear®’ design, by Brüel & Kjaer, where the centre post to which the assembly is clamped has an equilateral triangular or delta (Δ) cross section, meaning that the mechanical properties of the assembly are isotropic, with no preferential direction. The spring in this case is a cylindrical clamping spring, once again to maintain positive compressive forces between the masses, the elements and the centre post. Other shear designs exist, where the piezoelectric elements are clamped in one direction against a rectangular centre post, but this has the disadvantage of different transverse resonance frequencies in different directions. Another isotropic design uses cylindrical elements and masses, but these must then be cemented together, giving lower structural integrity, and temperature limitations.

The electrical circuit including the piezoelectric elements has very high impedance, and is subject to a number of problems, such as pickup of signals from electromagnetic radiation. The latter is minimised by using coaxial cables with an outer braided wire shield. The type of cable connector shown in Figure 1.6, a so-called ‘microdot’ connector, gives the best results for laboratory measurements, but the associated standard microdot cables are not very practical for regular measurements in the field. A more robust double shielded microdot cable solves some of these problems, but other types of cables with more robust TNC connectors or equivalent may be found preferable for regular monitoring, even if the repeatability and frequency range are degraded slightly.

If the transducers are connected directly to a voltmeter, the voltage corresponding to the generated charge is directly affected by the impedance of the circuit, and in particular the capacitance of the accelerometer cable, which varies with its length. For this reason, it is generally necessary to use accelerometers together with charge amplifiers, which convert a given charge at the high impedance input side to a proportional voltage on the low impedance output side. A typical design converts 1 pC at the input to 1 mV at the output. Cables can be very long on the output side without effect on sensitivity, and with negligible noise pickup. Another problem with the high impedance circuit is the sensitivity to ‘triboelectric noise’, or generation of a static electric charge by rubbing between the inner conductor and its sheath. Triboelectric noise is often minimised by using a low friction PTFE sheath and graphite lubricant between the inner conductor and the sheath.

The problems with the high impedance circuit can be reduced to a minimum by having the charge amplifier built into the transducer. Miniaturisation of electronic circuits has now made this possible, and it does solve many of the practical problems associated with special cables, electrical interference, etc. It does have one disadvantage, and that is that it is more difficult to detect overload of the input circuit. Separate charge amplifiers often have an overload indicator, and the possibility to change the gain either before or after filtration, integration etc. It should always be kept in mind that the input circuit has to cope with the full signal generated by the transducer, even that part outside the final frequency range selected by high- and low-pass filters. Piezoelectric accelerometers have low internal damping, and it is quite common for the resonant gain in Figure 1.5a to be 30 dB above the linear value in the operating range. The resonance frequency is typically 30 kHz, but in some machines, such as gas turbines, there can be significant excitation at that frequency. This might cause overload of the input amplifier, even if the signal is lowpass filtered at 10–20 kHz (the maximum valid frequency of measurement) in the amplifier.

The resonant frequency for transverse motion of the accelerometer is often lower than that in the main measurement direction, in particular for compression type accelerometers, and even though the transverse sensitivity is only a few percent, the signals can become distorted if the transverse resonance is excited. One way to solve this problem (and the excitation of the main axial resonance) where there can be strong excitation at such high frequencies, is to mount the accelerometer on a ‘mechanical filter’. This contains an elastomeric layer having a spring constant such that in combination with the mass of the transducer the mounted resonance is say 1/3 that of the transducer itself, at the same time providing good damping to reduce the resonance peak to between +3 and +5 dB. As opposed to scientific and laboratory measurements, for condition monitoring purposes it is most important that the frequency response of the transducer system is repeatable, rather than strictly linear, as the measurement in any case represents an external measurement of internal events, and the response of the transducer can be considered part of the response of the machine itself at that measurement point. Thus, it is quite common to use accelerometer signals up to 50–65% of the transducer resonance, where the deviation from linearity might be as high as 5 dB, even though the recommended range for linear measurements is typically 1/3 the resonance frequency. In the same way, the resonance frequency of the mechanical filter can be within the measurement range, as long as it is repeatable. In this connection it should be kept in mind that the stiffness properties of the elastomer will vary with temperature, so care should be taken if the temperature of the mounting point is subject to variation.

1.5.4.1 Frequency and Dynamic Ranges

One of the main advantages of accelerometers is the extremely wide range of both amplitude and frequency that they provide. The typical dynamic range of an accelerometer is 160 dB ($10^8 : 1$), although in conjunction with an amplifier this might be reduced to 120 dB ($10^6 : 1$) for a particular gain setting on the amplifier. As mentioned in the preceding section, a typical upper frequency limit for condition monitoring purposes is 10–20 kHz, while the lowest valid frequency is below 1 Hz.

It should be noted that such a low minimum frequency is an advantage of the shear design, since a major limiting factor at low frequency is given by the fact that the sensitivity of piezoelectric materials varies to some extent with temperature. Because of the thermal inertia of the transducer, its rate of temperature variation is limited, and there is no problem above a certain frequency. As will be seen in Figure 1.6, in the compression design the piezoelectric elements are in direct contact with the base, and can respond more rapidly to temperature change, whereas in the shear design, the elements are more isolated from the base. Another reason for generation of noise at low frequencies is ‘base strain’, where the larger bending deflections corresponding to a given acceleration at low frequency

can distort the piezoelectric elements, but much less in the shear design, once again because the triangular centre post is more isolated from the base. Thus, the lower limiting frequency of a compression type accelerometer might be as high as 5–10 Hz.

Figure 1.7 compares typical dynamic and frequency ranges for the three main types of transducers, and it is immediately evident that the accelerometer has much wider ranges than the other two. The dynamic range is shown in terms of the measured parameter, i.e. acceleration for an accelerometer, velocity for a velocity pickup, and relative displacement for a proximity probe. Superimposed on the original diagram are two possible ranges for an accelerometer and integrator, which produces a signal proportional to velocity. It assumes that the accelerometer has a dynamic range of 120 dB for one gain setting, but this can be moved by a further 30–40 dB by gain adjustment. Even though the integrator represents a lowpass filter, with slope –20 dB per frequency decade, the dynamic range of the combination is still >60 dB over a frequency range of three decades. It is illustrated how this three decade range can be simply switched from (for example) 10 Hz–10 kHz to 1 Hz–1 kHz, even with the same accelerometer and (external) amplifier. This is obviously much more flexible than

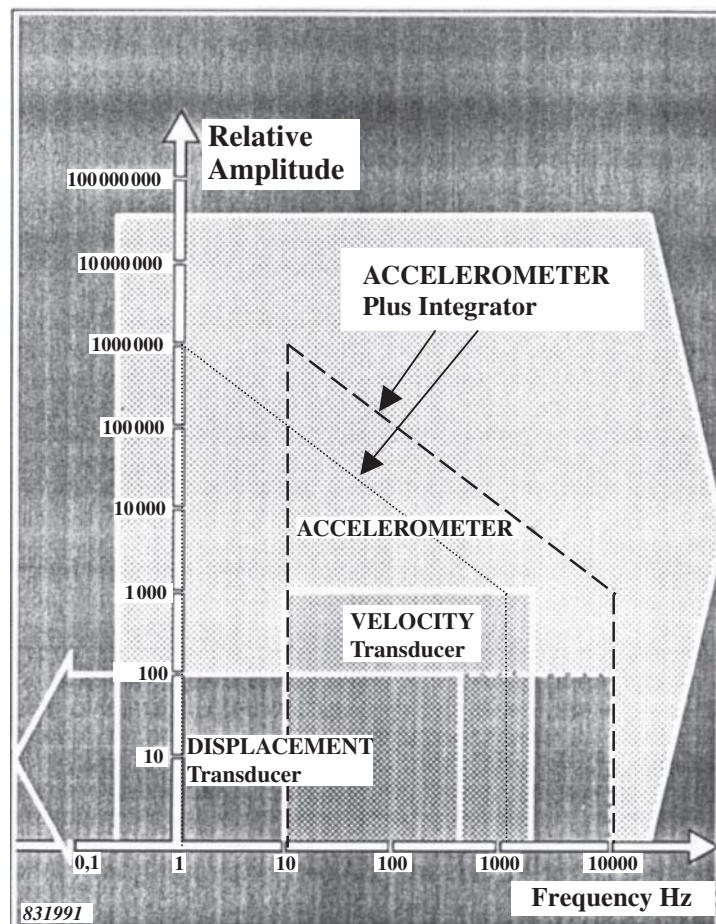


Figure 1.7 Typical frequency and dynamic ranges for the three main transducer types with superimposed ranges for an accelerometer and integrator. Source: Courtesy Brüel & Kjær.

the fixed approximate two decade range of the velocity pickup. The technical specifications of the combined accelerometer/integrator are also better; for example the phase distortion can be made negligible within a factor of 3 above the integration cutoff frequency.

Using an electronic integrator, such as built into some external charge amplifiers, was very valuable when the typical dynamic range of recorders and analysis systems was 50–80 dB, since it meant that this limited range could accommodate the velocity signal, which inherently occupied the minimum dynamic range (as explained above). With modern data acquisition systems, the dynamic range of amplifiers and analogue-to-digital (AD) converters is more typically 120 dB, and so signals recorded as acceleration can be integrated numerically to velocity if desired. This has the advantage that it can be done with non-causal integrators with zero phase distortion within the measurement range.

Velocity signals can be further integrated to displacement, either electronically or numerically, whereas the inverse operation of differentiating displacement signals to velocity, and velocity signals to acceleration tends to introduce problems by amplification of high frequency noise. It can be done successfully numerically, as long as the frequency components above the highest valid frequency (not dominated by noise) are first removed by lowpass filtration.

1.5.5 Dual Vibration Probes

Shaft vibration is normally measured by proximity probes, but this gives the motion relative to the housing. To obtain the absolute motion of the shaft, it is necessary to add this relative motion to the absolute motion of the housing, and so-called ‘dual probes’ are designed to do this. They contain both a proximity probe, and a seismic probe to measure the absolute motion of the housing. The seismic probe can either be a velocity transducer (with signal integrated to absolute displacement) or an accelerometer (with signal double integrated to absolute displacement). The overall frequency and dynamic ranges of the combination would normally be limited by the proximity probe, so it could be said that the accelerometer gives no particular advantage over the velocity probe, but on the other hand the accelerometer signal could be separately analysed in its own right, in which case it could give some advantage.

The ratio of relative to absolute vibration varies widely from machine to machine, and so where it is important to know how the shafts of adjacent machines are vibrating (because they have to be connected by a coupling for example) then this has to be on the basis of overall absolute motion, as produced by a dual probe.

1.5.6 Laser Vibrometers

In recent years there has been a rapid development of vibration transducers based on the laser Doppler principle. In this technique, a coherent laser beam is reflected from a vibrating surface, and is frequency shifted according to the absolute velocity of the surface (in the direction of the beam) by the Doppler Effect. The frequency shift is measured by an interferometer and converted to velocity. Note that because the frequency shift occurs at the reflection, the result is virtually independent of the motion of the transmitter/receiver; in other words, it does measure absolute rather than relative motion.

Laser vibrometers have the big advantage that they do not load the measurement object, and the measurement point can be changed easily and rapidly by deflecting the light beam. This is useful for making repeatable measurements over a grid in the minimum time possible. For this reason, they are now used extensively for modal analysis measurements, and perhaps to a lesser extent for operational

deflection shape (ODS) measurements. The latter can be very useful for diagnostic purposes, even though not discussed explicitly in this book, but because a scanning laser vibrometer system is so expensive (up to hundreds of thousands of dollars) they would only have a very limited application in machine monitoring. Even without the scanning system, the vibrometers are quite bulky and difficult to move around, so they could not at present be used for intermittent monitoring. It is possible that in the future they will be miniaturised to such an extent that they could be used for portable field measurements. The author has heard a presentation where the presenter mused that in the future they could be built into a hard hat, and the operator would just have to look in the direction of a machine, utter the ID of the machine into a microphone, and the laser and imaging system would locate the machine and take measurements at a prescribed number of monitoring points on it. Currently, however, they are not really a viable option for regular condition monitoring, even though they are used for example in production quality control measurements [26].

1.6 Torsional Vibration Transducers

Some failures in machines occur because of excessive torsional vibration. When the machine has only one compound shaft, for example a motor driven pump or a turbine driven centrifugal compressor, there is very little coupling between torsional vibrations and lateral vibrations as measured by either accelerometers or proximity probes, and so it is desirable to measure the torsional vibrations directly. When there is a gearbox in the train, there is some coupling, because input and output torques, and torque fluctuations, are different on either side of the gearbox, and the differences have to be supported by the housing and foundation, giving rise to lateral vibrations. However, it can still be an advantage to measure the torsional vibrations directly, in part to separate them from purely lateral vibrations from other sources.

Even when not representing potential failure in torsion, torsional vibrations sometimes carry significant diagnostic information as to machine condition, such as with reciprocating machines for example, where variations in torsional vibration indicate non-uniform torque inputs from different cylinders. This is explained in more detail in Sections 2.3.2, 4.3.3, and 7.4.2. Yet another example where it can be advantageous to measure torsional vibrations is in connection with gears, where the dynamic transmission error is effectively the difference in (scaled) torsional vibration on the input and output sides. The reason for the scaling is that the error is actually a linear displacement along the line of action of the meshing of the two gears, and thus represents a different rotational angle on each if the gear ratio is not 1 : 1. The use of transmission error as a diagnostic tool is explained in Section 7.2.2.

Thus, the various means of measuring torsional vibrations are now discussed.

1.6.1 Shaft Encoders

Shaft encoders are not a torsional vibration transducer as such, but information about torsional vibration (i.e. angular velocity variations) can be obtained by analysing shaft encoder signals. Shaft encoders give out a series of pulses at equal angular intervals, with typically 1024 per revolution. They are sometimes attached to the free end of a shaft (with the housing attached to the housing of the machine), but ‘through-shaft’ encoders also exist, which can be placed elsewhere on the shaft, possibly even between bearings. Mounting on the free end of a shaft would often be via a flexible coupling, which restricts the range of frequencies that can be transmitted to the encoder, but which means that low harmonics of shaft speed are faithfully reproduced. When flexible couplings are not used, even slight misalignment can give some distortion of low harmonic components, but this

may not be a problem if information is primarily desired about higher harmonics (e.g. toothmesh harmonics of gears).

There are two methods that can be used to extract torsional vibration information from shaft encoder signals. The first is to use phase and/or frequency demodulation of the encoder pulse frequency, as described in Sections 3.3.2, 4.3.3, 7.2.2, and 7.4.2. Phase demodulation obtains the torsional vibration information in terms of angular displacement, while its time derivative, frequency demodulation, expresses it in terms of angular velocity. It is possible to take a further differentiation to express it in terms of angular acceleration, but there is no equivalent modulation term. The second method is to use a very high frequency clock (typ. 80 MHz) to measure the time intervals between pulses from the encoder. The reciprocal of this can be scaled in terms of average angular velocity in the interval. One advantage is that a fixed number of samples is obtained per revolution, so there is no need to perform order tracking (see Section 5.1) to compensate for slow speed changes. Both methods are discussed and compared in [27].

1.6.2 Torsional Laser Vibrometers

Torsional laser vibrometers also exist, which have two laser beams directed at the surface of a rotating shaft [28]. The reflected signals are processed in such a way that everything is cancelled except the torsional motion expressed as angular velocity. It can be shown that this applies for arbitrary shape of the cross section of the shaft.

As for laser vibrometers which measure lateral motion, the major advantage is that the measurement is non-contact (though a section of the shaft must be exposed to view), but because of the dual beam they are even more expensive than single beam vibrometers. An example of the use of a torsional laser vibrometer to measure the speed fluctuations of the crankshaft of a large diesel engine is given in Section 8.4.1, but it is pointed out that equivalent results were obtained using a much cheaper shaft encoder. There was a slight difference in that the laser vibrometer measured the overall angular motion of the crankshaft, including rocking of the engine block, whereas the encoder measured the torsional fluctuations relative to the block, which in that case were more relevant, but the differences were small.

1.7 Condition Monitoring – The Basic Problem

It is worth discussing the way in which condition information can be extracted from vibration signals. Measured vibration signals are always a combination of source effects and transmission path effects. In general, as illustrated in Figure 1.8, a measurement at one point will be a sum of responses from a number of sources. Such a system is known as a multiple-input multiple-output (MIMO) system.

The contribution to the response at one measurement point from one source, in the time domain, is a convolution of the force signal with the impulse response function (IRF) of the transmission path from the source to the measurement point (see Section 3.2.6). In the frequency domain (and Laplace domain) this simplifies to a product of their respective spectra, the spectrum of the IRF being equal to the corresponding frequency response function (FRF). This can be represented symbolically as:

$$x_i = \sum_j s_j * h_{ij} \quad (1.1)$$

$$X_i = \sum_j S_j H_{ij} \quad (1.2)$$

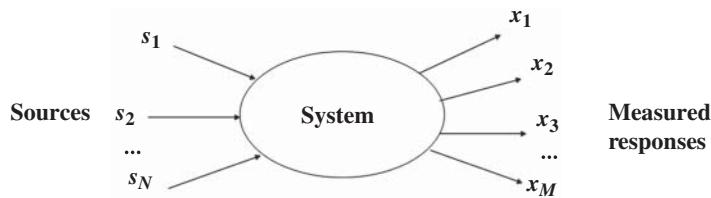


Figure 1.8 Illustration of a number of measured responses of a system due to a number of excitation sources.

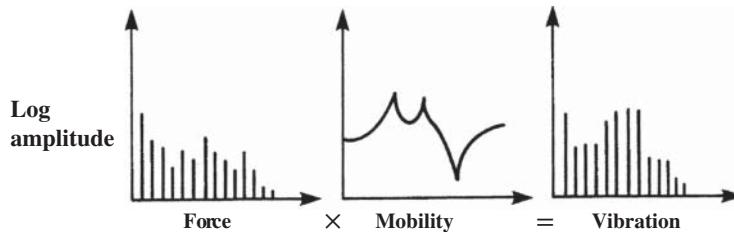


Figure 1.9 Combination of forcing function and transfer path to give response vibration for one source.
Source: Courtesy Brüel & Kjær.

where the upper-case letters in Eq. (1.2) represent the Fourier transforms of the lower-case symbols, and the asterisk represents the convolution operation (Section 3.2.6).

For an individual source (forcing function) in Eq. (1.2) the product is even further simplified to a sum by taking the logarithm. Since the FRF H_{ij} is complex, its logarithm has log amplitude ratio (log gain) as real part, and phase as imaginary part, meaning that the log amplitude of the response is the sum of the log amplitudes of the source and FRF, and the phase is the sum of the phases. This is illustrated in Figure 1.9 for the log amplitudes only. The ‘mobility’ is the FRF corresponding to force input and vibration velocity output. Note that the indicated multiplicative relationship is actually depicted as additive on log amplitude scales.

In Figure 1.9 the forcing function is depicted as consisting of discrete frequency components, which is typical for many machines running at constant speed. It illustrates that resonance frequencies do not appear in response spectra in such cases unless directly excited by a forcing frequency. For broadband noise excitation the response spectra will have peaks at resonance frequencies, but not discrete frequency peaks. These are usually recognisably different from discrete frequency components because of the broadness of the base, at least on log amplitude scales, and using a window function such as Hanning (Section 3.2.8.2). Figure 1.10 shows a numerically generated example combining three discrete frequency components at [400, 800, 1200] Hz with a narrowband resonance at 1000 Hz excited by white noise (averaged over 1000 spectra), and using Hanning weighting. The difference cannot be seen on linear amplitude scales.

Even in the MIMO case, where a general response spectrum is no longer the product of a single source spectrum and single FRF, but a sum of these, the appearance of the spectrum is still very similar, in particular on log amplitude scales. This is partly because the same discrete frequency components tend to appear at all measurement points with different strengths, and resonance frequencies are global properties of a structure, so tend to appear in all FRFs, once again with different strengths, so that one or two paths will tend to dominate for a particular measurement point and for a particular resonance peak.

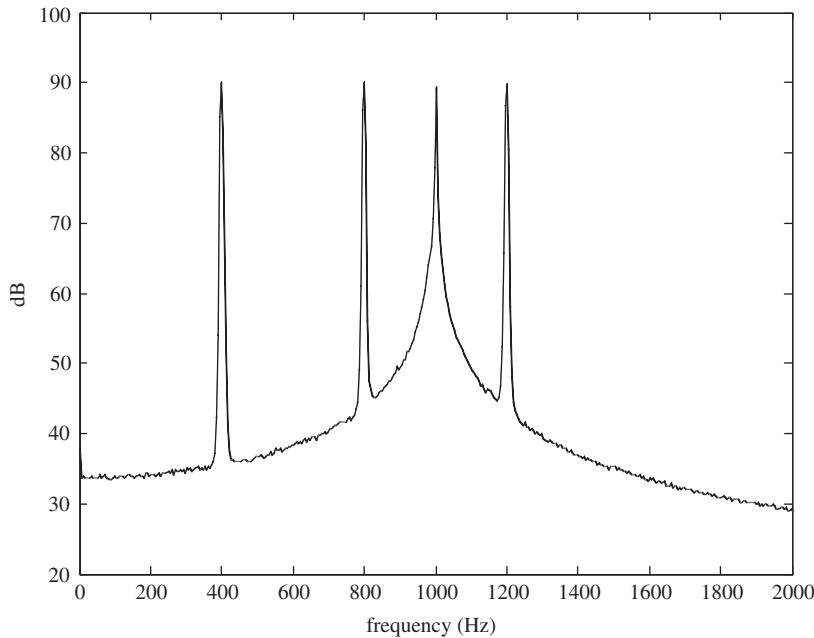


Figure 1.10 Comparison of the spectra of discrete frequency components at 400, 800, and 1200 Hz with that of a narrow band resonance at 1000 Hz.

The basic problem in diagnosing the reason for changes in response vibrations is to decide whether the change has occurred at the source(s) or in the structural transmission path.

Quite often, a change in condition results from a change at a source, such as an increase in unbalance force, or a change in the force between meshing gears. On the other hand, other types of faults may primarily result in changes in the structural response, such as a developing crack in a machine casing. Sometimes the two effects couple with each other, with the change in structural response giving a change in the forcing function. In one such case, a developing tooth root crack obviously affects the local structural properties, in particular the stiffness of that particular tooth (as discussed in Section 7.2.4), but in terms of responses at the bearings, over much of the frequency range this can primarily be interpreted as a change in the forcing function at the toothmesh. In the case of a crack in a shaft, as discussed in Section 2.2.1.2, if the crack is ‘breathing’, i.e. opening and closing with every revolution of the shaft, the character of the forcing function changes, giving rise to responses at the odd harmonics of the shaft speed, in contrast to the even harmonics primarily generated when the crack is permanently open.

Apart from a number of coupled cases as just mentioned, in a broad generalisation it can be said that for machines running at constant speed, sinusoidal components in response signals result from sinusoidal forcing functions at the same fundamental frequency, although because of structural nonlinearities, the responses will usually be distorted from sinusoidal and thus contain some level of harmonics of the fundamental frequency, even if the forcing function is relatively pure, such as a simple unbalance. This is the reason why frequency analysis is so powerful as a diagnostic tool, since families of harmonics with a given frequency spacing almost certainly result from a forcing function at that frequency. As discussed in Section 7.1, the presence of harmonics (perhaps only visible on a logarithmic amplitude scale) allows a much more accurate measurement of the fundamental

frequency, by fitting a finely tuneable ‘harmonic cursor’ to the family. In a similar way, some effects modulate other frequencies at a lower rate, and give rise to families of sidebands around the harmonics of the ‘carrier’ frequency, and a sideband cursor can find these modulating frequencies very accurately. An example is given by vibration signals from meshing gear teeth, where the harmonics of the toothmesh frequency (gear rotational speed times the number of teeth) are often modulated by the shaft speeds of the two gears in mesh (Section 2.2.2.2). Even where the carrier frequency is a random signal, but modulated by discrete frequency components, the latter can still be detected because of the so-called ‘cyclostationary’ properties of such a signal. A number of examples relating to bearing and engine diagnostics are given throughout the book.

The problem of deciding whether a change in a response signal is due to a change at the source or in the transmission path is one example of the more general problem of ‘blind source separation’ (BSS), and the related topic ‘blind system identification’. These are both areas of current research, which undoubtedly will have a considerable impact on machine diagnostics in the future, but it is a little early to include this topic in a book such as this. However, the interested reader may like to view the special issue on mechanical applications of BSS in Ref. [29].

A mechanical application of blind system identification is the topic of ‘operational modal analysis’, where inherent dynamic properties of structures are deduced from response measurements only. This is now a regular topic at conferences on more general modal analysis, and there is a special series of conferences, IOMAC (International Operational Modal Analysis Conference) specifically devoted to it. It is used for updating of simulation models of machines and faults in Chapter 8.

References

1. Rao, B.K.N. (2009). “Advances in diagnostic and prognostic strategies and technologies for failure-free maintenance of industrial assets”. *Comadem 2009*, San Sebastian, Spain (9–11 June).
2. Al-Najjar, B. and Alsouf, I. (2004). Enhancing a company’s profitability and competitiveness using integrated vibration-based maintenance: a case study. *Journal of European Operation Research* 157: 643–657.
3. Al-Najjar, B. (2007). The lack of maintenance and not maintenance which costs: a model to describe and quantify the impact of vibration-based maintenance on Company’s business. *International Journal of Production Economics (IJPPM)* 55 (8): 260–273.
4. Al-Najjar, B. and Jacobsson, M. (2013). A computerised model to enhance the cost-effectiveness of production and maintenance dynamic decisions; a case study at FIAT. *Journal of Quality in Maintenance Engineering* 19 (2): 114–127.
5. Vlok, P.J., Coetzee, J.L., Banjevic, D. et al. (2002). Optimal component replacement decisions using vibration monitoring and the PHM. *Journal of the OR Society* 53: 193–202.
6. Sundin, P.O., Montgomery, N. and Jardine, A.K.S. (2007) Pulp mill on-site implementation of CBM decision support software. Proceedings of International Conference of Maintenance Societies, Melbourne, Australia.
7. Business Wire. <https://www.businesswire.com/news/home/20200130005363/en/Global-Machine-Condition-Monitoring-Market-Size-Estimated>
8. Neale, M.J. and Woodley, B.J. (1978). “A Guide to the Condition Monitoring of Machinery” Report TRD 223 for the British Department of Industry.
9. IEEE (1999). IEEE draft standard P1438/D1.5. “Guide for Applications of Plant Condition monitoring for Hydroelectric Facilities”. Section 4.3 “Potential Benefits”.
10. McMillan, D. and Ault, G.W. (2008). Condition monitoring benefit for onshore wind turbines: sensitivity to operational parameters. *IET Renewable Power Generation* 2 (1): 60–72.
11. Braun, S. (ed.) (1986). *Mechanical Signature Analysis*. London: Academic Press.
12. Armor, A.F. (1983). On-line diagnostics for fossil power plants: the promise and the reality. In: *Proceedings of EPRI 1982 Conference and Workshop*, Hartford, CT, USA, 25–27 August, EPRI CS-2920, 1–5–1–26.
13. Raad, A., Zhang, F., Randall, B. (2003). “On the comparison of the use of AE and vibration analysis for early gear fault detection”. *The Eighth Western Pacific Acoustics Conference*, Melbourne, Australia.
14. Tan, C.K., Irving, P., and Mba, D. (2007). A comparative experimental study on the diagnostic and prognostic capabilities of acoustics emission, vibration and spectrometric oil analysis for spur gears. *Mechanical Systems and Signal Processing* 21: 208–233.

15. Lin, T.R., Kim, E., and Tan, A.C.C. (2013). A practical signal processing approach for condition monitoring of low speed machinery using Peak-Hold-Down-Sample algorithm. *Mechanical Systems and Signal Processing* 36: 256–270.
16. Robinson, J. C., Berry, J. E. (2001). “Description of PeakVue and illustration of its wide array of applications in fault detection and problem severity assessment”. *Emerson Process Management Reliability Conference 2001* (22–25 October 2001).
17. Randall, R.B. (2016). Modern envelope analysis for bearing diagnostics. *International Journal of COMADEM* 19 (3), July.
18. SPM Instrument AB. <https://www.spminstrument.com/products-and-services/transducers-and-transmitters/duotech-accelerometers/> (accessed 26 October 2020).
19. Zhu, J., Yoon, J.M., He, D., and Bechhoefer, E. (2015). Online particle-contaminated lubrication oil condition monitoring and remaining useful life prediction for wind turbines. *Wind Energy* 18: 1131–1149.
20. Peng, Y., Wu, T., Wang, S., and Peng, Z. (2017). Wear state identification using dynamic features of wear debris for on-line purpose. *Wear* 376-377: 1885–1891.
21. Sherwin, D.J. and Al-Najjar, B. (1999). Practical models for condition monitoring inspection intervals. *Journal of Quality in Maintenance Engineering* 5 (3): 203–220.
22. Childs, D. (1993). *Turbomachinery Rotordynamics*. NY: Wiley.
23. American Petroleum Institute (2000). *Machinery Protection Systems*”, API Standard 670, 4e. Washington, D.C.: API Publishing Services.
24. GE Power. Bently Nevada Application Note (2020). “GLITCH” Definition, Sources, and Methods of Correcting’. Available via http://www.gepower.com/prod_serv/products/oc/en/bently_nevada/application_notes.htm.
25. Rao, S.S. (2005). *Mechanical Vibrations*. Englewood Cliffs, NJ: Prentice Hall.
26. Vass, J., Šmíd, R., Randall, R.B. et al. (2008). Avoidance of speckle noise in laser vibrometry by the use of kurtosis ratio: application to mechanical fault diagnostics. *Mechanical Systems and Signal Processing* 22 (3): 647–671.
27. Sweeney, P.J. and Randall, R.B. (1996). Gear transmission error measurement using phase demodulation. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 210 (C3): 201–213.
28. Polytec GmbH. Available at: https://www.polytec.com/fileadmin/d/Vibrometrie/OM_PB_RLV-5500_E_42406.pdf (accessed 12 May 2020)
29. J. Antoni, S. Braun (Eds.), (2005). Blind source separation. *Mechanical Systems and Signal Processing* 19 (6) (Special issue).

2

Vibration Signals from Rotating and Reciprocating Machines

2.1 Signal Classification

As mentioned in Chapter 1, most machine components give rise to specific vibration signals that characterise them, and allow them to be separated from others, as well as distinguishing faulty from healthy condition. The distinguishing features may be because of different repetition frequencies, e.g. a garmesh frequency, which characterises a particular pair of gears, and different sideband spacings, which characterise the modulating effects of the two meshing gears on their common mesh frequency. Gear generated signals are usually at harmonics (integer multiples) of the associated shaft rotation speeds, whereas the characteristic frequencies of rolling element bearings are generally not at harmonics of the associated shaft speeds. Some signals, typically associated with fluid flow, such as turbulence or cavitation, have a random nature, but may have a characteristic distribution with frequency. These signals are often ‘stationary’, i.e. their statistical properties do not vary with time, but other random signals, characterised as ‘cyclostationary’, are often generated by machines, and have statistical properties which vary periodically. A typical example is the combustion signal in an internal combustion (IC) engine, where there is a combustion event in each cylinder each cycle (thus happening periodically), but with significant random variations from one cycle to another. With the growth in interest in variable speed machines, a new category has been proposed, viz. cyclo-non-stationary signals. Whereas cyclostationary signals arise from machines running at constant speed, cyclo-non-stationary signals arise from machines running at varying speed, so that modulating effects are no longer at constant frequency, which gives periodic modulating functions, but vary deterministically in concert with the machine speed, and can thus be extracted.

This chapter thus starts with the various categories into which vibration signals can be divided and thereby classified. The purpose in this chapter is mainly to categorise the various signals generated by machine components in healthy and faulty condition, but the type of signal also has a very large influence on the types of signal processing which can and should be applied to them, as described in Chapter 3. As mentioned, signals are often distinguished by the repetition frequencies of periodic events, and so one of the most fundamental ways of evaluating signals is in terms of their ‘frequency spectrum’, showing how their constitutive components are distributed with frequency. Mathematically, this is done with various forms of Fourier analysis, as described in great detail in Chapter 3, but at this stage it is sufficient to see how the various signal types manifest themselves in the time

and frequency domains. Some (non-stationary) signals have frequency content which varies with time, and once again this is discussed more rigorously in Chapter 3, but at this stage it is sufficient to realise that just as the human ear can recognise changes in frequency patterns with time (e.g. music), such patterns sometimes categorise certain types of machine faults. An example is given by faults in IC engines, where a trained mechanic can distinguish temporally changing frequency patterns resulting from ‘pinging’ and ‘bearing knock’. Many signal processing tools have been developed to try and replicate what can be distinguished by the human ear.

Figure 2.1 shows the basic breakdown into different signal types. The most fundamental division is into stationary and non-stationary, where, as mentioned above, stationary means that the statistical properties are invariant with time. For deterministic signals this basically means that they are composed entirely of discrete frequency sinusoids, and thus their frequency spectrum consists of discrete lines at the frequencies of those sinusoids. Once the frequency, amplitude, and initial phase (i.e. at time zero) of these components is known, the value of the signal can be predicted at any time in the future or past; hence the term ‘deterministic’.

Random signals are somewhat more complex, as their value at any time cannot be predicted, but for stationary random signals their statistical properties are unchanging with time. Individual random signals must be considered as realisations of a ‘random process’, where all realisations vary randomly, but are equally valid. The statistical properties can be obtained by averaging across an ‘ensemble’ of realisations, as illustrated in Figure 2.2. The conditions for stationarity, for measurements on a machine, are typically that the latter is operating at constant speed and load. If the function being averaged using the expectation operator $E[\cdot]$ is the signal itself, i.e. $f_x(t) = x(t)$, then the result of the average will be the mean value. If $f_x(t) = x^2(t)$, the result will be the mean square value.

One rarely has a large number of realisations of a process, and never an infinite number, so it is convenient to be able to perform the averaging along the record. This is valid if the signals are not only stationary but also ‘ergodic’. The fundamental meaning of this is that all realisations are statistically equivalent. The signals depicted in Figure 2.2 might, for example, be vibration signals measured on a number of vehicles driven at constant speed around a uniform test track. If the vehicles varied from small cars to large trucks, it is quite possible for the process to be stationary (i.e. the mean value at any time t to be constant), but for it to be ergodic, all the vehicles would have to be of the same type. It is then clear that averages along the record would have equal validity to averages across the ensemble.

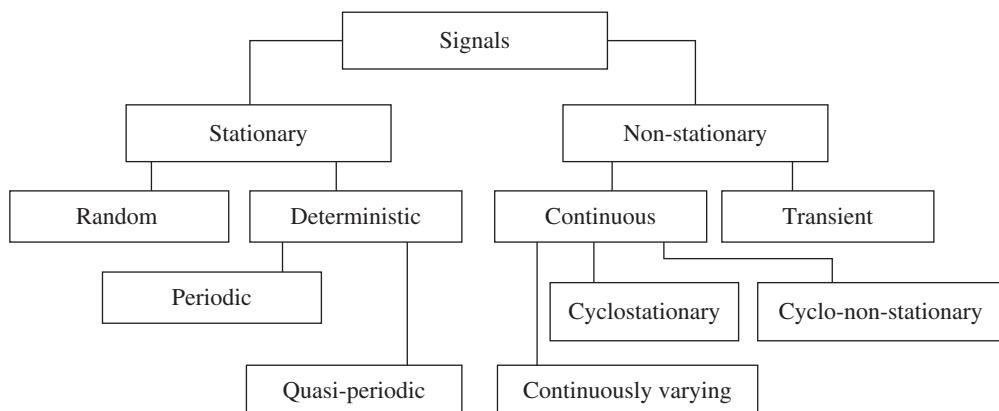


Figure 2.1 Signal types.

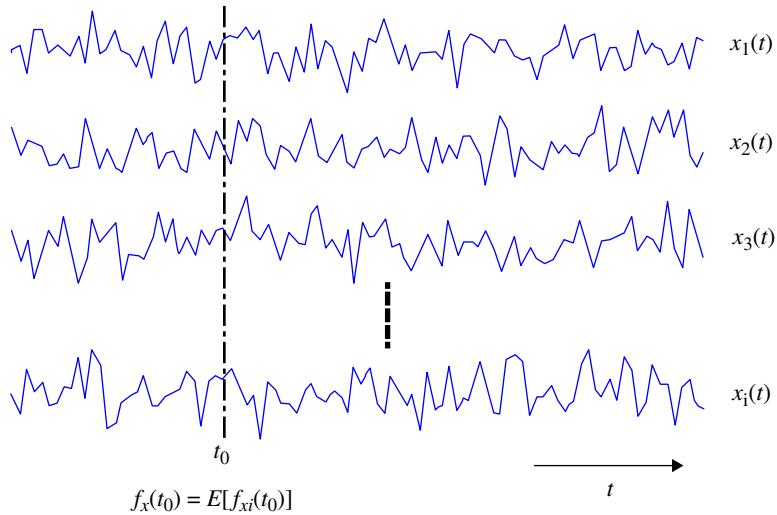


Figure 2.2 Ensemble averaging.

‘Non-stationary’ means anything which does not satisfy the conditions for stationarity, and once again it can be divided into two main classes, ‘continuously varying’ and ‘transient’. There is no hard and fast rule for distinguishing between these two types, but in general it can be said that transient signals only exist for a finite length of time, and are typically analysed as an entity. Once again, this requires clarification, since a decaying exponential function, for example, theoretically decays to infinity, but in practical terms it only has a measurable value for a finite time. The terms ‘energy’ and ‘power’ are used to distinguish between transient and continuous (stationary or nonstationary) signals. An analogy can be drawn with electrical signals in a resistive circuit, where the power $W = EI$, and E and I are the voltage and current, respectively. Since $E = IR$, where R is the resistance, the power is proportional to the square of the voltage or current, i.e. $W = I^2R = E^2/R$. Similarly, the true power associated with a vibration signal is related to the square of its amplitude through some sort of impedance or admittance function, and it is common to simply call the squared value the ‘power’. A transient signal has an instantaneous squared value or power at each point in time, but is characterised by the integral of this ‘power’ over its whole length in time, this being called its ‘energy’. A stationary random signal by definition has a constant power, and therefore infinite energy. Cyclo-stationary signals by definition have power (always positive) which varies periodically with time, and so their total energy is also infinite. Cyclo-non-stationary signals are similar except that their power varies with time in a deterministic way. Other nonstationary signals, such as vibration signals measured during the run-up or coast-down of a machine, also have a finite length, but are more likely to be considered as continually changing nonstationary signals, rather than transients, since they are typically analysed by being divided into short quasi-stationary sections, to see how their power varies with time (time/frequency analysis). Therefore, in this book a transient will be treated as a signal that is analysed as an entity, with finite energy, and not divided up into shorter sections. Typical examples would be the impulsive force corresponding to a hammer blow, and the impulse response of the structure to which the hammer blow is applied. Continuously varying nonstationary signals will often be treated by the techniques of time/frequency analysis as described in more detail in Chapter 3.

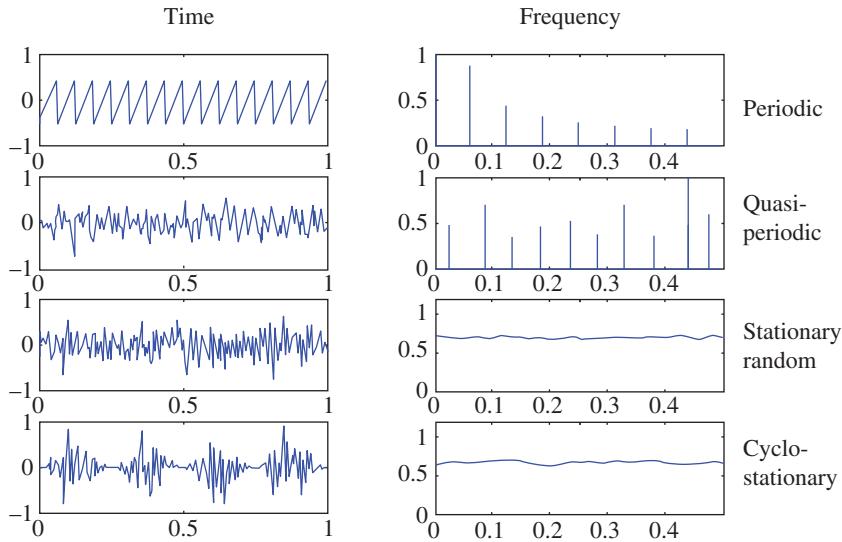


Figure 2.3 Typical signals in the time and frequency domains.

As mentioned above, the different types of signals have different characteristics in the time and frequency domains, and these are summarised in Figure 2.3 for stationary and cyclostationary continuous signals.

2.1.1 Stationary Deterministic Signals

The two first signals (periodic and quasi-periodic) are deterministic, and made up entirely of discrete sinusoidal components. For the periodic (in this case sawtooth) signal, these components are at integer multiples (i.e. harmonics) of the fundamental periodic frequency. For the quasi-periodic signal, the discrete frequencies are not all members of a harmonic series. Strictly speaking, this means that the frequency ratio between at least two components must be an irrational number, because otherwise a fundamental period could be found, but in practice it means that there is no direct relationship between at least two of the families of the constituent frequencies. A typical example is given by the vibration signals from a gas turbine engine, which has several independent shafts. Each shaft will normally generate families of harmonics, but the total signal will be quasi-periodic. In the time domain it is impossible to see that the quasi-periodic signal is made up entirely of periodic sine waves, but in the frequency domain it can be seen to only contain discrete frequency components. A quasi-periodic signal can be treated as a special case of a periodic signal where the period tends to infinity.

A particular type of periodic signal which is sometimes useful is a ‘pseudo-random’ signal. This can be considered as a section of random signal (sometimes quite long) which is repeated periodically. As for random signals, the phase relationship between adjacent frequency components is apparently random (although also repeating periodically, so also pseudo-random).

2.1.2 Stationary Random Signals

The third signal (stationary random) does not appear very different in the time domain from the quasi-periodic signal, but its spectrum is entirely different, with no discrete frequencies, and its

spectral power distributed continuously with frequency. The example shown is ‘white noise’, which over the frequency range considered has a uniform spectrum. To obtain such a smooth spectrum as depicted, it must be averaged over several realisations, not just the single one shown in Figure 2.3. This is discussed in more detail in Chapter 3.

As mentioned above, the properties of a random signal can only be described in terms of their statistical properties, such as mean value (first order), mean square value (second order) etc. The equation for the mean value is:

$$x_m(t) = E[x(t)] \quad (2.1)$$

and for the mean square value is:

$$x_{ms}(t) = E[x^2(t)] \quad (2.2)$$

To obtain a parameter with the same dimensions and units as the original signal, it is normal to take the square root of the mean square value to obtain the ‘root mean square’ or rms value, thus:

$$x_{rms}(t) = \sqrt{E[x^2(t)]} \quad (2.3)$$

An important second order statistic is the so-called ‘autocorrelation function’, which gives an indication of how well a signal correlates with a displaced version of itself. For a periodic signal, it is obvious that the correlation will be perfect every time it is displaced by an integer number of periods, and so the autocorrelation function is also periodic with the same period. The most fundamental definition of autocorrelation is:

$$R_{xx}(t, \tau) = E[x(t - \tau/2)x(t + \tau/2)] \quad (2.4)$$

which is evaluated at time t over an ensemble as in Figure 2.2. Since by definition for a stationary random signal the result is independent of time t , it is often written as a function of time displacement τ only as:

$$R_{xx}(\tau) = E[x(t - \tau/2)x(t + \tau/2)] \quad (2.5)$$

or equivalently:

$$R_{xx}(\tau) = E[x(t)x(t + \tau)] \quad (2.6)$$

As mentioned above, for ergodic stationary signals the averaging can be performed along a single record, so that the equivalents of Eqs. (2.5) and (2.6) are:

$$R_{xx}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t - \tau/2)x(t + \tau/2)dt \quad (2.7)$$

and

$$R_{xx}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t + \tau)dt \quad (2.8)$$

respectively, with Eq. (2.8) being the most commonly used.

2.1.3 Cyclostationary Signals

The fourth signal (cyclostationary) is an amplitude modulated white noise. It is shown in Chapter 3 that amplitude modulation of a signal (by a single frequency) results in pairs of sidebands in the spectrum, spaced around each modulated frequency component by an amount equal to the

modulation frequency. Since the spectrum of the white noise is uniform, the spectrum of the modulated signal is also uniform, even with the added sidebands, but there is a hidden structure in the spectrum which can be discovered by correlating the spectrum with itself (analogous to Eq. (2.4)), which gives non-zero correlation for (and only for) displacements equal to discrete multiples of the sideband spacing, the so-called cyclic frequency. This is treated in depth in Section 3.6.

2.1.4 Cyclo-non-stationary Signals

A typical cyclo-non-stationary signal is not depicted in Figure 2.3, but could for example be similar to the cyclostationary signal there except that the period of the modulating signal would not be constant but varying with time in a deterministic way. It would not be possible to see any difference in the spectrum, but information about the deterministic modulation can be extracted by the methods of cyclo-non-stationary analysis in Section 3.6.4.

2.2 Signals Generated by Rotating Machines

In condition monitoring, changes in vibration signals are ascribed to changes in condition, so it is important that other factors which cause changes in vibration signals are considerably reduced or eliminated. Vibrations tend to change with the speed and load of a machine, so this section primarily considers the signals generated by a rotating machine operating at constant speed and load, for which the signals will typically be stationary and/or cyclostationary. Occasionally, use can be made of non-stationary signals, such as those generated by a machine under run-up or coast-down conditions, but such signals will be discussed in conjunction with the appropriate analysis techniques, such as the time/frequency techniques treated in Section 3.5, or the cyclo-non-stationary techniques of Section 3.6.4.

2.2.1 Low Shaft Orders and Subharmonics

A number of faults manifest themselves at a frequency corresponding to the speed of the shaft in question, or its low harmonics and subharmonics. The following description, in particular of whirl, largely follows the insightful commentary of one of the pioneers of machine diagnostics, John Sohre, in his classic paper ‘Operating Problems with High Speed Turbomachinery, Causes and Corrections’ [1].

2.2.1.1 Unbalance, Misalignment, Bent Shaft

All three of these faults manifest themselves at shaft speed, and possibly the first few harmonics (multiples) of shaft speed.

2.2.1.1.1 Unbalance

Unbalance (imbalance) causes excitation by forces rotating at the shaft speed when the local centre of mass (CoM) of the cross section is not at the centre of rotation. The response depends on whether the inertias on the shaft are localised or distributed axially, and whether the shaft is running below or above its first critical speed. If the shaft is short and the inertia localised, there will basically be a

radial force rotating at shaft speed, which excites vibrations primarily in the two radial directions, but very little axially. Where the radial stiffness of the shaft and bearing supports is high, the response vibration will be stiffness controlled and so in terms of displacement it will be proportional to the unbalance force Mew^2 , where M is the mass of the rotor, e is the radial displacement of the CoM of the rotor from the centre of rotation, and ω is the rotational speed of the shaft. The shaft is most often axisymmetric, but the bearing supports usually have different stiffness in the horizontal and vertical directions, so that the vibration response will be different in the two directions. Even in this simplest of cases, the stiffness of the bearings is usually nonlinear, in particular for fluid film bearings, but even for rolling element bearings, so even though the unbalance force is only at shaft speed, the response will be distorted to some extent from sinusoidal, and so the spectrum of the response will contain harmonics of the shaft speed.

Where the inertia of the rotor is distributed axially, the CoM of each section is not necessarily the same, and thus the radial unbalance force changes in amplitude and direction along the rotor. If the rotor is rigid, all the unbalance forces can be combined into an equivalent single unbalance force at the global CoM of the rotor, and a moment about some axis through the CoM. Thus, the overall response is a combination of radial (cylindrical) motion, and rocking (conical) motion, once again with circles distorted to ellipses if the horizontal and vertical support stiffnesses are different, and generating higher harmonics if the bearing stiffnesses are nonlinear. Such rocking motions can give axial responses, even if the elemental unbalance forces are purely radial.

Rotor motion becomes even more complicated when the shaft is flexible and operating above its first critical speed. This applies to many turbomachines, such as turbogenerators and turbine-compressor units in the (petro-)chemical industry.

The simplest case is the so-called ‘Jeffcott’ or ‘Laval’ rotor, which has a single concentrated disc symmetrically located in the middle of the flexible shaft (modelled as a set of simple springs), but its behaviour throws some light on the more general case. Below the critical speed (which corresponds to the transverse bending natural frequency of the rotor as a beam) the motion is stiffness controlled as mentioned above (although the deflection has a feedback effect increasing the centrifugal unbalance force as speed increases). In principle the response at the critical speed would be infinite, but this takes an infinite time to build up, so if the speed is increased fast enough, it is possible to get through the critical speed. The vibration is now mass controlled (isolated from the supports), and there is a tendency for the disc to rotate about its CoM, meaning that the vibration tends to a limit given by the original eccentricity e . The motion of this simple rotor would be a ‘synchronous whirl’, i.e. it takes up a deformed shape which whirls around at shaft speed. In theory there is a ‘backward whirl mode’, where the centre of the rotor rotates at the critical frequency in the direction opposite to the rotation, but this is less likely to get excited. It could however be excited by ‘rubbing’ of the rotor on the stator.

Actual rotors are even further complicated by the following factors:

1. There are several discs on the rotor, in general with no symmetry, meaning that there are many critical speeds corresponding approximately to each of the mode shapes that the shaft would take up as a beam in bending.
2. Because this means that the individual discs tilt, gyroscopic effects come into play which make the natural frequencies vary with speed. The critical speeds become different for forward and backward whirl.
3. Many high speed rotors, in particular those which operate for long periods without shutting down, are supported in fluid film (hydrodynamic or journal) bearings, since in principle these are not subject to wear, except during startup and shutdown. The hydrodynamic bearings are very nonlinear, with their stiffness and damping properties varying with speed, radial load, and viscosity

(temperature) of the lubricant, meaning that the position of the centre of the shaft in the bearing varies with all these factors. The hydrodynamic bearings thus have considerable influence on the critical speeds and mode shapes.

For more details on these topics, the reader is directed to specialist books on rotor dynamics, such as [2–5].

To summarise, unbalance gives vibration responses at shaft speed and its low harmonics, and mainly in the radial direction, although rocking motions due to moment unbalance can give axial vibrations.

2.2.1.1.2 Misalignment

When a shaft has three or more bearings, for example when two machines are coupled together, there is a potential for misalignment, which can be parallel misalignment, meaning that one of the two shafts is displaced laterally, but still parallel to the other, or angular misalignment, where the axis of one is at an angle to that of the other. Such misalignment introduces into the shafts bending deflections, which are fixed spatially, but rotating with respect to the shafts. The induced bending moments thus depend on the bending stiffness of the shaft, and have to be counteracted by forces at the bearings and the foundations. Where the shaft stiffness varies with rotation angle, as caused for example by a keyway, the stiffness will typically vary twice per revolution and so the fixed displacements will give fluctuating moments and forces, and vibrations, varying at this rate.

Normally, flexible couplings are used to mitigate the effects of misalignment, and these can introduce vibrations according to their properties. Ref. [6] studies a particular type of coupling, a so-called disc coupling. More importantly, it gives an overview of the properties of other couplings, including the classic universal joint (Hooke's joint or Cardan joint), and the gear couplings much used for high speed turbomachines, as discussed by Bloch in [7]. The latter two types tend to give a response at twice shaft speed for the following reasons:

1. A universal joint with misalignment angle α (in radians) induces a torsional vibration into the driven shaft, with relative instantaneous angular velocity (with respect to the input speed), given for (input) shaft rotation angle θ_i by:

$$\omega_2/\omega_1 = \frac{\cos \alpha}{1 - \sin^2 \alpha \sin^2 \theta_i} \quad (2.9)$$

which will be seen to go through two periods of oscillation for each rotation of the shaft. For misalignment angles up to about 10° , this is dominated by the second harmonic of shaft speed, and can be approximated very accurately as:

$$\omega_2/\omega_1 = 1 - (\alpha^2/2) \cos 2\theta_i \quad (2.10)$$

For larger misalignment angles, the fourth and other even harmonics have more prominence.

The torsional vibration will introduce torques into the shafts, and these result in bending moments and consequential forces and vibrations at the bearings by two mechanisms. A pure torque T in the input shaft will give torque component $T \cos \alpha$ and bending moment $T \sin \alpha$ on the driven side (to see this consider the situation with a 90° bend at the coupling), and the bending moment must be resisted by lateral forces. Secondly, where there are gears in the system, with different input and output speeds, the unbalance between input and output torques must be accommodated by the bearings and foundations, and any torque fluctuations will give fluctuating forces and vibrations.

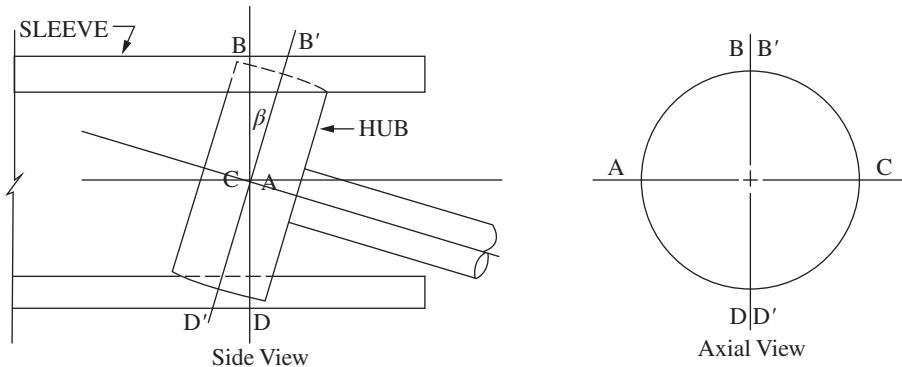


Figure 2.4 Operation of a gear coupling. Source: From [6].

2. A gear coupling, such as discussed in [6, 7] has gears on the two shaft ends to be coupled, which mate with an internal gear in a sleeve connecting them. It is very stiff in torsion, but allows for misalignment and relative axial motions of the shafts. Figure 2.4 (from [6]) illustrates one gear (hub) interacting with the sleeve for one rotation of the sleeve from A to B to C to D and back to A. There are several mechanisms producing interactions between torques and bending moments. There is the $T \sin \alpha$ bending moment as for a Hooke's joint, and another bending moment related to the fact that the torque is applied as tangential forces through teeth that are offset axially as shown in Figure 2.4. There are axial friction forces as the teeth slide against each other, and Bloch points out that for an individual tooth these give a moment varying twice per revolution. If all teeth had identical friction forces, these would cancel out, but in practice there are variations and therefore a residual moment varying twice per rev.

To summarise, misalignment tends to give vibrations at the low harmonics of shaft speed, with some coupling types favouring the even harmonics of shaft speed, in particular the second (the conventional way of referring to this is 2X).

It is often said that misalignment gives axial motion, but this is due to rocking motions associated with moments, and therefore in principle not very different from the rocking motions that can arise from couple unbalance.

Figure 2.5 illustrates the problem in distinguishing between unbalance and misalignment purely on the basis of the component at twice shaft speed. The measurements were made by the author on a gearbox between a gas turbine running at 85 Hz (5100 rpm) and a synchronous generator at 50 Hz (3000 rpm). Note that the higher noise levels at low frequency in the measurements 'Before repair' are because that signal was recorded as acceleration on an FM tape recorder with limited dynamic range, and integrated to velocity before analysis, whereas the signal 'After repair' was recorded directly as velocity.

The reduction in vibration levels at the two shaft speeds results entirely from an improvement in alignment, but the strongest components in terms of velocity are at the shaft speed for both shafts. The vibration levels of the two shafts before and after repair are given in Table 2.1.

As a result of the realignment, the vibration level at the first harmonic of the generator speed improved by 14 dB, and at the first harmonic of the turbine speed by 8 dB. There is an indication that misalignment was the cause of the initial high vibration levels, in that the second harmonic of the turbine speed has improved by 12 dB (compared with 8 dB for the first harmonic), but on the other

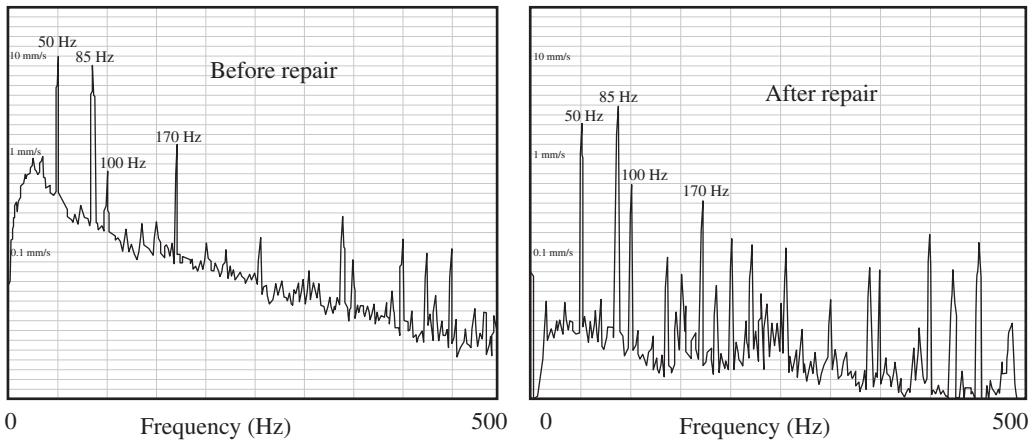


Figure 2.5 Example of the improvement given by realignment.

Table 2.1 Vibration reduction by realignment.

Frequency (Hz)	50 (1X)	100 (2X)	85 (1X)	170 (2X)
Level before repair (mm s^{-1})	10	0.5	9	1.4
Level after repair (mm s^{-1})	2	0.5	3.6	0.35
Improvement (dB)	14	0	8	12

hand higher unbalance could make fluid film bearings more nonlinear and thus increase the higher harmonics, so it is not a definite indication.

It is interesting that the second harmonic of the generator speed has not changed as a result of the realignment, but the explanation is probably to be found in the fact that 100 Hz is also the second harmonic of mains (line) frequency, and thus very likely dominated by electrical effects (see later under ‘Electrical Machines’).

2.2.1.1.3 Bent Shaft

If a shaft that is initially balanced acquires a permanent bow for some reason, the results will be a combination of the effects of unbalance and misalignment. If the reason for the bow is related to the unbalance, as in some cases of ‘thermal bow’, where ‘rubbing’ occurs between rotor and stator, causing a local hot spot and thermal expansion, there may be specific symptoms which help in the diagnosis, and a number of these are discussed in [3, 8].

2.2.1.2 Cracked Shaft

Development of a crack in a shaft is one of the most serious faults to be detected in condition monitoring, in particular of large rotors, such as in turbogenerators, and so it has been studied in depth. An immediate problem is that even a large crack has only a small effect on the natural frequencies of a shaft and almost none when the crack is closed. Cases have been recorded where a transverse crack in a rotor had progressed through 25% of the diameter and only changed the critical speed by 2.6% ([9]).

Three classic papers on the effects of transverse cracks in shafts were presented at the initial IMechE Conference on Vibrations in Rotating Machinery in Cambridge in 1976 [10–12]. They laid the foundations for most of the subsequent work. A short summary is that a crack which is permanently open increases vibration primarily at the first and second harmonics of shaft speed, whereas a ‘breathing’ crack, which opens and closes each revolution (e.g. because of sag due to gravity) also gives an increase at the third harmonic, which does give a better chance of distinguishing it from unbalance and misalignment. The effects of cracks are usually oriented differently than those of unbalance, and therefore the growth will often change both the amplitude and phase of the vibration at the low harmonics of shaft speed, so monitoring both is recommended. Sometimes amplitude will first reduce because of opposing phases, and this may be the first sign of a growing crack. Figure 2.6 from [13] illustrates this.

It is possible for a crack to be held closed during operation at normal speed (because of synchronous whirl with a deformation shape such that the crack is in compression) and then the symptoms are not apparent. A developing crack will often be made apparent by greatly increased response when passing through the critical speed during run-up or coast-down, even if the actual frequency of the critical speed is not greatly changed. Figure 2.7 shows an example from [14].

However, many turbomachines are expected to run for long periods without shutdown, so a method that could detect cracks while the machine is in operation is highly desirable. A case is described in [15] where a steam turbine in a nuclear power plant was subjected to regular ‘washing’ of the blades to clear depositions. This was done after unloading, but at full speed, and in the presence of a crack gave greatly increased 1X and 2X vibrations for several hours after the washing, which then reduced to normal. It was realised that this was explained by the fact that the washing cooled the rotor surface, causing the crack to open, but that it closed again after the temperature became uniform after several hours.

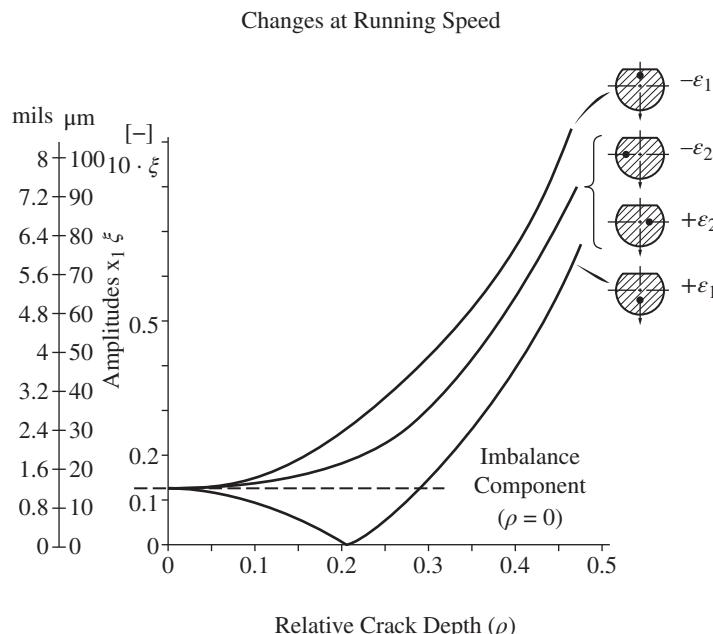


Figure 2.6 Effect of unbalance location on cracked rotor response (Ref. [13]).

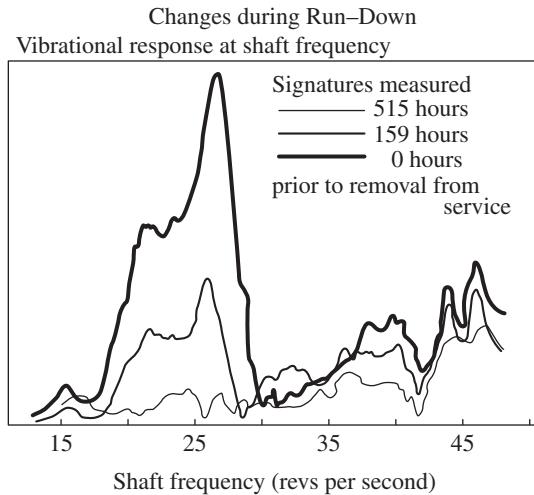


Figure 2.7 Increasing response at critical speed caused by growth of a crack (Ref. [14]).

The best way to obtain warning of crack development in a shaft, and distinguish this case from unbalance and misalignment, is to make a simulation model of the rotor and its supports, as described in the next section. An excellent summary of cracked rotors and their simulation is to be found in [16].

2.2.1.3 Rotor/Bearing/Foundation Models

From the discussion in Sections 2.2.1.1 and 2.2.1.2 it can be seen that unbalance, misalignment and cracked shaft can all give rise to changes at the shaft speed and its low harmonics, and simple rules cannot be given to distinguish between them. There is a tendency for unbalance to give a response primarily at 1X in the radial direction, and for misalignment to give more response at 2X and in the axial direction, but exceptions abound. A breathing crack will probably give greater changes at 3X, but this is not unique.

For the most critical machines it is becoming more common to develop mathematical models of the rotor/bearing/foundation system so as to be able to better predict the changes given by various types and levels of faults. Examples of such simulation models are given in [17–19]. Another good source of information about such simulation models is to be found in the Proceedings of two series of conferences on rotor dynamics, the IFTOMM International Conferences on Rotor Dynamics, and the IMechE conferences on Vibrations in Rotating Machinery, both of which are held every four years, with two years offset between them.

Finite element models of rotors have now advanced to the point where they are very accurate, with little need for updating, but the rotor support system is more difficult to model, and usually requires some form of updating based on measurements. This is particularly the case with rotors supported in journal bearings, since the alignment of the shaft is not defined by the alignment of the bearing centres (which can be measured externally). The properties of journal bearings are well described by Reynolds' equation, but this in principle requires a knowledge of oil viscosity (and thus temperature) at every point, and this is not always possible to measure or predict accurately. It also requires a knowledge of the extent of intact oil film, this often being broken by a zone of cavitation.

However, it is possible to greatly improve the accuracy of the rotor support models by making measurements of absolute and relative shaft/bearing vibration over a range of speeds, e.g. Ref. [20]. The procedure for determining the shaft configuration (alignment) is given in more detail in [21]. This is one situation where measurements with both proximity transducers and accelerometers are necessary. The properties of the fluid film bearings are related directly to the relative positions of the shaft and bearing, as measured by proximity probes (in addition to the vibrations around the mean positions).

For condition monitoring purposes, the simulation model does not have to predict the absolute values of measurement parameters completely accurately, as long as changes due to typical faults are well predicted.

2.2.1.4 Subharmonic and Non-synchronous Whirl

A number of phenomena cause the centre of the shaft to whirl (either forwards or backwards) at a frequency different from the rotation speed.

2.2.1.4.1 Oil Whirl and Oil Whip

Oil whirl is a phenomenon that can occur under certain conditions in fluid film bearings, in particular when they are lightly loaded. It is characterised by a forward whirl at a frequency just less than half the shaft speed (typically 42–48%). It is explained by Sohre [1] as the shaft ‘surfing’ on a wave running around in the bearing clearance. Because the oil adjacent to the shaft travels at shaft speed and that at the bearing surface has zero velocity, the mean velocity of the lubricant is approximately half the shaft speed, but slightly less in the critical pressure zone (supporting the bearing load) because the pressure gradient causes a backward flow. Instability can occur when the system is able to extract more energy from the rotation than is dissipated by damping mechanisms.

Oil whirl that could become destructive should be eliminated at the design stage for normal operation. One countermeasure is to use non-cylindrical bearings, such as ‘lemon bearings’, where the joint faces are machined off so that the vertical diameter is smaller than the horizontal. The most effective measure is often to use tilting pad bearings, with a number of individual bearing sectors that can tilt to adjust the pressure distribution. For condition monitoring, it can usually be assumed that the bearings have been designed to eliminate oil whirl under normal operating conditions, but it can still arise because of changes to these operating conditions, such as changes in alignment due to settling of the foundations, thermal anomalies etc.

Oil whip is the term often given to the coincidence of oil whirl frequency and the critical speed of the shaft, which sometimes results from the fact that many industrial turbomachines run at approximately twice their first critical speed. It would normally be fixed by eliminating the oil whirl.

2.2.1.4.2 Hysteresis Whirl

Hysteresis whirl is characterised by the whirl frequency being at the shaft’s critical speed independent of the actual shaft speed. It is due to friction forces generated by relative movement of rotor components, giving a consumption of energy in a hysteresis loop for every cycle of the difference frequency between the shaft speed and natural frequency. It is often initiated on passing through the critical speed, and remains at the natural frequency as the shaft speed increases up to running speed. It results from the fact that the friction forces have a component that is tangential to the whirl motion. Professor Stephen Crandall of MIT has given a simple explanation using the analogy of a conical pendulum, which in its simplest form is a compound pendulum suspended in a ball joint, and

whirling in a conical motion at the same frequency as it would have in planar motion. In the absence of friction it would whirl with constant amplitude. If a spinning bowl of viscous liquid (soup) were raised so as to engulf the lower end of the pendulum, the whirl would continue at approximately the same frequency (for light damping), but its change in amplitude would depend on whether the spin speed were lower or higher than the natural whirl frequency. If the spin speed were lower, the tangential friction forces would tend to reduce the whirl amplitude, whereas if it were higher they would tend to increase it. In the same way, friction forces in a rotor can lead to increased whirl or instability at shaft speeds above the critical speed (extracting energy from the shaft rotation to more than make up for the friction losses).

2.2.1.4.3 Exact Subharmonic Whirl

Some conditions give rise to exact subharmonic whirl (i.e. at exactly $\frac{1}{2}$ or $\frac{1}{3}$ of shaft speed). This is typical of ‘parametric excitation’, where the parameters of the governing differential equations vary periodically with time. An example of such an equation is Hill’s equation, where for example the stiffness of a mass/spring/damper system varies periodically, or its simplest form, the Mathieu equation, where it varies sinusoidally. Thus, the equation of the (single degree-of-freedom or SDOF) system is:

$$\frac{d^2q}{dt^2} + \omega_n^2[1 + f(t)]q = 0 \quad (2.11)$$

where the ‘pumping’ term $f(t)$ involves periodic variations in either the natural frequency or damping of the system (<http://www.answers.com/topic/parametric-oscillator>). This term can be moved to the right hand side of the equation as a forcing function or parametric excitation, and gives a forced response at this frequency. Since the driving term is the product of the response q and the pumping term $f(t)$, the excited frequencies will be the sum and difference frequencies of the two terms (see Chapter 3 for an explanation). If the frequency ω_1 of q is approximately (but not exactly) ω_n , and the frequency of $f(t)$ double this or $2\omega_1$, the response frequencies will be at ω_1 and $3\omega_1$. The higher frequency will not elicit much response, but the lower one, being near ω_n will be amplified. Thus, if the excitation frequency $2\omega_1$ is the rotation speed of a shaft, because the stiffness varies once per revolution, the exact half order subharmonic ω_1 will be amplified, and lock onto exactly half the shaft speed rather than the natural frequency.

The reason for periodic variation in stiffness can be a loss of contact stiffness once per rev because of looseness, or conversely an increase in contact stiffness due to ‘rubbing’. Loose assembly of bearings has been known to give such an excitation of exact half order components (e.g. [4]) and Figure 2.8 shows an example measured by the author in a chemical plant before and after a maintenance shutdown of a compressor.

2.2.1.4.4 Dry Friction Whirl and Whip

Dry friction whirl is a phenomenon that can occur through contact between the rotor and stator (bearings, seals, or casing). It is described in detail by Childs in Ref. [2] with additional possibilities discussed in [3], both influenced by the pioneering analysis by Black [22]. Where the contact occurs without slip, a backward precessing whirl is excited at a fixed proportion of the shaft speed determined by geometric factors. This occurs up to a limiting frequency defined by ‘the natural frequency of the coupled rotor-stator system’ (Childs [2]). For shaft speeds above this limiting frequency the phenomenon is called ‘dry friction whip’. Childs states that ‘Dry friction *whirl* and *whip* are only likely to occur for contact of a small-diameter shaft at a (very) large clearance’.

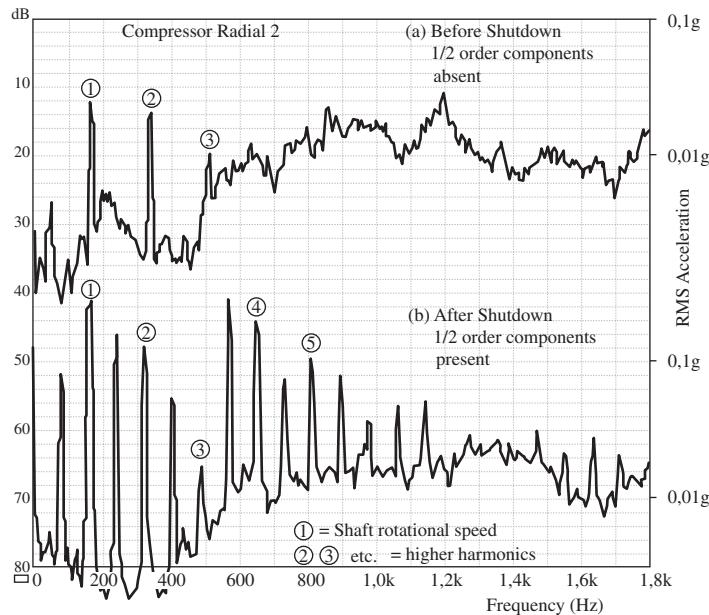


Figure 2.8 Generation of exact half order components (and harmonics) due to loose assembly of a journal bearing in a centrifugal compressor. Source: Courtesy Brüel & Kjær.

2.2.2 Vibrations from Gears

Gears are very widely used in machines to transmit power from one shaft to another, usually with a change in speed and torque. The majority of gears have conjugate profiles such that kinematically there will be a constant output speed for a constant input speed [23]. The most common gear tooth profile is involute, as the speed ratio is then insensitive to small variations in centre distance, although the ‘pressure angle’ does change. The pressure angle is the angle between the direction of the normal force between the mating teeth and the common tangent to the pitch circles of each gear as illustrated in Figure 2.9. For spur gears, the ‘line of action’ traces out the path of the point (actually an axial line) of contact between each meshing tooth pair as they move between entry into and exit from meshing. As also shown in Figure 2.9, the line of action is tangent to the ‘base circles’ of the two gears, these defining the base of the involute curves of the tooth profiles. For helical gears, the lines of contact still fall in the plane tangent to the base cylinders, but are oblique rather than axial.

In practice, the situation is not so ideal, as the teeth deform under load, introducing a ‘meshing error’ or ‘transmission error’ (TE), even when the tooth profiles are perfect. In addition there are geometric deviations from the ideal profiles, both intentional and unintentional. The intentional deviations are typically due to ‘tip relief’, where metal is removed from the tip of each tooth with a maximum at the tip, and gradually reducing to zero at some distance down the tooth, but before the pitch circle. This allows each tooth to come into mesh without impact, which otherwise would occur because the adjacent teeth supporting the load are deflected from their ideal positions. For a given load, TE, vibration and noise of a gearset can be minimised by using the ideal amount of tip relief, but this can of course only apply for a particular load.

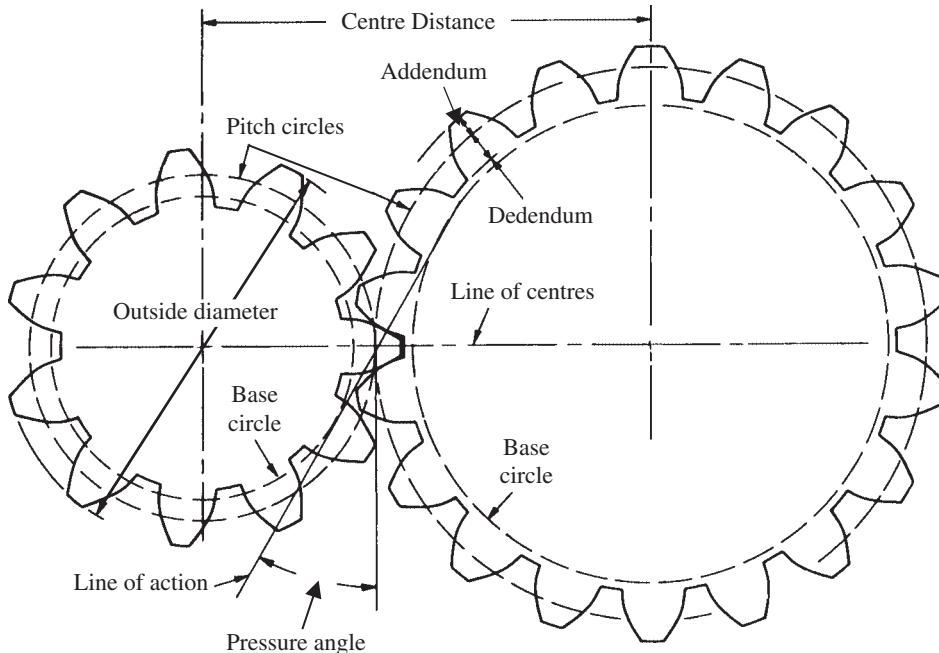


Figure 2.9 Basic dimensions of a pair of spur gears.

Overall transmission error thus has a load dependent component due to tooth deflection, and a non load dependent component due to geometric deviations from the ideal profile. Three different types of transmission error can be distinguished; unloaded static transmission error, loaded static transmission error, and dynamic transmission error. Unloaded static TE is what is measured for example when a test gear is meshed with an ideal master gear under a very light load, sufficient only to keep the teeth in contact. It would also be measured by gear profile measuring machines and coordinate measuring machines. It is dominated by high points in the profile, and thus not very representative of TE under load. Loaded static TE includes the tooth deflection due to a constant load torque, and can be measured under slow roll conditions. Local Hertzian deformation will tend to even out high spots and reduce their dominance. Because of the local high stresses, such high spots will often wear away during running-in. Gears are often lapped, i.e. run together with a grinding compound, as a final finishing stage in manufacture so as to aid this process. Dynamic TE is the actual TE in operation, where dynamic effects cause fluctuations in the torque transmitted by the gearset, varying for different frequencies. It is described in Section 7.2.2 how this may be useful as a condition monitoring parameter.

Because the TE varies with tooth deflection, which in turn varies with load, the amplitude of the resulting vibration at the toothmeshing frequency varies directly with the load fluctuations in service, and can be considered as an amplitude modulation effect. The fact that the vibration amplitude varies with the mean load also means that vibration measurements should only be compared for condition monitoring purposes for the same load each time. Sometimes the only fixed load that can be relied upon is zero load, but in general this is not a good choice for monitoring purposes, because the teeth can lose contact and give rise to chaotic vibrations which are not very repeatable, and which do not necessarily respond to faults in the gears.

Mark made a classic analysis of gear vibrations and their relationship to static TE in the late 1970s [24, 25]. For each gear, the vibrations were divided up into the mean component over all teeth, thus repeating at the toothmesh frequency (gear rotational speed times the number of teeth on the gear) and what he called the ‘random’ component, this being the deviations from the mean for each tooth. Such deviations are of course pseudo-random, because they repeat each revolution of the gear. Mark found it convenient to decompose the overall deviation for each tooth into a set of Legendre polynomials of orders from zero and up, as illustrated in Figure 2.10. Separate polynomials are required to describe the deviations along the profile ($l = 1, 2, \dots$) and axially ($k = 1, 2, \dots$). Both k and $l = 0$ corresponds to a parallel displacement, i.e. tooth spacing error, while k and $l = 1$ correspond to linear deviations in the respective directions, and k and $l = 2$ to quadratic deviations, and so on. Since even the toothmeshing frequency is already quite high in general, the low order polynomials up to order 3 or 4 would normally describe adequately the errors that manifest themselves within the measured frequency range, and which are not attenuated by the ‘mesh transfer functions’ described below.

Since the error component corresponding to each value of k or l for each tooth is a single number, the series for the whole gear can be considered as a set of digital values sampled at the toothmesh frequency. Mark uses this fact to explain the typical form of the spectra of gear vibrations as corresponding to a ‘discrete Fourier transform’ or DFT, which repeats periodically in frequency at intervals equal to the sampling frequency, in this case the toothmesh frequency. To understand this it is necessary to use some of the Fourier theory developed in Chapter 3, where it is shown that the DFT is periodic in both time and frequency domains.

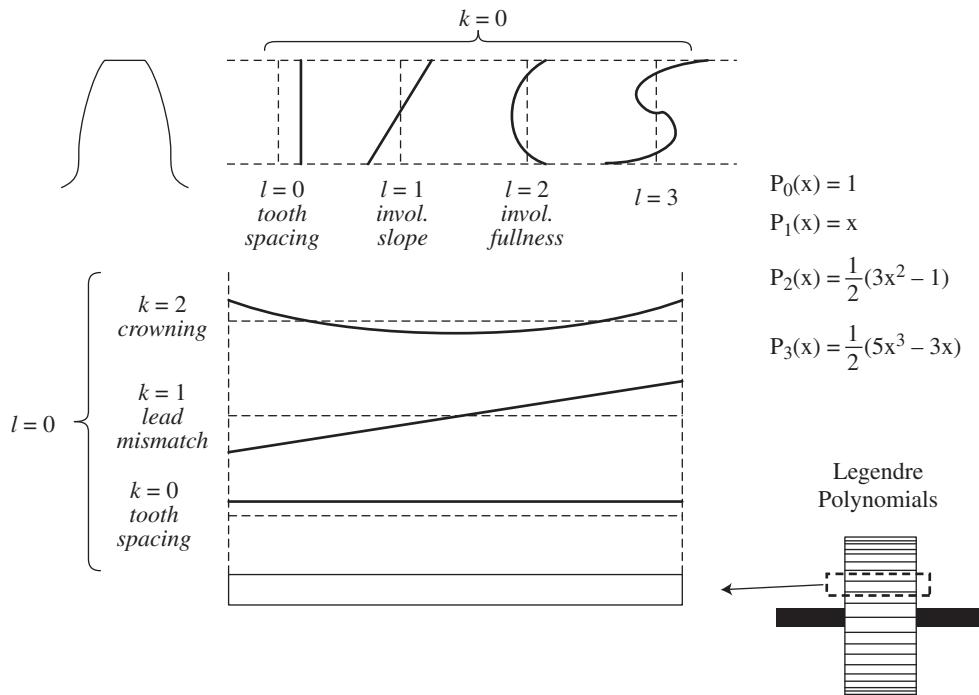


Figure 2.10 Illustration of Mark’s division of TE into elemental functions in terms of Legendre polynomials, both axially, and along the tooth profile.

Actual gear vibration spectra do not repeat periodically, and Mark explains this as due to the abovementioned ‘mesh transfer functions’ with which the original periodic spectrum is multiplied. The mesh transfer function is related to the smoothing and lowpass filtration effects given by both local and global tooth deflections as the load is distributed across each tooth flank and between the teeth in mesh.

Figure 2.11a shows a typical ‘DFT’ spectrum for the tooth spacing error on a particular gear, and Figure 2.11b the corresponding ‘mesh transfer function’.

A simple example of the action of a mesh transfer function is given by the effect of tooth contact ratio, which can be used to reduce gear vibration. Contact ratio (CR) can be understood as the average number of teeth in contact throughout a meshing cycle. For spur gears it is typically about 1.5, meaning that two tooth pairs share the load for half the time, and only one pair for the other half. The stiffness of a single tooth pair is approximately constant during the mesh cycle [23], so for

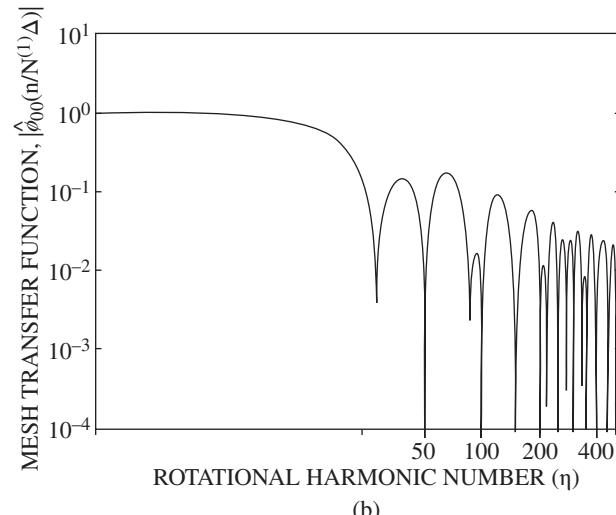
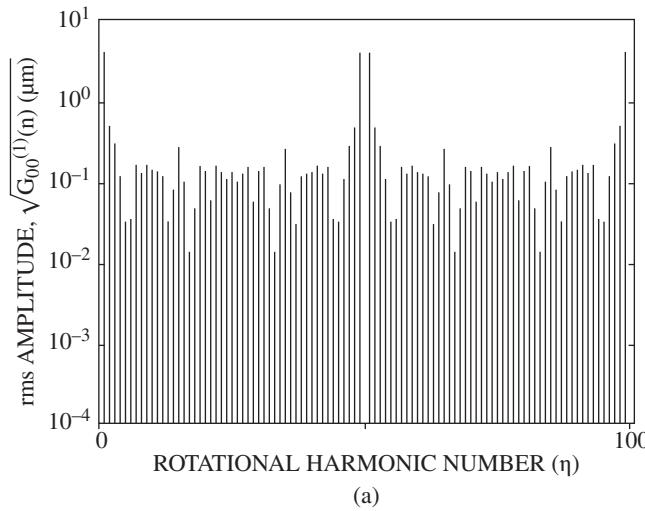


Figure 2.11 Typical individual error components for tooth spacing error [25] (a) Error spectrum (b) Corresponding mesh transfer function Reprinted by permission of the Acoustical Society of America.

constant load the deflections tend to double in passing from double tooth pair to single tooth pair contact. This gives a strong parametric excitation at the toothmesh frequency. If the CR can be made an integer, say 2, then there are always two pairs in contact, and in principle no excitation at the garmesh frequency. Mark's 'mesh transfer function' is related to CR, and for this simple case is a lowpass filter with zeros at multiples of the toothmesh frequency divided by the CR. Thus, if the CR is an integer, the zeros will coincide with all harmonics of the garmesh frequency and eliminate this excitation.

For helical gears, there is also an axial contact ratio (or overlap ratio) because of the number of teeth in contact in the axial direction along the gear. This gives an additional set of 'mesh transfer functions', with additional zeros in the spectrum, and a double lowpass filtering effect, so that in general helical gears generate less vibration, in particular at higher harmonics of the garmesh frequency.

An alternative way of describing the sidebands found in gear vibration spectra around the harmonics of toothmesh frequency is as modulation sidebands caused partly by the amplitude modulation mentioned above, but also by frequency modulation [26]. The author (and others) initially made an error in interpreting the frequency modulation as directly due to the changes in rotating speed of the gears, but the most significant part of it comes from the modulation in space of the tooth contact point, so that the frequency modulation would be present even if the gears were connected to infinite inertias so that their speed remained constant. This is discussed in more depth in Section 7.2, because it depends on some of the theory of Chapter 3.

The vibrations generated by gears in both healthy and deteriorated condition can be classified as follows:

2.2.2.1 Mean Effects for All Tooth Pairs

These manifest themselves at the toothmesh frequency and its harmonics and can be subdivided into:

- Tooth deflection due to the mean torque.
- The mean part of initial profile errors resulting from manufacture, including intentional profile modifications.
- Uniform wear over all teeth.

Condition monitoring attempts to distinguish the third effect from the first two. As depicted in Figure 2.12, uniform wear tends to give a double-scalloped wear pattern on each tooth, because there

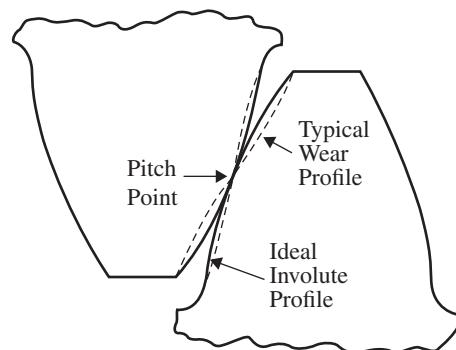


Figure 2.12 Typical double-scalloped wear pattern from sliding on either side of the pitch circles. Source: Courtesy Brüel & Kjær.

is a pure rolling action for contact at the pitch circles and sliding for contact on either side. Thus, the initial indication of wear will usually be an increase in the second harmonic of the toothmesh frequency, since the effect at the first harmonic must become greater than that due to tooth deflection to become apparent. As wear proceeds, the profile will deteriorate more generally and all harmonics of toothmesh frequency will increase.

In some cases of high load and poor lubrication, the lack of sliding at the pitch circles can lead to local breakdown of the lubricant film and consequent ‘pitchline pitting’ [27]. Because of its impulsiveness, this would give an increase in all high harmonics of toothmesh frequency.

2.2.2.2 Variations from the Mean

Since these repeat for each rotation of each gear, they manifest themselves at the harmonics of each gear’s rotational speed, including sidebands around the common garmesh frequency. The spacing of these harmonics and sidebands identifies the gear which has produced them (equal to the rotational speed).

The variations can be further subdivided as follows:

- Slow variations, e.g. runout, distortion, non-uniform wear. Low harmonics and sidebands around toothmesh are affected.
- Local faults, e.g. tooth root cracks, spalls. A wide distribution of harmonics and sidebands results.
- Random errors, e.g. random tooth spacing error. A wide distribution of harmonics and sidebands results.
- Systematic errors, e.g. ‘Ghost components’, from the gear cutting machine (see below).

Figure 2.13 illustrates the differences given by distributed and local effects in the frequency domain. It should be kept in mind that what is illustrated is the effect close to the source, and that the actual measured response spectra will be affected by the transfer functions from the source to the

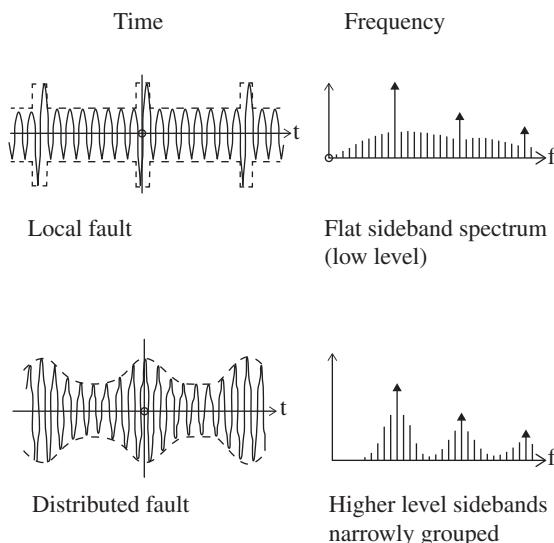


Figure 2.13 Comparison of the effects of local and distributed faults in gears in the time and frequency domains. Source: Courtesy Brüel & Kjær.

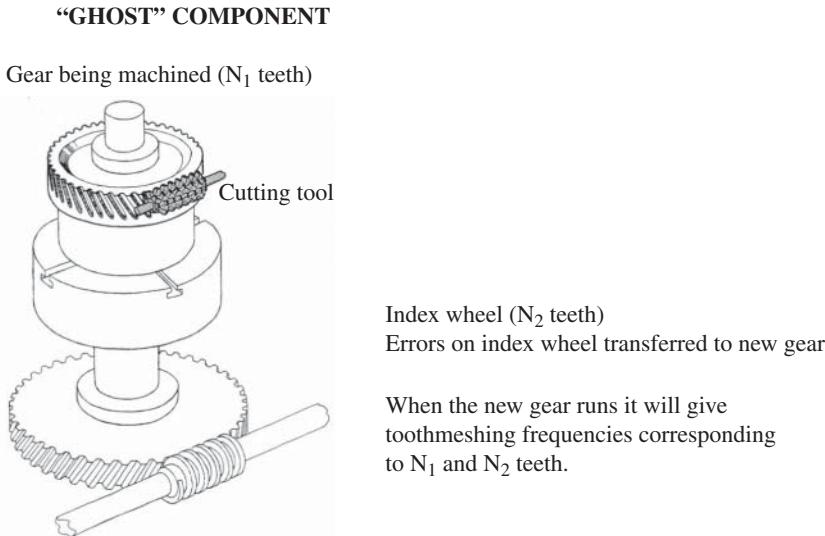


Figure 2.14 Illustration of the generation of ghost components. Source: Courtesy Brüel & Kjær.

measurement points. However, the spectral effects in Figure 2.13 can be interpreted as the *changes* in the frequency spectra.

‘Ghost’ components are systematic errors resulting from the manufacturing process. The gear being cut (hobbed, shaped, or ground) is turned by another gear (index wheel) as shown in Figure 2.14. This means that any errors in this gear are transferred to all new gears being manufactured, and can produce vibrations corresponding to a different number of teeth (the number on the index wheel) when the gear runs in service.

Ghost components are sometimes found in the vibrations of high quality gears, because they define the limit of precision to which the gears can be manufactured. They can even dominate the vibrations, in particular at low load. As with all geometric errors, the vibrations produced are not very sensitive to load, and this can help to identify them. Figure 2.15 shows an example from the gearbox of a gas turbine driven ship, at <10% load and at full load.

The ghost component dominates even at full load, but it has increased by only 6 dB in going from <10% load to full load, whereas the toothmesh component has increased by 20 dB. The second harmonic of the ghost component stayed about the same, while the second harmonic of toothmesh increased by 7 dB at full load.

Ghost components can be used diagnostically, because if anything they are most likely to reduce as a result of wear. Figure 2.16 shows an example where an increase in misalignment over a month has caused two changes, both indicative of wear.

1. A 6 dB increase at the toothmesh frequency.
2. A similar decrease of a ghost component.

Ghost components must appear at a harmonic of the rotational speed of the gear in question, since they have to correspond to an integer number of teeth. It is rarely possible to trace them back to the machine on which the gear was manufactured, but they can often be recognised by the characteristics described in this section. In general, they do not pose a problem in condition monitoring, as they are most likely to reduce as a result of lapping, running-in and wear.

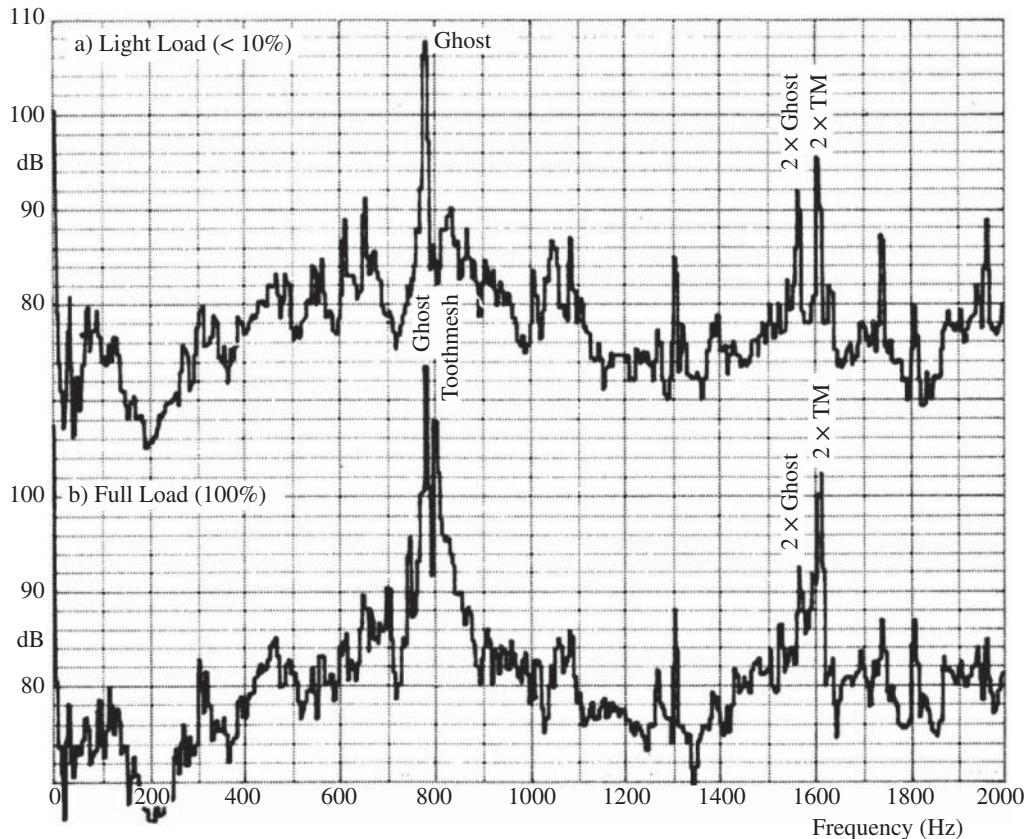


Figure 2.15 Effect of load on ghost and toothmesh components. Source: Courtesy Brüel & Kjær.

2.2.3 Rolling Element Bearings

Rolling element bearings are one of the most widely used elements in machines and their failure one of the most frequent reasons for machine breakdown. However, the vibration signals generated by faults in them have been widely studied, and very powerful diagnostic techniques are now available as discussed in Section 7.3.

Figure 2.17 shows typical acceleration signals produced by localised faults in the various components of a rolling element bearing, and the corresponding envelope signals produced by amplitude demodulation. It will be shown that analysis of the envelope signals gives more diagnostic information than analysis of the raw signals. The diagram illustrates that as the rolling elements strike a local fault on the outer or inner race, a shock is introduced that excites high frequency resonances of the whole structure between the bearing and the response transducer. The same happens when a fault on a rolling element strikes either the inner or outer race. As explained in [28], the series of broadband bursts excited by the shocks is further modulated in amplitude by two factors:

- The strength of the bursts depends on the load borne by the rolling element(s), and this is normally modulated by the rate at which the fault is passing through the load zone.
- Where the fault is moving, the transfer function of the transmission path varies with respect to the fixed positions of response transducers.

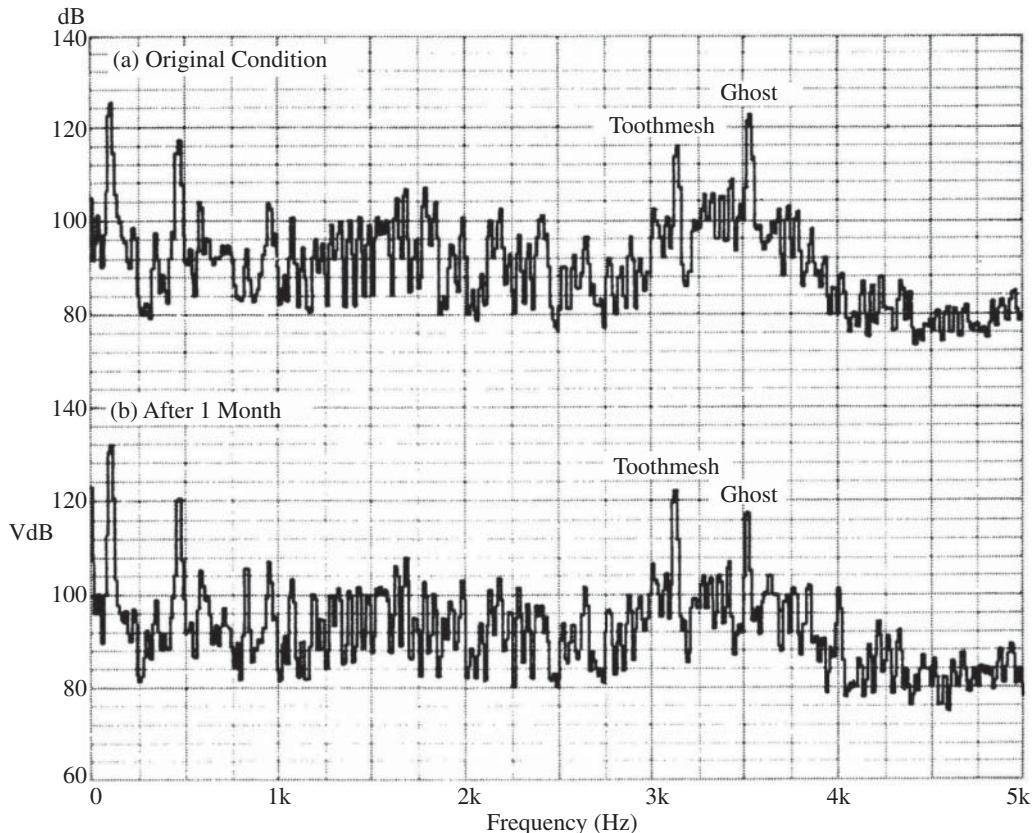


Figure 2.16 Effects of wear on toothmesh and ghost components. Source: Courtesy Brüel & Kjær.

Figure 2.17 illustrates typical modulation patterns for unidirectional load on the bearing, and fixed outer race; at shaft speed for inner race faults, and cage speed for rolling element faults. The formulae for the various frequencies shown in Figure 2.17 are as follows:

$$\text{Ballpass frequency, outer race : } BPFO = \frac{n f_r}{2} \left\{ 1 - \frac{d}{D} \cos \phi \right\} \quad (2.12)$$

$$\text{Ballpass frequency, inner race : } BPFI = \frac{n f_r}{2} \left\{ 1 + \frac{d}{D} \cos \phi \right\} \quad (2.13)$$

$$\text{Fundamental train frequency (cage speed) } FTF = \frac{f_r}{2} \left\{ 1 - \frac{d}{D} \cos \phi \right\} \quad (2.14)$$

$$\text{Ball (roller) spin frequency : } BSF(RSF) = \frac{f_r D}{2d} \left\{ 1 - \left(\frac{d}{D} \cos \phi \right)^2 \right\} \quad (2.15)$$

where f_r is the shaft speed, n is the number of rolling elements, and ϕ is the angle of the load from the radial plane. Note that the ballspin frequency (BSF) is the frequency with which the fault strikes the same race (inner or outer), so that in general there are two shocks per basic period. Thus, the even harmonics of BSF are often dominant, in particular in envelope spectra.

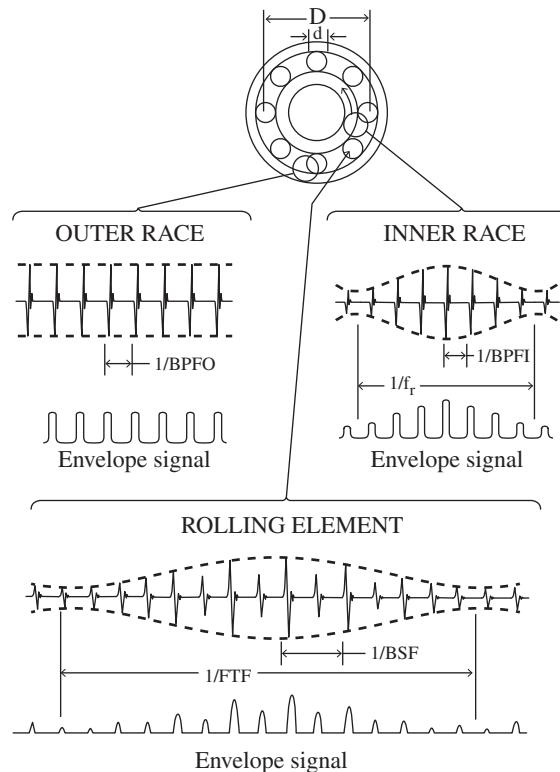


Figure 2.17 Typical signals and envelope signals from local faults in rolling element bearings.

These are however the kinematic frequencies assuming no slip, and in actual fact there must virtually always be some slip for the following reasons; the angle ϕ varies with the position of each rolling element in the bearing, as the ratio of local radial to axial load changes, and thus each rolling element has a different effective rolling diameter and is trying to roll at a different speed. The cage ensures that the mean speed of all rolling elements is the same, by causing some random slip. This is typically of the order of 1–2%, both as a deviation from the calculated value and also as a random variation around the mean frequency. This random slip, while small, does give a fundamental change in the character of the signal, and is the reason why envelope analysis extracts diagnostic information not available from frequency analyses of the raw signal. It means that bearing signals can be considered as (almost) cyclostationary (see Chapters 3 and 7). This also allows bearing signals to be separated from gear signals with which they are often mixed, as discussed in Sections 5.3, 7.2, and 7.3.

It should be noted that the argument about variation of rolling diameter with load angle applies equally to spherical roller bearings, since by virtue of their kinematics, the ratio of roller diameter to race diameter varies with the axial position, and so there is only one position where there is no slip. The slip on either side of this position is in opposite directions, and generates opposing friction forces which balance, but the location of the no-slip diameter is strongly influenced by the point of maximum pressure between the rollers and races, and is thus dependent on the ratio of axial to radial load, which varies with the rotational position of the roller in the bearing. For taper roller bearings the situation is not so clear, since because of the conical geometry the ratio of roller diameter to pitch diameter (d/D) should be a constant independent of the axial position along the roller. However, if

the load angle does not correspond to the inclination angle of the roller, the load centres on the inner and outer race will not be diametrically opposed and the ratio d/D will still vary with the position of the roller in the bearing. The same argument cannot be made for cylindrical roller bearings, which are unable to sustain an axial load, but on the other hand, they would rarely have negative clearance, and the rollers are only compelled to roll in the load zone. Thus, when they enter the load zone, they will tend to have a random position in the clearance of the cage, and the repetition frequency would have a stochastic variation as for other bearing types, even if the deviation of the mean value from the kinematic frequency is less. Moreover, because of elasto-hydrodynamic effects at the interfaces there would not be a pure non-slip rolling action between rollers and races.

The basic reason why there is often no diagnostic information in the raw spectrum is illustrated in Figure 2.18. This shows acceleration signals from a simulated outer race fault, with and without random slip. Spectra are shown for both the raw signal and the envelope. Only one resonance is shown, but this could be the lowest of a series. As is quite common, the resonance frequency is two orders of magnitude higher than the repetition frequency of the impacts. Because the frequency response of the fault pulses is measured in terms of acceleration, the spring line at low frequencies is an ω^2 parabola, with zero value and zero slope at zero frequency. Thus, the low harmonics of the repetition frequency have very low magnitude and are easily masked by other components in the spectrum. If the signal were perfectly periodic, the repetition frequency could be measured as the separation of the harmonic series in the vicinity of the resonance frequency, but as illustrated in Figure 2.18e, the higher harmonics smear over one another with even a small amount of slip (here 0.75%). However, the envelope spectra show the repetition frequency even with the small amount of slip, despite the fact that the higher harmonics in the latter case are slightly smeared.

As mentioned, the lowest resonances significantly excited are often, but not universally, very high with respect to the bearing characteristic frequencies. It would for example not be the case for gas turbine engines, where the latter are often in the kHz range. Even so, the low harmonics of the bearing characteristic frequencies are almost invariably strongly masked by other vibration components, and it is generally easier to find wide frequency ranges dominated by the bearing signal in a higher frequency range. The advantage of finding an uncontaminated frequency band encompassing several harmonics of the characteristic frequency is that bearing fault signals are generally impulsive, but

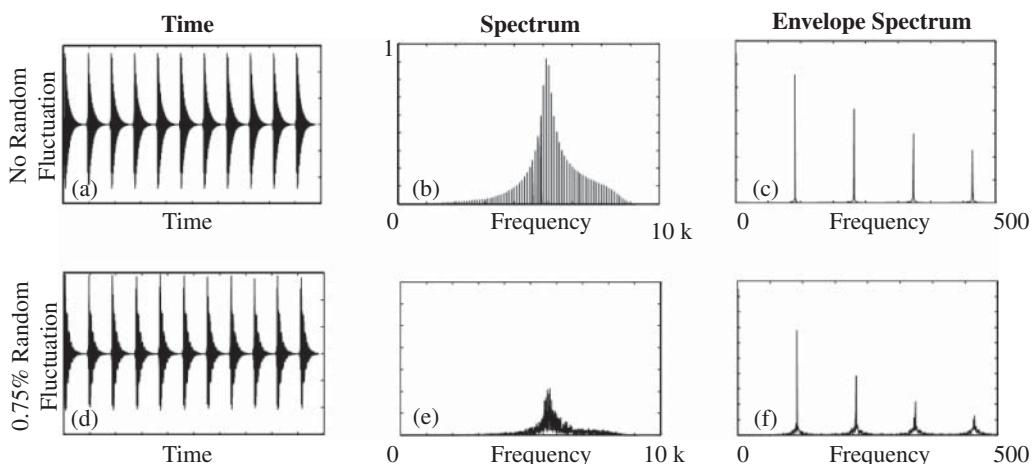


Figure 2.18 Bearing fault pulses with and without random fluctuations (a, d) Time signals (b, e) Raw spectra (c, f) Envelope spectra.

cannot be recognised as such unless the frequency range includes at least 10 or so harmonics. If a pulse train is lowpass filtered between the first and second harmonics of the repetition frequency, the result is a sinewave, with no impulsivity at all. The most powerful bearing diagnostic techniques depend on detecting and enhancing the impulsiveness of the signals, and so the fact that low harmonics of the bearing characteristic frequencies can sometimes be found in raw spectra is ignored in the models based on impulse responses to sharp impacts. As a counter example, a paper by the author [29] was the first to use the cepstrum to diagnose bearing faults, this relying on being able to find separated harmonics of the bearing frequency over a reasonably wide frequency range. It was a high-speed machine (an auxiliary gearbox running at 3000 rpm), and the first 20 or 30 harmonics were separated and gave a component in the cepstrum. On the other hand, the primary method recommended in this book, envelope analysis, performed equally well if not better, and does not require the harmonics to be separated, as illustrated in Figure 2.18, so the cepstrum method has little application.

Bearing faults usually start as small pits or spalls, and give sharp impulses in the early stages covering a very wide frequency range (even in the ultrasonic frequency range to 100 kHz). However, for some faults such as brinelling, where a race is indented by the rolling elements giving a permanent plastic deformation, the entry and exit events are not so sharp, and the range of frequencies excited not so wide. They would still generally be detected by envelope analysis, however, as described in Section 7.3.

Since the publication of the first edition, it has become apparent that when faults become physically larger, there can be measurable low harmonics of the bearing frequencies, even in terms of acceleration. These can be ascribed to the fact that the forces exciting responses are due to geometric errors in the races and rolling elements, and are not fixed forces as such. They must give additive components in the relative displacement of the races, but the amount measured on the casing (directly connected to the outer race) will depend on the distribution between the two races, and thus the relative kinetics of the rotor and stator of the individual machine. It can however, explain why the low harmonics, when measurable, do not necessarily follow the ω^2 pattern of the impulse response model.

Cases have been encountered where faults have not been detected while small, and the spalls have become extended and smoothed by wear. Although not necessarily generating sharp impacts any more, this type of fault can often be detected by the way in which it modulates other machine signals, such as the garmesh signal generated by gears supported by the bearings. Figure 2.19 illustrates the case of an extended inner race spall, where the garmesh signal is modulated by the type of signal shown, a mixture of a deterministic (local mean) part, and an amplitude modulated noise as the rough section of the race comes into the load zone. It should be kept in mind that the rolling elements are on a different part of this rough surface for every revolution of the inner race. As discussed in Sections 3.6 and 7.3, this is a different type of cyclostationary signal, and once again can be distinguished from gear signals because the latter are deterministic.

2.2.4 Bladed Machines

Many machines such as rotating fans, pumps, compressors and turbines, have a number of uniformly spaced blades or vanes on the rotor, which interact with components on the stator to give a periodic excitation of the casing. The basic frequency is the blade-pass frequency, or number of blades (vanes) on the rotor multiplied by the rotational speed of the shaft. When the machine is in good condition, the interactions are generally quite small, in the sense that the flow of fluid from the rotor is guided to have the correct angle, so that it corresponds with the angle of any guide vanes or blades on the stator

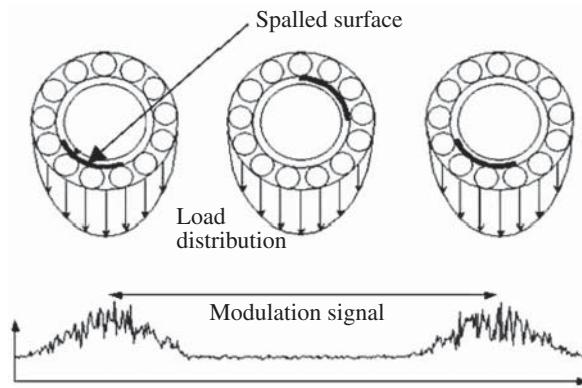


Figure 2.19 Typical modulating signal from the effect of an extended inner race fault on a gear signal.

(e.g. stator blades in turbines). If the fluid flow angles change, for example because of a damaged blade or build-up of fouling, more impulsive interactions will occur, giving a change in the vibration signals.

An example was given in [30] from measurements made on steam turbines at the French Electrical Authority EDF. Some turbines were prone to fatigue failure of the blades, and a serious failure was sometimes preceded by breakage of one or two small blades. If this could be detected, the more serious failure could be avoided. The misdirected steam from one broken blade caused an impulsive interaction with stator blades, and was picked up by an externally mounted accelerometer as an impulsive event repeating once per revolution of the shaft. This resulted in a considerable increase in shaft speed harmonics in a mid frequency range with low response under normal conditions. This case is given as an example of the use of cepstrum analysis to detect harmonic patterns in Chapter 6.

There are many similarities with gear diagnostics, in that errors can be divided into a mean part, the same for the passage of all blades, and variations between blades. The former gives changes at the blade-pass frequency and its harmonics, while the latter gives changes at other harmonics of shaft speed including sidebands around the harmonics of blade-pass frequency (analogous to Figure 2.13).

There is one difference from gears, at least for signals measured on the casing, in that the interaction between the moving blades and the casing is via a fluid, so that the generation of interactive forces is not entirely deterministic. The fluid pressure will generally be affected by random fluctuations from turbulence, giving a small random modulation around the various harmonics, which makes the signal cyclostationary.

2.2.5 Electrical Machines

Electrical machines, motors and generators, have vibrations induced by electromagnetic forces in addition to the usual forces from mechanical effects such as unbalance, misalignment, etc. Only AC (alternating current) machines are treated here, as they are by far the most prevalent, even for variable speed applications since the development of power electronics.

There is a fundamental difference between synchronous motors (and generators), where the rotor runs at synchronous speed, and induction (asynchronous) motors where the rotor runs at slightly less than synchronous speed. In both cases there is a rotating magnetic field in the stator, whose speed depends on the number of poles. A two-pole machine (one north pole and one south pole) gives

the highest speed, corresponding to the mains frequency (US line frequency), since considering one mains cycle the positive voltage peak, say, can be associated with the north pole and the negative peak with the south pole. Thus, in one mains period, both poles have rotated one full revolution around the stator. In a four pole motor the passage from one north pole to the next involves only half a revolution, so the speed is halved to half mains frequency, and so on. It should be noted that independent of whether the mains voltage is positive or negative, the torque produced is always in the same direction (rectified) so that the mechanical effects of electrical signals are usually at multiples of twice mains frequency. If components at mains frequency are found in measured vibration signals, suspicion should immediately fall on the measurement system, as they would normally come from electrical interference effects such as ‘ground loops’.

In synchronous motors the rotor also has magnetic poles, either permanent magnets or electromagnets, and the rotor locks on to the rotating field. Increased torque load gives a phase lag of the rotor, but it rotates at the synchronous speed. Torque fluctuations result in a phase modulation of the rotor speed. In an induction motor, eddy currents are induced in the rotor and their magnetic field interacts with the rotating field in the stator causing the rotor to attempt to follow the rotating field, but with a difference or ‘slip’ frequency which is proportional to the torque load. Thus, torque fluctuations in an induction motor give a frequency rather than phase modulation of the rotor speed.

The vast majority of generators (alternators) are synchronous machines, whose properties are similar to motors except that the rotor is driven mechanically and has a phase advance rather than lag, so that electrical currents are induced in the field windings. Induction machines can also be used as generators, in which case the rotor is driven at above the synchronous speed.

In this book, ‘slip frequency’ is taken as the difference between synchronous speed and shaft speed, although electrical engineers sometimes have a different definition.

2.2.5.1 Vibrations of Synchronous Machines

The rotating magnetic fields in the stator induce mechanical distortions that are independent of whether it is a north or south pole, and so at a fixed point on the stator, these fluctuations occur at the ‘pole-pass frequency’ or twice mains frequency (100 Hz in Europe; 120 Hz in the USA), independent of the number of poles. Any anomaly in the stator would give a change in the rotating flux patterns and thus an increase in the twice mains frequency component.

With two-pole synchronous machines, twice mains frequency is also twice shaft speed, so it can be difficult to separate electrical from mechanical effects such as misalignment. This was discussed in connection with Figure 2.5. There are two ways in which the effects can be separated. The first is by varying the load on the machine, which should affect the electrically based excitation, but the mechanical excitation in only a minor way. Figure 2.20 illustrates how the two effects might dominate the second harmonic component in different load ranges.

The other way to distinguish electrical from mechanical effects is to disconnect the power and observe the resulting vibrations. Electrical effects should die away very quickly, whereas mechanical effects such as misalignment would change more slowly and would be locked to shaft speed as it decayed.

2.2.5.2 Vibrations of Induction Machines

Stator faults in induction motors will show up at the pole-passing frequency (twice mains frequency) as for synchronous machines, but they are now separated from mechanical faults because the rotor is running slower than synchronous speed by an amount equal to the slip frequency. The difference

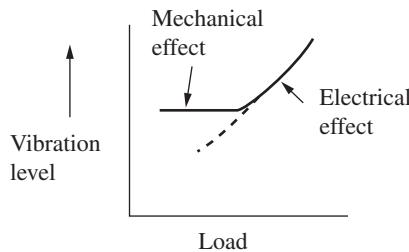


Figure 2.20 Load variation to distinguish the cause of an observed effect.

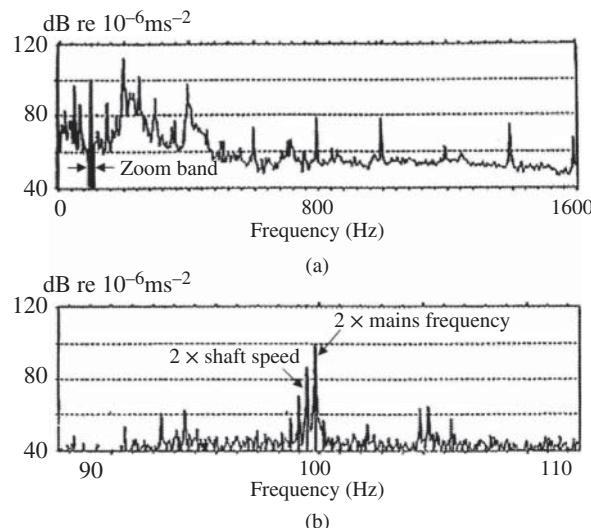


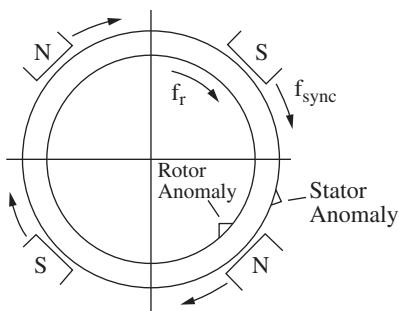
Figure 2.21 Use of FFT zoom to separate the harmonics of shaft speed from those of mains (line) frequency in an induction motor vibration spectrum. (a) Baseband spectrum with zoom band around 100 Hz highlighted (b) Zoom spectrum showing that twice mains frequency dominates over twice shaft speed.

may only be a fraction of a Hertz, but can generally be separated by ‘zoom analysis’ (see Section 3.2.3). An example is given in Figure 2.21, which shows that a strong component at approximately 100 Hz in the vibration signal from a 2-pole induction motor driving a screw compressor, is shown by zoom analysis to be dominated by twice mains frequency rather than twice shaft speed, and thus has an electrical origin.

A fault on the rotor rotates at shaft speed, and will generally show up in the vibration at shaft speed. However, it can generally be differentiated from other shaft speed responses, such as unbalance, by the strong modulation that usually accompanies rotor faults [31]. This occurs at the rate at which the rotating field poles pass a given point on the rotor (the anomaly). This is equal to the slip frequency multiplied by the number of poles. This is summarised in Figure 2.22, including the formulae for the various frequencies involved in induction motor diagnostics.

For example, for a four-pole induction motor operating in the USA with 60 Hz mains frequency, the synchronous frequency is 30 Hz. If the shaft speed is 29.75 Hz, it means that the slip frequency is 0.25 Hz. The modulating frequency for a rotor fault would be $4 \times 0.25 = 1$ Hz. Every four seconds

INDUCTION MOTOR DIAGNOSTICS



$$\text{Mains frequency} = f_{\text{mains}}$$

$$\text{Number of poles} = N_p$$

$$\text{Synchronous frequency } f_{\text{sync}} = 2 * f_{\text{mains}} / N_p$$

$$\text{Rotor speed} = f_r$$

$$\text{Slip frequency } f_{\text{slip}} = f_{\text{sync}} - f_r$$

$$\text{Stator fault frequency} = \text{pole passing frequency} = 2 * f_{\text{mains}}$$

$$\text{Rotor fault frequency} = f_{\text{slip}} * N_p$$

Figure 2.22 Illustration of the various frequencies associated with a fault on the stator or rotor of an induction motor.

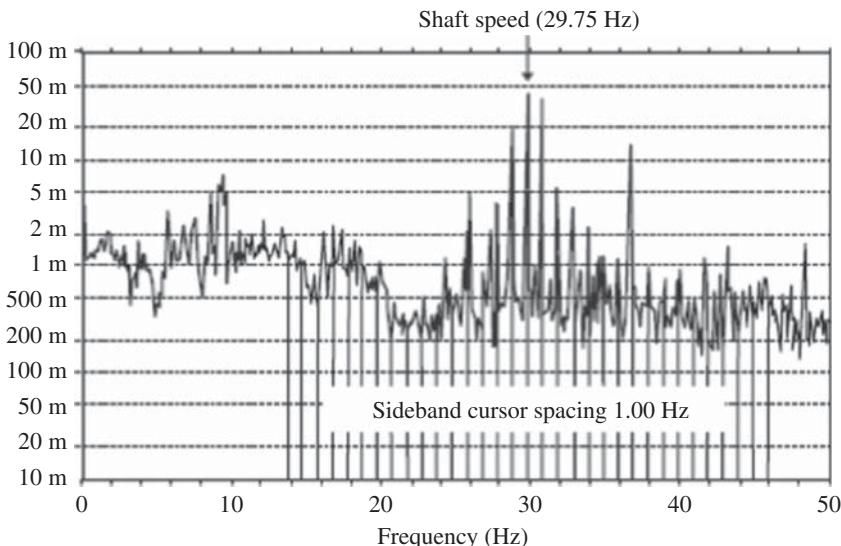


Figure 2.23 Example of a rotor fault on an induction motor, showing modulation sidebands around the shaft speed component of 29.75 Hz (spacing 1 Hz).

the rotating field would make 120 revolutions while the shaft would make 119 revolutions. The difference of one revolution means that four poles have passed every point on the rotor in four seconds, or in other words one per second or 1 Hz. Figure 2.23 illustrates just such a case, where a rotor fault gives a component at shaft speed 29.75 Hz, but surrounded by strong sidebands spaced at the modulation frequency 1 Hz.

Squirrel cage induction generators have been widely used in conjunction with wind turbines, and since it is beneficial to adjust the rotational speed according to wind velocity, this is usually done by changing between 4-pole and 6-pole operation, to give one speed 50% higher than the other.

It is even more efficient to have continuously variable speed, and later wind turbine systems have used so-called ‘doubly fed induction generators’, where a rotating field is supplied to both the stator and the rotor. The rotation of the rotor field can be forwards or backwards with respect to the stator field (which is usually supplied directly from the mains), meaning that the final rotor speed can be higher or lower than the normal synchronous speed, typically by an amount of $\pm 30\%$. The name is somewhat unfortunate as the machine is actually a synchronous generator, with no slip between the rotating fields on the rotor and stator. In fact, if the field supplied to the rotor is non-rotating, it becomes a normal synchronous generator.

It should perhaps be mentioned that there are a number of other mechanical effects associated with electrical machines, which affect the rotor dynamics, but these require specialist analysis of the type discussed in Section 2.2.1.3. An example is the fact that the rotor is attracted to the stator with (very nonlinear) magnetic forces, thus acting as a nonlinear negative spring in the rotor dynamic equations. Ref. [32] is an example of a publication dealing with such matters.

2.3 Signals Generated by Reciprocating Machines

This section deals mainly with signals from internal combustion (IC) engines, such as diesel and spark ignition engines, but also includes other reciprocating machines such as pumps and compressors.

Much more than with rotating machines, the vibration signals from reciprocating machines are a series of responses to impulsive events in the machine cycle, such as combustion, piston slap, bearing knock, valves opening and closing, etc. The four phases of an IC engine cycle are induction – compression – combustion – exhaust. In a four-stroke engine each of these phases occupies one stroke, or half a revolution, so that a complete cycle is two revolutions, and the cyclic frequency is half crankshaft speed. For two-stroke engines, all four phases are achieved in two strokes, or one revolution, so that the cyclic frequency is equal to crankshaft speed.

Reciprocating compressors are divided into ‘single-acting’, where gas is compressed in only one direction of the piston motion, and ‘double-acting’, where compression is achieved during both forward and backward strokes of the piston.

Figure 2.24 is a cross-section through a gas engine/compressor that will be used to demonstrate the basic concepts of reciprocating machine vibrations. Such a gas engine/compressor, in a 12- or 16-cylinder version, is often used to pump natural gas through pipelines, using some of the gas as fuel. It is a spark ignition engine, and in fact has two spark plugs in each cylinder for security. Each ‘throw’ of the crankshaft has two engine pistons in a ‘V’ arrangement, and one compressor piston. Note that the piston rod moves parallel to the cylinder, so as to facilitate double-acting operation, and the compressor connecting rod is connected to the piston rod at a sliding ‘crosshead’.

2.3.1 Time-Frequency Diagrams

Since the signals from reciprocating machines vary in both time and frequency, it is convenient to represent them in some form of ‘time-frequency diagram’. A number of different varieties of such diagrams are discussed in Chapter 3, but the type used here for illustration purposes is related to the STFT (short time Fourier transform) of Chapter 3, except that because of the cyclic

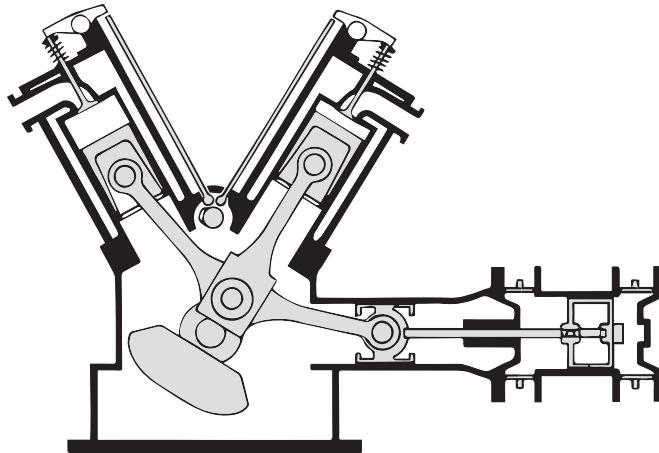


Figure 2.24 Cross-section through a gas engine/compressor.

operation, averages can be made over several cycles, this tending to give reproducible diagrams for a given machine under the same operating conditions. The way in which such diagrams are generated is illustrated in Figure 2.25. This shows a typical vibration (acceleration) signal measured on the head of a diesel engine, and including several cycles. The largest impulsive events in the cycle are due to combustion, and are biggest for combustion in one cylinder because the accelerometer happens to be closest to that cylinder. A once-per-cycle (i.e. every two revolutions) trigger signal has been recorded at the same time as the vibration signal to provide a timing reference.

Figure 2.25c shows how a short ‘time window’, with a fixed time delay after each trigger pulse, is able to select out an equivalent short section of acceleration signal from a series of successive engine cycles. The sine squared window displayed is known as a ‘Hanning’ window and is discussed in detail in Chapter 3, where it is shown that multiplying the signal by such a smooth window is better than simply using a rectangular window with abrupt steps at the beginning and end. Figure 2.25d,e illustrate the spectra of each successive windowed section (for different time delays), and at the right the averages over about 30 cycles, which tend to be reproducible. When this is repeated for more overlapping window positions over a complete cycle, a time-frequency diagram can be produced which has time (interpretable as crank angle) on one axis, and frequency on the other.

Two examples of such diagrams are given in Figure 2.26 for measurements made on the compressor of Figure 2.24 in both single-acting and double-acting mode. The trigger signal was based on engine cycles and so encompassed two revolutions of the crankshaft, or two compression cycles. Note that the amplitude scale of the spectra is in dB (with respect to a reference acceleration level) and so covers a wide dynamic range of 60 dB. In single-acting mode there are large sections of the diagram where not much is going on, and in contrast something is detected in these areas in double-acting mode. The greater strength of the signals for compression in the outward moving direction is partly because of the position of the accelerometer on that head.

As an illustration of the advantages of dividing up the signal in both the time and frequency directions, Figure 2.27 shows averaged frequency spectra (over all time) for the same two cases. Very little difference can be seen between them, because the additive components from compression in the reverse direction are several dB smaller, and thus add little to the averages.

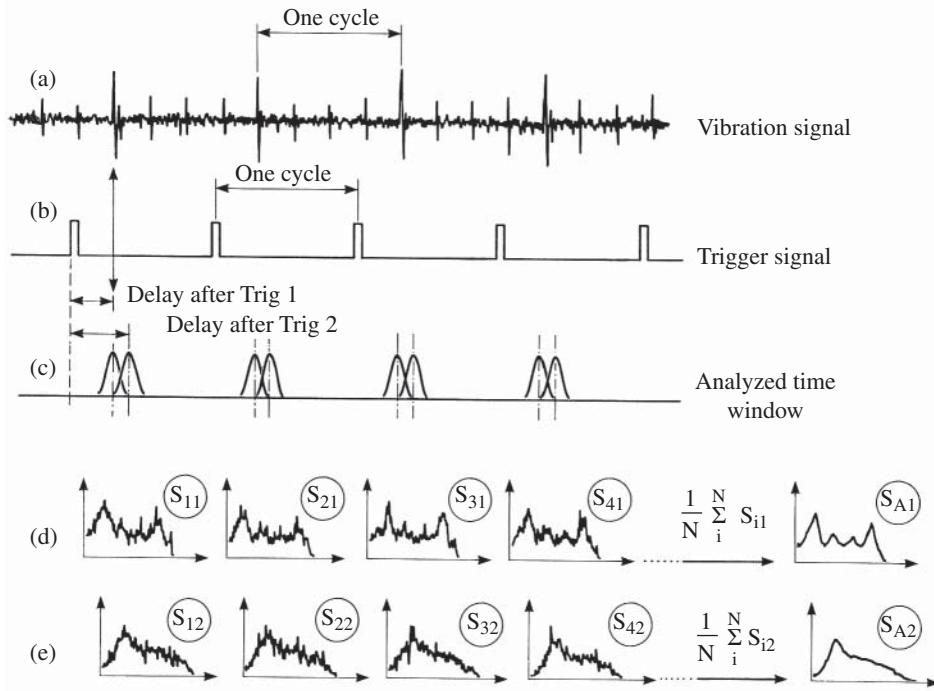


Figure 2.25 Time-frequency analysis procedure. Source: Courtesy Brüel & Kjær.

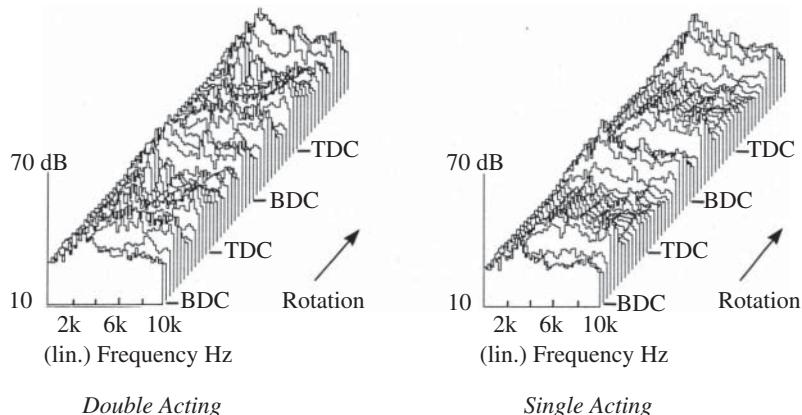


Figure 2.26 Time-frequency diagrams for the compressor of Figure 2.24 in single- and double-acting modes. BDC = Bottom dead centre; TDC = top dead centre. Source: Courtesy Brüel & Kjær.

Measurements made on the engine casing are also illustrative, as they indicate the best mounting point for an accelerometer to capture the most important events in the engine cycle. Figure 2.28 shows the time-frequency diagrams for four different measurement points. It is interesting that the dominant events detected are to do with opening and closing of valves, and the combustion is not very much in evidence, in contrast to diesel engines as in Figure 2.25. This is presumably because the compression ratio in the gas engine is quite low, and the combustion relatively gentle.

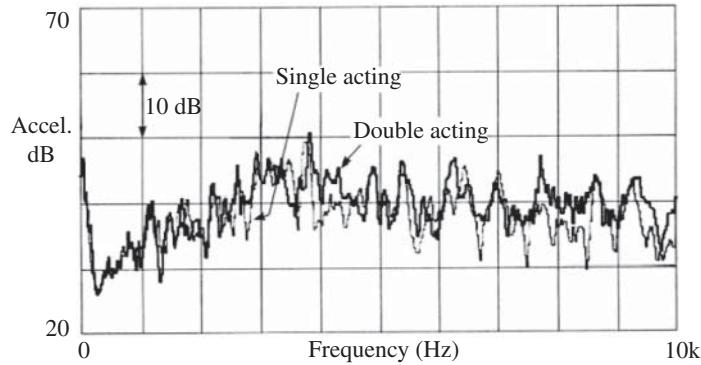


Figure 2.27 Averaged spectra for the two signals of Figure 2.26. Source: Courtesy Brüel & Kjær.

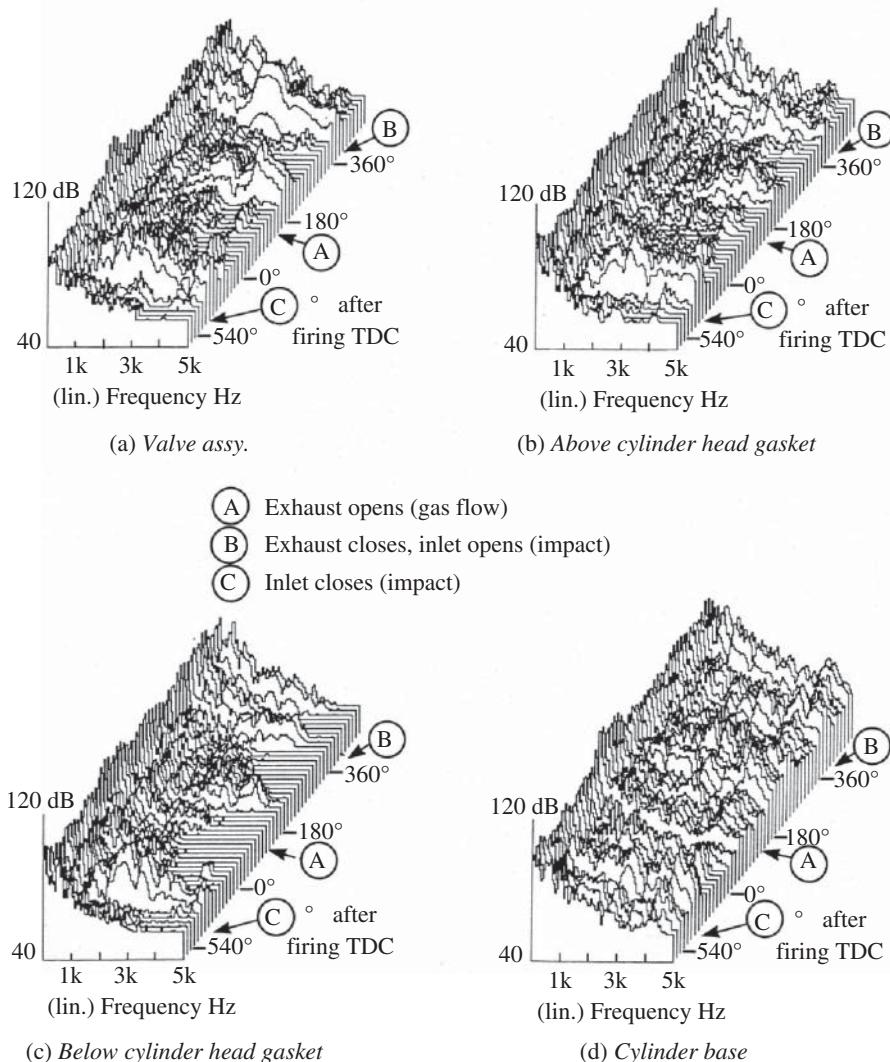


Figure 2.28 Time-frequency diagrams for different measurement positions on an engine. Source: Courtesy Brüel & Kjær.

It is seen that, even though strongest when measured near the valve assembly, the signals from the valves opening and closing are detectable below the cylinder head gasket. On the other hand, an event occurring after event B, and measured most strongly at the base of the cylinder, is also detected below the cylinder head gasket, but not so well at the top of the engine. It would thus appear that the position below the gasket would be the best for capturing all events for this particular engine.

Such a trial and error procedure is one way of determining the best transducer positions for measurements of this type. In Section 7.4 there is a discussion of the means of determining the best accelerometer positions for representing the combustion signals in diesel engines, and in particular trying to reconstruct combustion pressure signals from external acceleration measurements.

2.3.2 Torsional Vibrations

Torsional vibrations, i.e. variations in angular velocity of the crankshaft, result directly from the fluctuating torque given by the combustion events in an engine, or even the varying cylinder pressure in a reciprocating compressor or pump. Torsional vibration represents a frequency and/or phase modulation of an otherwise uniform rotational speed, and is treated in detail in Chapter 3 in conjunction with demodulation in general. It can be easily measured using shaft encoders attached to the shaft in question, or with torsional laser vibrometers (Chapter 1).

For smaller engines, with effectively rigid crankshafts, the torsional response of the crankshaft follows the torque generated by each cylinder in an engine, and is a good indicator of combustion uniformity.

Figure 2.29 shows the angular velocity for a 6-cylinder spark ignition engine with a complete misfire on one cylinder [33]. In normal operation there are six uniform fluctuations in angular velocity each cycle, coming from six uniform torque pulses. With a complete misfire in one cylinder, as in Figure 2.29, the speed drops dramatically, and has to be gradually built up again by the ensuing five remaining torque pulses.

For larger multiple cylinder engines, where the torsional vibration modes of the crankshaft fall within the frequency range of the excitation, the situation is more complicated, and the response

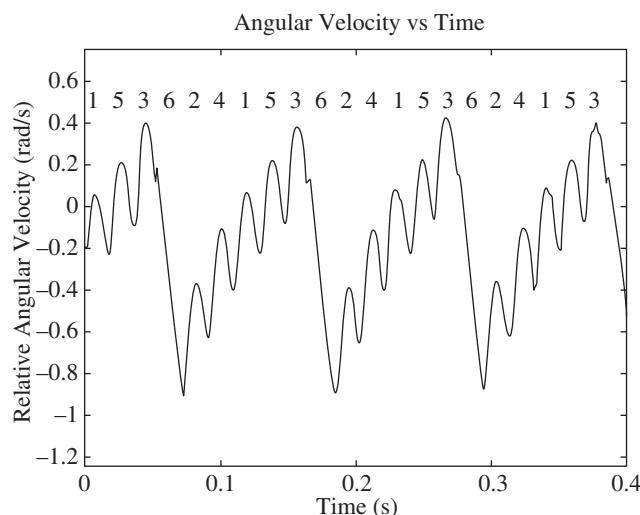


Figure 2.29 Angular velocity fluctuations for a misfire in one cylinder.

measured in one position is no longer directly representative of the whole crankshaft. Section 8.4.1.1 includes a discussion of how this can be taken into account by using a dynamic model of the crankshaft to compensate for such effects.

References

1. Sohre, J. (1968). Operating problems with high speed turbomachinery, causes and corrections. *ASME Petroleum Mechanical Engineering Conference*.
2. Childs, D. (1993). *Turbomachinery Rotordynamics*. NY: Wiley.
3. Gasch, R., Nordmann, R., and Pfützner, H. (2002). *Rotordynamik* (in German). Berlin: Springer.
4. Rao, J.S. (1996). *Rotor Dynamics*, 3e. New Delhi: New Age International Publishers.
5. Genta, G. (2005). *Dynamics of Rotating Systems*. Springer.
6. Dewell, D.L. and Mitchell, L.D. (1983). Detection of a misaligned disc coupling using spectrum analysis. *MSA Session, ASME Conference*, Dearborn, MI (11–14 September). Publisher American Society of Mechanical Engineers, New York, NY.
7. Bloch, H.P. (1976). How to uprate turbomachinery by optimised coupling selection. *Hydrocarbon Processing* 55 (1): 87–90.
8. Bachschmid, N., Pennacchi, P., and Vania, A. (2007). Thermally induced vibrations due to rub in real rotors. *Journal of Sound and Vibration* 299 (4–5): 683–719.
9. Sanderson, A.F.P. (1992). The vibration behaviour of a large steam turbine generator during crack propagation through the generator rotor. In: *IMechE International Conference on Vibrations in Rotating Machinery*, Bath, UK, paper C432/102, 263–273.
10. Henry, R.A. and Okah-Avae, B.E. (1976). Vibrations in cracked shafts. Paper C162/76, *IMechE Conference on Vibrations in Rotating Machinery*, Cambridge.
11. Mayes, I.W. and Davies, W.G.R. (1976). The vibrational behaviour of a rotating shaft system containing a transverse crack. Paper C168/76, *IMechE Conference on Vibrations in Rotating Machinery*, Cambridge.
12. Gasch, R. (1976). Dynamic behaviour of a simple rotor with a transverse crack. Paper C178/76, *IMechE Conference on Vibrations in Rotating Machinery*, Cambridge.
13. Baumgartner, R.J. and Ziebarth, H. (1982). Vibration monitoring criteria for early discernment of turbine rotor cracks. In: *Proc. EPRI 1982 Conference and Workshop*, Hartford CT, August 25–27, 3.1–3.18.
14. Perratt, D.W. (1982). Development of condition monitoring equipment for CEGB generating plant. In: *Proc. EPRI 1982 Conference and Workshop*, Hartford CT, August 25–27, 2.88–2.102.
15. Kottke, J.J. and Menning, R.H. (1981). Detection of a transverse crack in a turbine shaft – the oak creek experience. *ASME Conference Paper 81-JPGC-Pwr-19*.
16. Bachschmid, N., Pennacchi, P., and Tanzi, E. (2009). *Cracked Rotors*. Springer.
17. Kicinsky, J. (2006). *Rotor Dynamics*. Gdansk, Poland: IFFM Publishers.
18. Bachschmid, N., Pennacchi, P., and Vania, A. (2002). Identification of multiple faults in rotor systems. *Journal of Sound and Vibration* 254 (2): 327–366.
19. Bachschmid, N., Pennacchi, P., and Vania, A. (2006). A model based identification method of transverse cracks in rotating shafts suitable for industrial machines. *Mechanical Systems and Signal Processing* 20 (8): 2112–2147.
20. Chen, P.Y.P., Feng, N., Hahn, E.J., and Hu, W. (2005). Recent developments in turbomachinery modeling for improved balancing and vibration response analysis. *ASME Journal of Engineering for Gas Turbines and Power* 127 (July): 646–653.
21. Feng, N., Hahn, E.J., and Hu, W. (2004). A comparison of configuration state identification techniques based on sensitivity to measurement error. In: *Proc. ISMA 2004*, 2525–2536. Leuven: KUL.
22. Black, H.F. (1968). Interaction of a whirling rotor with a vibrating stator across a clearance annulus. *Journal of Mechanical Engineering Science* 10 (1): 1–12.
23. Smith, J.D. (1983). *Gears and their Vibration*. NY: Marcel Dekker.
24. Mark, W.D. (1978). Analysis of the vibratory excitation of gear systems: basic theory. *Journal of Acoustical Society of America* 63: 1409–1430.
25. Mark, W.D. (1979). Analysis of the vibratory excitation of gear systems. II: Tooth error representations, approximations, and application. *Journal of Acoustical Society of America* 66: 1758–1787.
26. Randall, R.B. (1982). A new method of modeling gear faults. *ASME Journal of Mechanical Design* 104: 259–267.
27. Drosjack, M.J. and Houser, D.R. (1977). An experimental and theoretical study of the effects of simulated pitch line pitting on the vibration of a geared system. *ASME DET Conference*, Chicago, IL (26–30 September).

28. McFadden, P.D. and Smith, J.D. (1984). Model for the vibration produced by a single point defect in a rolling element bearing. *Journal of Sound and Vibration* 96 (1): 69–82.
29. Bradshaw, P. and Randall, R.B. (1983). Early detection and diagnosis of machine faults on the Trans Alaska Pipeline. *MSA Session, ASME Conference*, Dearborn MI (September 11–14, 1983). Publisher American Society of Mechanical Engineers, New York, NY.
30. Sapy, G. (1975). Une Application du Traitement Numérique des Signaux au Vibratoire de Panne: La Détection des Ruptures d'Aubes Mobiles de Turbines. *Automatisme, Tome XX* (10): 392–399.
31. Maxwell, J. H. (1983). Induction motor magnetic vibration. *Proc. Vibration Institute Meeting*, Houston (19–21 April).
32. Pennacchi, P. (2008). Computational model for calculating the dynamical behaviour of generators caused by unbalanced magnetic pull and experimental validation. *Journal of Sound and Vibration* 312: 332–353.
33. Jenner, L. (1995). Ford IC Engine Diagnostics. BE Thesis, School of Mechanical and Manufacturing Engineering, UNSW, Australia.

3

Basic Signal Processing Techniques

3.1 Statistical Measures

3.1.1 Probability and Probability Density

In Chapter 2, random signals were discussed with respect to their appearance in the time and frequency domains. Another way of characterising random signals is the way in which their instantaneous value is distributed. This can be expressed in terms of the ‘probability distribution’, or in other words the probability that their value is less than or equal to a specified value [1].

Figure 3.1 shows a section of a typical random signal $x(t)$ with minimum value x_{min} and maximum value x_{max} within the selected window. It is first considered sampled at small uniform intervals Δt . The fraction of these samples that are less than or equal to a particular value of x can be used to define a probability distribution function $P(x)$, which is the probability that the value of a particular sample is less than or equal to x . This can be extended to continuous functions, which for a time function means that can be interpreted as the fraction of time that $x(t)$ is less or equal to than x .

Thus,

$$P(x) = \Pr[x(t) \leq x] \quad (3.1)$$

$P(x)$ must have the form shown in Figure 3.2, which states that $x(t)$ is certain to be less than or equal to the maximum value x_{max} (i.e. $P(x_{max}) = 1$) and it can never be less than the minimum value x_{min} (i.e. $P(x_{min}) = 0$). The probability that $x(t)$ is between $x + \Delta x$ and x is obviously $P(x + \Delta x) - P(x)$ as also shown in Figure 3.2.

The probability density $p(x)$ is defined as

$$p(x) = \lim_{\Delta x \rightarrow 0} \frac{P(x + \Delta x) - P(x)}{\Delta x} = \frac{dP(x)}{dx} \quad (3.2)$$

Since $p(x) = \frac{dP(x)}{dx}$ and in the general, case $P(\infty) = 1$ while $P(-\infty) = 0$, it is evident that $\int_{-\infty}^{\infty} p(x)dx = \int_{-\infty}^{\infty} dP(x) = [P(\infty) - P(-\infty)] = 1$, i.e. the total area under the probability density curve must always be 1.

For so-called Gaussian random signals with a ‘normal distribution’, the probability density function (pdf) is given by the formula [1]:

$$p(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3.3)$$

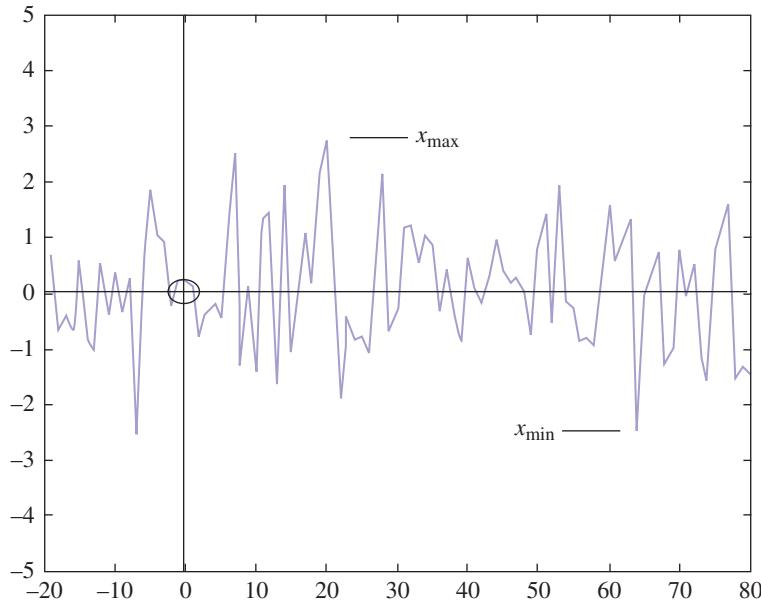


Figure 3.1 Random signal.

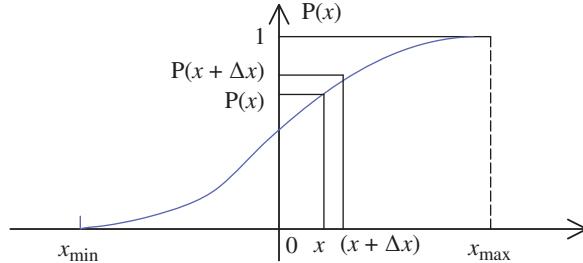


Figure 3.2 Probability distribution for a random signal with maximum value x_{\max} and minimum value x_{\min} .

which is basically an e^{-x^2} curve centred on the mean value μ , scaled in the x-direction in terms of the standard deviation σ , and in the y-direction so as to make the total integral unity. It is depicted in Figure 3.3 (for zero mean value μ).

3.1.2 Moments and Cumulants

The statistical parameters of a signal can be obtained from the pdf by taking various moments, e.g. the mean value

$$\mu = \int_{-\infty}^{\infty} x p(x) dx \quad (3.4)$$

This is the first moment of the pdf, and because the area under the curve is unity, it defines its centre of gravity. Evidently, for any symmetrical function, such as the Gaussian function of Eq. (3.3), the mean value will be at the line of symmetry.

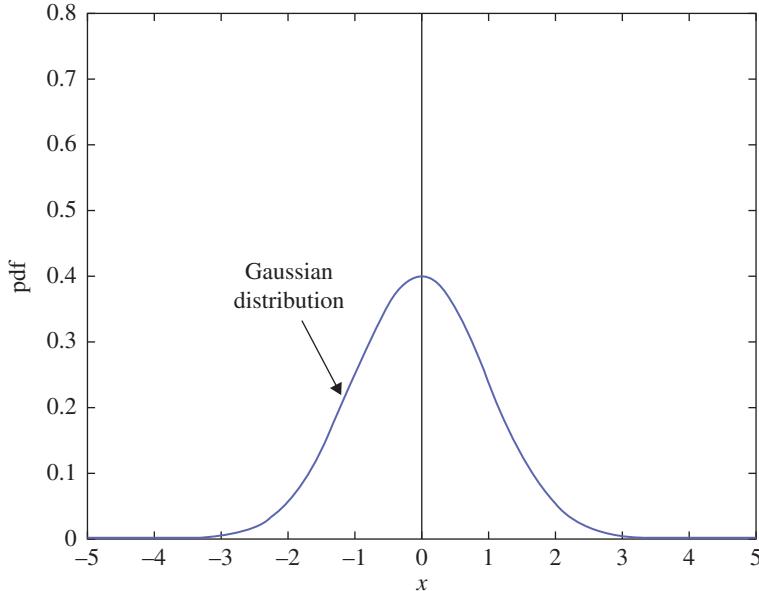


Figure 3.3 Probability density function for the Gaussian distribution.

Similarly, the variance is given by the second moment about the mean value (called the ‘centred moment’), or:

$$\sigma^2 = \int_{-\infty}^{\infty} [x - \mu]^2 p(x) dx \quad (3.5)$$

which corresponds to the ‘moment of inertia’ about the mean value, and whose square root σ (having the same dimensions as x) is known as the standard deviation.

The third centred moment gives a parameter called the ‘skewness’, which is zero for symmetrical functions, and large for asymmetrical functions, while the fourth moment is often called the ‘kurtosis’, and is large for ‘spiky’ or impulsive signals, because of the considerable weighting given to local spikes by taking the fourth power. The skewness and kurtosis are usually normalised by dividing by the appropriate power of the standard deviation, and are thus given by Eqs. (3.6) and (3.7), respectively:

$$S = \frac{\int_{-\infty}^{\infty} [x - \mu]^3 p(x) dx}{\sigma^3} \quad (3.6)$$

$$K = \frac{\int_{-\infty}^{\infty} [x - \mu]^4 p(x) dx}{\sigma^4} \quad (3.7)$$

The moments can also be understood as the expected value of the signal raised to different powers, such that (in the notation of [2]), the r^{th} order moment

$$\mu_{x(r)} = E(x^r) \quad (3.8)$$

with $E(\cdot)$ being the expected value. It is convenient to define a moment generating function $\Phi_x(v)$, or ‘first characteristic function’, as:

$$\Phi_x(v) = E[e^{jvx}] = E \left[1 + jvx + \frac{(jvx)^2}{2} + \frac{(jvx)^3}{3!} + \dots \right] = \sum_0^{\infty} \mu_{x(r)} (jv)^r / r! \quad (3.9)$$

so that the r^{th} order moment (multiplied by j^r) can be obtained by evaluating the r^{th} derivative at $v = 0$. $\Phi_x(v)$ can also be obtained as the (inverse) Fourier transform (FT) of $p(x) = p_x(u)$, (see Section 3.2 for details of the forward and inverse Fourier transform).

$$\Phi_x(v) = \int_{-\infty}^{\infty} e^{jvu} p_x(u) du \quad (3.10)$$

Cumulants are related to moments, and have some advantageous properties. They can be derived from the ‘second characteristic function’, $\Psi_x(v)$, which is the natural logarithm of $\Phi_x(v)$ [2, 3].

$$\Psi_x(v) = \log \left[\int_{-\infty}^{\infty} e^{jvu} p_x(u) du \right] = \sum_0^{\infty} c_{x(r)} (jv)^r / r! \quad (3.11)$$

expressed in terms of its Taylor series expansion, from which the cumulants $c_{x(r)}$ can be derived as the coefficients of the r^{th} derivative evaluated at the origin, in a similar manner to the moments from Eq. (3.9).

The cumulant of order r can be estimated from the moments of order r and lower orders, but contains no cumulant of lower order. For the first few cumulants, the relationships are [2]:

$$\begin{aligned} c_{x(1)} &= \mu_{x(1)} \\ c_{x(2)} &= \mu_{x(2)} - \mu_{x(1)}^2 \\ c_{x(3)} &= \mu_{x(3)} - 3\mu_{x(1)}\mu_{x(2)} + 2\mu_{x(1)}^3 \\ c_{x(4)} &= \mu_{x(4)} - 4\mu_{x(3)}\mu_{x(1)} - 3\mu_{x(2)}^2 + 12\mu_{x(2)}\mu_{x(1)}^2 - 6\mu_{x(1)}^4 \end{aligned} \quad (3.12)$$

so that for centred variables, where $\mu_{x(1)} = \mu = 0$, $c_{x(1)} = \mu_{x(1)}$, $c_{x(2)} = \mu_{x(2)} = \sigma^2$, $c_{x(3)} = \mu_{x(3)}$, but $c_{x(4)} = \mu_{x(4)} - 3\mu_{x(2)}^2$, so that when normalised by σ^4 , (as in Eq. (3.7)), the normalised fourth cumulant (also called ‘kurtosis’, but represented here by K_c) is given by:

$$K_c = \mu_{x(4)} / \sigma^4 - 3 = K - 3 \quad (3.13)$$

This often causes great confusion, so when the term ‘kurtosis’ is used it should be specified if this is moment-based or cumulant-based. In the remainder of this book, the cumulant version will be used unless otherwise specified. It can be shown that all cumulants of order higher than 2 are equal to zero for Gaussian random signals, which is another advantage of cumulants, and forms the basis of a lot of ‘higher order statistics’, for example used in ‘blind source separation’ as treated in [2, 3].

A related property of cumulants applies to ‘circular’ complex variables [2], such as arise from Fourier transforms of real variables. Because of the relationship between the real and imaginary parts of such variables, the cumulant-based kurtosis is derived from the moment-based one by subtraction of 2 instead of 3, as will be seen with respect to ‘spectral kurtosis’ in Section 5.4.1 of Chapter 5.

A significant advantage of cumulants vs moments arises from the log transformation of Eq. (3.11), which converts a product to a sum. Because the pdf of a sum of independent random variables is the convolution of the individual pdfs [2, 3], this is transformed to a product, by the Fourier transformation to the first characteristic function, in Eq. (3.10), (see the detailed discussion of convolution and the convolution theorem in Section 3.2.6 below). This product thus becomes a sum in the second characteristic function of Eq. (3.11), which means that the cumulant of a sum of independent variables is the sum of the cumulants. This relationship, converting a convolution to a sum, is virtually the same as that given by the cepstrum as discussed in detail in Chapter 6.

3.2 Fourier Analysis

3.2.1 Fourier Series

The basic concept of Fourier analysis is to express signals as a summation of sinusoidal components, and with few exceptions virtually all signals can be decomposed in this way. Fourier's original analysis (now called Fourier Series) was applied to finite length signals (or periodic signals, as the resulting solution is periodic with the finite length as period). In machine vibration analysis it is used primarily for periodic signals, as produced by a machine rotating at constant speed. Thus, for any periodic signal $g(t)$ of period T for which $g(t) = g(t + nT)$, where n is any integer, it can be shown that:

$$g(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(k\omega_0 t) + \sum_{k=1}^{\infty} b_k \sin(k\omega_0 t) \quad (3.14)$$

where ω_0 is the fundamental angular frequency in rad/s ($= 2\pi / T$). The fundamental frequency in Hz (f_0) equals $1/T$. The coefficients of the cosine and sine terms can be obtained by correlating the latter with, $g(t)$ as follows:

$$a_k = \frac{2}{T} \int_{-T/2}^{T/2} g(t) \cos(k\omega_0 t) dt \quad (3.15)$$

$$b_k = \frac{2}{T} \int_{-T/2}^{T/2} g(t) \sin(k\omega_0 t) dt \quad (3.16)$$

For a given periodic signal, the division into sine and cosine components depends on an arbitrary assignment of zero time, but the total component at frequency ω_k ($= k\omega_0$) is given by:

$$a_k \cos(\omega_k t) + b_k \sin(\omega_k t) \quad (3.17)$$

which can alternatively be written as:

$$C_k \cos(\omega_k t + \phi_k) \quad (3.18)$$

where $C_k = \sqrt{a_k^2 + b_k^2}$ and $\phi_k = \tan^{-1}(b_k/a_k)$. This makes it clearer that the sinusoid has a constant amplitude, with the phase angle being that existing at the arbitrarily defined time zero. A different time zero would only affect the initial phase ϕ_k .

Expression (3.18) can again be represented as:

$$\frac{C_k}{2} \left\{ \exp[j(\omega_k t + \phi_k)] + \exp[-j(\omega_k t + \phi_k)] \right\} \quad (3.19)$$

which can be interpreted as two rotating vectors, each of length $C_k/2$, one rotating at angular frequency ω_k with initial phase ϕ_k and the other rotating at angular frequency $-\omega_k$ with initial phase $-\phi_k$, as illustrated in Figure 3.4 [4].

Using this interpretation of Fourier analysis as representing $g(t)$ as a sum of rotating vectors leads to the alternative version of Eq. (3.14) as:

$$g(t) = \sum_{k=-\infty}^{\infty} A_k \exp(j\omega_k t) \quad (3.20)$$

where the coefficients A_k are now complex and incorporate the phase shift in the form

$$A_k = \frac{C_k}{2} \exp(j\phi_k) \quad (3.21)$$

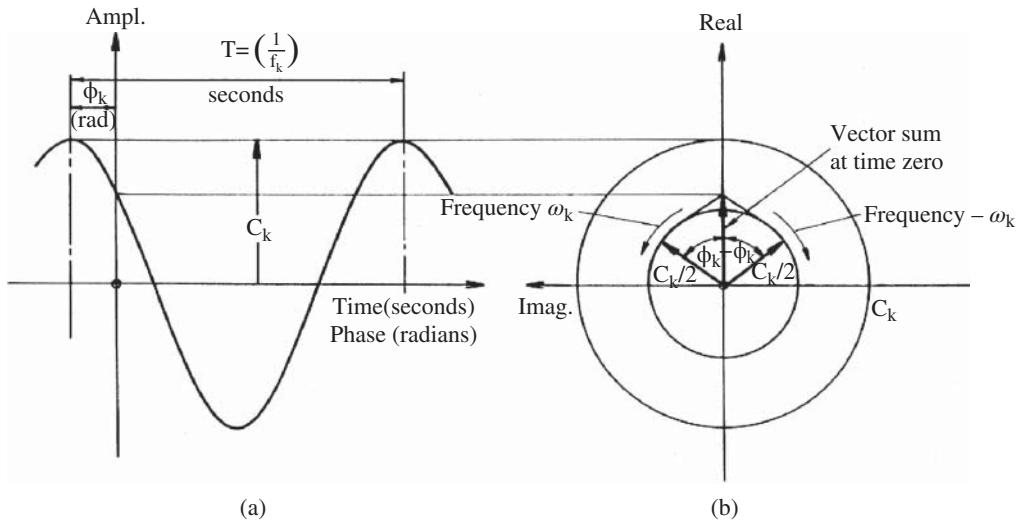


Figure 3.4 Representation of a sinusoid as a sum of two rotating vectors.

The equation for calculating the coefficients A_k (equivalent to Eq. (3.15) and (3.16)) now becomes:

$$A_k = \frac{1}{T} \int_{-T/2}^{T/2} g(t) \exp(-j\omega_k t) dt \quad (3.22)$$

This has the simple physical interpretation that multiplication by $\exp(-j\omega_k t)$ subtracts angular frequency ω_k from each component, meaning that the one originally rotating at ω_k is stopped in the position it had at time zero (this then being extracted by the integral) while all other components still rotate at some other multiple of ω_k (either positive or negative) and thus integrate to zero over the periodic time. Thus each frequency component A_k represents the position (and value) of the rotating vector at time zero, so that to obtain its position at any other time t it is necessary to cause it to rotate at angular frequency ω_k by multiplying by $\exp(j\omega_k t)$. Summing then over all frequency components gives Eq. (3.20). Note that Figure 3.4b directly shows the spectrum components (the values at time zero).

Before leaving this interpretation of Fourier series (FS) analysis as a sum of rotating vectors each of amplitude half that of the corresponding sinusoid (i.e. $C_k/2$), but having a two-sided spectrum in that each positive frequency component is accompanied by its complex conjugate at negative frequency, it can be seen that the same result can be achieved by retaining the positive frequency components only, but doubling their length to C_k and then taking the projection of each vector on the real axis. This is illustrated in Figure 3.5. A signal with a one-sided spectrum like this is complex (and known as an ‘analytic signal’) but it will later be shown that the projection on the imaginary axis is the Hilbert transform of the real part. This applies equally to the vector sum of a number of components at different frequencies as to each individual component.

3.2.2 Fourier Integral Transform

Signals other than periodic can also be expressed as a sum of complex exponentials, in particular transients, to which the so-called Fourier transform applies. The Fourier transform can be derived

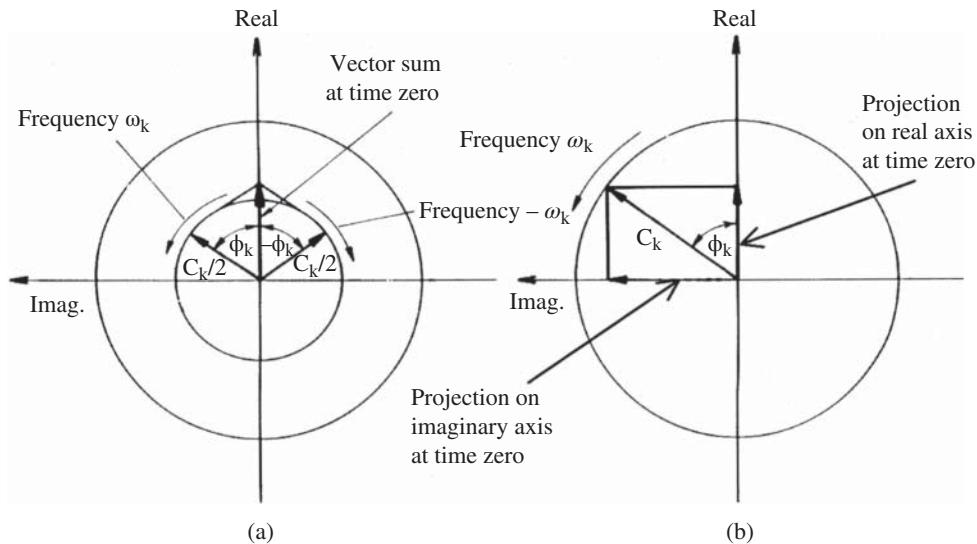


Figure 3.5 Equivalence of the vector sum of positive and negative frequency components, and projection on the real axis of a positive frequency component.

from the Fourier series by allowing the periodic time to tend to infinity and at the same time removing the division by T because transients have finite energy rather than finite power. Eqs. (3.22) and (3.20) then become:

$$G(f) = \int_{-\infty}^{\infty} g(t) \exp(-j2\pi ft) dt \quad (3.23)$$

$$g(t) = \int_{-\infty}^{\infty} G(f) \exp(j2\pi ft) df \quad (3.24)$$

respectively, where angular frequency ω_k in rad/s has been replaced by the continuous frequency function f expressed in Hz. Equations (3.23) and (3.24) are known as the forward and inverse Fourier (integral) transforms, respectively.

Note that the Fourier transform of Eq. (3.23) is related to the Laplace transform, defined by:

$$G(s) = \int_0^{\infty} g(t) \exp(-st) dt \quad (3.25)$$

where s is a complex variable, which can be represented as $\sigma + j\omega$ in terms of its real and imaginary parts. Thus, for impulse response functions (IRFs), which are necessarily causal (i.e. they do not exist for negative time) their Fourier transform, known as the frequency response function (FRF) is equal to their Laplace transform (transfer function) for the special case that s is restricted to the imaginary axis ($s = j\omega = j2\pi f$).

The forward and inverse transforms of Eqs. (3.23) and (3.24) are almost symmetrical, the only difference being the sign of the exponent. This means that results which apply to the forward transform most often also apply to the inverse transform, for example the convolution theorem to be discussed below. This is particularly the case for real, even functions, for which it makes no difference if time or frequency run forwards or backwards.

3.2.3 Sampled Time Signals

All signals which are to be processed digitally have to be digitised or discretely sampled. As will be seen in Figure 3.6c, this is the inverse case of the Fourier series (Figure 3.6b), where the spectrum is discretely sampled, and the symmetry of the Fourier transform means that the spectrum of a sampled time signal is periodic. The corresponding versions of the forward and inverse transforms are:

$$G(f) = \sum_{n=-\infty}^{\infty} g(t_n) \exp(-j 2\pi f t_n) \quad (3.26)$$

$$g(t_n) = \frac{1}{f_s} \int_{-f_s/2}^{f_s/2} G(f) \exp(j 2\pi f t_n) dt \quad (3.27)$$

where $t_n = n\Delta t = n/f_s$

3.2.4 The Discrete Fourier Transform (DFT)

The sampled time signals in Section 3.2.3 are in principle of infinite length, but when the record length is finite, this leads to the same situation as with the Fourier series in that the spectrum is discrete and the time record implicitly periodic. As seen in Figure 3.6d, this leads to a combination of the cases of Figure 3.6b,c so that both the time record and frequency spectrum are discretely sampled and periodic. The continuous infinite integrals of the Fourier transform become finite sums, usually expressed as:

$$G(k) = 1/N \sum_{n=0}^{N-1} g(n) \exp(-j 2\pi k n / N) \quad (3.28)$$

$$g(n) = \sum_{k=0}^{N-1} G(k) \exp(j 2\pi k n / N) \quad (3.29)$$

This version, known as the discrete Fourier transform (DFT), corresponds most closely to the Fourier series in that the forward transform is divided by the length of record N to give correctly scaled Fourier series components. If the DFT is used with other types of signals, e.g. transients or stationary random signals, the scaling must be adjusted accordingly as discussed below. Note that with the very popular signal processing package Matlab®, the division by N is done in the inverse transform. This means that the forward transform is more similar to the Fourier integral, but it then requires scaling in every case, as it must be multiplied by the discrete equivalent of dt , i.e. the sample interval Δt , to scale and correctly dimension the integral.

Note also that for convenience, the time and frequency zero positions have been shifted from the centre of the record to the beginning, but because of the implicit periodicity this just means that the second half of each record represents the negative axes (see Figure 3.6d).

The forward DFT operation can be understood as the matrix multiplication:

$$\mathbf{G}_k = \frac{1}{N} \mathbf{W}_{kn} \mathbf{g}_n \quad (3.30)$$

where \mathbf{G}_k represents the vector of N frequency components, the $G(k)$ of Eqs. (3.28) and (3.29), while \mathbf{g}_n represents the N time samples $g(n)$. \mathbf{W}_{kn} represents a square matrix of unit vectors $\exp(-j 2\pi k n / N)$ with angular orientation depending on the frequency index k (the rows) and time sample index n (the columns). This is illustrated graphically in Figure 3.7.

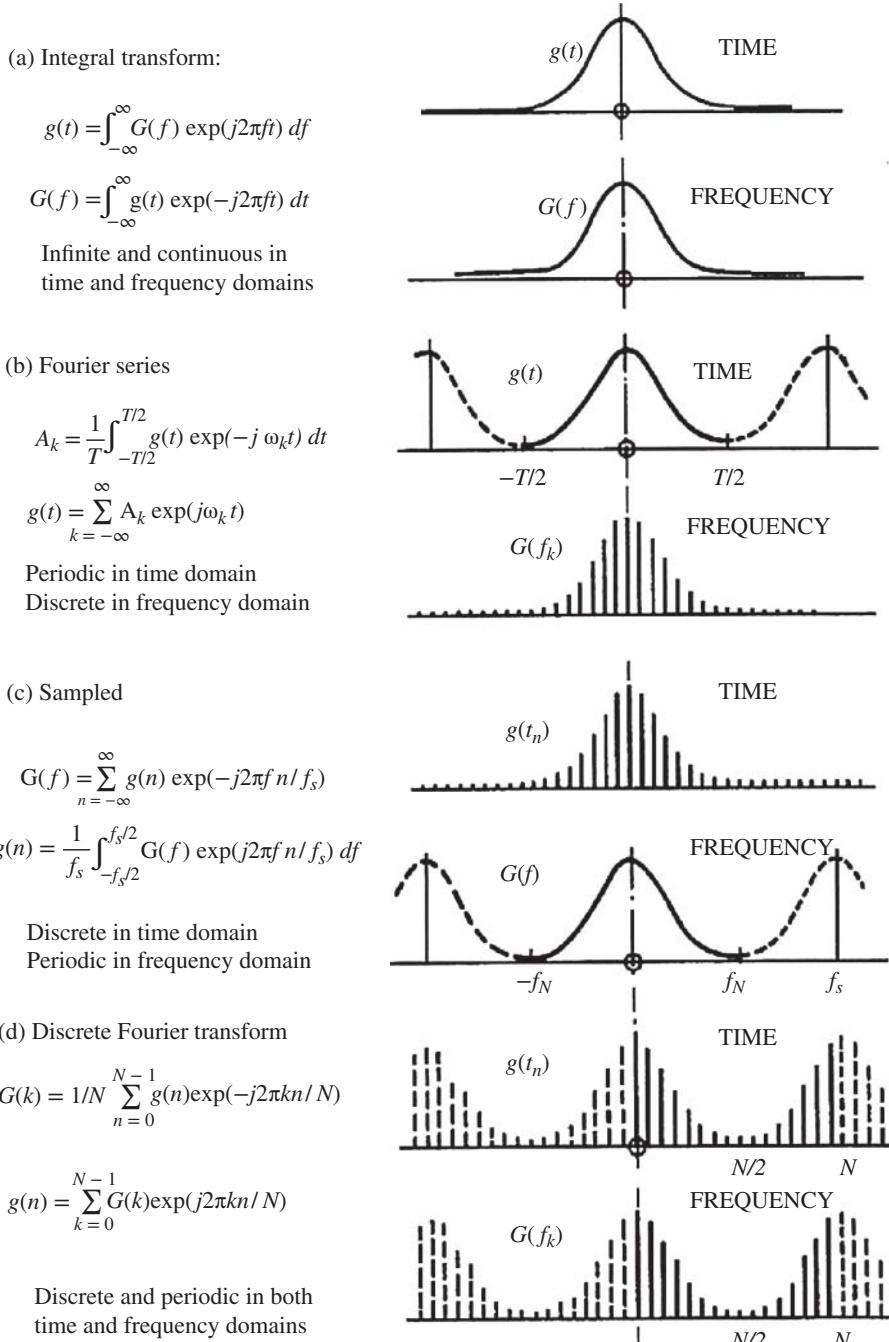


Figure 3.6 Various forms of the Fourier transform (a) Fourier integral transform (b) Fourier series (c) Sampled functions (d) Discrete Fourier transform.

$$\begin{Bmatrix} G_0 \\ G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \\ G_6 \\ G_7 \end{Bmatrix} = \frac{1}{8} \begin{bmatrix} \uparrow & \uparrow \\ \uparrow & \nearrow & \rightarrow & \nwarrow & \downarrow & \nearrow & \leftarrow & \nwarrow \\ \uparrow & \rightarrow & \downarrow & \leftarrow & \uparrow & \rightarrow & \downarrow & \leftarrow \\ \uparrow & \nwarrow & \leftarrow & \rightarrow & \downarrow & \nwarrow & \rightarrow & \leftarrow \\ \uparrow & \downarrow & \uparrow & \downarrow & \uparrow & \downarrow & \uparrow & \downarrow \\ \uparrow & \nearrow & \rightarrow & \nwarrow & \downarrow & \nearrow & \leftarrow & \nwarrow \\ \uparrow & \leftarrow & \downarrow & \rightarrow & \uparrow & \leftarrow & \downarrow & \rightarrow \\ \uparrow & \nwarrow & \leftarrow & \rightarrow & \downarrow & \nwarrow & \rightarrow & \leftarrow \end{bmatrix} \begin{Bmatrix} g_0 \\ g_1 \\ g_2 \\ g_3 \\ g_4 \\ g_5 \\ g_6 \\ g_7 \end{Bmatrix}$$

Real
Imag

Figure 3.7 Matrix representation of the DFT (note the rotated real and imaginary axes).

For $k = 0$ the zero frequency value $G(0)$ is simply the mean value of the time samples $g(n)$ as would be expected. For $k = 1$ the unit vector rotates $-1/N^{\text{th}}$ of a revolution for each time sample increment, resulting in one complete (negative) revolution after N samples. For higher values of k the rotation speed is proportionally higher. For $k = N/2$ (half the sampling frequency, the so-called ‘Nyquist frequency’) the vector turns through $-\pi$ for each time sample, but it is not possible to see in which direction it has turned. For $k > N/2$ the vector turns through more than π (in the negative direction) but is more easily interpreted as having turned through less than π (in the opposite direction) and thus if the time signal has been lowpass filtered at half the sampling frequency (as should always be the case) the second half of \mathbf{G}_k will contain the negative frequency components ranging from minus one-half the Nyquist frequency to just below zero.

3.2.5 The Fast Fourier Transform (FFT)

The fast Fourier transform (FFT) [5] is simply a very efficient algorithm for calculating the DFT Eqs. (3.28) and (3.29). Starting with the matrix version (3.30), in the simplest form (the so-called radix 2 algorithm), the FFT is based on N being a power of 2, and factorises a modified version of the matrix \mathbf{W}_{kn} into $\log_2 N$ matrices each with the property that multiplication by them only requires N complex operations as compared with the N^2 operations required for direct multiplication by \mathbf{W}_{kn} . Thus the total number of complex operations is reduced from N^2 to $N \log_2 N$, a saving by a factor of more than 100 for the typical case where $N = 1024 (= 2^{10})$.

The modified version (here called Matrix \mathbf{B}) of \mathbf{W}_{kn} (here called Matrix \mathbf{A}) has the rows arranged in ‘bit-reversed order’ as illustrated in Figure 3.8. This means that the most significant bit is indexed rather than the least significant, and so the phase jumps, for example, go from coarse to fine with an increasing row number. Multiplication by \mathbf{B} means that the results are also in bit-reversed order, but reshuffling to the correct address is a simple operation (which can be done pair-wise) taking negligible time compared with the multiplications. Figure 3.9 shows for $N = 8$ (and $\log_2 N = 3$), how matrix \mathbf{B} can be factorised into 3 matrices, \mathbf{X} , \mathbf{Y} , \mathbf{Z} , in each row of which there are only two nonzero elements, one of which is unity. Thus, multiplication by each factor matrix requires only N complex multiplications and additions. From the pattern of the factor matrices it is obvious how the decomposition can be extended to higher powers of 2. The factor matrices contain progressively finer rotations (in this case $1/2$, $1/4$, $1/8$ rotations), and the top left partitioned sub-matrix is always of

Row number in B									Row number in A
0 0 0 (0)	↑	↑	↑	↑	↑	↑	↑	↑	0 0 0 (0)
0 0 1 (1)	↑	↓	↑	↓	↑	↓	↑	↓	1 0 0 (4)
0 1 0 (2)	↑	→	↓	←	↑	→	↓	←	0 1 0 (2)
0 1 1 (3)	↑	←	↓	→	↑	←	↓	→	1 1 0 (6)
1 0 0 (4)	↑	↗	→	↖	↓	↗	←	↖	0 0 1 (1)
1 0 1 (5)	↑	↖	→	↗	↓	↖	←	↗	1 0 1 (5)
1 1 0 (6)	↑	↖	←	↗	↓	↖	←	↗	0 1 1 (3)
1 1 1 (7)	↑	↖	←	↗	↓	↖	←	↗	1 1 1 (7)

Figure 3.8 Modified matrix \mathbf{B} , with rows shifted to bit-reversed address.

$$\begin{array}{c}
 \left[\begin{array}{ccccccc} \uparrow & \uparrow & 0 & 0 & 0 & 0 & 0 \\ \downarrow & \downarrow & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \downarrow & \rightarrow & 0 & 0 & 0 \\ 0 & 0 & \uparrow & \leftarrow & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \uparrow & \rightarrow & 0 \\ 0 & 0 & 0 & 0 & \uparrow & \leftarrow & 0 \\ 0 & 0 & 0 & 0 & 0 & \uparrow & 0 \end{array} \right] \left[\begin{array}{ccccccc} \uparrow & 0 & \uparrow & 0 & 0 & 0 & 0 \\ 0 & \uparrow & 0 & \uparrow & 0 & 0 & 0 \\ 0 & 1 & \uparrow & 0 & 0 & 0 & 0 \\ 0 & 0 & \uparrow & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \uparrow & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \uparrow & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \uparrow & 0 \end{array} \right] \left[\begin{array}{ccccccc} \uparrow & 0 & 0 & 0 & \uparrow & 0 & 0 \\ 0 & \uparrow & 0 & 0 & 0 & \uparrow & 0 \\ 0 & 0 & \uparrow & 0 & 0 & 0 & \uparrow \\ 0 & 0 & 0 & \uparrow & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \uparrow & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \uparrow & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \uparrow \end{array} \right] = \left[\begin{array}{ccccccc} \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \uparrow & \downarrow & \rightarrow & \leftarrow & \uparrow & \downarrow & \leftarrow \\ \downarrow & \uparrow & \leftarrow & \rightarrow & \downarrow & \uparrow & \rightarrow \\ \uparrow & \leftarrow & \uparrow & \rightarrow & \downarrow & \leftarrow & \uparrow \\ \downarrow & \uparrow & \rightarrow & \leftarrow & \downarrow & \uparrow & \leftarrow \\ \uparrow & \leftarrow & \uparrow & \rightarrow & \downarrow & \leftarrow & \uparrow \end{array} \right] \\
 \mathbf{X} \qquad \mathbf{Y} \qquad \mathbf{Z} \qquad = \qquad \mathbf{B}
 \end{array}$$

Figure 3.9 Matrix \mathbf{B} factorised into three factor matrices $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$.

the form:

$$\begin{bmatrix} \mathbf{I} & \mathbf{I} \\ \mathbf{I} & -\mathbf{I} \end{bmatrix}$$

Further savings can be made in special cases, such as the radix 4 and radix 8 transforms, and a similar but not so effective gain can be made for factorisation other than in powers of 2, but the main point is that the properties of the FFT are those of the DFT.

3.2.6 Convolution and the Convolution Theorem

One of the reasons for performing Fourier analysis is that it converts convolution in one domain into a multiplication in the other domain (as does the Laplace transform). Not only does this simplify the solution of many problems, but it is also useful in graphical illustrations of many relationships. Convolution is the operation by which the output (response) of a linear system is obtained from the input (forcing function) and the transfer properties of the physical system, in the time domain represented by its ‘impulse response function’. The IRF of a system is its output when excited by a unit impulse (delta function) at time zero.

3.2.6.1 Delta Functions

Mathematically, delta functions are quite complex, but here they will be treated heuristically and without full mathematical rigour, for illustrative purposes. In mechanics, the term ‘impulse’ is used to mean the integral over time of force, and over the time of application of that force, the impulse is equal to the change in momentum of the object to which the force is applied. If the force is applied over a very short time (e.g. a hammer blow) the change in momentum (and thus velocity for a fixed mass) is very sudden. A unit impulse (in mechanical terms) is defined as the limit of such an impulsive force as the time of its application tends to zero, while maintaining a constant (unit) value of impulse, and thus a constant velocity change. More generally, a unit impulse or delta function $\delta(t)$ is defined by the integral

$$\lim_{\epsilon \rightarrow 0} \int_{-\epsilon/2}^{\epsilon/2} \delta(t) dt = 1 \quad (3.31)$$

Thus, a unit (force) impulse applied at time zero gives a sudden step change in velocity at time zero, and by analogy the integral of a delta function is a step function changing from value zero to one at the origin. This requires infinite acceleration and thus force (in the mechanical application), and so is not physically realisable, but is an idealised situation that is a reasonable model when the time of application is very short with respect to the response time of the system to which it is applied.

3.2.6.2 Convolution

When a forcing function $f(t)$ is applied to a physical system, the effect between time t and $t + dt$ can be considered as an impulse of value $f(t)dt$, giving an impulse response starting at time t and scaled by the strength of the impulse (i.e. proportional to $f(t)$). The total response over time will thus be the sum of all these scaled impulse responses initiated at different times in the past, and can be represented by the Duhamel integral

$$x(t) = \int_{-\infty}^{\infty} f(\tau)h(t - \tau)d\tau \quad (3.32)$$

which is said to be the convolution of $f(t)$ and $h(t)$ and is represented symbolically as

$$x(t) = f(t) * h(t) \quad (3.33)$$

Note that the operation is commutative, so that $f(t)$ and $h(t)$ can be exchanged in (3.32).

The convolution operation is quite complex, and can be seen to consist of four stages:

- 1) One function $h(\tau)$ is reversed to $h(-\tau)$.
- 2) It is then displaced by an amount t to $h(t - \tau)$
- 3) It is then multiplied by the other function $f(\tau)$ to give $f(\tau)h(t - \tau)$
- 4) Finally, this product is integrated over the dummy variable τ to give the total output at time t .

The reason for the reversal in step 1) will be understood by reference to the application illustrated in Figure 3.10 [4]. This shows the operation of two averaging circuits, which can be used to give the running average of a signal. Here, the signal being averaged, $y(\tau)$, is the square of a signal for which the local ‘mean square value’ is sought, this being the average of the squared value over a defined time in the immediate past. The impulse response of the two averaging circuits is in this case represented by $g(\tau)$. The circuit on the left, (a)–(e), would give the running linear average over the

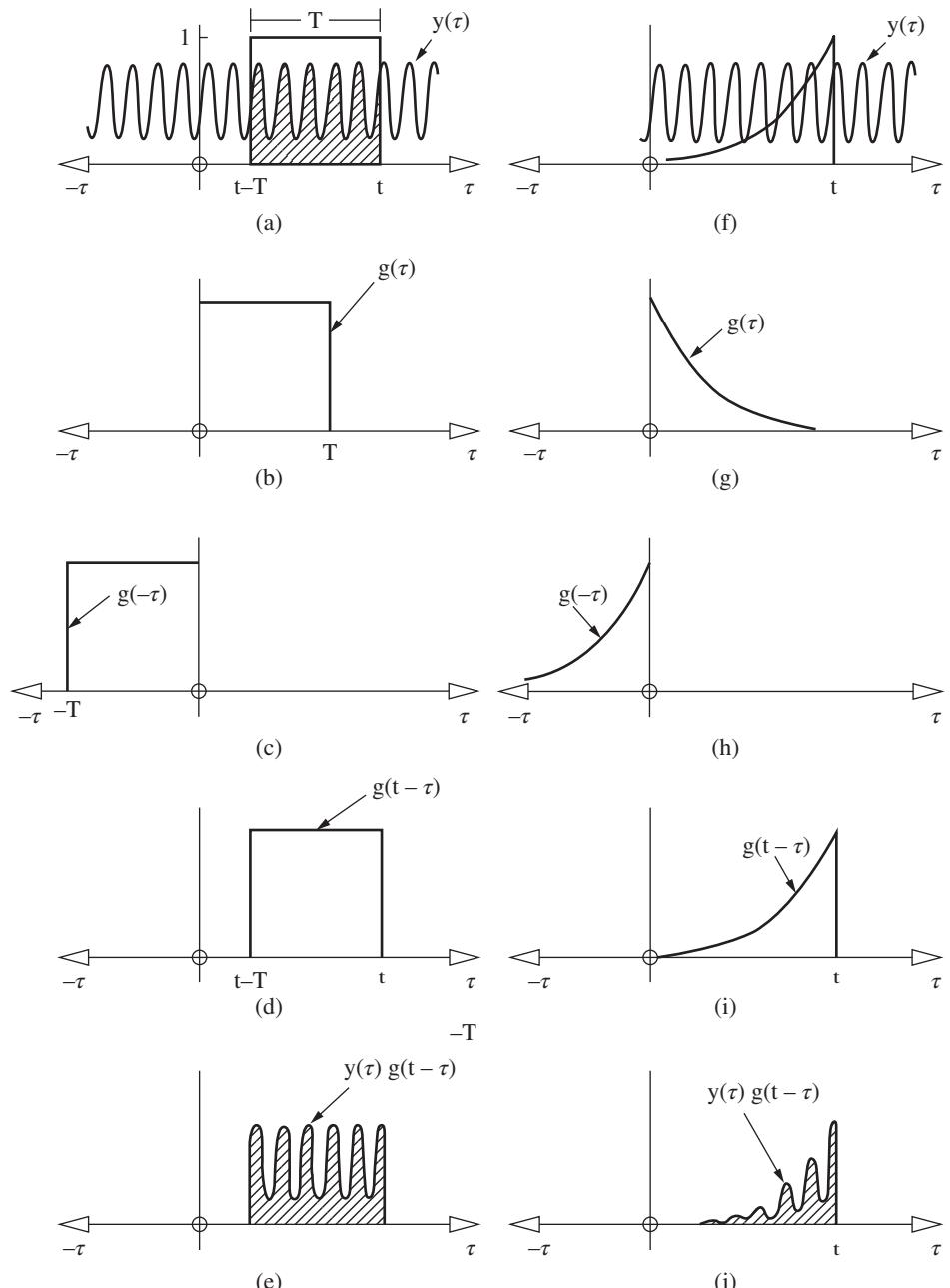


Figure 3.10 Running average as a convolution. Source: From [4]. (a)–(e) Linear weighting (f)–(j) Exponential weighting.

previous T seconds, by outputting the value of the integral $\int_{t-T}^t y(\tau)g(t-\tau)d\tau$. Scaling this result by dividing by T will evidently give the mean square value over the previous T seconds, and is seen to correspond to a scaled version of the convolution $y(t)^*g(t)$ when $g(\tau)$ is the impulse response of the averaging circuit. It is extremely difficult to produce a circuit with the impulse response of Figure 3.10b, but the circuit on the right, (f)–(j), is a simple RC smoothing circuit, with impulse response a decaying exponential function defined by:

$$\begin{aligned} g(t) &= e^{-\sigma t}, \quad t \geq 0 \\ g(t) &= 0, \quad \text{otherwise} \end{aligned} \quad (3.34)$$

The convolution represented by the integral of Figure 3.10j still gives an averaging or smoothing of the signal $y(\tau)$, but is a weighted average with most weight on the most recent value of the signal, and decreasing weight on previous values going backwards in time. This explains why the impulse response of Figure 3.10g must be reversed in the convolution operation.

As a matter of interest, the scaling of the two results (for a stationary input) will be the same when the area under the two IRFs is the same. This occurs when the peak value of the exponential function is twice that of the linear function, and when the time constant ($1/\sigma$) of the exponential function is equal to $T/2$. For this value of σ the effective averaging time of the exponential circuit is the same as for the linear circuit, in the sense that the reduction of ripple is the same for both circuits. This is because they can both be interpreted as lowpass filters, and the noise bandwidth of the filters is the same when $1/\sigma = T/2$. This is the subject of exercises in this chapter.

Note the similarity of the convolution operation of Eq. (3.32) with the autocorrelation function of Eq. (2.8) in Chapter 2. It can be seen that the autocorrelation function of a single transient, where the division by T would not be carried out (as opposed to a stationary signal where an average over time is desired) can be considered as the convolution of a function with itself reversed. This is treated in more detail in Section 3.2.6.5.

Even though the convolution operation is quite complex in general, it becomes relatively simple when one of the functions is a delta function. In this case, the result of the convolution is that the delta function is replaced by the convolving function, and scaled by its value (when it is not a unit impulse). The scaling factor can be a complex number.

Rather than proving this mathematically, it will be demonstrated by the case where the circuit is a delay line. A delay line is a circuit which delays any input signal by a fixed delay time τ . Obviously, an impulse applied at time zero will emerge at time τ and so the impulse response $h(t)$ is a delayed delta function as depicted in Figure 3.11a. A signal applied at time zero will also be delayed by τ and so as depicted in Figure 3.11a the delta function has effectively been replaced by the convolving function.

As illustrated in Figure 3.11b, a signal with echoes can be generated as the convolution of the original signal with a unit delta function at the origin and scaled delta functions at the delay times.

In Figure 3.11c it is shown how a periodic signal can be generated as the convolution of the basic signal (for one period) convolved with a train of unit delta functions spaced at the periodic time. It is shown below how this can be used to derive Fourier series from the Fourier transform of the basic signal.

3.2.6.3 The Convolution Theorem

As alluded to above, the convolution theorem states that the Fourier transform converts a convolution in one domain into a product in the other domain. A proof is given here for the forward transform (see inset), but because of the symmetry of the Fourier transform the same applies to the inverse transform.

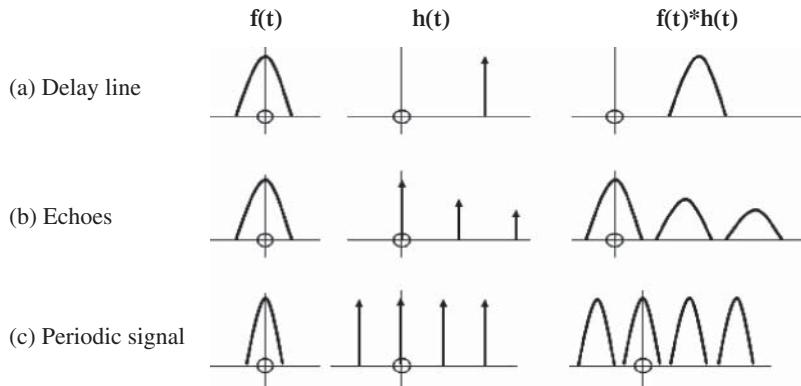


Figure 3.11 Convolution with delta functions.

The Convolution Theorem

It is to be shown that if $x(t) = f(t) * h(t)$, i.e.

$$x(t) = \int_{-\infty}^{\infty} f(\tau)h(t - \tau)d\tau \quad (a)$$

then

$$X(f) = F(f).H(f) \quad (b)$$

where the upper case variables are the Fourier transforms of the lower case variables. By applying the Fourier transform of Eq. (3.17)

$$X(f) = \int_{-\infty}^{\infty} x(t) \exp(-j2\pi ft)dt \quad (c)$$

$$F(f) = \int_{-\infty}^{\infty} f(t) \exp(-j2\pi ft)dt \quad (d)$$

$$H(f) = \int_{-\infty}^{\infty} h(t) \exp(-j2\pi ft)dt \quad (e)$$

Substituting (a) in (c) gives

$$X(f) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(\tau)h(t - \tau)d\tau \right] \exp(-j2\pi ft)dt \quad (f)$$

which by reversal of the order of integration gives

$$\begin{aligned} X(f) &= \int_{-\infty}^{\infty} f(\tau) \left[\int_{-\infty}^{\infty} h(t - \tau) \exp(-j2\pi ft)dt \right] d\tau \\ &= \int_{-\infty}^{\infty} f(\tau) \left[\int_{-\infty}^{\infty} h(u) \exp(-j2\pi f(u + \tau))du \right] d\tau \\ &= \int_{-\infty}^{\infty} f(\tau) \left[\int_{-\infty}^{\infty} h(u) \exp(-j2\pi fu)du \right] \exp(-j2\pi f\tau)d\tau \end{aligned} \quad (g)$$

$$X(f) = F(f).H(f)$$

QED (h)

Thus, a convolution in the time domain is transformed into a product in the frequency domain, but by the same token a product in the time domain is transformed into a convolution in the frequency domain.

An example of the latter is shown graphically in Figure 3.12, where the Fourier transform of the square of a cosine signal ($\text{Acos}(2\pi f_0 t)$) is derived graphically by convolving the spectrum of the cosine with itself. Note that this is done without full mathematical rigour, as the Fourier transform of Eq. (3.17) only applies strictly to transient signals, for which the integral of the modulus from $-\infty$ to ∞ is finite. For a sinusoidal signal, the Fourier transform can be taken as delta functions with values (areas) equal to the Fourier series (FS) components. In this case the FS components are equal to $A/2$ at $\pm f_0$, the frequency of the cosine. In the time domain (on the left in Figure 3.12), the square gives a raised cosine of maximum value A^2 and minimum value zero, and with double the original frequency (i.e. $2f_0$). This can be considered to be the sum of the constant (mean) value, $A^2/2$ and a cosine of frequency $2f_0$ and amplitude $A^2/2$. As shown on the right, the same result is obtained by convolving the spectrum of the original cosine with itself. Because the spectrum is symmetric, the first step of reversing the function has no effect.

In the second step, for zero displacement, the product of the two functions gives $A^2/4$ at $\pm f_0$, which when summed (integrated over all frequency) gives the value $A^2/2$ as the zero frequency value of the convolution.

As one spectrum is displaced to the right, the product is zero until a displacement of $2f_0$, when the positive frequency component of one spectrum coincides with the negative frequency component of

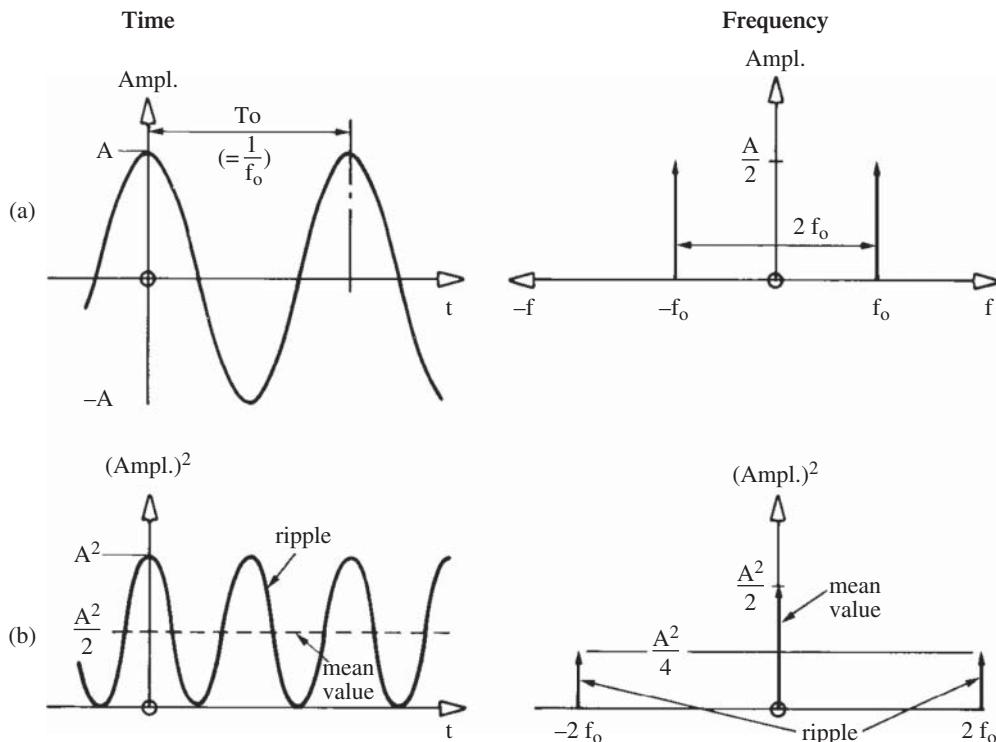


Figure 3.12 Spectrum of a squared cosine by convolution. Source: From [4].

the other. The product of the two spectra is then $A^2/4$, as is the total integral over all frequencies. The same applies for a negative displacement of one spectrum. Note that the ripple of amplitude $A^2/2$ in the time domain splits into two components each of amplitude $A^2/4$ in the frequency domain. Note also that the mean square value ($A^2/2$) of the original cosine can be obtained by lowpass filtration of the squared signal with a lowpass filter that removes the ripple component at $\pm 2f_0$.

The next example, shown in Figure 3.13, obtains the spectrum of a half cosine pulse from the spectra of a continuous cosine and a rectangular function of appropriate length (from which the half cosine pulse is obtained by multiplication). It can easily be shown (see exercises for this chapter) that the Fourier transform of a symmetric rectangular pulse of height A and length T is:

$$AT \frac{\sin(\pi f T)}{\pi f T} \quad (3.35)$$

and this is used in the figure.

It can thus be seen that the Fourier transform of the half cosine pulse $g(t)$ of length T and height A is:

$$G(f) = \frac{AT}{2} \left[\frac{\sin(\pi f T - \pi/2)}{(\pi f T - \pi/2)} + \frac{\sin(\pi f T + \pi/2)}{(\pi f T + \pi/2)} \right] \quad (3.36)$$

with a peak value of $2AT/\pi$ at $f=0$, the first zero crossing at $1.5/T$ and subsequent zero crossings spaced at $1/T$.

Another useful example of the application of the convolution theorem is the generation of the FRF of a single degree-of-freedom (SDOF) system (e.g. a mass/spring/damper system) as the Fourier transform of the IRF. It is well-known that the latter is an exponentially damped sinewave starting at time zero, the frequency of the sinewave being the damped natural frequency of the system. Thus:

$$h(t) = e^{-\sigma t} \sin 2\pi f_d t, \quad t \geq 0 \quad (3.37)$$

which is seen to be the product of the damped exponential function $f(t)$ with a continuous sinewave $g(t)$, as depicted in Figure 3.14.

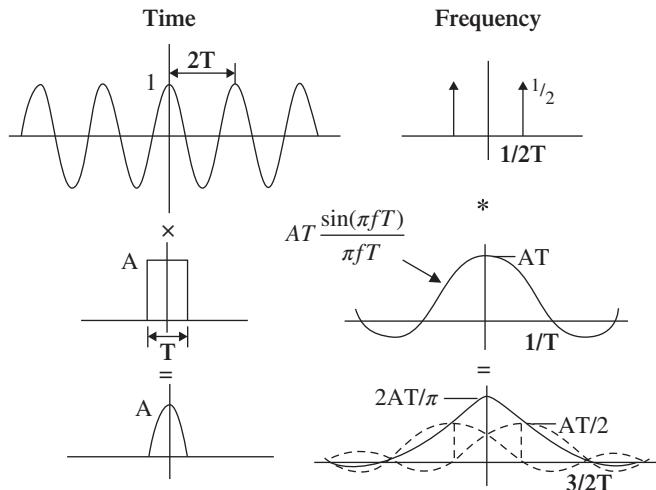


Figure 3.13 Spectrum of a half cosine pulse by convolution.

A convenient way to obtain the Fourier transform of $e^{-\sigma t}$ is to make use of its Laplace transform $\frac{1}{s+\sigma}$ (from any table of Laplace transforms) and substitute $j\omega$ for s . Thus

$$F(f) = \frac{1}{\sigma + j2\pi f} \quad (3.38)$$

In Figure 3.14 it is seen that $F(f)$ has maximum amplitude $1/\sigma$ at zero frequency and decreases to zero as $f \rightarrow \pm\infty$. The -3 dB points, where the amplitude is reduced by the factor $1/\sqrt{2}$, occur when $f = \pm\sigma/2\pi$. At zero frequency, the FRF is real and positive, and thus has zero phase. As $f \rightarrow \infty$, the phase tends to that of $1/j$, i.e. $-\pi/2$, and similarly as $f \rightarrow -\infty$, the phase tends to $+\pi/2$.

In Figure 3.14 it is seen that the spectrum of the sinewave $g(t)$ consists of two delta functions, one of value $-j/2$ at $+f_d$ and one of value $+j/2$ at $-f_d$. The convolution to obtain the Fourier transform of the product in the time domain thus consists in replacing these two delta functions with the scaled version of $F(f)$. This is one case where the scaling of the delta functions is a complex number.

Note that the resulting complex spectrum is compatible with the representations shown in Figure 3.15 ([6]) for the FRF of a SDOF system (keeping in mind the orientation of the real and imaginary axes), including the Nyquist plot (real vs imaginary) which is a circle.

3.2.6.4 Fourier Series from the Fourier Transform

As shown in Figure 3.11c, any periodic function can be generated by convolving one period (of length T) with an infinite train of unit delta functions with spacing T . Since the Fourier transform (actually Fourier series) of such a train of delta functions is another train of delta functions with

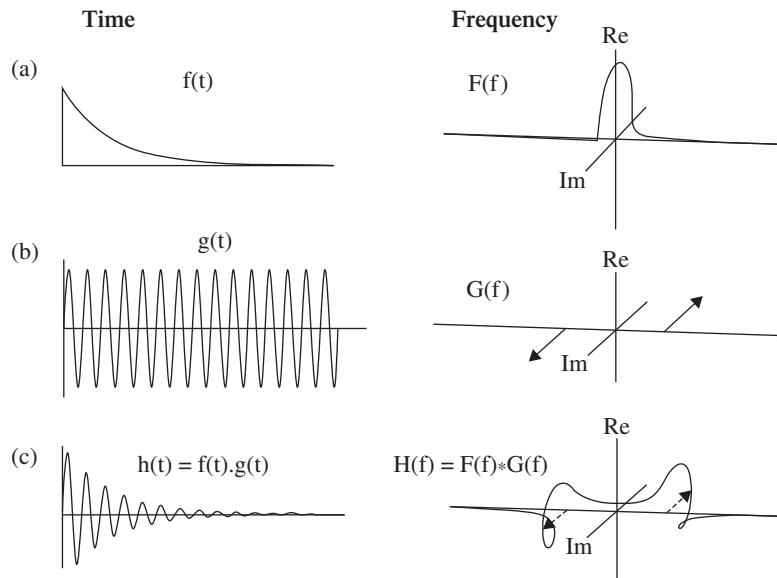


Figure 3.14 FRF of a SDOF system by convolution.

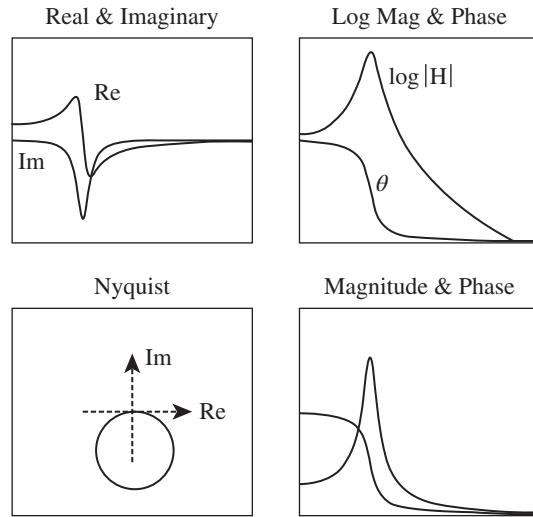
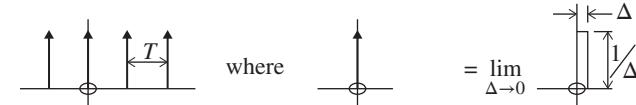


Figure 3.15 Representations of the FRF of a SDOF system.

scaling $1/T$ and spacing $1/T$ (see inset), the spectrum of the periodically repeated transients is the product of the spectrum of the single transient with the scaled train of delta functions, thus sampling it at intervals of $1/T$, and multiplying these values by $1/T$. Thus, the Fourier series spectrum for a transient $g(t)$ with Fourier transform $G(f)$, repeated with a period of T , can be calculated as:

Fourier Series for a Train of Unit Delta Functions



$$\begin{aligned}
 G(k) &= \lim_{\Delta \rightarrow 0} \frac{1}{T} \int_0^T g(t) \exp(-j2\pi f_k t) dt \\
 &= \lim_{\Delta \rightarrow 0} \frac{1}{T} \int_0^{\Delta} \frac{1}{\Delta} \exp(-j2\pi f_k t) dt \\
 &= \lim_{\Delta \rightarrow 0} \frac{1}{T} \left(\frac{1}{-j2\pi f_k \Delta} \right) [\exp(-j2\pi f_k t)]_0^\Delta \\
 &= \lim_{\Delta \rightarrow 0} \frac{1}{T} \left(\frac{1}{-j2\pi f_k \Delta} \right) \left[1 + (-j2\pi f_k \Delta) + \frac{(-j2\pi f_k \Delta)^2}{2!} + \dots - 1 \right] \\
 &= \frac{1}{T}
 \end{aligned}$$

that is, a train of harmonics of value $1/T$ independent of frequency

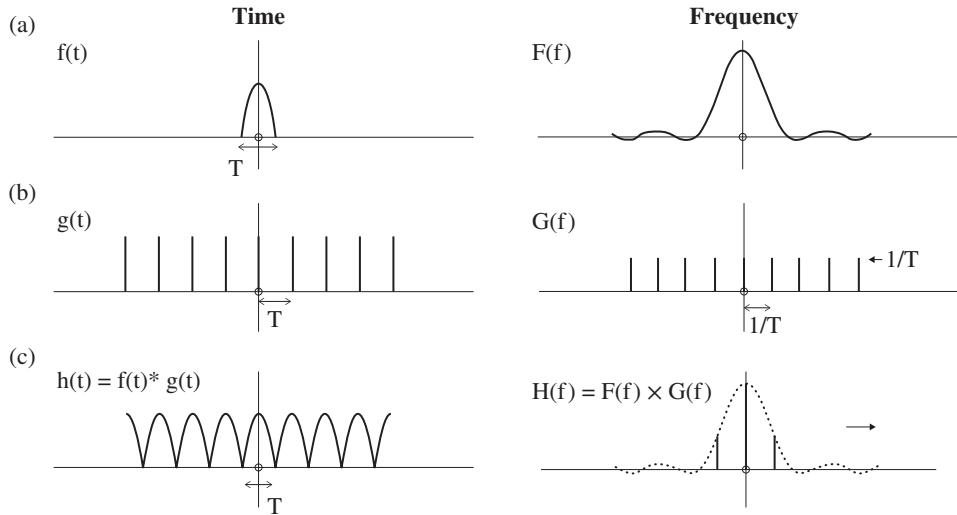


Figure 3.16 Fourier series of a half-wave rectified cosine from the Fourier transform of a half-cosine pulse.

$$G(f_k) = \frac{1}{T} G\left(\frac{k}{T}\right) = \frac{1}{T} \int_{-T/2}^{T/2} g(t) e^{-j2\pi kt/T} dt \quad (3.39)$$

This is illustrated in Figure 3.16 for a full-wave rectified cosine $h(t)$, which can be modelled as the convolution of a half-cosine pulse $f(t)$ of length T with a train of unit delta functions $g(t)$ of spacing T . The spectrum of the half cosine pulse, $F(f)$, is generated in Figure 3.13 and given by Eq. (3.36).

Note that to generate the Fourier series for a half-wave rectified cosine, exactly the same procedure can be used except that the spacing of the unit delta functions becomes $2T$, and the Fourier series components have spacing $1/2T$ and value $1/2T$. Because most of the resulting additional Fourier series components of the half-wave rectified cosine coincide with zero crossings, the only extra components are at frequencies $\pm 1/2T$.

3.2.6.5 Relationship Between Autocorrelation and Autospectrum

As mentioned in Section 3.2.6.2, the autocorrelation function corresponds to the convolution of a signal with itself reversed in time, at least for a single transient. For a single rotating vector, reversal in time reverses the phase angle, resulting in the complex conjugate, and so its spectrum (value at time zero) is also the complex conjugate. This applies for all frequency components, so if $X(f) = \Im\{x(t)\}$ then

$$\Im\{x(-t)\} = X^*(f) \quad (3.40)$$

Thus, for a single realisation, having finite energy as opposed to power, the autocorrelation function can be defined as:

$$\begin{aligned} R_{xx}(t) &= \int_{-\infty}^{\infty} x(\tau)x(t+\tau)d\tau \\ &= x(t)^*x(-t) \end{aligned} \quad (3.41)$$

It thus follows that the Fourier transform of the autocorrelation function is the product of the spectrum with its complex conjugate, this being the autospectrum of the transient.

$$\Im\{R_{xx}(t)\} = X(f)X^*(f) = |X(f)|^2 \quad (3.42)$$

This relationship can be generalised, and it has been shown (the Wiener-Khinchin relations) that for ergodic stationary random signals the Fourier transform of the autocorrelation function (defined as Eqs. (2.5) to (2.8)) is equal to the power spectrum. For an elegant proof of this, reference can be made to [7].

Figure 3.17 makes use of the convolution theorem to derive the autocorrelation functions for band-limited noise and narrow band noise. In both cases, it is seen that the effective correlation length is of the order of $1/B$, where B is the total bandwidth in frequency. Use is made of this elsewhere in the book.

3.2.7 Zoom FFT

The basic DFT transform of Eq. (3.28) extends in frequency from zero to the Nyquist frequency and has a resolution equal to the sampling frequency f_s divided by the number of samples N . Sometimes it is desired to analyse in more detail in a limited part of the frequency range, in which case use can be made of so-called ‘zoom analysis’.

Since resolution $\Delta f = f_s/N$, the two ways to improve it are:

Increase the length of record N . Some analysers include this option in the form of ‘non-destructive zoom’ [4], which makes use of an algorithm to perform a transform of size $m \times N$ by combining the results of m undersampled transforms of size N . This was useful when hardware restrictions limited the size of transform which could be performed, but in modern analysers, and in signal processing packages such as Matlab there is virtually no restriction on

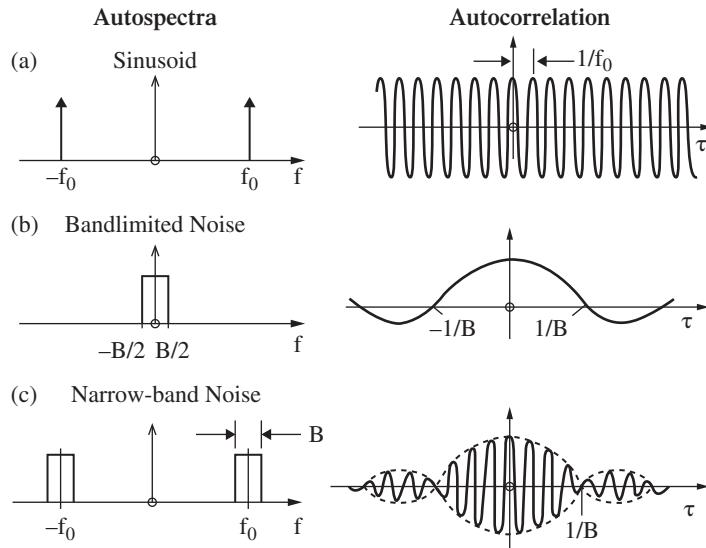


Figure 3.17 Autocorrelation vs autospectrum for three signals. Note that spectrum of (c) is the convolution of (a) and (b).

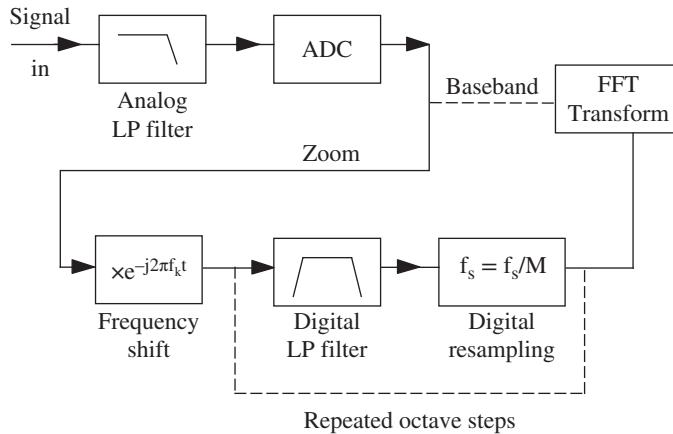


Figure 3.18 Schematic diagram of FFT zoom process.

transform size, and so zoom can be achieved by performing a large transform and then viewing only part of the result.

Reduce the sampling frequency f_s . This can be done if the centre of the desired zoom band is shifted to zero frequency so that the zoom band around the centre frequency can be isolated by a lowpass filtration. The highest frequency is then half the zoom band and the sampling frequency can be reduced accordingly without aliasing problems. This process is illustrated in Figure 3.18. The lowpass filtering and resampling process is usually done in octave (2 : 1) steps, as a digital filter will always remove the highest octave, relative to the sampling frequency, and halving the sampling frequency simply means discarding every second sample. This type of zoom is normally done in real-time by a specialised hardware processor, the advantage being that the sampling rate is reduced before signals have to be stored. The zoom process is a useful precursor to demodulation, even where the further processing is to be done in a computer, and this is discussed further in Section 3.3. Note that the time signal output from the zoom processor is complex, as the corresponding spectrum is not conjugate even.

3.2.8 Practical FFT Analysis

3.2.8.1 Pitfalls of the FFT Process

The so-called pitfalls of the FFT are all properties of the DFT and result from the three stages in passing from the Fourier integral transform to the DFT. The first step is digitisation of the time signal which can give rise to ‘aliasing’; the second step is truncation of the record to a finite length, which can give rise to ‘leakage’ or ‘window effects’, while the third results from discretely sampling the spectrum, which can give rise to the ‘picket fence effect’. Figure 3.19 shows these three steps graphically, using the convolution theorem [4].

In Figure 3.19a–c the infinite continuous time signal is sampled as in Figure 3.6c producing a periodic spectrum with a period equal to the sampling frequency f_s . It can be seen that if the original signal contains any components outside the range $\pm f_N$, where f_N is the ‘Nyquist frequency’ or half the sampling frequency, then these will overlap with the true components giving ‘aliasing’ (higher frequencies represented as lower ones). Once aliasing is introduced it cannot be removed, so it is

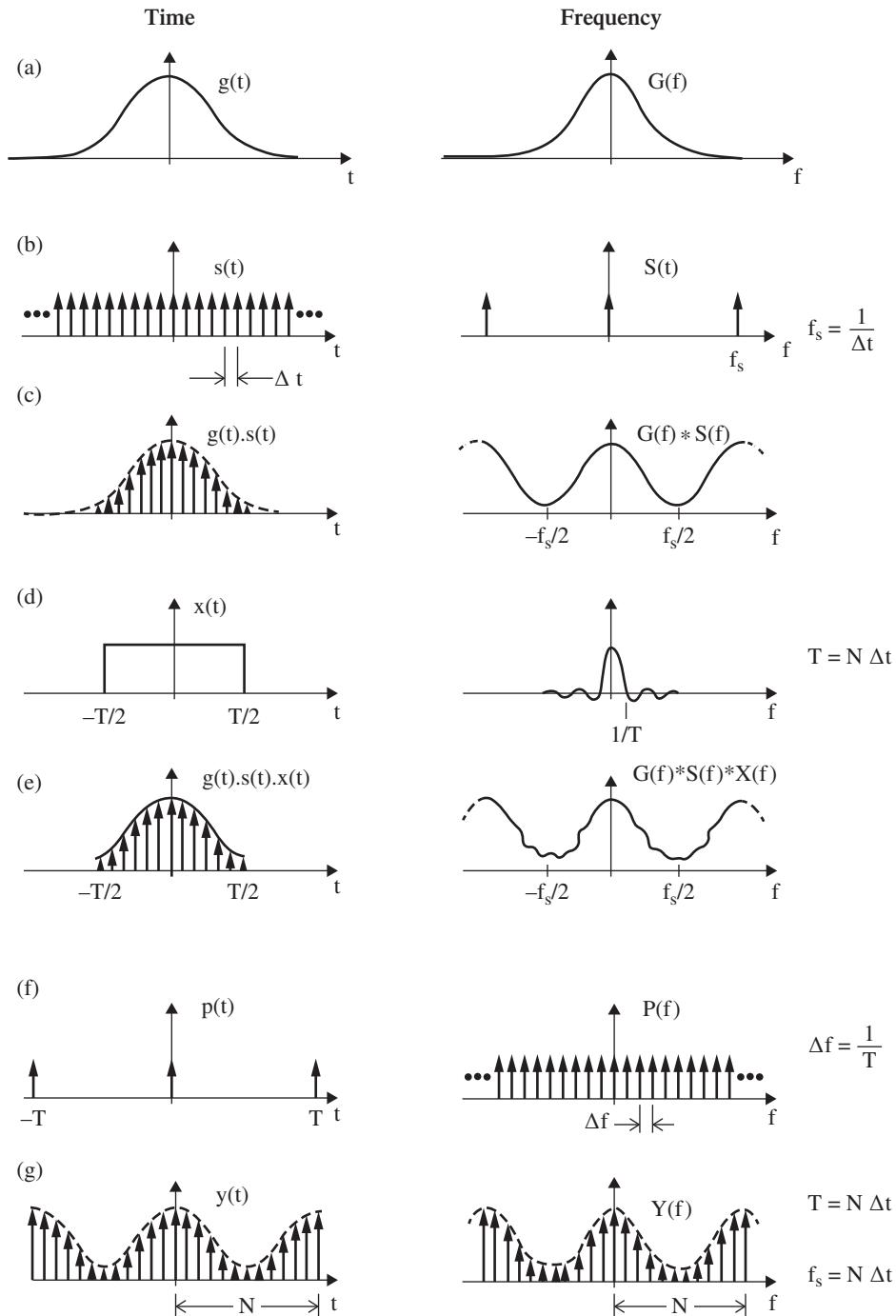


Figure 3.19 Three steps in passing from the FT to the DFT (a-c) Time sampling (d, e) Truncation (f, g) Frequency sampling.

important to use appropriate analogue lowpass filters before digitising any time signal for processing. After initial correct digitisation, digital lowpass filters can be used to permit resampling at a lower sampling rate. In Figure 3.19d,e the signal is truncated to length T by multiplying it by a finite (rectangular) window. The spectrum is thus convolved with the Fourier transform of the window, which acts as a filter characteristic. Energy at a single frequency is spread into adjacent frequencies in the form of this characteristic, hence the term ‘leakage’. Finally, in Figure 3.19f,g the continuous spectrum is discretely sampled in the frequency domain, which corresponds in the time domain to convolution with a train of delta functions of spacing T , making the time signal periodic. The spectrum is not necessarily sampled at peaks; hence the term ‘picket fence effect’; it is as though the spectrum is viewed through the slits in a picket fence.

To avoid aliasing it is virtually always necessary to use an antialiasing filter with a very steep roll-off. It has become fairly standard to use filters with a roll-off of 120 dB/octave, allowing approximately 80% of the calculated spectrum to be used. Thus, with a 1K (1024 point) transform, spectrum line number 512 is at the Nyquist frequency, and higher frequencies fold back towards the measurement range. Line number 624 folds back into the top of the desired measurement range (line number 400), and is only 64% of an octave above it, and so is attenuated by 77 dB, taking it below the typical dynamic range. The antialiasing filters typically result in considerable distortion of the time signal, and are thus usually not included in digital oscilloscopes (which thus should not be used for digitisation of signals for further processing). Figure 3.20 (from [8]) illustrates how antialiasing filtering corrects the spectrum while distorting the time signal, and vice versa.

The effects of leakage are influenced by the final spectrum sampling, and Figure 3.21 illustrates that for a rectangular window (whose FT is a $\frac{\sin(x)}{x}$ or sinc(x) function) if the window contains an integer number of periods of a sinusoid, even though each spectral line is associated with a sinc function, these are sampled at the zeros and are thus not apparent. On the other hand, in the worst case of a residual half period, the effective filter characteristic is very poor. Use is made of this

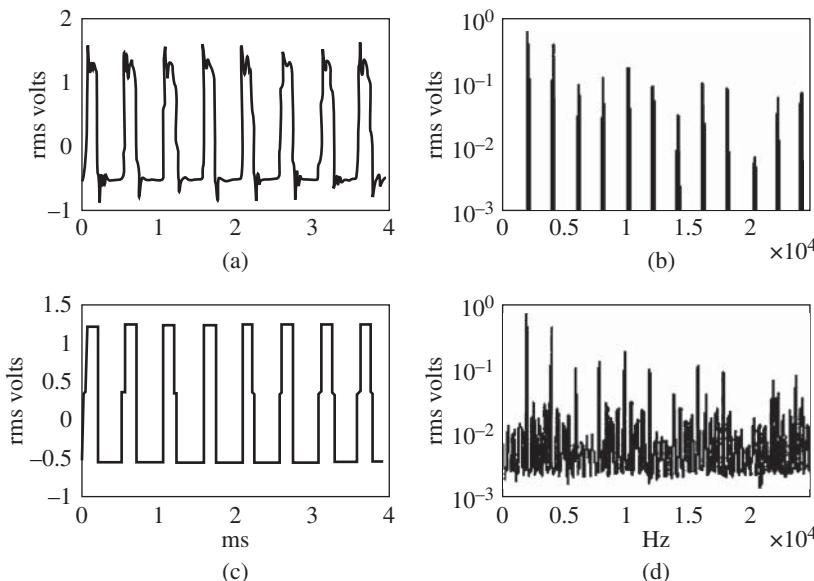


Figure 3.20 (a) Time signal correctly lowpass filtered (b) Spectrum of (a) (c) Time signal without lowpass filtration (d) Spectrum of (c) (note aliasing).

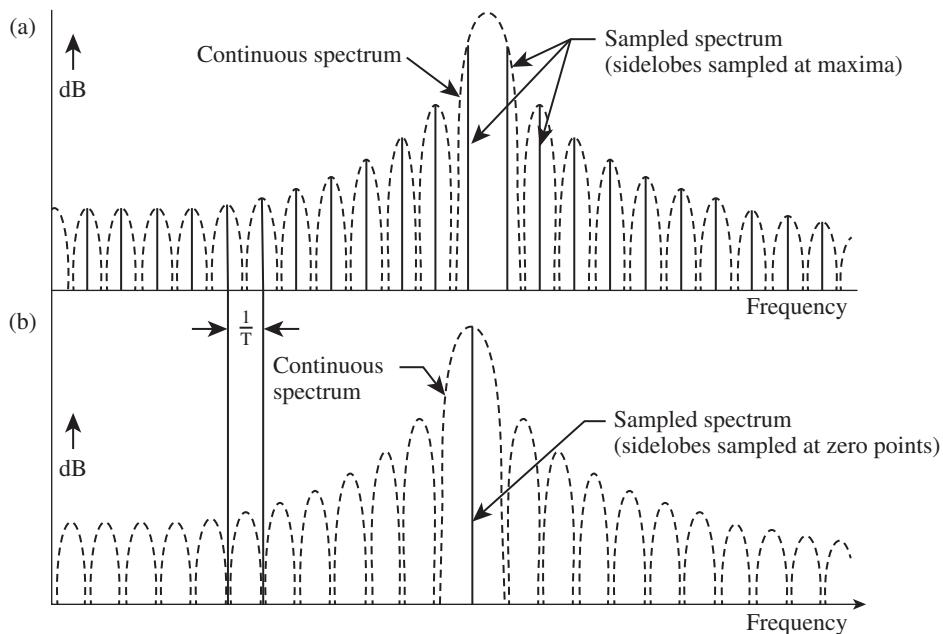


Figure 3.21 How frequency sampling affects the apparent filter characteristic of a window. (a) Extra half period in record; (b) Integer number of periods in record (Courtesy Brüel & Kjær)

phenomenon in ‘order tracking’ of machines (Section 5.1), where the signal sampling is synchronised with machine speed and it can be arranged that there is an integer number of periods of (all harmonics of) the rotational frequency within the record length, in which case a rectangular window can be used. Otherwise, for continuous signals it is usually necessary to choose a data window other than rectangular to achieve a better filter characteristic. This is discussed below.

Because of the picket fence effect, spectral functions are not necessarily sampled at their peaks and the ‘picket fence error’ is the difference between the true value and the value of the maximum spectral line. For rectangular weighting this can be as much as 3.9 dB, and most other windows have a reduced value.

3.2.8.2 Data Windows

3.2.8.2.1 Continuous Signals

For continuous signals, a major function of the window is to reduce the effect of the discontinuity which usually arises when a random section of signal is made periodic. Practically, it means minimising the sidelobes in the filter characteristic, both the highest and the remaining ones (by maximising their rate of rolloff). To improve enhancement of discrete frequency components with respect to broadband noise it is desirable to minimise the noise bandwidth of the characteristic, but on the other hand, attention must also be paid to minimising the picket fence effect. Table 3.1 gives a comparison of the properties of the most common windows applied to stationary signals, and Figure 3.22 (from [9]) compares their worst case filter characteristics. The Hanning window, which can be considered as one period of a sine squared function, is a good general purpose window, with picket fence effect limited to 1.4 dB, noise bandwidth 1.5 (times Δf the line spacing) and desirable characteristics

Table 3.1 Properties of various windows.

Window	Noise bandwidth	Highest sidelobe (dB)	Sidelobe rolloff (dB/decade)	Picket fence effect (dB)
Rectangular	1.0	-13	20	3.9
Hanning	1.5	-33	60	1.4
Kaiser-Bessel	1.8	-60	20	0.8
Flat top	3.8	-70	20	< 0.1

with respect to overlap averaging, to be discussed below. The best window with respect to separating adjacent components of widely differing levels is probably the Kaiser-Bessel, but the same can usually be achieved by simply analysing with more resolution (zoom analysis or a larger transform). The flat top window is specifically designed to minimise the picket fence effect, and is thus usually the best choice when calibrating measurements with a calibration signal whose frequency can fall anywhere between two analysis lines. It can also be useful when analysing a signal dominated by one or more families of harmonics, since as long as they are resolved (keeping in mind that the noise bandwidth is $3.8 \Delta f$) there is no need to compensate the indicated values of the various harmonics.

Figure 3.23 shows how compensation can be made for both picket fence error and frequency error when using a Hanning window. Provided frequencies are stable along the record length, the difference in dB between the two highest samples around a frequency peak (ΔdB) determines the errors. As mentioned above, the Hanning window has the desirable property that the effective weighting in overlap averaging (see later) can be made completely uniform with an overlap of 2/3, 3/4, etc. With 50% overlap, the weighting varies by 2 : 1, but this is not a problem with stationary signals. When finding the averaged spectrum of a long transient signal (longer than the transform size), a uniform weighting is preferable.

3.2.8.2.2 Transient Signals

Analysis of transient signals is common in impact measurements for modal analysis. The force signal is always short, and a rectangular window is suitable, although this may be tailored to just more than the length of the force pulse in order to exclude noise. For the response signal it must be ensured that it has died away (i.e. by 50 to 60 dB) by the end of the record. This can be achieved by extending the record length (i.e. by zoom or a larger transform size), but where this is restricted by other constraints it is common to apply an exponential window, starting just before the response signal, which attenuates the signal sufficiently. This is equivalent to applying additional damping, which is known very precisely, and so can be subtracted from the resulting measurements. A short taper, typically of a half-Hanning shape, can be added to both the leading and trailing edges of a transient window and to the leading edge of an exponential window, to make the transitions less abrupt.

3.2.8.3 Application in the Frequency Domain

Since multiplication by a window function corresponds in the frequency domain to a convolution with its Fourier transform (FT), this is sometimes the most efficient way to apply them. Examples of where this is advantageous are firstly where the FT of the window is very simple, such as with the Hanning function, secondly where only a part of the spectrum is required, as with zoom spectra, and thirdly where several different windows can be applied to exactly the same FT. The basic principle can be explained using the Hanning window, which when repeated periodically (as happens

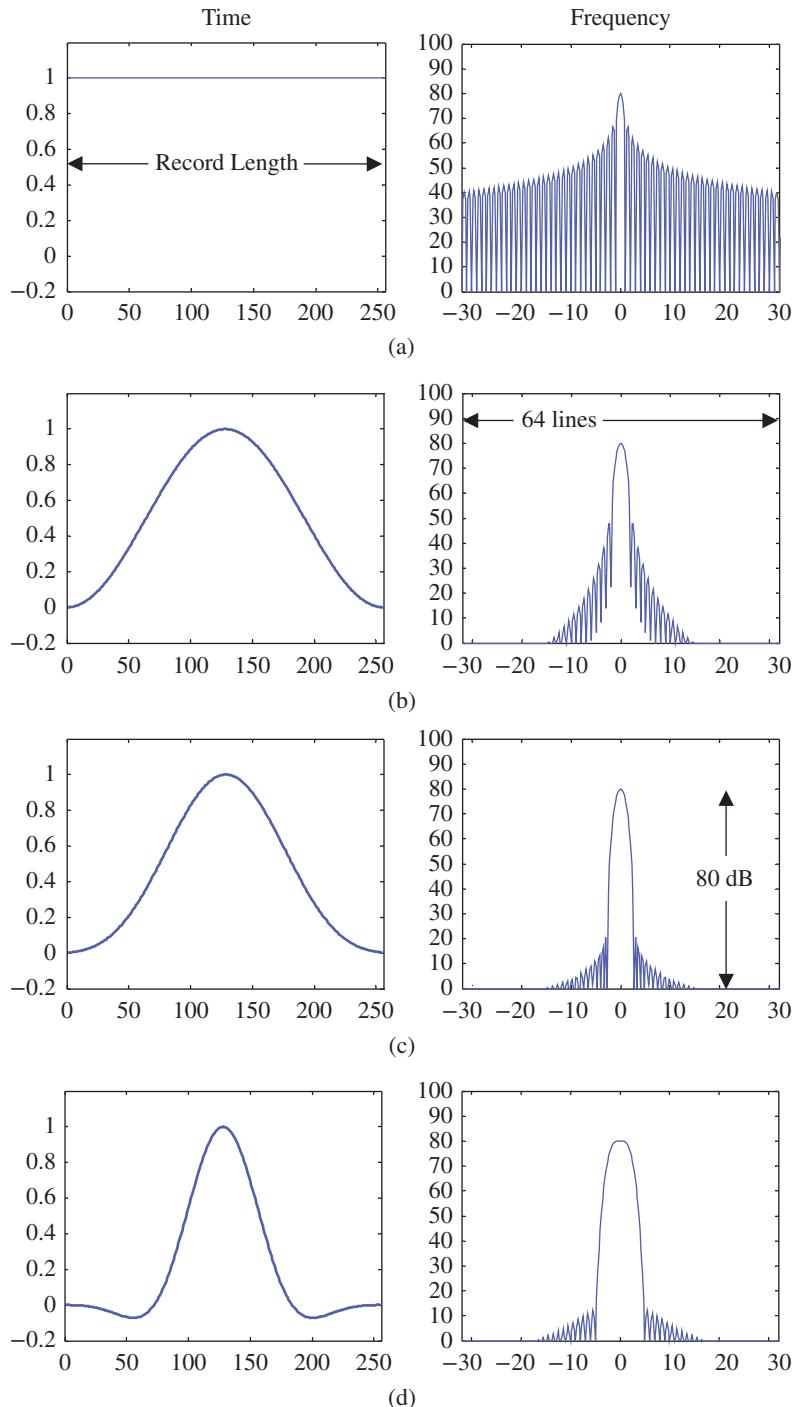


Figure 3.22 Data windows for continuous signals ([9]). (a) Rectangular (b) Hanning (c) Kaiser-Bessel (d) Flat top.

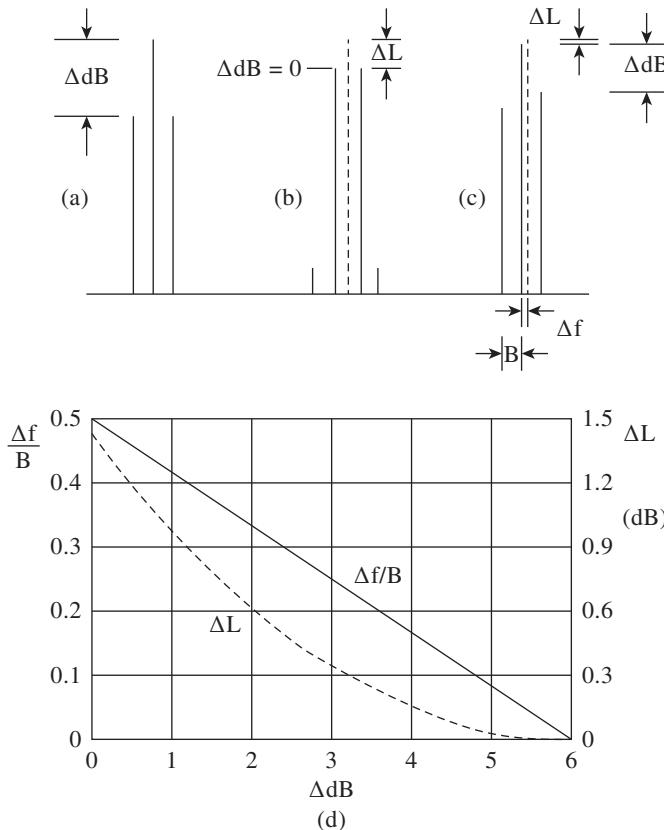


Figure 3.23 Compensation for the picket fence effect with a Hanning window ([4]). ΔdB = difference between two highest spectrum samples. ΔL = picket fence error. Δf = frequency error (a), (b), (c) Minimum, maximum, and intermediate error cases (d) Error nomogram.

implicitly with the DFT) can be represented as $\sin^2 \theta$ or $\frac{1}{2} - \frac{1}{2} \cos 2\theta$ which has the convolution coefficients $[-\frac{1}{4}, \frac{1}{2}, -\frac{1}{4}]$ (keeping in mind that the frequency corresponds to one period along the record length, and thus to one line spacing). Convolution with such a simple function is often more efficient than direct multiplication in the time domain, in particular with binary arithmetic, where the multiplications by the coefficients correspond to a lateral shift. Note that the coefficients as stated are for a window with maximum value one, and are usually modified to scale the result (see below under ‘Scaling’).

3.2.8.4 Spectrum Averaging

The need for averaging of FFT spectra is determined by whether the signal contains random components or not. Averaging should always be done in terms of signal power (i.e. amplitude squared) as it is this which is conserved independent of phase. The DFT spectra of discrete frequency components always have the same amplitude, and therefore little is achieved by averaging the squared amplitudes, although this can be useful for clarifying which components are discrete frequency and

which are random. Where analysis is being done for diagnostic purposes (e.g. a zoom spectrum to measure a single frequency or family of sidebands very accurately) it is preferable not to average, as in practice machine speeds vary slightly with time and averaging results in a smearing of frequencies and in particular disguises frequency spacings.

When meaningful spectra are to be obtained from random signals, which in mechanical signals are typically caused by fluid flow (turbulence, cavitation), road roughness etc., it is necessary to average a number of power spectrum estimates. The number of averages required is determined by the desired accuracy, as the relative standard deviation of the result (for Gaussian signals) is given by [4]:

$$\epsilon = \frac{1}{2\sqrt{n}} \quad (3.43)$$

where n is the number of independent averages. Thus, for $n = 16$, $\epsilon = 12.5\%$ or 1 dB, meaning that there is a 68% probability that the result will be within ± 1 dB, 95% probability that it will be within ± 2 dB and 99.7% probability that it will be within ± 3 dB. To halve the error it is necessary to make four times as many averages, etc.

With a rectangular window, ‘independent’ means non-overlapping, but with other windows such as Hanning, advantage can be gained by overlapping, as information is lost near the two ends where the weighting is near zero. In fact, very little is lost statistically by overlapping 50%, and so this is recommended for stationary random signals, since twice as many effective averages can be obtained from a given length of signal. As mentioned above, the overall weighting is not uniform in that case, as at the point where the successive Hanning windows overlap, their amplitude weighting is $\frac{1}{2}$ and thus their power weighting $\frac{1}{4}$, meaning that the final weighting of that part of the signal in the overlap average is $\frac{1}{2}$ and the total power weighting along the signal varies between $\frac{1}{2}$ and unity. This gives no problem for stationary signals, but to extract all information from a given length of record, in particular if it is non-stationary, it is advisable to overlap by a factor of at least $\frac{2}{3}$ although with typical FFT record lengths in powers of 2 it is often simpler to overlap by $\frac{3}{4}$. In the latter case the effective number of averages to insert in Eq. (3.43) is half the actual number.

3.2.8.5 Spectrum Scaling

3.2.8.5.1 Discrete Frequency Signals

As stated above, the DFT operations of Eqs. (3.28) and (3.29) result in a correctly scaled Fourier series spectrum for the forward transform and exact reconstitution of a (periodic) time signal for the inverse transform. Note that this means that the resulting value A_k at the positive frequency ω_k has an amplitude half that of the corresponding sinewave (C_k). To obtain the equivalent RMS (root mean square) value, $|A_k|$ must be multiplied by $\sqrt{2}$ to take account of the power in the negative frequency component. The RMS value is also equal to $C_k/\sqrt{2}$. If the signal contains discrete frequency components which do not have an integer number of periods along the record length, then a window function such as Hanning will generally be used to reduce leakage. However, if the window is scaled to a maximum value of unity, the average ‘power’ (i.e. mean square value) of the signal will obviously be reduced, and it is necessary to compensate for this. From Section 3.2.8.3 it will be seen that the convolution coefficients for a Hanning window scaled to a maximum value of 2 are $[-\frac{1}{2}, 1, -\frac{1}{2}]$ meaning that a single component in one line would be replaced by three components of which the central one would have the same value (i.e. the peak value would be scaled correctly). This is usually the best scaling to use for discrete frequency components as it means that the maximum value around a spectral peak can be read off directly after correction for picket fence error (e.g. using Figure 3.23). For windows other than Hanning, the same effect can be achieved by scaling the window such that

its central convolution coefficient is unity. Note that because of the extraneous sidebands introduced, the total ‘power’ in the spectrum has been increased as discussed below.

At this stage it is convenient to introduce ‘Parseval’s theorem’ which in broad terms states that the total power (or energy) in a signal can be obtained by integrating over all time or all frequency, and in both domains is related to amplitude squared. For a stationary signal with finite power, the frequency spectrum will either contain discrete frequency components whose amplitude squared directly represents the power at each frequency, or for random signals the squared amplitude spectrum is continuously distributed over frequency and represents ‘power spectral density’ (PSD), which has to be integrated over a finite bandwidth to give finite power. In both cases the equivalent ‘power’ in the time domain is the mean square value, obtained by integrating the instantaneous squared value (instantaneous power) over a sufficiently long time and dividing by that time. For transient signals with finite ‘energy’ (integral of ‘power’ over time) the squared amplitude of its Fourier transform represents ‘energy spectral density’ (ESD) which when integrated over all frequency gives the same total energy as integrating the instantaneous power of the signal over all time.

Henceforth, the terms power and energy will be used without inverted commas to represent signal amplitude squared and its time integral, respectively, as these are generally related to physical power and energy by an impedance or admittance function (e.g. electrical power = $I^2R = V^2/R$ in terms of current I , voltage V , and resistance R). Thus for a signal with units U (where U represents m, ms^{-1} , ms^{-2} , g, N, etc.) power has units U^2 , energy has units $U^2\text{s}$, PSD has units U^2/Hz ($= U^2\text{s}$), ESD has units $U^2\text{s}/\text{Hz}$ ($= U^2\text{s}^2$).

3.2.8.5.2 Stationary Random Signals

Each signal record transformed will be treated by the DFT algorithm as a periodic signal, but the power in each spectral line can be assumed to represent the integral of the PSD over the frequency band of width $\Delta f (= 1/T)$, and thus the average PSD is obtained by multiplying the squared amplitude by T . The required averaging over a number of records does not change this scaling. How well the average PSD represents the actual PSD depends on the width of peaks (and valleys) in the spectrum. The width of such a peak is typically determined by the damping associated with a structural resonance excited by the broadband random signal, and the 3 dB bandwidth is given by twice the value of σ (of Eqs. (3.37) and (3.38)) expressed in Hz. The PSD will be sufficiently accurate if the 3 dB bandwidth is a minimum of five analysis lines.

If a window such as Hanning has been used to reduce leakage, and if it is scaled so as to read the peak value of discrete frequency components (as recommended above in this section) the calculated PSD value will have to be divided by the ‘noise bandwidth’ indicated in Table 3.1 to compensate for the extra power given by the spectral sidebands. This noise bandwidth is the sum of the squares of the convolution coefficients (for Hanning it is $0.5^2 + 1^2 + 0.5^2 = 1.5$). When integrating over several frequency lines (e.g. to convert a constant bandwidth spectrum to constant percentage bandwidth, see Section 3.4) the total power in each integrated band must be divided by the noise bandwidth of the window because of the extra power associated with each line; this is the same as integrating the PSD over the required bandwidth.

Note that discrete frequency components cannot be represented (except as delta functions) on a PSD scale as they are concentrated in an infinitely narrow bandwidth and thus have infinite PSD. Note also that their power is independent of the analysis bandwidth Δf used to analyse them, whereas the power of spectral lines of random signals varies directly with the analysis bandwidth. Zoom analysis is therefore sometimes used to make discrete frequency components stand out from random background noise.

3.2.8.5.3 Transient Signals

Transient signals are also treated as being one period of a periodic signal, so not only does the power in a spectral line have to be converted to an average spectral density by dividing by Δf , but also the average power must be converted to energy per period by a further multiplication by T , altogether a multiplication by T^2 to obtain a result scaled as ESD. Generally, transient signals will be shorter than the transform length and thus a rectangular window will be used, and if the signal has decayed to near zero at the end of the record (e.g. following the recommendations of Section 3.2.8.2) the signal bandwidth will be sufficiently greater than the analysis bandwidth for the average ESD to represent the true ESD. The extra damping given by an exponential window will genuinely give a reduction in signal energy.

3.3 Hilbert Transform and Demodulation

3.3.1 Hilbert Transform

The Hilbert transform can be said to be the relationship between the real and imaginary parts of the Fourier transform of a one-sided function [10]. For example, any IRF is causal and thus one-sided in the time domain, and this means that the real and imaginary parts of the corresponding frequency function (e.g. that shown in Figure 3.15) are related by a Hilbert transform. That there should be a relationship becomes evident when it is considered that a causal function is made up of even and odd components which are identical for positive time, and thus cancel for negative time, as shown in Figure 3.24. Thus:

$$x(t) = x_e(t) + x_o(t) \quad (3.44)$$

and

$$x_e(t) = x_o(t) \times \text{sgn}(t) \quad (3.45)$$

where $\text{sgn}(t)$ is the sign function. Since the even part of a time function transforms to the real part of its Fourier transform, and the odd part to the imaginary part [4], by applying the convolution theorem to Eq. (3.45) it can be seen that:

$$X_e(f) = X_o(f) * \Im\{\text{sgn}(t)\} \quad (3.46)$$

The Fourier transform of the sign function is the imaginary hyperbolic function $1/j\pi f$ so that the final expression relating the real part of the Fourier transform ($X_R(f) = X_e(f)$) to the imaginary part ($X_I(f) = X_o(f)/j$), and writing out the convolution in full, is given by:

$$X_R(f) = \frac{1}{\pi} \int_{-\infty}^{\infty} X_I(\phi) \cdot \frac{1}{(f - \phi)} d\phi \quad (3.47)$$

The equivalent equation for the Hilbert transformation of a time function $x(t)$ is:

$$\tilde{x}(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} x(\tau) \frac{1}{(t - \tau)} d\tau \quad (3.48)$$

Taking the Fourier transform of Eq. (3.48) gives:

$$\tilde{X}(f) = X(f) \cdot (-j\text{sgn}(f)) \quad (3.49)$$

which shows that a Hilbert transform can be achieved more simply by transforming into the frequency domain, shifting the phase of positive frequency components by $-\pi/2$ and of negative frequency components by $+\pi/2$, and then transforming back to the time domain.

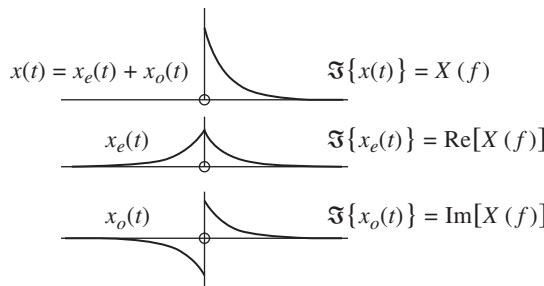


Figure 3.24 Decomposition of a causal signal into even and odd components, and relationships in the time and frequency domains.

A one-sided frequency spectrum can similarly be divided into conjugate even and conjugate odd components which transform by the inverse Fourier transform to real and imaginary time signals, respectively, which are related by a Hilbert transform. The sum of these two components is known as an ‘analytic signal’, which can be formed from a given real time signal by adding j times its Hilbert transform. Alternatively, it can be obtained more simply by transforming the real time signal into the frequency domain, obtaining the equivalent one-sided spectrum by multiplying by $2H(f)$, where $H(f)$ is the Heaviside or unit step function, and transforming back to the time domain. This is also a very efficient way of performing a Hilbert transform.

As a corollary, it is worth pointing out that when working with a signal processing package such as Matlab, modifying frequency spectra and then transforming back to the time domain, it is not necessary to adjust all negative frequency components in the same way (but complex conjugate) as the positive frequency components; it is much simpler to multiply the positive frequency components by 2 (but not the zero frequency or Nyquist frequency components), set the second half of the spectrum (the negative frequency components) to zero, perform an inverse transform to an analytic signal and simply take the real part (this last operation is usually necessary even when working with 2-sided spectra, as the program does not know that the answer is supposed to be real, and will usually calculate a very small imaginary part). This process is illustrated in Figure 3.25.

3.3.2 Demodulation

Modulation occurs when an otherwise sinusoidal signal, a so-called carrier signal, has its amplitude or frequency made to vary with time. In the first case it is known as amplitude modulation, and in the second it can be considered as a frequency or phase modulation. Phase modulation is the deviation in phase (angular displacement) from the linearly increasing phase of the carrier, while frequency modulation is the difference in instantaneous frequency (angular velocity) from the constant carrier frequency. Thus, frequency modulation is the derivative of phase modulation. A direct mechanical example of phase/frequency modulation is shaft torsional vibration, as discussed in Section 2.3.2, which when expressed in terms of shaft angle is a phase modulation, and when in terms of shaft speed is a frequency modulation. There is no modulation term for the angular acceleration obtained by further differentiation. A mechanical example of amplitude modulation is the variation in vibration amplitude at the meshing frequency in a gearbox, as the increase in tooth deflection with load gives an increasing departure from ideal involute profiles, and often tooth load varies periodically with the rotation of the gears (see Figure 2.13 of Chapter 2).

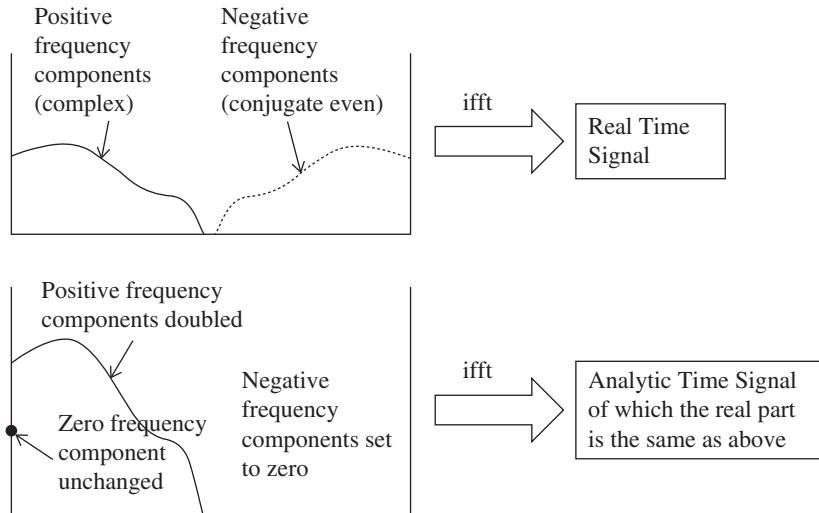


Figure 3.25 Manipulation of the positive frequency spectrum to obtain a real time signal.

The signals produced by faults in rolling element bearings are a series of high frequency bursts as resonance frequencies are excited by near periodic impacts, as discussed in Section 2.2.3. The diagnostic information is contained in the repetition frequency, not in the resonance frequencies excited, but spectra obtained by direct Fourier analysis are dominated by the latter, and the important information is disguised by smearing of the high order harmonics. Such signals can be modelled as an amplitude modulation of a carrier signal at the resonance frequency by a near periodic series of exponential pulses (though in general there will also be a jump in phase at the start of each new pulse). In so-called ‘envelope analysis’, as discussed in Section 7.3, the signal envelope is extracted by amplitude demodulation, and frequency analysed to reveal the repetition frequencies even when these have a small random fluctuation.

Thus, a generally modulated signal can be represented by:

$$A_m(t) \cos(2\pi f_c t + \phi_m(t)) \quad (3.50)$$

where $A_m(t)$ represents the amplitude modulation function (plus DC offset to ensure that $A_m(t)$ is always positive), and $\phi_m(t)$ represents the phase modulation function in radians.

The corresponding frequency modulating function (in Hz) is $\frac{1}{2\pi} \frac{d\phi_m(t)}{dt}$.

Expression (3.50) will be seen to be the real part of the rotating vector:

$$A_m(t) \exp\{j(2\pi f_c t + \phi_m(t))\} \quad (3.51)$$

whose modulus is the amplitude modulating function plus DC offset, and whose phase is the phase modulating function plus the linear carrier component. Thus, if it is desired to demodulate a real signal such as (3.50), it is desirable to find the corresponding imaginary part so as to form the complex expression (3.51).

These relationships can be interpreted graphically by reference to Figure 3.5, where a sinusoid was obtained from an analytic signal with just one (positive) frequency component. If the amplitude of this rotating vector is allowed to vary with time, for example sinusoidally at a lower frequency than the rotation, its projection on the real axis will be amplitude modulated, as will its projection on

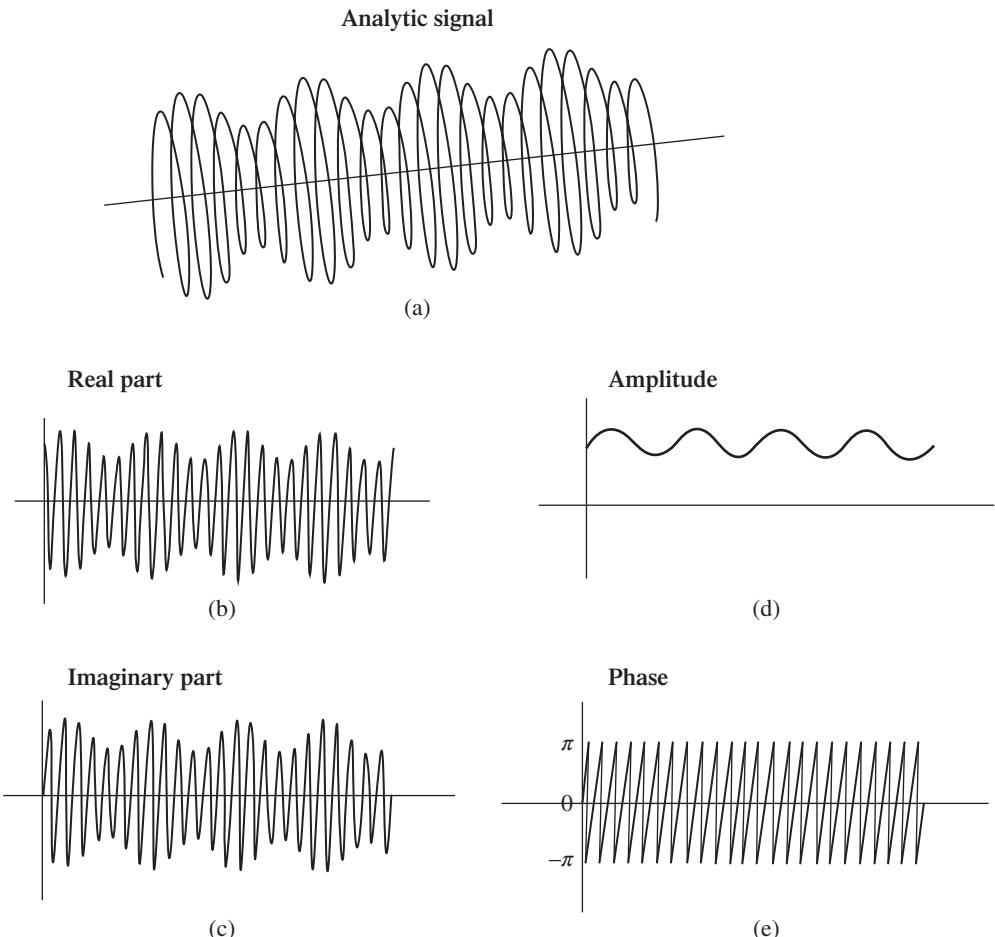


Figure 3.26 Analytic signal for an amplitude modulated cosine [4].

the imaginary axis. The analytic signal for amplitude modulation by a single frequency is shown in Figure 3.26 [4], along with the projections on the real and imaginary axes and the extracted amplitude and phase. Note that the phase is just the linear phase function of the carrier ($2\pi f_c(t)$) but is represented between $\pm\pi$ because this is all that can be determined from the real and imaginary components. In general, it is necessary to ‘unwrap’ the phase, removing the jumps over 2π , to make the phase a continuous function of time. Phase unwrapping is discussed below.

In a similar manner, if the speed of rotation of the vector is allowed to vary from a constant (the carrier frequency) the frequency deviation represents frequency modulation, and its integral represents phase modulation. The rotating vector will have the form of expression (3.51). For pure phase/frequency modulation where the amplitude is constant, modulation by a single frequency will give an analytic signal as shown in Figure 3.27 ([4]).

Once again the phase is depicted between $\pm\pi$, but if unwrapped it becomes a straight line (the dotted line in Figure 3.27e) with superimposed sinusoidal fluctuation around it (the solid line in Figure 3.27e). This sometimes leads and sometimes lags the carrier.

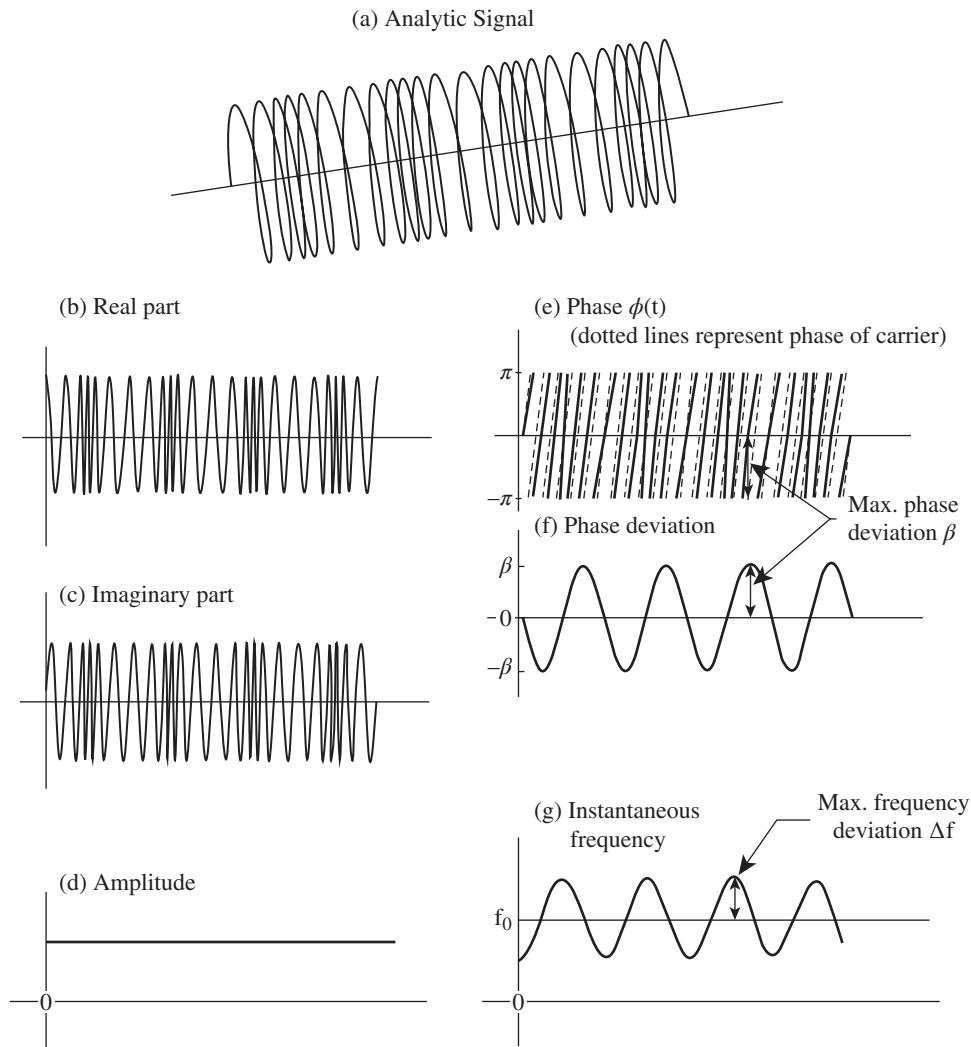


Figure 3.27 Analytic signal for a phase/frequency modulated cosine [4].

3.3.2.1 Modulation Sidebands

The spectrum of a modulated signal has sidebands spaced at the modulating frequency. In a coordinate system rotating with the carrier, the upper sideband rotates in the positive direction at the modulating frequency, while the lower sideband rotates at the same frequency in the opposite direction. In Figure 3.28 it is shown that for amplitude modulation, the phase relationships of the sidebands must be such that their vector sum is always aligned with the carrier component, and serves to change its amplitude sinusoidally. Thus, an amplitude modulated sinusoid can be interpreted as the sum of three constant amplitude vectors (the result of Fourier analysis) or as a single vector rotating at the carrier frequency with sinusoidally varying amplitude.

Figure 3.29 shows that if one of the sidebands has its phase reversed, the vector sum of the two sidebands is now orthogonal to the carrier and thus primarily gives a phase modulation.

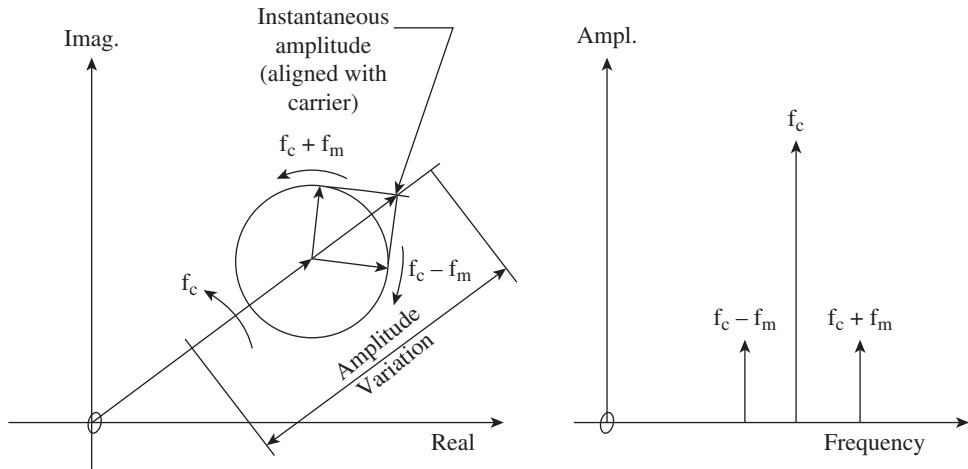


Figure 3.28 Phase relationships of the sidebands for amplitude modulation [4].

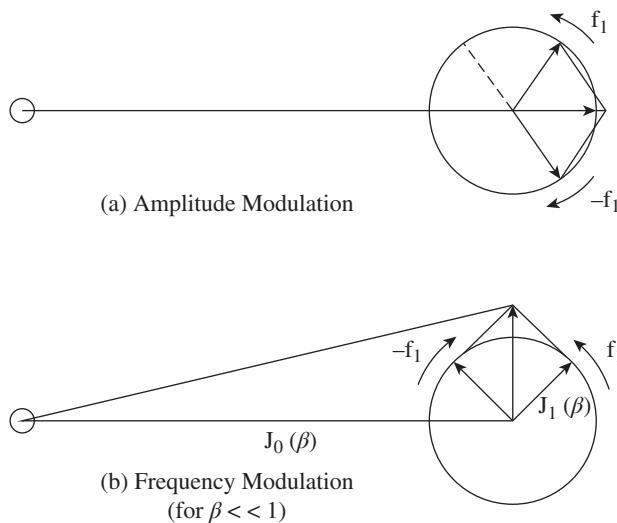


Figure 3.29 Phase relationships of the sidebands for phase modulation [4].

Note that in this case, the amplitude of the resulting vector does change (the hypotenuse of the right-angled triangle) twice per period of the modulating frequency, but the change is very small for small phase modulations. For phase modulation amplitudes up to about 1 rad, the changing amplitude can be compensated by a second pair of sidebands at twice the modulating frequency, whose phase is arranged for amplitude modulation. For even greater phase modulation amplitudes, even more sidebands are required, even for sinusoidal modulation at one frequency, but this is rarely the case for torsional vibration, which is normally less than a degree. It is relevant for other applications of phase/frequency demodulation, such as order tracking (Section 5.1) and determination of machine speed. This simple example shows that the sidebands from amplitude

and phase modulation usually have different phase on either side of the carrier, and when there is a combination of the two, they will often reinforce on one side and partially cancel on the other. This is taken up again in Section 5.4 when discussing the modulating effects of gearmeshing.

The conditions for which the Hilbert transform can be used for demodulation will now be discussed. Provided the fluctuating part of (3.51),

$$A_m(t) \exp\{j\phi_m(t)\} \quad (3.52)$$

has a half bandwidth less than the carrier frequency f_c , the spectrum of (3.51) will be one-sided, and (3.51) will be an analytic signal. In this case the required imaginary part is the Hilbert transform of the real part, and the methods of Section 3.3.1 can be used. As the spectra of the two parts of (3.52) are convolved, the total bandwidth is less than the sum of the individual bandwidths. The bandwidth of the amplitude part is directly that of $A_m(t)$, and even though that of $\exp\{j\phi_m(t)\}$ is theoretically infinite, if the maximum phase deviation is less than 1 rad, the effective bandwidth (within the dynamic range) is less than twice that of $\phi_m(t)$.

3.3.2.2 Practical Demodulation

The actual demodulation can be carried out in two ways using FFT analysis. Both involve shifting the carrier frequency to zero. As illustrated in Figure 3.30, this can be achieved by postprocessing of a time signal using FFT transforms, although the first one will have to be large to accommodate the high carrier frequency (high sampling frequency), while being long enough to contain sufficient periods of the lower modulating frequencies (Figure 3.30a). Where phase demodulation is required, the centre of the demodulation band will have to be shifted to zero frequency, and negative frequency components shifted to the other end of the frequency record (Figure 3.30b). However, for amplitude demodulation the result is unaffected by the frequency shift, and it is more convenient to shift the left hand end of the band to zero frequency, and pad the negative frequency side with zeros, thus maintaining an analytic signal (Figure 3.30d). In either case, there should be at least as many contiguous zeros in the spectrum as components, since the modulus is the square root of the amplitude squared, and the latter corresponds in the frequency domain to the convolution of the spectrum with its complex conjugate reversed end-for-end. The zeros prevent extraneous wrap-around errors in the convolution operation. Such errors may result if the situation depicted in Figure 3.30c is attempted, and zero padding is not used.

Note that the zoom processor depicted in Figure 3.18 virtually achieves the same thing in the time domain, by subtracting the carrier frequency from all components in the signal (provided the zoom centre frequency is set at the carrier frequency). This changes the formula of the signal (related to the modulation of a single carrier) from Expression (3.51) to (3.52), and thus gives a complex time signal whose amplitude is the amplitude modulation signal, and whose phase is the phase modulation signal. At the same time the lowpass filtering associated with the zoom operation can be used to select just that part of the spectrum corresponding to the single modulated carrier, and the downsampling gives a frequency range appropriate to the modulating frequencies, much lower than the carrier frequency in general.

3.3.2.3 Phase Unwrapping

In general, the phase will be calculated from the complex number as the arctangent of imaginary over real value, and this is only defined between $\pm\pi$ as mentioned above. In general this will have to be

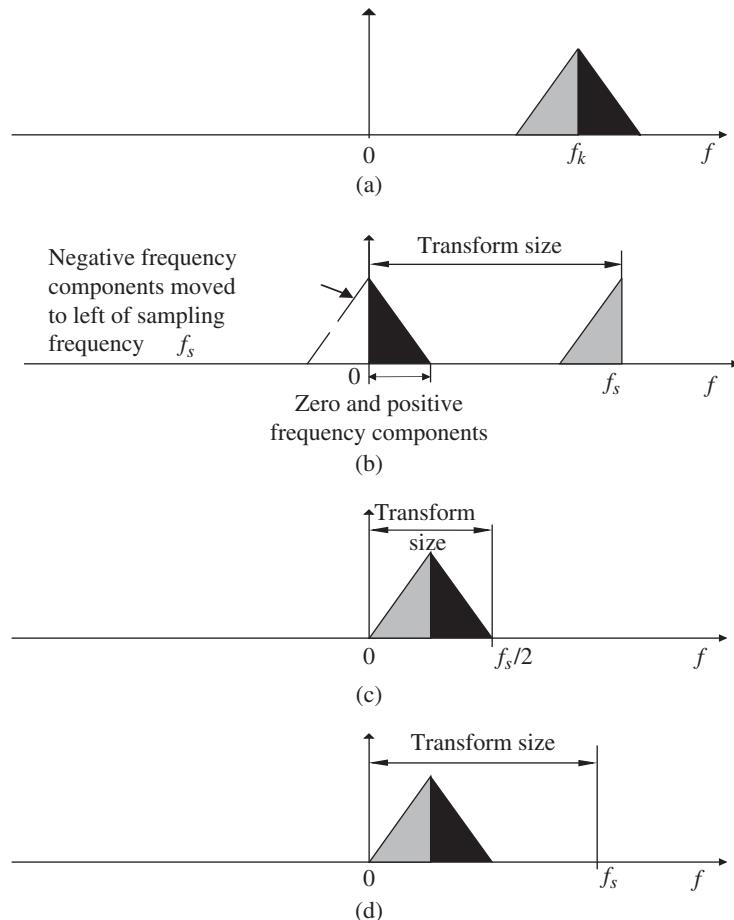


Figure 3.30 Procedure for demodulation by shifting the carrier frequency and reducing the transform size.

unwrapped to a continuous function of time. Perhaps the most reliable phase unwrapping algorithm is that proposed by Tribolet [11], but for most practical phase demodulation problems, a somewhat simpler algorithm can be used which simply checks whether the next sample has changed by more than π from the previous value (indicating a jump over 2π which has to be removed). Care has to be taken in cases where the slope of the phase function is large, and the sampling frequency too low, as the phase jump in that case may be genuinely $> \pi$. A trial and error process can be used to resample the signal with a finer and finer sample spacing (doubling the sampling frequency at each step) until a stable unwrapping result is obtained. Halving the sample spacing will mean that a jump just $> \pi$ will become just $> \pi/2$ and thus not be detected as a phase jump. The interpolation can be very simply achieved by repeatedly doubling the size of the spectrum that is inverse transformed (Figure 3.30b).

Note also that in choosing the carrier frequency, this can only be done in discrete steps equal to the frequency resolution Δf of the spectrum ($1/T$ Hz, where T is the record length in seconds). Choosing the wrong centre frequency will result in the addition of a linear slope to the actual phase function. Since Δf corresponds to one rotation along the record, the maximum slope resulting from choosing

the nearest line to the actual carrier will give a slope of $\pm\pi$ along the record. Such a slope can easily be removed by a detrend operation if desired.

3.4 Digital Filtering

As made clear in Section 3.2, the FFT provides a very efficient way of obtaining frequency spectra on a linear frequency scale with constant bandwidth, and this is most often advantageous for diagnostic purposes. However, for generating spectra with constant percentage bandwidth (CPB) (i.e. $1/n^{\text{th}}$ octave) on a logarithmic frequency scale, digital filters give considerable advantage, in particular recursive IIR (infinite impulse response) filters. Digital filters are similar to analogue filters in that the output signal is convolved with the impulse response of the filter, and operate directly in the time domain on continuous (though sampled) signals as opposed to the blockwise treatment of the FFT process. The coefficients that define the filter properties give a characteristic which is defined in relation to the sampling frequency. Thus, 18 sets of filter coefficients will define the 1/18th octave filters in one octave, but halving the sampling frequency will produce the equivalent filters one octave lower. Before halving the sampling frequency, the signal must be lowpass filtered by a filter that removes the upper octave of frequency information, but this can also be done by a digital filter with the same coefficients for every octave.

Figure 3.31 illustrates that when the sampling frequency is repeatedly halved for each octave, the total number of samples to be treated per unit time = $M(1 + 1/2 + 1/4 + 1/8 + \dots) = 2M$ samples/s, so that if the digital filter processor is capable of operating twice as fast as necessary for the highest octave, any number of lower octaves can be processed in real-time. This feature was mentioned in conjunction with the zoom processor of Figure 3.18.

If the digital filtering cannot be done in real-time, a very large number of data will have to be stored in advance. As an example, to produce 1/18th octave filters over three decades in frequency (frequency range 1000 : 1), each estimate of a spectrum value would have to encompass at least the impulse response time of the filter, approximately 30 periods of the centre frequency for a 1/18th octave filter. For the lowest filter in the lowest octave there would have to be six samples per period, and since the sampling frequency would have to be decimated by a factor of 500 from the highest to the lowest decade, this corresponds to almost 100 000 samples in the original record. To achieve a result with only 10 averages would thus require of the order of 10^6 samples in the original record.

CPB spectra can also be obtained by conversion from FFT spectra, as illustrated in Figure 3.32, where each decade is converted separately. The bandwidth of the individual lines in the original FFT spectra (including the effect of any window) must be less than the desired percentage bandwidth at the lowest frequency in the FFT band. The conversion is achieved by calculating the lower and upper cutoff frequencies of each constant percentage band, and then integrating up the power in the FFT lines (and parts of lines) between the limits. The method indicated in Figure 3.32 gives a

.....	M samples/s
.....	M/2 samples/s
.	M/4 samples/s
.	M/8 samples/s

Figure 3.31 Effect of repeatedly halving the sampling frequency.

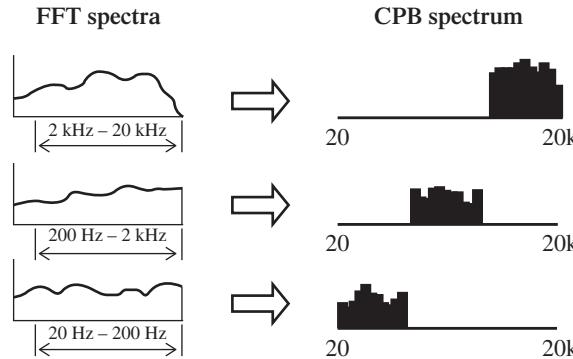


Figure 3.32 Conversion from FFT spectra to a CPB spectrum.

large change in filter characteristic at the junction between decades, but this is not likely to be such a problem with machine vibration analyses as with acoustic spectra. To reduce the latter problem, some FFT analysers do the conversion on an octave rather than a decade basis.

3.4.1 Realisation of Digital Filters

The intention here is not to give a full treatise on digital filtering, but just to give an impression of how digital filters work, and how they relate to other topics treated in the book, such as the linear prediction discussed in Section 5.3.3 and adaptive filters in Sections 5.3.4 and 5.3.5.

The simplest type of digital filter is a ‘finite impulse response’ (FIR) filter, where the impulse response of the filter (of finite length) is convolved with the signal according to the equation:

$$y_i = \sum_{k=0}^M b_k x_{i-k} \quad (3.53)$$

Here x_i represents the input signal, y_i represents the output signal, and the b_k represent the convolution weights or samples of the impulse response. Eq. (3.53) is a digitised, finite length version of the convolution Eq. (3.32) and is effectively a ‘moving average’, giving rise to the term MA model. Applying a z-transform to Eq. (3.53), which is the equivalent of a Laplace transform for discrete time signals, the convolution becomes the product:

$$Y(z) = \sum_{k=0}^M b_k z^{-k} X(z) = B(z)X(z) \quad (3.54)$$

from which the transfer function can be seen to be:

$$B(z) = \sum_{k=0}^M b_k z^{-k} = b_0 \prod_{k=0}^M (1 - z^{-1} z_k) \quad (3.55)$$

which has no poles and is thus an ‘all-zero’ model. This type of model is obviously most efficient when the effective length of the impulse response is short, meaning that it is highly damped and thus without sharp spectral peaks.

IIR filters can be achieved by forming each output value as a weighted sum of previous output values and the current input, using the equation:

$$y_i = - \sum_{k=1}^N a_k y_{i-k} + x_i \quad (3.56)$$

where in principle $N \rightarrow \infty$ but can be truncated when the terms become sufficiently small. After z-transformation this gives:

$$Y(z)A(z) = X(z) \quad (3.57)$$

from which the transfer function can be seen to be:

$$1/A(z) = 1 / \sum_{k=0}^N a_k z^{-k} = 1 / \prod_{k=1}^N (1 - z^{-1} p_k) \quad (3.58)$$

which has no zeros and is an all-pole model. IIR filters are most efficient where there are sharp spectral peaks (low damping), which would require a very long FIR filter.

A simple example of an IIR filter is an exponential averaging algorithm, where α times each input is added to $(1 - \alpha)$ times the previous weighted average. For example, if $\alpha = 0.1$, the weighting on previous samples, going backwards in time, will be $0.1^*(1.0, 0.9, 0.9^2, 0.9^3, \dots)$, which is seen to be the digitised equivalent of the decaying exponential function in Figure 3.10g.

3.4.2 Comparison of Digital Filtering with FFT Processing

The primary use of digital filters is to provide real-time processing, such as required in telecommunications, speech analysis, and automatic control. Real-time analysis requires causal processing, with filters, etc. having causal impulse responses, giving delays of the order of the filter response time, and corresponding phase shifts in the FRFs. Filter characteristics are far from ideal (i.e. rectangular in the frequency domain), and the deviations are defined in terms of filter ‘roll-off’, typically expressed in terms of dB per octave, outside the passband, and with ‘ripple’ inside the passband. Filters can be designed to have zero phase shift, but only by making them non-causal and applied by post-processing. A typical example is the FILTFILT operation in Matlab, where the signal is first processed in the forward (e.g. time) direction, and then in the reverse direction, so that the phase shift introduced in the first operation is cancelled in the second, but the filter order is doubled (with filter roll-off and ripple corresponding to that order).

Real-time processing is virtually never required in machine health monitoring, as information is usually being sought days, weeks, or months ahead of when the condition becomes serious, but even for permanent monitoring of critical machines, the actual delay time between true causal real-time processing and non-causal processing, for example by FFT analysis, is only of the order of a second, or less, and still normally negligible in terms of the time required to shut a machine down, or otherwise react to an emergency.

FFT analysis is inherently non-causal, since as pointed out in Section 3.2.4 the second half of each time record represents negative time, in the same way that the second half of each spectrum represents negative frequency. On the other hand, this means that filtering in the frequency domain, by windowing a frequency band with a rectangular window, is almost ‘ideal’, in the sense that large local (discrete) frequency components can be excluded by setting the band limit a couple of lines away. There is a small window effect coming from the (usually small) sidelobes from the time window used, but this extends at most over a few lines and is much less than the roll-off effects of causal filters

(analogue or digital). This is one of the main reasons why the ‘Hilbert’ method of envelope analysis (see Section 7.3.2) is vastly superior for bearing diagnostics than the older analogue (and derived digital) techniques formerly used. Non-causal processing is always to be preferred for demodulation, in particular phase/frequency demodulation, where phase shifts are critical.

3.5 Time/Frequency Analysis

In theory the Fourier transform requires integration over all time, but we are all aware that the ear can detect changes in frequency with time (for example music), and so an analysis technique has been sought which matches the ear’s ability to follow changing frequency patterns.

3.5.1 The Short Time Fourier Transform (STFT)

A simple approach is to move a short time window along the record and obtain the Fourier spectrum as a function of time shift. This is called the short time Fourier transform (STFT). However, the uncertainty principle means that the frequency resolution is the reciprocal of the effective time window length, and this does not seem to accord with the ear’s appreciation of a tonal quality of a note even if it lasts for a short time. Even so, the STFT is sometimes useful for tracking changes in frequency with time, even with the restriction of resolution. It is described by the formula:

$$S(f, \tau) = \int_{-\infty}^{\infty} x(t)w(t - \tau) \exp(-j2\pi ft)dt \quad (3.59)$$

where $w(t)$ is a window which is moved along the record. Normally, the amplitude squared $|S(f, \tau)|^2$ is displayed on a time vs frequency diagram, in which case it is sometimes known as a spectrogram. The window could be of finite length such as a Hanning window, or theoretically infinite such as a Gaussian window, but in practice of course it must be truncated.

3.5.2 The Wigner-Ville Distribution

The Wigner-Ville distribution (WVD) seems to violate the uncertainty principle in appearing to give better resolution than the STFT, but suffers from interference components between the actual components. The original Wigner distribution [12] was modified by Ville [13] who proposed the analysis of the corresponding analytic signal so as to eliminate interference between positive and negative frequency components. The WVD is one of the so-called ‘Cohen’s class’ of time-frequency distributions [14], most of which have been proposed to improve on the WVD in some way. Even the STFT falls into this class. Cohen’s class may be represented by the formula:

$$C_x(t, f, \phi) = \Im\{R(t; \tau)\} \quad (3.60)$$

where $R(t; \tau)$ is a weighted autocorrelation-like function defined by:

$$R(t; \tau) = \int_{-\infty}^{\infty} x(u + \frac{\tau}{2})x * (u - \frac{\tau}{2})\phi((t - u), \tau)du \quad (3.61)$$

and $\phi(u, \tau)$ is a kernel function used to smooth the WVD (with $\phi = 1$ the WVD is obtained). The ‘pseudo WVD’ is a finite windowed version of the WVD and the ‘smoothed pseudo WVD’ suppresses interference in both the time and frequency directions. Figure 3.33 compares the WVD and

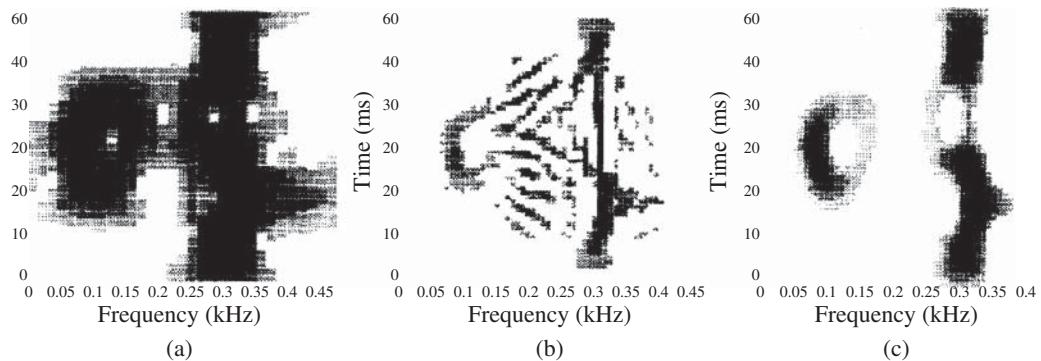


Figure 3.33 Comparison of time-frequency distributions for a diesel engine vibration signal (a) STFT (b) Wigner-Ville distribution (c) Smoothed pseudo Wigner-Ville.

the smoothed pseudo-WVD against the STFT for a vibration signal from a portion of a diesel engine cycle, and shows that at least in this case the smoothing gives a simultaneous resolution in both directions that is better than the STFT, while still suppressing the major interference components. In Ref. [15] the proposal is made to use various smoothing techniques to locate the interference components, and then remove them from the unsmoothed WVD to retain optimum resolution.

3.5.3 Wavelet Analysis

Another approach to time-frequency analysis is to decompose the signal in terms of a family of ‘wavelets’ which have a fixed shape, but can be shifted and dilated in time. The formula for the wavelet transform is:

$$W(a; b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi * \left(\frac{(t-b)}{a} \right) dt \quad (3.62)$$

where $\psi(t)$ is the mother wavelet, translated by b and dilated by factor a . Since this is a convolution, the wavelets can be considered as a set of impulse responses of filters, which because of the dilation factor have constant percentage bandwidth properties. In principle, they are not very different from $1/n^{\text{th}}$ octave filters, but with zero phase shift because the mother wavelet is normally centred on zero time. The dilation factor a is known as scale, but represents log frequency, as for constant percentage bandwidth filters. Wavelets give a better time localisation at high frequencies, and for that reason can be useful for detecting local events in a signal. Many authors have described their use for detecting local faults in gears and bearings (e.g. [16, 17]), and Ref. [18] is an extensive overview of the applications of wavelet analysis for machine condition monitoring and fault diagnostics.

Wavelets can be orthogonal or non-orthogonal, and continuous or discrete [19]. Examples of orthogonal wavelets are the Daubechies dilation wavelets [20], which are compact in the time domain, but in principle infinite in the frequency domain. They tend to have irregular shapes in the time domain. Newland [19] describes complex harmonic wavelets, which are compact in the frequency domain, but infinite in the time domain. They have the appearance of windowed sinusoids (harmonic functions) and are typically of one octave bandwidth, although they can be narrower. The advantage of complex wavelets is that the imaginary part of the wavelet is orthogonal to the real part (sine rather than cosine) and thus the overall result is not sensitive to the position (phasing) of the event being transformed (it may be centred on a zero crossing of the real part, but this would

be a maximum of the imaginary part). The local sum of squares of the real and imaginary parts is a smooth function. Harmonic wavelet transforms can be efficiently produced using FFT methods [19].

Orthogonal wavelets are the most efficient to use when analysis/synthesis is to be performed (e.g. after denoising), or when the significant features of the signal are to be represented with a minimum number of parameters (e.g. as inputs to artificial neural networks). However, it is possible to obtain complete reconstruction of a signal using non-orthogonal wavelets, as long as there is some redundancy or overlap [19]. For analysis purposes, non-orthogonal wavelets such as Morlet wavelets are often more convenient, and in any case it is generally preferable to use redundant ‘lapped’ transforms to aid visual interpretation [19].

3.5.3.1 Wavelet Packets

As mentioned above, the frequency bands corresponding to the various scales for normal wavelets are octave-based, since the lower half band is repeatedly split into two for each scale. Wavelet packets are derived by splitting each upper half band in the same way, so that the number of bands at each scale is doubled and the bandwidth halved at each step, meaning that the number of (equal) bands into which the frequency range is split at each stage is $n = 1, 2, 4, 8 \dots$ etc. This means that the resulting spectrum at each level is constant bandwidth with constant spacing, and each could in principle be produced by an STFT with window length 2^n samples.

3.5.3.2 Wavelet Denoising

One of the primary applications of wavelets in machine diagnostics is in denoising of signals in both time and frequency domains simultaneously. Most wavelet denoising is an extension of the work of Donoho and Johnstone [21], who defined two types of thresholding to remove noise, this being defined as any components with amplitude less than a certain threshold value. In so-called ‘hard thresholding’ the retained components are left unchanged, but in ‘soft thresholding’, the noise estimate (the threshold value) is subtracted from them also (symmetrical treatment of positive and negative values). More advanced methods are continually being developed.

Figure 3.34 shows the result of denoising acceleration signals from a gearbox with a simulated tooth root crack, using a proposed new ‘dual-tree complex wavelet transform (DTCWT)’ and ‘Neigh-Coeff shrinkage’ for thresholding, and compares it with other wavelet transforms [22].

3.5.3.3 Morlet Wavelets

Morlet wavelets are Gaussian windowed sinusoids, which are non-orthogonal but suitable for analysing many machine vibration signals since they are similar in appearance to narrow band impulse responses (although non-causal). They can conveniently be made with bandwidth equal to $1/n^{\text{th}}$ octaves, to correspond to different damping. The original real Morlet wavelets were windowed cosines, but as for harmonic wavelets it is convenient to use the complex version (with a one-sided spectrum, so that the imaginary part is the Hilbert transform of the real part, i.e. a windowed sine function).

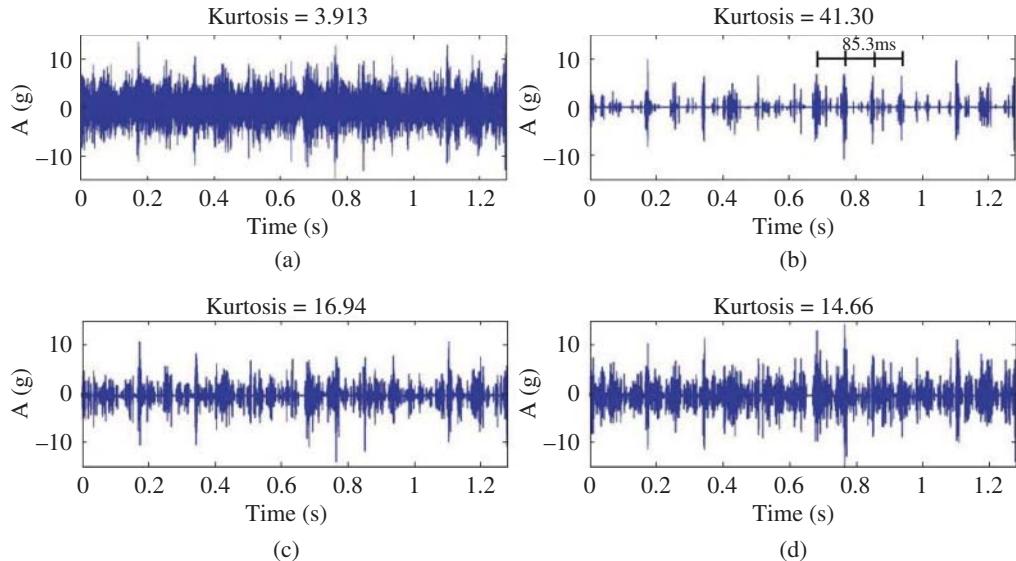


Figure 3.34 Example of advanced wavelet denoising (a) Raw vibration signal; denoised signal using Neigh-Coeff shrink based on (b) DTCWT, (c) DWT (discrete wavelet transform) and (d) SGWT (second-generation wavelet transform). Source: From [22].

Thus, the complex Morlet wavelet is defined in the time domain as a complex exponential wave multiplied by a Gaussian function, and it also has the shape of a Gaussian window in the frequency domain as follows:

$$\psi(t) = \frac{\sigma}{\sqrt{\pi}} e^{-\sigma^2 t^2} e^{j2\pi f_0 t} \quad (3.63)$$

$$\Psi(f) = \Psi * (f) = e^{-(\pi^2/\sigma^2)(f-f_0)^2} \quad (3.64)$$

where $\Psi(f)$ is the Fourier Transform of $\psi(t)$. Since $\Psi(f)$ is real, $\Psi(f) = \Psi^*(f)$ where $*$ denotes the complex conjugate. f_0 is the centre frequency of the window and σ determines its width. Since the real and imaginary parts of $\psi(t)$ are Hilbert transforms, it is analytic and its Fourier transform $\Psi(f)$ is one-sided with positive frequencies only. This is modified slightly by the truncation at low frequencies as shown in Figure 3.35, though this is small for small σ .

The advantage of the complex Morlet wavelets, compared with the real version using windowed cosines, is that the imaginary parts (sines) have their maxima when the cosines have zero crossings, so that the squared amplitude of the wavelet coefficients is not sensitive to the phasing of local features in the time signals.

The wavelet transform can be achieved by multiplication in the frequency domain of the signal spectrum with that of the Morlet wavelet given in (3.64) and then inverse Fourier transformation. In the example shown here in Figure 3.35, the ‘bandwidth’ of the wavelets was taken to be one octave.

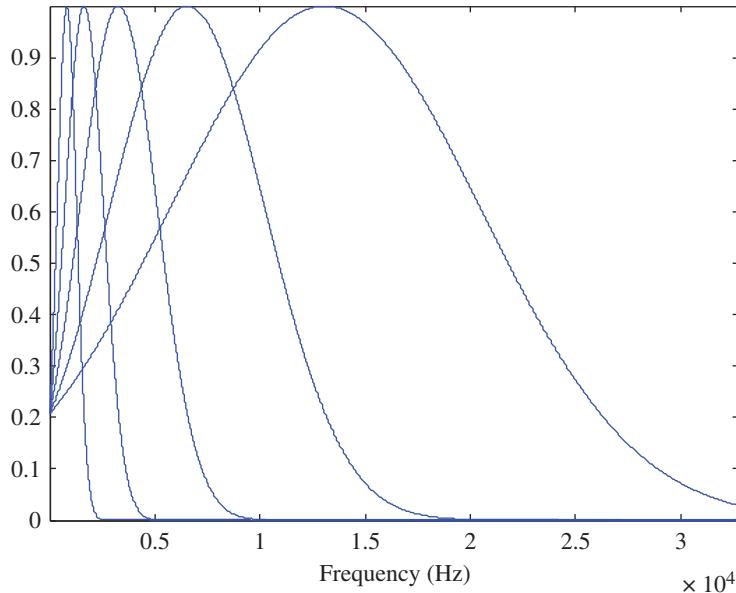


Figure 3.35 Filter characteristics of the octave band Morlet wavelet filters.

3.5.3.4 Choice of Wavelets

In general, wavelets should be chosen to have the greatest similarity to the features in the signal which are to be extracted. Some work has been done on ‘adaptive wavelets’ (e.g. [23, 24]) whose properties are modified to maximise some criterion. However, the analyst can often choose suitable wavelet types and properties based on a knowledge of the signals to be extracted, or sometimes by trial and error. Figure 3.36 shows an example of the advantage given by choosing the ‘impulse wavelets’ of [24] vs Morlet wavelets for analysis of a local fault in a gear [25]. Note that the vertical ripple is presumably because of the choice of real rather than complex wavelets. It is evident that the impulse wavelets conform better to the signal given by a local crack in this case. At the same time, this example emphasises the problems associated with the use of wavelets for analysis and diagnostics. Even if the eye can see localised effects in the diagram, it would still require some sort of image analysis to characterise them.

3.5.4 Empirical Mode Decomposition

Empirical mode decomposition (EMD) is not a time/frequency analysis technique as such, but a way of dividing a signal up into so-called intrinsic mode functions (IMFs), which may be a more compact way of describing them, than say Fourier analysis. The IMFs are constrained to be mono-components, each comprising a single carrier tone, which is amplitude and frequency modulated [26]. Thus, the phase of each IMF must be continuous, and the number of zero crossings must be equal to the number of extrema (maxima plus minima) or differ at most by one in the whole record.

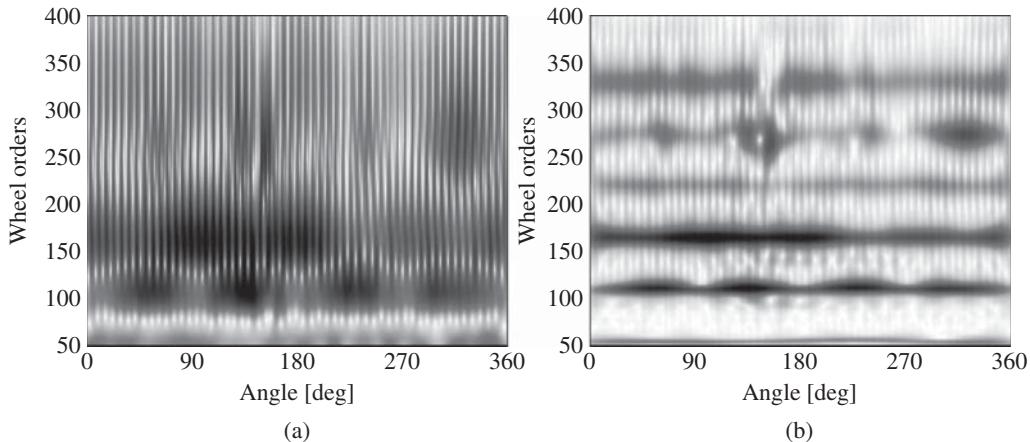


Figure 3.36 Comparison of two wavelet types for the analysis of a small crack in a gear (a) Real Morlet wavelets (b) Impulse wavelets. Source: From D'Elia [25].

3.5.4.1 Extraction of IMFs

The IMFs are extracted by an iterative ‘sifting’ process, with the following steps:

- 1) The local extrema of the signal are found and the upper and lower envelopes formed by fitting cubic spline curves to the maxima and minima respectively.
- 2) The mean of the upper and lower envelopes is found, and subtracted, leaving the first estimate of IMF1, say IMF1.1. For this to be a true IMF, the upper and lower envelopes would have to be symmetric around a mean value of zero, but this is rarely the case for a first iteration.
- 3) The process is repeated on IMF1.1, giving IMF1.2, IMF1.3 ... etc., until symmetry is achieved, after which this is denoted IMF1 and subtracted, leaving a residual signal
- 4) This is treated in the same way to find IMF2, IMF3 ...etc., until the residual is no longer oscillatory, after which the process is stopped.

This is illustrated in Figure 3.37, adapted from a presentation by Flandrin [27], for a mixture of two components, a constant frequency tone and a chirp of increasing frequency, but everywhere of higher frequency than the tone.

3.5.4.2 Problems with EMD and Solutions

Ref. [28] is a review of the applications of EMD in fault diagnosis of rotating machinery, and includes a good discussion of the problems associated with EMD in general, such as different criteria for terminating the sifting, as well as some solutions.

One of these problems is end effects, which can be seen in the results of Figure 3.37, even for a relatively simple case with no added noise. IMF1 is seen to be dominated by the chirp signal, but with errors at the ends; similarly for IMF2, which is dominated by the tone. IMFs 3–7 (not uniformly scaled) are primarily required to compensate for these end effects. The end effects are basically due to the fact that the spline fits near the ends are simply extrapolations from the nearest local maxima

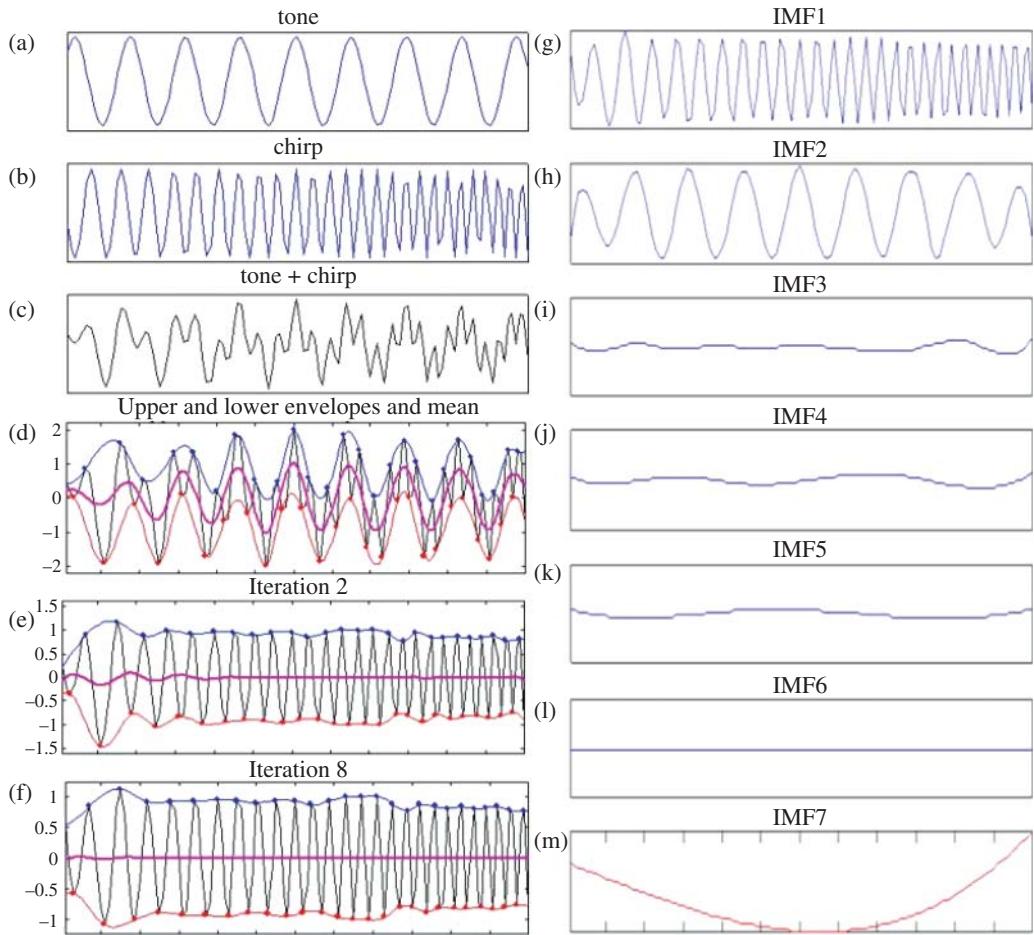


Figure 3.37 Extraction of IMFs of a tone plus chirp. Source: From Flandrin [27]. (a)–(c) components plus signal (d)–(f) Extraction of IMF1 (g)–(m) IMFs 1–7.

and minima within the captured record, and it is likely that actual extrema in the extended curve lie just outside this.

Another major problem is ‘mode mixing’, where the extracted IMFs may jump between two or more actual modes where they intersect or approach each other. This often occurs because of noise affecting the results where the amplitude is small. The suggested solution [29] is to add a range of different white noise signals to the actual signal and obtain the ‘true’ IMFs as an average over an ensemble of estimates, and this approach is called the ‘ensemble empirical mode decomposition (EEMD)’.

3.5.4.3 Applications in Machine Diagnostics

Ref. [28] cites a large number of papers proposing the application of EMD and EEMD for various machine diagnostic problems, but without much critical discussion of the claims of each paper. Three

main areas are discussed, namely the applications to bearing diagnostics, gear diagnostics and rotor dynamic problems. Despite listing large numbers of references, a major point in the Discussion section of the review paper is ‘Making comparisons on applications of EMD for different diagnosis objects, i.e. rolling element bearings, gears, and rotors, it is found that EMD performs better in extracting fault characteristics of rotors than both bearings and gears’. This is attributed to the fact that rotor signals are measured in terms of (relative) displacement, with the signals being relatively simple, while those from bearings and gears are usually measured in terms of acceleration, and are complicated and noisy.

In the author’s opinion, the application to rolling element bearings is very dubious, since it is difficult to see how the signals from bearing fault responses can be expressed in terms of mono-components with continuous amplitude and phase. Except for outer race faults, the series of impulse responses are amplitude modulated by a function consisting of a series of Stribeck functions (approximately half cosine) separated by sections, at least as long, equal to zero, where the fault is outside the load zone. Moreover, at high frequency the impulse responses contain overlapping modes, giving beats between the close modal frequencies, and a characteristic of beats is that the phase jumps by π at every zero crossing of the envelope. It is difficult to see how mode-mixing can be avoided, even by using EEMD. Moreover, the diagnostic information in bearing signals is almost entirely contained in the envelope (amplitude modulation) part of the signal, but in EMD a lot of computational effort is expended in determining (possibly non-existent) well-behaved carrier signals, which have little diagnostic content. It makes much more sense to use cyclostationary analysis tools, such as spectral correlation and spectral coherence (Sections 3.6.1 and 3.6.2), which separate the modulation information from the (often random) carrier information. This same advantage can be extended to cyclo-non-stationary analysis (Section 3.6.4), for variable speed machines.

The application to gear diagnostics does have some justification, and in one example at least; that presented in Ref. [30], the information about a local tooth root crack is primarily contained in two IMFs, IMF1 and IMF2. It is shown that the energy in IMF2 trends very well with the growth of the crack. However, it is a simple single stage gearbox. In Ref. [31] EMD analysis was used as a means of dividing the signal from a complex wind turbine gearbox into mono-components, to use for Teager analysis for both frequency and amplitude demodulation (see Section 5.2.2), and in this case gave somewhat similar results to [30] for a local fault in the high-speed part of the gearbox (second parallel gear stage). Since the gearbox speed was constant it is not clear whether the separation was any better than given by bandpass filters, but it is conceivable that it would be better for variable speed, where the different frequency regions (from different gear meshes for example) might overlap when spectra are taken over the whole record. It is not clear whether the separation would be so good in the lower frequency range, where for example the planetary garmesh frequency is often close to the high speed output shaft speed (both being of the order of 100 times the input speed).

3.6 Cyclostationary Analysis and Spectral Correlation

Section 5.3 discusses the separation of deterministic and random signals, but cyclostationary analysis gives the possibility to separate a further category of signals, namely those with cyclostationary properties, and in particular those with cyclostationarity of second order. Moreover, cyclostationary signals with different cyclic frequencies can be separated from each other, as well as from deterministic and stationary random signals.

An n^{th} order cyclostationary signal is one whose n^{th} order statistics are periodic. Thus a first order cyclostationary signal has a periodic mean value (e.g. a periodic signal plus noise), while a second

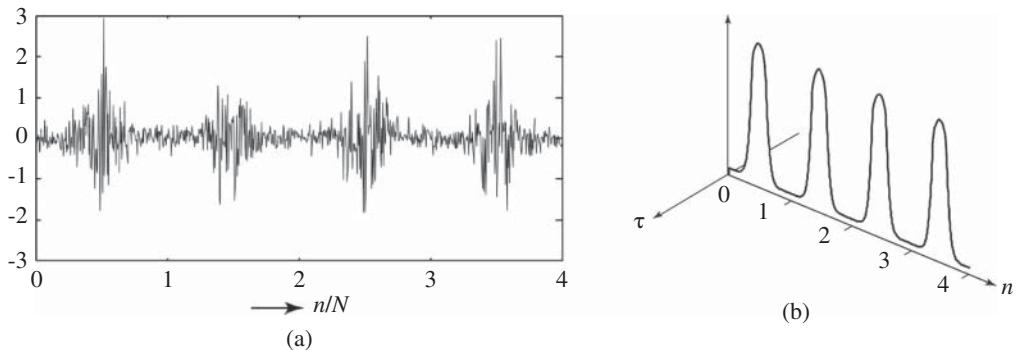


Figure 3.38 Example of amplitude modulated white noise (from Antoni [32]) (a) time signal over four periods of cyclic frequency (b) Two-dimensional autocorrelation function vs time (sample) n and time lag τ .

order cyclostationary signal has a periodic autocorrelation function (e.g. white noise amplitude modulated by a periodic signal as in Figure 2.3 of Chapter 2). Another example of the latter taken from Antoni [32] is shown in Figure 3.38a. The theoretical two-dimensional autocorrelation function for this case is shown in Figure 3.38b. It could be calculated using Eq. (2.4) of Chapter 2 for an infinite number of realisations. For convenience, the equation is repeated here as (3.65).

$$R_{xx}(t, \tau) = E[x(t - \tau/2)x(t + \tau/2)] \quad (3.65)$$

Note that for time lag $\tau = 0$, this gives the mean square value, which is the same as the variance, since the mean value of the white noise is zero. For values of time lag τ other than zero the autocorrelation is zero, because with white noise the inverse transform of the spectrum is a delta function at zero time lag. Thus, in this case the autocorrelation function is only non-zero for the slice at $\tau = 0$. If the modulated noise were bandlimited, then as illustrated in Figure 3.17 the autocorrelation would be a sinc(x) function scaled by the local variance. For large bandwidth, it would be very localised at small values of τ .

3.6.1 Spectral Correlation

If a two-dimensional Fourier transform is performed on the two-dimensional autocorrelation function, the so-called ‘spectral correlation’ is obtained. The reason for the name will become obvious later. Thus,

$$S_{xx}(\alpha, f) = \lim_{W \rightarrow \infty} \frac{1}{W} \int_R \int_{-W/2}^{W/2} R_{xx}(t, \tau) e^{-j2\pi(f\tau + \alpha t)} dt d\tau \quad (3.66)$$

where $S_{xx}(\alpha, f)$ is the spectral correlation density as a function of normal frequency f and cyclic frequency α . As with one-dimensional autocorrelations, frequency f is obtained from transformation with respect to time lag τ , but cyclic frequency α is obtained from transformation with respect to time t . Because in this case the autocorrelation function is periodic with respect to time t , this Fourier transform must be interpreted as a Fourier series, giving discrete components in cyclic frequency, but because it is a transient in the time lag τ direction, $S_{xx}(\alpha, f)$ is continuous in the normal frequency direction. The result for the signal of Figure 3.38 is given in Figure 3.39 in two stages. In the first stage, (Figure 3.39a) the transformation from τ to f is performed, giving the variation in power

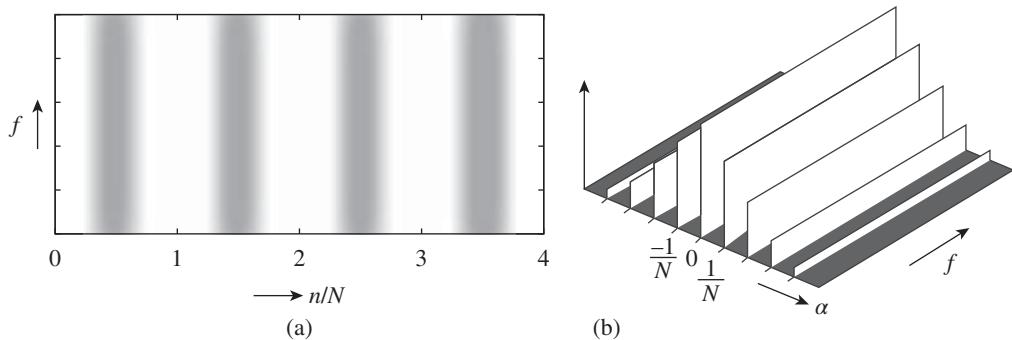


Figure 3.39 Spectral correlation for the case of Figure 3.38. Source: From Antoni [32]. (a) First transformation from τ to f (b) Result of two transforms.

spectrum as a function of time. This is closely related to a WVD, and is discussed below under ‘Wigner-Ville spectrum’ (WVS). The result in this case (expressed as a contour) shows the white spectrum varying in amplitude proportionally to the local value of the variance (the curve at zero time lag in Figure 3.38b). After the second transform from t to α the amplitude of the white noise spectra in Figure 3.39b, for each discrete multiple of the fundamental cyclic frequency, is proportional to the corresponding harmonic of the Fourier series of the variance curve of Figure 3.38b.

The same result can be obtained by correlating the spectrum with itself, hence the name ‘spectral correlation’. This can be explained by considering that for an amplitude modulated noise signal as in Figure 3.38, each spectral line has a pair of sidebands spaced at each modulating frequency or harmonic (from convolution of each line with the spectrum of the modulating function). These are not visible because the basic spectrum is white. However, when the spectrum is correlated with itself, this correlation becomes apparent whenever the spectrum shift corresponds to a multiple of the cyclic frequency, suddenly giving a finite value, while being zero for other frequency shifts. Another way of explaining it derives from the definition of Eq. (3.67). When the first transform is made with respect to τ , the result at each time is the instantaneous power spectrum (i.e. the product of a complex spectrum with its complex conjugate). When these products are transformed with respect to t , it gives a convolution in the frequency domain, and as we have seen in Section 3.2.6.5 a convolution of complex conjugates (inversion of the frequency axis) corresponds to a correlation.

Thus, the alternative way to calculate the spectral correlation can be expressed as:

$$S_{xx}(\alpha, f) = \lim_{T \rightarrow \infty} E\{X_T(f + \alpha/2)X_T^*(f - \alpha/2)\} \quad (3.67)$$

In practice it is usually best to use the frequency domain method to estimate spectral correlations, as it does not require prior knowledge of the cyclic frequency (which would be required for ensemble averaging of the autocorrelation function). In some applications the cyclic frequency is not precisely known, and in particular in the case of localised faults in rolling element bearings, it has some random variation. In this latter case, the signals are not strictly cyclostationary, but the term ‘pseudo-cyclostationary’ has been coined, as it is still convenient to use the methods of cyclostationary analysis to treat them [33]. This is taken up in more detail in Section 7.3.1 of Chapter 7.

Antoni gives detailed information in two excellent tutorials [32, 34] on how to estimate and interpret spectral correlations and other functions of interest in cyclostationary analysis. More recently, higher speed methods of performing these calculations have been published [35, 36].

Note that the autocorrelation function of periodic signals is also periodic vs time lag τ (e.g. Figure 3.17a) and so the spectral correlation function is discrete in both f and α directions (a ‘bed of nails’). This is illustrated in Figure 3.40 (from [32]) for the case of a simulated gearbox vibration signal, where the gearmesh signal is modulated by a periodic signal from a gear (first order cyclostationary) and a signal from an extended inner race spall in a bearing. The latter is a mixture of first order (the local mean) and second order (amplitude modulated noise) cyclostationarity. The spectral correlation in (c) shows the continuous components (vs frequency f) from the second order part, and the discrete (vs frequencies f and α) components from the first order part. Thus, if all discrete frequency components are first removed from the signal (e.g. by DRS, Section 5.3.6), only the second order components will be left in the spectral correlation. This means of differentiating gear and bearing faults is described in more detail in Section 7.3.2 of Chapter 7.

3.6.2 Spectral Correlation and Envelope Spectrum

It can be shown (see inset) that the integral of the spectral correlation over all frequency f is the Fourier transform of the expected value of the squared signal, and so is effectively the spectrum of the squared envelope. An example is shown in Figure 3.41 of the application to an inner race bearing fault. Section 7.3.1.2 of Chapter 7 contains a discussion of when the full spectral correlation gives advantages over the envelope spectrum.

Integration of Spectral Correlation over f

$$\begin{aligned} \int S_{xx}(\alpha, f) df &= \int \int E[x(t + \tau/2)x * (t - \tau/2)] e^{-j2\pi\alpha t} \left(\int e^{-j2\pi ft} df \right) dt d\tau \\ &= \int \int E[x(t + \tau/2)x * (t - \tau/2)] \delta(\tau) d\tau e^{-j2\pi\alpha t} dt \\ &= \int E[x(t)x * (t)] e^{-j2\pi\alpha t} dt \\ &= \Im_{t \rightarrow \alpha} \{E[|x(t)|^2]\} \end{aligned}$$

the spectrum of the squared envelope

3.6.3 Wigner-Ville Spectrum

A distribution similar to the WVD is obtained by performing a Fourier transform with respect to time lag τ of the autocorrelation of Eq. (2.4). It is known as the Wigner-Ville spectrum (WVS). The difference is the averaging given by taking the expected value, instead of performing the Fourier transform on a single realisation. Antoni [37] has shown that for second order cyclostationary signals, the interference components average to zero, while the resolution remains unchanged. Figure 3.42 compares the WVS with the WVD for the impulse response of a nonlinear SDOF system. In the WVS the decreasing natural frequency with amplitude is shown very clearly (with the same resolution as the WVD) whereas the interference terms in the latter cloud the issue. The system was excited by a burst random sequence, which is a typical second order cyclostationary signal.

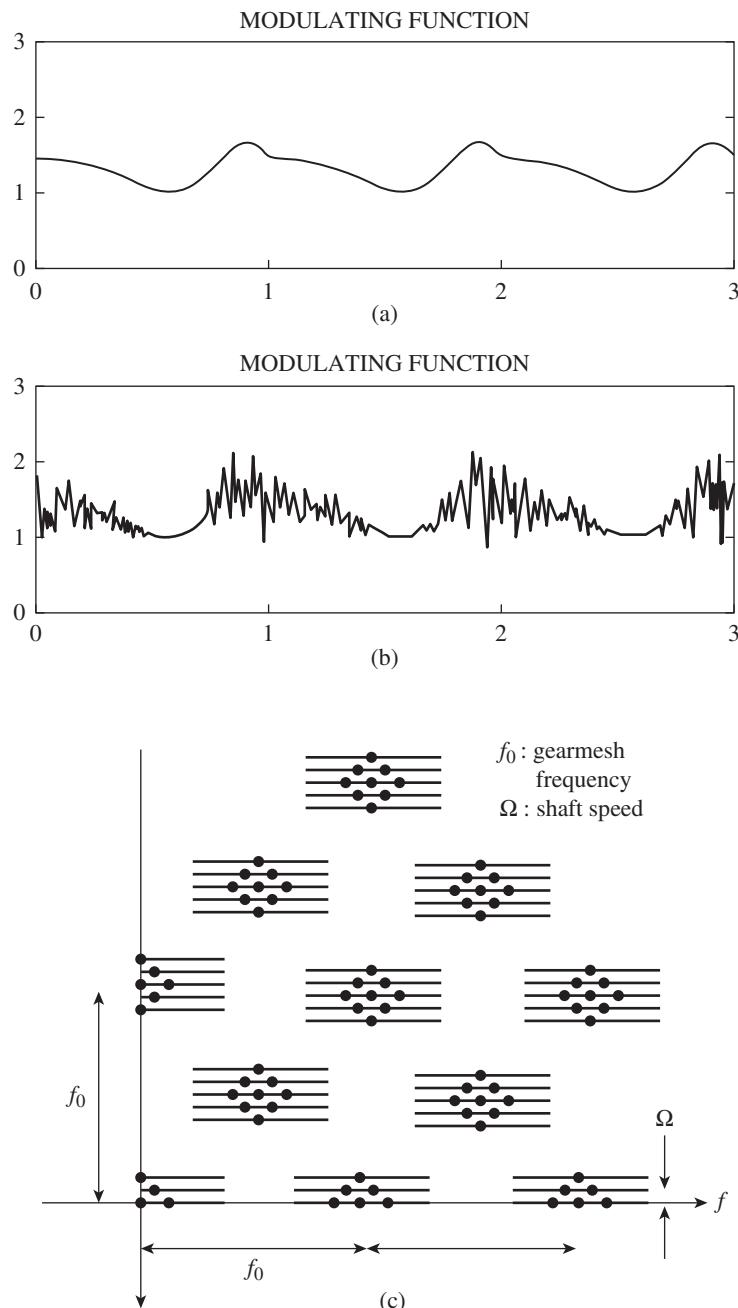


Figure 3.40 Spectral correlation for a mixture of first and second order cyclostationarity illustrated using modulation by gear and bearing signals (a) gearmesh modulation by a gear signal (b) gearmesh modulation by an extended inner race bearing fault (c) spectral correlation for case (b). Source: From [32].

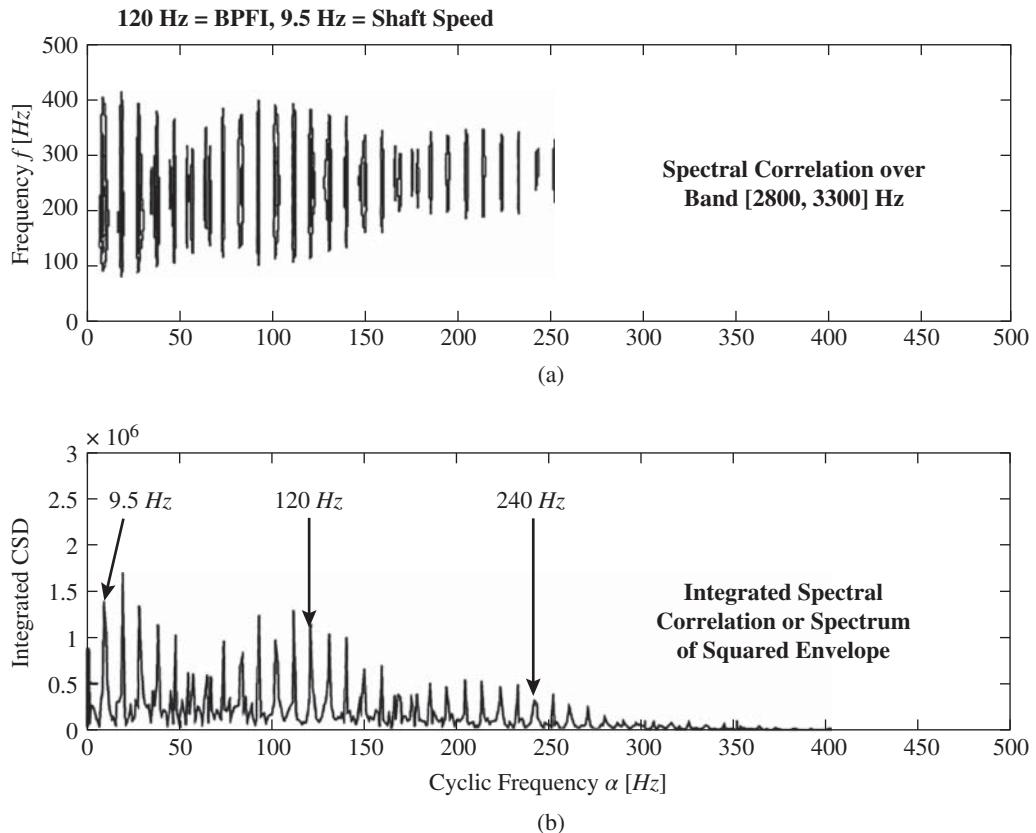


Figure 3.41 Spectral correlation and spectrum of the squared envelope for a local bearing fault.

3.6.4 Cyclo-non-stationary Analysis

Cyclostationary signals arise when stationary signals, including stationary random, are modulated by a periodic signal, so that their statistics are periodic. Thus, they are only produced by constant speed machines. When the speed of a machine varies, the response signals are no longer cyclostationary, but if the speed variation is deterministic, they can be classified as ‘cyclo-non-stationary’ if the cyclic excitations are periodic in rotation angle rather than time [38]. This is often the case with variable speed rotating machines. A major problem then is that the spacing of shaft related events, such as the meshing of gearteeth, varies with speed, whereas excited responses such as impulse responses, with constant natural frequency and damping, have constant length despite the variations in spacing.

Figure 3.43, taken from Ref. [39], shows the basic problem, when impulse responses, in this case from bearing fault signals, are repeated at a rate dependent on shaft speed. In the time domain, the pulses have the same length, but varying spacing. If this is transformed into the angle domain, by order tracking (see Section 5.1), the pulses are equally spaced, but have different length. Both the natural frequency and the damping, which have a fixed ratio, are affected in the same way. In this case, the change in apparent natural frequency is not important as it carries no diagnostic information, and is removed by enveloping, but the change in the exponential decay does mean that the repeated

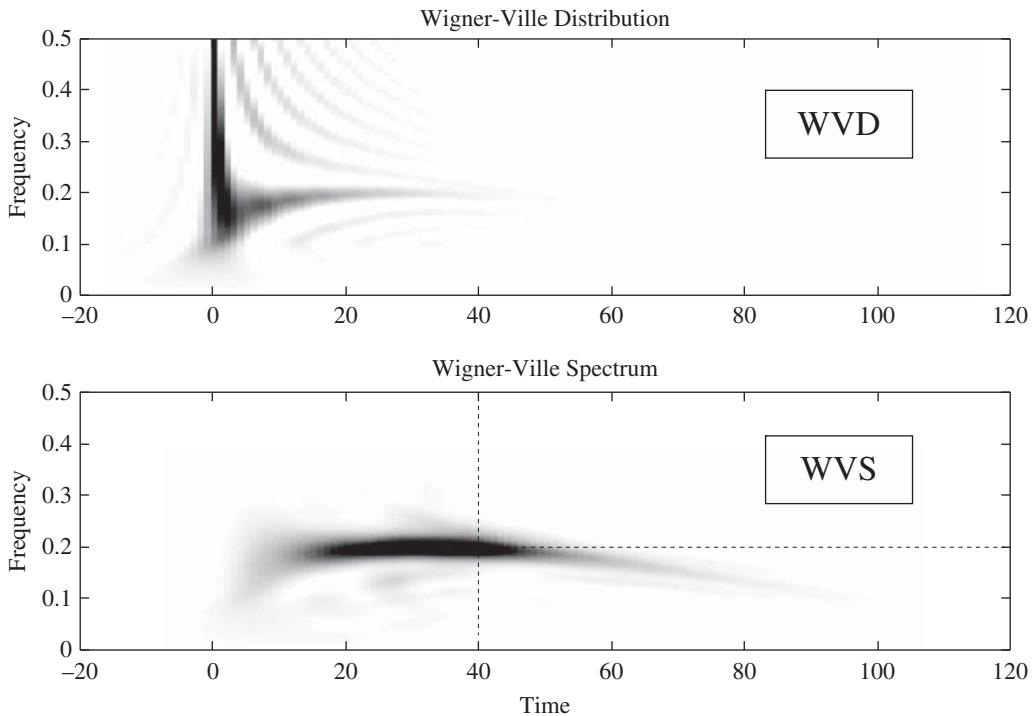


Figure 3.42 Impulse response of a nonlinear SDOF system excited by a burst random signal, expressed as a Wigner-Ville spectrum (WVS) and compared with the Wigner-Ville distribution (WVD) with interference components [37].

pulses are no longer periodic. This only has a minor effect on the envelope spectrum (in the order domain), as a small change in the number and strength of the harmonics, and there would also be a small smearing of the frequency, as it would be based on the centre of gravity of the pulses, rather than their starting points, even if these were equally spaced. However, as shown in Section 7.3.1, this is smeared in any case because of small random variations in spacing, even at constant speed.

A more formal way of dealing with the problem is developed in Refs. [38, 40]. For cyclo-non-stationary signals, a two-dimensional autocorrelation function can be generated, similar to that depicted in Figure 3.38b, but with the time t -axis replaced with angle $\theta(t)$, while retaining the τ -axis in terms of time. If the spectral correlation were directly calculated on the time signal, the resonance frequencies of the carrier would be correctly represented, but the modulating ‘cyclic frequencies’ would be smeared. If the spectral correlation operation were carried out on the order tracked signal, the ‘cyclic frequencies’ would be correct, but the carrier resonances smeared.

The generation of a frequency/order spectral correlation is illustrated in Figure 3.44, and requires a knowledge of angle θ as a function of time t . This relationship can be determined in the same way as in Section 5.1.3 for order tracking, if it is not known as an analytical equation. If the instantaneous speed $\omega(t)$ is known, then $\theta(t)$ can be obtained by integration with respect to time. Figure 3.44a shows the relationship of $\theta(t)$ with t (in this case corresponding to a linear speed ramp), and an example of a series of impulse responses at intervals of 2π in angle. Figure 3.44b, shows the corresponding autocorrelation function (asymmetric equation) with time on both axes. It is evident that a Fourier transform in the t direction would give smearing because of the non-uniform spacing. Figure 3.44c

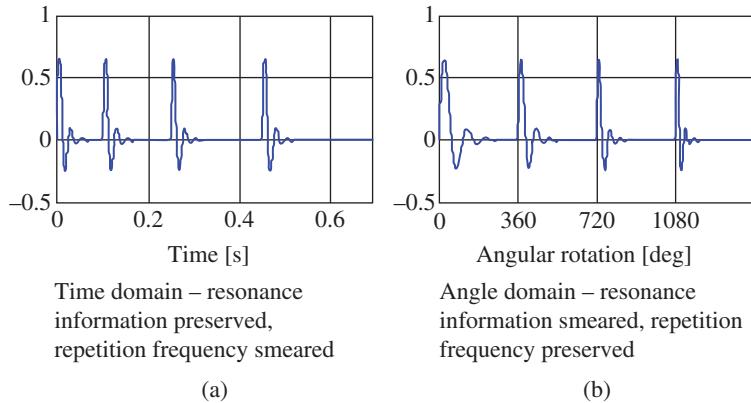


Figure 3.43 Bearing fault responses for varying machine speed [39]. (a) In time domain (b) In angle domain.

shows the angle-time autocorrelation, denoted $\mathfrak{R}_{2X}(\tau, \theta)$, where the spacings are now uniform along the θ axis. $\mathfrak{R}_{2X}(\tau, \theta)$ is given by the equation

$$\mathfrak{R}_{2X}(\tau, \theta) = E[X(t(\theta)X * (t(\theta) - \tau)] \quad (3.68)$$

and is shown in Appendix A of [40] to have a Fourier series expansion in terms of θ as long as the rate of change of speed is not too high.

Figure 3.44d shows the result of a Fourier transform in the τ direction, giving a series of PSD functions, continuous in transform frequency f but periodic with intervals of 2π in the θ direction. The Fourier transform is given by:

$$S_{2X}(f, \theta) = \int_{-\infty}^{\infty} \mathfrak{R}_{2X}(\theta, \tau) e^{-j2\pi\tau f} d\tau \quad (3.69)$$

Because of the periodicity in the θ direction, a Fourier series can be carried out transforming θ to cyclic ‘frequency’ (actually order) α , as in the equation:

$$S_{2Y}(f, \alpha) = \lim_{W \rightarrow \infty} \frac{1}{\Phi(W)} \int_{\Phi(W)} S_{2Y}(f, \theta) e^{-j\alpha\theta} d\theta \quad (3.70)$$

Once again, the background for this is given in Appendix A of [40]. It is perhaps a bit confusing that the spectra along the θ axis in Figure 3.44d seem very similar to those along the α axis in 3.44e, but where in (d) they are constrained by the periodicity to be uniformly scaled, the scaling in (e) represents the harmonic strength, which is only uniform in this diagram because the spectra in (d), sliced at a given frequency in the θ direction, are very impulsive, and localised along the θ axis, and thus have uniform harmonics.

Refs. [38, 40] refer to a couple of earlier papers giving some of the initial ideas leading to this analysis. They also give a number of examples of application of these techniques, including rolling element bearing diagnostics and gear rattle noise.

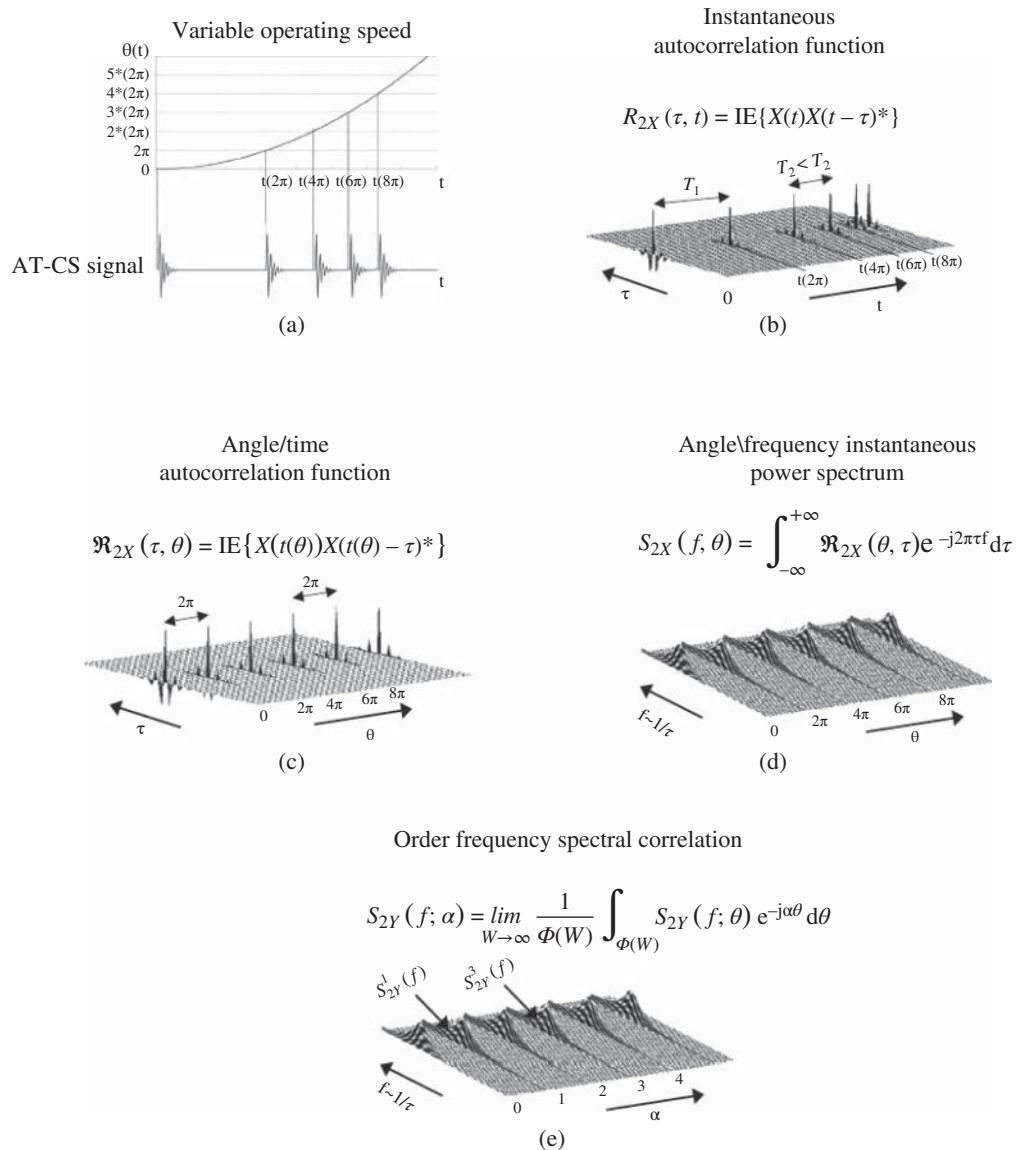


Figure 3.44 Development of order-frequency spectral correlation [40].

References

1. Papoulis, A. and Unnikrishna Pillai, S. (2002). *Probability, Random Variables and Stochastic Processes*, 4e. McGraw-Hill.
2. Laccoume, J.-L., Amblard, P.-O., and Comon, P. (1997). *Statistiques d'ordre supérieur pour le traitement de signal*. Paris: Masson.

3. Zarzoso, V. and Nandi, A.K. (1999). Blind source separation, Ch. 4. In: *Blind Estimation Using Higher Order Statistics* (ed. A.K. Nandi), 167–252. Springer Dordrecht.
4. Randall, R.B. (1987). *Frequency Analysis*, 3e. Copenhagen: Brüel & Kjaer.
5. Cooley, J.W. and Tukey, J.W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation* 19 (90): 297–301.
6. Døssing, O. (1984). Notes on modal analysis, Brüel & Kjaer, Copenhagen, Denmark.
7. Shin, K. and Hammond, J.K. (2008). *Fundamentals of Signal Processing for Sound and Vibration Engineers*. Wiley.
8. Randall, R.B. (2002). Vibration analyzers and their use, Chapter 13. In: *Shock and Vibration Handbook*, 5e (eds. C.M. Harris and A.G. Piersol). McGraw-Hill.
9. Randall, R.B. (2007). Noise and vibration data analysis. Chapter 46. In: *Handbook of Noise and Vibration Control* (ed. M. Crocker), 549–564. Wiley.
10. Bendat, J.S. (1985). The hilbert transform and applications to correlation measurements. Report prepared for Brüel & Kjaer, Copenhagen, Denmark.
11. Tribolet, J.M. (1977). A new phase-unwrapping algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-25: 170–177.
12. Wigner, E.P. (1932). On the quantum correction for thermodynamic equilibrium. *Physical Review* 40: 749–759.
13. Ville, J. (1948). Théorie et applications de la notion de signal analytique. *Cables et Transmission* 2A: 61–74.
14. Cohen, L. (1995). *Time-Frequency Analysis*. NJ: Prentice-Hall.
15. Chiollaz, M. and Favre, B. (1993). Engine noise characterisation with Wigner-Ville time-frequency analysis. *Mechanical Systems and Signal Processing* 7 (5): 375–400.
16. Staszewski, W.J. and Tomlinson, G.R. (1994). Application of the wavelet transform to fault detection in a spur gear. *Mechanical Systems and Signal Processing* 8 (3): 289–307.
17. Rubini, R. and Meneghetti, U. (2001). Application of the envelope and wavelet transform analysis for the diagnosis of incipient faults in ball bearings. *Mechanical Systems and Signal Processing* 15 (2): 287–302.
18. Peng, Z.K. and Chu, F.L. (2004). Application of the wavelet transform in machine condition monitoring and fault diagnostics: a review with bibliography. *Mechanical Systems and Signal Processing* 18: 199–221.
19. Newland, D.E. (2007). Wavelet analysis of vibration signals, Chapter 49. In: *Handbook of Noise and Vibration Control* (ed. M. Crocker), 585–597. Wiley.
20. Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory* 36: 961–1005.
21. Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81 (3): 425–455.
22. Wang, Y., He, Z., and Zi, Y. (2009). Enhancement of signal denoising and multiple fault signatures detecting in rotating machinery using dual-tree complex wavelet transform. *Mechanical Systems and Signal Processing* 24: 119–137.
23. Lin, J. and Zuo, M.J. (2003). Gearbox fault diagnosis using adaptive wavelet filter. *Mechanical Systems and Signal Processing* 17 (6): 1259–1269.
24. Schukin, E.L., Zamaraev, R.U., and Schukin, L.I. (2004). The optimisation of wavelet transform for the impulse analysis in vibration signals. *Mechanical Systems and Signal Processing* 18: 1315–1333.
25. D'Elia, G. (2008). Fault detection in rotating machines by vibration signal processing techniques. PhD Thesis. Università degli Studi di Bologna.
26. Huang, N.E., Shen, Z., Long, S.R. et al. (1971). The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis. *Proceedings of the Royal Society of London A* 454: 903–995.
27. Flandrin, P. (2007). ‘The way emd works’. Powerpoint presentation, available at <http://perso.ens-lyon.fr/patrick.flandrin/emd.html> (accessed 11 June 2018).
28. Lei, Y., Lin, J., He, Z., and Zuo, M.J. (2013). A review on empirical mode decomposition in fault diagnosis of rotating machinery. *Mechanical Systems and Signal Processing* 35: 108–126.
29. Wu, Z.H. and Huang, N.E. (2009). Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Advances in Adaptive Data Analysis* 1: 1–41.
30. Loutridis, S.J. (2004). Damage detection in gear systems using empirical mode decomposition. *Engineering Structures* 26: 1833–1841.
31. Antoniadou, I., Manson, G., Staszewski, W.J. et al. (2015). A time–frequency analysis approach for condition monitoring of a wind turbine gearbox under varying load conditions. *Mechanical Systems and Signal Processing* 64–65: 188–216.
32. Antoni, J. (2007). Cyclic spectral analysis in practice. *Mechanical Systems and Signal Processing* 21: 597–630.
33. Antoni, J. and Randall, R.B. (2002). Differential diagnosis of gear and bearing faults. *ASME Journal of Vibration and Acoustics* 124: 165–171.
34. Antoni, J. (2009). Cyclostationarity by examples. *Mechanical Systems and Signal Processing* 23: 987–1036.
35. Antoni, J., Xin, G., and Hamzaoui, N. (2017). Fast computation of the spectral correlation. *Mechanical Systems and Signal Processing* 92: 248–277.

36. Borghesani, P. and Antoni, J. (2018). A faster algorithm for the calculation of the fast spectral correlation. *Mechanical Systems and Signal Processing* 111: 113–118.
37. Antoni, J. (2003). ‘On the benefits of the Wigner-Ville spectrum for analysing certain types of vibration signals’. *Wespac8 Conference*, Melbourne.
38. Abboud, D., Baudin, S., Antoni, J. et al. (2016). The spectral analysis of cyclo-non-stationary signals. *Mechanical Systems and Signal Processing* 75: 280–300.
39. Borghesani, P., Ricci, R., Chatterton, S., and Pennacchi, P. (2013). A new procedure for using envelope analysis for rolling element bearing diagnostics in variable operating conditions. *Mechanical Systems and Signal Processing* 38: 23–35.
40. Abboud, D. and Antoni, J. (2017). Order-frequency analysis of machine signals. *Mechanical Systems and Signal Processing* 87: 229–258.

4

Fault Detection

4.1 Introduction

Fault detection is the first step in the overall process of Detection, Diagnostics, and Prognostics. Since all signals have to be processed to determine whether a significant change has occurred, the techniques employed must be considerably more efficient than those which might be used for the latter two processes. It is typical for there to be significant changes in only about 2% of the cases analysed. Where frequency spectra are involved, the type of frequency analysis used also has greatly different requirements than for diagnostics. For a start, the faults may show up at any frequency over a wide range, whereas once they are detected it is possible to concentrate in detail on the relevant frequency range. Moreover, since the faults are detected by changes in level, the measurement accuracy is somewhat more important than in diagnosis, where frequency accuracy and frequency patterns are more important.

Faults reveal themselves in very different ways in rotating and reciprocating machines, so each type is discussed separately.

4.2 Rotating Machines

4.2.1 Vibration Criteria

The simplest, though not the most reliable way, to detect faults in machines is to compare their vibration levels with standard criteria for Vibration Severity. There are a number of these, based on the original Rathbone and Yates charts, which were determined largely empirically. Rathbone [1] asked experienced engineers to judge the vibration severity of a large number of operating machines, using their finger to gauge the vibration level. He made measurements at the same time with a simple vibration meter, and in plotting the average results in terms of vibration displacement found that equal severity was represented by sloping curved lines. These represented constant velocity over the major part of the frequency range, but tended towards constant displacement at low frequencies and constant acceleration at high frequencies. Yates [2], working with the British Navy, realised that a constant velocity criterion was justified theoretically, and so his charts (based largely on marine steam turbines) were set on this basis. He argued using dimensional analysis that for a given geometric shape, an object would have the same stress when vibrating at the same velocity level in a given mode at its natural frequency, independent of size.

Of course, any vibration is a continuous alternation between potential (strain) energy, proportional to the square of stress, and kinetic energy, proportional to the square of velocity, and so vibration velocity corresponds most closely to stress over a wide range of machine sizes and speeds, and also over a frequency range with a given machine. The maximum strain energy and kinetic energy are not at the same location; with a cantilever beam, for example, the maximum strain is at the built-in end with permanent zero velocity, while the maximum velocity is at the free end, with permanent zero strain, and the maxima occur 90° out-of-phase in the vibration cycle.

The equivalence of maximum stress and velocity is the fundamental reason why virtually all vibration criteria are now expressed in terms of vibration velocity, although not all are in agreement in detail. In the original investigations it was somewhat fortuitous that the human finger tends to evaluate vibration level in terms of velocity, but it did turn out that this had justification in terms of internal stresses.

A commonly used criterion chart, in particular in the USA, is the General Machinery Criterion Chart, Figure 4.1, proposed by the company IRD Mechanalysis, which is for displacement or velocity measurements at the bearing cap. It has its origins in the Rathbone and Yates criteria, and is reported in [3]. The lines of constant severity, ‘Smooth’, ‘Very good’, ‘Good’, etc. are lines of constant velocity, with the conversion from displacement assuming that the vibration in terms of displacement is dominated by the value at the shaft rotating speed. The values are given as peak-to-peak for displacement, and zero-to-peak for velocity, these being respectively $2\sqrt{2}$ and $\sqrt{2}$ times the root mean square (RMS) value for sinusoidal signals. Most instruments would actually measure an approximation of RMS value, but the assumption is tacitly made that the assumed signal at the shaft speed is sinusoidal.

Note that there is always a factor of two involved in going from one severity class to the next, i.e. a constant interval of 6 dB, and that logarithmic axes are employed. This means that a ‘linear’ progression in severity corresponds to a linear change on a log scale, and thus an exponential change in linear values. Somewhat illogically, this point was rarely taken into account in plotting trends of measured values, except in the cases where the criteria were expressed in terms of velocity decibels (VdB). The latter was done from an early stage by the US and Canadian navies (e.g. [4, 5]).

The Rathbone and Yates criteria were based largely on the vibration of steam turbines, the most critical machines at the time, though Yates’ machines, mounted in somewhat more flexible ship structures, allowed slightly higher values than Rathbone’s largely land-based machines. In 1957 the German Engineers’ Association, Verein deutscher Ingenieure (VDI), produced a set of criteria in the recommendation VDI 2056, which adjusted levels according to the size and type of support of the machines. This gave different sets of levels for four machine classes, Small, Medium, Large on a rigid foundation, and Turbomachines (large on a flexible foundation). In 1974 these recommendations were incorporated into the ISO standard ISO 2372, and the latter was first replaced by ISO 10816, for which the major component (Part 1) still had the same structure, as reported in the first edition of this book. It was superseded by ISO 20816 in 2016, where significant changes were made to the general chart for Part 1, reproduced in Figure 4.2 [6]. It has been recognized that there can be significant differences between different machines, of different types, and even within the same type, so rather than specifying fixed limits for the different severity zones (A, B, C, D), the recommendation is now that the borders between the zones should lie within a certain range. When used as acceptance criteria they should be agreed between the manufacturer and customer. More specific limits are found for different machine types in different sub-parts of the standard.

It is obvious that there will be differences in detail between a set of criteria that makes no distinction between machine classes (Figure 4.1) and one that does (Figure 4.2), but what is more important are the points on which they agree. The first is that equal changes in severity are represented by equal changes on a log amplitude scale. There is a larger number of severity zones (9) in the General Machinery Criterion Chart of Figure 4.1, with ratios of 2 (6 dB) between them, whereas for any

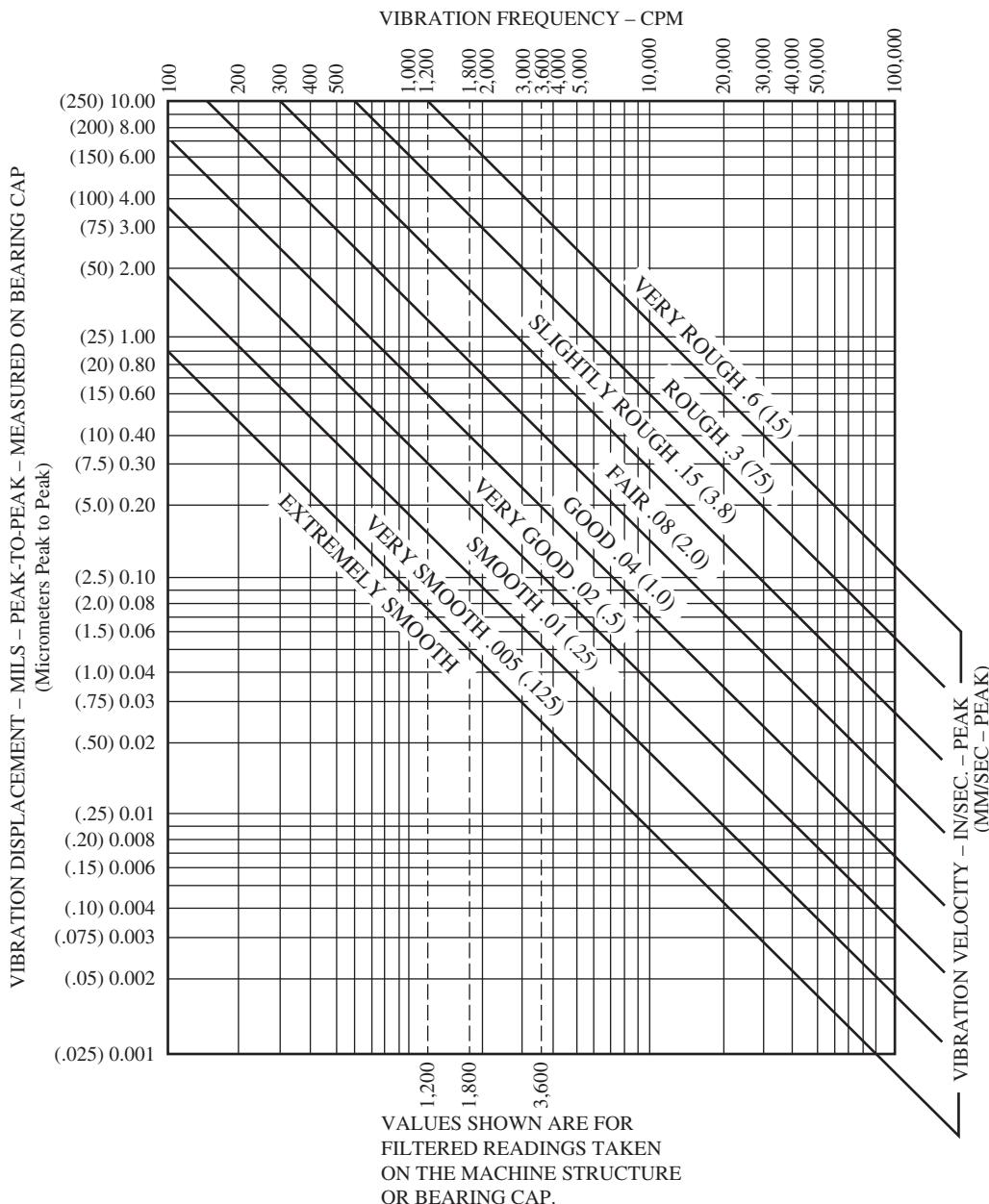


Figure 4.1 General machinery criterion chart [3].

given type of machine there are only four zones in ISO 20816, but these are separated by a factor of 2.5 (8 dB). However, in both sets of criteria a change by a factor of 10 (20 dB) would go from 'good' to 'not permissible'. It can thus be concluded that a change by a factor of 2 to 2.5 (6–8 dB) is significant, and that a change by a factor of 10 (20 dB) is serious. Other reasons for evaluating vibration levels on a logarithmic or dB scale are as follows:

Range of typical zone boundary values for non-rotating parts r.m.s. vibration velocity mm/s			
0,28			0,28
0,45			0,45
0,71			0,71
1,12	Zone boundary A/B 0,71 to 4,5		1,12
1,8			1,8
2,8	Zone boundary B/C 1,8 to 9,3		2,8
4,5			4,5
7,1	Zone boundary C/D 4,5 to 14,7		7,1
9,3			9,3
11,2			11,2
14,7			14,7
18			18
28			28
45			45

NOTE 1 This table only applies to machines for which specific International Standards have not been developed and for which there is no suitable experience available.

NOTE 2 Small machines (e.g. electric motors with power up to 15 kW) tend to lie at the lower end of the range and large machines (e.g. prime movers with flexible supports in the direction of measurement) tend to lie at the upper end of the range.

Figure 4.2 Table C.1 from ISO Standard 20816-1 [6] giving the recommended ranges of boundary values for different vibration severity zones, in terms of RMS velocity measured on non-rotating parts of rotating machines in different classes. By permission of Standards Australia on behalf of ISO under Licence CL2020rbr.

- Vibrations are external manifestations of internal forces and stresses. The user is actually interested in the latter, not the former. A change by a certain number of dB at a source will give the same dB increase at all measurement points (from that component alone) independent of the transfer path, or the original vibration level at that measurement point.
- A change by a certain factor, or number of dB, corresponds to a certain design safety factor based on maximum stress, once again independent of the measurement point.

Even so, individual machines can vary a lot from the average. As far back as the late 1960s, Woods found in his PhD work [7], from which many results were published in an ASME paper with his supervisor Prof. Downham [8], that 14 machines in a petrochemical plant had bearing (acceleration) impedances varying over a range of 1000 : 1 (60 dB), even though they would all have been classified as the same type according to VDI 2056 or ISO 2372. These results are illustrated in Figure 4.3. This would mean that the same externally measured vibration level would imply that the internal forces varied over that range, or alternatively that the same internal forces on different machines would give a correspondingly large range of vibration responses. It is possible that the machines had been purchased from different suppliers over a number of years, and represented different design philosophies, but this would be the case for machines in general to which it is hoped to apply criteria such as ISO 20816.

These criteria are often used in purchasing specifications, and over the years would likely have given rise to more uniformity over machines of the same class, in particular for machines such as electric motors where there is a very large market and much competition, but it is still unwise to place too much credibility in whether a particular machine, measured at a particular point, meets certain criteria or not.

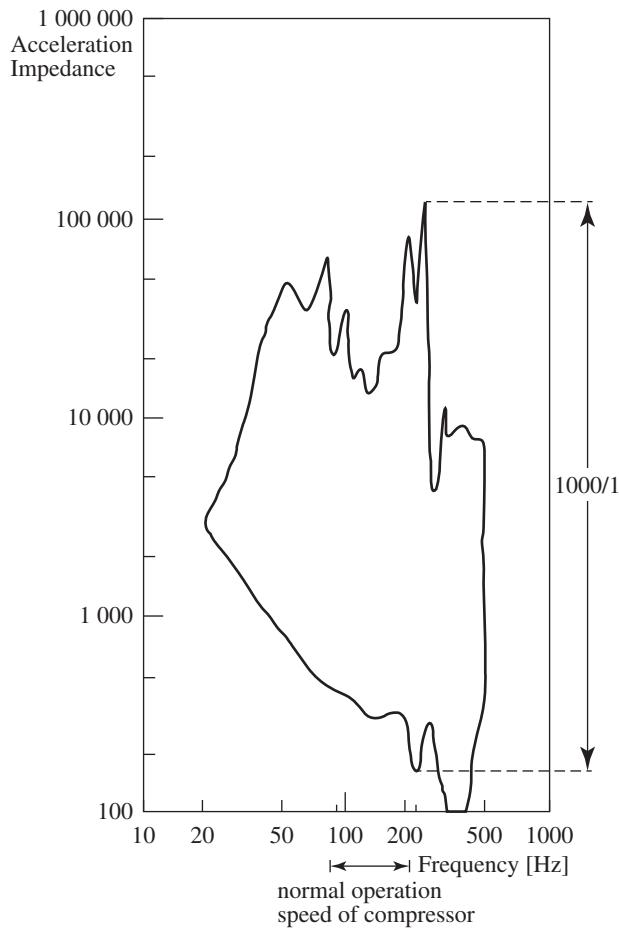


Figure 4.3 Overall impedance envelope measured on 14 machines in an ethylene plant [8].

Thus, as recommended by Downham and Woods, a more reliable way to detect faults is to detect changes in the vibration level at each measurement point on each machine (and this is now given as an alternative in ISO 20816). When they are in good condition, the vibration levels will normally remain stable (when operated at the same load and speed), and it is changes in level that are indicative of changing condition.

Note that most criteria are based on total (RMS) vibration velocity levels covering a range from 10 to 1000 Hz, but this is mainly because they were initially based on measurements using a vibration meter with a velocity transducer limited to this range. There is no technical reason (except for greater variability in the potential sources) why the arguments for using vibration velocity should not apply equally over the range 10 Hz–10 kHz, for example. This is quite easily achieved using accelerometers with electronic or numerical integration (see Section 1.5 of Chapter 1).

4.2.2 Use of Frequency Spectra

The use of velocity means that there is a better chance than otherwise that changes at any frequency will affect the overall RMS levels, but it is still evident that monitoring of frequency spectra, rather than overall levels, will have a better chance of detecting changes at whatever frequency they should

occur. Cases abound (see examples below) where significant changes in individual frequency components only affect the overall vibration levels at the very last stages, if ever. It thus appears that it is advisable in general to detect changes in vibration spectra, rather than overall RMS levels, though this does engender additional problems.

The first problem is that faults may occur at any frequency over a very wide range, typically three decades or more. With fluid film bearings the range can extend down to 40% of the lowest shaft speed (e.g. oil whirl) up to at least the 400th harmonic of the highest shaft speed (e.g. harmonics of gearmesh and bladepass frequencies). Rolling element bearings often have fault indications at frequencies of the order of 1000 or more times the shaft speed. The linear frequency scale and constant frequency resolution of Fast Fourier Transform (FFT) analysis means that the highest decade occupies 90% of the scale, and the highest two decades, 99%. To adequately cover the full frequency range of three decades or more, it would generally be necessary to perform three FFT analyses, each separated by a decade, with full scale frequencies of for example 100 Hz, 1 kHz, and 10 kHz. On the other hand, a constant percentage bandwidth (CPB) spectrum, with 3% bandwidth (1/24th octave), can cover three decades with only 240 filters.

Another advantage of using CPB spectra, with logarithmic frequency axes, derives from their uniform resolution along those axes. For example, a speed change of 3% would correspond to the same lateral shift of the spectrum at all frequencies. With FFT spectra, a constant percentage speed change is very difficult to compensate for, as it represents different numbers of lines at different frequencies. A 3% change in an 800-line spectrum, for example, is 24 lines at full scale, 12 lines at half scale, and so on.

Direct digital comparison of spectra is not straightforward, mainly because of ‘undersampling’ of discrete frequency peaks. This does not mean that information is lost; just that if the speed changes less than the amount required to shift a peak from one line to the next, the samples along the flanks of discrete frequency components can vary by large amounts, often exceeding 10 dB. If a direct digital comparison is made it will appear that corresponding changes in the spectra have occurred, even though the peak values are almost unchanged. A typical example is shown in Figure 4.4 for a comparison of FFT spectra from a machine in unchanged condition, where the speed was stable to within 0.25%.

This applies equally to CPB or FFT spectra. A simple way of preventing this happening is to make comparisons with a ‘mask’ spectrum, obtained from the original reference spectrum by displacing it to each side, and taking the envelope. This is illustrated in Figure 4.5. In this case the dynamic range of the reference spectrum has also been limited (to 40 dB). This reflects the fact that frequency modulation (FM) tape recordings typically had a dynamic range of this order, but with modern data acquisition systems the dynamic range is typically doubled, and it may no longer be necessary to limit the dynamic range of the reference spectrum. An alternative way of obtaining a mask spectrum is to use the upper envelope of all measured spectra over a long enough period that all normal minor speed and load variations have been encountered, but not so long that the condition has changed. This approach would be most viable for permanent monitoring systems where there is the potential to collect large amounts of data. The mask value would then normally represent the mean value plus two to three standard deviations.

4.2.3 CPB Spectrum Comparison

A very efficient way of performing spectrum comparison for fault detection of rotating machines was presented in Refs. [9, 10]. It is based on using CPB spectra of 3 or 4% bandwidth (1/24 or

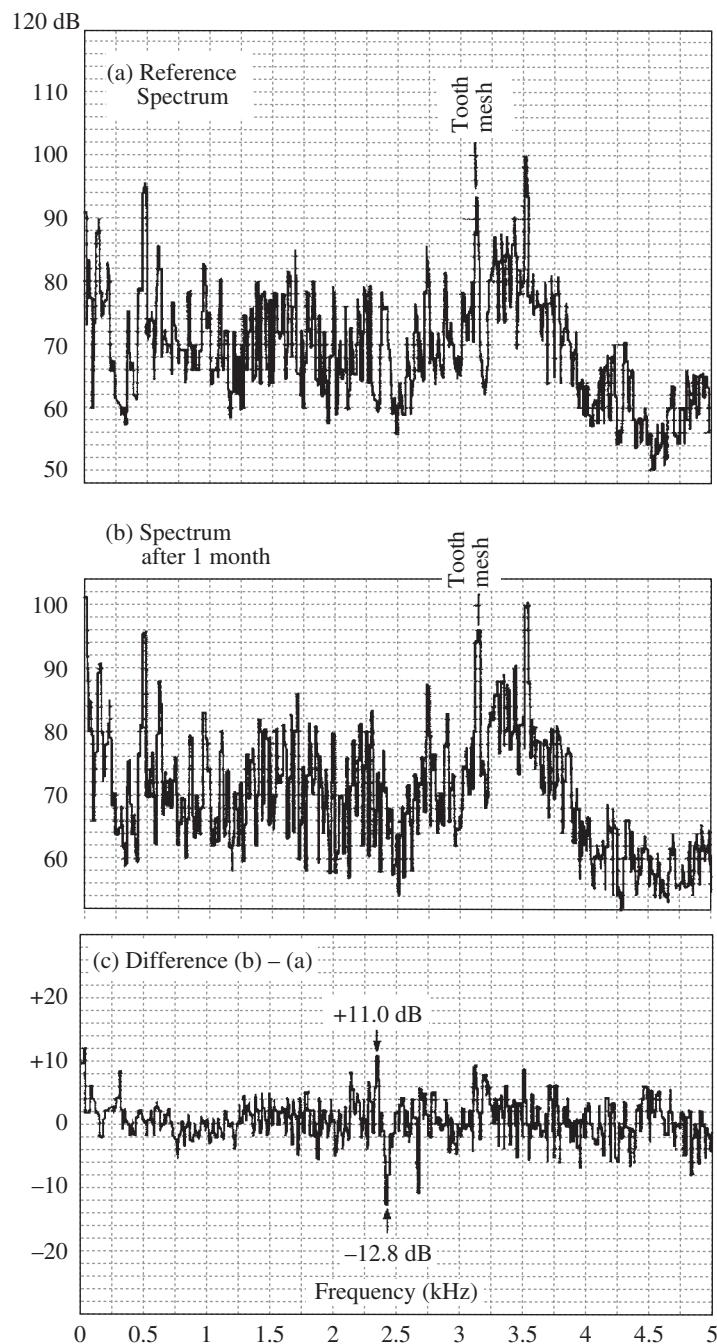


Figure 4.4 Direct digital comparison of two spectra with no change in machine condition.

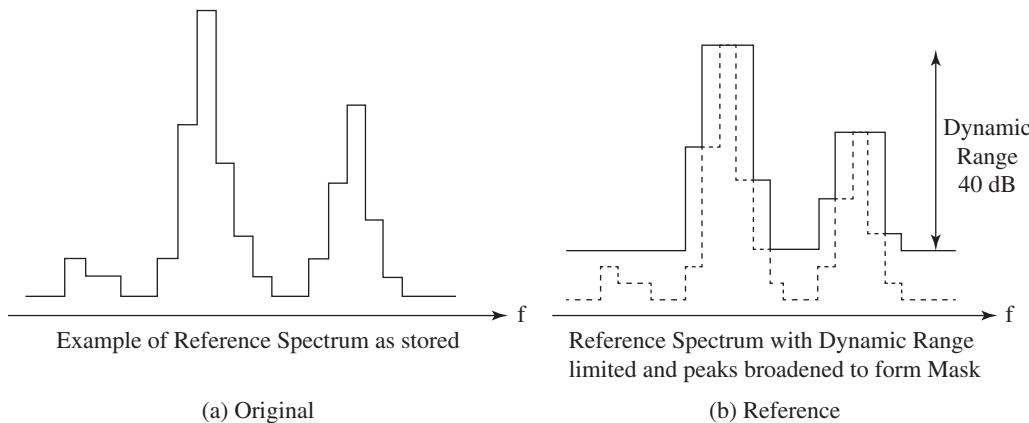


Figure 4.5 Generation of a mask spectrum from an original measured spectrum.

1/18-octave), and covering three or more decades in frequency. It is simplest to express the spectra in terms of VdB, with a suitable reference level for the dBs being $1 \times 10^{-8} \text{ m s}^{-1}$, but they can be expressed in velocity units (e.g. mm/s) on a logarithmic scale. Because reference levels are not universal, they must be stated. For the reasons given above it is best to express the spectra in terms of velocity, but since each frequency is now treated separately, it is possible to use acceleration directly, for example expressing the spectra in acceleration decibels (AdB), with a typical reference value of $1 \times 10^{-6} \text{ m s}^{-2}$, even if the spectra are sloping.

Speed changes up to half the bandwidth (<1.5%) will be accounted for by the use of a spectrum mask as described in the previous section, but even larger changes can be compensated by a lateral shift of the spectrum by rounding to the nearest integer number of bandwidths. This is often valid for changes up to about 10%, but is signal dependent. The compensation referred to is for harmonics of the dominant shaft speed, and will primarily apply to low harmonics of shaft speed and higher harmonics such as gearmesh frequencies, bladepass frequencies, and slot-pass frequencies in electric motors. Spectral peaks dominated by resonances will not change with shaft speed, even as individual harmonics pass through them, and this sets a limit on how much compensation can be made without generating a new reference spectrum for the altered conditions.

Figure 4.6 shows a comparison of FFT spectra for the same machine as Figure 4.4, but some months later when there was not only a change in condition, but also a speed change of the order of 4%. It is now evident that a direct spectrum comparison is meaningless, primarily because the same components (e.g. toothmesh frequency) are in different spectral lines.

Figure 4.7 shows a comparison for the same machine over a longer period, using CPB spectra, and with a lateral shift of one line in the last case (corresponding to Figure 4.6). An automatic readout lists changes greater than 6 dB (considered significant). It is measured on the gearbox between an electric motor running at approximately 50 Hz and a centrifugal compressor at 121 Hz.

Figure 4.7a shows the original CPB spectrum to be used as reference and Figure 4.7b the mask spectrum derived from it in the way described in Figure 4.5. Figure 4.7c shows the same comparison as Figure 4.4, and it is seen that the correct result of 'no significant change' is obtained. Figure 4.7d and e from the following two months show the gradual growth of some frequency components representing a change in condition. In the case of Figure 4.7d the maximum change (of the high speed shaft component) is by about 14 dB, and is thus more than significant.

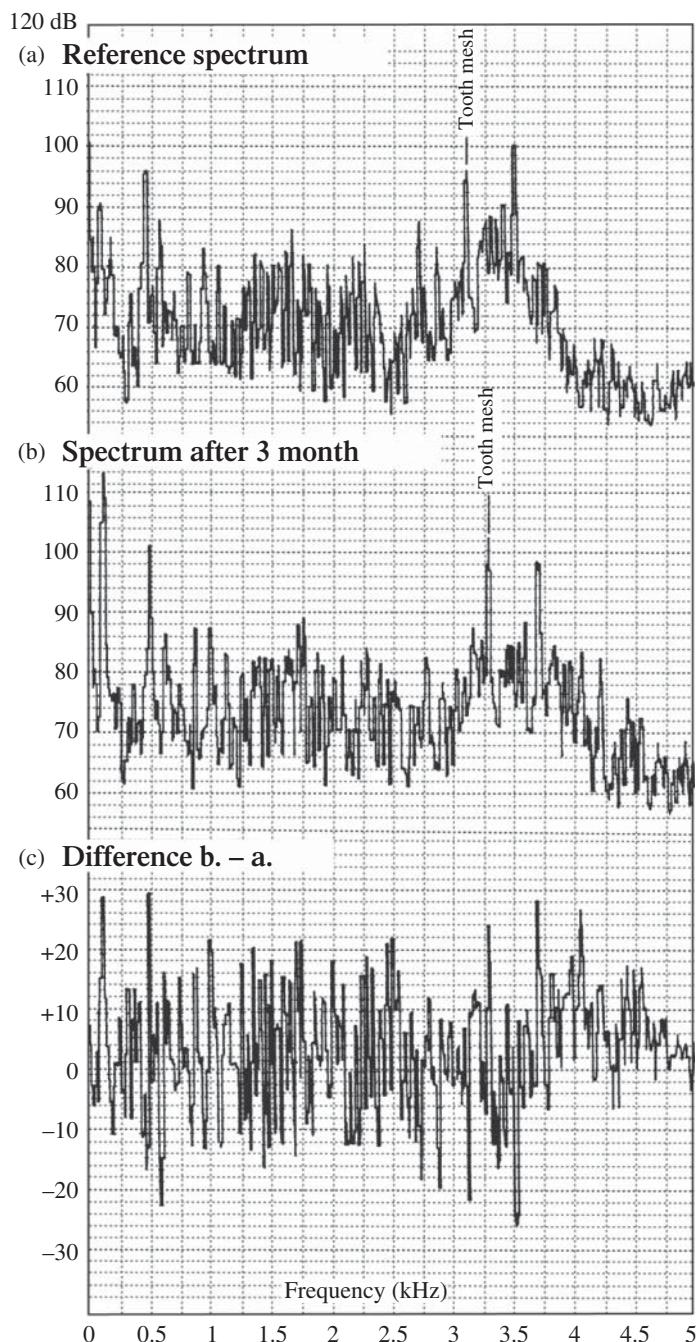


Figure 4.6 FFT spectrum comparison with a small speed change.

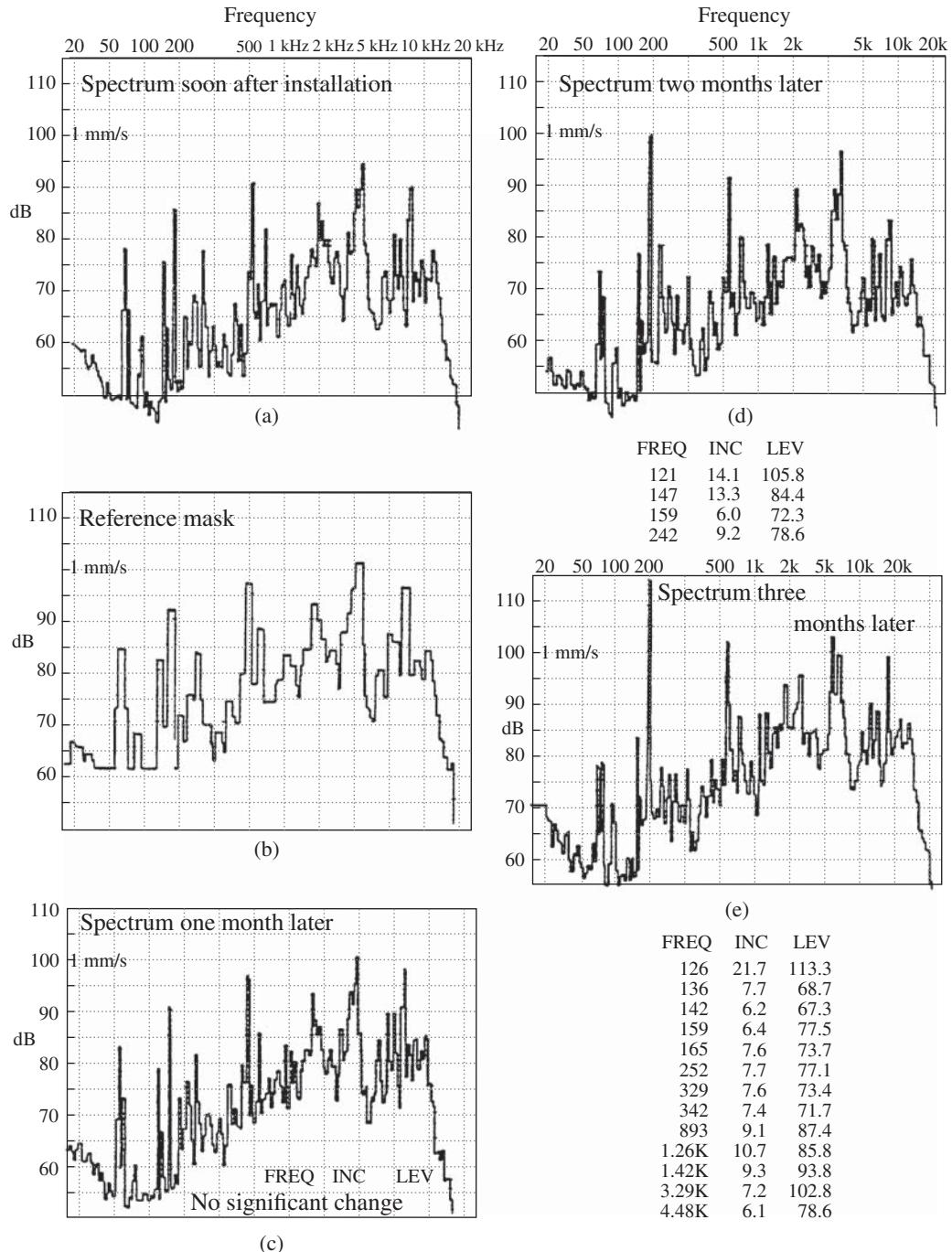


Figure 4.7 Spectrum comparison using CPB spectra.

Figure 4.7e represents the same comparison as Figure 4.6, and it is seen that once the compensation for speed change is made, the truly significant spectral components become apparent. It should perhaps be mentioned that because the drive was by an induction motor, the speed did not in fact change by 4%, but the last recording was made using an FM tape recorder with low battery voltage, so that it was running much slow. It does serve as an example of how genuine speed changes can be compensated for. The major change (at about 123 Hz) is now more than 20 dB, so could be considered quite serious. After diagnostic analysis, it was confirmed that it coincided with the speed of the compressor shaft, and that the second and fourth harmonics were quite high.

The machine was shut down for repair shortly after, and it was found that the increases in vibration were due to increasing misalignment, brought about by failure of the grouting in which the gearbox was set. This was in turn due to the fact that the fourth harmonic of the compressor shaft excited a resonance of the coupling between the gearbox and compressor, this being confirmed by a resonance tap test. The bolted rim flanges of the coupling were excited by a hammer blow in the axial direction (with the machine stationary), and the dominant resonance frequencies extracted from the response spectrum.

Note that the overall vibration level would only have changed significantly at the time of the last measurement, and then only by <10 dB.

Another couple of examples will serve to show the versatility of the technique, as well as the advantage given by the comparison of spectra instead of overall levels.

Figure 4.8 shows an example [11] from measurements made on an auxiliary gearbox mounted on a gas turbine driven oil pump on the Trans Alaska Pipeline.

It shows the comparison of a spectrum with the mask formed from the original reference spectrum, and the resulting spectrum of exceedances. This comparison is for the situation four months after the first detection of the fault. The maximum change of 20 dB is quite serious, but stabilised at

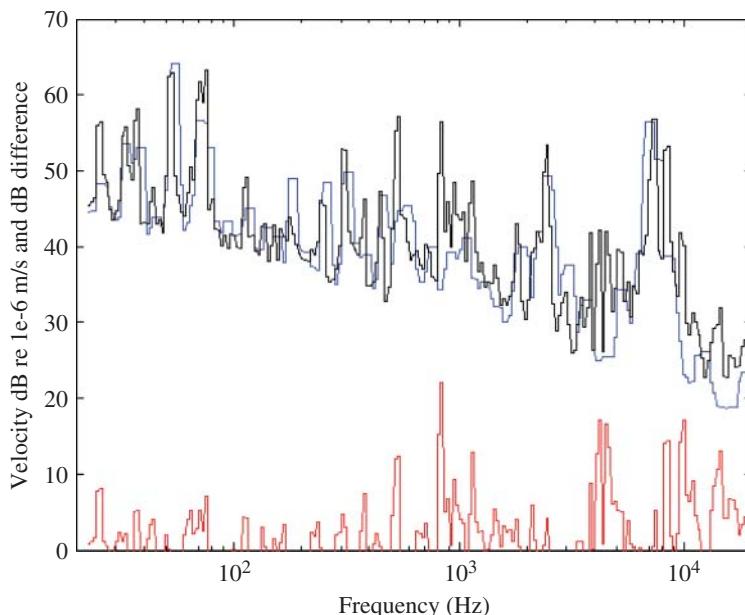


Figure 4.8 (Upper) Comparison of new spectrum with mask, velocity dB re 1×10^{-6} m s $^{-1}$. (Lower) dB difference spectrum.

COMPARISON SPECTRUM AND DIFFERENCE.

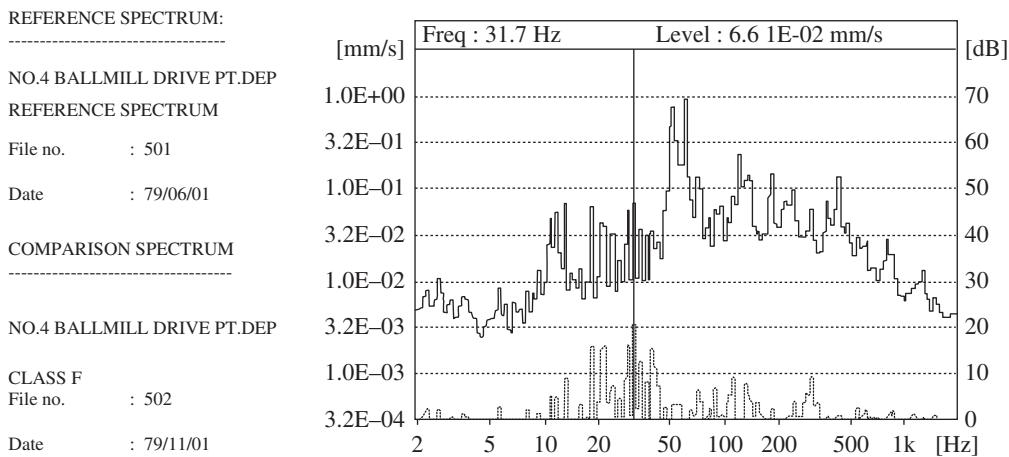


Figure 4.9 Spectrum comparisons for a ball mill in a copper mine.

this level, and the machine was allowed to run for a further five months before being repaired at a convenient time. Note that despite the significant change of a number of frequency components, the overall RMS value of the signal would not have changed, because of the masking effect of strong adjacent components. The fault was in a bearing, but the spectrum was dominated locally by strong gear-related components.

Figure 4.9 shows an example from a much lower speed machine, a ball mill in a copper mine. The drive pinion speed is 2.65 Hz, so the frequency range monitored was adjusted to the three decades 2 Hz–2 kHz. This was one of ten identical ball mills being monitored with a relatively new monitoring system. On one occasion (1st November, 1979) two mills (No. 4 and 8) showed significant changes of up to 20 dB and 14 dB, respectively, but because of lack of experience with the system, the machines were allowed to continue. Two and a half weeks later, No. 8 had a serious breakdown due to cracked gear teeth on the drive pinion. One separated tooth passed through the mesh and caused serious damage (although the biggest potential cost, the ring gear, was able to be repaired). No further monitoring had been done, so it is not known how great the exceedances just prior to breakdown were.

On the failure of Mill No. 8, No. 4 was immediately stopped and found to have similar cracking of a number of teeth. From that point on, much more credence was given to the output of the monitoring system. No diagnostic information can be obtained from the CPB spectra, but once the fault had been detected, FFT analyses could be made in an appropriate frequency band. Figure 4.10 shows a comparison of FFT spectra for one of the mills in original condition, and prior to failure.

It is seen that the spectral changes are primarily at widely distributed harmonics of a frequency (which corresponds to the speed of the input pinion), most of which can be interpreted as sidebands around the first two harmonics of the gearmesh frequency. Such a wide distribution of sidebands is typical of a localised fault on a gear as explained in Section 2.2.2.2 (Figure 2.13).

Once again, the gear failed without any increase at all in overall vibration levels, since the latter were dominated by components at the harmonics of the toothmesh frequency.

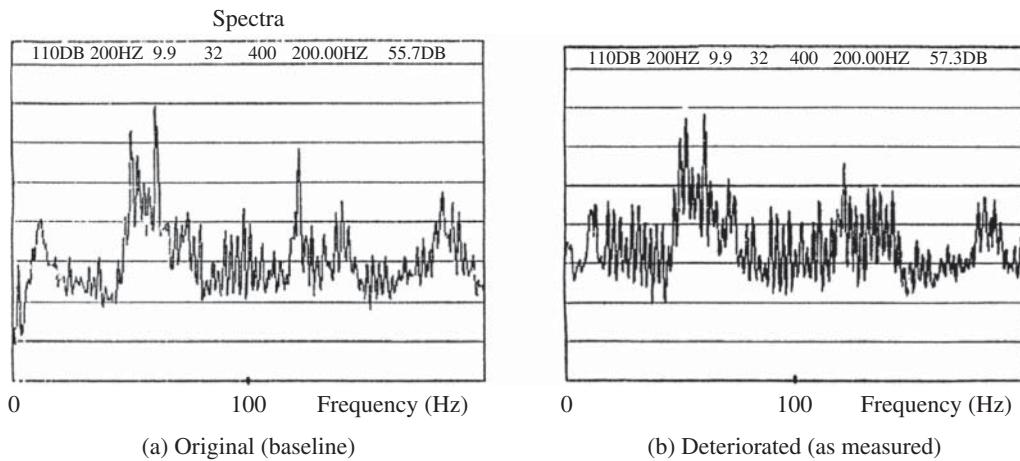


Figure 4.10 FFT spectra for the ball mill drive in original and deteriorated condition.

4.3 Reciprocating Machines

In Section 2.3.1 it was explained how the signals from reciprocating machines differ from those of rotating machines in that they vary in both time (crank angle) and frequency. Another point of difference for fault detection is that decreases may be just as important as increases (e.g. in the case of a misfire). For the rotating machines, the use of a mask virtually ensures that decreases will be found at some frequencies, but faults invariably give increases somewhere in the spectrum (or at least are not characterised only by spectrum decreases).

Even so, as discussed in the next section, vibration criteria do exist for reciprocating machines treated in essentially the same way as rotating machines.

4.3.1 Vibration Criteria for Reciprocating Machines

ISO Standard 10816 Part 6 is entitled ‘Reciprocating machines with power ratings above 100 kW’ and has a somewhat similar format to ISO 10816 – Part 1 for rotating machines. One major difference is that limits are given for all three parameters, displacement, velocity, and acceleration. Figure 4.11 is an extract from ISO 10816 Part 6 [12] which shows the major recommendations. This standard has an amendment published in 2015. Note that it does not cover reciprocating compressors, which were originally covered by ISO 10816 Part 8, which has now been replaced by ISO 20816 Part 8.

There is no definition of the ‘Machine Vibration Classification Number’ used in the table, but the remark is made that ‘many industrial and marine diesel engines may be classified in either classification number 5, 6, or 7’. It appears that the classification number is to be agreed between the parties using the standard.

The Standard makes the point that the limiting values are intended not only to protect the machine itself, but also auxiliary equipment mounted on it, although since the latter may be mounted in positions not represented by the preferred measurement points, special analyses may be required to avoid resonant amplification, etc. at those mounting positions. The point is also made that the vibration measurements will give little information on torsional vibrations, which may be the primary indicator of some types of faults.

Vibration severity grade	Maximum values of overall vibration measured on the machine structure			Machine vibration classification number						
	Displacement μm (r.m.s)	Velocity mm/s (r.m.s)	Acceleration m/s^2 (r.m.s)	1	2	3	4	5	6	7
Evaluation zones										
1,1	17,8	1,12	1,76							
1,8	28,3	1,78	2,79	A/B						
2,8	44,8	2,82	4,42		A/B					
4,5	71,0	4,46	7,01			A/B				
7,1	113	7,07	11,1	C						
11	178	11,2	17,6		C					
18	283	17,8	27,9			C				
28	448	28,2	44,2	D			C			
45	710	44,6	70,1		D			C		
71	1125	70,7	111			D			C	
112	1784	112	176				D		C	
180								D		D

Key to zones

A: The vibration of newly commissioned machines would normally fall within this zone.

B: Machines with vibration within this zone are normally considered acceptable for long-term operation.

C: Machines with vibration within this zone are normally considered unsatisfactory for long-term continuous operation. Generally, the machine may be operated for a limited period in this condition until a suitable opportunity arises for remedial action.

D: Vibration values within this zone are normally considered to be of sufficient severity to cause damage to the machine.

NOTE — Vibration values for reciprocating machines may tend to be more constant over the life of the machine than for rotating machines. Therefore zones A and B are combined in this table. In future, when more experience is accumulated, guide values to differentiate between zones A and B may be provided.

Figure 4.11 Table A.1 from ISO Standard 10 816-6 [12] giving recommended vibration limits for reciprocating machines. By permission of Standards Australia on behalf of ISO under Licence CL2020rbr.

4.3.2 Time/Frequency Diagrams

Just as comparison of frequency spectra greatly increases the chances of detecting faults in rotating machines, the comparison of time/frequency diagrams for reciprocating machines increases the effectiveness with respect to both overall vibration levels, and also spectrum comparison alone. For example, Figures 2.26 and 2.27 of Chapter 2 show the advantages of time/frequency diagrams over simple spectrum comparison.

The question arises as to the best way to make the comparison of the time/frequency diagrams. Arguments can be made for linear frequency scale averaged power spectra, as in Figure 2.26, or logarithmic frequency scale CPB spectra as for rotating machines. The difference is that each spectrum in the time/frequency diagram is time windowed, with the consequent limitation in minimum resolution bandwidth. An FFT-based spectrum will thus always have the maximum resolution possible (except perhaps for the Wigner-Ville spectrum [WVS]). On the other hand, since most events in a reciprocating machine cycle are impulsive (and thus in any case of short time duration), the spectra are relatively broadband, and for roughly constant damping factor most efficiently expressed on a log frequency scale where resonances will tend to have the same width (and same number of lines for CPB spectra).

For most of the examples shown in this section, the time/frequency diagrams have a linear frequency scale coming from the FFT-based analysis of the windowed time records (as shown in Figure 2.25), but the actual comparisons of the diagrams are made on the basis of spectra converted to 1/3-octave bandwidth, by the same process as indicated in Figure 3.32 (but for one decade). The use of 1/3-octave bandwidth (23%) allows for minor speed changes without having to use a mask, and thus allows the detection of decreases as well as increases in sections of the diagram. An FFT-based spectrum can be converted to 1/3-octave over $1\frac{1}{2}$ decades, or in other words, 15 lines.

Comparisons are made of logarithmic amplitude values for the same reasons as in Section 4.2.1, and differences are expressed in dBs. As in the case of rotating machines, a significant change is considered to be 6 dB, and increments after that are in 3 dB steps. From experience with this approach, it appears that local increases of 20 dB are not as serious as they would be in spectrum comparison of rotating machines.

Figure 4.12 is a repeat of Figure 2.26 with the inclusion of the results of comparison, taking the double-acting condition as the reference. In this case, because of the absence of signals for the reverse stroke, the differences are all negative (indicated by letters as opposed to numbers).

Figure 4.13 is a similar comparison for the machine of Figure 2.28. Measurements were made at the optimum position determined from that diagram, namely below the cylinder head gasket. The fault in this case was a misfire in the cylinder on which the measurements were made. It is interesting that since the cylinders were physically separated, there is no contamination by the signals from adjacent cylinders, which were functioning normally. For security, there were two spark plugs in each cylinder, and both had to be disconnected to cause the misfire. Once again, the only differences are negative, because of the absence of an event in the normal cycle. Interestingly, the major effect appears to be the absence of noise from gas flow through the exhaust valve, at around 90° crank angle, rather than the absence of a pressure pulse from the ignition of the air/gas mixture, which would be around the reference position of 0° . This is presumably because the combustion in such gas engines, with a low pressure ratio, is relatively gentle. Another smaller difference (though the one that dominated spectra averaged over the complete cycle) corresponds to the closing of the exhaust valve just before 270° crank angle. This apparently gives a larger impact with the higher exhaust pressure.

Figure 4.14 gives an example that really shows the benefits of this approach. It is a comparison of signals from a small 4-cylinder diesel engine with normal and under-size pistons. The reduction in piston diameter was only just outside tolerance and was intended to represent a small increase in piston slap. In this case there are increases of between 6 and 12 dB, in the vicinity of each top dead centre (TDC), where the combustion pulses are also located. By inspection of the diagram, it can be seen that the piston slap excites a greater response at around 10 kHz, while the unchanged combustion effect is at lower frequencies.

Figure 4.15 shows a comparison of sound signals recorded for the same case, where no effective difference was found. This indicates that it is unlikely that a mechanic would be able to hear the difference, and thus the method using vibration measurements is more sensitive. Presumably, for larger increases in piston clearance, the difference would also have been seen in the sound signals.

In cases such as these, it can be said that the fault detection also can be used as a basis for diagnostics, since there is so much information contained in a time/frequency diagram. With some experience, it could be established that changes in a particular part of a time/frequency diagram were most likely to correspond to a particular fault. The author considers that time/frequency comparisons of the type shown in Figures 4.12–4.14, based on 1/3-octave spectrum resolution, are simple enough to be automated, and applied fairly easily to large numbers of signals. Once a fault has been detected in this way, more detailed analysis could be carried out to refine the diagnosis. This might for

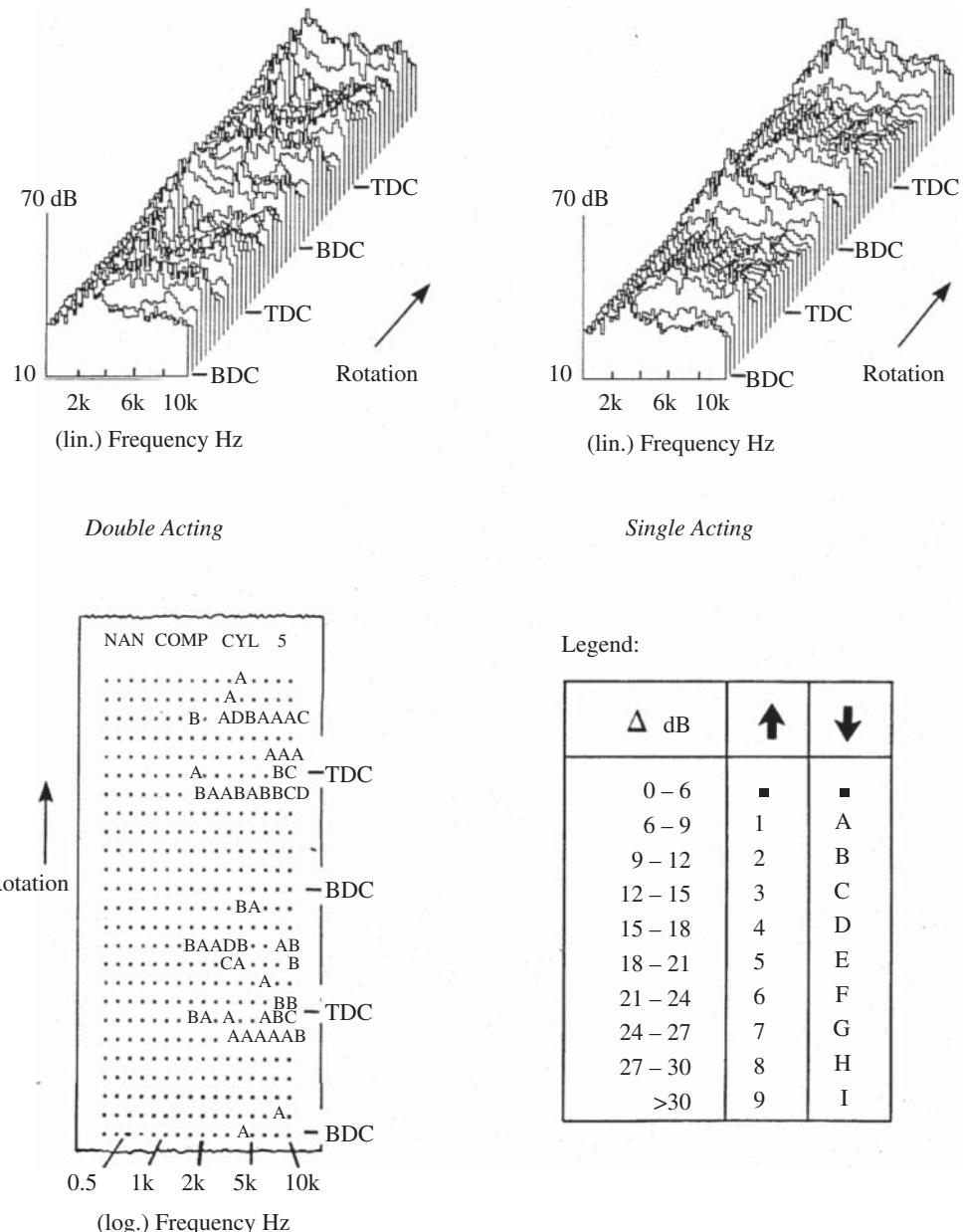


Figure 4.12 Time-frequency diagrams for the compressor of Figure 2.24 in single- and double-acting modes. The comparison diagram at bottom left shows maximum decreases of the order of 20 dB. (BDC = Bottom dead centre; TDC = top dead centre).

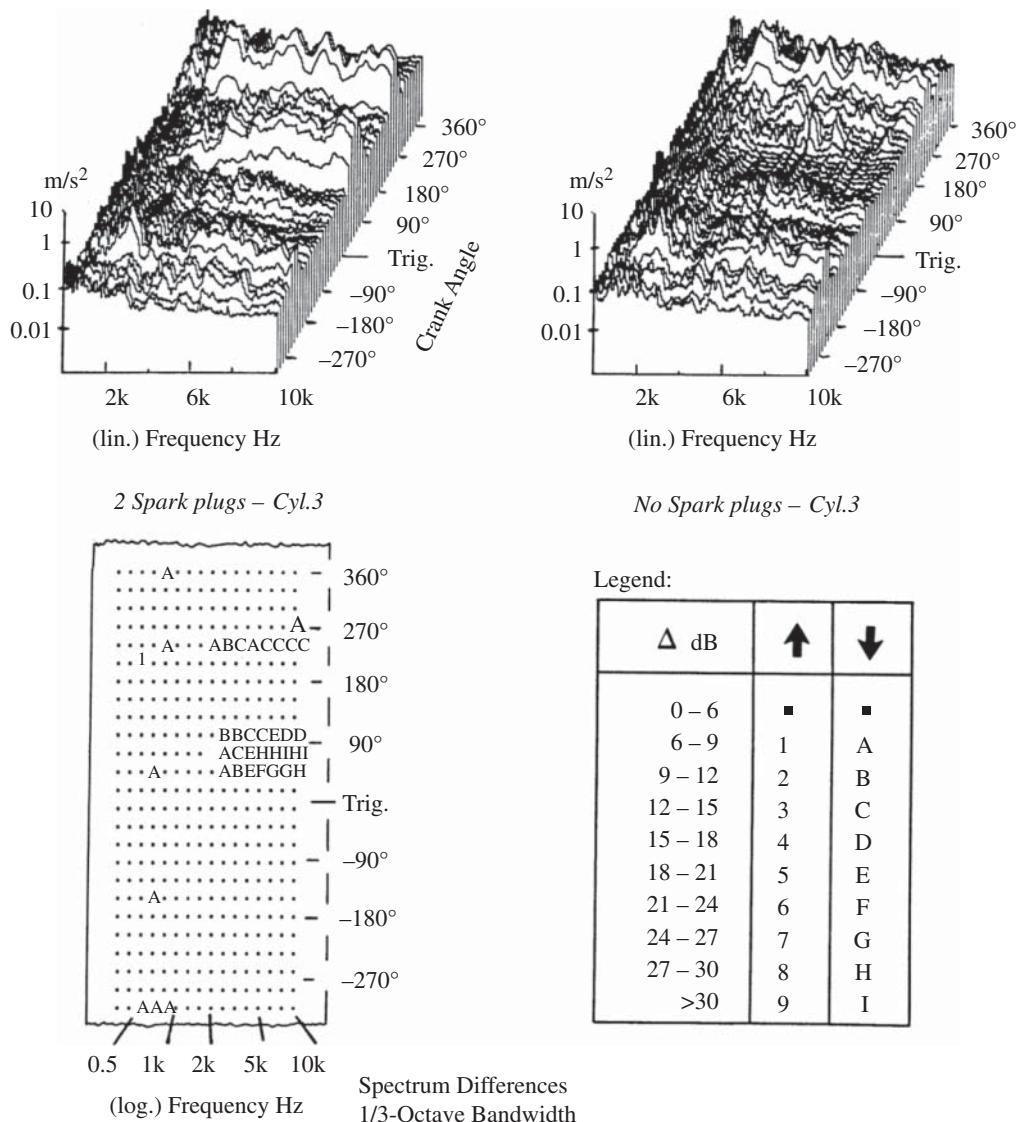
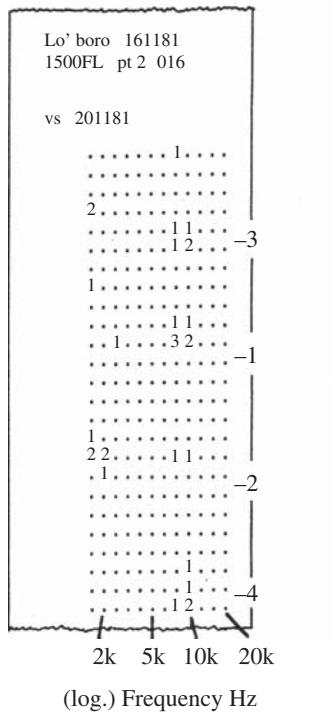
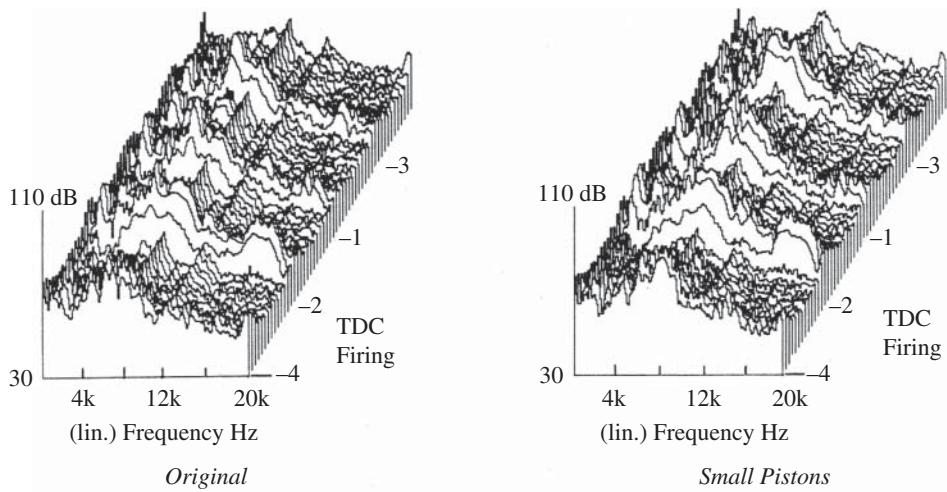


Figure 4.13 Effect of a misfire in a gas engine.

example involve the application of the WVS as defined in Section 3.6.3 and applied to the diagnostics of reciprocating machines in Section 7.4.1.

4.3.3 Torsional Vibration

For engines with effectively rigid crankshafts, the measurement of torsional vibration at one point is a very simple and effective way to monitor the instantaneous rotational speed and thus detect any non-uniformities in combustion (or cyclic pressure for machines such as pumps or compressors).



Legend:

Δ dB	\uparrow	\downarrow
0 – 6	■	■
6 – 9	1	A
9 – 12	2	B
12 – 15	3	C
15 – 18	4	D
18 – 21	5	E
21 – 24	6	F
24 – 27	7	G
27 – 30	8	H
>30	9	I

Figure 4.14 Effect of increased piston clearance in a diesel engine.

As described briefly in Section 2.3.2, the torsional vibration can be measured by frequency demodulation of a shaft encoder signal from an encoder mounted on the crankshaft. For the purposes of detection, the encoder can be replaced by a proximity probe detecting the passage of teeth on the ring gear used for starting the engine. Such a probe is shown in Figure 4.16, mounted in the bell housing of an automotive engine at the axial location of the ring gear. Another proximity probe detects the

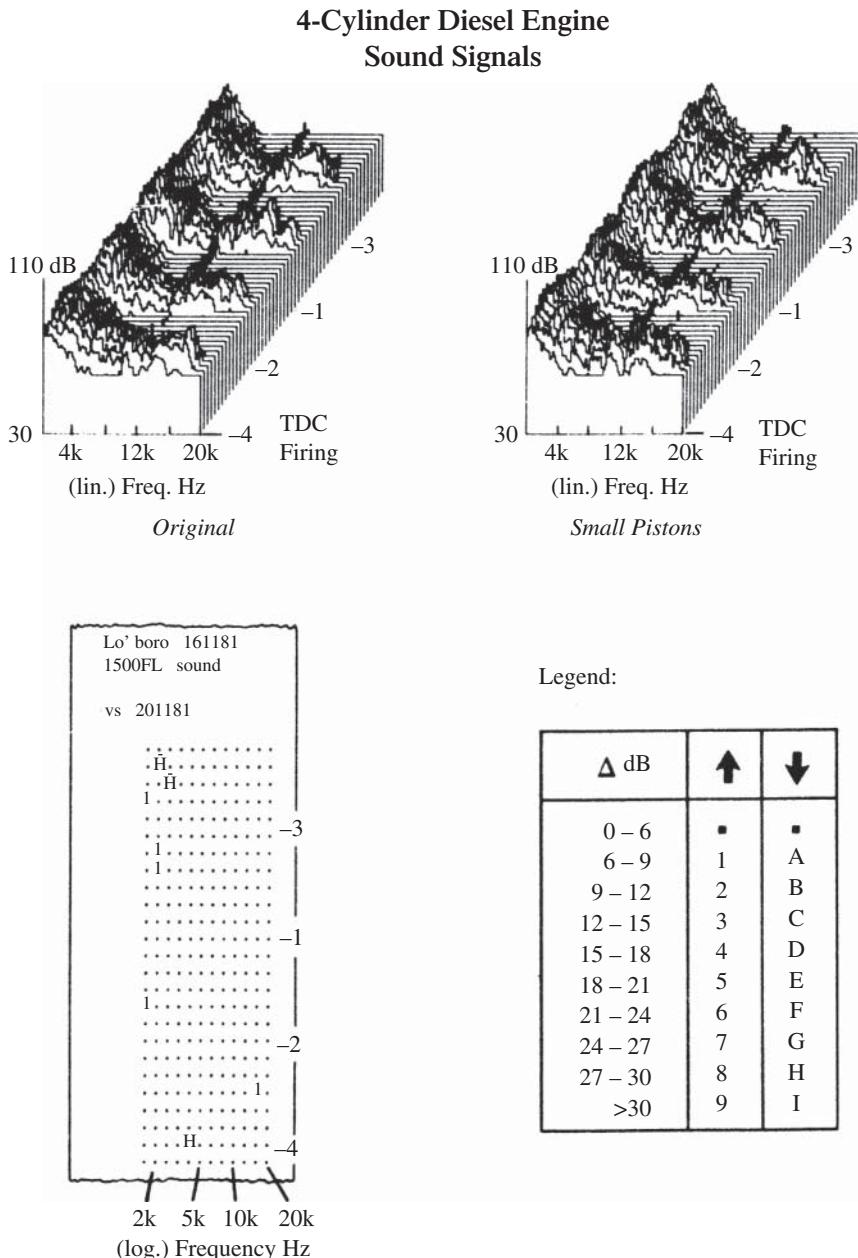


Figure 4.15 Comparison of sound signals for the same case as Figure 4.14.

passage of a screw head, giving a once-per-rev pulse. The number of teeth on the ring gear is typically of the order of 160, and this will allow demodulation of signals containing significant harmonic components up to one quarter of this (see Sections 1.3 and 7.4.2.2). Some examples in this section will illustrate that this is normally valid.

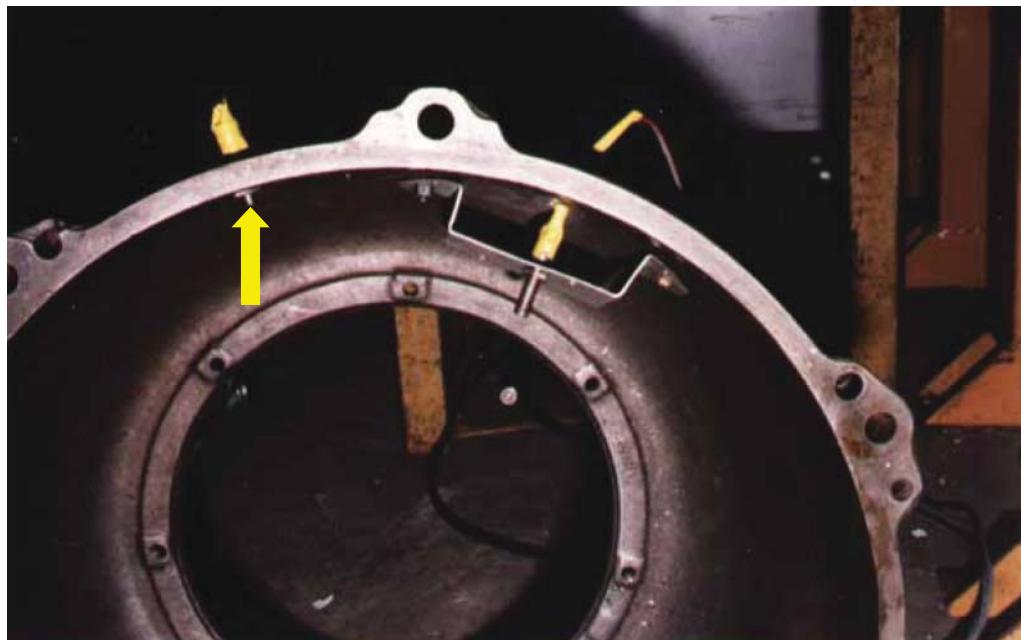


Figure 4.16 Proximity probe (indicated) used to detect the passage of ring gear teeth.

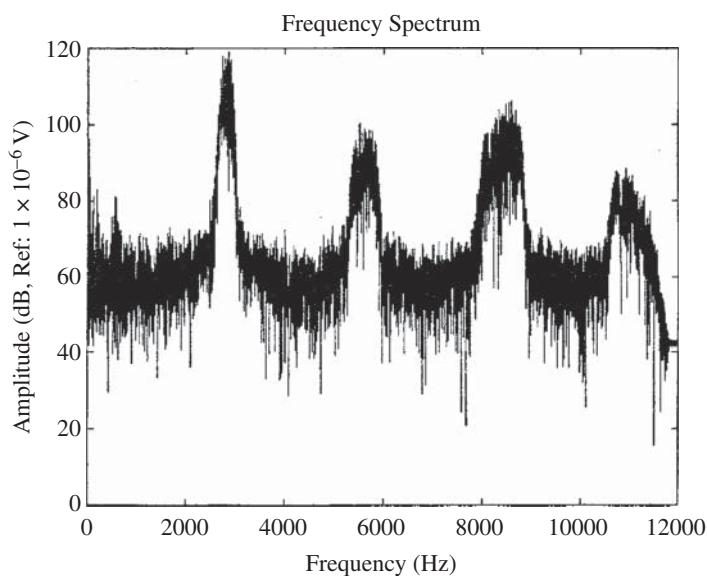


Figure 4.17 Spectrum of ring gear signal for a spark ignition engine with a complete misfire in one cylinder.

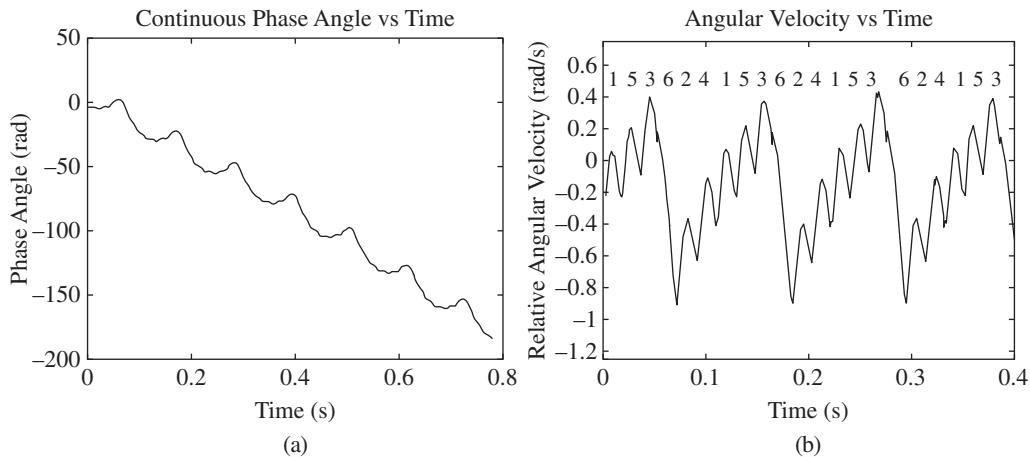


Figure 4.18 Demodulation of the first harmonic of the encoder signal in Figure 4.17 (a) Phase demodulation. (b) Frequency demodulation (angular velocity).

As discussed in Sections 7.2.2 and 7.4.2.2 of Chapter 7, there are basically two ways in which the encoder signals can be processed to obtain the crankshaft angular velocity. One is first to do a phase demodulation, and then differentiate the phase signal to obtain angular velocity.

The case illustrated in Figure 2.29 of Chapter 2 is repeated here with some further details of the procedure [13]. The ring gear had 157 teeth, and the spectrum of the ring gear signal is shown in Figure 4.17.

Since the speed variation gives a fixed fluctuation in terms of time of arrival for the encoder pulses, the corresponding phase variation is proportional to the order of the harmonic to be demodulated. For example, if the first harmonic of the encoder signal is demodulated, then the resulting demodulated phase will have to be divided by 157 to obtain a result in terms of shaft rotation angle. If the second harmonic is demodulated the dividing factor will be 314, and so on. Thus in principle the same result will be obtained from each harmonic, so the decision should be made on the basis of the one with the highest signal/noise ratio, and where there is no overlap of sidebands with an adjacent harmonic. The latter condition is satisfied by all three harmonics shown, but the dynamic range is best with the first harmonic, so this was chosen.

Figure 4.18a shows the result of phase demodulation, estimating the carrier frequency by eye as the ‘centre of gravity’ of the extended band with sidebands (it could be calculated from the mean speed over the record times the number of teeth). The slope of the demodulated phase is because there was a small error in the choice of centre frequency. This is not a big problem as long as it does not impede unwrapping of the phase (Section 3.3.2.1).

Figure 4.18b shows the corresponding angular velocity, obtained by taking the derivative of the phase. It is the same as Figure 2.29 of Chapter 2. If the derivative is obtained by simple differencing (e.g. the Matlab® function DIFF) then high frequency noise will be amplified. It is preferable to perform the differentiation by $j\omega$ operations in the frequency domain so that it can be combined with bandpass filtration.

Figure 4.19 shows the (amplitude) spectrum of a typical phase signal similar to that in Figure 4.18a. It is shown on a logarithmic amplitude scale over 80 dB. By eye it can be judged that the useful (discrete frequency) components extend to about line no. 200, so a suitable cutoff frequency in this case would be say line no. 250. Frequency components above this could be set to zero before multiplying the remaining (complex) spectrum values by $j\omega$ (to effect the differentiation)

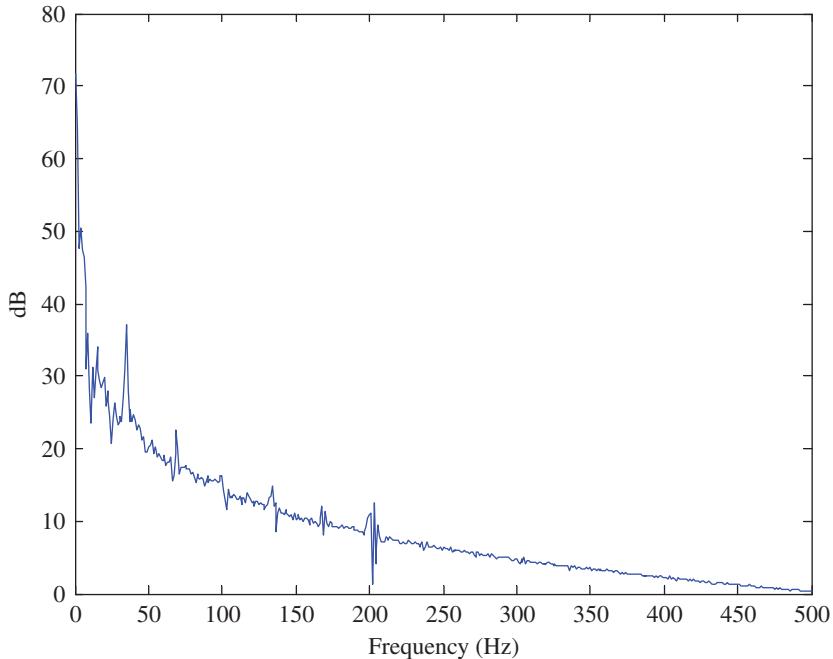


Figure 4.19 Spectrum of typical phase modulation signal.

before transforming back to the angular velocity signal in the time domain. ω is of course the frequency axis values scaled in radians/s (note that line no. 1 corresponds to zero frequency). Before performing frequency domain operations such as this, it is a good idea to remove the residual slope as seen in Figure 4.18a, to minimise the discontinuity when the signal is joined into a loop, and remember that the values near the ends of the record will be affected by any such discontinuity, and may have to be discarded. It is convenient to perform such operations on the positive frequency part of the spectrum only, using the Hilbert transform relations as shown in Figure 3.25 of Chapter 3.

The noise in Figure 4.19 is falling off approximately as $1/\omega$ so after differentiation it would be almost uniform. In the example shown, the Nyquist frequency was at line no. 2049, so the lowpass filtration removes 7/8 of the noise.

Figure 4.18 is for a complete misfire in cylinder No. 6. Figure 4.20 shows that even partial misfires can be detected using this technique. The partial misfire resulted from a loose spark plug in cylinder 6. It is seen that the regain of angular velocity for this cylinder is less than for the others.

Figure 4.21a through c, for a similar engine [14], show how misfires occurring for different reasons manifest themselves in the torsional vibrations.

Figure 4.21a is for a complete misfire from removal of the spark plug lead (as for Figure 4.18). In Figure 4.21b, the misfire is caused by a leak in the inlet manifold adjacent to one cylinder, causing a change in the mixture. Figure 4.21c is for a simulated ‘burnt valve’, this being simulated by adjusting the tappets to keep an exhaust valve slightly open.

It can be seen that even though both partial and full misfires can be detected, it is not possible to say anything about the cause of the misfire. Partial misfires give a smaller change in angular velocity corresponding to firing in that cylinder. It is most likely that the type of detailed analysis described in Section 7.4 on diagnostics, and Section 8.4 on fault simulation of reciprocating machines will give

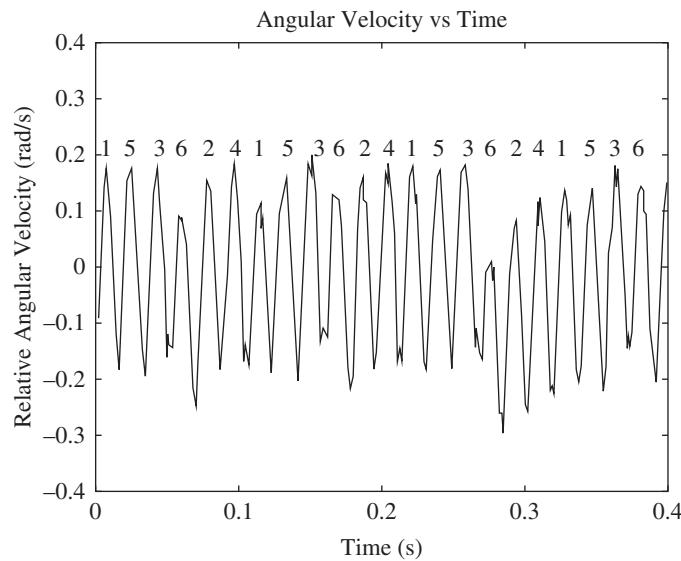


Figure 4.20 Partial misfire caused by a loose spark plug in Cyl. 6.

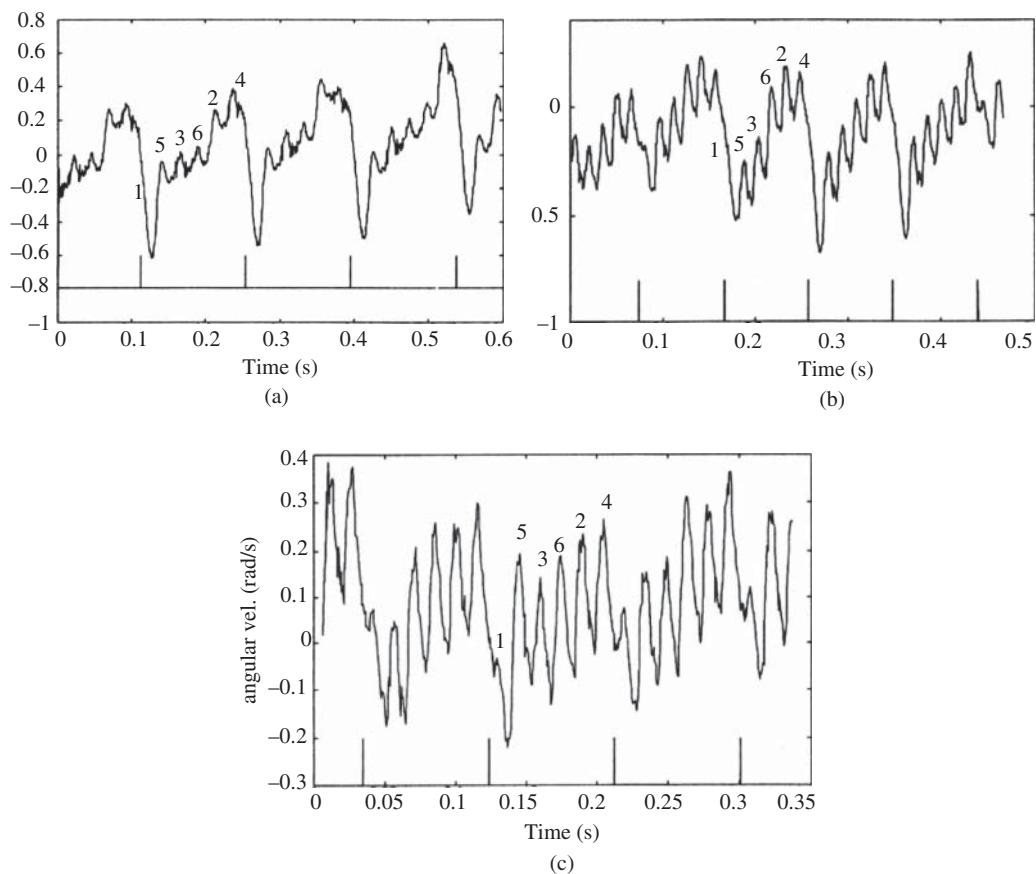


Figure 4.21 Effects of different full and partial misfires in cyl. 1 on the torsional vibration for a 6-cylinder spark ignition engine (a) spark plug disconnected (b) leak in inlet manifold (c) simulated 'burnt valve'.

more diagnostic information once the original fault detection has been made. It is also shown there that misfires can be detected from the rotational acceleration of the engine block.

One advantage of the method based on torsional vibration, despite its simplicity, is that in addition to detecting changes in the forcing function (cylinder pressure), it should also react to changes in the structural dynamic properties of the torsional vibration system. As mentioned in Section 4.3.1 such effects are less likely to manifest themselves in lateral vibrations of non-rotating components.

Section 8.4 shows how simulation of various engine faults can be used to aid diagnostics and prognostics, and Section 9.3.4 how simulated engine fault data can be used to train neural networks to recognise equivalent real faults.

References

1. Rathbone, T.C. (1939). Vibration tolerance. *Power Plant Engineering* 43: 721.
2. Yates, H.G. (1949). Vibration diagnosis in marine geared turbines. *Transactions – North East Coast Institution of Engineers and Shipbuilders*. D35–D50 (January 28): 225–261.
3. Mitchell, J.S. (1981). *Machinery Analysis and Monitoring*. Pennwell Publishing.
4. Chapman, R.N. (1967). Vibration analysis applied to machinery maintenance. *Naval Engineers Journal* 79 (3): 431–437.
5. Glew, C.A.W. and Watson, D.C. (1971). ‘The octave band vibration analyser as a machinery defect analyser’. *ASME Paper 71-DE-47*, American Society of Mechanical Engineers.
6. ISO 20816, Part 1 (2016). *Mechanical Vibration - Measurement and evaluation of Machine Vibration - Part 1: General Guidelines*. International Standards Organisation.
7. Woods, R. (1968). An investigation into vibration criteria for rotating machinery, PhD Thesis, University of Aston in Birmingham, UK.
8. Downham, E. and Woods, R. (1971). ‘The rationale of monitoring vibration on rotating machinery in continuously operating process plant’. *ASME Paper 71-Vibr-96*, American Society of Mechanical Engineers.
9. Randall, R.B., (1979). ‘Efficient machine monitoring using a calculator-based system’. *First Conference on Condition Monitoring in the Process Industries*. Grand Hotel, Manchester (27–28 November 1978).
10. Randall, R.B. (1985). Computer aided vibration spectrum trend analysis for condition monitoring. *Maintenance Management International* 5: 161–167.
11. Bradshaw, P. and Randall, R.B., (1983). ‘Early fault detection and diagnosis on the trans alaska pipeline’. *MSA Session, ASME Conference*, Dearborn, pp 7–17, American Society of Mechanical Engineers.
12. ISO 10816, Part 6 (1995). *Mechanical Vibration - Evaluation of Machine Vibration by Measurements on Nonrotating Parts - Part 6: Reciprocating Machines with Power Ratings Above 100 kW*. International Standards Organisation.
13. Jenner, L., (1994). ‘Ford IC engine diagnostics’, BE Thesis, School of Mechanical and Manufacturing Engineering, UNSW, Australia.
14. Lam, S.T., (1995). ‘Internal combustion engine diagnostics using torsional vibration analysis’, BE Thesis, School of Mechanical and Manufacturing Engineering, UNSW, Australia.

5

Some Special Signal Processing Techniques

5.1 Order Tracking

In analysing rotating machine vibrations, it is often desirable to have a frequency x-axis based on harmonics or ‘orders’ of shaft speed. This can be to avoid smearing of discrete frequency components due to speed fluctuations, or can be to see how the strength of the various harmonics changes over a greater speed range, for example as they pass through resonances. If a constant amplitude signal, which is synchronous with the rotation of a shaft for example, is sampled a fixed number of times per revolution, the digital samples are indistinguishable from those of a sinusoid, and thus give a line spectrum, whereas if normal temporal sampling is used the spectrum spreads over a range corresponding to the variation in shaft speed. Thus, for order analysis it is necessary to generate a sampling signal from a signal synchronous with shaft speed. This can be a tacho or shaft encoder signal, giving respectively one or several pulses per rev, or can sometimes be extracted by filtering the response signal itself.

5.1.1 Comparison of Methods

It is sometimes possible to use a shaft encoder mounted on the shaft in question to directly provide a sampling signal, but more often the latter has to be generated electronically or numerically. Formerly, this was done using a phase-locked loop to track the tacho signal and then generate a specified number of sampling pulses per period of the tracked frequency. However, an analogue phase-locked loop has a finite response time and cannot necessarily keep up with random speed fluctuations such as occur with an internal combustion engine from cycle to cycle. The best method is to digitally resample each record based on the corresponding period of the tacho signal, so as to achieve sampling for uniform increments in shaft rotation angle. This is known as ‘angular resampling’, and the procedure used to accomplish it is called computed order tracking (COT). The advantage of COT is that the signal is available in both the time and angle domains, and once the time/phase map relating them is established, it is possible to transform between them. It will be found that when analysing signals from machines with varying speed, it can often be an advantage to transform back and forth several times between the two domains.

5.1.2 Computed Order Tracking (COT)

In computed order tracking there are actually two separate interpolation procedures required. The first is to establish the map of phase vs time of a reference signal, usually sampled at uniform time intervals, which is periodic in the angle domain with a period that corresponds to a known factor (or order) of the desired basic reference shaft speed. In a gearbox, for example, this might be any of the shaft speeds, for which the speeds of all other shafts are known from the gear ratios. Once the phase/time map is established with an acceptable error, the times corresponding to uniform increments of phase (i.e. rotation angle) can be determined by interpolation, and the actual response signals resampled at those times, using a second interpolation process. Resampling of a phase/time map is shown in Figure 5.1.

In the general case, true information about phase will only be known at certain times. For example, if the reference signal is a series of once-per-rev tacho pulses, it will be known only every 2π radians, or if it is an N-per-rev shaft encoder it will be known every $2\pi/N$ radians. If it is a frequency modulated sinusoid, such as a shaft harmonic isolated using a bandpass filter, the zero crossings will occur at phase increments of π radians. Note that this will be independent of any simultaneous amplitude modulation (AM), as long as the AM signal is everywhere positive.

McFadden [1] made an extensive study of the errors involved in polynomial interpolation of increasing orders. He showed that n^{th} order interpolation corresponds to convolving the series of samples (delta functions) with an appropriate n^{th} order polynomial having a value of 1 at zero, and a value of zero at multiples of the sample spacing (so that the resampled values at the original samples remain unchanged). This is illustrated in Figure 5.2 for linear interpolation (first order polynomial). It corresponds to convolution with a triangle (linear function between samples).

Zero order interpolation consists of taking the nearest sample to the calculated interpolated value, and to be useful would require the sample rate to be first increased by a large factor (say 10), before selecting the nearest sample. This method is known (e.g. in Figure 5.4) as ‘sample-and-hold’, and corresponds to convolution of the samples with a rectangle of width equal to the sample spacing (but preferably centred on the sample).

Increasing the sample rate by an integer factor can be achieved in two ways. In the time domain, it can be done by inserting the appropriate number of zeros in between each actual sample, and then applying a digital lowpass filter to limit the frequency range to the original maximum, thus smoothing the curve (it will also require rescaling proportional to the resampling factor). Resampling by a factor of 4 is illustrated in Figure 5.3.

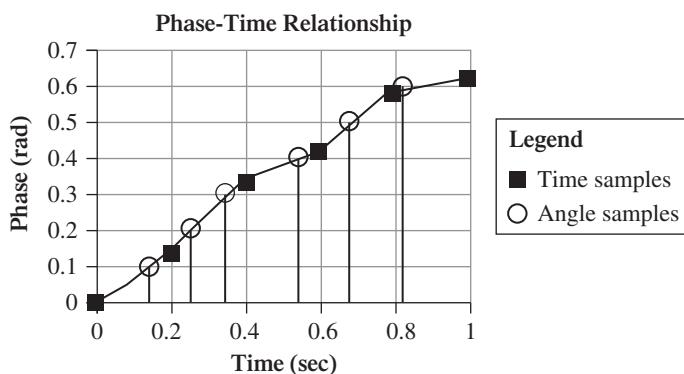


Figure 5.1 Resample times for equal angle increments.

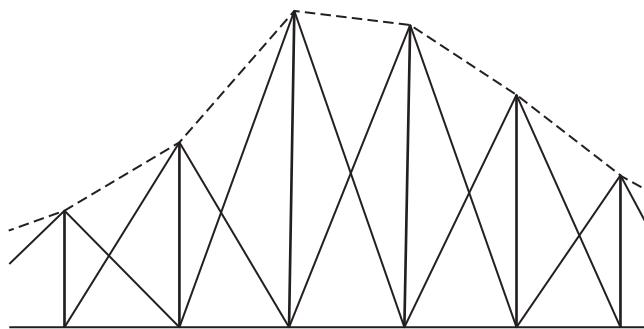


Figure 5.2 Linear interpolation by convolution of samples with a triangle.

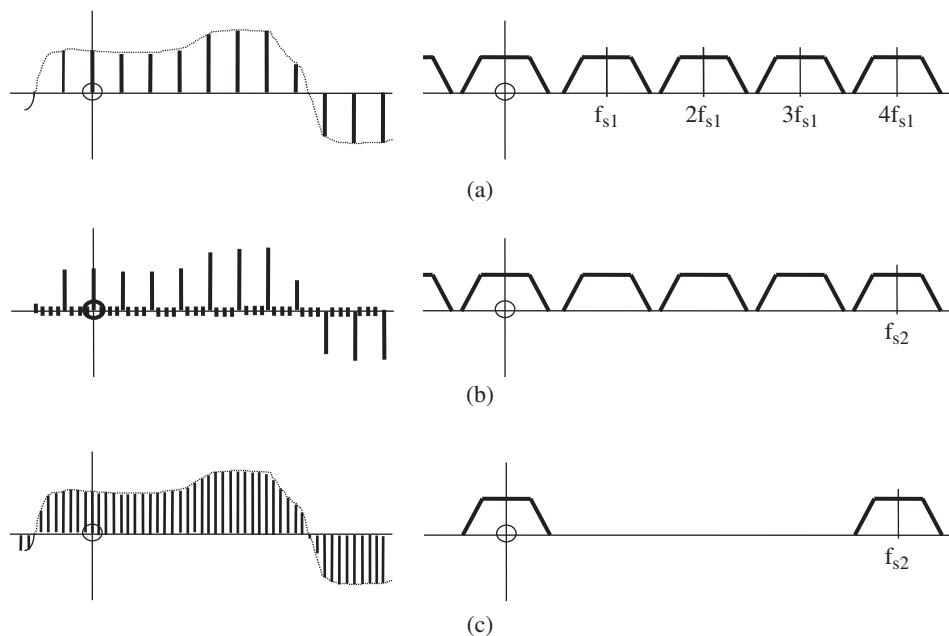


Figure 5.3 Digital resampling with four times higher sampling frequency. (a) Signal sampled at f_{s1} and its spectrum (b) Addition of zeros which changes sampling frequency to f_{s2} (c) Lowpass filtration and rescaling.

The same result can be achieved in the frequency domain by padding the FFT spectrum with zeros in the centre (i.e. around the Nyquist frequency, equivalent to Figure 5.3c, zero to f_{s2}) and then inverse transforming the increased (two-sided) spectrum to the same increased number of time samples. Note that the record length in seconds is the reciprocal of the frequency line spacing in Hz which is not affected by the zero padding. This latter procedure can also be used to resample a record consisting of an integer number of samples to another (though greater) integer number, and is the basis of the Matlab® function INTERPFT.

In general, more accurate interpolation will be achieved using higher order polynomials and calculating their values at the interpolated positions. The accuracy of the interpolation can be judged by considering that the interpolation in the time domain corresponds to a multiplication in the frequency

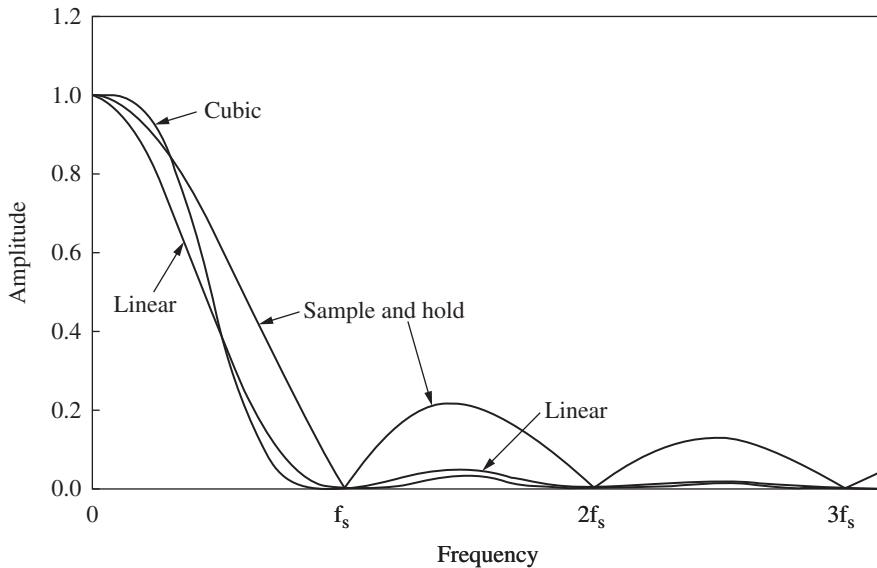


Figure 5.4 Comparison of frequency characteristics for interpolation at different orders. Source: from [1].

domain by a filter characteristic which is the Fourier transform of the convolving function. Figure 5.4 (from [1]) compares these filter characteristics for orders zero, 1 and 3, (Sample and hold, Linear and Cubic spline).

There are two sources of error in the spectra multiplied by these frequency characteristics:

- 1) The periodic spectra centred on higher multiples of the sampling frequency (for example in Figure 5.3a or c) are multiplied by the sidelobes of the characteristic and alias down into the measurement range. The signal is already sampled, and so they can't be removed by lowpass filtering. Thus, it is an advantage for the sidelobes to be as small as possible.
It is also an advantage for the sampling frequency to be as high as possible, since the sampling frequency f_s in Figure 5.4 would correspond to f_{s1} in Figure 5.3a and f_{s2} in Figure 5.3c.
- 2) All the functions have a lowpass filter characteristic in the frequency range up to the maximum useful value (defined by an anti-aliasing filter), normally $0.4 f_s$. Once again, it is an advantage for the sampling frequency to be as high as possible, since increasing it by a factor of 4 would mean the useful frequency content would only correspond to $0.1 f_s$.

In Figure 5.4, the spectrum of the Sample and hold case (rectangle) is a $\sin(x)/x$ ($\text{sinc}(x)$) function, with rather large sidelobes. The triangular function depicted in Figure 5.2 is the convolution of the rectangular function with itself, so its spectrum is $[\text{sinc}(x)]^2$. Even though this has much lower sidelobes than the $\text{sinc}(x)$ function, its lowpass filter effect is stronger. The spectrum of the cubic function has even smaller sidelobes, but initially has less lowpass filter effect than zero and first order interpolation. In fact, if the sample rate is initially doubled before resampling, the lowpass filter effect will be negligible up to the maximum useful frequency of $0.2 f_s$.

Note that one of the first resampling methods published [2], achieves quadratic resampling of the signal. The method was formalised in [3]. It assumes the angular acceleration of the signal is constant between samples, meaning that the angular velocity is linear and the angular displacement

(phase) is quadratic. It is not depicted in Figure 5.4, but gives intermediate results between linear and cubic resampling. It can thus be concluded that cubic resampling is the most efficient and accurate method of polynomial resampling, and can be used to resample actual response signals, as well as to determine the phase/time curve between the original reference samples.

5.1.3 Phase Demodulation Based COT

The most accurate way of determining the phase/time curve is by phase demodulating the reference signal as described in [4]. This is because phase demodulation gives the true phase/time curve, to any desired degree of resolution, provided the spectrum of the bandlimited mono-component carrier frequency is completely isolated in the frequency domain, both from noise and any other shaft speed related components, including other harmonics of the reference signal.

This is very similar to the case of a signal sampled so as to obey the Nyquist sampling theorem, i.e. more than two samples per period. All information about the time signal is contained in this bandlimited spectrum, and it can be regenerated to any degree of resolution by increasing the sampling frequency by zero padding around the Nyquist frequency in the frequency domain (see for example Figure 5.3a,c). In the case of phase demodulation, it corresponds to making the sampling frequency f_s , in Figure 3.30b, arbitrarily large, to increase the resolution. The only difference is that the modulation sidebands for phase demodulation theoretically extend to infinity; however, only a limited number can be considered significant [4], and can be isolated using a rectangular window in the frequency domain. Incidentally, this corresponds to interpolation by a $\text{sinc}(x)$ function in the time domain, but since this is infinitely long, it would have to be truncated if used for convolution there, whereas inverse transformation from the frequency domain gives the true interpolated curve.

In [4] the authors have shown that for practical purposes, ‘bandlimited’ can be taken as separated down to -40 dB (with respect to the largest component in the band) from adjacent unrelated frequency components. In [4] it is also shown that this 40 dB separation can just be achieved between the first and second harmonics of a pulse tacho signal, for a maximum speed variation of $\pm 33\%$ of the fundamental component, in other words from $2/3$ to $4/3$ of it, or a maximum possible speed range of $2 : 1$. This applies only for very low modulation frequencies and the range is even lower if the rate of frequency change is higher [4]. The amount of maximum frequency deviation against modulation frequency, both expressed as a percentage of the carrier frequency, is given in Figure 5.5 for three different criteria for what constitutes significant sidebands. The most restrictive is the 40 dB dynamic range (DR) criterion mentioned above, but curves for 20 dB DR and Carson’s rule are also given [4].

Figure 5.6a shows the phase modulation sidebands for a simulated tacho signal with two harmonics of the speed, with frequency sweep less than the maximum of $\pm 33\%$. It shows that the spread of the sidebands is proportional to the order, and a simple explanation of this is given in Figure 5.6b, which shows the PDF of a sinusoid. The sideband spacing in Figure 5.6a is equal to the modulating frequency, and it can be understood that as this tends to zero, the modulated spectrum will tend to become continuous, and equal to the PDF. The second harmonic sweeps over twice the frequency range of the first harmonic, and thus the spread of sidebands in this case does not go outside the frequency sweep range, and is proportional to the harmonic order. Ref. [4] shows that for a tacho signal with any number of harmonics (e.g. a series of pulses), the sidebands around the higher harmonics do not overlap into the intersection of the first two, as long as those are separated.

The allowable sweep range is approximately inversely proportional to the harmonic order if a higher harmonic is being demodulated, for example being approximately $\pm 10\%$ for demodulation of a third harmonic. Ref. [5] was the first paper (in English) to propose the phase demodulation method for obtaining a phase/time map, and in fact to obtain it from a vibration response signal, but

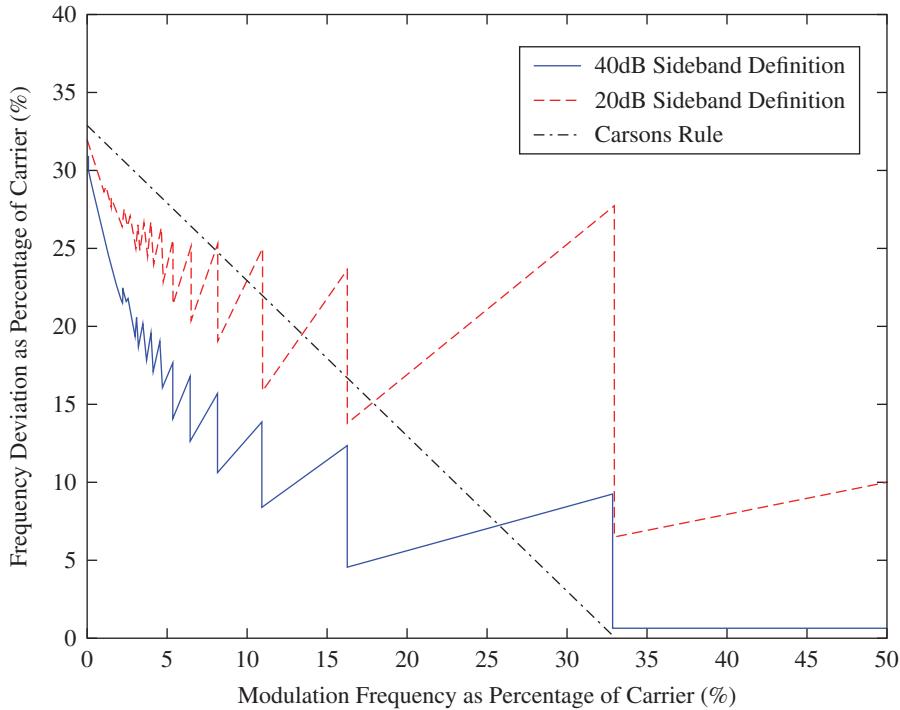


Figure 5.5 Envelope of permissible combinations of frequency deviation and modulation frequency.
Source: from [4].

the reference signals used were high order harmonics of shaft speed (gears mesh frequencies) and so it was limited to rather small speed variations.

Figure 5.7 shows two signals from a gearbox with a $\pm 10\%$ speed variation. One is a once-per-rev tacho signal and the other an acceleration response signal. Both time signals are shown along with the low frequency part of their spectra. For the tacho signal the first three harmonics are separated and could be demodulated, but for the response signal only the third is strong enough, though still separated. The difference in the spectra of the third harmonic can be explained by the significant amplitude modulation (AM) of the acceleration signal, but as mentioned above, this has no influence on the phase signal. Figure 5.8a,b show the result of using the two phase/time maps to order track the tacho signal, and even though that from the acceleration signal is a little noisier, it can still be used. In fact, when the order tracked acceleration signal was analysed for a bearing inner race fault, it could be diagnosed as shown in Figure 5.8c.

5.1.3.1 Improvement by Iteration

An advantage of the phase demodulation method is that the phase/time map can be improved by iteration. For example, if the speed variation is large it may only be possible to separate the first harmonic from higher harmonics, but after the first iteration much higher harmonics can be separated, and a further phase demodulation will give a phase correction which can be added into the first, and so on as higher harmonics become separated. For example, in Figure 5.8b the 9th harmonic seems well

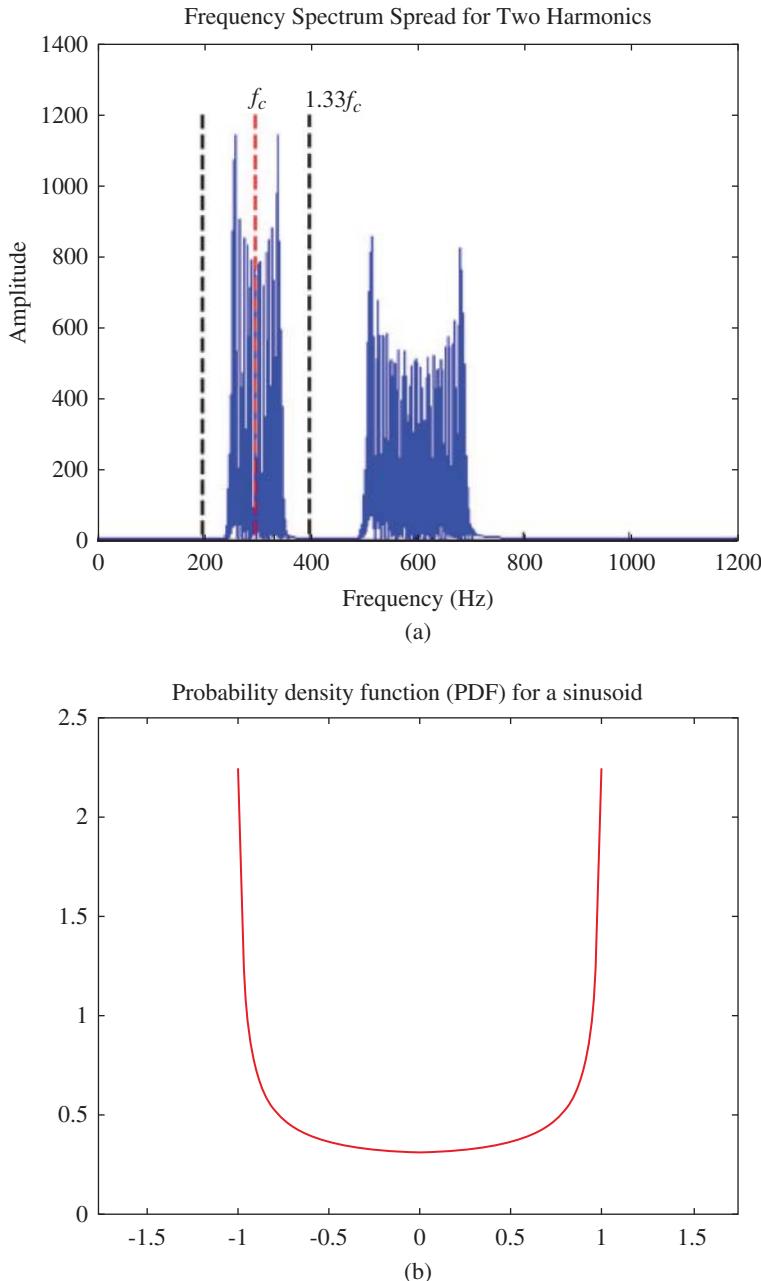


Figure 5.6 (a) Phase modulation sidebands for a tacho signal with two harmonics of the fundamental frequency (b) PDF for a sinusoid. Source: from [4].

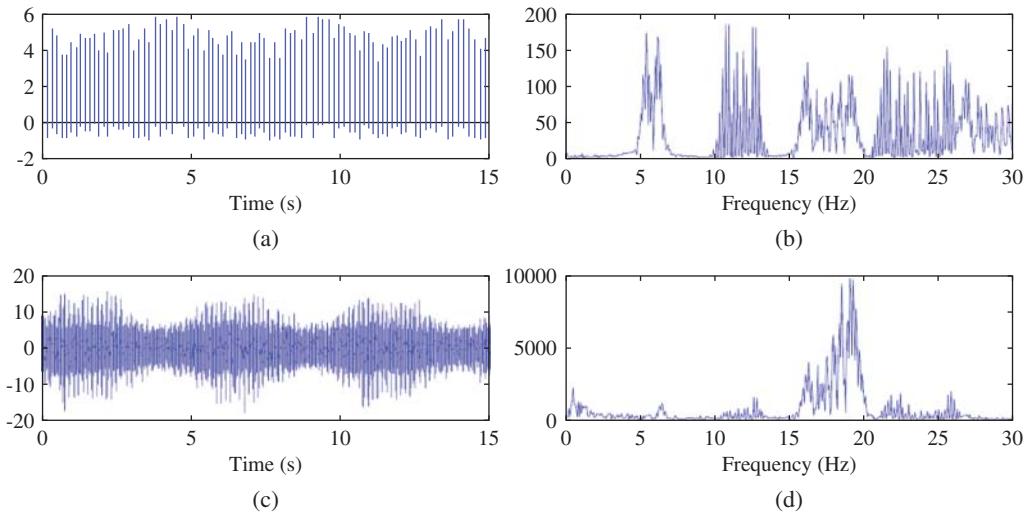


Figure 5.7 Signals with $\pm 10\%$ speed variation: (a, b) tacho, (c, d) acceleration, (a, c) time signals, (b, d) spectra. Source: from [4].

separated and could have been used for a further iteration. In that case it was not deemed necessary, as the aim was to perform bearing diagnostics, where the frequencies can be smeared by 1–2% in any case.

In [4], four stages of iteration were carried out in the demodulation of the tacho signal for the gearbox, with $\pm 10\%$ speed variation, using the 1st, 5th, 21st, and 151st harmonics in sequence, and the envelopes of the (harmonic) spectra are shown in Figure 5.9. The first three iterations give considerable improvement, revealing better estimates of the amplitudes of progressively higher harmonics (up to >800), but the fourth iteration gives little further improvement.

As will be seen from the example given in Section 7.1.3, of combination of order tracking with a harmonic cursor, even two iterations can often give extremely accurate identification of high order harmonics, in that case allowing blind identification of the number of teeth in a mating gear pair.

5.1.3.2 Using a Response Signal as Reference

A response signal can only be used as a speed reference (as in the example of Figure 5.8) when the frequency of the response corresponds to the speed of the machine at the time of measurement. The transfer function from the point of excitation to the response measurement point may involve a time lag, meaning that the frequency of the response corresponds to the speed at the time of excitation, a little earlier, when the speed might have been slightly different, and this can vary over the time of a measurement record, giving a smearing of the orders. The time lag corresponds to phase shifts in the transfer function, and is only significant in the vicinity of resonances. If a particular harmonic remains on a pure spring line (or indeed mass line) during the whole measurement record involving a speed change, the response will be immediate. This problem is studied in some depth in Ref. [6], to which the reader is referred.

Another aspect of demodulating response signals is that they are likely to be both frequency modulated, and amplitude modulated, and so the spread of sidebands is more than that due to the

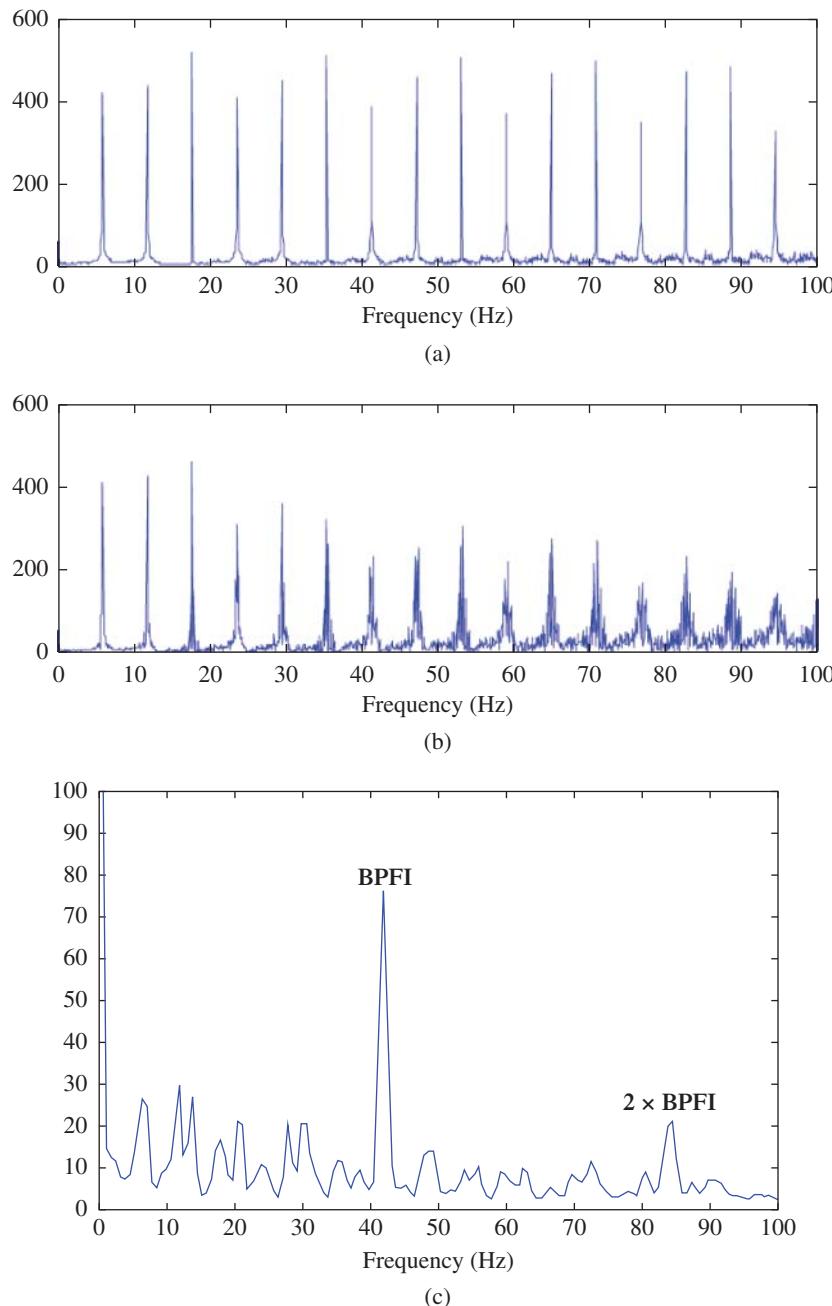


Figure 5.8 (a), (b) Spectra of order tracked tacho signal using third harmonic of (a) tacho signal (b) acceleration signal; (c) Envelope spectrum of order tracked acceleration signal showing two harmonics of BPFI (ballpass frequency, inner race). Source: from [4].

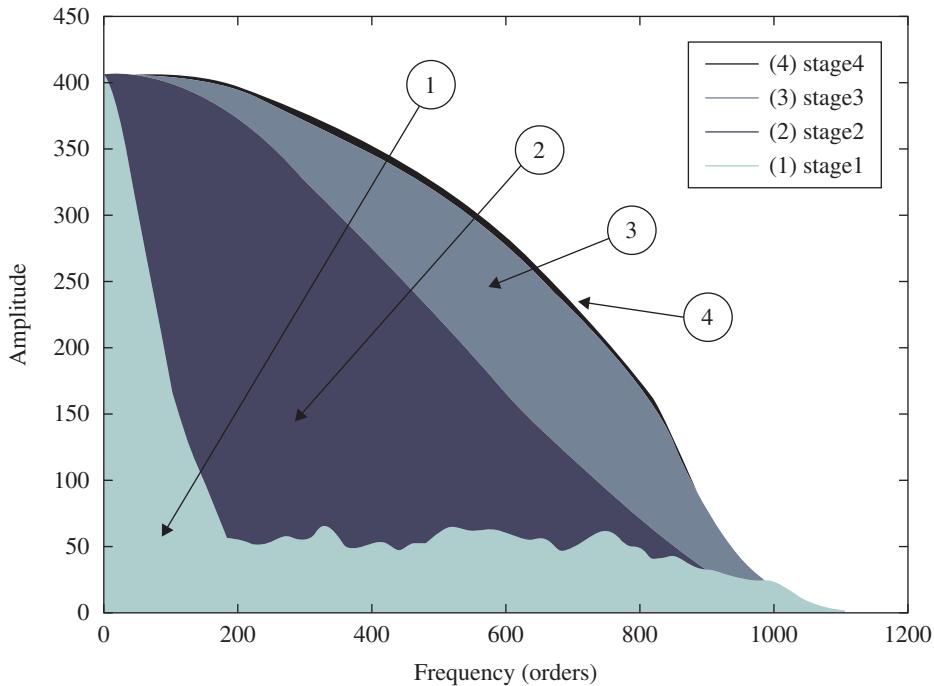


Figure 5.9 Comparing spectra of order-tracked tacho from all four stages for $\pm 10\%$ case.

FM (PM) alone. A generally modulated mono-component carrier can be expressed as $A(t)e^{j\phi(t)}$, whose spectrum is the convolution of the spectra of the AM and PM components.

In [4] it is discussed how this increases the width of the band to be demodulated (to maintain the 40 dB separation) and though this is not by the full amount of the bandwidth of $A(t)$ alone, it can still restrict the amount of allowable frequency deviation for a given modulation frequency. Because of the difficulty of predicting the valid demodulation bandwidth, it is probably easiest to determine it by inspection.

5.1.4 COT Over a Wide Speed Range

When the speed range exceeds the maximum allowable for the phase demodulation method, of 2 : 1 or less, there are a number of options available. One is to make a first correction using some other method, to bring the range within that suitable for phase demodulation, and then make further iterations to improve the phase/time map.

This is similar to what is suggested in Ref. [7] although there it was considered to be sufficient to use only one step of phase demodulation. The method is actually proposed there for the purpose of determining the instantaneous speed, but this can be integrated to provide a phase/time curve for order tracking. The first step extracts an estimate of the instantaneous frequency of a suitable harmonic as selected from a spectrogram, even if the one chosen as a reference overlaps with another over the full time span. At every point in time, the peak value corresponding to the selected harmonic is detected in a narrow band around the line.

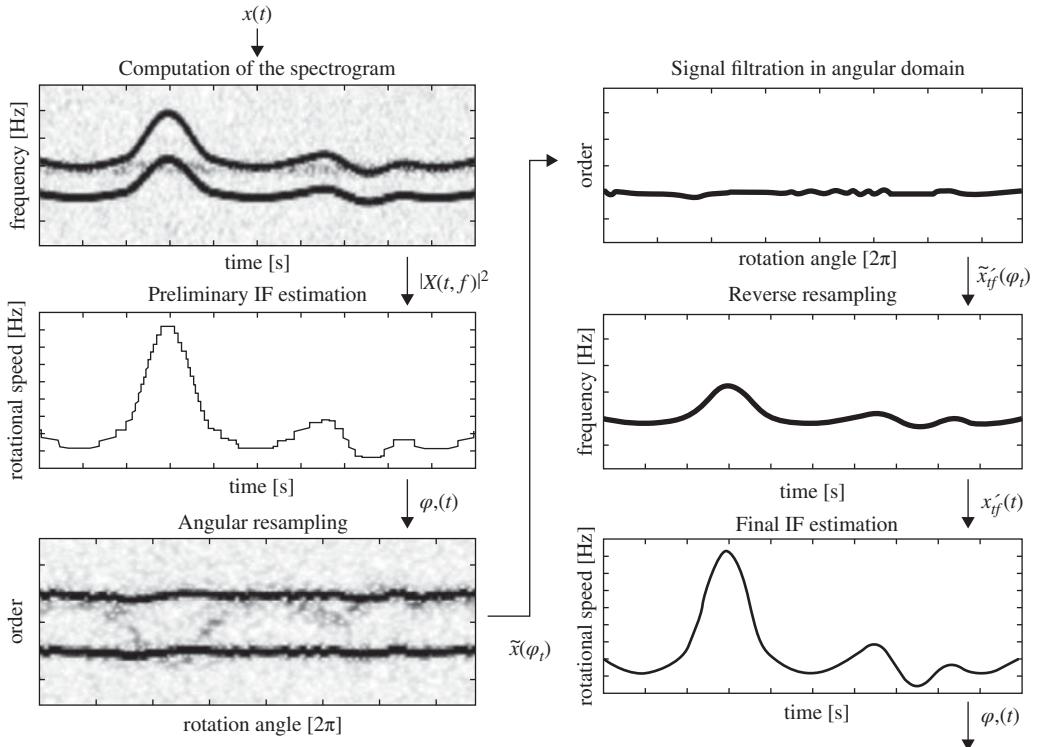


Figure 5.10 Scheme of the algorithm of the method of Ref. [7].

The rough speed vs time curve is used for angular resampling of the signal, after which the two overlapping components can be separated and the reference one filtered and smoothed.

When it is converted back to the time domain by reverse angular sampling it gives a much improved estimate of the speed curve. This is illustrated in Figure 5.10 (from [7]).

5.1.4.1 Extension of Phase Demodulation Method

It is possible to use the phase demodulation method, as long as the signal is divided into overlapping sections, in each of which the speed range is allowable. This is described in Ref. [8]. If the overlapping sections are extracted by appropriate windowing, a smooth transition can be made between the segments to obtain the order-tracked version of the whole signal. Since the segments are defined by a maximum speed ratio, it is not possible to extend down to zero frequency, but four or five segments will normally extend down to a speed below which there is nothing of interest in the signal.

To maintain a uniform weighting of the recombined signal after treatment, the weighting of each segment in the overlap area should be similar to half Hanning functions, as illustrated in Figure 5.11. They overlap at the point where each equals 0.5, and add to 1 everywhere in the overlap zone. Even though the transformation from equal time to equal phase intervals distorts the x-axis of the windows, this distortion is identical for both segments so that they still add to 1 everywhere, and each equals 0.5 at the intersection point, originally in the centre of the overlap region in the time domain. It is

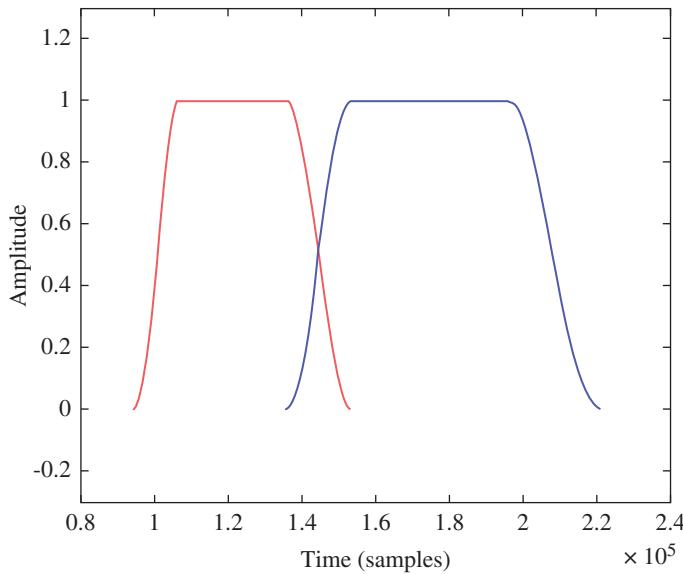


Figure 5.11 Windowing functions for two adjacent segments.

suggested that these (break) points should be made the reference points for lining up the phase axis for the whole record after resampling.

In some cases, it has been found critical that the reference signal used for order-tracking is not windowed. Windowing should have no effect on the phase of a reference signal, which is its only property of interest in the order-tracking process, and the phase can become indeterminate when the amplitude gets too small. If it is desired to have an order tracked version of the reference signal, a copy can be made and treated in the same way as other measured signals. It is not necessary to window the extreme ends of the highest and lowest speed segments, as the ends should be discarded in any case. For this reason, a section of the signal at constant speed should be included at the upper end so that a portion can be discarded finally without loss of information, and the same for a portion at the lowest speed end where the vibrations are not of much interest. The whole order-tracked signal after re-assembly can be weighted with a new window function for further processing, such as order spectrum analysis.

It is suggested that the overlap should be expressed as a percentage of the length of the lower speed segment, for example the one on the left in Figure 5.11. This means that the overlap will in general be different at the two ends of each segment, and be smaller than the nominal percentage at the upper end of each segment. The nominal overlap used for the results given below was $\pm 5\%$.

Since the sampling frequency is proportional to the reference frequency, which varies over the record, it should be ensured that the signal is appropriately lowpass filtered so as to always satisfy the Nyquist condition (> 2 samples per period of the highest retained frequency, which in each segment must everywhere be greater than the highest order to be retained). The resampling frequency can in fact be specified as a certain number of samples per period of the reference frequency (usually corresponding to the shaft speed). In the highest speed record, avoidance of aliasing can be ensured by doubling the sampling frequency before order tracking, since the maximum speed range is 2 : 1. For a run-down, where the speed varies over a wider range, it will be necessary to lowpass filter the

signal before resampling to even lower frequencies. The filter cut-off frequency should be lower than half the speed-related resample frequency at the lowest speed in the segment, as given by:

$$f_{LP} < FRf_{st} \quad (5.1)$$

where f_{st} is the current temporal sample rate in that segment, F is the proportion of the sample rate up to which the LP filter characteristic is uniform, and R is the ratio of lowest to highest speed in the segment. F would be 0.4 for typical antialiasing filters, but can be just less than 0.5 for an ideal filter (see below). R is obviously a minimum of 0.5. f_{LP} should simultaneously be greater than the highest order to be retained at the highest speed in the segment, as given by:

$$f_{LP} > Nf_{r\max} \quad (5.2)$$

where N is the minimum number of orders to be retained everywhere, and $f_{r\max}$ is the highest value of the reference frequency in the segment.

Before performing the resampling of each signal segment to a variable rate, it is most efficient to resample to a fixed rate that is higher than required to meet the Nyquist condition at the highest speed in the segment, at least if this is less than half the current rate. This is because the easiest way to reduce sample rate is in steps of 2 : 1 by decimation (including lowpass filtration, e.g. using Matlab DECIMATE). Thus, if the value given by Eq. (5.2) is less than half the value given by (5.1), then f_{st} can be halved by decimation (and updated).

Note that the most efficient way to perform the lowpass filtration is simply to set frequency lines above that defined by (5.2) to zero in the (complex) spectrum, and transform back to the time domain, for example using the 1-sided spectrum and Hilbert transform principles as shown in Figure 3.25b. Not only does this give zero phase shift, but also provides an ideal filter, meaning that the cut-off frequency need only be one line below half the resample frequency as mentioned above.

The general procedure is demonstrated in Ref. [8] using a signal generated to resemble the response on the casing of a machine such as a gearbox during a run-up from low speed at a linearly increasing rate. It is in fact a gearbox casing (without internals) but excited by a force signal, applied through a shaker, with ten harmonics of equal amplitude (at least in the control signal sent to the shaker amplifier) with fundamental frequency increasing from zero to 200 Hz over 30s (sweep rate 6.67 Hz/s). Thus, the highest excitation was 2000 Hz, but because of distortion, response harmonics are present over the full valid frequency range of 6400 Hz (order 32 at the highest speed). The sampling frequency was $2.56 \times$ the highest valid frequency or 16 384 Hz. Figure 5.12 shows spectrograms of the force and response acceleration signals. The radial lines emanating from the bottom left origin represent the various harmonic orders, and horizontal bands in (b) represent constant resonance frequencies.

Regardless of whether the signal represents a run-up (as in this case) or a run-down, it is recommended to start the analysis at the highest speed and work down.

The details for this case are given in [8], but to determine the maximum frequency deviation for each segment, and thus the breakpoints between windows, it is desirable to use a diagram similar to Figure 5.5, but adjusted for the fact that the modulation is by a constant sweep rate rather than a constant frequency. Figure 5.13a is a zoom on the low frequency part of Figure 5.5, because sweep rates usually correspond to very low modulation frequencies. Figure 5.13b is a diagram indicating how an equivalent modulation frequency can be determined. If this is judged as the maximum rate of change of a sinusoid, the sweep rate corresponds to $2\pi f_d f_m$ as shown, but the average slope of $4f_d f_m$ was found to be adequate in this case. Carrier frequency f_c (used to normalize the deviation and modulation frequencies) will be the mean of the upper and lower frequencies of the segment, and the deviation itself half the difference between these frequencies. It was found that the deviation

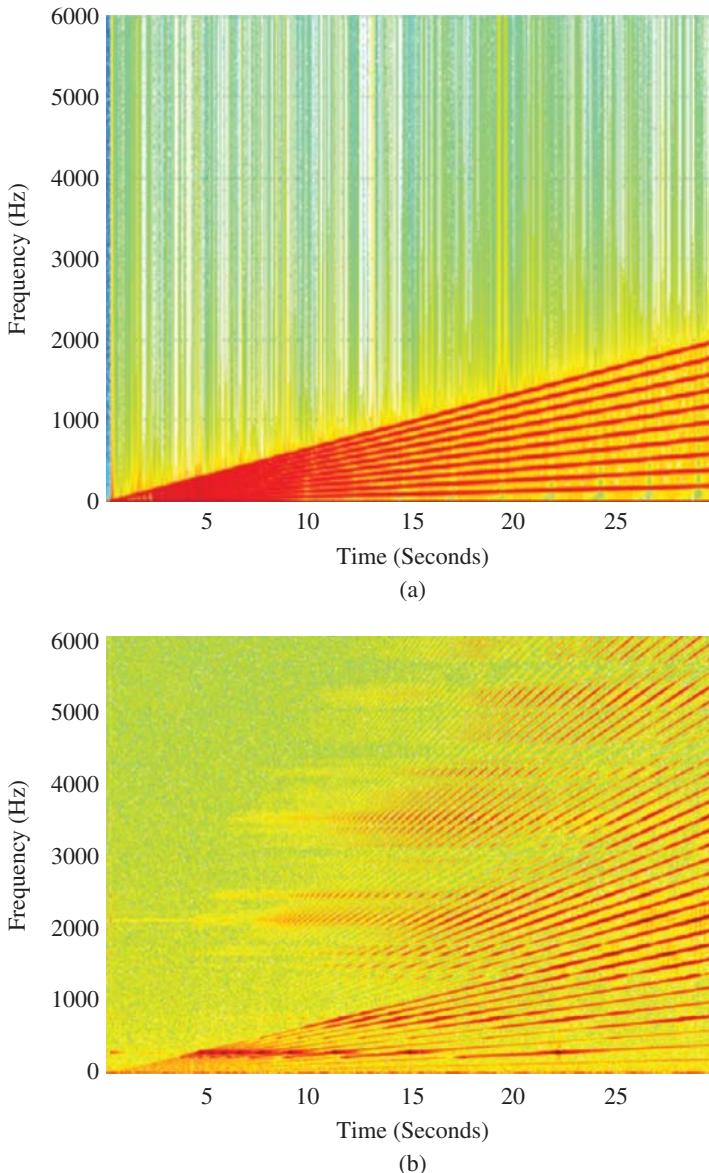


Figure 5.12 Spectrograms (a) Force signal (b) Response acceleration.

could have been as high as $\pm 30\%$ in the high frequency segment and $\pm 25\%$ in the lowest frequency segment, but $\pm 18\%$ was chosen for the demonstration here. Five segments were chosen on this basis, meaning that the lowest frequency of segment 1 (the lowest speed) was 16% of the highest speed, or 32 Hz.

The phase vs time curve was determined for each segment by phase demodulation of the first order of the non-windowed tacho signal. Figure 5.14 compares the spectra of these segments for the lowest (segment 1) and highest (segment 5) speeds. Note that even though they are numbered from

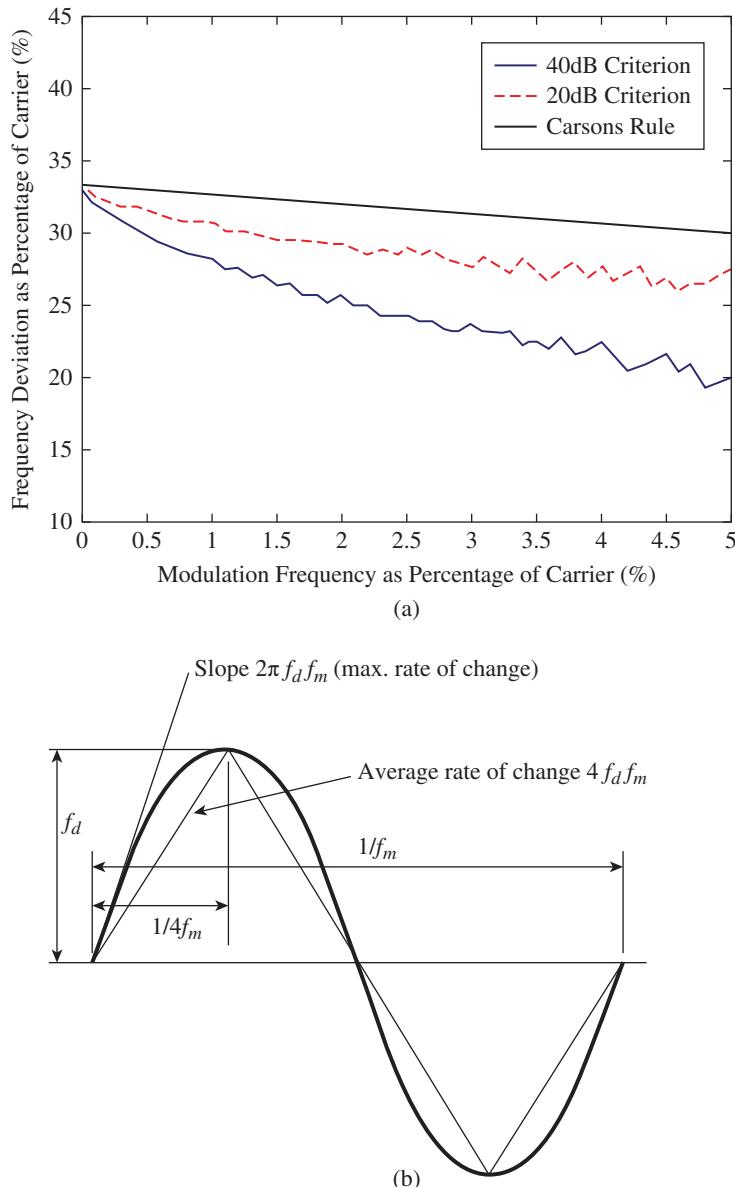


Figure 5.13 (a) Low frequency part of Figure 5.5 (b) Equivalent modulation frequency.

lowest to highest, the order of processing was in the inverse order. Because a smaller deviation was chosen, it can be seen that even the second order is almost separated. To aid in the demodulation and unwrapping of the resulting phase curve, the carrier frequency f_c is removed (shifted to zero) as part of the demodulation process, but it is known precisely, and can be added back in to obtain the total phase $2\pi f_c t + \phi(t)$ where $\phi(t)$ is the variation in phase (around the carrier) found by the demodulation process. For each segment, the initial phase at the beginning of the record is typically

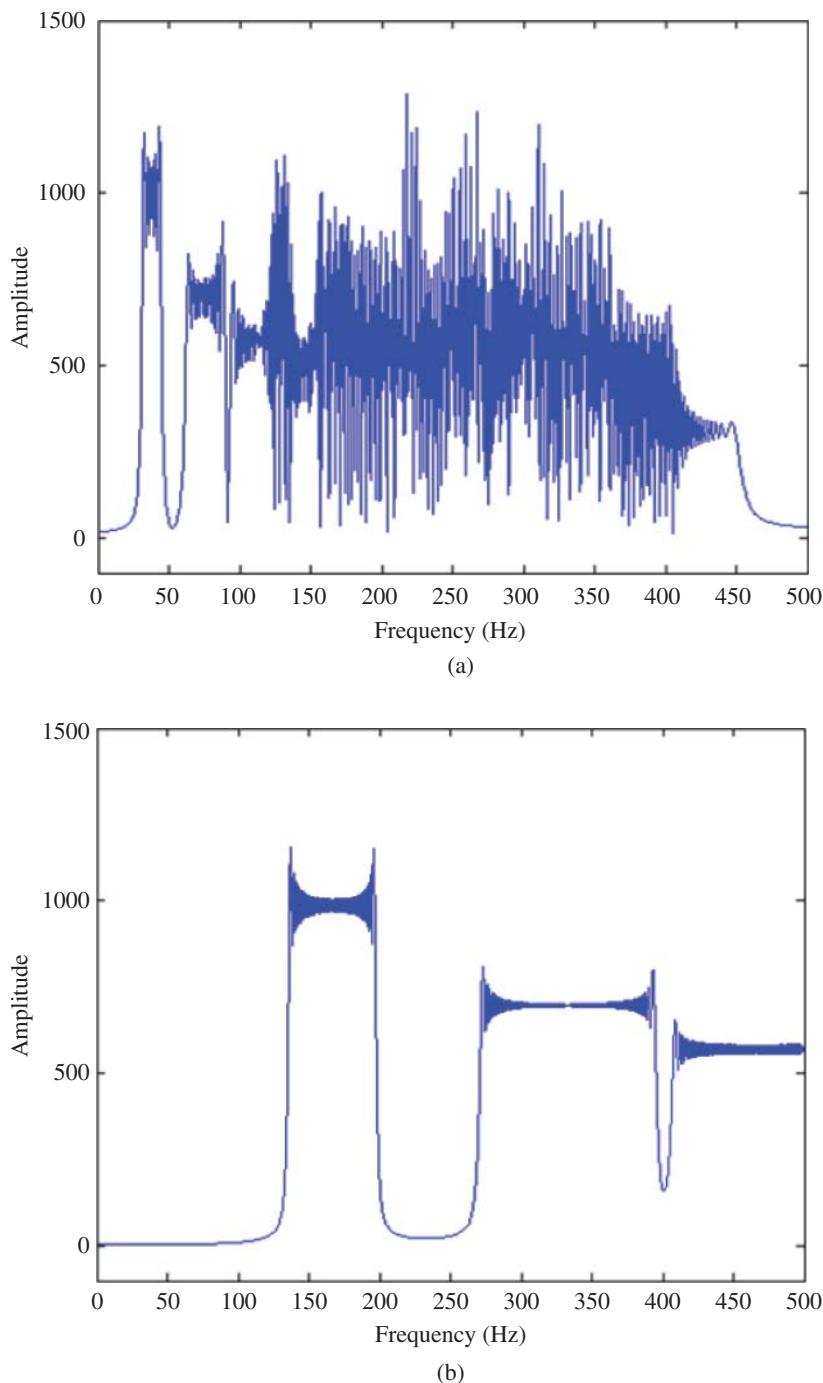


Figure 5.14 Spectra of segment tacho signals (a) Segment 1 (b) Segment 5.

near zero, but must be raised to match the phase of the adjacent segment at the common breakpoint. It is suggested that the matching should be done at these breakpoints, where there is equal accuracy in both segments. Thus, even though the demodulation is done from right to left, the phase at the lowest frequency of the lowest segment can be set to zero, and corrections finally added to each higher segment so as to attain a match at each breakpoint.

Figure 5.15 shows order spectra after order tracking of the lowest and highest speed segments (1 and 5). It will be noticed that orders 5 to 10 are a little smeared in (a), but since the corresponding orders of the tacho signal were not smeared, it is evident that this is a result of amplitude modulation.

Figure 5.16 shows the overlaid order tracked signals (both tacho and response) which became continuous after addition, and Figure 5.17 the corresponding order spectra.

Finally, Figure 5.18 shows spectrograms of the order tracked signals, both tacho and response. The resample frequency was chosen to be 120 samples per period of the fundamental order, so that the Nyquist ‘frequency’ corresponded to order 60. It is seen that the response order spectrum encroaches on this at the lowest speed in each segment, but that the antialiasing filter cut-off is much gentler in the highest speed segment because it is an analogue filter, with ideal filters in the lower speed segments. The retained highest order is everywhere >32 .

5.2 Determination of Instantaneous Machine Speed

Condition monitoring of variable speed machines is becoming more important, with the increase in application to machines such as wind turbines, mobile mining equipment, marine drives, etc. Quite apart from correcting for speed variations, using order tracking as in the previous section, it is often useful to have an exact measure of instantaneous rotational speed, and this section shows how this can be done using both tachometer signals and vibration response signals, with an estimate of accuracy. Since the so-called Teager Kaiser Energy Operator (TKEO) has been used for this purpose, the opportunity is taken to discuss this, and other energy operators, in some detail.

First, the most accurate method of finding the instantaneous frequency of a mono-component signal (single carrier frequency, modulated in both frequency and amplitude), the derivative of the instantaneous phase, is discussed.

5.2.1 Derivative of Instantaneous Phase

A mono-component signal (Eq. (3.51) of Chapter 3) can be written as the analytic signal,

$$x_a(t) = A(t) \exp(j\phi(t)) \quad (5.3)$$

where the instantaneous phase $\phi(t)$ includes the linear carrier frequency term $2\pi f_c t$. As explained in [9], the instantaneous frequency will then be given by the derivative of the phase

$$\dot{\phi}(t) = \omega(t) = 2\pi f(t) \quad (5.4)$$

but only on the condition that $x_a(t)$ is analytic (positive frequencies only), as for all signals treated in Section 5.1.3.

Since $x_a(t)$ is the product of the amplitude and phase modulation components, its spectrum is the convolution of the two spectra, with a bandwidth less than the sum of the two bandwidths, and centred on the carrier frequency. The bandwidth of $A(t)$ is directly that of its spectrum, and thus constant. Even though the bandwidth of $\exp(j\phi(t))$ is theoretically infinite, the sidebands are only significant over a band roughly corresponding to the frequency sweep (from minimum to maximum

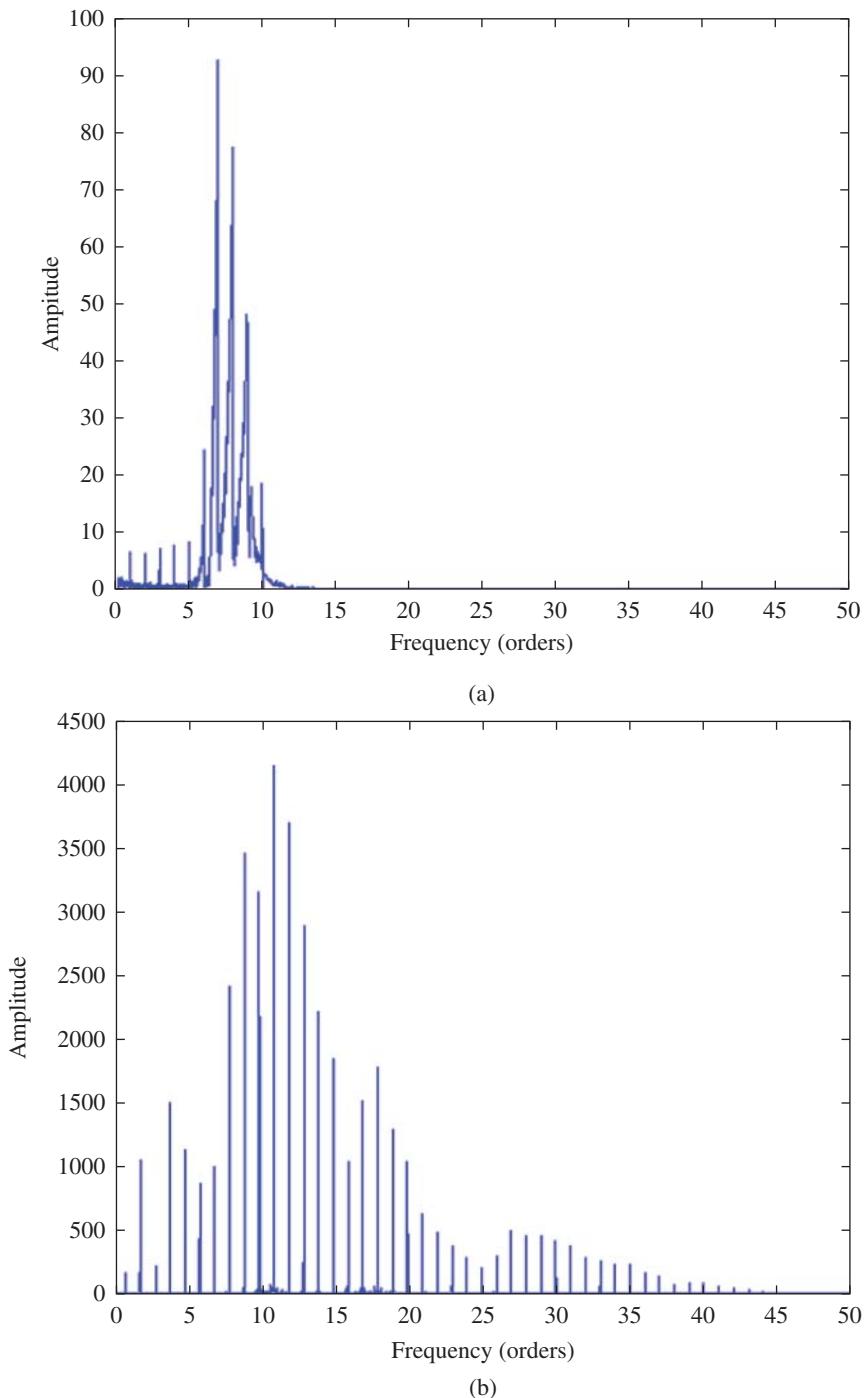


Figure 5.15 Typical order spectra of response signal after order tracking (a) Segment 1 (b) Segment 5.

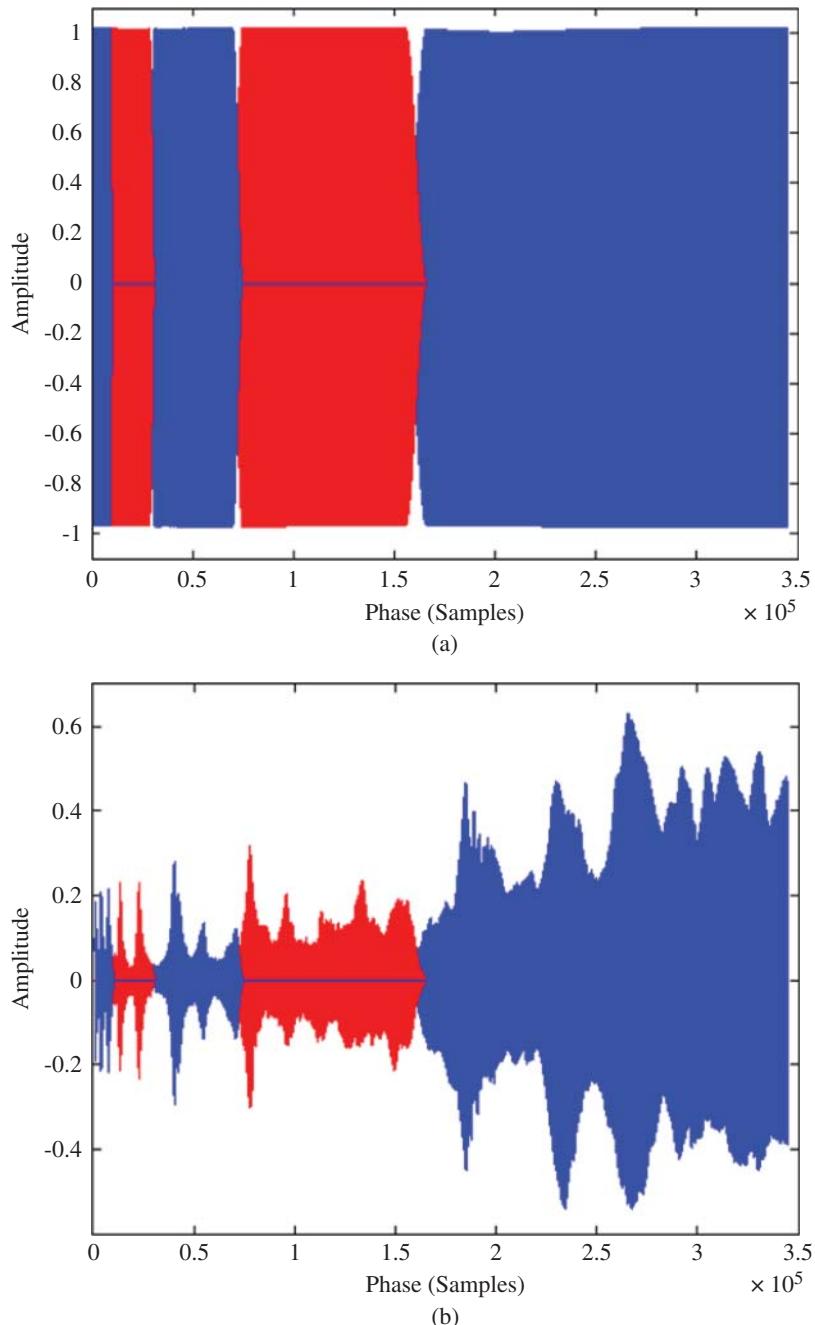


Figure 5.16 Overlaid order tracked signals for the five segments (a) Tacho (b) Response.

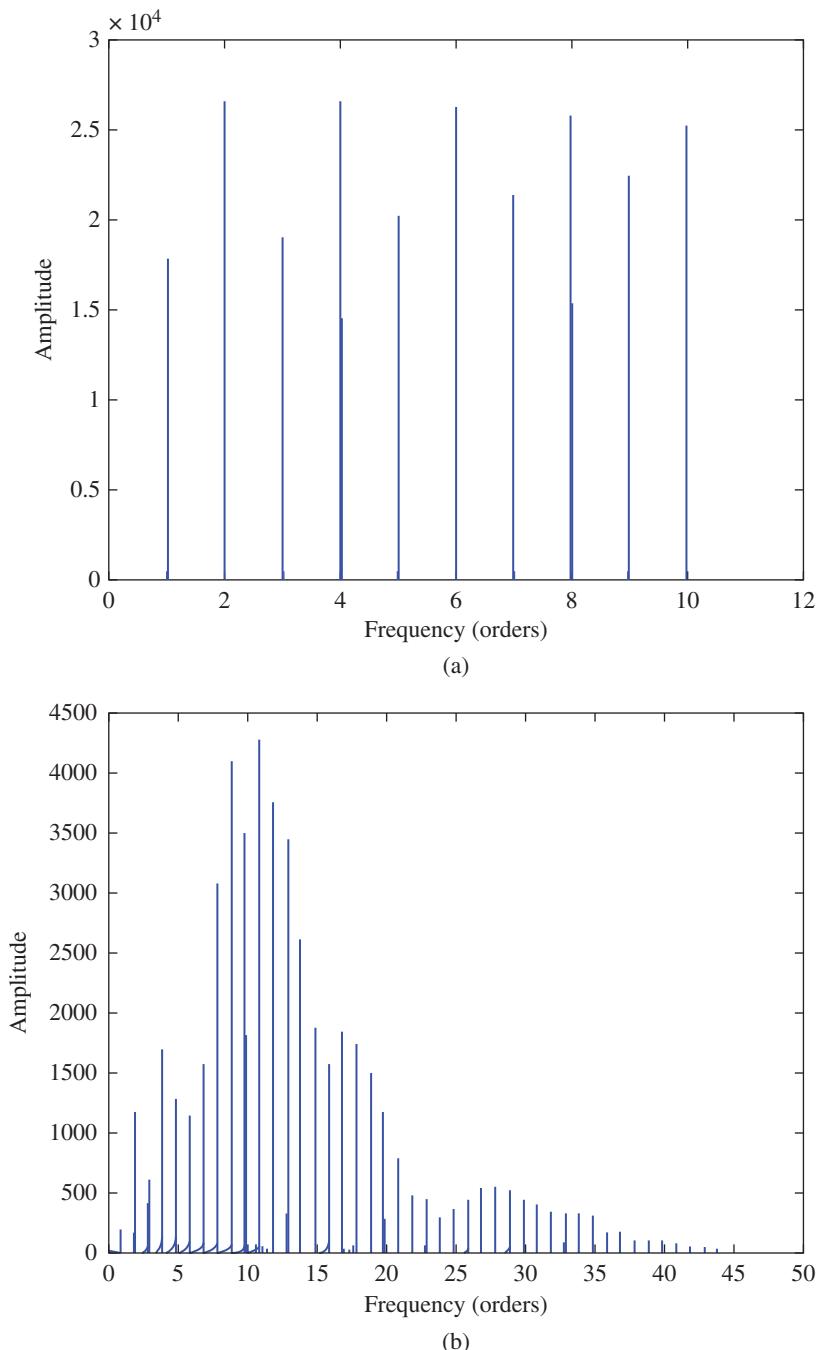


Figure 5.17 Order spectra of the re-combined signals (a) Tacho (b) Response.

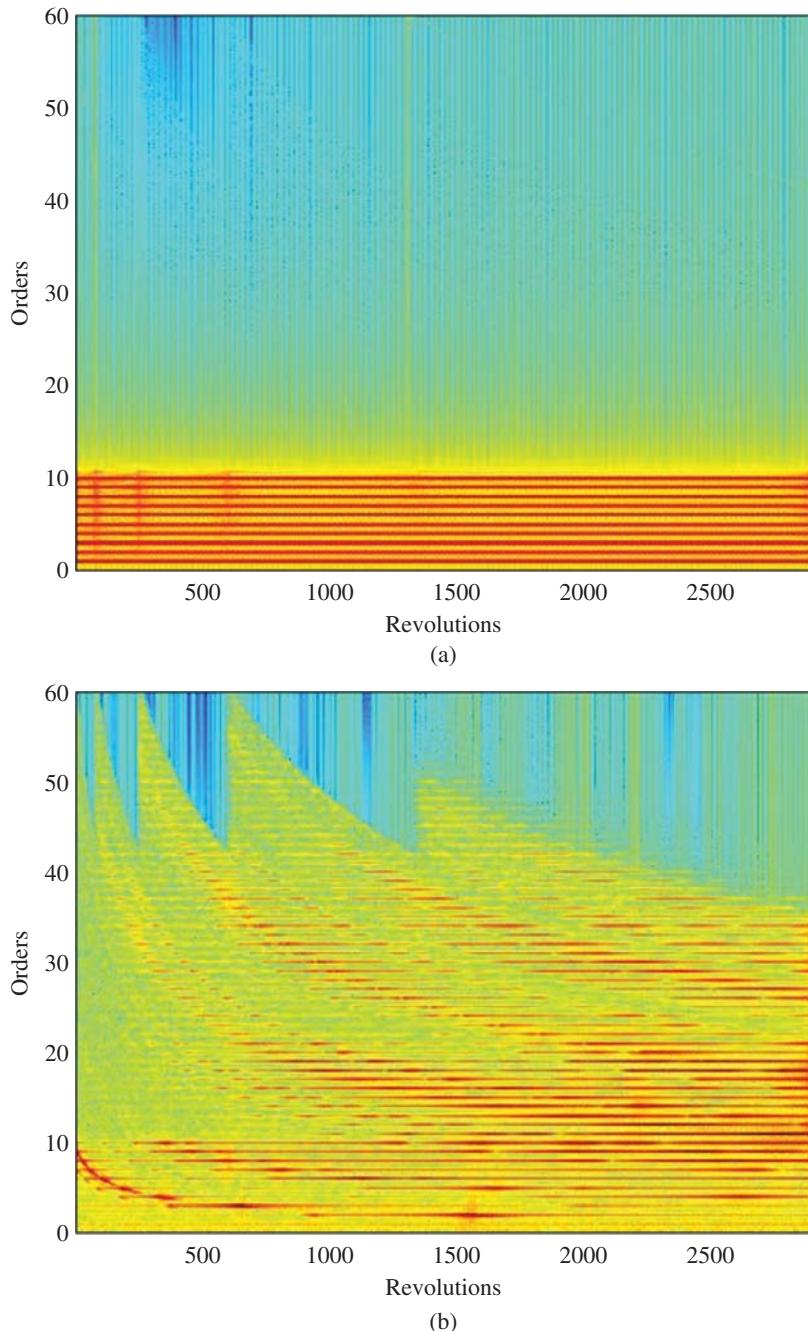


Figure 5.18 Spectrograms for the order tracked signals (a) Tacho (b) Response signal.

in the frequency modulation) and will thus give an analytic signal if the difference between the minimum frequency and the carrier frequency is less than the carrier. More accurate estimates are given in Section 5.1.3 and [9].

In [9] it is shown that the derivative of the instantaneous phase of an analytic signal is:

$$\omega(t) = \dot{\phi}(t) = \frac{x(t)\dot{\hat{x}}(t) - \hat{x}(t)\dot{x}(t)}{A_x^2(t)} = \text{Im} \left[\frac{\dot{x}_a(t)}{x_a(t)} \right] \quad (5.5)$$

where $\hat{x}(t)$ is the Hilbert transform of $x(t)$, so that $x_a(t) = x(t) + j\hat{x}(t)$.

Equation (5.5) can be very efficiently evaluated via the frequency domain, since $x_a(t)$ has a one-sided spectrum of limited bandwidth, and the spectrum of its derivative $\dot{x}_a(t)$ is obtained by multiplying by $j\omega$ over this bandwidth. Note that this is an exact differentiation for those frequencies, in contrast to numerical differentiation in the time domain, which always involves errors. Inverse transforming the spectra of both $\dot{x}_a(t)$ and $x_a(t)$ back to the time domain, the imaginary part of their quotient gives the instantaneous angular velocity (which for an analytic signal must necessarily be positive everywhere). The disadvantage of differentiating in the frequency domain comprises the wraparound errors at the ends of records, caused by the circularity of the FFT operation, but these can usually be removed by truncation.

5.2.2 Teager Kaiser and Other Energy Operators

The TKEO has been used for amplitude and frequency demodulation of mono-component signals, even for machine diagnostic applications, and so it is discussed in some detail here, primarily to show that in general it gives no advantage in machine diagnostics, and instead gives several disadvantages. On the other hand, some other energy operators, similar to the TKEO, can be validly used for some diagnostic purposes. The main advantage claimed for the TKEO is that it can be calculated very efficiently, effectively in real-time in the time domain, and that in fact is a major advantage in its original application of speech analysis. It does mean, however, that to realise this advantage all other processing used in conjunction with it must also be real-time, or causal, introducing many unnecessary errors in the results.

For machine diagnostics, and in particular for machine condition monitoring, there is no advantage whatsoever in real-time processing, and on the other hand considerable advantage in using non-causal processing (often via the frequency domain), for example to achieve ideal, zero phase shift bandpass filtration, as well as the error-free differentiation/integration mentioned above. This is because condition monitoring is usually seeking information on faults which will develop over a period of days, weeks or months. Even for continuous online monitoring, with a view to shutting machines down very quickly, the difference between real-time and non real-time processing (usually less than one second) would be immaterial, since most often an operator has to decide whether to shut the machine down, and even when tripped, many machines would continue to have high speed for much longer than a second.

The idea for the TKEO was first put forward by Teager and then formalised by Kaiser [10] for use in speech analysis. It was supposed to represent the ‘total energy’ i.e. kinetic (KE) plus potential (PE) with the formula given in continuous and discrete forms by:

$$\Psi_c(x(t)) = [\dot{x}(t)]^2 - x(t)\ddot{x}(t) \quad (5.6)$$

$$\Psi_d(x(n)) = [x(n)]^2 - x(n+1)x(n-1) \quad (5.7)$$

References. [11, 12] give a thorough study of the errors involved in the continuous and discrete versions. It can be seen that Eq. (5.7) uses only three adjacent samples, and can thus be evaluated

efficiently in effective real time, as appropriate to speech analysis. Ref. [11] also gives an important result that is best derived as follows, for the case where the rates of change of the amplitude and instantaneous frequency are low and can be neglected. Assuming the amplitude and frequency modulated mono-component can be expressed as:

$$x(t) = A(t) \cos \phi(t), \quad \text{where } \omega(t) = \dot{\phi}(t) \quad (5.8)$$

$$\dot{x}(t) = -\dot{\phi}(t)A(t) \sin \phi(t) + \dot{A}(t) \cos \phi(t) \approx -\omega(t)A(t) \sin \phi(t) \quad (5.9)$$

then

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

and

$$\Psi_c[x(t)] = [x(t)]^2 - x(t)\ddot{x}(t) \approx [\omega(t)]^2[A(t)]^2(\sin^2 \phi(t) + \cos^2 \phi(t)) = [\omega(t)]^2[A(t)]^2 \quad (5.10)$$

Kaiser used the analogy of a mass on a spring, to which energy was slowly being added or subtracted, to claim that the first term of (5.6) is proportional to the KE in the mass and the second term to the PE (strain energy) in the spring, and that the TKEO thus represented the total instantaneous energy. However, this only applies when $x(t)$ is displacement, since if it were velocity the total energy would be given directly by the squared envelope of the signal, this being the sum of squares of the signal and its Hilbert transform. If the measured signal is acceleration, the TKEO is the squared envelope of the jerk. It can easily be shown that the Hilbert transform of $\dot{x}(t)$ in (5.9) is $\omega(t)A(t)\cos\phi(t)$, meaning that the squared envelope of $\dot{x}(t)$ is given by:

$$A_{\dot{x}}^2(t) = \dot{x}^2(t) + \hat{\dot{x}}^2(t) = [\omega(t)]^2[A_x(t)]^2 \approx \Psi_c[x(t)] \quad (5.11)$$

where $A_x(t)$ is now used for the envelope of the signal. From (5.11) it immediately follows that:

$$\omega^2(t) = \frac{A_{\dot{x}}^2(t)}{A_x^2(t)} \quad (5.12)$$

This result was presented in [13] where it was not realised that it gives a more accurate result than using the (time domain) TKEO, in particular if the squared envelopes and differentiation are estimated using the Hilbert transform via the frequency domain, as suggested for Eq. (5.5).

In fact, it is not possible to use the TKEO directly to get an estimate of the instantaneous frequency of the signal. Formulas for frequency and amplitude demodulation of a mono-component signal using the TKEO, are derived in [12] and give:

$$(a) \omega(t) \approx \sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}}, \quad (b) A(t) \approx \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \quad (5.13)$$

and by comparing (5.13a) with (5.12) it can be seen that (5.13a) actually gives the squared frequency of the derivative of the signal, since its denominator is the numerator of (5.12). It will be shown this gives a systematic error.

For that reason, even though the squared envelope of the derivative of a signal was called the ‘TKEO’ in [13], it was decided to rename it the ‘Frequency Domain Energy Operator’, or FDEO, when calculated using Hilbert methods via the frequency domain. On doing some research it was realised that this relationship had been recognised earlier, leading to the definition of the ‘Frequency Weighted Energy Operator’, or FWEO in [14], though we believe it is something of a misnomer,

since the frequency weighting is a direct result of the differentiation, which is an intrinsic part of the definition of the TKEO. In any case, it was not suggested in [14] that it should be estimated via the frequency domain, and the authors proposed use of an ‘envelope–derivative operator’, still enacted in the time domain. They do remark that this is less ‘real-time’ than the TKEO, but that their application to EEG signals does not require real-time processing.

5.2.3 Comparison of Time and Frequency Domain Approaches

As mentioned already, the main advantage of the time domain version of the TKEO is the fact that it is real-time, as appropriate for its original application of speech analysis, but this is no advantage for machine diagnostic applications, and in fact a disadvantage since it enforces the use of causal signal processing. The latter means that filters have poor characteristics, with phase shifts near low and high cut-off frequencies, and a limited rate of roll-off outside the pass-band. Numerical differentiation gives problems with amplification of high frequency noise as well as phase shifts.

On the other hand, non-causal signal processing, often just involving simple windowing in FFT operations, can give close to ideal (i.e. rectangular) filters with no phase shift. Differentiation and integration can be achieved by $j\omega$ operations in the frequency domain, with virtually zero error in the processed band (selected with an ideal filter), and Hilbert transforms are also most efficiently carried out via the frequency domain.

The disadvantage of non-causal processing is the circularity of FFT processing, forcing time records to be periodic, giving discontinuities when the ends are joined into a loop. However, these end effects are limited if the discontinuity is limited, which is generally the case when the speed range in one record is limited to $<2:1$, and the amplitude variation not too great. As shown by the examples in this section, the end effects do not extend very far and can be truncated if a slightly longer record is processed.

5.2.3.1 Examples of Accuracy Using Different Algorithms

The following example shows that differentiating a signal usually changes the mixture of amplitude modulation (AM) and phase/frequency modulation (PM/FM), so both cannot represent the true FM of the signal if it changes. This becomes obvious by reference to Figure 7.16 of Section 7.2.2 of Chapter 7. This shows an example of a signal with pure AM and zero PM/FM, modified by a transfer function with a slope in the vicinity of the AM spectrum (carrier with a symmetric pair of sidebands). At some time, which without loss of generality is set to zero here, the carrier and the sidebands must be aligned so that the sidebands add to a vector in line with the carrier, and as time moves away from zero, the two sidebands, rotating in opposite directions with respect to the carrier, add to a component aligned with the carrier and changing its length sinusoidally at a frequency corresponding to the sideband spacing from the carrier. When this spectrum is multiplied by the transfer function shown (even without phase change), it makes the sidebands uneven, but they can be decomposed into a symmetric pair, giving AM, and an asymmetric pair which in fact primarily give PM/FM, when the sidebands are small with respect to the carrier. This is illustrated in Figure 3.29 of Chapter 3. Since FM is the derivative of PM (angular velocity vs angular displacement) the FM is scaled by multiplication by ω_m , the modulation frequency in rad/s.

The true instantaneous frequency of an analytic signal is given by Eq. (5.5), but this can be developed [15] to give the squared value:

$$\omega^2(t) = \frac{A_x^2(t)}{A_x^2(t)} - \left(\frac{\dot{A}_x(t)}{A_x(t)} \right)^2 \quad (5.14)$$

which is seen to be consistently smaller than the value given by Eq. (5.12), and independent of $\dot{\omega}(t)$. When the squared frequency is evaluated using Eq. (5.13a), based on the TKEO from Eq. (5.7), there are two further sources of error, the first being that it gives the squared frequency of the derivative of the signal, and the second arising from the numerical (double) differentiation implied by Eq. 5.13a).

Figure 5.19 shows a numerical example for a case of pure AM ($\pm 20\%$) of a 500 Hz carrier, modulated at 10 Hz. The ‘exact’ result (Eq. (5.5)) has no error, while the ‘Ratio of SEs’ (squared envelopes) has the small error given by the second term in (5.14), but the ‘Exact (next deriv)’ and ‘Ratio of SEs (next deriv)’ show that the error involved in differentiation is at least an order of magnitude greater, while the error from the additional high frequency ripple in the ‘TKEO’ result is about 50% greater again. The formula for the error purely resulting from differentiation is seen to be relatively small when the spread of sidebands is small as in this case. This would seem to indicate that the error could be significantly greater for FM, where the spread of sidebands can be much greater, even for modulation by a single frequency. The AM sidebands would convolve with the FM sidebands across their whole range. However, Eq. (5.14) shows that the result of a single differentiation does not give an error in the frequency of a pure FM signal (since it depends only on the rate of change of amplitude), but this does then introduce an amplitude modulation, which will lead to errors in subsequent differentiations. In any case, most machine signals will have some AM in addition to FM, and will also give frequency errors for a single differentiation.

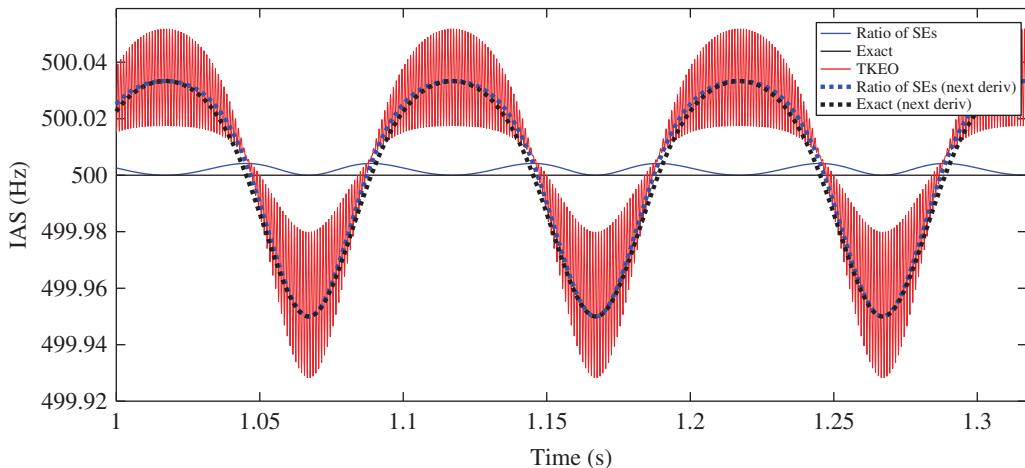


Figure 5.19 Effect of differentiation for a signal with pure AM, and carrier frequency 500 Hz.

The question then arises as to the effect on the instantaneous frequency of a vibration response signal as compared with the rotational speed. If, as is likely, the frequency of internal forces is most closely related to the machine speed, the frequency of response signals will be affected by passage through a transfer function between the force and the response. If the demodulated frequency band is in the vicinity of a resonance, with phase shift, the instantaneous frequency errors can be quite large because of time delays, as discussed in detail in [6]. However, even if the band lies on a spring line or mass line of the FRF, where response is immediate, the frequency will depend on the measured parameter and whether the band is on a spring line or mass line. This is illustrated in Figure 5.20, which shows the resulting errors in instantaneous frequency for a pure FM signal (carrier frequency 500 Hz, frequency sweep ± 100 Hz) when lying on the spring line of a resonance (at 1500 Hz), or on the mass line of a resonance (at 150 Hz). The error given by the exact equation is compared with that given by the TKEO, Eq. (5.13a), the latter being much greater. When the demodulated band is on a mass line, which is almost constant for measured acceleration, the error using Eq. (5.5) is very small, but when it lies on a spring line (effectively double differentiated) the error is almost 0.05%. It could be corrected by double integrating as part of the frequency estimation process.

Figure 5.21a shows an example from [15, 16] of the instantaneous speed of a gearbox, estimated using the FDEO from a 2 per rev tacho signal on the output shaft, with a mean tacho frequency of about 74 Hz, and the acceleration signal of the input shaft, with a mean frequency of about 20 Hz (gear ratio 1.84), both scaled for the speed of the input shaft. For both of these signals the band encompassed by the first harmonic was free of interference from other components. Both curve estimates were smoothed as discussed below, but each had about the same amount of noise before smoothing. There is no absolute measure of the correct result, though the tacho could be expected to be more accurate. Even so, outside the end effect zones, the maximum difference is 0.24% and standard deviation 0.03%.

There was some noise in the demodulated results, and this was smoothed using a zero phase shift moving average filter (Matlab function FILTFILT) of length 100 samples. Figure 5.21b compares the unsmoothed and smoothed results near the right hand end of the speed profile obtained from the acceleration signal. This is scaled in samples, but corresponds to the section from 41–45 seconds in Figure 5.21a. It is seen that the extent of the effects of both the wraparound error and the smoothing filter is of the order of the smoothing filter (100 samples), and could be removed by truncation. It can be seen in Figure 5.21a that the extent of the end effect is longer (in seconds) for the response derived curve, but this is because of the lower sampling frequency of the demodulated signal (related to the signal frequency). The end effect for the higher frequency tacho signal was about the same number of samples.

Figure 5.22a shows the estimated speed using the TKEO method (Eq. (5.13a)) on the same data as Figure 5.21. To simulate a real-time result, the demodulated band was extracted using an IIR Butterworth filter. The spectrum of the extracted signal is shown in Figure 5.22b. As expected, the TKEO estimate was much noisier than the FDEO result of Figure 5.21, and even after smoothing using the same (non-causal) filter, it is still noisier.

As a matter of interest, the spectrum of the response signal in the demodulation band around 20 Hz was studied to see if it were on a spring line or mass line. An exponential lifter [17] was applied to the response spectrum (effectively that of Figure 5.22, but over a much wider frequency range) to greatly de-emphasise the forcing functions and enhance the modal response (though with added damping). In Figure 5.23 it can be seen that this does appear to find the dominant resonances in the transfer path. However, in the vicinity of the demodulated band, even though it appears to be gently rising, the slope over the octave from 14–28 Hz is much less than the 12 dB which would correspond to a spring line, whereas the region from 300–600 Hz is very close to a spring line.

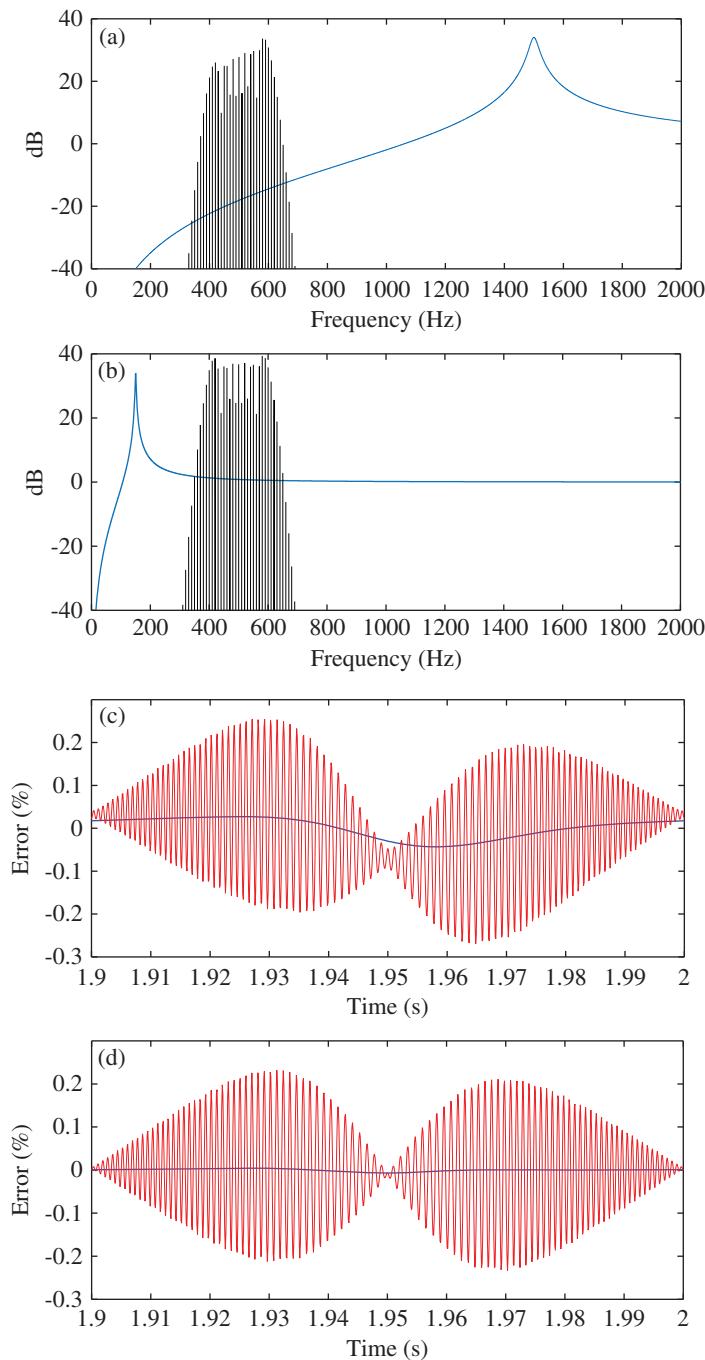


Figure 5.20 Effect of transfer function on errors in IAS based on response signal. (a, b) Signal spectrum and FRF; (c, d) % Errors, (oscillating): TKEO, (smooth): Exact; (a, c) Signal spectrum on spring line; (b, d) Signal spectrum on mass line

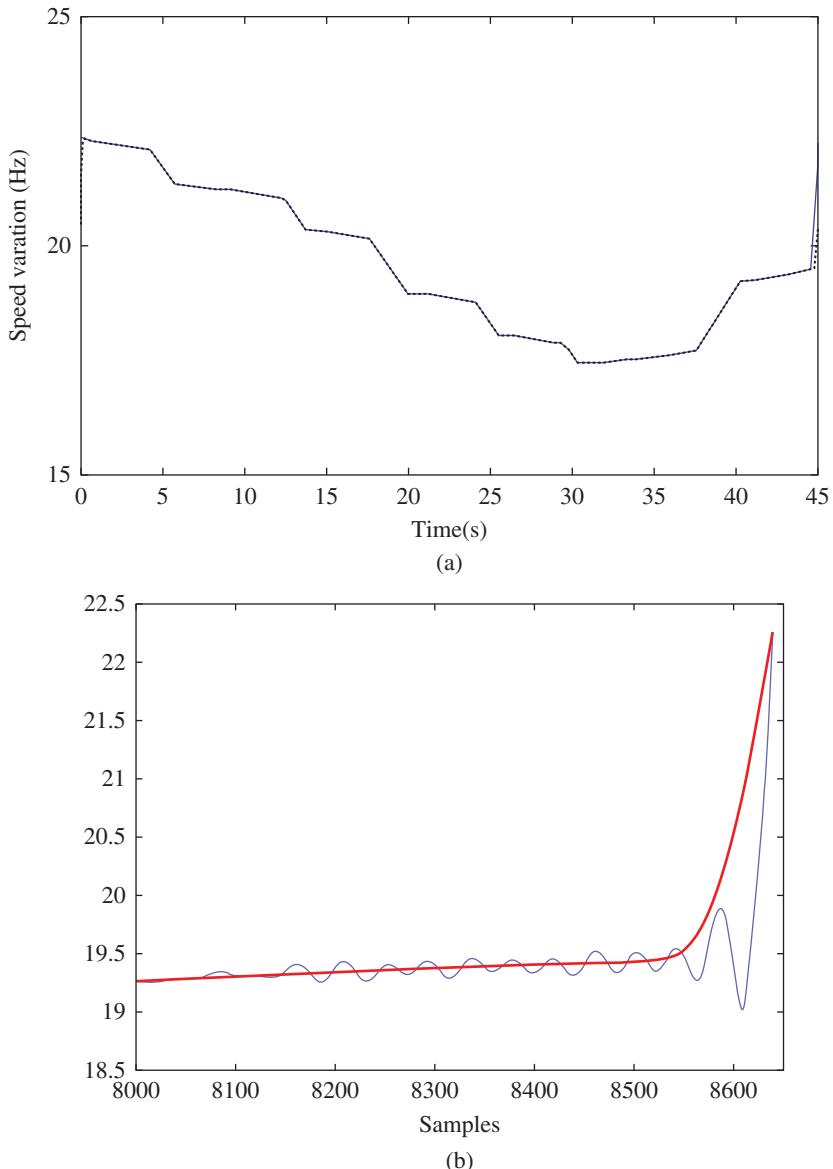


Figure 5.21 (a) Comparison of speed estimates from acceleration (solid) and tacho (dotted) adjusted for ratio
(b) Zoom on end effects for acceleration signal wraparound error (light), smoothed result (solid).

5.2.4 Other Methods

One widely used method, though not very accurate, and often followed up by a later stage of improvement by another method, is based on tracking peak values in a spectrogram, as in Ref. [7] and illustrated in Figure 5.10 of Section 5.1.4, where its use for order tracking over a wide speed range is described.

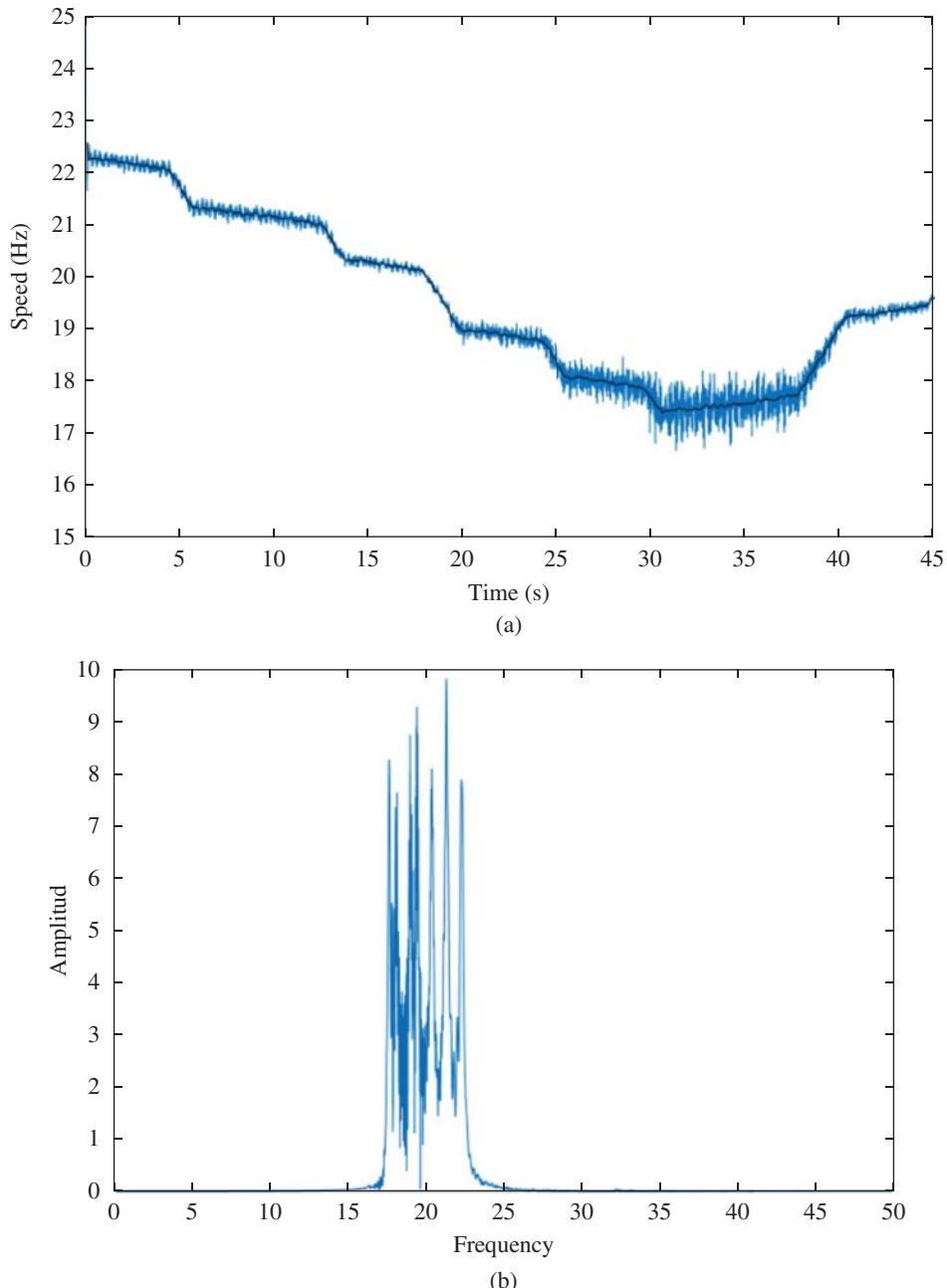


Figure 5.22 Speed estimates using TKEO (Eq. (5.13a)) on a bandpass filtered signal encompassing the first harmonic of the acceleration signal. (a) Raw speed estimate and smoothed version (b) Spectrum of bandpass filtered signal.

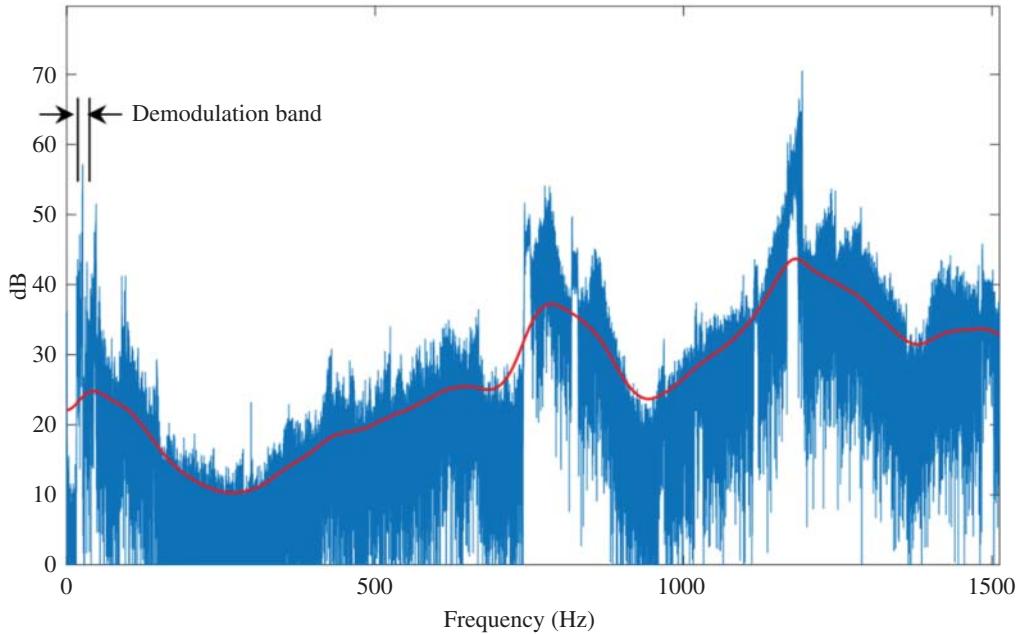


Figure 5.23 Use of an exponential lifter to extract the modal part (solid line) of the response acceleration signal.

Probably the most accurate method, in the presence of masking noise, is that presented in Ref. [18], based on the probability density functions over a number of harmonics, not only of a particular shaft, but possibly over the harmonics of a number of shafts whose geared ratios are all known with respect to a reference shaft in a complex gearbox. The method starts from a spectrogram showing the variation of all harmonics to be taken into account. To avoid distortion and excessive weighting being given to individual harmonics as they pass through resonances, the first step is to whiten the base noise level in the spectrum, on the assumption that the latter provides an indication of the typical modal weighting at each time (speed). In this paper it was done by using a local average of the noise between discrete order components, and unifying it throughout the diagram to obtain pre-whitening of the background noise. In principle, it is no different from achieving the same result using an exponential lifter to determine the modal response, and then subtracting it, for example as done in Figure 6.22 of Section 6.3.2 of Chapter 6.

The result of this first pre-whitening step is illustrated in Figure 5.24a (from [18]), which is based on the data from the *CMMNO'14 diagnosis contest* mentioned in the title of the paper. The data is from a wind turbine gearbox, with initial planetary stage followed by two parallel gear stages. All details are given in [18]. It is seen that all discrete orders protrude from a uniform noise base.

The method is fully explained in [18], but can briefly be described as follows. At each time slice in the spectrogram the spectrum value in a narrow band around each order is treated as the PDF of the estimated frequency. To combine the estimates for several harmonics of the same fundamental order, the frequency values are divided by the harmonic order, and the combined PDF for that fundamental order is obtained by multiplying the individual estimates together. The same principle is applied to obtain the PDFs in terms of a particular reference fundamental frequency, (called the IAS, or instantaneous angular speed) using the known geared ratios between them.

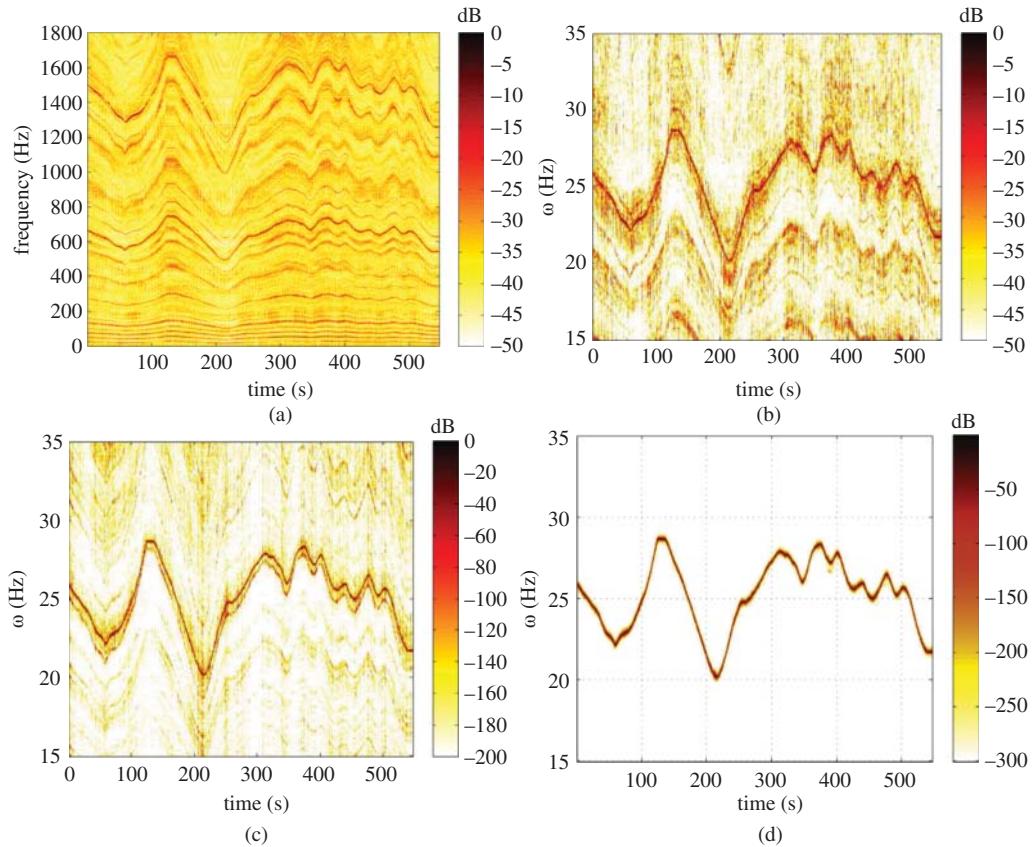


Figure 5.24 (a) Spectrogram with pre-whitened noise base (b) PDF of the IAS at each time step, considering fundamental meshing orders (c) PDF of the IAS at each time step, including 10 first harmonics of each fundamental order (d) Final estimate obtained by smoothing (c).

Figure 5.24b shows the result of combining the estimated IAS (at each time step) using only the six fundamental meshing orders, and Figure 5.24c the improvement given by including the first 10 harmonics of each. The result of Figure 5.24c is still only valid for the individual times at which it has been estimated, and may have some distortion from the rate at which the speed is changing at each time.

The final step is to smooth the results at each time step, but as opposed to simple lowpass filtration the approach applied uses probabilistic methods to ensure continuity of the IAS while smoothing the PDFs. The final result is shown in Figure 5.24d, and has a deviation mostly less than 0.1% of the ‘true’ value given by the reference angle encoder, with a small number of larger deviations up to 0.4%.

5.3 Deterministic/Random Signal Separation

A large part of condition monitoring consists of separating the mixed signals at each measurement point into the components coming from individual sources. An important part of this consists in

separating deterministic signals (for example from gears) from random signals. The latter include general background noise including measurement noise, noisy source signals such as from fluid flow in a pump or turbine, and cyclostationary signals, for example from rolling element bearings (see Sections 2.1.3, 2.2.3, 3.6, and 7.3). Even with machines running at nominally constant speed, the signals directly phase locked to shaft speeds are not truly deterministic unless any small random speed variations are removed by order tracking, as discussed in Section 5.1. In particular, order tracking must be used for methods based on time synchronous averaging, which is a classic method of separating deterministic and random signals. There are a number of other methods, however, not all requiring order tracking if the speed is reasonably constant, and all having different pros and cons, as discussed in Ref. [19]. Most of those methods, viz linear prediction, adaptive noise cancellation, self adaptive noise cancellation, and discrete/random separation, are discussed below, while a new method based on cepstrum analysis is discussed in Chapter 6.

5.3.1 Time Synchronous Averaging

The classic way of separating periodic signals from background noise (and everything else not periodic with a particular fundamental frequency) is by time synchronous averaging (TSA). It has been used over many years for extracting the vibration signal corresponding to a particular gear in a gearbox.

In practice it is done by averaging together a series of signal segments each corresponding to one period of a synchronising signal. Thus:

$$y_a(t) = 1/N \sum_{n=0}^{N-1} y(t + nT) \quad (5.15)$$

This can be modelled as the convolution of $y(t)$ with a train of N delta functions displaced by integer multiples of the periodic time T , which corresponds in the frequency domain to a multiplication by the Fourier transform of this signal, whose amplitude can be shown to be given by the Expression (5.16) (from [20]); note that the expression given in [20] for the ‘modulus’ does not have the modulus signs on the right-hand-side of the equation):

$$|C(f)| = |(1/N) \sin(N\pi Tf)/ \sin(\pi Tf)| \quad (5.16)$$

The filter characteristic corresponding to this expression is shown in Figure 5.25 for the case where $N = 8$, and is seen to be a comb filter selecting the harmonics of the periodic frequency. The greater the value of N the more selective the filter, and the greater the rejection of non-harmonic components. The noise bandwidth of the filter is $1/N$, meaning that the improvement in signal/noise ratio is $10 \log_{10} N$ dB for additive random noise. For masking by discrete frequency signals, it should be noted that the characteristic has zeros which move with the number of averages, so it is often possible to choose a number of averages which completely eliminates a particular masking frequency.

The above characteristic is for an infinitely long time signal $y(t)$, and in Ref. [21] it is shown that for the practical situation of a finite length of signal with finite sampling frequency, the above simplified model is not exactly true for discrete frequency ‘noise’, but that it is often still possible to find an optimum number of averages to completely remove a discrete masking signal.

For good results the synchronising signals should correspond exactly with samples of the signal to be averaged, as one sample spacing corresponds to 360° of phase of the sampling frequency, and thus to a possible error of 144° of phase at 40% of it, which is a typical maximum signal frequency. Moreover, even a 0.1% speed fluctuation would cause a jitter of the same order as the sample spacing

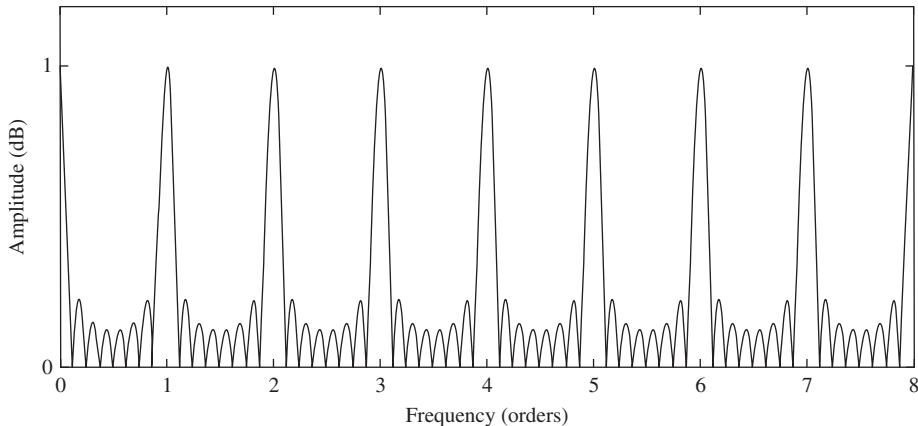


Figure 5.25 Filter (amplitude) characteristic corresponding to 8 synchronous averages. Source: from [20].

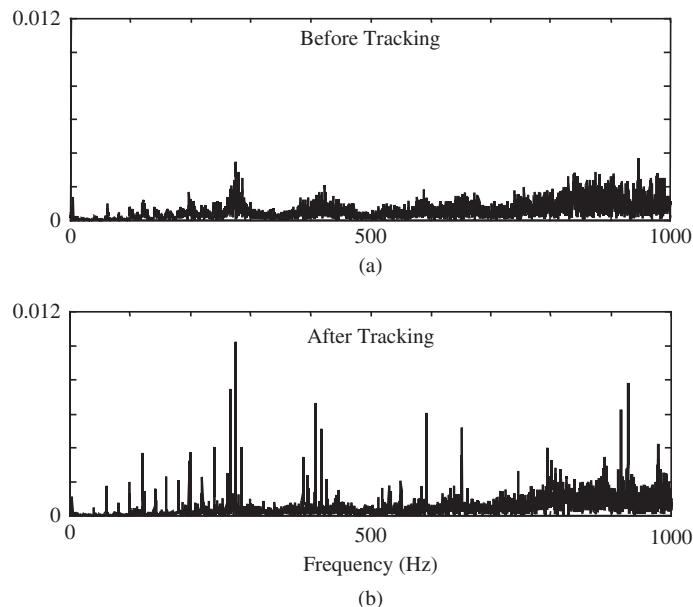


Figure 5.26 Use of tracking to avoid smearing of shaft speed related components.

in a (typical) 1K record, with respect to the first, and thus an even greater loss of information at the end of the record, after averaging. Order tracking as described in Section 5.1, solves both these problems and should always be applied before TSA.

Figure 5.26 shows the effect of order tracking on the spectrum of a signal from the gearbox of a mining shovel, with a reasonable variation in speed over the cycle. Without the order tracking, no discrete frequency components are visible in the spectrum. Figure 5.27 shows the results of using synchronous averaging on the data of Figure 5.26.

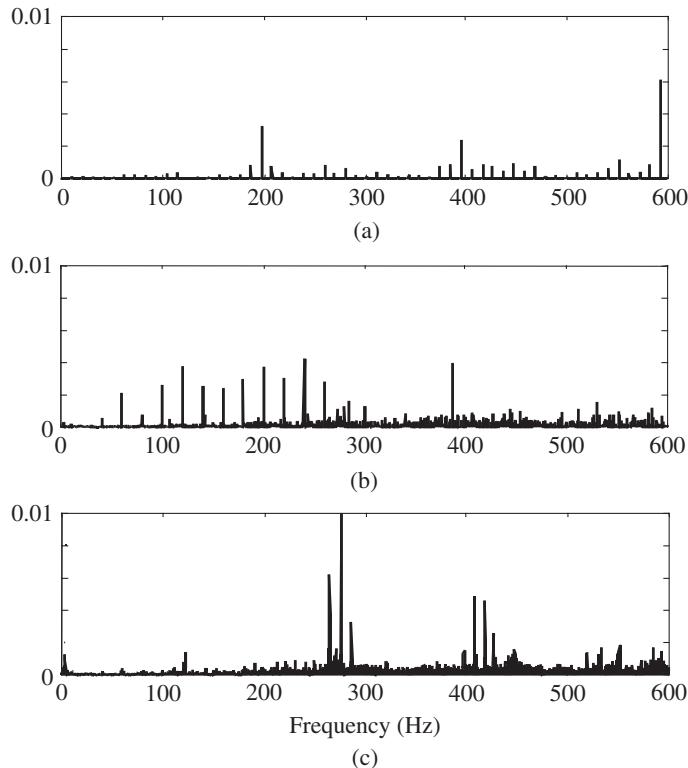


Figure 5.27 Application of synchronous averaging to data of Figure 5.26. (a) Spectrum synchronous with low speed gear (b) Spectrum synchronous with high speed gear (c) Spectrum dominated by bearing fault after effects of two gears removed.

The order tracked data was arranged to have an integer number of samples per period of the low speed gear, which allowed determination of the harmonics of this gear speed by synchronous averaging. The spectrum of this signal is shown in Figure 5.27a. After a periodic repetition of this signal was subtracted from the overall tracked signal (Figure 5.26b) the data was resampled to have an integer number of samples per period of the high speed gear, after which its harmonics could be determined in the same way (Figure 5.27b). Finally, after subtraction of this periodic signal from the data, the remaining signal was dominated by the effects of an inner race bearing fault (Figure 5.27c).

Obtaining the synchronous average signal for the individual planet gears and sun gear in a planetary gearbox is discussed in Section 7.2.1 of Chapter 7.

5.3.2 Linear Prediction

Linear prediction is basically a way of obtaining a model of the deterministic (i.e. ‘predictable’) part of a signal, based on a certain number of samples in the immediate past, and then using this model to predict the next value in the series. The residual (unpredictable) part of the signal is then obtained by the subtraction from the actual signal value.

The model used for linear prediction is an ‘autoregressive’ or AR model as described by the following equation:

$$\hat{x}(n) = - \sum_{k=1}^p a(k) \cdot x(n-k) \quad (5.17)$$

where the predicted current value $\hat{x}(n)$ is obtained as a weighted sum of the p previous values.

The actual current value is given by the sum of the predicted value and a noise term:

$$x(n) = \hat{x}(n) + e(n) \quad (5.18)$$

The $a(k)$ weighting coefficients in Eq. (5.17) can be obtained by a linear operation from the auto-correlation function $r_{xx}(n)$ of the time series $x(n)$, for which biased estimates can be obtained from:

$$\hat{r}_{xx}[k] = \frac{1}{N} \sum_{n=0}^{N-1} x[n]x[n-k], \quad 0 \leq k \leq p-1 \quad (5.19)$$

The $a(k)$ are obtained using the Yule-Walker equations, with the matrix form given in Eq. (5.20), often using the so-called Levinson-Durbin Recursion (LDR) algorithm [22]:

$$\begin{bmatrix} r_{xx}[0] & r_{xx}[-1] & \cdots & r_{xx}[-p+1] \\ r_{xx}[1] & r_{xx}[0] & \cdots & r_{xx}[-p+2] \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}[p-1] & r_{xx}[p-2] & \cdots & r_{xx}[0] \end{bmatrix} \begin{bmatrix} a[1] \\ a[2] \\ \vdots \\ a[p] \end{bmatrix} = - \begin{bmatrix} r_{xx}[1] \\ r_{xx}[2] \\ \vdots \\ r_{xx}[p] \end{bmatrix} \quad (5.20)$$

They can also be obtained using Burg’s maximum entropy method (MEM) [23].

Note that Eqs. (5.18, 5.19) can be combined and written as:

$$x(n) + \sum_{k=1}^p a(k) \cdot x(n-k) = e(n) \quad (5.21)$$

which is essentially the same as Eq. (3.53), (with $x(n) = y_i$ and $e(n) = x_i$), so that the equivalent of Eq. (3.54) is:

$$X(z)A(z) = E(z) \quad (5.22)$$

or

$$X(z) = \left(\frac{1}{A(z)} \right) E(z) \quad (5.23)$$

which can be considered (in the z-domain) as the output $X(z)$ of a system with transfer function $1/A(z)$ when excited by the forcing function $E(z)$. As shown in Eq. (3.55) the transfer function is an all-pole filter with poles p_k .

The forcing function $E(z)$ is white, containing stationary white noise and impulses, and its time domain counterpart $e(n)$ is said to be ‘prewhitened’. Prewhitening is one of the applications of linear prediction used elsewhere in this book.

Another application of AR modelling is for frequency analysis, as the transfer function $1/A(z)$ can be said to have the same amplitude spectrum as the signal being analysed when the excitation is white. Thus, the poles of the transfer function represent the discrete frequency components contained in the signal, and in general with better resolution than obtained by Fourier analysis from the same record [24] (though with poorer amplitude accuracy). Such spectral analysis is often done by the MEM mentioned above, but the advantage is not necessarily as marked as may first appear. As an example, Figure 5.28 shows the results of applying maximum entropy analysis to a very short

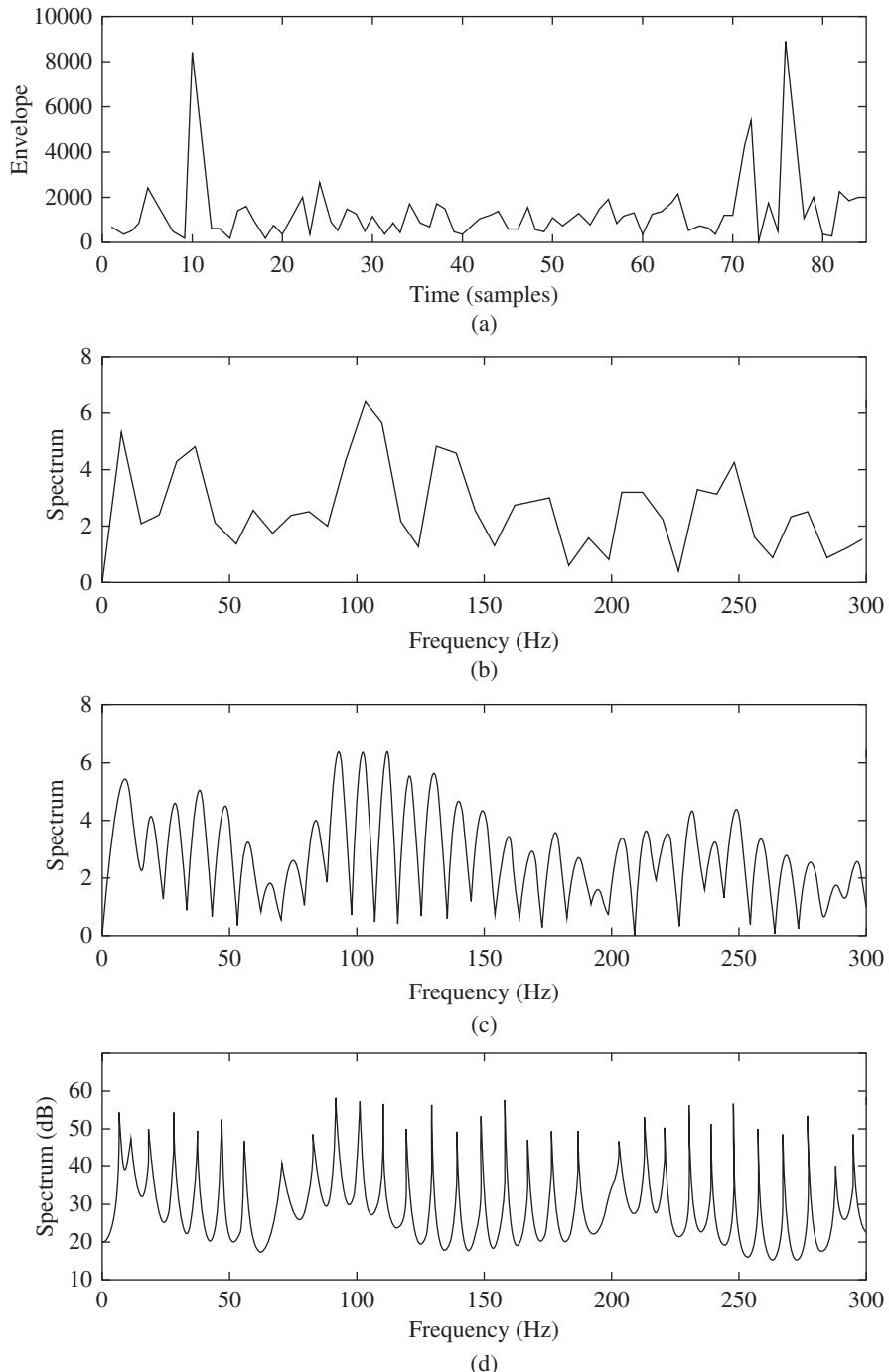


Figure 5.28 Envelope spectra of short time recording (1.29 revolutions) just spanning the first two sections where the inner race fault passes the loading zone [25]. (a) Envelope signal after bandpass filtration in frequency range $2.7 \sim 3.3$ kHz, (b) Envelope spectrum without interpolation, (c) Envelope spectrum with interpolation, (d) Maximum entropy envelope spectrum

record of envelope signal from a bearing with an inner race fault [25]. The record length comprised only 1.29 revolutions of the shaft speed which determined the spacing of modulation sidebands in the envelope spectrum. The maximum entropy spectrum of Figure 5.28d appears to give very good resolution of the sidebands, but Figure 5.28c shows that Fourier analysis can give almost as much information provided a sufficient degree of spectrum interpolation is used. The spectrum interpolation was achieved by padding the data record with zeros to seven times its original length. Note that the maximum entropy spectrum had to be represented on a logarithmic amplitude scale because of the much wider range of amplitude values than for the Fourier analysis cases.

An important consideration in AR modelling is the choice of p , the model order. When the application is to separate discrete frequency components from stationary white noise, a standard approach is to use the Akaike information criterion (AIC) [26]. However, in a number of other applications, such as where impulsive events dominate the whitened signal, other criteria may be more appropriate, and are discussed in this book in the relevant section.

5.3.3 Adaptive Noise Cancellation

Adaptive noise cancellation (ANC) is a procedure where a (primary) signal containing two uncorrelated components can be separated into those components by making use of a (reference) signal containing only one of them. The reference signal does not have to be identical to the corresponding part of the primary signal, just related to it by a linear transfer function. The ANC procedure adaptively finds that transfer function, and can thus subtract the modified reference signal from the primary signal, leaving the other component (Widrow and Stearns [27]). A typical ANC filter process, applied to the separation of gear and bearing signals, is shown in Figure 5.29. The adaptive filter adjusts its parameters to minimise the variance of the error signal ϵ . Because the two components are uncorrelated, the variance of the total signal will be the sum of the variances of the two constituents, and thus the separation will be achieved when the variance of the difference signal is minimised, meaning that it then contains no part of the reference signal.

5.3.4 Self Adaptive Noise Cancellation

When one of the two components to be separated is deterministic (discrete frequency) and the other random, the reference signal can be made a delayed version of the primary signal, because if the delay is longer than the correlation length of the random signal, the adaptive filter will not recognise

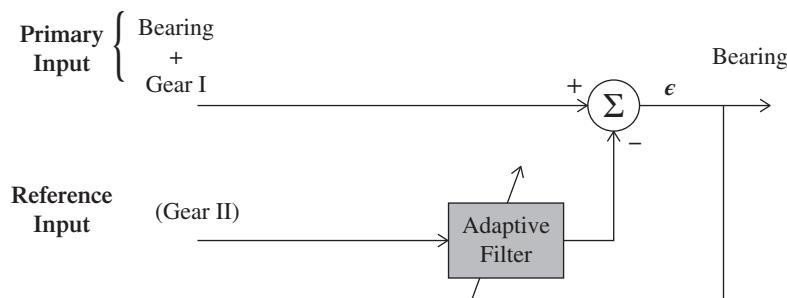


Figure 5.29 Schematic diagram of Adaptive Noise Cancellation.

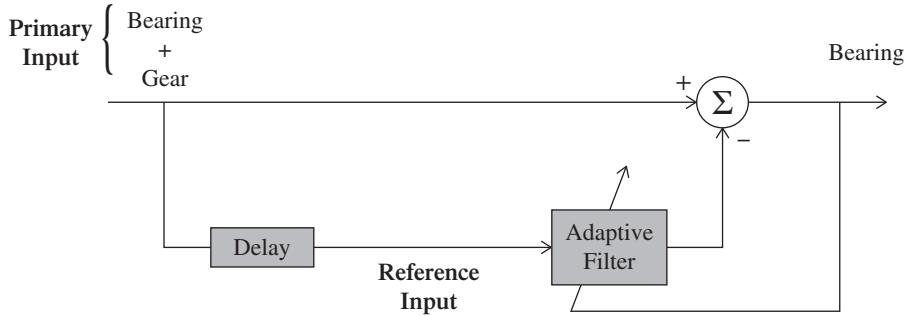


Figure 5.30 Schematic diagram of Self-Adaptive Noise Cancellation used for removing periodic interference (Widrow and Stearns [27]).

the relationship, and will find the transfer function between the deterministic part of the signal and the delayed version of itself. Thus, the separation can be achieved using one signal only, and the procedure is called self adaptive noise cancellation (SANC). This is illustrated in Figure 5.30, once again applied to the separation of gear and bearing signals, where the gear signal is deterministic.

The adaptive filter in Figure 5.30 is a recursive filter (see Section 3.4.1) with a number of weights to be determined, but which also updates at each step which means it can cope with slow changes to the signal or system properties. The recursive algorithm is the so-called least mean squares (LMS) algorithm (Widrow and Stearns [27]) and can be expressed as follows:

$$\mathbf{W}_{k+1} = \mathbf{W}_k - \mu \nabla_k \quad (5.24)$$

where gradient vector

$$\nabla_k = \frac{\partial E[\varepsilon_k^2]}{\partial \mathbf{W}_k} \quad (5.25)$$

and μ is a convergence factor, which should be chosen carefully to avoid divergence on the one hand, but not give rise to excessive adaptation time on the other.

Ho [28] shows that a conservative approximation for Eq. (5.24) is:

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \frac{2\mu_n \varepsilon_k \mathbf{X}_k}{(L+1)\hat{\sigma}_k^2} \quad (5.26)$$

where

\mathbf{W}_k	= Vector of weight coefficients of the adaptive filter at the k^{th} iteration.
μ_n	= Normalized convergence factor: $0 < \mu_n < 1$
μ	= Convergence factor: $\mu = \frac{\mu_n}{(L+1)\hat{\sigma}_k^2}$
ε_k	= Output error at k^{th} iteration
\mathbf{X}_k	= Vector of input values at k^{th} iteration
L	= Order of the adaptive filter
$(L+1)$	= Number of filter coefficients
$\hat{\sigma}_k^2$	= Exponential-averaged estimate of the input signal power at the k^{th} Iteration

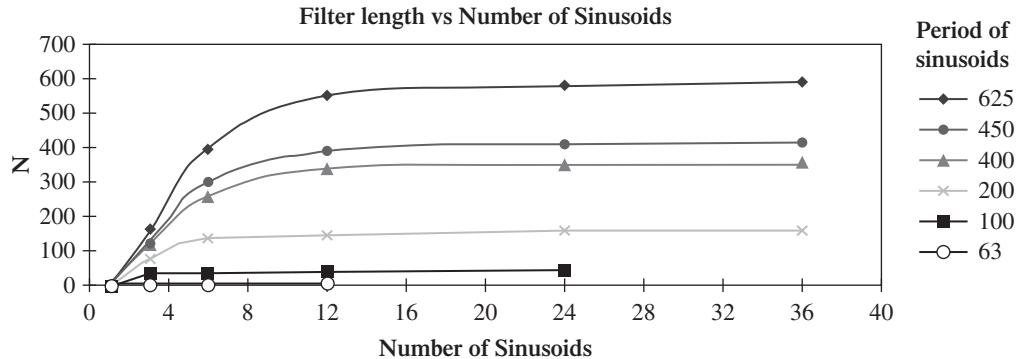


Figure 5.31 Minimum filter order vs number of discrete spectrum components.

Thus, there are three factors to be chosen for a successful result. Perhaps the most important is L , the order of the adaptive filter, which for mechanical systems, and in particular the separation of gear and bearing signals, is quite large, typically in the hundreds and even thousands.

Ho made an empirical study of the optimum choice of filter order [29], and the basic results (for a single family of equally spaced harmonics or sidebands in the band being treated) are given in Figure 5.31. Ref. [29] also contains recommendations for the situation where there is more than one family, and then the minimum frequency spacing plays a role.

In Ref. [30] Antoni likened SANC to prediction theory, and pointed out that the asymptotic result in Figure 5.31 is in agreement with the analytical result that the bandwidth of the SANC filter is the reciprocal of the period of the frequency being separated, and the filter characteristic is almost identical to that for the equivalent comb filter for synchronous averaging (e.g. Figure 5.25). In [30] Antoni also pointed out that it can be advantageous to vary the convergence factor exponentially, to start with a larger value and reduce it as the convergence proceeds.

With respect to delay, this should be chosen so as to be longer than the correlation length of the random part of the signal, but shorter than that of the deterministic part. In principle, the latter is infinite, but in practice the correlation does decrease with increasing delay, so the delay should be as short as possible. Some guidance is therefore needed to determine the correlation length of the random part. This will normally be governed by the bandwidth of (resonance) peaks in the band being processed, the narrowest corresponding to the longest correlation length, and this can sometimes be judged by inspection, or by knowledge of typical damping factors of the system. Figure 3.17 shows that for narrow band noise of bandwidth B , the correlation length (from the autocorrelation function which is the inverse Fourier transform of the autospectrum) is of the order of $1/B$. Thus, for a 3 dB bandwidth of 1% it would correspond to 100 periods of the centre frequency. The actual delay should be made perhaps three times this correlation length.

Even so, in most cases the Discrete/Random separation technique, discussed in the next section, would generally give similar results to SANC, but more efficiently and without the possibility of divergence.

5.3.5 Discrete/Random Separation (DRS)

This method [31] was proposed in a companion paper to [30] since it achieves virtually the same result, but much more efficiently because it is based in the frequency domain, and can take advantage of the speedup given by the FFT.

In contrast to SANC, it does not require adaptation as such, as the required filter to remove discrete frequency components is first determined, possibly from the whole length of data, and then applied to the same data. For this reason, it does require the discrete frequency components to be very stable, and so order tracking might be necessary as a pre-processing step.

The basic principle is to obtain the transfer function between the signal and a delayed version. The process used is the same as that used to obtain the H_1 Frequency Response Function (FRF) as used in modal analysis [32], where the cross spectrum from input to output is divided by the autospectrum of the input. Thus:

$$H_1(f) = \frac{E[G_b(f)G_a^*(f)]}{E[G_a(f)G_a^*(f)]} = \frac{G_{ab}(f)}{G_{aa}(f)} \quad (5.27)$$

where $G_a(f)$ is the spectrum of the input, $G_b(f)$ is the spectrum of the output, $G_{ab}(f)$ is the cross spectrum, and $G_{aa}(f)$ is the input autospectrum. Ideally, this would give a value of unity at the frequencies of the discrete components (where the signals are correlated) and zero at frequencies where there is noise, but the values actually depend on the signal/noise ratio (SNR).

H_1 is typically used where most noise is in the output signal, since noise is averaged out of the cross spectrum. In the current application there is an equal amount of noise in both the original and delayed signals, but H_1 is still the optimal filter. Antoni [31] shows that the amplitude of the separation filter is:

$$\frac{(\rho N/2)|W(f)|^2}{(\rho N/2)|W(f)|^2 + 1} \quad (5.28)$$

where $\rho = \text{SNR}$, N is the transform size, and $W(f)$ is the Fourier transform of the window used, scaled to a maximum value of 1 in the frequency domain. Even for a SNR as low as 10^{-2} (-20 dB), this gives a value of 0.7 for $N = 512$. This is the same as for the equivalent SANC filter. The filter characteristic is somewhat poorer than for the equivalent SANC, in terms of sidelobes for a rectangular window, or noise bandwidth if a window such as Hanning is used, but on the other hand, the DRS is so much more efficient that a longer filter (larger value of N) can be used to give better resolution.

Once the filter is determined in the frequency domain, it can be used to filter the signal using a fast convolution procedure, also using FFTs in the frequency domain, by the ‘overlap/add’ method. The latter is necessitated by the circular convolution nature of the FFT (and DFT), but even though it more than doubles the computation time, it is still vastly quicker than direct convolution in the time domain. Note that if the amplitude spectrum of the filter is used, this gives zero phase shift of the separated signal. The filter is then non-causal. Note also that the frequency domain method can be used on the one-sided spectra of analytic signals, giving a further advantage when used in combination with demodulation, such as in bearing diagnostics.

A typical example of the application of DRS is given in Figure 5.32, which shows part of the spectrum of an acceleration signal from a helicopter gearbox, containing harmonics of a number of shaft speeds. Note that the processing was done on the full bandwidth signal, with normalised frequency extending from 0 to 0.5, and the zoomed sections shown are for illustrative purposes. Figure 5.32a shows the original spectrum, and Figure 5.32b the filter characteristic derived by the DRS process. It is seen that this has a value close to 1 where there are discrete frequency components, and where the SNR is high, and close to 0 elsewhere. Where the discrete components do not protrude so much from the noise, the value is lower (e.g. about 0.5 for normalised frequency approx. 0.085). Figure 5.32c is the spectrum of the extracted deterministic part, and it is seen that the SNR has typically improved by about 20 dB, while Figure 5.32d shows the spectrum of the original noise.

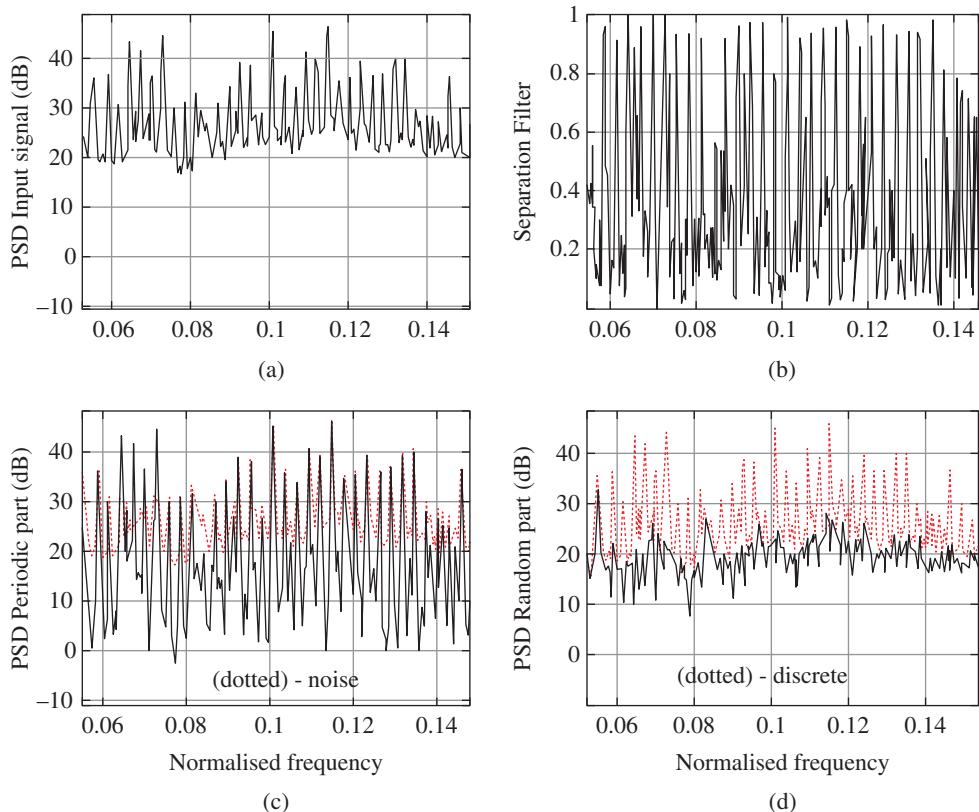


Figure 5.32 Application of DRS to a helicopter gearbox vibration signal (a) Original spectrum (zoomed) (b) Amplitude characteristic of filter (c) Spectrum of deterministic part (d) Spectrum of random part.

5.4 Minimum Entropy Deconvolution

Many vibration signals measured externally on machines are considerably distorted by the transmission paths from the source to the transducer. This is particularly the case for impulsive type signals that are typically the result of internal sharp impacts, for example from local spalls in gears and bearings. Many of the diagnostic tools described in later sections depend on being able to identify a train of response pulses arising from such impacts, and using for example envelope analysis to determine their frequency of repetition. However, this will only be possible if the impulse response functions (IRFs) are shorter than the spacings between them, and this is not always the case for high speed machines.

The ‘minimum entropy deconvolution’ (MED) method is designed to reduce the spread of IRFs, to obtain signals closer to the original impulses that gave rise to them. It was first proposed by Wiggins [33] to sharpen the reflections from different subterranean layers in seismic analysis. The basic idea is to find an inverse filter that counteracts the effect of the transmission path, by assuming that the original excitation was impulsive, and thus having high kurtosis. The name results from the fact that increasing entropy corresponds to increasing disorder, whereas impulsive signals are very structured, requiring all significant frequency components to have zero phase simultaneously at the

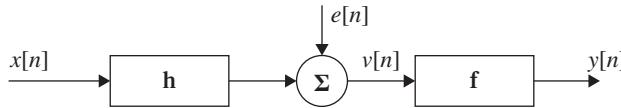


Figure 5.33 Inverse filtering (deconvolution) process for MED.

time of each impulse. Thus, minimising the entropy maximises the structure of the signal, and this corresponds to maximising the kurtosis of the inverse filter output (corresponding to the original input to the system). The method might just as well be called ‘maximum kurtosis deconvolution’ because the criterion used to optimise the coefficients of the inverse filter is maximisation of the kurtosis of the inverse filter output. The MED method was applied to gear diagnostics by Endo and Randall [34], and to bearing diagnostics by Sawalhi, Randall, and Endo [35].

Figure 5.33 illustrates the basic idea. The forcing signal $x[n]$ passes through the structural filter \mathbf{h} whose output is mixed with noise $e[n]$ to give the measured output $v[n]$. The inverse (MED) filter \mathbf{f} produces output $y[n]$, which has to be as close as possible to the original input $x[n]$. Of course the input $x[n]$ is unknown, but is assumed to be as impulsive as possible.

The filter \mathbf{f} is modelled as an FIR filter with L coefficients such that:

$$y[n] = \sum_{l=1}^L f[l]v[n-l] \quad (5.29)$$

where $f[i]$ has to invert the system IRF $h[i]$ such that:

$$f[i] * h[i] = \delta(i - l_m) \quad (5.30)$$

The delay l_m is such that the inverse filter can be causal. It will displace the whole signal by l_m but will not change pulse spacings.

The method adopted in [34] is the objective function method (OFM) given in [36], where the objective function to be maximised is the kurtosis of the output signal $y[n]$, by varying the coefficients of the filter $f[l]$. This kurtosis is given by:

$$O_k(f[l]) = \sum_{i=1}^N y^4[i] / \left[\sum_{i=1}^N y^2[i] \right]^2 \quad (5.31)$$

and the maximum is found by finding the values of $f[l]$ for which the derivative of the objective function is zero, i.e.:

$$\partial(O_k(f[l]))/\partial(f[l]) = 0 \quad (5.32)$$

Ref. [36] describes how this can be achieved iteratively, when the filter coefficients of $f[l]$ converge within a specified tolerance.

In both [34] and [35], the MED process is combined with AR linear prediction filtering (Section 5.2.2), the total process being called ARMED. The AR operation achieves considerable whitening of the spectrum, but because it uses the autocorrelation function, it has no phase information, and in contrast the MED process achieves phase alignment to maximise the impulsivity of the filtered signal. An example from [34] is given in Figure 5.34.

In this case, the AR linear prediction was used for removing the regular toothmesh signal, but leaving the once-per-rev pulse from the spall, which has now become visible in (b). Note that the

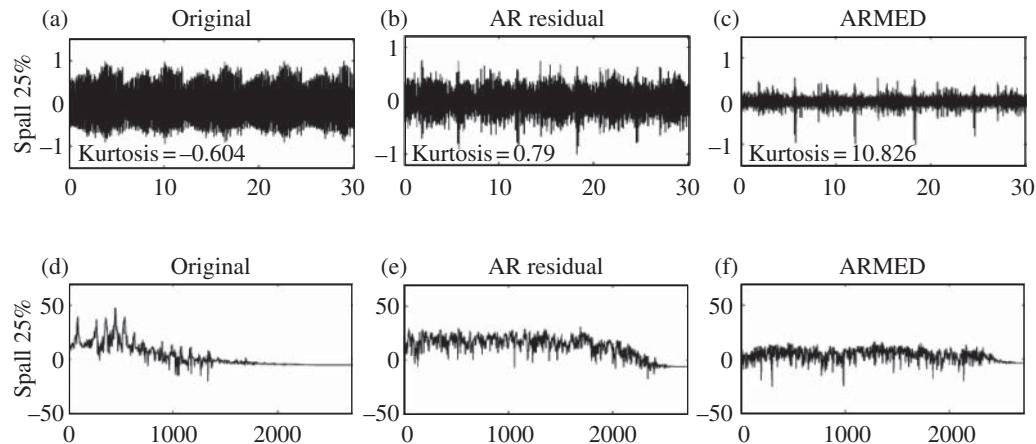


Figure 5.34 The effect of processing gear signals for the case of a spall on one tooth (Upper) – time signals (Lower) – Autospectra

MED filter has given considerably more enhancement of the fault pulses, even though the autospectra of (b) and (c) are almost equally white.

Another example, of the application to bearing diagnostics, is given in Figure 5.35, from [35]. As reported in [35], the bearing under test is a high speed bearing similar to those used in gas turbine engines, but mounted in a test rig, where a real spall was induced. In this case the AR filtering, in addition to prewhitening, was used to remove the discrete frequency components related to the harmonics of the shaft speeds of the test machine.

Once again, the AR filtering has made the pulses visible, but because of the high speed (12 000 rpm), the impulse responses excited have a length comparable to their spacing, and tend to overlap. The kurtosis has only improved from -0.40 to 1.25 in the AR operation, but increased to 38.6 after applying MED. This fault is at a fairly advanced stage, but using MED meant that it could be detected much earlier.

5.5 Spectral Kurtosis and the Kurtogram

Spectral kurtosis (SK) provides a means of determining which frequency bands contain a signal of maximum impulsivity. It was first used in the 1980s for detecting impulsive events in sonar signals [37]. It was based on the STFT (short time Fourier transform, Section 3.5.1) and gave a measure of the impulsiveness of a signal as a function of frequency. It was revived in the 1990s as a tool to be used in blind source separation (BSS), to determine whether this could be done in the frequency domain [38]. This was based on the fact that narrow band filtration tends to make a signal more Gaussian, and BSS usually relies on the components in a mixture being non-Gaussian to enable their separation. Transformation into the frequency domain converts a convolutive mixture into a simple (multiplicative) mixture, much easier to separate by the techniques of independent components analysis, and so the SK was useful to indicate whether this would be possible. Another group at the same institution (INPG Grenoble in France, now GIPSA Lab) later developed the use of the SK for other purposes [39], such as distinguishing between sinusoidal and very narrow band noise signals, with SK values of -1 and zero, respectively.

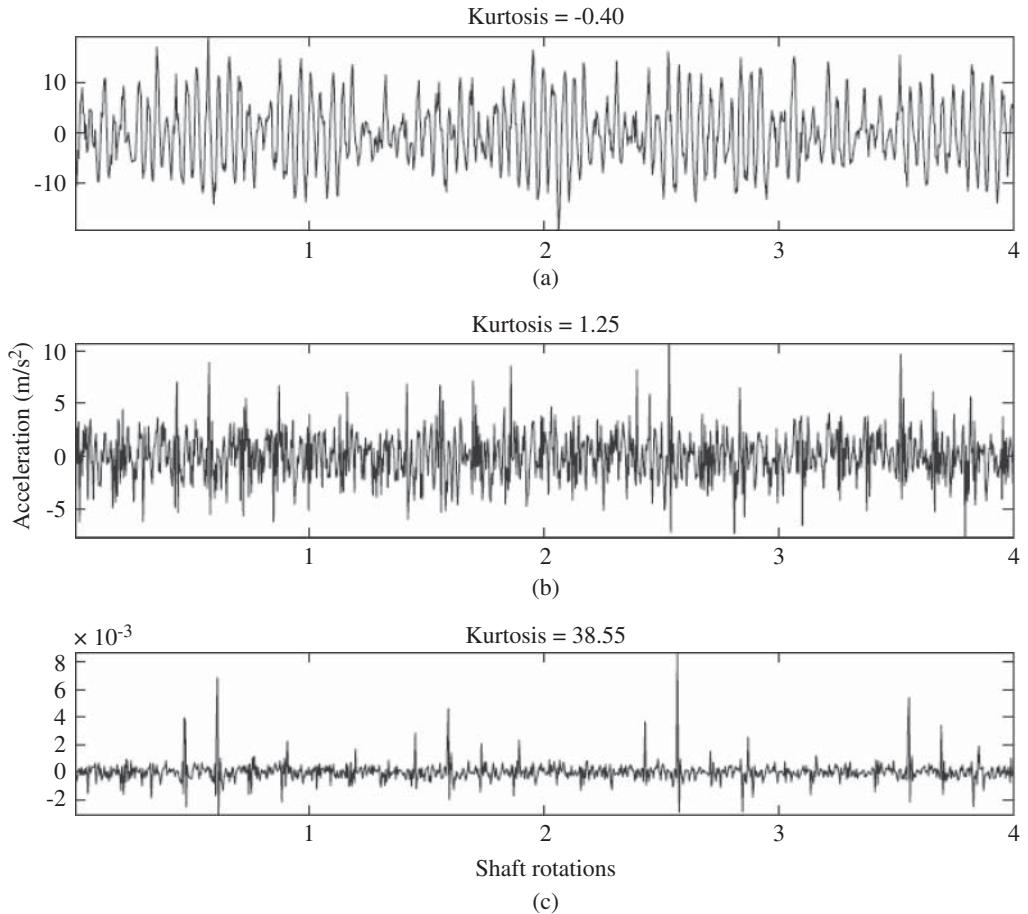


Figure 5.35 Example of applying both AR and MED filtering to bearing signals with an inner race fault in a high speed bearing (a) Original time signal (b) After application of AR filtering (c) After additional MED filtering.

The main application in this book is to the diagnostics of machine faults that give rise to a series of impulse responses. Kurtosis has long been used as a measure of the severity of machine faults, since its proposal by Stewart et al in the 1970s, (e.g. [40]) but there was only a vague suggestion that clearer results might be achieved by using filtering in frequency bands, typically octaves, and the concept of spectral kurtosis was not really developed. The application of SK to bearing faults was first outlined by Antoni [41, 42], who made a very thorough study of the definition and calculation of the SK for this purpose.

5.5.1 Spectral Kurtosis – Definition and Calculation

For the case of a series of impulse responses $g(t)$ modelling a rolling element bearing signal, excited by impulses X at times τ_k the response $Y(t)$ (as shown in Figure 5.36a) is given by:

$$Y(t) = \sum_k g(t - \tau_k)X(\tau_k) \quad (5.33)$$

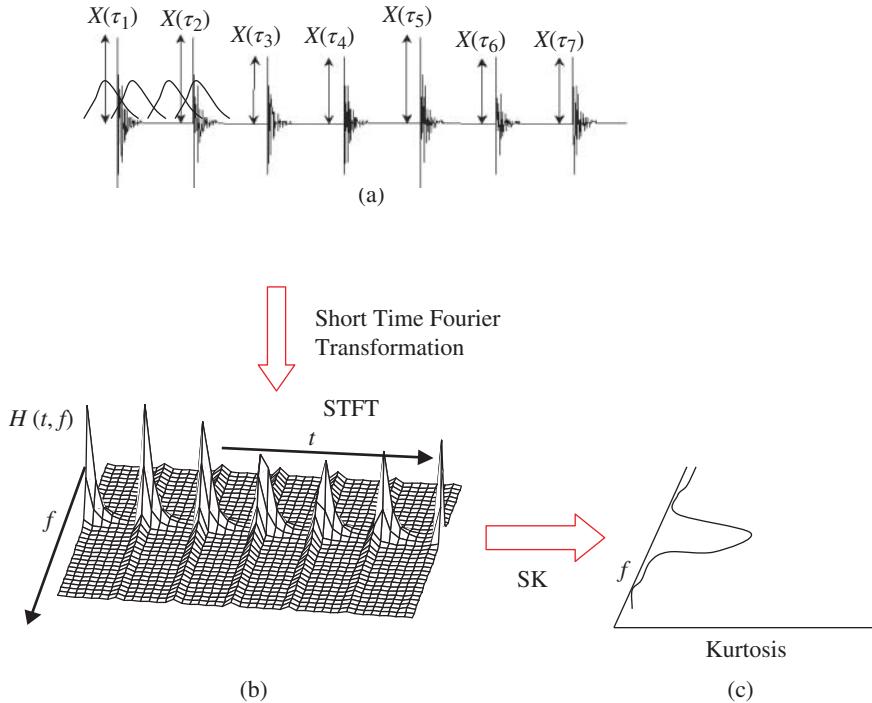


Figure 5.36 Calculation of SK from the STFT for a simulated bearing fault signal (a) simulated time signal (b) STFT (c) SK as a function of frequency.

The STFT obtained by shifting a time window along the record, is here represented in terms of the amplitude envelope function $H(t,f)$, but its square represents the power spectrum values at each time position. The average of all these short time power spectra (equivalent to the Welch method) would be the power spectrum of the whole record, or in other words at each frequency the mean square value of the output of a filter corresponding to that frequency line.

The kurtosis for each frequency f can be calculated by taking the fourth power of $H(t,f)$ at each time and averaging its value along the record, then normalising it by the square of the mean square value. It can be shown that if 2 is subtracted from this ratio, as given in Eq. (5.34), the result will be zero for a Gaussian signal [39, 41] (i.e. definition in terms of cumulants).

$$K(f) = \frac{\langle H^4(t,f) \rangle}{\langle H^2(t,f) \rangle^2} - 2 \quad (5.34)$$

The results obtained from Eq. (5.34) depend on the parameters chosen for the STFT. The denominator of Eq. (5.34) is independent of the window length chosen, but the numerator is affected by it. To obtain a maximum value of kurtosis, the window must be shorter than the spacing between the pulses, but longer than the individual pulses. The choice of optimum window length is discussed in detail below, using the application to rolling element bearing faults.

Figure 5.37 shows a comparison of the spectral kurtosis with the dB spectrum difference caused by an inner race fault in a ball bearing. Note that the shape of the SK curve is very similar to the dB difference, and thus conveys virtually the same information without requiring historical data. It is also interesting that the actual values of SK and dB difference are very similar, though this is

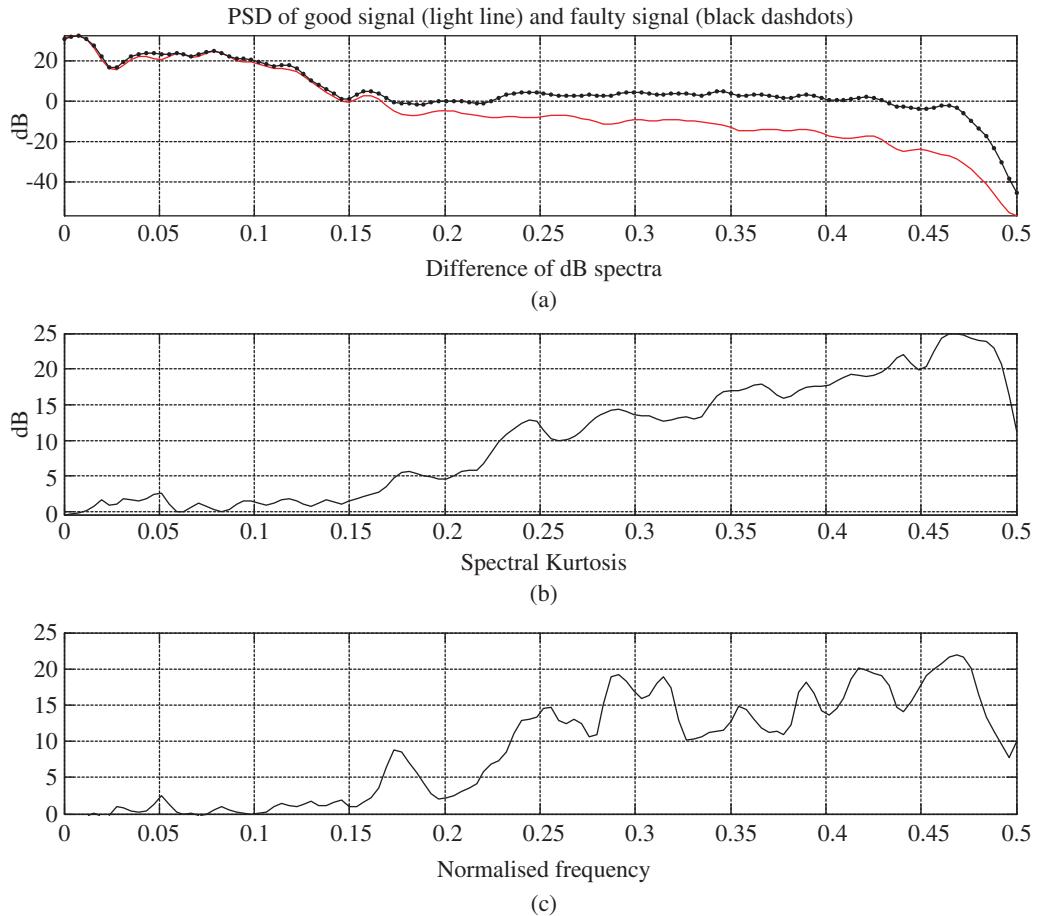


Figure 5.37 Comparison of SK with dB spectrum difference for an inner race bearing fault. Frequency scale normalised to sampling frequency 48 kHz. (a) dB spectrum comparison over frequency range 0-24 kHz with and without the fault (b) dB spectrum difference (c) Spectral kurtosis.

somewhat fortuitous because the scaling of the SK is dependent on the choice of window parameters, as stated above.

5.5.2 Use of SK as a Filter

Since the SK is large in frequency bands where the impulsive bearing fault signal is dominant, and effectively zero where the spectrum is dominated by stationary components (see Figure 5.37) it makes sense to use it as a filter function to filter out that part of the signal with the highest level of impulsiveness. For the hypothetical case of a series of impulses mixed with stationary noise, Antoni in [42] shows that the optimum Wiener filter (maximising the similarity between the filtered component and the true noise-free signal) is the square root of the SK. He also shows that the optimum matched filter (maximising the SNR of the filtered signal, without regard to its shape) is a narrow band filter at the maximum value of SK. Figure 5.38 shows the result of applying the optimum Wiener and matched filters to a signal for a weak outer race bearing fault, masked by gear noise.

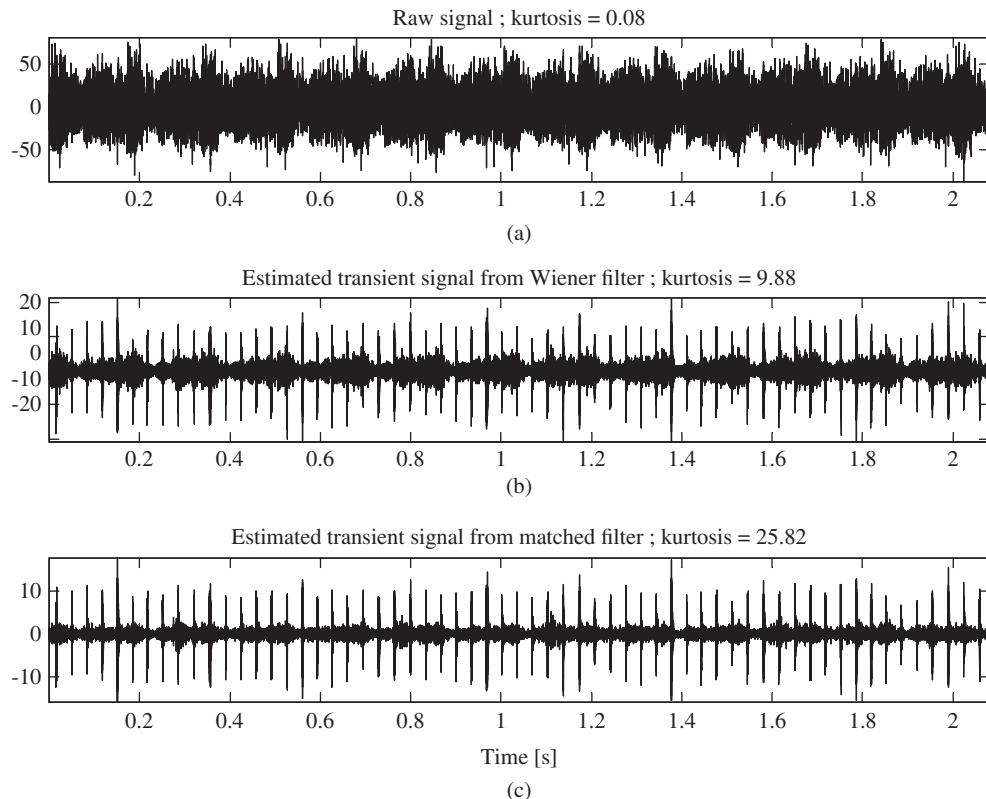


Figure 5.38 Use of SK as a filter on a signal from a gear test rig with a weak outer race bearing fault. (a) Total signal (b) Output from Wiener filter (c) Output from matched filter. Source: from [42].

In this case the matched filter produces the best result, but the result of envelope analysis in either case clearly reveals the outer race fault frequency visible in the filtered time signal.

However, the optimum result in a given case may vary with both the centre frequency and bandwidth of the filter, and in [42], a display showing the optimum combination was called the ‘kurtogram’.

5.5.3 The Kurtogram

Figure 5.39 shows the kurtogram for the outer race fault of Figure 5.38. Figure 5.40a shows the optimum bandpass filter derived from it, compared against the SK for all frequencies and finally Figure 5.40b shows the filtered time signal resulting from this optimum filter. It has a kurtosis of 46.7, compared with 9.9 and 25.8, respectively, for the Wiener and matched filters of Figure 5.38.

5.5.3.1 The Fast Kurtogram

Computation of the full kurtogram covering all combinations of centre frequency and bandwidth is very costly, and so a number of more efficient alternatives have been proposed. The most detailed study is presented by Antoni in [43], who proposes the ‘fast kurtogram’, based on a series of digital

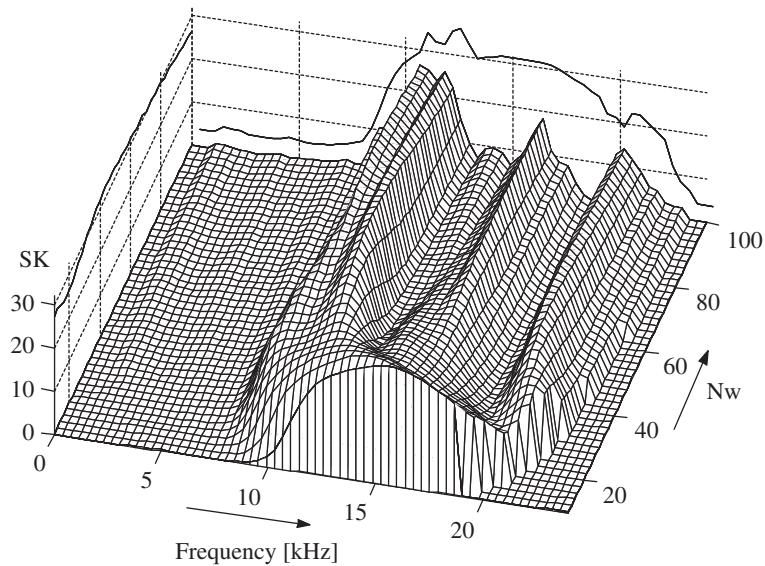


Figure 5.39 Kurtogram for signal of Figure 5.38. Source: from [42]. N_w is window length defining spectral resolution. SK is spectral kurtosis. Maxima are projected onto each plane.

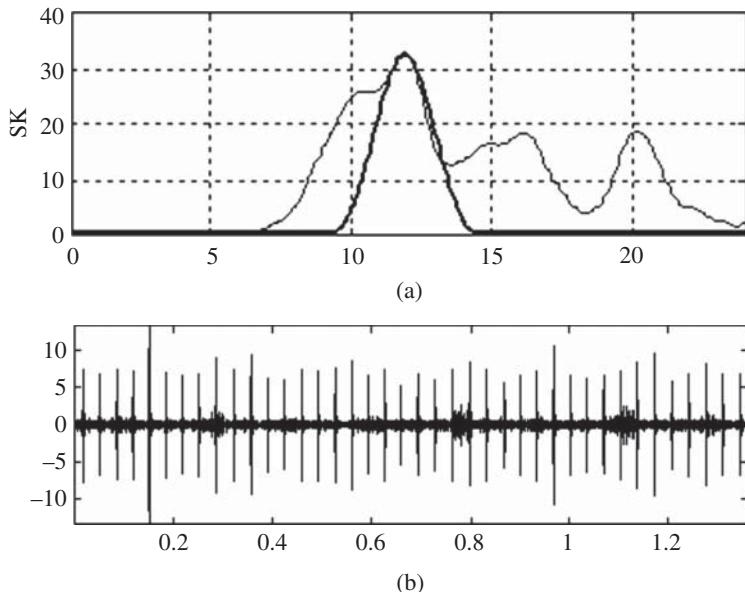


Figure 5.40 (a) Optimal bandpass filter compared with SK (b) Outer race fault signal obtained using the filter of (a). Source: from [42].

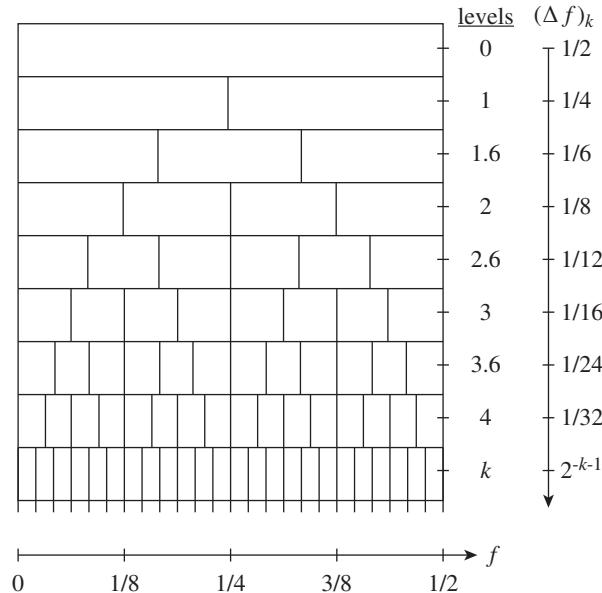


Figure 5.41 Combinations of centre frequency and bandwidth for the 1/3-binary tree kurtogram estimator [43].

filters rather than the STFT. The gains in computational speed are based on a dyadic decomposition, rather similar in principle to the FFT, and even more similar to the ‘discrete wavelet packet transform’ (DWPT). In the most basic version, the frequency range is progressively split into bands that are one half the width of the previous stage (or scale), a so-called binary tree. The version recommended, however, is the ‘1/3-binary tree’, where the split includes divisions into 1/3, so that the overall division is in the sequence 1/2, 1/3, 1/4, 1/6, 1/8, 1/12, etc. The resulting combinations of centre frequency and bandwidth are shown in Figure 5.41. Figure 5.42 compares the fast kurtogram of a signal from loose parts monitoring in a nuclear plant with the full kurtogram equivalent. It is seen that virtually the same choice of centre frequency and bandwidth would be made in each case. However, it is evident that for coarse frequency resolution, there is a limitation on choice of centre frequency which may not always give the optimum result. Ref. [44] shows that an initial result obtained from the fast kurtogram can sometimes be improved using a genetic algorithm to allow a variation of parameters around the constricted values given by the fast kurtogram.

Antoni points out that the discrete wavelet transform occupies fewer combinations than the fast kurtogram, being limited to constant percentage bandwidth, and that the DWPT gives a poorer frequency characteristic of some filters, as well as being limited to the binary tree.

Nonetheless, since the kurtogram is used to detect series of impulse responses, such as from bearing faults, and these tend to have an approximately constant damping ratio, which manifests itself in the frequency domain as a constant percentage bandwidth structure, it can also be argued that some of the combinations given by the 1/3-binary tree are unlikely, and that a 1/nth-octave wavelet analysis is adequate for seeking filter bandwidth/centre frequency combinations. For this reason, Sawalhi and Randall [45] have proposed a ‘wavelet kurtogram’ based on non-orthogonal complex Morlet wavelets. These can have any desired bandwidth, but the sequence (in terms of octaves) 1/1, 1/2, 1/3, 1/4, 1/6, 1/8, 1/12 ... is often used. A number of examples are given in Section 7.3 on rolling element bearing diagnostics.

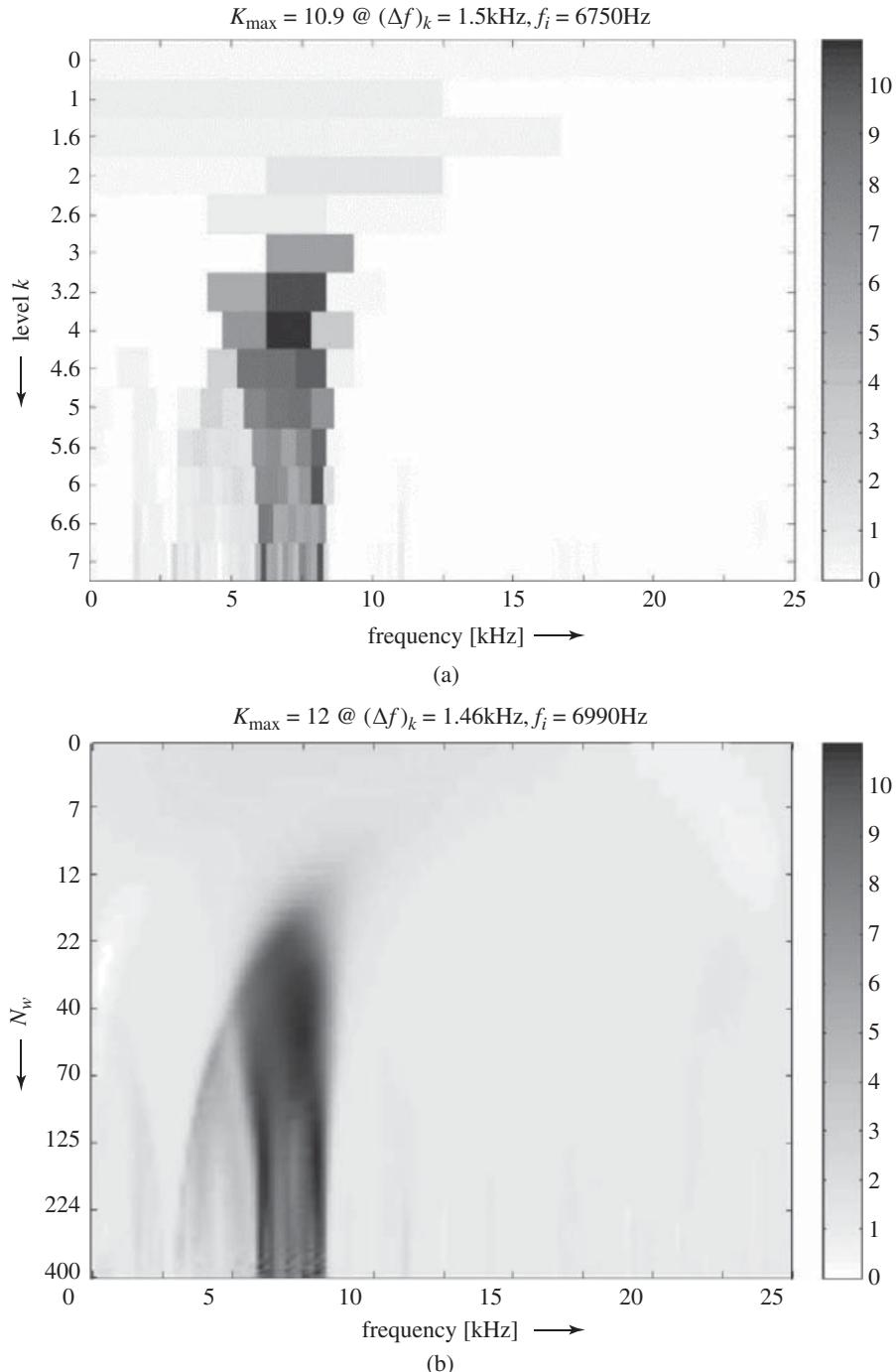


Figure 5.42 Comparison of (a) the fast kurtogram with (b) the full kurtogram for an impulsive signal from loose parts monitoring [43].

References

1. McFadden, P.D. (1989). Interpolation Techniques for Time Domain Averaging of Gear Vibration. *Mechanical Systems and Signal Processing* 3 (1): 87–97.
2. Potter, R., and Gribler, M. (1989). ‘Computed order tracking obsoletes older methods’, SAE Paper 891131.
3. Fyfe, K.R. and Munck, E.D.S. (1997). Analysis of computed order tracking. *Mechanical Systems and Signal Processing* 11 (2): 187–205.
4. Coats, M.D. and Randall, R.B. (2014). Single and multi-stage phase demodulation based order-tracking. *Mechanical Systems and Signal Processing* 44 (1-2): 86–117.
5. Bonnardot, F., El Badaoui, M., Randall, R.B. et al. (2005). Use of the acceleration signal of a gearbox in order to perform angular resampling (with limited speed fluctuation). *Mechanical Systems and Signal Processing* 19: 766–785.
6. Borghezani, P., Pennacchi, P., Randall, R.B., and Ricci, R. (2012). Order tracking for discrete-random separation in variable speed conditions. *Mechanical Systems and Signal Processing* 30: 1–22.
7. Urbanej, J., Barszcz, T., and Antoni, J. (2013). A two-step procedure for estimation of instantaneous rotational speed with large fluctuations. *Mechanical Systems and Signal Processing* 38: 96–102.
8. Coats, M.D. and Randall, R.B. (2014). Order tracking under run-up and run-down conditions. In: *IUTAM Rotor Dynamics conference*, Milan, September. Springer.
9. Feldman, M. (2011). Hilbert transform in vibration analysis. *Mechanical Systems and Signal Processing* 25: 735–802.
10. Kaiser, J.F. (1990). On a simple algorithm to calculate the ‘energy’ of a signal. In: *Int. Conf. on Acoustics, Speech, and Signal Process.*, Albuquerque, New Mexico, April, IEEE-ICASSP-90, 381–384.
11. Maragos, P., Kaiser, J.F., and Quatieri, T.F. (1993). On amplitude and frequency demodulation using energy operators. *IEEE Transactions on Signal Processing* 41 (4): 1532–1550.
12. Maragos, P., Kaiser, J.F., and Quatieri, T.F. (1993). Energy separation in signal modulations with application to speech analysis. *IEEE Transactions on Signal Processing* 41 (10): 3024–3051.
13. Randall, R.B. (2016). A new interpretation of the Teager Kaiser energy operator. In: *Vibrations in Rotating Machinery*. IMechE, Manchester, September 2016.
14. O'Toole, J.M., Temko, A., and Stevenson, N. (2014). Assessing instantaneous energy in the EEG: a non-negative, frequency-weighted energy operator. In: *Proc. 36th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, 3288–3291. Chicago IL, 26–30 August, doi: <https://doi.org/10.1109/EMBC.2014.6944325>: IEEE.
15. Randall, R.B. and Smith, W.A. (2019). Uses and mis-uses of energy operators for machine diagnostics. *Mechanical Systems and Signal Processing*, 133 Ref: 106199.
16. Randall, R.B. and Smith, W.A., (2016). ‘Use of the Teager Kaiser Energy Operator to estimate machine speed,’ PHM Europe conference, Bilbao (July 2016).
17. Randall, R.B. (2017). A history of cepstrum analysis and its application to mechanical problems. *Mechanical Systems and Signal Processing* 97: 13–18.
18. Leclère, Q., André, H., and Antoni, J. (2016). A multi-order probabilistic approach for Instantaneous Angular Speed tracking: debriefing of the CMMNO’14 diagnosis contest. *Mechanical Systems and Signal Processing* 81: 375–386.
19. Randall, R.B., Sawalhi, N., and Coats, M.D. (2011). A comparison of methods for separation of deterministic and random signals. *International Journal of Condition Monitoring* 1 (1): 11–19.
20. Braun, S. (1975). The extraction of periodic waveforms by time domain averaging. *Acustica* 32 (2): 69–77.
21. McFadden, P.D. (1987). A revised model for the extraction of periodic waveforms by time domain averaging. *Mechanical Systems and Signal Processing* 1 (1): 83–95.
22. Kay, M.S. and Marple, S.L. (1981). Spectrum analysis – a modern perspective. *Proceedings of the IEEE* 69 (11): 1380–1419.
23. Burg, J.P., (1975). Maximum entropy spectral analysis, PhD Dissertation, Stanford University, Stanford, CA, USA.
24. Braun, S. and Hammond, J.K. (1986). Parametric methods. In: *Mechanical Signature Analysis* (ed. S. Braun). London: Academic Press.
25. Randall, R.B. (2002). ‘State of the art in monitoring rotating machinery.’ *Tutorial, ISMA 2002*, PMA Dept., KUL, Leuven, Belgium.
26. Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* 21: 243–247.
27. Widrow, B. and Stearns, S. (1985). *Adaptive Signal Processing*, 349–351. Englewood Cliffs NJ: Prentice-Hall.
28. Ho, D. (2000). ‘Bearing diagnostics and self adaptive noise cancellation.’ PhD Dissertation, UNSW.
29. Ho, D. and Randall, R.B. (1997). ‘Effects of time delay, order of fir filter and convergence factor on self adaptive noise cancellation,’ *ICSV5*, Adelaide.
30. Antoni, J. and Randall, R.B. (2004). Unsupervised noise cancellation for vibration signals: Part I – evaluation of adaptive algorithms. *Mechanical Systems and Signal Processing* 18: 89–101.

31. Antoni, J. and Randall, R.B. (2004). Unsupervised noise cancellation for vibration signals: Part II – a novel frequency-domain algorithm. *Mechanical Systems and Signal Processing* 18: 103–117.
32. Mitchell, L.D. (1982). Improved methods for the FFT calculation of the frequency response function. *Journal of Mechanical Design* 104: 277–279.
33. Wiggins, R.A. (1978). Minimum entropy deconvolution. *Geoexploration* 16, Elsevier Scientific Publishing, Amsterdam: 21–35.
34. Endo, H. and Randall, R.B. (2007). Enhancement of autoregressive model based gear tooth fault detection technique by the use of minimum entropy deconvolution filter. *Mechanical Systems and Signal Processing* 21 (2): 906–919.
35. Sawalhi, N., Randall, R.B., and Endo, H. (2007). The enhancement of fault detection and diagnosis in rolling element bearings using minimum entropy deconvolution combined with spectral kurtosis. *Mechanical Systems and Signal Processing* 21 (6): 2616–2633.
36. Lee, J.Y. and Nandi, A.K. (1999). Extraction of impacting signals using blind deconvolution. *Journal of Sound and Vibration* 232 (5): 945–962.
37. Dwyer, R.F. (1983). Detection of non-Gaussian signals by frequency domain kurtosis estimation. In: *Int. Conf. On Acoustics, Speech, and Signal Processing*, Boston, 607–610. IEEE_ICASSP.
38. Capdevielle, V., Servière, C., and Lacoume, J. (1996). Blind separation of wide-band sources: application to rotating machine signals. In: Proc. of the 8th European Signal Processing Conf., vol. 3, 2085–2088.
39. Vrable, V.D., Granjon, P. and Servière, C. (2003) ‘Spectral kurtosis: from definition to application’, *IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, Grado, Italy (8–11 June 2003).
40. Dyer, D. and Stewart, R.M. (1977). Detection of rolling element bearing damage by statistical vibration analysis’ *ASME Transactions-. Journal of Mechanical Design* 100 (2): 229–235.
41. Antoni, J. (2006). The spectral kurtosis: a useful tool for characterising nonstationary signals. *Mechanical Systems and Signal Processing* 20 (2): 282–307.
42. Antoni, J. and Randall, R.B. (2006). The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines. *Mechanical Systems and Signal Processing* 20 (2): 308–331.
43. Antoni, J. (2006). Fast computation of the kurtogram for the detection of transient faults. *Mechanical Systems and Signal Processing* 21 (1): 108–124.
44. Zhang, Y. and Randall, R.B. (2009). Rolling element bearing fault diagnosis based on the combination of genetic algorithms and fast kurtogram. *Mechanical Systems and Signal Processing* 23: 1509–1517.
45. Sawalhi, N. and Randall, R. B., (2005) ‘Spectral kurtosis optimization for rolling element bearings,’ *ISSPA Conference*, Sydney, Australia (August 2005).

6

Cepstrum Analysis Applied to Machine Diagnostics

6.1 Cepstrum Terminology and Definitions

6.1.1 Brief History of the Cepstrum and Terminology

The cepstrum was originally proposed by Bogert, Healy, and Tukey [1] as a better alternative than the autocorrelation function for detecting echo delay times, specifically for seismic signals. At that time, it was defined as the power spectrum of the logarithm of the power spectrum. The ‘power cepstrum’ was later redefined [2] as the inverse Fourier transform of the log power spectrum, partly because it is more logical to use the inverse transform between a function of frequency and a function of time, and partly because it is then reversible to the power spectrum (e.g. after editing). The question arises as to why the cepstrum was first defined in that way, but the apparent answer is that Ref. [1] was published two years before the FFT, despite having a common author (Tukey), and software was readily available for power spectra, but not complex Fourier transforms, even just two years before the publication of the FFT.

Also, not long after the publication of the FFT, the ‘complex cepstrum’ was defined by Oppenheim and Schafer as the inverse Fourier transform of the complex logarithm of the complex spectrum [2], this being reversible to a time function, and for example permitting removal of echoes from a time signal. Because the cepstrum involves a Fourier transform of a spectrum it is sometimes called a ‘spectrum of a spectrum’ and this is in fact the reason for the name ‘cepstrum’ and a number of related terms coined in the original paper by Bogert, Healy, and Tukey, and discussed later in this section. However, the autocorrelation function is the inverse Fourier transform of the corresponding autospectrum (see Section 3.2.6.5) and so is equally a spectrum of a spectrum. What really distinguishes the cepstrum is the logarithmic conversion of the spectrum before the second transform. In response spectra, this converts the multiplicative relationship between the forcing function and transfer function (from force to response) into an additive one which remains in the cepstrum. This gives rise to one of the major applications of the cepstrum. For SIMO (single input, multiple output) systems, the addition in the cepstrum corresponds to a convolution in the time domain of the forcing function and impulse response function. Note that this does not apply to MIMO (multiple input, multiple output) systems, as each response is then a sum of convolutions.

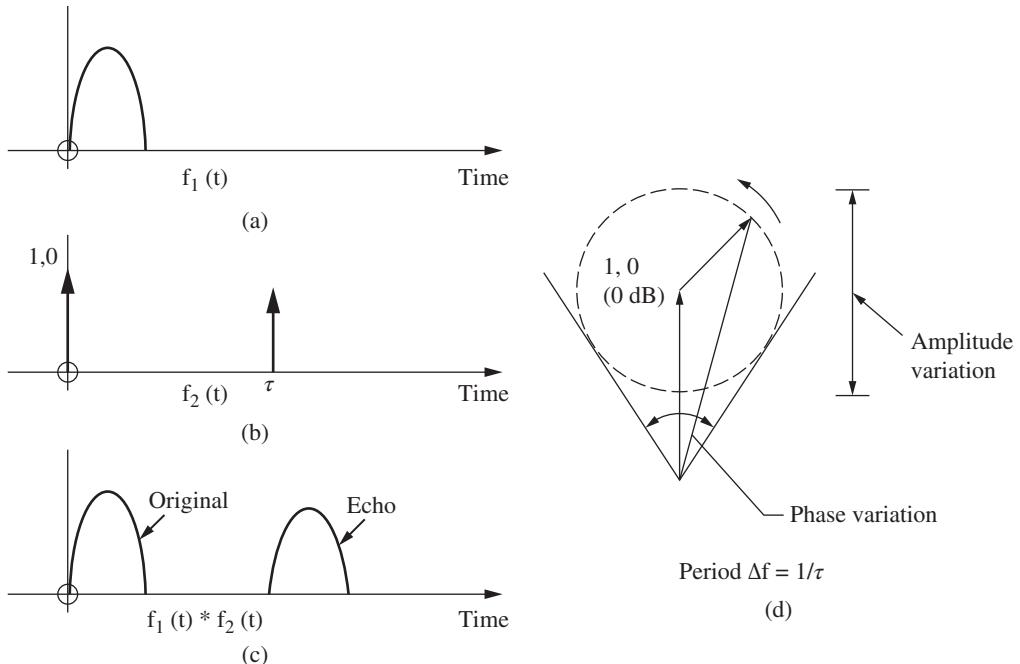


Figure 6.1 Showing that an echo gives a periodic component in the signal spectrum.

Another situation where convolution in the time domain is converted into an addition is the case of echoes. As illustrated in Figure 6.1 (see also Figure 3.11), a signal with an echo can be modelled as the convolution of the original signal (called $f_1(t)$) with a signal (called $f_2(t)$) consisting of a unit delta function at the origin and a scaled delta function at the delay time τ .

$$f_2(t) = \delta(t) + a\delta(t - \tau) \quad (6.1)$$

with spectrum

$$F_2(f) = 1 + a \exp(-j2\pi f \tau) \quad (6.2)$$

As shown in Figure 6.1(d) [3], the latter is the sum of a stationary vector and a smaller rotating one executing one period for every advance of $1/\tau$ along the frequency axis. Thus, both its amplitude (and also log amplitude) and phase are periodic in frequency with this period. After taking logs, the log spectrum of the total signal is the sum of the original spectrum and an additive periodic component with frequency period $\Delta f = 1/\tau$. The inverse Fourier transform gives the cepstrum of the original function $f_1(t)$ plus a series of discrete components at multiples of τ corresponding to the Fourier series components of the periodic component. This has application in detecting echoes and measuring their delay time, the original application of the power cepstrum, but they can also be removed using the complex cepstrum.

This is illustrated in Figure 6.2 [3]. It shows that because the cepstrum is shorter than the impulse response, it is simple to remove echoes that are overlapping. This is a numerically generated case, but echoes can be removed in real signals as well [3]. Using the real cepstrum, the effects of echoes can be removed from the log spectrum, by similar editing of the cepstrum. Note that in the complex

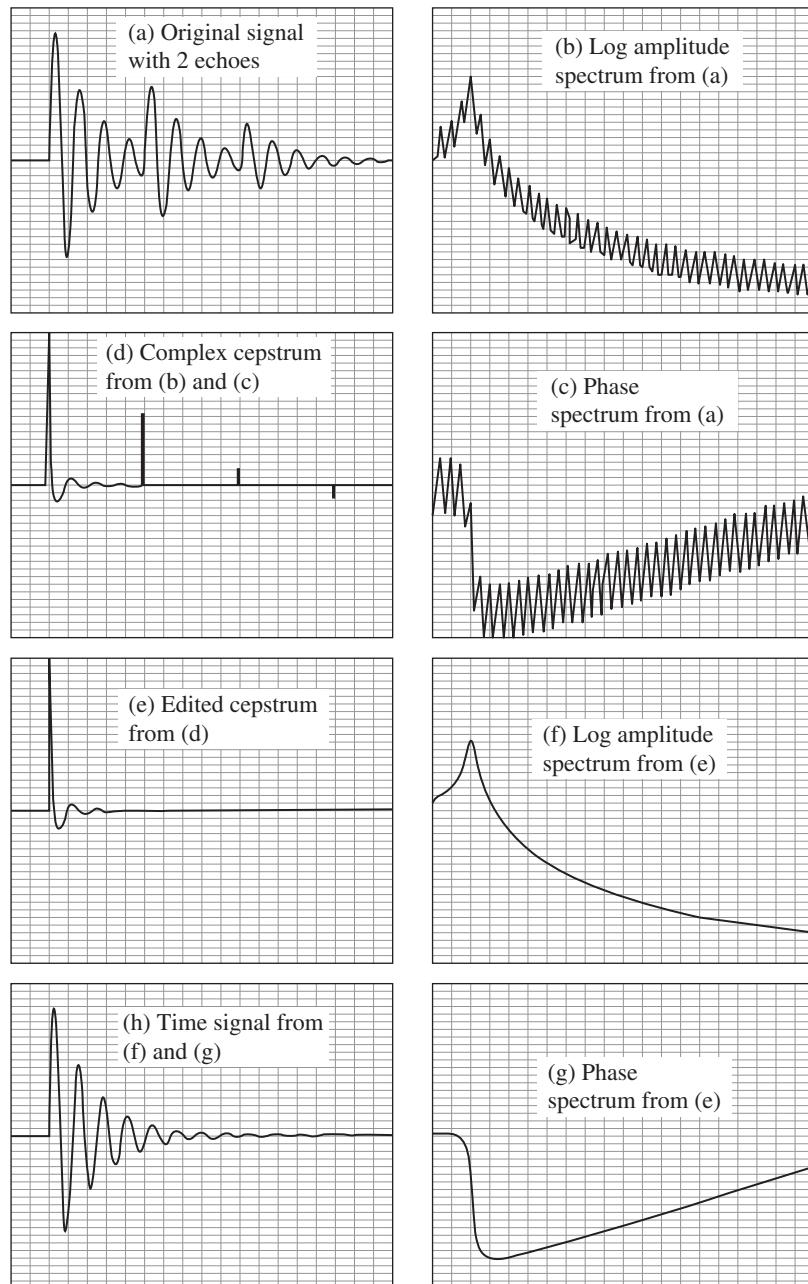


Figure 6.2 Echo removal using the complex cepstrum [3].

cepstrum, the scaling of the log amplitude spectrum should be in nepers (natural log of the amplitude ratio) to correspond with radians in the phase spectrum.

Oppenheim and Schafer in [4] give an interesting account of the early history of the cepstrum, with applications largely in echo detection and deconvolution in geophysics, seismology, telecommunications and speech analysis. The latter application utilised two separate properties of the cepstrum, measuring the spacing of harmonics in the log spectrum, to determine voice pitch, and to deconvolve the excitation (voiced or unvoiced) from the transfer function, whose resonances define the formants of phonemes. These two application areas apply equally to machine diagnostics, and the purpose of this chapter is to discuss those applications, of which the first, on gear diagnostics, was published in 1975 [5]. Reference [6] is a historical survey of this development, so only a summary is given here.

As discussed in Section 6.2, perhaps the main application of cepstrum analysis in machine condition monitoring is for signals containing families of harmonics and sidebands (of uniform spacing) where it is the whole family, rather than individual frequency components, that characterises the fault. Examples are given by localised faults in gears (see Figure 2.13 of Chapter 2) giving rise to harmonics and/or sidebands throughout the spectrum, and much more evident in the log amplitude rather than the linear amplitude spectrum (vastly reducing the effect of the somewhat arbitrary structural transfer functions between the source and measurement points). This is taken up in more detail in Section 7.2 of Chapter 7. The cepstrum can also be used for the harmonics generated by bearing faults, but only if they are well separated. This was mentioned in Section 2.2.3 of Chapter 2, where it is pointed out that it is often better to use envelope analysis, as it does not have that restriction.

6.1.1.1 Terminology

As mentioned above, in the original paper [1], the authors coined the word ‘cepstrum’ by reversing the first syllable of ‘spectrum’, the justification being that it was a ‘spectrum of a spectrum’. Similarly, the word ‘quefrency’ was obtained from ‘frequency’, and the authors also suggested a number of others, including:

rahmonic	from	harmonic
lifter	from	filter
short-pass lifter	from	low-pass filter
long-pass lifter	from	high-pass filter
gamplitude	from	magnitude
saphe	from	phase
quefrency alanysis	from	frequency analysis

Of these, quefrency, rahmonic, and lifter are useful in clarifying that the operations or features refer to the cepstrum, rather than the spectrum or time signal, and are still regularly used in the literature as well as in this book. The usefulness of the other terms is somewhat more dubious.

6.1.2 Cepstrum Types and Definitions

The original definition of the (power) cepstrum was:

$$C_p(\tau) = |\Im\{\log(F_{xx}(f))\}|^2 \quad (6.3)$$

where $F_{xx}(f)$ is a power spectrum, which can be an averaged power spectrum or the amplitude squared spectrum of a single record.

The complex cepstrum is defined as:

$$C(\tau) = \mathfrak{F}^{-1}[\log(X(f))] \quad (6.4)$$

where

$$X(f) = \mathfrak{F}[x(t)] = A(f) \exp(j\phi(f)) \quad (6.5)$$

in terms of its amplitude and phase

$$\text{so that} \quad \log(X(f)) = \ln(A(f)) + j\phi(f) \quad (6.6)$$

Despite its name, since $\ln(A(f))$ is even and $\phi(f)$ is odd, the complex cepstrum is real-valued.

The new power cepstrum is given by:

$$C_p(\tau) = \mathfrak{F}^{-1}\{\log(F_{xx}(f))\} \quad (6.7)$$

which for the spectrum of a single record (as in [3]) can be expressed as:

$$C_p(\tau) = \mathfrak{F}^{-1}\{2\ln(A(f))\} \quad (6.8)$$

Note that by comparison, the autocorrelation function can be derived as the inverse transform of the power spectrum, or:

$$R_{xx}(\tau) = \mathfrak{F}^{-1}[|X(f)|^2] = \mathfrak{F}^{-1}[A^2(f)] \quad (6.9)$$

The so-called real cepstrum is obtained by setting the phase to zero in Eq. (6.6):

$$C_r(\tau) = \mathfrak{F}^{-1}\{\ln(A(f))\} \quad (6.10)$$

which is seen to be simply a scaled version of (6.8).

Note that before calculating the complex cepstrum the phase function $\phi(f)$ must be unwrapped to a continuous function of frequency, and this is often difficult, so that it is easier to use the real cepstrum.

Another type of cepstrum which is useful in some cases is the ‘differential cepstrum’, which is defined as the inverse transform of the *derivative* of the logarithm of the spectrum [7]. It is most easily defined in terms of the Z-transform (which can replace the Fourier transform for sampled functions) as:

$$C_d(n) = Z^{-1} \left\{ z \frac{d/dz(H(z))}{H(z)} \right\} \quad (6.11)$$

where n is the quefrency index, and it can be directly calculated from a time signal as:

$$C_d(n) = Z^{-1} \left[\frac{Z\{n x(n)\}}{Z\{x(n)\}} \right] \quad (6.12)$$

Among other things this has the advantage that the phase of the (log) spectrum does not have to be unwrapped.

Where the frequency spectrum $X(f)$ in Eq. (6.4) is a frequency response function (FRF) which can be represented in the Z-plane by a gain factor B and the zeros and poles inside the unit circle, a_i and c_i , respectively, and the zeros and poles outside the unit circle, $1/b_i$ and $1/d_i$, respectively,

(where $|a_i|, |b_i|, |c_i|, |d_i| < 1$) then it has been shown by Oppenheim and Schafer [8] that the complex cepstrum is given by the analytical formulae:

$$\begin{aligned} C_h(n) &= \ln(B) & , \quad n = 0 \\ C_h(n) &= -\sum_i \frac{a_i^n}{n} + \sum_i \frac{c_i^n}{n} & , \quad n > 0 \\ C_h(n) &= \sum_i \frac{b_i^{-n}}{n} - \sum_i \frac{d_i^{-n}}{n} & , \quad n < 0 \end{aligned} \quad (6.13)$$

in terms of quefrency index n .

Since the cepstrum is real, the complex exponential terms in (6.13) can be grouped in complex conjugate pairs so that a typical pair of c_i terms, for example, can be replaced by $\frac{2}{n} C_i^n \cos(n\alpha_i)$ where $C_i = |c_i|$ and $\alpha_i = \angle c_i$. This represents an exponentially damped sinusoid, further damped by the hyperbolic function $2/n$. Figure 6.3 compares the cepstrum with the impulse response function for a SDOF (single degree of freedom) system which has one pair of poles and no zeros. On a logarithmic amplitude scale, zeros of the FRF (antiresonances) are like inverted poles (resonances) so it is no surprise that the corresponding terms in the cepstrum have inverted sign.

Taking the derivative of the log spectrum in the Z-domain to obtain the differential cepstrum results in multiplication by n in the cepstrum, so that a typical term becomes $2A_i^n \cos(n\alpha_i)$, an exponentially damped sinusoid without the hyperbolic weighting, so that its form is similar to that of the impulse response function. This is useful in that techniques which have been developed to curve fit parameters to the impulse response function can be directly applied to the differential cepstrum. This is the second advantage of the differential cepstrum, but mainly in cases where it is being used for operational modal analysis [9]. The so-called ‘mean differential cepstrum’ [10] has additional advantages in this application.

Note that for functions with minimum phase properties, which applies to FRFs for many physical structures, there are no poles or zeros outside the unit circle (the b_i and d_i vanish) and thus there are no negative quefrency terms in Eq. (6.13), so that the cepstrum (and differential cepstrum) are causal. By normal Hilbert transform relationships (see Section 3.3) this means that the real and imaginary parts of the corresponding Fourier transform, the log amplitude and phase of the spectrum, are related by a Hilbert transform, and only one has to be measured. It also means that the complex cepstrum

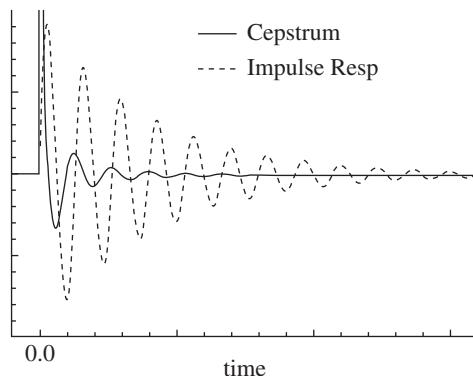


Figure 6.3 Comparison of the complex cepstrum with the impulse response function for a SDOF system.

can be obtained from the corresponding real cepstrum (which is real and even) by doubling positive quefrency terms and setting negative quefrency terms to zero. In this case also, the phase of the spectrum does not have to be measured or unwrapped, so the real cepstrum can be derived from a smoothed power spectrum resulting from excitation by a broadband random signal.

6.2 Typical Applications of the Real Cepstrum

6.2.1 Practical Considerations with the Cepstrum

There are a number of artefacts that can arise in the calculation of the cepstrum, and care has to be taken that they do not unduly affect the results. A number of such effects are illustrated in Figure 6.4.

In 6.4a it is shown that the noise level in the spectrum affects the detection of a series of harmonic components. It is obvious that if the harmonics are completely immersed in noise, they will not be detected at all in the cepstrum. A number of techniques can be used to cause discrete frequency components to stand out more from background noise. One is to use a narrower bandwidth, as mentioned in Section 3.2.8.5, and zoom analysis can be used for this. Another is to use synchronous averaging or other means to reduce noise relative to discrete frequency components. Note that synchronous averaging can only be used for harmonic components, and then only with one basic period. It cannot be used to enhance sidebands with equal spacing if the sideband family does not pass through zero

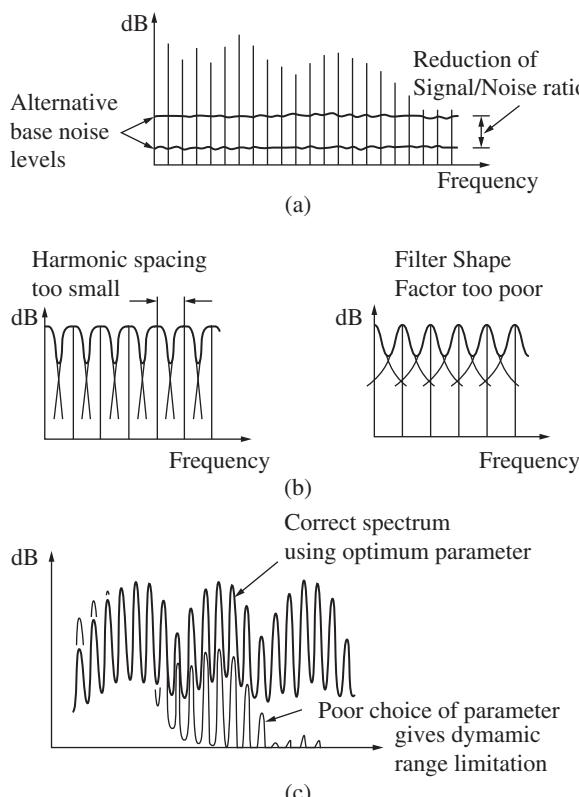


Figure 6.4 Artefacts affecting the cepstrum. (a) Effect of noise (b) Effect of filter characteristic (c) Effect of vibration parameter.

frequency (i.e. when the sidebands are also harmonics). A number of alternative techniques for separating discrete frequency and random signals are discussed in Section 5.3. In any case, the scaling of rahmonic families in the cepstrum is dependent on the signal/noise ratio, and so it is generally only valid to compare cepstra obtained using the same analysis parameters, and care should be taken that noise levels in the measurements have not changed.

The result of the cepstrum is dependent on the filter characteristic used in the original spectrum analysis, and Figure 6.4b illustrates some situations where this can directly give faulty results. It is normally assumed that if a family of harmonics or sidebands rises in the log spectrum, then this will give an increase in the corresponding components in the cepstrum. However, if the spectrum components ‘bridge over’ a fixed number of dB below the peaks, as illustrated in Figure 6.4b, then the high quefrency components will not change. Two potential reasons for the bridging are that the components are not sufficiently resolved (separated) in the spectrum, or that the filter characteristic is very poor (e.g. FFT with rectangular window). The filter characteristic of a Hanning window is quite good, and will give separation of adjacent components over a 50 dB range if they are separated by at least eight analysis lines. This would normally be sufficient for the noise to dominate in the bridging between adjacent components. If a rectangular window is used, a separation of 20 lines is preferable. In interpreting the values of different components in the cepstrum (resulting from series of harmonics in the spectrum) it should be remembered that it is analogous to frequency analysing a series of pulses of constant width but different spacing. The further apart the components, the greater the number of rahmonics that will be generated, and the smaller the amplitude of individual rahmonics for a given protrusion (in dB) of the harmonic family from the noise level in the log spectrum. Once again, it is only valid to compare cepstrum values obtained using the same analysis parameters, and even then difficult to make comparisons between components at different quefrequencies.

Another practical point illustrated in Figure 6.4c is the effect of the choice of vibration parameter (velocity or acceleration; displacement would rarely be used). Since the difference between them in principle is a moderate difference in slope of the log spectrum (corresponding to the integration from acceleration to velocity) this is a very ‘low quefrency’ effect and would not influence the high quefrency values containing diagnostic information. However, the latter can be influenced considerably if the spectrum values fall below the dynamic range of the analysis. Thus, in general it is best to choose that parameter which has the most uniform spectrum levels over the frequency range of interest. This would most often be velocity, but may occasionally be acceleration. While on the subject of dynamic range, it should be kept in mind that large negative deviations in the spectrum (in terms of dB) have just as much effect on the cepstrum as positive deviations, but may have little physical meaning. It may sometimes therefore be a good idea to limit negative deviations by imposing an artificial ‘noise level’ corresponding to the valid dynamic range of the measurement (80–100 dB). The reference value for the logarithms should also be chosen with care. In principle it has no effect other than on the zero quefrency value in the cepstrum, and so is optimally chosen to be in the mid range of the spectrum logarithmic values. If it is chosen as the average dB value in the spectrum, the zero quefrency value will be zero, and there will be minimal effect on the useful part of the cepstrum. However, since there is a small, but finite, chance that a spectrum value in a single FFT might be close to zero, which corresponds to $-\infty$ dB, it is probably better to use the median, rather than the mean dB value as reference level.

Because of the required good resolution of harmonic/sideband components in the spectrum, it is often advantageous to acquire the latter using zoom. Selection of a portion of the spectrum can also be done to exclude unwanted components. The latter might include low harmonics of shaft speed affected by phenomena such as misalignment, so as to separate them from modulation sidebands around garmesh frequencies. When performing cepstrum analysis on a zoom spectrum, the left hand

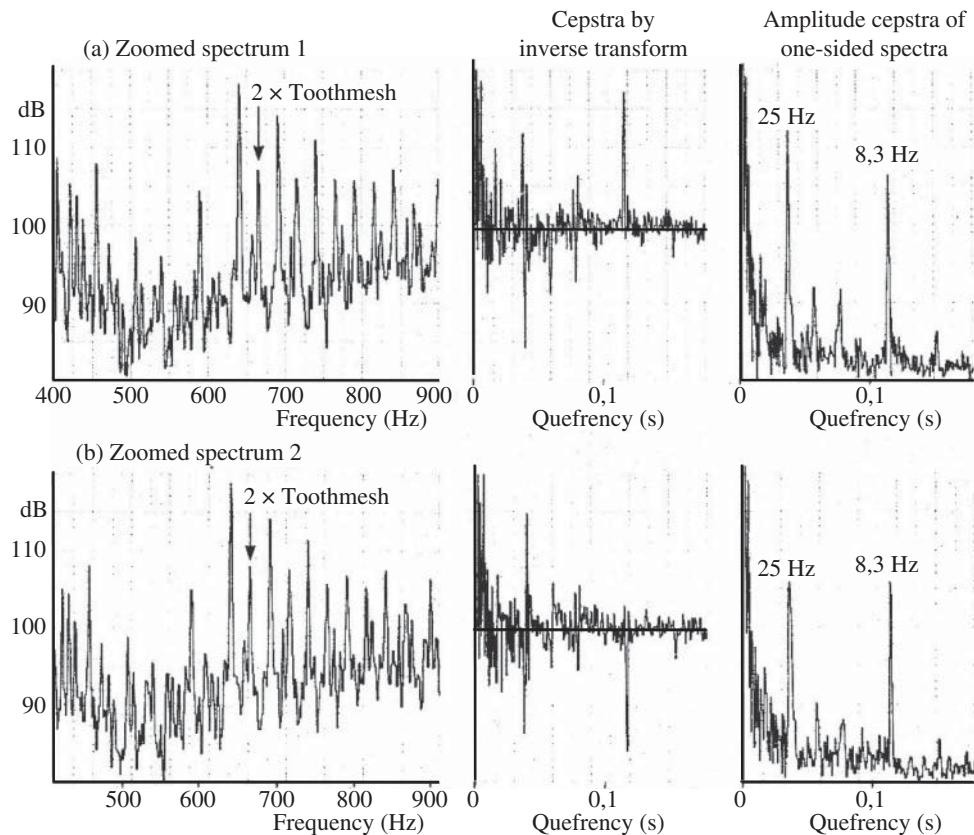


Figure 6.5 Advantage of the analytic cepstrum for zoom spectra.

end of the spectrum no longer corresponds to zero frequency, and the resulting cepstrum can give confusing results. Figure 6.5 shows an example where two slightly shifted zoom spectra have been used to obtain the cepstrum encompassing the sidebands around the second harmonic of a garmesh frequency. Because the sideband families no longer pass through the effective ‘zero frequency’ at the left hand end of the spectrum, the rahmonics in the real cepstrum are no longer positive peaks corresponding to the sideband spacings. The quefrency corresponding to the 25 Hz spacing is close to a zero crossing in both cases. The quefrency corresponding to the 8.3 Hz spacing has a positive peak in one case and a negative in the other. This problem can be solved by making use of Hilbert transform principles, since the true spacing will be indicated by the peak in the amplitude of the complex (analytic) signal obtained by inverse transforming the one-sided log spectrum (zero padded to replace the negative frequency components). For proper scaling, the one-sided spectrum (in dB) should be doubled in amplitude. It is convenient to call such a cepstrum an ‘analytic cepstrum’, to distinguish it from the complex cepstrum, which is real.

Exactly the same phenomenon will be encountered whenever a sideband family does not pass through zero frequency, even the genuine zero frequency. For normal parallel gears, the garmesh frequencies are a harmonic of both shaft speeds, and so modulation by these shaft speeds gives sideband families that are also harmonics, and thus pass through zero frequency. However, in planetary

gears, not all modulation frequencies are submultiples of the garmesh frequency, so sideband families do not necessarily pass through zero frequency.

Note that the use of the analytic cepstrum will typically be for the applications described in Section 6.2.2, where it is being used primarily as an aid to interpreting spectra, and little editing (of cepstra in particular) is being done. For the applications of Sections 6.2.3 and all of 6.3, it is best to use rectangular windows to avoid distortion.

6.2.2 Detecting and Quantifying Harmonic/Sideband Families

Figure 6.6 compares the cepstrum and the autocorrelation function for the case of a fault in a high speed bearing. This illustrates the difference made by taking the logarithm of the power spectrum before the inverse transform to the cepstrum. In the log power spectrum in 6.6a there is a family of harmonics of the BPFO (ballpass frequency, outer race), with a spacing of 206 Hz. These are not even visible in the power spectrum, since they are all below the -20 dB line, which corresponds to 1% of full-scale on the linear (amplitude squared) scale. Thus, the cepstrum (Figure 6.6c) is dominated by rahmonics corresponding to the BPFO (spaced at 4.84 ms), whereas the autocorrelation function (Figure 6.6b) simply exhibits a beat between the two largest components in the spectrum, which have no relation to the bearing fault. Note that the accuracy given by the quefrency of the cepstrum component is very good, because it represents the average of the spacings in the whole spectrum. Note also that the cepstrum can only be used for bearing fault diagnosis when the fault generates discrete harmonics in the spectrum. This is often the case for high speed machines, where resonances excited by the fault represent a relatively low harmonic order of the ballpass frequencies involved, but is usually not the case for slow speed machines, where this order may be in the hundreds or even thousands, and these high harmonics are typically smeared together. It should be noted that ‘envelope analysis’, where the envelope obtained by amplitude demodulation of the bandpass filtered signal is frequency analysed, can be used in either case.

Figure 6.7 shows a similar example for the case of a steam turbine with missing blades [11]. It is from work done by the French Electrical Authority EDF, where loss of one or a small number of small blades on a large turbine had been experienced, with little effect on the overall vibration level (the loss of one blade sometimes gave rise to the ‘wiping’ of a blade on the other side, thus returning the turbine to approximate balance). The misdirected steam from a faulty blade caused an interaction with a succession of stator blades, which was detected as an impulsive event once-per-revolution of the rotor by an accelerometer mounted externally on the casing. This gave a considerable increase of harmonics of shaft speed (50 Hz) in the mid frequency range used to detect the phenomenon.

The specific application of cepstrum analysis to gear diagnostics is treated in more detail in Section 7.2.3 of Chapter 7.

It is important to realise that because of the linear abscissae of both the spectrum and cepstrum, they only cover a limited frequency range (the upper decade is from 10% to 100% of full scale). Thus, from exactly the same signal, it is often necessary to obtain three different spectra with different frequency ranges, and then the three corresponding cepstra will also contain different information. This is typically the case for helicopter and wind turbine gearboxes, with an overall speed ratio of about 100 : 1.

Figures 6.8–6.10 (from [12]) illustrate this for signals from a wind turbine gearbox, in healthy and faulty condition, from a round robin conducted by NREL in the US [13]. Figure 6.8 compares spectra in the range up to 100 Hz, and the corresponding cepstra.

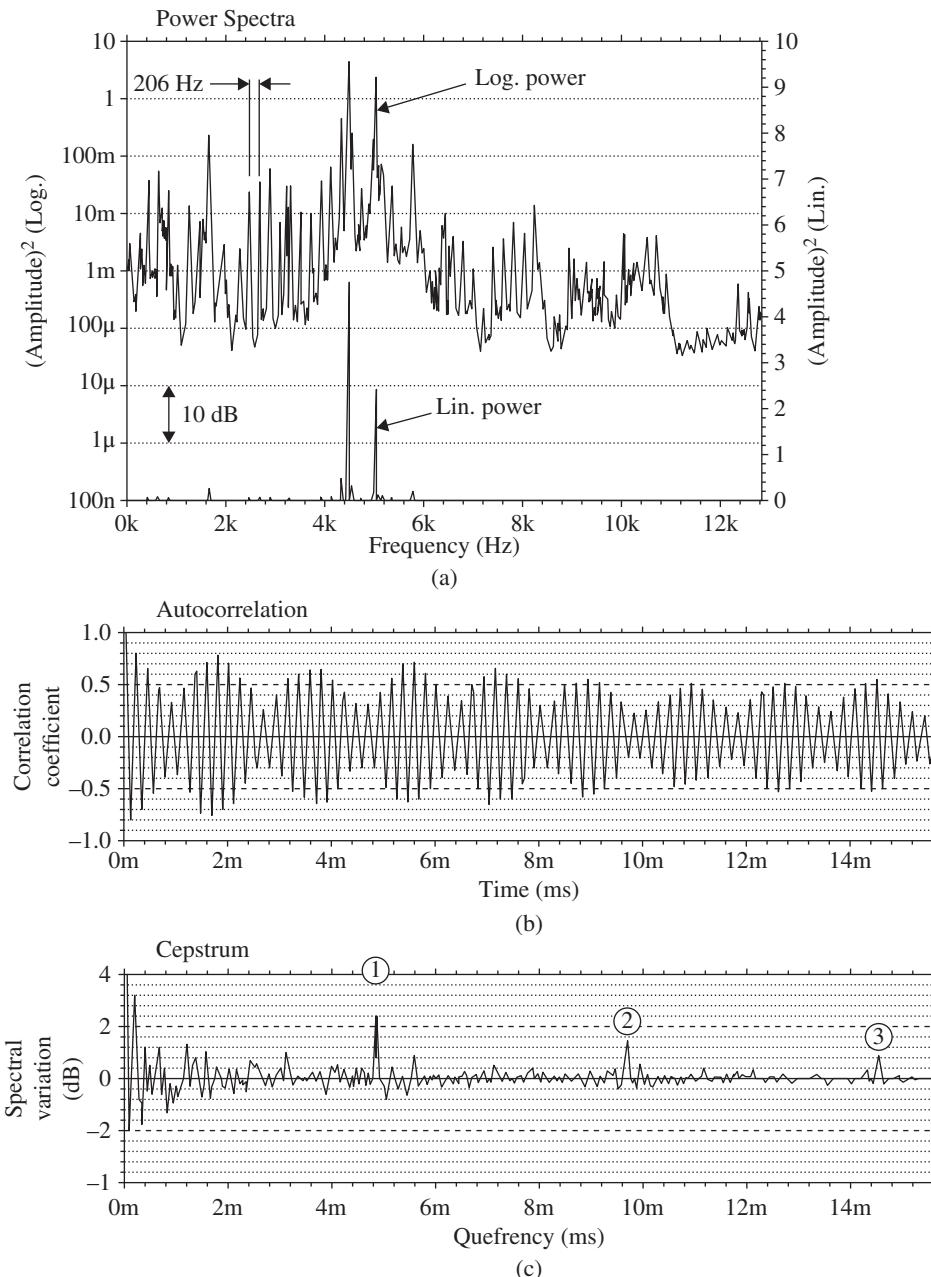


Figure 6.6 Comparison of cepstrum and autocorrelation function for the case of a bearing fault.

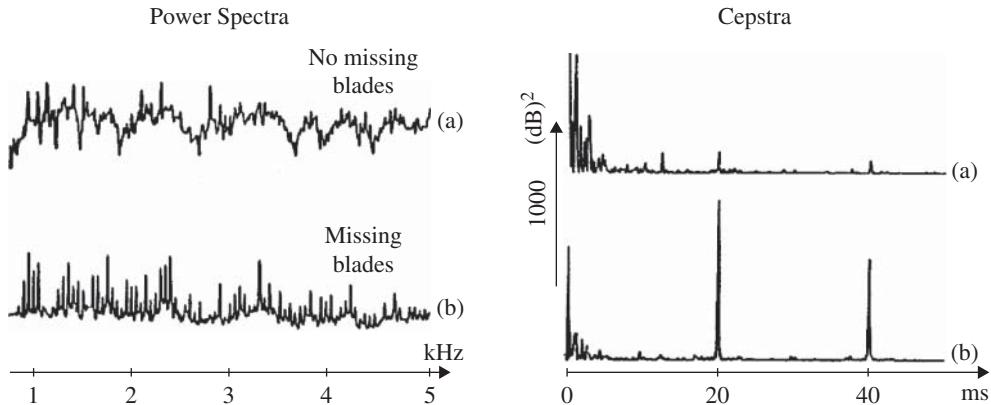


Figure 6.7 Use of the cepstrum to detect missing blades in a steam turbine [11].

It is seen that this frequency range encompasses two harmonics of the epicyclic gearmesh frequency, which in the faulty case are surrounded by sidebands spaced at the planet pass frequency ($3 \times$ the planet carrier speed). It can be seen that the growth of the gearmesh harmonics (corresponding to uniform wear over the whole gear) is clearly evident in the spectra, but the cepstra give more information about the growth of modulation sidebands at the rate at which the meshing planets are passing the measurement point. The second harmonic in Figure 6.8d corresponds to $1 \frac{1}{2}$ times the carrier speed and is unexplained.

Figure 6.9, for the frequency range up to 500 Hz, contains completely different information. There is seen to be some growth in the low harmonics of the high speed shaft, but little growth in modulation sidebands with the fault development, as evidenced by the small change in the cepstra, with only a small increase in the first harmonic of the intermediate shaft speed.

Figure 6.10, for the frequency range up to 2000 Hz, contains very different information again, with the spectra encompassing several harmonics of the intermediate shaft gearmesh (highlighted with a harmonic cursor), and two of the high speed shaft gearmesh. Comparing the spectra in 6.10a,c it is seen that there has been a considerable increase in the harmonics of the high speed gear mesh (HSGM), in particular of the second which has increased from 80 to nearly 110 dB. There has also been a considerable increase of all the harmonics of the intermediate shaft (IS) gear mesh (highlighted by the harmonic cursor). This indicates uniformly distributed wear. The spectrum of 6.10a, however, is complicated by the growth in multiple sidebands, mainly spaced at the speed of the high speed shaft (HSS) around the harmonics of the HSGM. Comparing the cepstra in 6.10b,d, gives a much clearer picture of the sideband structures, where spacings at both the HSS and ISS are apparent in deteriorated condition. Localised spalls were found on the HSS gear, which greatly increased the sidebands with this spacing. This situation is complicated by the very poor design of this gear-set with an exact 4 : 1 ratio (88 : 22 teeth), and the two sets of sidebands would be even better separated in the more normal case of a hunting tooth design (which did apply to the intermediate gearmesh).

To summarise, the (log) spectrum gives most information about faults distributed over the whole gear, which change the harmonics of gearmesh frequency, while the cepstrum gives a very compact impression of the complex sidebands which result from localised faults on individual gears.

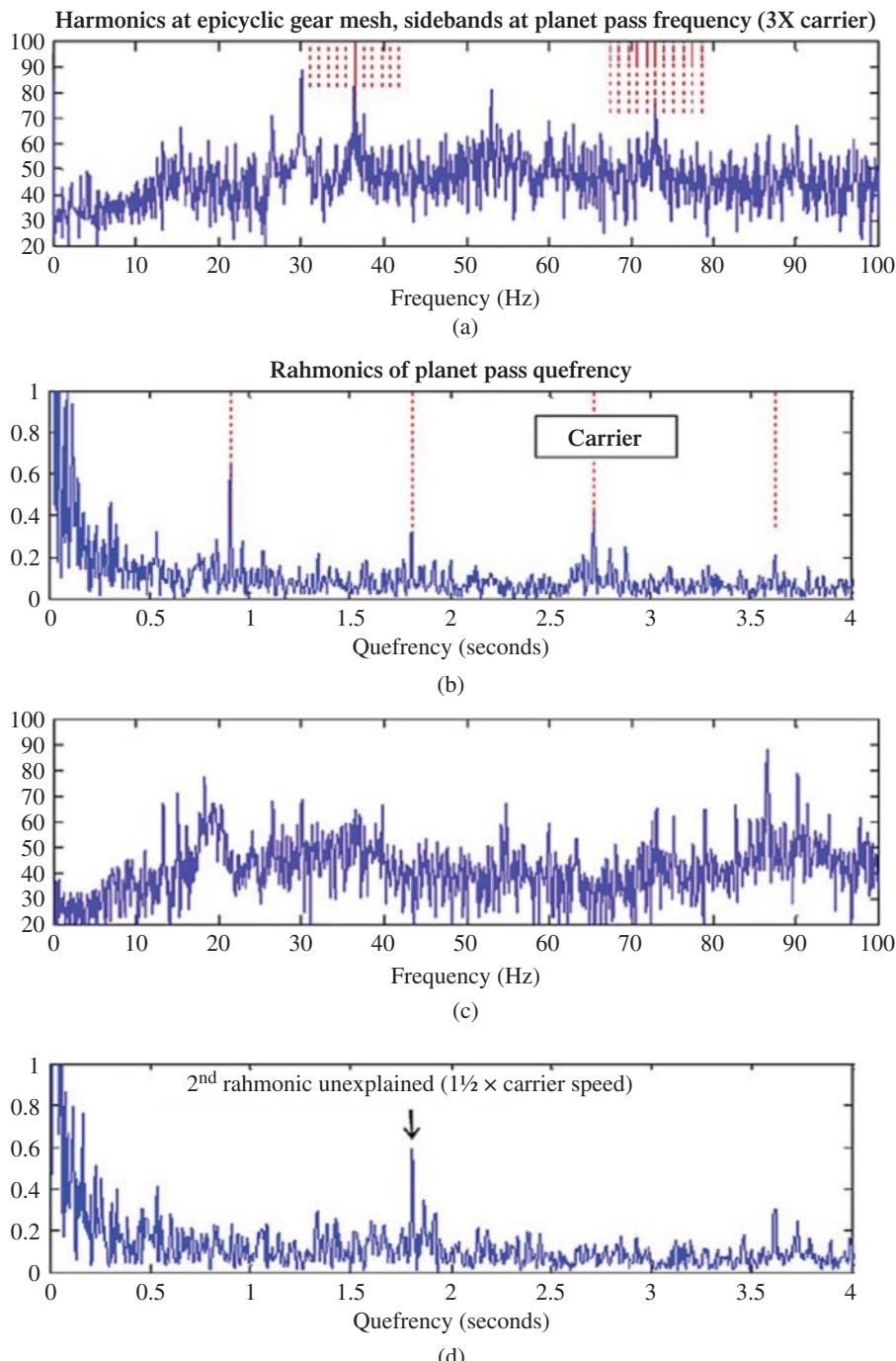


Figure 6.8 Comparison of spectra for low frequency range and of the corresponding cepstra. (a, b) Faulty (c, d) Healthy (a, c) Spectra (b, d) Cepstra.

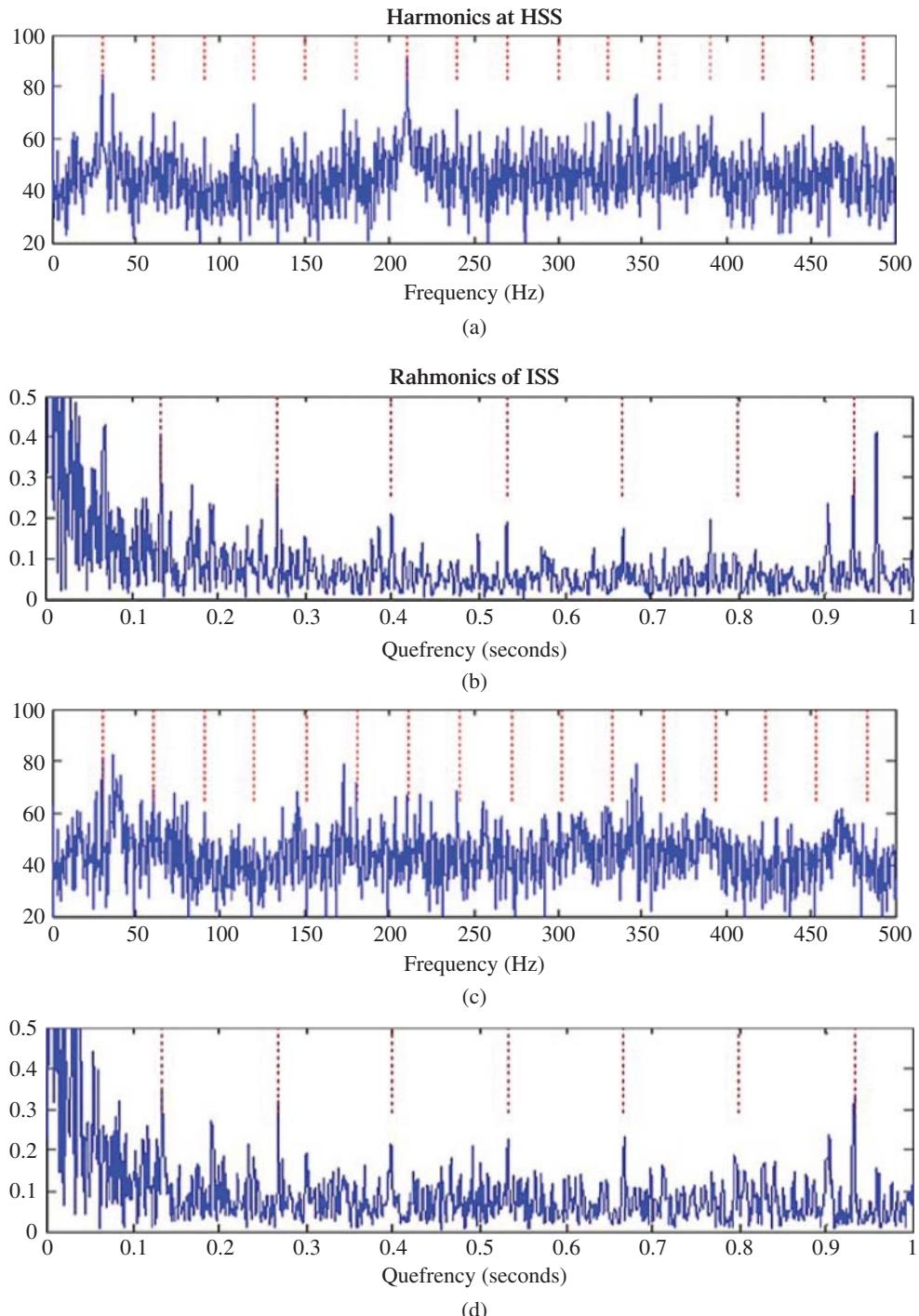


Figure 6.9 Comparison of spectra for intermediate frequency range and of the corresponding cepstra.
 (a, b) Faulty (c, d) Healthy (a, c) Spectra (b, d) Cepstra.

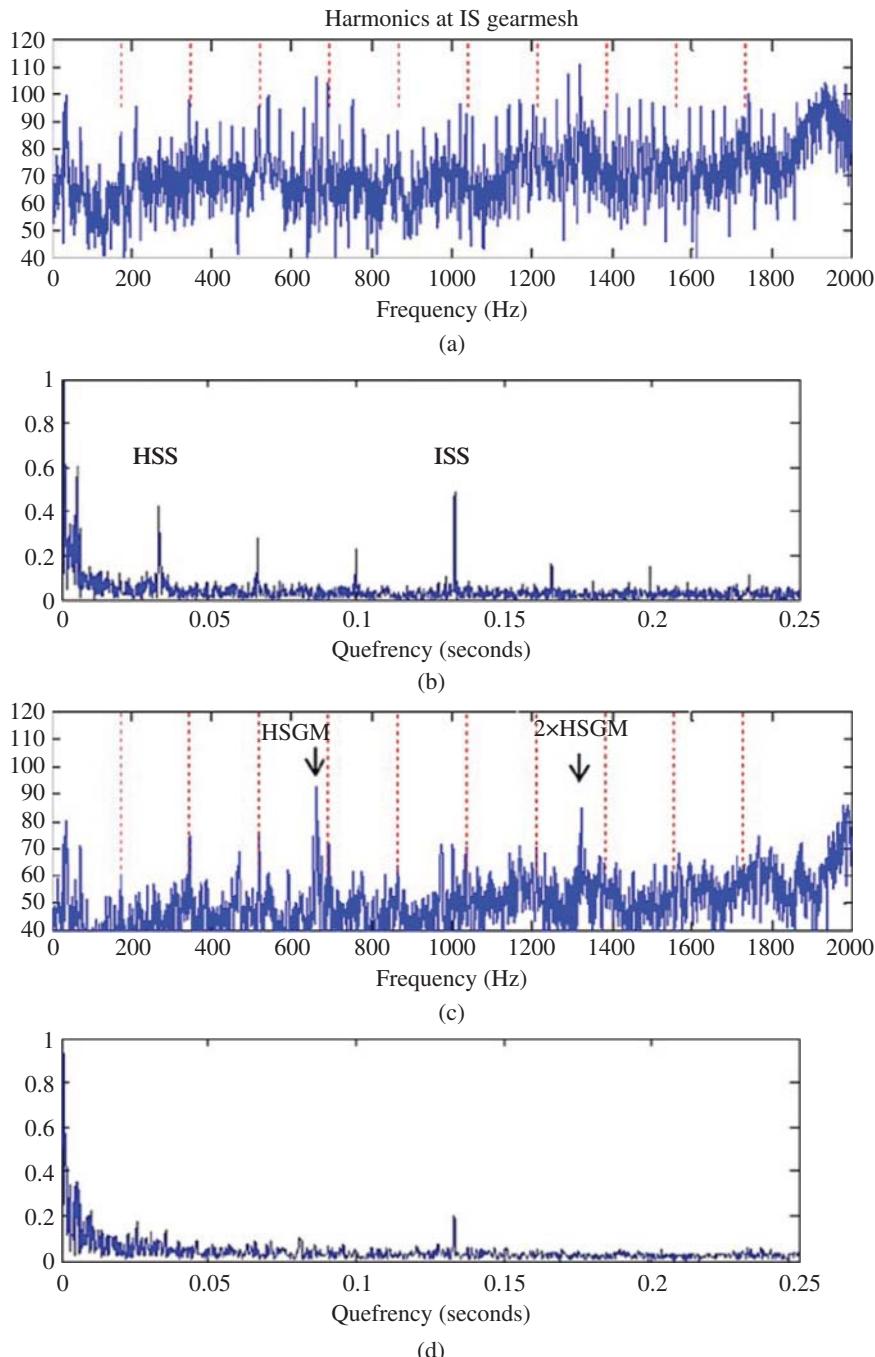


Figure 6.10 Comparison of spectra for high frequency range and of the corresponding cepstra. (a, b) Faulty (c, d) Healthy (a, c) Spectra (b, d) Cepstra.

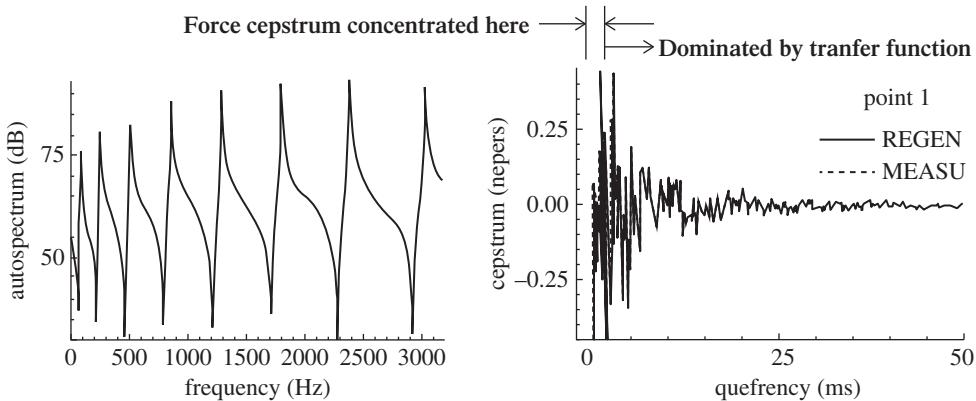


Figure 6.11 Extraction of the complex cepstrum of a transfer function from the response of a beam (a) driving point response autospectrum (b) measured and regenerated cepstra [14].

6.2.3 Separation of Forcing and Transfer Functions

As mentioned in the introductory Section 6.1.1, the cepstrum of a response signal (at least for SIMO systems) is the sum of the cepstra of the forcing function and the structural transfer function. Not only are they additive, but also often largely separated, in particular when the forcing function is broadband (either impulsive or broadband random) so that its log spectrum is slowly varying (low quefrency) and cepstrum very short. All quefrency components above this low quefrency are associated with the transfer function. This is illustrated in Figure 6.11, from [14]. The true cepstrum of the driving point FRF was reconstructed from the poles and zeros extracted from the cepstrum of the response to a hammer blow, curve fitted in the quefrency range above that affected by the impulsive force.

However, it is common that the quefrency range dominated by such transfer functions is lower than the range corresponding to more complex forcing functions, with multiple families of closely spaced harmonics and sidebands. This is illustrated in Figure 6.12, obtained from a measurement on a gas turbine, with multiple blade rows, auxiliary gearboxes and other accessories. This shows the use of an exponential short pass lifter to remove most of the forcing functions (except for the almost white noise associated with randomness in gear meshing, gas flow through the engine, etc.). From the discussion in connection with Eq. (6.13) it is apparent that multiplication of the cepstrum by an exponential window $\exp(-\sigma_0 t)$ will add damping equal to σ_0 to every pole and zero in the transfer function. In Figure 6.12 the value of σ_0 was made equal to the damping of the lowest apparent mode of the system (at about 2 kHz) meaning that its damping would be doubled, but since relative damping tends to be constant, the effect of the window on higher frequency modes is correspondingly reduced.

There are many diagnostic situations where only the resonance frequencies are important, and then it immaterial whether the system is SIMO or MIMO, since the poles are global properties, even if the zeros are local. Figure 6.13 (from [15]) shows an example from the case of a ball mill drive gearbox, with and without cracked teeth on the pinion. Liftering of response spectra corresponding to the two conditions was used to separate the forcing function(s) from the modal response. This was considered important because for example an increase in the second harmonic of the garmesh frequency might be due to tooth wear (normally in two patches on each tooth, one on each side of the pitchline, where there is a rolling rather than sliding action). However, it might be because a natural frequency of the structure or internals has reduced so as to coincide with the harmonic in question, because of a developing crack, and this situation might have a very different prognosis.

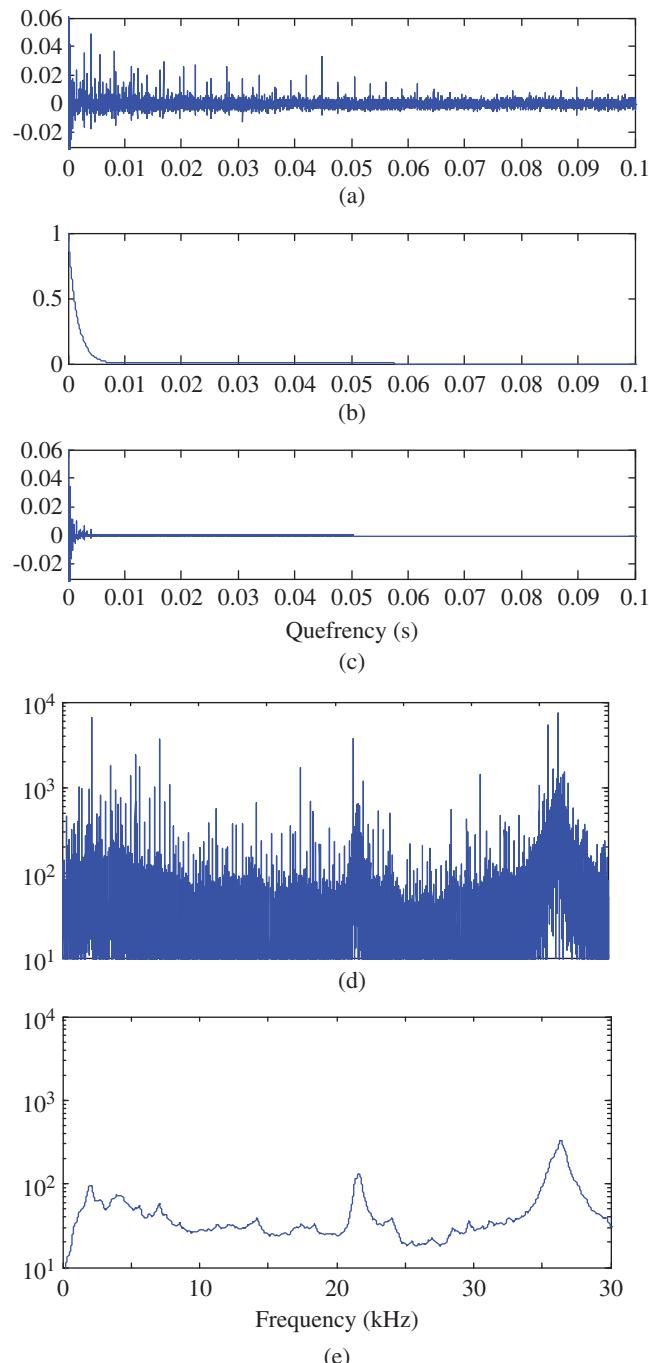


Figure 6.12 Use of an exponential lowpass lifter to enhance modal information (a) Full cepstrum (of (d)) (b) Exponential lifter (c) Liftered cepstrum (d) Original spectrum (e) Liftered spectrum.

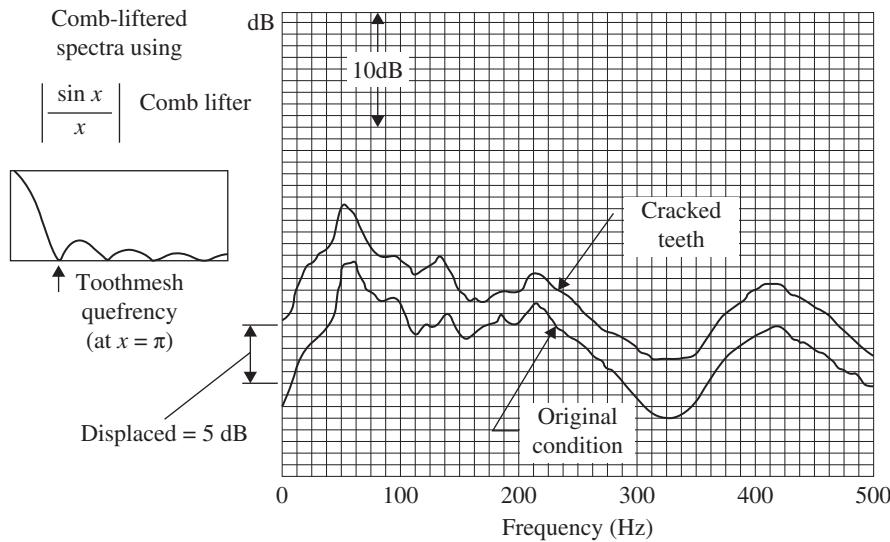


Figure 6.13 Use of lifting in the cepstrum to remove the forcing function component from the cepstra for a gear with and without cracked teeth, leaving the part dominated by the structural transfer functions [15].

The lifter used, shown in the figure, combined a comb notch lifter, tuned to the harmonics of the gearmesh frequency, with a shortpass lifter. This was done many years ago, and a combination of an exponential shortpass lifter with a rectangular comb notch lifter would probably be used now to give better (separate) control of the two filtering functions, as described in Section 6.3.

The result of the lifting shows that the modal response has changed little as a result of the developing cracks, confirming that the change was in the forcing function. Even though a modal analysis was not carried out, it is obvious that the natural frequencies did not change appreciably.

6.3 Modifying Time Signals Using the Real Cepstrum

Since the first edition of this book was published, there has been a significant new development, which introduces the possibility of editing time signals using the real cepstrum. It was previously thought that this could only be done using the complex cepstrum, but this is not possible for stationary signals and so a large number of applications are excluded. Stationary signals are composed of two types of signals:

- 1) discrete frequency components, where the phase (as a function of frequency) is unspecified between them, and
- 2) stationary random signals, whose phase is random.

In neither case can the phase be unwrapped, so it is not possible to obtain a complex cepstrum from them.

However, it was realised [16] that there are many situations where editing could be carried out by modification of the amplitude only, which can be achieved using the real cepstrum. The modified amplitude spectrum can then be combined with the original phase spectrum of each record to generate an edited time record.

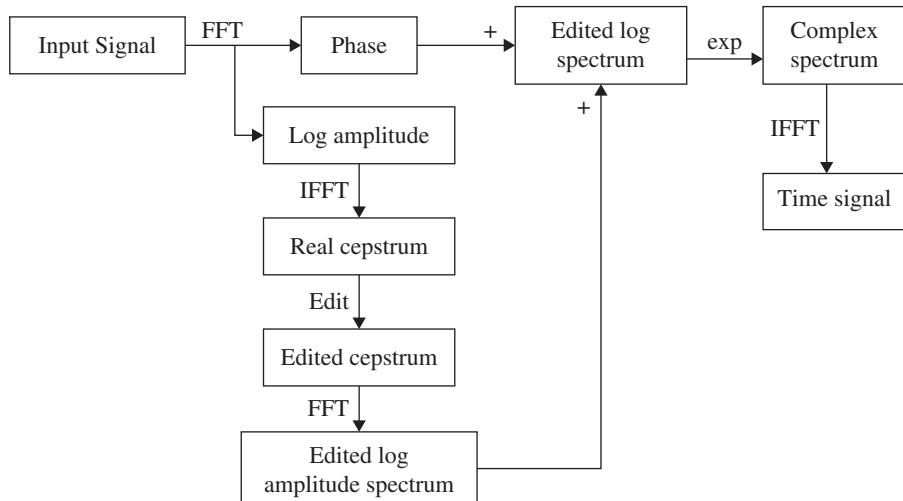


Figure 6.14 Schematic diagram of the cepstral method for editing time signals using the real cepstrum and the original phase spectrum.

The basic procedure is shown in Figure 6.14, for individual time records. These can be very long (millions of samples). They can also be woven together, after editing, by the same procedure as illustrated in Figure 5.11. After each record is transformed to a complex spectrum, of which the complex log is taken, the phase is then saved, and the log amplitude spectrum processed using the real cepstrum. The edited log amplitude spectrum is combined with the original phase, and the complex log spectrum exponentiated to a complex spectrum, which can then be inverse transformed to the edited time signal. The errors for the most important applications are small, as discussed in the next two sections.

One important application is where large numbers of discrete frequencies are to be removed (Section 6.3.1). Another is where modal properties are to be removed or enhanced, as done for the log amplitude spectrum, for example in Figure 6.13. These two applications are discussed in detail in the following two sections.

6.3.1 Removing Harmonic/Sideband Families

Whole families of uniformly spaced harmonics or sidebands can be removed by removing a small number of rahnmonics in the cepstrum. Removing a discrete frequency really means setting the value at that frequency to the expected value of the noise, of which the best guide are the frequency components on either side of the discrete frequency, but usually at a much lower level. Setting the value of the discrete rahnmonics to zero automatically smooths over the amplitude of the log spectrum in the vicinity of the corresponding harmonics and/or sidebands (since notches in the log spectrum would equally give non-zero cepstrum components as for peaks) so at least the amplitude estimate would be correct. The phase in the modified spectrum at the position of the removed components would still be the same as that for the original discrete frequencies, and thus in general incorrect, but it should be kept in mind that these are typically spaced by 20 lines or more, and thus often negligible once their amplitude is reduced to that of the noise components, which in any case have random phase. However, there is at least one case where the discrete frequency components can dominate over

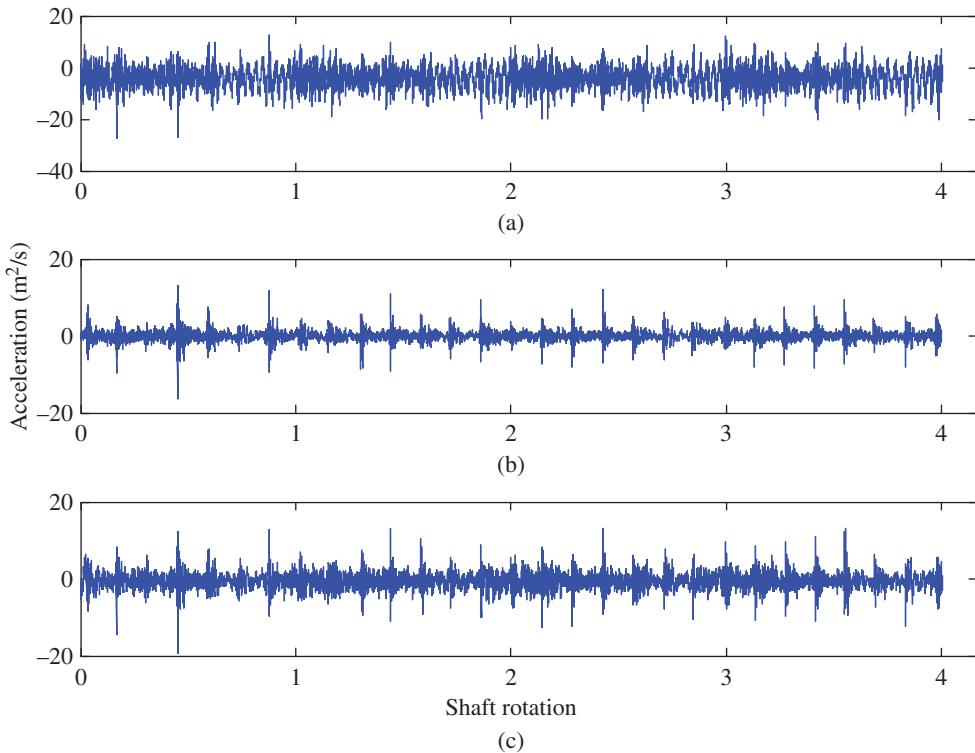


Figure 6.15 Time domain signals for gearbox test rig Source: From [16]. (a) Raw signal (b) Residual signal after removing the synchronous average. (c) Residual signal after editing the cepstrum to remove the shaft rahmonics.

noise, and that is when the whole spectrum is set to be white. This was originally called cepstrum pre-whitening, and is discussed in Section 6.3.3.

Figure 6.15 (from [16]) shows a typical application of discrete frequency removal to signals from a gearbox with a faulty bearing. It shows the original time signal, and compares it with the result of removing all harmonics of the gear signal by two methods (the gear ratio is 1 : 1 so only one set of harmonics had to be removed). The first method, Figure 6.15b, removed the Time Synchronous Average (TSA), while the cepstrum method results are shown in Figure 6.15c. Even though the result of the cepstral method is slightly noisier than for the TSA, the subsequent envelope analysis of the edited signals for the bearing diagnosis was just as successful (see Figure 6.16).

Another example in Ref. [16], for a bladed test rig with a faulty bearing, shows that only the cepstral method completely removed higher ‘harmonics’ of the shaft speed, since these corresponded to blade pass frequencies, which are not perfectly periodic because the signal from a passing blade is transmitted to the casing via a turbulent fluid, giving a small random amplitude and frequency modulation. TSA removed low order harmonics, which were almost discrete frequency, but left narrow band noise peaks at the higher harmonics, which were slightly more smeared. The cepstral method removed all frequency components with a uniform spacing, independent of whether they were discrete frequency or narrow band noise peaks.

In Ref. [17] the question is discussed as to the type of (comb) notch lifter to use for removal of families of harmonics or sidebands. It is highly unusual for the rahmonics to be at discrete quefrequencies,

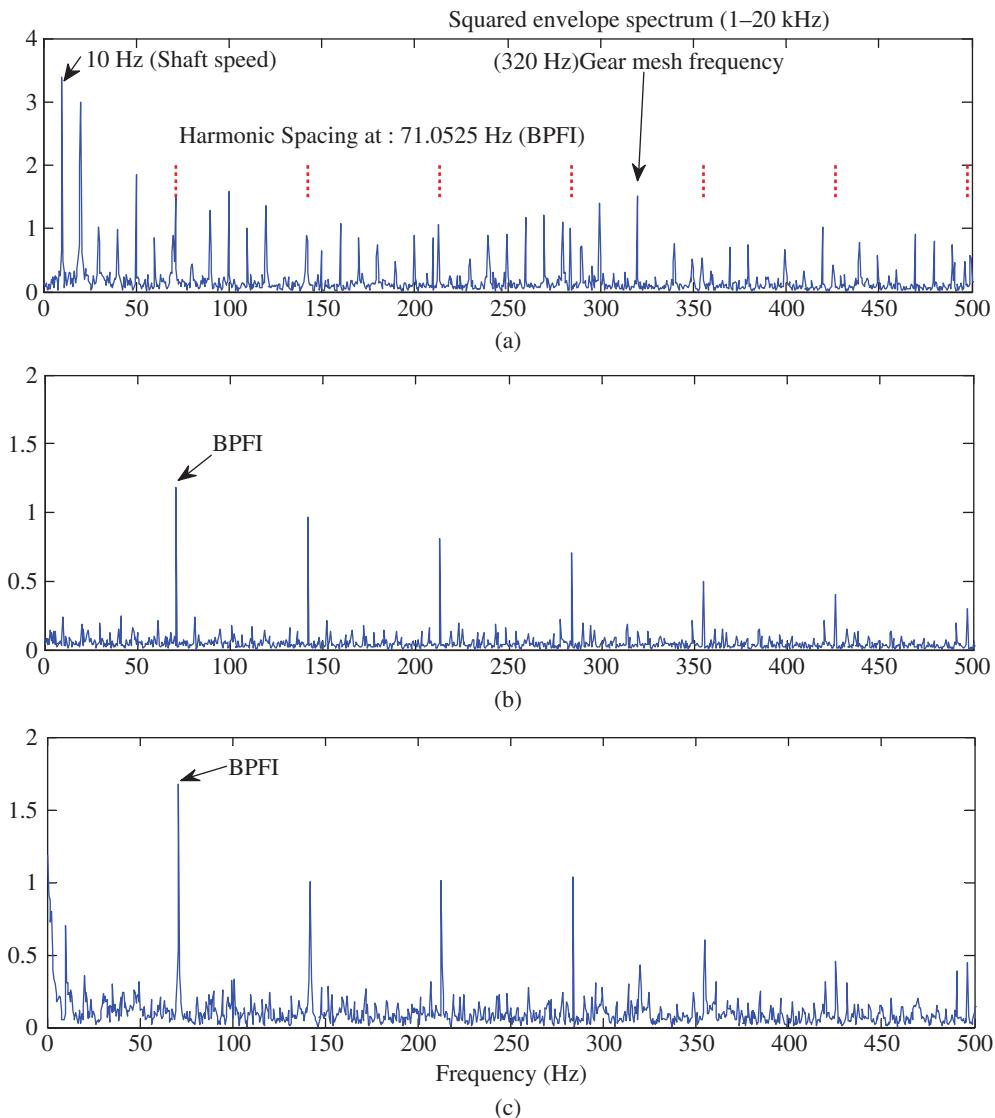


Figure 6.16 Squared envelope spectra (1–20 kHz) Source: From [16]. (a) raw signal (b) Residual signal (after removing the synchronous average) (c) Residual signal after editing the cepstrum to remove the shaft rahmonics.

since this would require the pattern of uniformly spaced components in the log spectrum to be perfectly periodic over the frequency range of the spectrum, i.e. for the frequency spacing to be an integer submultiple of the sampling frequency, and for all components to be protruding the same number of dB from a smooth (and ideally uniform) noise level. It is much more likely for this ideal pattern to be effectively multiplied by a window in the frequency domain, meaning that the corresponding rahmonics in the cepstrum would be convolved with the (inverse) Fourier transform of this window.

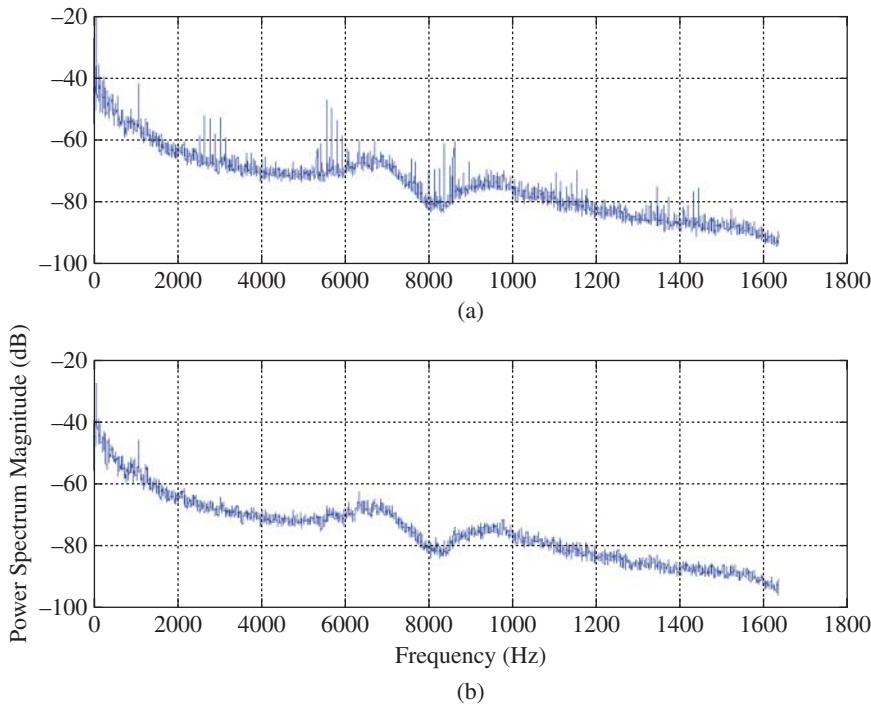


Figure 6.17 Log spectra from a wind turbine (a) before and (b) after removal of sidebands.

Figure 6.17a shows a case where a number of groups of modulation sidebands were located around harmonics of a garmesh frequency in a wind turbine spectrum. The modulation frequency (and thus the sideband spacing) is 120 Hz, twice the line frequency in the USA, and thus having an electromagnetic origin. It meant however that the sidebands were not harmonics of any fundamental frequency, and so could not be removed by TSA. However, with an appropriate comb notch lifter applied to the cepstrum, the sidebands were removed, as shown in Figure 6.17b.

Figure 6.18 shows the cepstra corresponding to Figure 6.17. Since the number of sidebands in each group in 6.17a is 6–8, the notch width had to correspond to the reciprocal of this, viz. $\pm 15\%$ of the rahmonic spacing. This was originally found by trial and error, but corresponds to the theoretical value explained above. In this example, only the modified spectrum is shown, but the equivalent edited time signal could have been obtained by the procedure shown in Figure 6.14.

In the two examples so far, the notch width was kept constant (corresponding to the Fourier transform of a fixed frequency window), and in Ref. [17] this is called a ‘Type 1’ notch lifter. There are other situations, however, where it is advantageous to increase the width of the notch with quefrency. One is where the ‘harmonics’ in the spectrum increase in width with frequency, e.g. the spectrum of a series of impulses with a small amount of random frequency modulation, such as the case of bladepass harmonics in a bladed machine, as mentioned above. Another is the case of the impulse responses from a bearing fault, where the ‘harmonics’ are increasingly smeared with increasing order. This is treated in detail in Ref. [17], where it is shown that even though the width of harmonics in the spectrum increase almost linearly with frequency, the width of the corresponding rahmonics increases at a somewhat slower rate. Even so, a ‘Type 2’ notch lifter was proposed in [17] for such cases, where the notch width was made proportional to local quefrency. One advantage of this arrangement is that the width increases to the point where the notches overlap, so that the

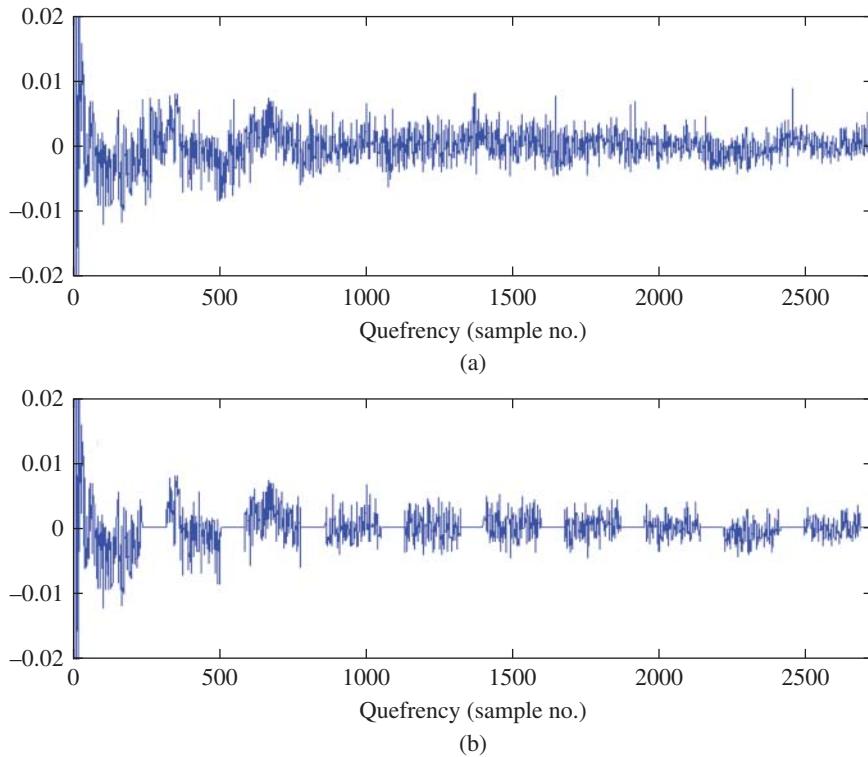


Figure 6.18 Cepstra corresponding to Figure 6.17. (a) original cepstrum (b) Cepstrum after filtering with a comb notch lifter of notch width $\pm 15\%$.

overall lifter becomes a ‘shortpass lifter’ above this quefrency. For example, a lifter of width $\pm 5\%$ of the quefrency spacing, would be $\pm 5\%$ at the first harmonic, but would remove everything above the 10th harmonic.

A Type 2 notch lifter was used in the same case as discussed in connection with Figures 6.8–6.10. In the discussion of Figure 6.10 it was pointed out that the spectrum is completely dominated by the sidebands corresponding to local faults on the high speed pinion, but also that the gear ratio of this final stage was exactly 4 : 1. This makes it difficult to remove the harmonics of this shaft, since they are also harmonics of the intermediate speed shaft (ISS). However, after forming the TSA of the ISS, it could be divided into four equal sections, each corresponding to a revolution of the HSS, which were averaged to obtain the best estimate of the HS pinion TSA. This could then be repeated four times and subtracted from the total signal to give the best estimate of the ISS TSA alone. The results are shown in Figure 6.19.

It is seen that the spectrum of the residual is dominated by the harmonics of the IS pinion gearmesh (23 teeth), but without modulation sidebands, although the noise level is quite high because of the small number of averages. These mesh harmonics were also much higher compared with the original condition, and the lack of sidebands is compatible with the fact that the wear was distributed uniformly by the hunting tooth design, a point remarked on in the inspection report [13].

It is shown in Ref. [17] that a better result can be obtained using cepstral filtering, as shown in Figures 6.20 and 6.21.

The cepstra before and after the application of the notch lifter are shown in Figure 6.20, where it is seen that the Lifter Type 2 has been used, with increasing notch width. Note that the cepstra were

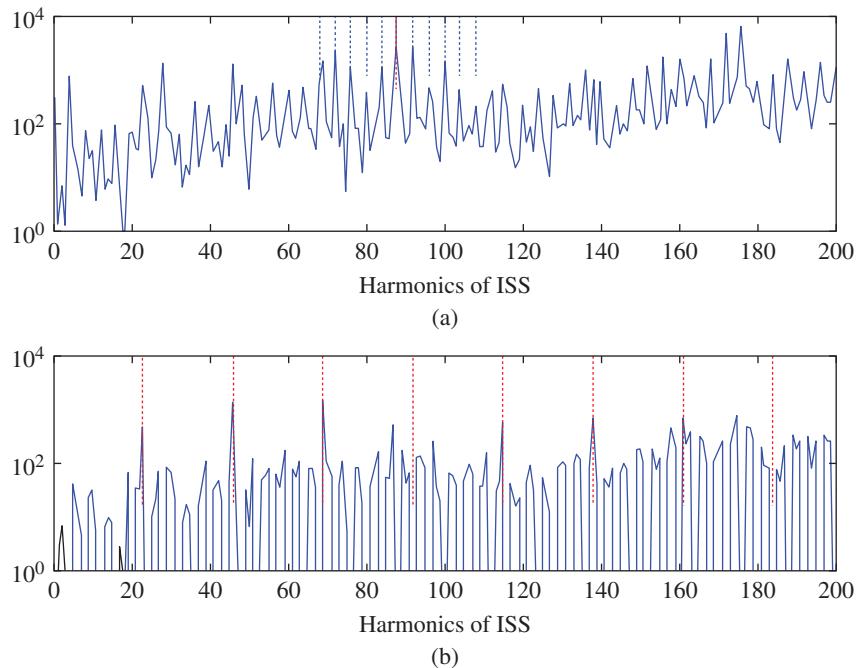


Figure 6.19 Spectra of IS TSA average (a) Original including four rotations of the HSS. Harmonics at 88 (HS mesh), sidebands at 4 (HS shaft) (b) Residual after removal of the HSS average. Harmonics at 23 (IS pinion mesh). Source: From [12].

obtained from the raw signals, without the requirement for order tracking or TSA, because the speed was reasonably stable.

Figure 6.21 shows the log spectra corresponding to the cepstra of Figure 6.20, and can be compared with Figure 6.19. It is scaled in Hz rather than harmonic order, but the HS shaft speed is about 30 Hz (and IS shaft speed 7.5 Hz). It is seen that the results are very similar, but the cepstral method gives better resolution and better definition of the base noise level.

6.3.2 Enhancing/Removing Modal Properties

In the years since the first edition of this book, it has become much more common to have to process signals from variable speed machines, such as wind turbines and mining machines. Another case in point is rail vehicles, where the (torque) load is relatively low at constant speed, except perhaps for the highest speeds of very high speed trains, and where maximum load is experienced during acceleration and deceleration.

This had led to a much better appreciation of the contributions of speed related components and fixed resonance frequencies in machine vibration signals, as discussed in connection with cyclo-non-stationary signals, Section 3.6.4. It means that certain processing operations (to do with modal properties) are best done in the time/frequency domains, while others (shaft speed related) are best done in the angle/order domains. Even so, while gear vibrations are dominated by components at discrete orders, such as garmesh frequencies, the responses are subject to variable gain, as they move through fixed resonance frequencies, even for a given condition. Thus, it would be

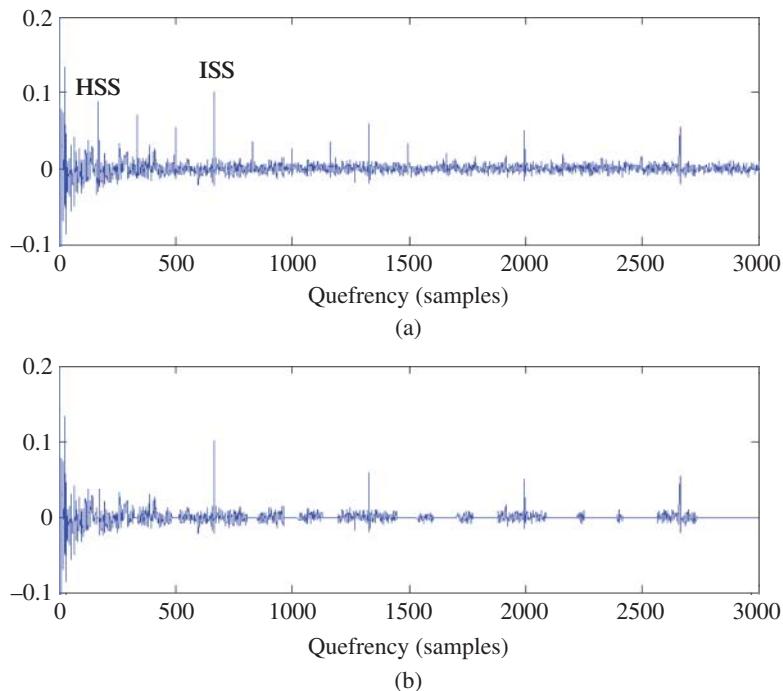


Figure 6.20 Cepstra of the IS shaft signal before (a) and after (b) the application of a notch lifter Type 2 to remove HSS harmonics.

desirable to remove the influence of these modal properties to obtain the ‘intrinsic’ forcing function, independent of speed. On the other hand, bearing fault information tends to be carried by fixed resonance frequencies (although the dominant resonances can change with change in the physical size of the faults), and these resonances are best found and isolated in the frequency domain, even though bearing characteristic frequencies are shaft speed related, so that order tracking has to be used for the subsequent envelope analysis. These points are discussed in Ref. [6], based on three conference publications [18–20].

Firstly considering gears, even if the torque were constant, the response vibrations at different measurement points would be different, but their amplitude would vary with speed, as gearmesh harmonics and associated sidebands pass through resonance frequencies. Applying order tracking to such responses removes frequency modulation, but not amplitude modulation, so in contrast to constant speed measurements, it is no longer possible to remove the gear signals by extracting them based on a synchronous average. The latter would have the ‘average’ amplitude for the whole record, and when repeated and subtracted, would still leave the (deterministic) variations around the mean amplitude. In [18, 20] it was proposed that the amplitude variations due to modal effects could be removed by first determining the latter by exponential shortpass filtering, and then subtracting them from the log amplitude spectrum, or directly in the real cepstrum. Strictly speaking, this does not remove phase changes due to the varying transfer functions, but even so a large improvement can be achieved as shown in Figures 6.22 and 6.23 (from [20]).

Figure 6.22 shows the spectrum of a gear response signal, for one pattern of speed variation (random $\pm 20\%$ around 22 Hz), before and after removing the modal information extracted using an exponential lifter in the real cepstrum. The subtraction was actually done directly in the cepstrum.

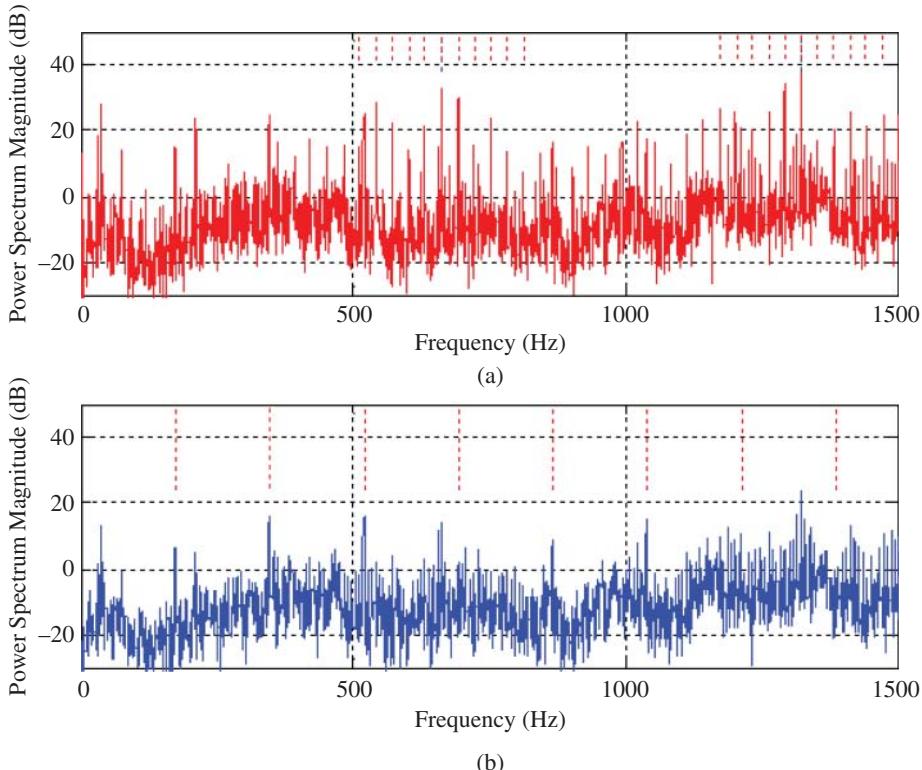


Figure 6.21 Log spectra corresponding to the cepstra of Figure 6.20 (a) Original signal. Harmonics at HS gearmesh. Sidebands at HS shaft speed. (b) Liftered signal. Harmonics at IS pinion mesh frequency.

It is seen that the liftered spectrum in 6.22b is devoid of modal peaks, and thus can be considered to represent the intrinsic forcing function. This liftered spectrum, and the equivalent one from another pattern of speed variation (random $\pm 20\%$ around 15 Hz), were used to regenerated time signals using the procedure of Figure 6.14. These were then corrected for the different speed variation patterns by order tracking, and the resultant order spectra are compared in Figure 6.23 for the original and liftered cases. It is seen that after liftering the spectra for the two cases are very similar.

The only significant differences are in harmonics 2 and 3 of the gearmesh (GM) frequency, but these are seen in 6.23a to be very weak and affected by noise at that measurement point.

A complete removal of transfer function effects, including phase shifts, would require a full modal model of the system, and to be accurate, with true natural frequencies for the operational conditions, would have to be obtained by operational modal analysis (OMA), as discussed in [9]. It would normally also mean that an analytical model, such as a finite element model, would have to be updated to agree with the OMA model. This is a much more complicated approach than that based on removal of amplitude modulation effects alone, for example requiring a true blind source separation to be able to treat each path from source to response separately [9].

There are other explanations for amplitude modulation, other than modal amplification, such as torque variations due to acceleration/deceleration, and these are discussed in Section 7.2.5 on diagnostics of gears with varying speed and load.

For bearing diagnostics, it is more likely that the cepstral separation would be used for the opposite purpose than for gears, retaining the modal response (which carries most of the bearing information)

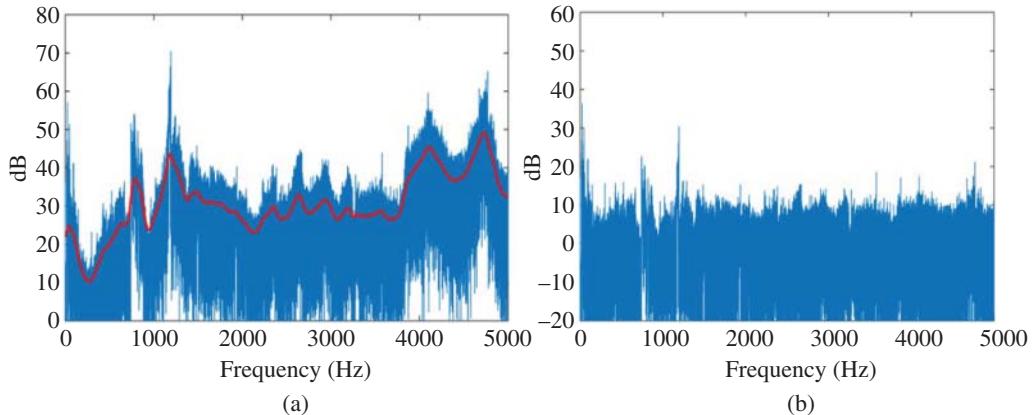


Figure 6.22 Spectra of the original (a) and ‘liftered’ (b) signal. The thicker line in (a), representing the modal response, which was subtracted in the cepstrum, was produced using an exponential lifter in the cepstrum [20].

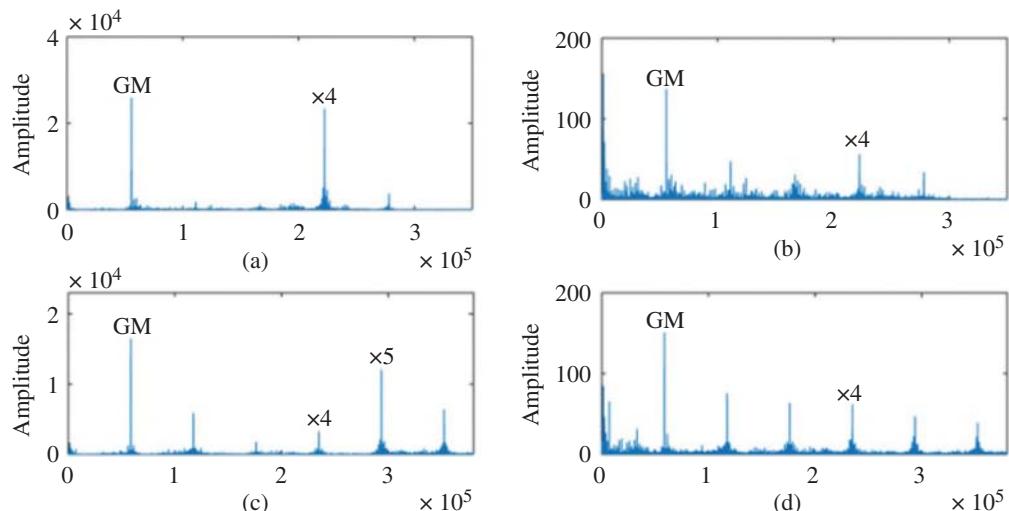


Figure 6.23 Effect of liftering on order tracked spectra (a, b) Signal varying around 22 Hz (c, d) Signal varying around 15 Hz (a, c) Original (b, d) Liftered [20].

and discarding the shaft related forcing functions, which are more likely to be dominated by gears and other non bearing related sources [18]. This is discussed in Section 7.3.5, on diagnostics of bearings with varying speed and load, where the cepstral method is compared with other alternatives.

6.3.3 Cepstrum Pre-whitening

One of the first developments from the method proposed in [16] was termed ‘cepstrum prewhitening’, and consisted in setting the whole of the real cepstrum (except possibly for the zero quefrency value) to zero, making the log amplitude spectrum completely white, and then combining that with the original phase spectrum to generate time signals [21]. The basis for this was the realisation that by

Parseval's theorem, a white spectrum meant that all frequency bands of equal width would have the same RMS value, but any bands containing impulsive components (with higher crest factor) would have higher peak values, and thus be more likely to dominate the latter in the time signal, even without any bandpass filtering. This is what is meant in Section 6.3.1 by the statement 'discrete frequency components can dominate over noise', even when their amplitude is reduced to the same level as adjacent noise components, because the discrete harmonics which produce periodic impulses are required to be cosine functions, which align in phase at the start of each new period, and therefore dominate over random components whose combined amplitude is proportional to the square root of the number.

Ref. [21] demonstrates a case with signals from a helicopter gearbox, where the cepstrum pre-whitening gives almost as good a result as a method similar to the benchmark method (see Section 7.3.2) involving separation of gear signals using DRS (Section 5.3.5) followed by linear prediction (Section 5.3.2). The effect on the time signals is shown in Figure 6.24, where the filtered signals have many similarities, even though the increase in kurtosis is somewhat higher for the DRS-LP approach.

Figure 6.25 compares the envelope spectra for the signals of 6.24b,c, but with (b) further processed using spectral kurtosis (SK, Section 5.5) for two different frequency ranges, one giving evidence of

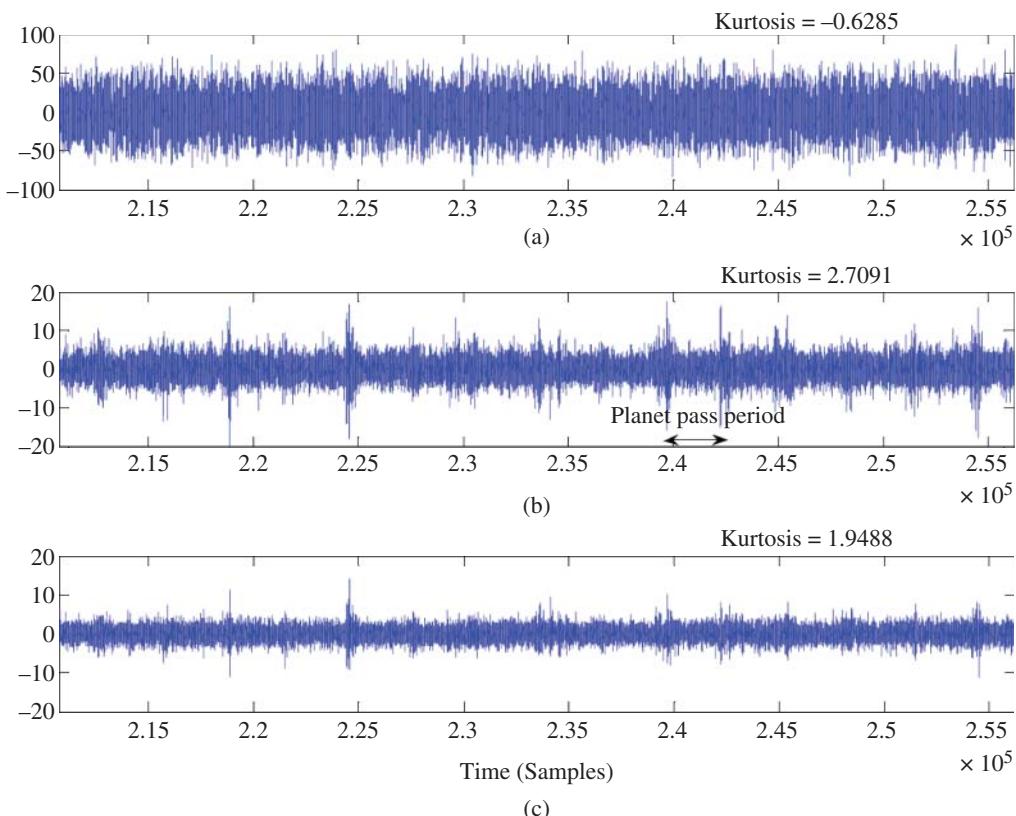


Figure 6.24 (a) Raw signal (b) DRS-LP (c) Signal processed by cepstrum pre-whitening. Source: From [21].

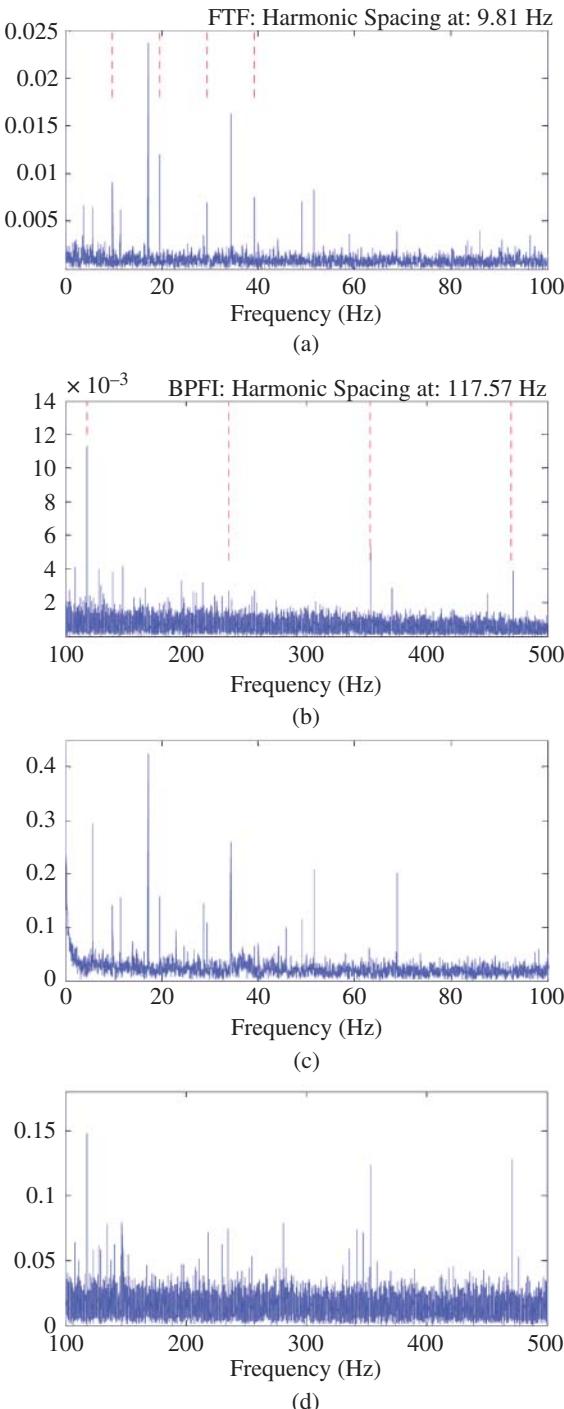


Figure 6.25 Envelope spectra for different processed signals and frequency ranges. (a)–(b) DRS-LP-SK (c)–(d) Cepstral whitening. Source: From [21].

a planet bearing roller fault, by modulation at the cage frequency (FTF), and the other of a planet bearing inner race fault, by an increase at BPFI. Harmonic cursors show the two frequency spacings, which can be seen in both filtered signals, though more strongly for the FTF components in the DRS-LP-SK case.

It should be noted that, as with all cases where modal information is removed or enhanced using cepstral liftering, the scaling is no longer meaningful, and can be set arbitrarily, possibly based on other criteria.

It was quickly realised that the process of cepstral whitening could be achieved without using the cepstrum, by simply dividing the complex Fourier spectrum of a signal by its modulus or amplitude. This would automatically set the amplitude of the spectrum to unity (dimensionless) corresponding to setting the zero quefrency value to zero, but in either case the result could be re-scaled as desired.

In Ref. [22] it was shown that the ‘cepstral pre-whitening’ procedure, even though produced by the alternative method, was very valuable as a pre-processing technique for signals from a rail vehicle transmission under acceleration/deceleration, because of its ability to largely suppress the effects of gears. Order tracking was performed after the pre-processing. A major claimed advantage, of course, is that the procedure can be applied blind, without prior knowledge of the machine structure, such as gear parameters.

References

1. Bogert, B.P., Healy, M.J.R., and Tukey, J.W. (1963). The quefrency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking. In: *Proc. of the Symp. on Time Series Analysis* (ed. M. Rosenblatt), 209–243. New York, NY: Wiley.
2. Childers, D.G., Skinner, D.P., and Kemerait, R.C. (1977). The cepstrum: a guide to processing. *Proceedings of the IEEE* 65 (10): 1428–1443.
3. Randall, R.B. (1987). *Frequency Analysis*, 3e. Copenhagen: Brüel & Kjaer.
4. Oppenheim, A.V. and Schafer, R.W. (2004). From frequency to quefrency: a history of the cepstrum. *IEEE Signal Processing Magazine* 106: 95–99.
5. Randall, R.B., (1973). “Cepstrum analysis and gearbox fault diagnosis”, *Brüel and Kjaer Application Note* No. 13-150, Copenhagen.
6. Randall, R.B. (2017). A history of cepstrum analysis and its application to mechanical problems. *Mechanical Systems and Signal Processing* 97: 3–19.
7. Polydoros, A. and Fam, A.T. (1981). The differential cepstrum: definitions and properties. In: *Proc. IEEE Int. Symp. Circuits Systems*, 77–80. IEEE.
8. Oppenheim, A.V. and Schafer, R.W. (1989). *Discrete Time Signal Processing*. New Jersey: Prentice-Hall.
9. Randall, R.B., Antoni, J., and Smith, W.A. (2019). A survey of the application of the cepstrum to structural modal analysis. *Mechanical Systems and Signal Processing* 118: 716–741.
10. Antoni, J., Guillet, F. and Danière, J., (2000). “Identification of non-minimum phase transfer functions from output-only measurements” *ISMA25 Conference*, KUL, Leuven Belgium.
11. Sapy, G. (1975). Une Application du Traitement Numérique des Signaux au Diagnostic Vibratoire de Panne: La Détection des Ruptures d’Aubes Mobiles de Turbines. *Automatisme*, Tome XX (10): 392–399.
12. Sawalhi, N., Randall, R.B., and Forrester, D. (2014). Separation and enhancement of gear and bearing signals for the diagnosis of wind turbine transmission systems. *Wind Energy* 17 (5): 729–743.
13. Sheng, S. (Ed.) (2012). Wind turbine gearbox condition monitoring round robin study – vibration analysis. National Renewable Energy Laboratory, Technical Report NREL/TP-5000-54530 (July 2012). doi: <http://www.nrel.gov/docs/fy12osti/54530.pdf>.
14. Gao, Y. and Randall, R.B. (1996). Determination of frequency response functions from response measurements—I. extraction of poles and zeros from response cepstra. *Mechanical Systems and Signal Processing* 10: 293–317.
15. Randall, R.B. (1997). Advanced machine diagnostics. In: *The Shock and Vibration Digest*, vol. 29(1), 6–30. Willowbrook, IL, USA: Vibration Institute.
16. Sawalhi, N., Randall, R.B., (2011). “Editing time signals using the real cepstrum”. *MFPT Conference*, Virginia Beach (May 2011).

17. Randall, R.B., Sawalhi, N. (2013). "Cepstral removal of periodic spectral components from time signals", *CMMNO Conference*, Ferrara, Italy (8–10 May).
18. Randall, R.B., Smith, W.A., Coats, M.D. (2014). "Bearing diagnostics under widely varying speed conditions". *CMMNO Conference*, Lyon, France (15–16 December).
19. Randall, R.B., Coats, M.D., Smith, W.A. (2014). "Gear diagnostics under widely varying speed conditions". *CMMNO Conference*, Lyon (15–16 December).
20. Randall, R.B., Smith, W. (2016). "New cepstral methods for the diagnosis of gear and bearing faults under variable speed conditions". *ICSV 2016 – 23rd International Congress on Sound and Vibration*, Athens, Greece.
21. Sawalhi, N., Randall, R.B. (2011). "Signal pre-whitening using cepstrum editing (liftering) to enhance fault detection in rolling element bearings". *Comadem Conference*, Stavanger, Norway (May/June).
22. Borghesani, P., Pennacchi, P., Randall, R.B. et al. (2013). Application of cepstrum pre-whitening for the diagnosis of bearing faults under variable speed conditions. *Mechanical Systems and Signal Processing* 36: 370–384.

7

Diagnostic Techniques for Particular Applications

7.1 Harmonic and Sideband Cursors

7.1.1 Basic Principles

The diagnostic capability of frequency analysis is vastly increased by adding harmonic and sideband cursors to the toolkit. A harmonic cursor is a set of markers indicating all members of a specific harmonic family with a very fine resolution. The diagnostic power resides in the fact that the harmonics are exact integer multiples of the fundamental frequency, and so if a high order harmonic can be located, the accuracy of determination of the harmonic spacing is increased in proportion to the harmonic order. Moreover, if when zooming in a high frequency range, a number N of harmonics of the same frequency spacing can be selected at the same time, the accuracy is increased in proportion to the number N . This applies despite the fact that the individual cursor lines may be limited to the spectral resolution of the FFT analysis, but means that the adjustment of the harmonic spacing must have a resolution considerably finer (by at least a factor N) than the spectral resolution in the zoom band.

A sideband cursor similarly depicts a family of sidebands with a given spacing around a specified central ‘carrier’ frequency. The accuracy is not the same as for a harmonic cursor, since the family is not constrained to pass through zero frequency, but the same improvement in accuracy can be obtained by simultaneous selection of a number of sidebands in the same family. Moreover, in some cases the sidebands are also harmonics, such as in gear vibrations where the toothmesh frequency is an integer multiple of both shaft speeds (the number of teeth on that gear) and the modulations are at multiples of the shaft speeds, meaning that the resulting sidebands are also harmonics. In this case, a harmonic cursor could be used to determine the sideband spacing even more accurately. The centre (carrier) frequency of the sideband family should be represented in the sideband cursor family with the same accuracy as obtained from a harmonic cursor (the carrier would most often be a harmonic) rather than at the centre frequency of the analysis line closest to the harmonic in question.

7.1.2 Examples of Cursor Application

Simple examples are given by the separation of the harmonics of shaft speed from those of mains (line) frequency in induction motor vibrations, as illustrated by Figure 2.21, in which case a very high degree of accuracy can be achieved from a harmonic cursor using the fact that both are true series of harmonics. Similarly, an example of the use of a sideband cursor is given by Figure 2.23, for confirming that the sideband spacing is at slip frequency times the number of poles. In this case, the slip frequency could be determined very accurately from harmonic cursors applied to the shaft speed and (double) mains frequency components, and used to set the sideband cursor. In fact, the difference in frequency between twice mains frequency and the nearest shaft harmonic would always represent the sideband spacing for rotor faults as for the example in Figure 2.23.

An even more powerful application applies to gears, in that a harmonic cursor can be used for the blind determination of the numbers of teeth on a gear pair, as long as this represents a ‘hunting tooth’ design. If these numbers are m and n , then a hunting tooth design means that there are no common factors. This is considered good design practice, since it means that each tooth on one gear contacts every tooth on the mating gear, and, for example, a fault on one tooth will be smeared over every tooth on the other gear rather than causing more localised effects. If m and n are both divisible by 3 for example, it means that groups of three teeth on one gear would form a compound ‘tooth’, which always meshed with similar compound groups of three teeth on the other gear. As illustrated in Figure 7.1, for a hunting tooth design the toothmesh frequency is the first common harmonic frequency for the two gear rotational speeds. The closest the harmonics can approach elsewhere is $\frac{1}{m \times n}$. Also shown in Figure 7.1 is the result of applying harmonic cursors to the two sets of harmonics in the acceleration spectrum of a gear pair. It requires that the machine speed is stable to about 1 : 20000, but this is not unusual, and where it is not the case, order tracking (Section 5.1) can be used to achieve this degree of stability. It also requires that a reasonable number of harmonics of both shaft speeds are present in the spectrum (on a logarithmic amplitude or dB scale) in particular the low orders and as sidebands in the vicinity of the toothmesh frequency, but once again this is not unusual. A harmonic cursor is set up on each shaft in succession, first on the low orders, and then progressively adjusted by zooming in higher frequency bands. This will usually determine the fundamental frequency to the required accuracy. If lists of the two harmonic series are then compared, the toothmesh frequency corresponds to where they match to better than 1 : 20000. In Figure 7.1, this is found to be about 3357.7 Hz where the agreement is 1 : 33000, and indicates that the numbers of teeth are 34 and 135, respectively. In this case, the smallest alternative difference is at approx. 3258 Hz, where the discrepancy is approx. 1 : 5000, in agreement with $m \times n$. Note that in the figure, the shaft speeds are only indicated to 3-figure accuracy, but can be obtained to 5-figure accuracy by dividing the highest adjusted harmonic frequency by the corresponding order. Even for non hunting tooth combinations, where both tooth numbers are divisible by 2 or 3, for example, it is still often possible to divine the actual numbers of teeth, even though the harmonic patterns would coincide at 1/2 or 1/3 and 2/3 the garmesh frequency, respectively. One case is where the apparent number of teeth on the pinion would be too small (normally a minimum of 13). Another applies when there is an inspection port, which allows the tooth pitch to be measured roughly, even with a tape measure. It would then be obvious if the actual numbers of teeth were two or three times greater than the apparent numbers.

7.1.3 Combination with Order Tracking

Even for machines with variable speed, this procedure can be used in combination with order tracking, as long as the speed variation does not give rise to too much amplitude modulation (which would

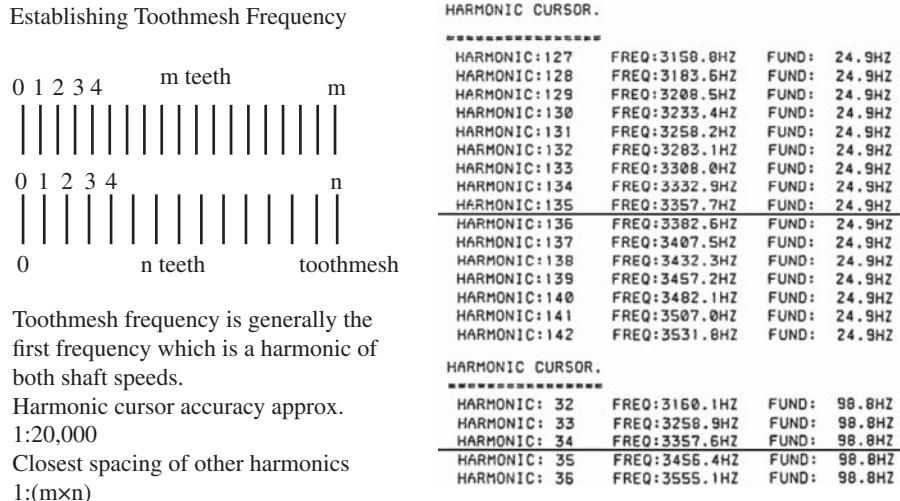


Figure 7.1 Harmonic cursor used to find the number of teeth in a gear pair. Source: Courtesy Brüel & Kjær.

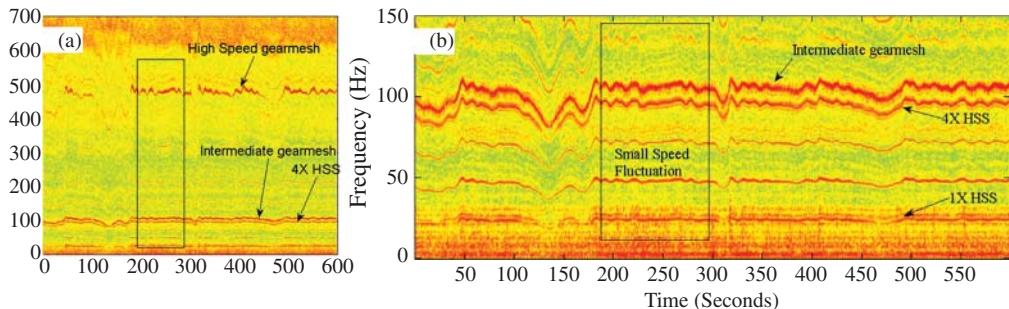


Figure 7.2 Spectrograms in two frequency ranges for wind turbine gearbox. (a) Including 1st harmonic of HS gearmesh. (b) Including 1st harmonic of IS gear mesh.

smear the order tracked harmonics). This was demonstrated in Ref. [1], where the numbers of teeth in the two parallel gear sections of a wind turbine gearbox were identified from signals obtained when the machine was operating at variable speed.

Figure 7.2 shows two spectrograms, in different frequency ranges, of a typical signal measured on the gearbox. Figure 7.2a shows both the high speed (HS) gearmesh, and the intermediate speed (IS) gearmesh, while Figure 7.2b zooms on a range retaining the latter. A section between 190 and 290 seconds has been selected where the speed variation is limited (to about $\pm 2\%$). Figure 7.2b also shows four harmonics of the HS shaft speed (HSS), with the fourth being quite close to the IS gearmesh.

Figure 7.3 shows FFT spectra of the selected section of signal in two frequency ranges. Figure 7.3a, b show the spectra of the raw signal up to 4000 Hz and 500 Hz, respectively. Figure 7.3c, d show the order tracked spectra in orders of the HS shaft (achieved through two stages of iteration [2]). According to this the HS gearmesh frequency is at order 20.

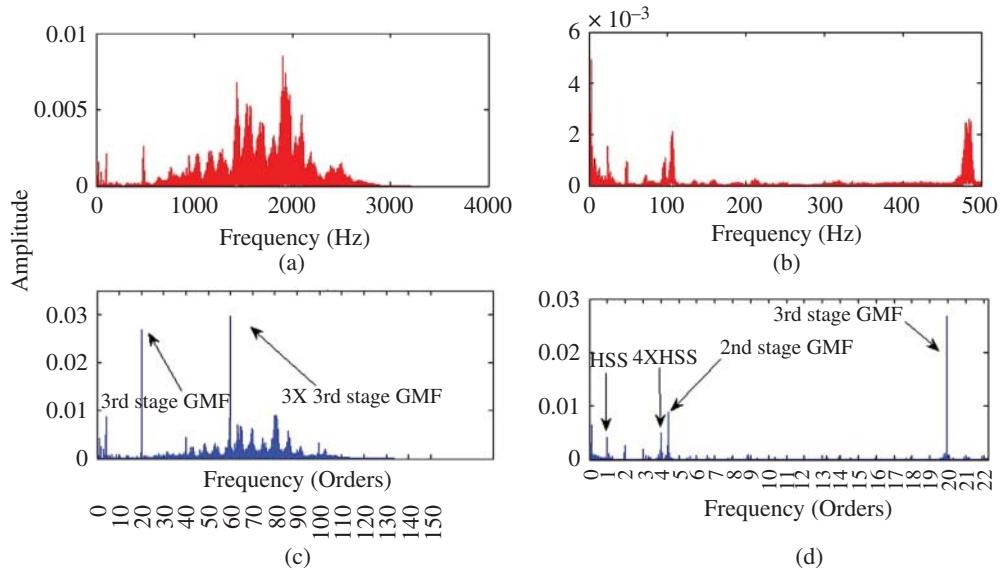


Figure 7.3 FFT spectra of the raw and order tracked signals: (a) Raw signal (0–3500 Hz). (b) Order tracked raw signal (corresponding 0–500 Hz in orders of the HSS). (c) Raw (0–500 Hz). (d) Order tracked (corresponding 0–500 Hz in orders of the HSS).

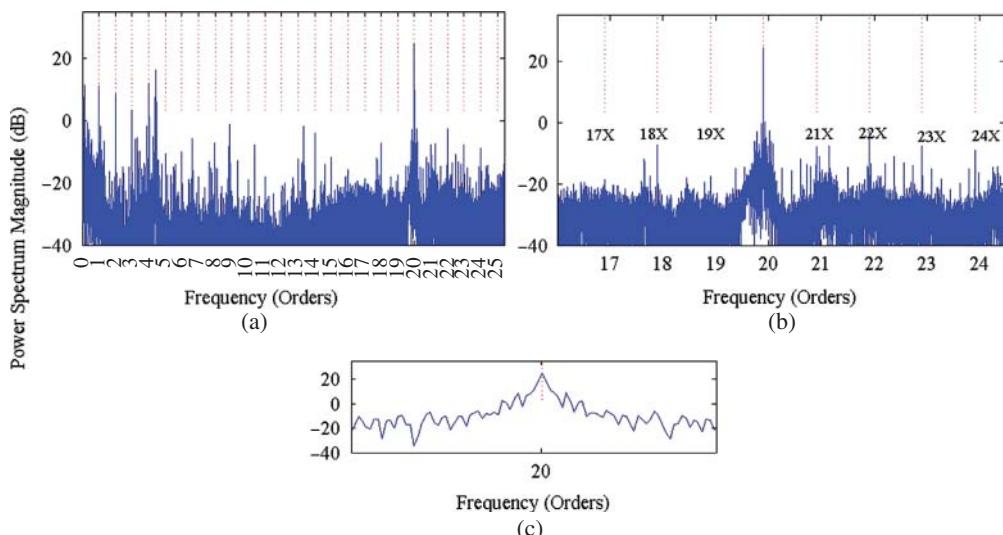


Figure 7.4 HSS harmonics at various levels of zoom (a) the first 25 harmonics (b) A zoom showing harmonics 17–24 (c) A zoom around the 20th HS harmonic (3rd stage GMF).

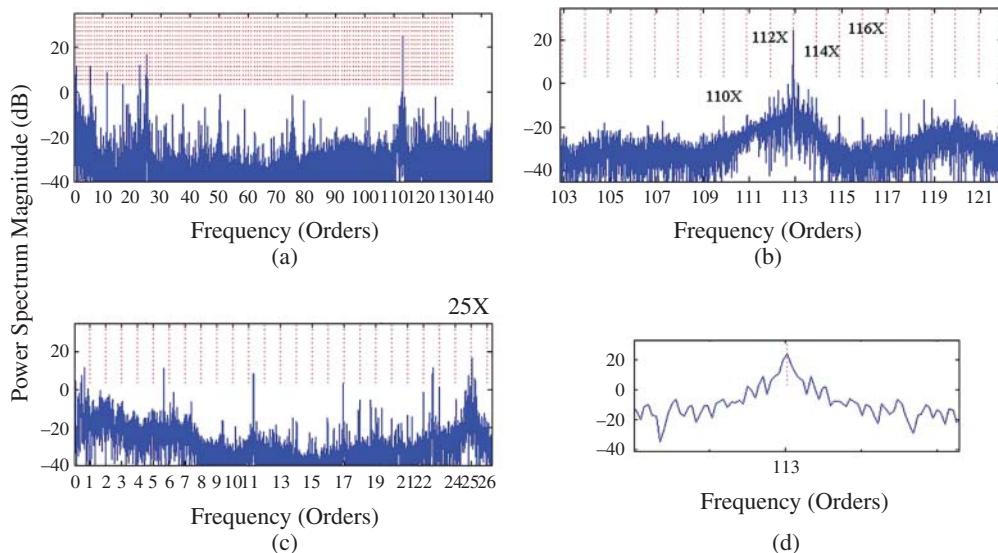


Figure 7.5 ISS harmonics at various levels of zoom (a) the first 130 harmonics (b) A zoom showing harmonics 103–122 showing the HS GM frequency at order 113 (c) A zoom on the first 26 IS harmonics, showing the IS GM frequency at order 25 (d) A zoom around the 113th IS harmonic.

This is confirmed in Figure 7.4, where a finely tuned harmonic cursor is shown to align with the harmonic pattern even with very high levels of zoom. The harmonic cursor was in fact aligned with the third harmonic of HS gearmesh (which can be seen in Figure 7.3b).

The harmonic cursor was then adjusted to a spacing corresponding to the IS shaft speed, and the result is shown in Figure 7.5. Figure 7.5b shows that the harmonic series coincides exactly with the 113th harmonic of the IS shaft, confirmed by Figure 7.5d, and Figure 7.5c that it also coincides exactly with the 25th harmonic of the same shaft (the IS gearmesh).

In Table 1 of [1] (not reproduced here), the equivalent frequencies of the two different sets of harmonic orders of the HS and IS shafts are compared to establish where they coincide. Of course, the actual frequency is varying, but comparisons were made in terms of the mean frequency of the HS shaft over the length of signal processed, this being found to be 23.60618104 Hz (to 10 significant figures). The 20th order of the HS shaft was found to be 472.12362080 Hz, and the 113th harmonic of the IS shaft 472.12362082 Hz, thus agreeing to 10 significant figures. This is unusually accurate, but obviously much closer than the closest separation for a hunting tooth pair of $1/(20 \times 113) = 0.00044$, showing that the tooth numbers do correspond to 20 and 113, and at the same time represent a hunting tooth pair.

In a similar manner, a harmonic cursor tuned to the harmonics of the sun gear speed, whose shaft contained the other gear contributing to the IS gearmesh frequency, found that the latter corresponded to the 71st order, not agreeing quite as well as for the HS mesh, but still several orders of magnitude better than the closest separation of $1/(25 \times 71) = 0.00056$, confirming the numbers of teeth in that mesh as 25 and 71, and at the same time that it was also a hunting tooth combination.

This example shows the diagnostic power given by a finely tuneable harmonic cursor in combination with finely tuneable order tracking (using multiple iterations).

7.2 Gear Diagnostics

The basic generation of vibrations by gears is described in Chapter 2, and includes the vibrations generated while in normal condition and those due to various faults. It is pointed out that the vibrations tend to be deterministic, since the same tooth profiles mesh in the same way each time. Randomness can enter the picture because of speed fluctuation, but then the order tracking procedures of Section 5.1 can be used to compensate for them, and should always be used as a precursor to time synchronous averaging (TSA). Note that this actually changes the ‘time’ axis to rotation angle, but the term TSA will be used in this book as long as the machine has nominally constant speed. Randomness can also occur because of random load (for example with a rock crusher or wind turbine) and then the situation becomes more difficult. Where the load varies with time, but only changes relatively slowly, it is often possible to make recordings when the load is between certain specified limits (perhaps best determined by trial and error) and then only make comparisons for the same load condition. Data for a number of different load ranges could be kept as references. The following discussion primarily covers the situation where compensation can be made for minor speed variations, and where the range of load variation is limited, but the case of varying speed and load is treated in Section 7.2.5.

7.2.1 Techniques Based on the TSA

Gear diagnostics has been studied over many years, and as early as the 1970s, Stewart [3] proposed a number of powerful diagnostic tools that have become the benchmarks for gear diagnostics, in particular for helicopter gearboxes, which represent a particularly complicated case. They were based on the TSA for each gear, and its spectrum, from which were extracted a number of ‘figures of merit’. Probably the most powerful of these was FM4, defined as the kurtosis of the ‘residual signal’ obtained by removing the regular garmeshing pattern, which tended to obscure local variations in the signal. Originally the residual signal was obtained by subtracting the known garmesh harmonics from the spectrum and returning to the time domain. It was later realised that this often left quite large modulation effects at the first and second harmonics of rotational speed, which were not related to local faults, and so one or two pairs of sidebands around each toothmesh harmonic might also be removed to obtain the residual signal.

More recently, it was suggested by Wang and Wong [4] of the Australian Defence Science and Technology (DST) Group that a more flexible way of removing the regular toothmesh pattern was by using linear prediction, as described in Section 5.3.2. Figure 7.6 [4] shows the improvement given by using linear prediction compared with their traditional method (note that since the disturbance has two periods along the record, the second order sidebands have probably not been removed).

Peter McFadden, formerly of DST Group and later at Oxford University, introduced a number of improvements to the methods based on TSA (as well as to the order tracking and TSA operations themselves, see Sections 5.1 and 5.3.1). A significant development was a method to obtain the TSA for the planet and sun gears in a planetary (epicyclic) gearbox. For transducers mounted externally on the casing (typically on or near the annulus gear), the signal is heavily weighted by the proximity of a particular planet to the measurement point. McFadden proposed to make use of this by windowing out a short section of signal corresponding to the passage of just one tooth past the transducer [5]. This is illustrated in Figure 7.7. The next time the same planet passes the transducer, another tooth on the planet gear (and on the sun gear) will be in mesh, but from the kinematics it is known exactly which ones, so the windowed signal can be allocated to a different ‘bin’ for each tooth and each gear. After a sufficient number of passages, not only every tooth on each planet gear and on the sun gear will be encountered, but will be encountered several times, so that averaging of the windowed

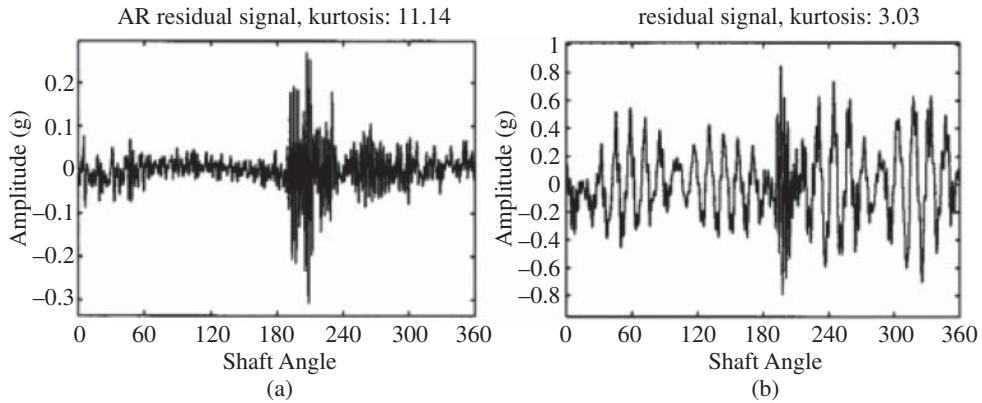


Figure 7.6 Comparison of residual signals using different methods [4] (a) Linear prediction (b) Editing spectrum.

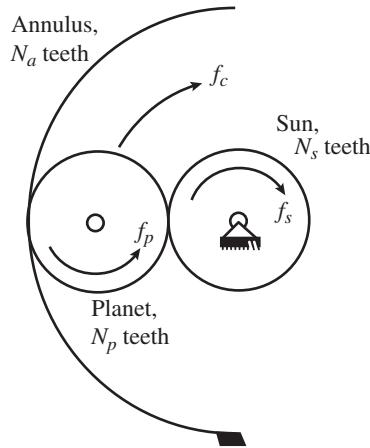


Figure 7.7 Illustration showing that as a planet gear passes a particular point on the annulus gear, a particular tooth is in contact with the annulus, and another with the sun gear [5].

sections for each tooth will finally give a (uniformly weighted) TSA signal for each tooth, and then by joining them together, for the whole gear. This applies to each of the planet gears, and for the sun gear (for which the signals for meshing with all of the planet gears can contribute to the TSA for each tooth). In the original paper [5], McFadden used a rectangular window corresponding exactly to the period of one toothmesh, but in later papers (e.g. [6]) he recommended a Hanning window (which he called a Tukey window in some papers) of total length twice the toothmesh period. There was little visible difference in the TSA waveforms, but a considerable improvement in the frequency spectra of the latter.

A considerable improvement of that procedure was made by Forrester and Blunt [7], also of DST Group, who recognised that it was not necessary to restrict the averaging to just one or two teeth each time, but to a whole series of consecutive teeth, always meshing in the same sequence. This

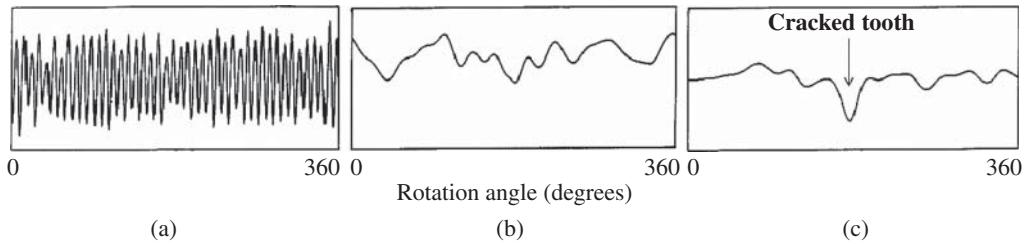


Figure 7.8 Demodulation of the second harmonic of the toothmesh frequency for a cracked gear [9]. (a) Time signal after TSA (b) Amplitude modulation signal (c) Phase modulation signal.

made the averaging process much faster, and did not require such long signals. The improved method was only published in a couple of conference papers (e.g. [7]), but was patented by DST, somewhat limiting access to the method.

It should be noted that the toothmesh frequency (and thus the toothmesh period) is the same for every meshing pair in a (single stage) planetary gearbox. For fixed annulus gear (the most common situation) it is equal to the planet carrier speed times the number of teeth on the annulus or ring gear. It is thus a harmonic of the planet carrier speed, but not of the sun gear speed. Note, however, that the meshing of the individual planet gears can either be synchronous or sequential, depending on the numbers of teeth on the various components. In the latter case this can be used, in conjunction with accurate phase measurements, to indicate the individual planet on which a fault is located, requiring much shorter records than the TSA method just described [8].

Another major contribution to gear diagnostics was the use of Hilbert transform techniques to demodulate the TSA signal [9] (with a harmonic of the gearmesh frequency as carrier) to expose local variations not visible to the eye (in particular phase modulation to which the eye is not sensitive). Figure 7.8 (from [9]) shows a typical example where the TSA signal for a cracked gear was demodulated. The gear actually failed 103 service hours after this signal was recorded, and the fault was even more evident before the final failure, but only using signal analysis techniques that were not developed at the time of the failure. In this case the second harmonic of the toothmesh frequency was demodulated because in terms of acceleration it was considerably stronger than the first, and thus gave a better signal/noise ratio.

It should be noted that this technique is limited by the bandwidth of the modulating signal, as the maximum bandwidth that can be demodulated corresponds to \pm half the toothmesh frequency, independent of which harmonic is demodulated. This is illustrated below in Section 7.2.2, where not only acceleration signals but also transmission error (TE) signals are demodulated, in order to see the interaction of amplitude and phase modulation for these two cases. In cases where the fault excites very high frequency resonances it is likely that the conditions for demodulation of the toothmesh signal will not be met. Such an example from a wind turbine gearbox is given in Section 7.2.5. That same example demonstrates that even synchronous averaging cannot be used in all cases, depending on the frequency range over which faults manifest themselves, as discussed in Section 7.2.5.

7.2.2 Transmission Error as a Diagnostic Tool

Gear diagnostics suffers to some extent from the modification of the signal from source to measurement point, and so a measurement right at the gearmesh can be valuable. Such a possibility is given by the measurement of transmission error (TE), which is possible if accurate shaft encoders can be

attached to the shafts on which the gears are mounted, in particular the free (non-drive) ends, which in any case are normally most accessible. If there is no drive torque in the shaft connecting the gear to the encoder, it will accurately follow the torsional vibration of the gear up to a very high frequency.

7.2.2.1 TE Measurement by Phase Demodulation and Pulse Timing

As described in Section 2.2.2, transmission error represents the difference between the angular motion of a driven gear and that which it would have if the transmission were perfectly conjugate, i.e. constant speed out for constant speed in. Though often measured in terms of angular deviation of one gear or the other, it only corresponds for the gear pair as a linear deviation along the common tangent to the base circles of the two gears (i.e. along the line of action), and thus must be scaled by the gear ratio to achieve a common basis for the two gears. It can thus be measured by measuring the torsional vibration of each gear, scaling one by the gear ratio to make it equivalent to the other in terms of linear motion, and then subtracting it from the other. It would usually be interpreted as output minus input, but could be expressed in terms of angular motion of either gear or as linear motion along the line of action. The scaling is purely of amplitudes, since the time axis is independent of which gear is being considered.

Figure 7.9 shows the generation of the TE signal from the torsional vibrations of two shafts in a gearbox with a simulated tooth root crack on one gear. The ratio is 1 : 1 (32 teeth on each gear) so that the torsional vibration signals did not have to be scaled before subtraction. The torsional vibration was measured by phase demodulation of the signal from a shaft encoder on each shaft using the procedure illustrated in Figure 3.30b. Note that the phase of each gear exhibits some random speed variation (the local slope of the phase curve), but the TE, as the difference, is very regular and periodic with the rotational speed. Note also the difference in scale of the TE compared with the

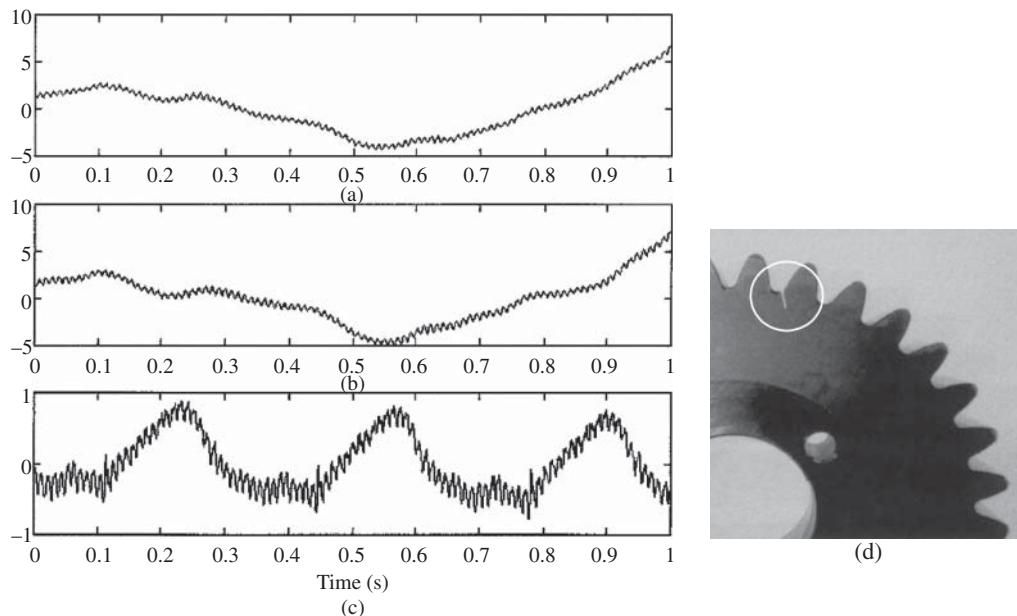


Figure 7.9 Transmission error for a gear with a simulated tooth root crack (a) Phase of Gear 1 (b) Phase of Gear 2 (c) TE by subtraction (d) Simulated crack.

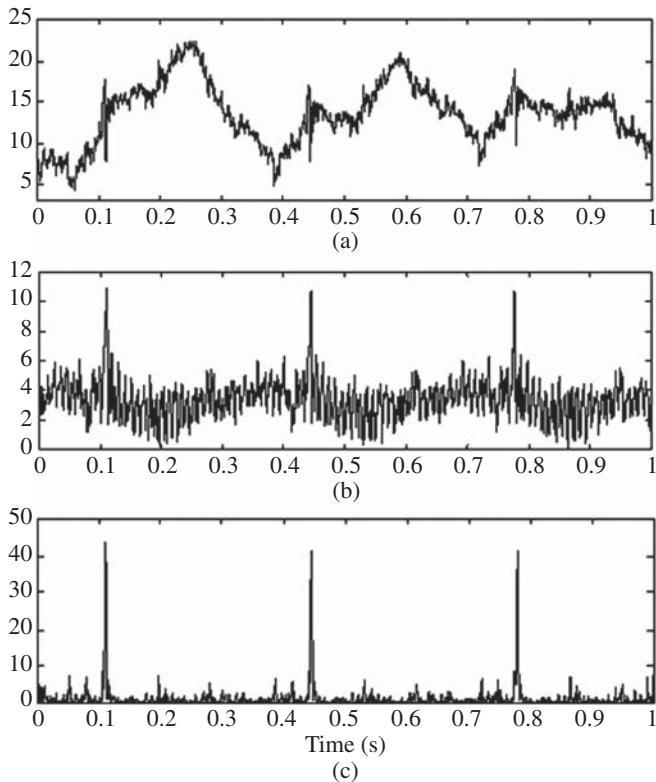


Figure 7.10 Enhancement of the fault indication from Figure 7.9c. (a) Forming residual by removing toothmesh harmonics (b) Removing low frequency drift by highpass filtration at half mesh frequency (c) Squared envelope after (a) and (b).

original torsional vibration signals. The speed is 3 Hz, so that there are three rotations in the one second records.

Note that the toothmesh signal is very clear (32 periods per revolution), and also that a local deviation is apparent (the first at about 0.11 second). In this case the local fault was apparent directly in the TE signal (this was not the case for all speeds and loads), but was considerably enhanced by three stages of signal processing, as shown in Figure 7.10.

This method of measuring the TE of gears was first published in [10] as part of the PhD thesis work of P.J. Sweeney. The phase demodulation of the encoder signals (and the frequency analysis of the result) was done using the zoom processor method described in Section 3.3.2.2, using a Brüel & Kjaer Analyzer Type 2035. In Figure 7.11, the results are compared with the ‘pulse timing’ method developed by Sweeney, using a 100 MHz clock to time the intervals between encoder pulses. This is based on the very simple principle that the time intervals between encoder pulses give a measure of the times for constant increments in phase angle (and can thus be used for angular resampling as an alternative to the order tracking methods described in Section 5.1). The reciprocal of the time intervals can be calibrated in terms of instantaneous angular speed and thus frequency modulation, and have the advantage that there are a fixed number of samples per rotation, so that order tracking is not required. This method of measuring torsional vibration can also be valuable for reciprocating machines and so is mentioned again in Section 7.4. At the time of Sweeney’s thesis work it was very

difficult to use such a high frequency clock, and phase demodulation was much more practical, but commercial data acquisition systems are now available with an 80 MHz clock, and have made the pulse timing technique much more viable.

It will be noted that the results obtained by the two methods are virtually identical. The precision with which the measurements are made will be found impressive, since it is primarily based on the

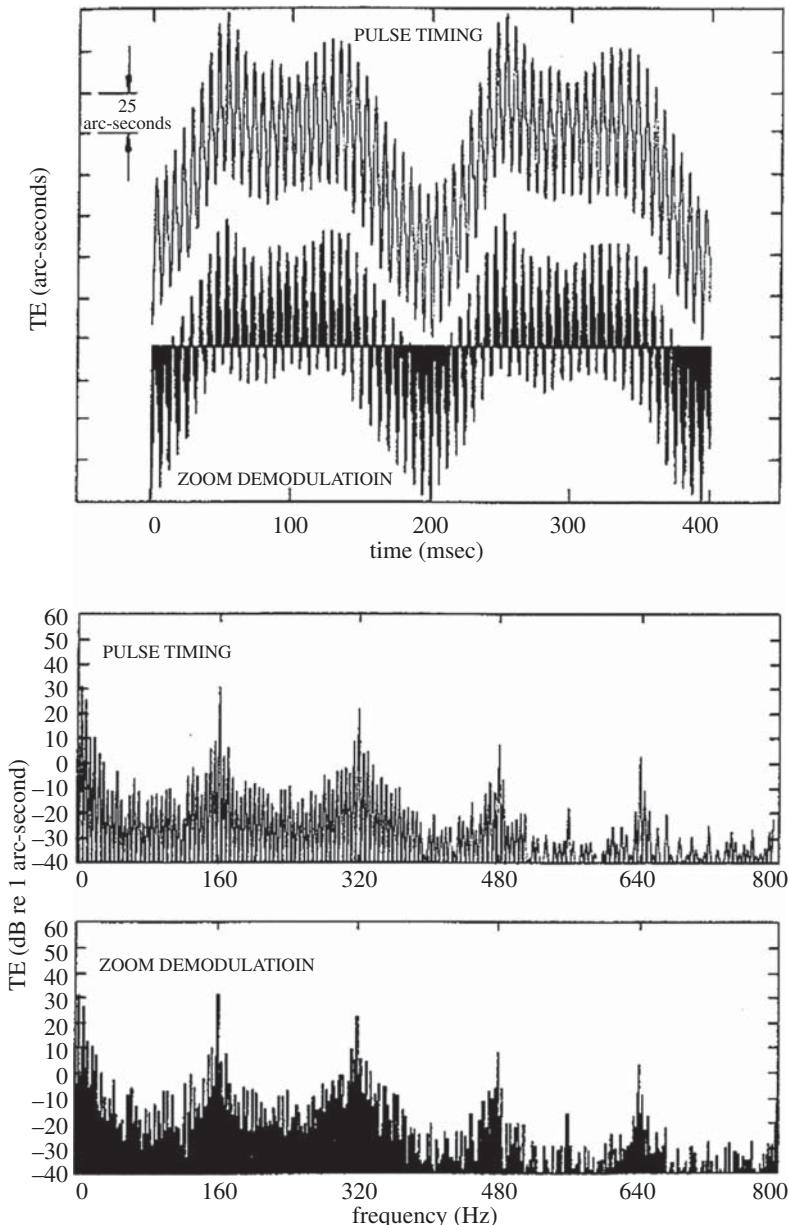


Figure 7.11 Measurement of gear transmission error by phase and frequency demodulation of shaft encoder signals from each gear by zoom demodulation and pulse timing [10]. (Upper) Time signals (Lower) Frequency spectra.

accuracy of the sampling and timing clocks involved, and is not affected by the accuracy of analogue instrumentation (as are phase meters previously used for gear TE measurement). The encoders used here, Heidenhain type ROD260, have a high and documented accuracy, with errors of the order of 3–5 arcseconds. However, from the results it appears that even these small errors are concentrated at the low harmonics of rotation (probably dominated by minor eccentricity of the inscribed disc) and are even lower at the higher harmonics that would correspond to garmesh frequencies. In this particular case, the measured low harmonics of the TE, and the value at toothmesh frequency are of the order of 50 arcseconds, each arcsecond being equivalent to about 0.025 μm of linear motion at the pitch circle.

Shaft encoders with documented error are expensive, so Shu Du [11] investigated ways to measure and correct for encoder error, in particular for cheaper encoders. In [11] it is explained that by making two measurements, and swapping the two encoders between them, it was possible by taking the difference to obtain a measure of the combined encoder error, and the mean of the two results gave the best estimate of the true result. This simple procedure applies to gears with 1 : 1 ratio, so that the result is approximately the same for the two measurements and the error given directly.

Figure 7.12 shows a typical result for 1 : 1 ratio pairs of spur and helical gears on the same test rig (32 and 29 teeth respectively). It was ensured that the alignment of the two encoders was the same for the starting position of both sets of measurements. Two revolutions are shown. Note that the scaling of the TE is 10 times greater than that of the error. It is seen that even though the TE at the toothmesh frequency for the helical gears is much less than for the spur gears (although the low frequency components due to tooth spacing error etc. are much the same), the encoder error is virtually the same in the two measurements, thus helping to validate the method. Further validation is given by the results shown in Figure 7.13, for measurements on a spur gear pair of non 1 : 1 ratio (32 : 49). In this case, the two encoders are rotating at different speeds in the two measurements where they are swapped, and the difference of the two measurements contains two versions of the combined encoder error, one for the rotation period of each gear. It can however be extracted by synchronous averaging with respect to the period for one gear, and this is shown in Figure 7.13. Note that the fundamental period of the TE would now correspond to 49 revolutions of the 32 tooth gear (only two shown). The encoder error extracted by synchronous averaging with respect to the 32 tooth gear is very similar to that shown in Figure 7.12, though smoothed by the averaging. Its spectrum shows that the error is concentrated in the low harmonics of the rotational speed, and is not much greater than for the more expensive encoders.

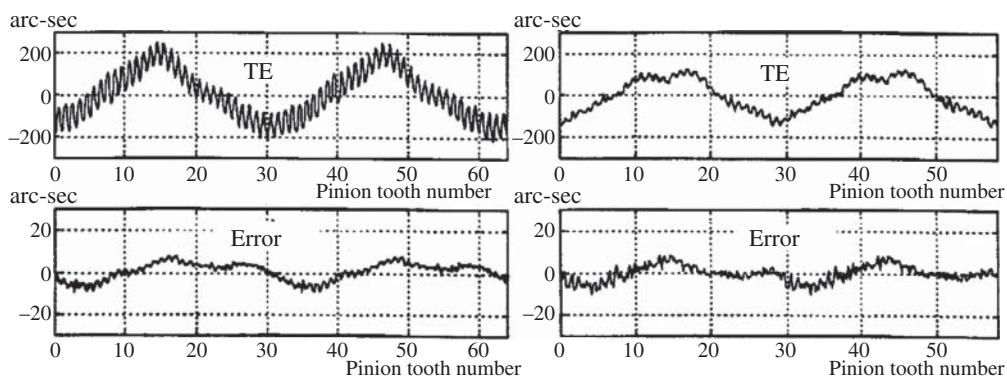


Figure 7.12 Measurement of TE and combined encoder error for 1 : 1 ratio gears [11] (Left) Spur gears (32 teeth) (Right) Helical gears (29 teeth).

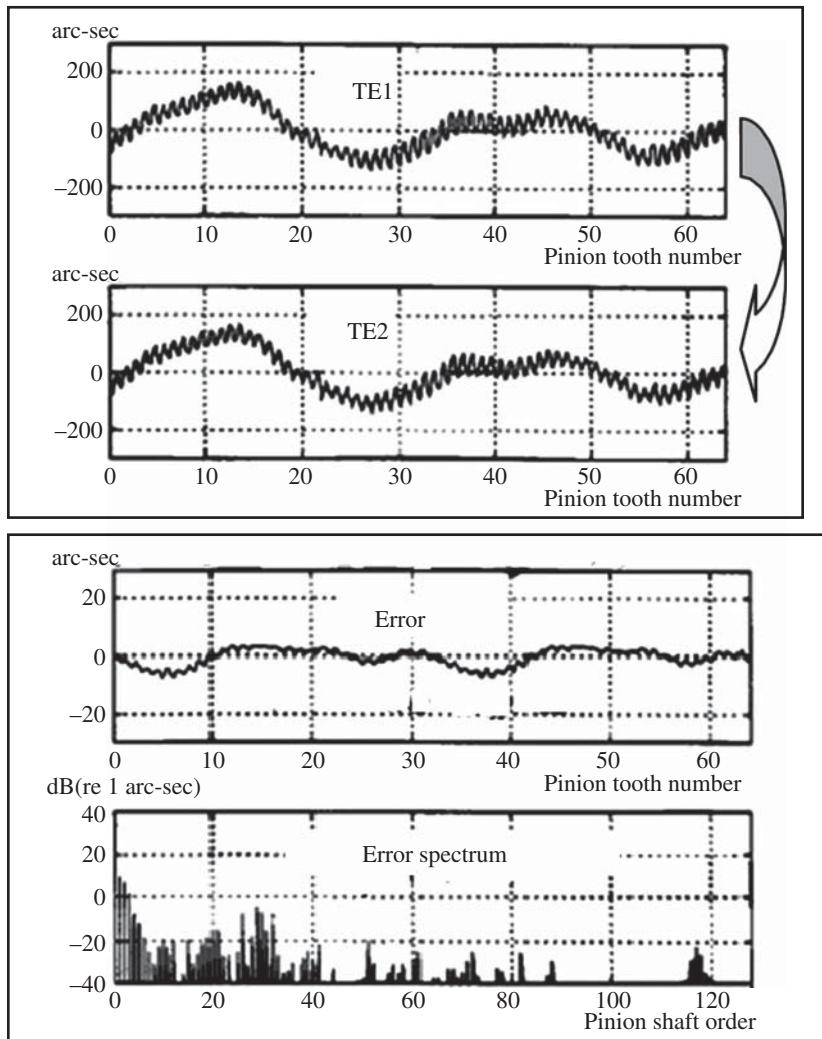


Figure 7.13 Measurement of encoder error for gear ratio 32 : 49 [11]. (Upper) TEs measured with swapped encoders for two rotations of the 32 tooth gear. (Lower) Time record and its spectrum of the combined encoder error, extracted by synchronous averaging with respect to the 32 tooth gear.

The errors measured with the lower cost encoders (Heidenhain ROD426) were sufficiently small that Du did not find it necessary to compensate for them in the presentation of his thesis work, some of which was reported in [12].

It is interesting to compare the results of demodulating TE with those from acceleration signals measured at the same time. Figure 7.14 shows the results for the TE on the same gears as Figure 7.9. It is seen that the demodulated amplitude shows the fault very clearly, whereas the phase does not. Figure 7.15 shows the equivalent results for demodulation of the acceleration signal (but includes the demodulated amplitude of the TE for comparison). It is evident that for the acceleration it is the phase modulation signal that carries more information about the crack. It can be speculated that this can partly be explained by the fact that the amplitude of TE directly gives variations in torque,

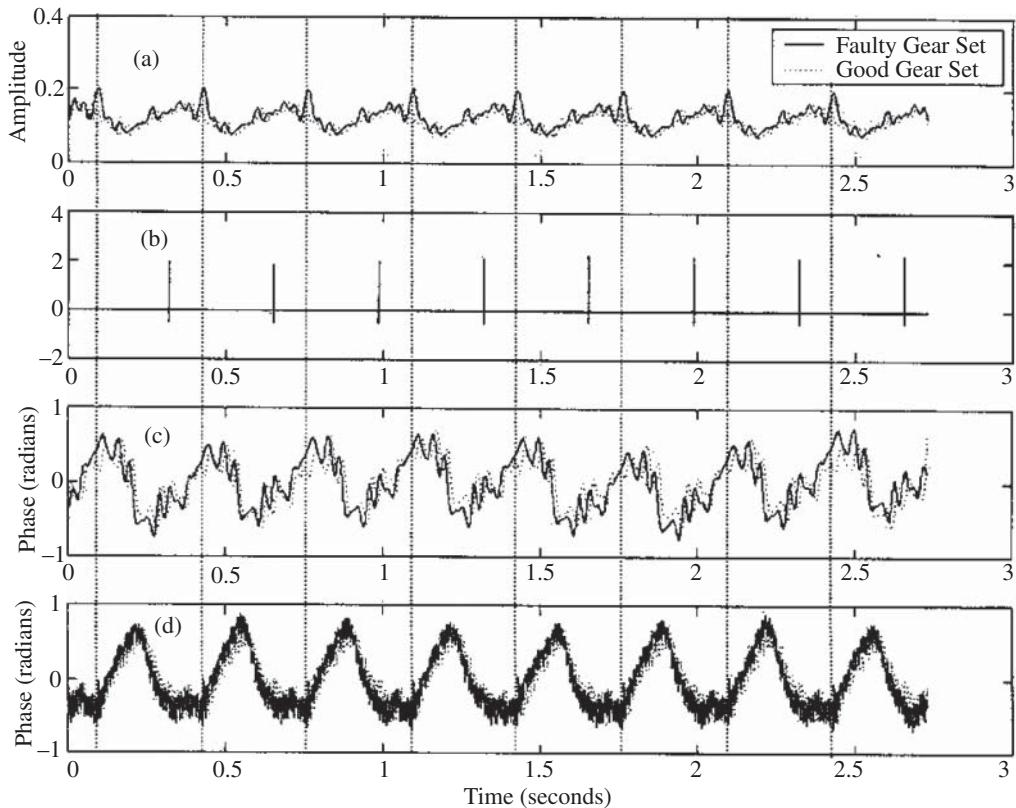


Figure 7.14 Results of demodulation of the TE signal for a simulated tooth root crack (a) Demodulated amplitude (b) Tacho signal (c) Demodulated phase (d) TE signal.

which might explain changes in phase modulation, but it should be kept in mind that the acceleration response will depend on the transfer function between force at the mesh and acceleration at the measurement point.

Figure 7.16 shows a simple example where a pure amplitude modulation at the source can be changed by a transfer function into a mix of amplitude and phase modulation. It is based on the fact that as shown in Figure 3.29, the primary difference between amplitude and phase/frequency modulation is in the phase relationships of the sidebands on either side of the carrier. The phase variation of the transfer function (not shown in Figure 7.16) would add even more to this change between amplitude and phase modulation.

Thus, one possible advantage of using TE as a diagnostic parameter is that it would allow more to be inferred from the division between amplitude and phase modulation.

It is illustrative to investigate how the spectra are demodulated around one of the harmonics of garmesh for the data shown in Figures 7.14 and 7.15.

Figure 7.17 shows the spectra for the TE and Acceleration signals. For the TE signal of Figure 7.17a the sidebands overlap to some extent outside the region of 0.5–1.5 times the toothmesh (TM) frequency which has been demodulated, and this will result in a small amount of distortion. As shown in Section 5.1.3, the non overlap region is actually 0.7–1.3 times TM frequency, but here the second harmonic is much smaller than the first. For the acceleration signal of Figure 7.17b, the

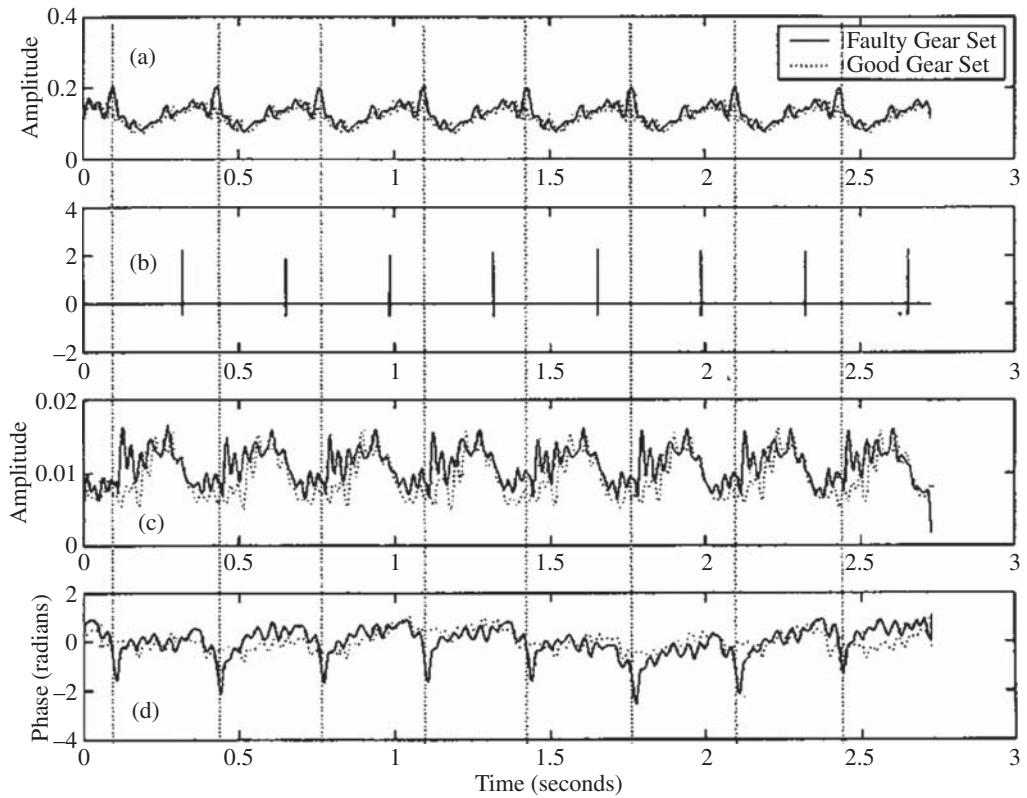


Figure 7.15 Comparison of demodulation results for TE and acceleration signals (a) Amplitude demodulated TE (b) Tacho signal (c) Amplitude demodulated acceleration (d) Phase demodulated acceleration.

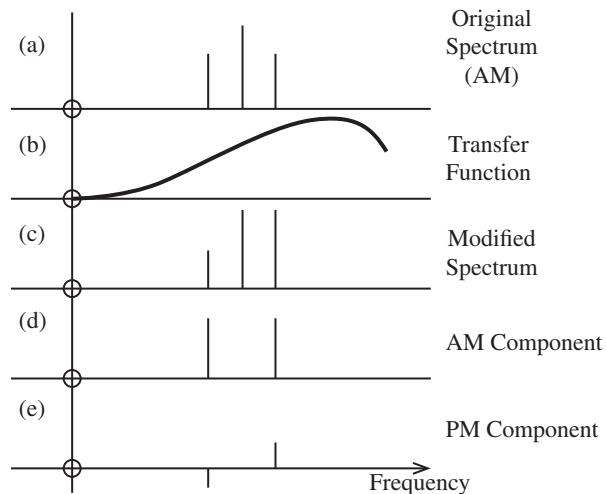


Figure 7.16 Illustration of how a transfer function can change the distribution of amplitude and phase modulation.

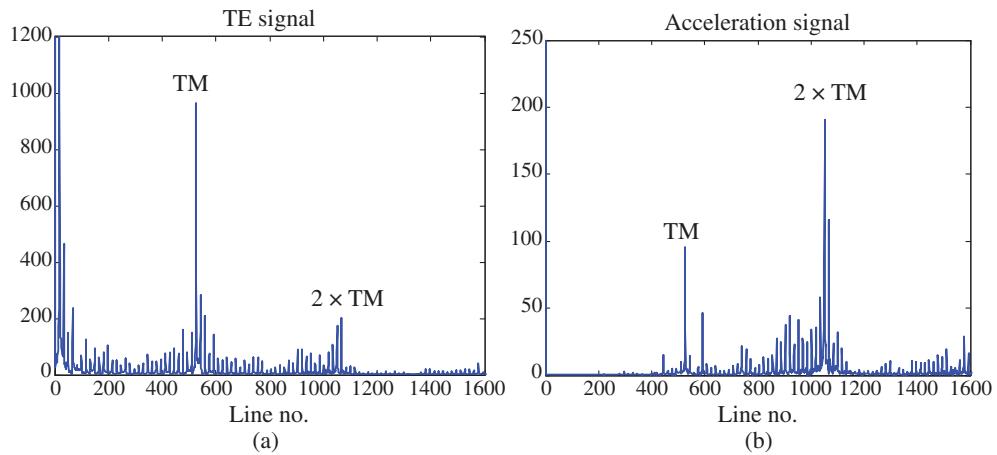


Figure 7.17 Spectra of (a) the TE signal and (b) the Acceleration signal.

same applies to the sidebands around the second harmonic of toothmesh (range 1.33–2.66 times TM) but 1.5–2.5 times TM was demodulated in this case. Since the acceleration signal had been truncated to the same length (in seconds) as the TE signal, the TM frequency corresponded to line number 526 in both cases.

At the time of writing the first edition of this book, it was thought that the use of TE for diagnostics would be mainly of interest for laboratory investigations, but with the development of the Internet of Things, it is now much more likely that shaft encoders will be built into commercial machines, because of their contributions to operations and control as well as condition monitoring. This is particularly the case for variable speed and remotely operated machines such as wind turbines.

7.2.2.2 Advantages of TE as a Diagnostic Tool

Further developments in the use of TE as a diagnostic tool have recently been made, including the measurement of the three main types of TE, geometric transmission error (GTE), determined by geometric deviations of both mating teeth, by measuring at low speed and low load, static transmission error (STE), which includes the elastic deflections of the tooth pair, by measuring at low speed and higher load, and dynamic transmission error (DTE), where inertial and damping effects change the dynamic tooth loads, by measuring at higher speed and higher load. This has opened up new areas of gear diagnostics, including the measurement of total tooth wear, not possible with vibration measurements. Four of these new developments are briefly described in the following.

7.2.2.2.1 Development of Uniform Wear and Pitting

As originally published in [13], a series of tests were done on mild steel gears, which were run for a long period (nearly 50 hours) during which time they developed surface pitting fairly uniformly distributed around the gears. Results are shown for the pinion, which had more wear because of the smaller number of teeth, and were obtained from the overall TE using synchronous averaging with respect to the pinion.

Figure 7.18 shows measurements of TE at no load (GTE) and 20 Nm pinion load (STE), for increasing wear with time. DTE could not be measured, as the encoders used were included in slip

rings, and had a low torsional resonance frequency. On the left are shown images of the worn surface of a typical tooth, taken from moulds of the surface in situ, as described in [14]. The first three rows (0–6 h) are for the period where the pitting did not extend across the whole tooth face along the line of contact, and thus did not greatly affect the basic involute profile of the tooth. It is interesting that even though the TE increases monotonically with wear, it decreases with load during the initial period, but increases slightly with load after that. This was initially puzzling, but can be explained as shown in Figure 7.19 (figure courtesy of P. Borghesani).

GTE is very sensitive to ‘high-spots’, which may be easy to deform under relatively light load. It is suspected that these have distorted what may be considered as the true GTE (for a relatively light load) in the two cases illustrated; first where the load makes the tooth contact closer to that between involute surfaces, and after some wear allows the mating tooth to penetrate more into deviations from the involute surface. This point is taken up again in Chapter 8, where the modelling of GTE in series with a spring representing the gearmesh is discussed.

7.2.2.2 *Tooth Root Crack on One Tooth*

A quite different case was described in [15], with measurements made on basically the same test rig, but with hardened and ground gears, with different numbers of teeth. A simulated tooth root crack, at 45° and extending to the tooth centreline (50%), was seeded by EDM machining at the root of one tooth of the pinion. Another major difference was that the original slip rings were replaced by high quality encoders (Heidenhain type ROD426) so that DTE could be measured up to a pinion speed of at least 20 Hz. Measurements were made at pinion speeds varying from 2 to 20 Hz and loads from zero to 20 Nm. Figure 7.20 displays an extract from the results for two rotations of the pinion. It is scaled directly in μm . A prominent feature of the unfiltered measurements (a–d) is a strong component at the rotational speed of the pinion, which is consistent in all measurements, and presumed to be due to an eccentricity of the pinion, so did not give much diagnostic information. Another feature is that the component at the gearmesh (GM) frequency (27 teeth) grows with load at both low and high speed, but is considerably stronger at 20 Hz. This was found to be due to the fact that the GM frequency was close to a resonance (540 Hz) at 20 Hz shaft speed, and it can be seen that the increased values are dominated by the first harmonic of the gearmesh. It was found, however, that the most important information about the cracked tooth was contained in additive components, below the GM frequency, and not in modulation effects, or excitation of the resonance. The TE was therefore filtered to include only the shaft harmonics from 3 to 13 (just below half the GM frequency) to exclude modulation sidebands, and the eccentricity, and these are the results plotted in the lower half of the figure (e–h). The effect of the cracked tooth could now be seen, but only for low loads (only the lowest shown here). This was initially puzzling, but it was realised that it resulted from an initial GTE that was counteracted by the tooth deflection under increasing load. It meant that the tooth spacing had actually closed as a result of machining the crack, presumably because of the relief of residual stresses from machining a slot in the hardened gear. This is the opposite of what is expected in the case of an actual (natural) crack because, as described in Section 7.2.4, and Ref. [16], it is common for the tooth spacing to increase at the location of a crack, because there is always plastic deformation at the crack tip, giving a permanent tilt of the tooth. Loading the tooth would then increase the STE from the initial GTE.

This theory was largely confirmed by examining in detail the differential TE at the location of the peak as a function of the load, and the results are shown in Figure 7.21a for the 2 Hz low speed case. The peak value was taken as the location where the maximum effect of the cracked tooth was experienced, i.e. with a single tooth pair in mesh including the cracked tooth. Figure 7.21b shows the corresponding differential deflection vs load at this point, for which the differential compliance of the mating pair is given by the slope of the deflection vs load regression line.

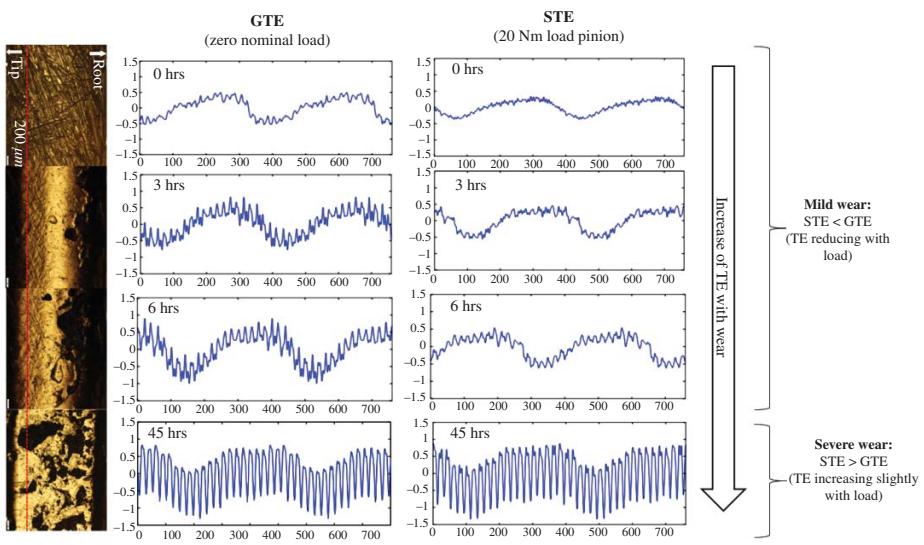


Figure 7.18 Variation of TE with wear and load for case of tooth pitting.

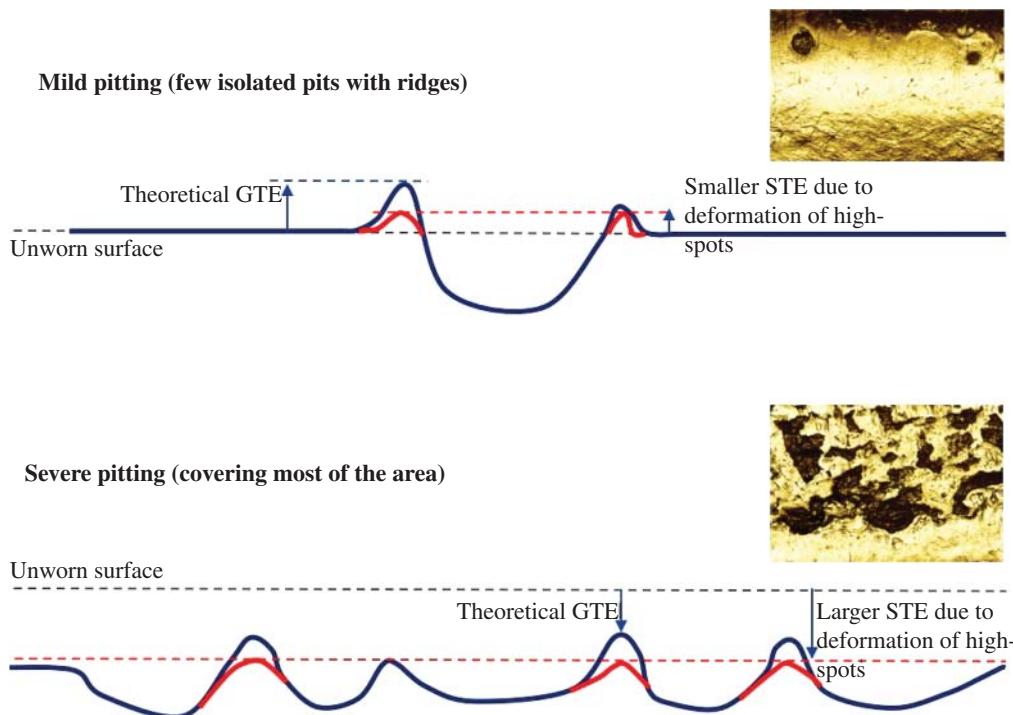


Figure 7.19 Effect of ‘high-spots’ on TE under load with mild and severe pitting.

A rough calculation showed that this estimate was very much in line with the expected change in stiffness as a result of a normal tooth meshing with one with a 50% tooth root crack (i.e. extending to the tooth centreline) for which a 33% increase in compliance had been predicted by FEM analysis. Taking the standard mesh stiffness of a steel tooth pair as $1.4 \times 10^{10} (\text{N m}^{-1}) \text{ m}^{-1}$ [17], this corresponds for these gears to a compliance of $14 \mu\text{m (kN)}^{-1}$, of which 33% is $4.7 \mu\text{m (kN)}^{-1}$, very close to the measured $5.2 \mu\text{m (kN)}^{-1}$, in particular considering that it is a slot, not a crack.

Ref. [15] shows that even the DTE results at 20 Hz, corresponding to Figure 7.20f,h, gave a very similar result with a differential compliance of $6.9 \mu\text{m (kN)}^{-1}$, showing that TE is much less sensitive than vibration to operating parameters such as speed and resonance characteristics, and can give a correctly scaled measure of the actual deflection.

7.2.2.2.3 Measurement of Absolute Wear

If a phase reference is available for each of the gears in a meshing pair, it is shown in [18] that it is possible to measure the ‘absolute TE’ (i.e. total change from an original reference), which can give a measure of total relative angular displacement between the gears, and thus the total wear (removal of material from the tooth pair), see Figure 7.22.

The paper explains how it is possible to obtain a common phase reference for the mating of two teeth (one on each gear) by making use of the known ‘hunting tooth period’ (HTP), after which these two teeth re-enter into mesh. This is found by comparing the relative positions of the two sets of tacho pulses, one from each gear, for all possible rotations of the driven gear, within an HTP, after order tracking of both signals with respect to the driving gear. Even though such order tracking ensures

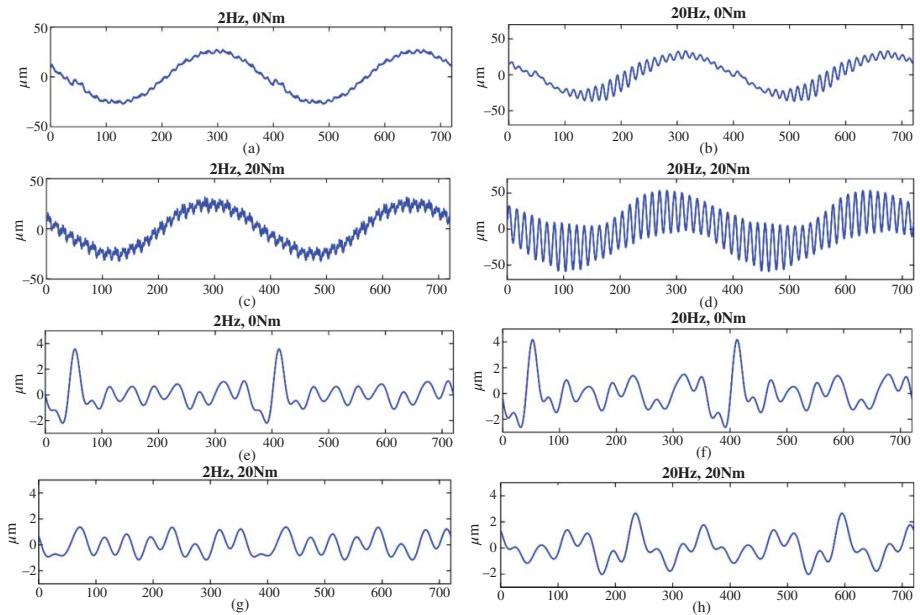


Figure 7.20 Original and filtered TE for a range of speeds and loads (shown). (a–d) Original (e–h) Filtered (Left) 2 Hz (Right) 20 Hz.

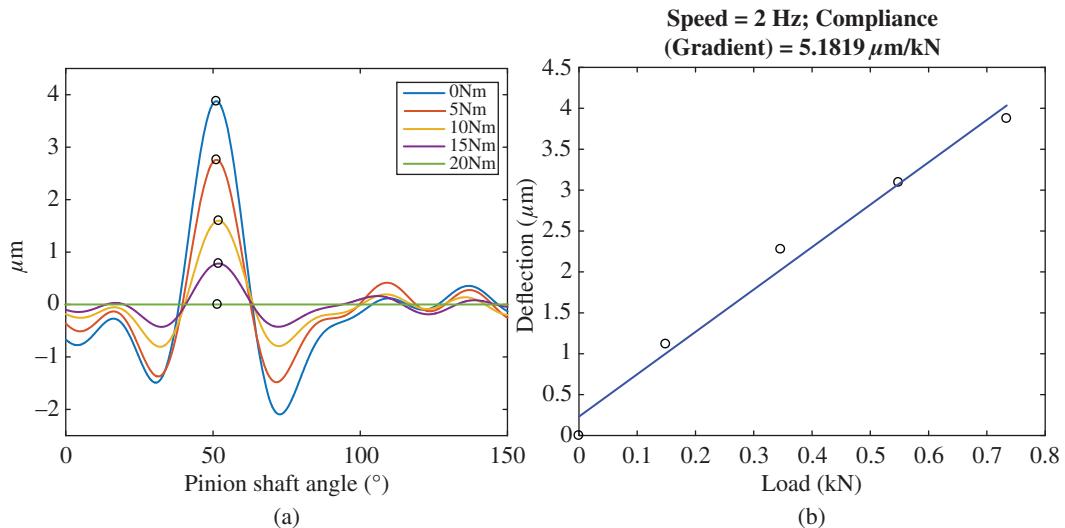


Figure 7.21 (a) Differential TE with increasing load at 2 Hz (b) Corresponding differential deflection vs load, and mesh compliance.

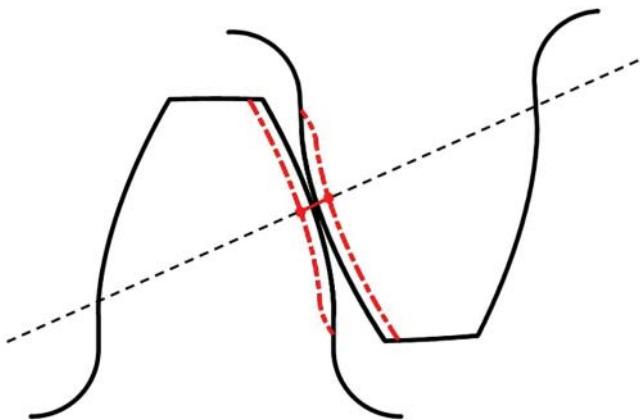


Figure 7.22 Illustration of wear depth as absolute geometric transmission error.

that the signals are phased consistently with respect to the driving gear, the angle of the driven gear at the beginning of the signal is still arbitrarily shifted in the HTP. To obtain the DC transmission error, both healthy and worn cases must start at the same point in the HTP so that the signals are phased consistently with respect to *both* shafts. The correct phasing corresponds to where the displacement due to wear is small, obviously much less than the tooth pitch, whereas all other possibilities would be spaced by a multiple of the tooth pitch. A circular shift of each HTP record can then align them to give a consistent starting phase for each gear.

In fact, the loss of material through wear can be estimated reasonably accurately using only a once-per-rev tacho pulse from each gear, but if shaft encoder signals are also available, the total or absolute TE can be measured for each tooth pair, including the variations around the shifted mean

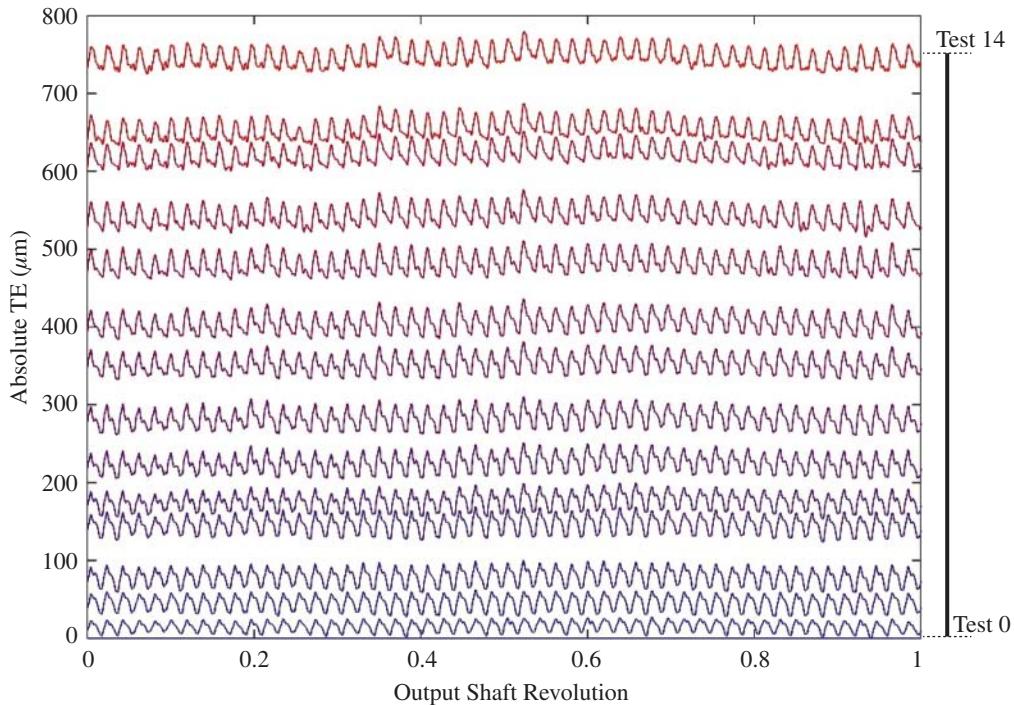


Figure 7.23 Absolute TE (low speed, low load) at different wear stages with respect to healthy gear condition (Test 0-14).

involute curve for each tooth (Figure 7.22). Note that these variations are the only part of the TE that gives rise to vibrations, which cannot therefore give any information about total wear.

Ref. [18] used measurements from a dry wear case on the same test rig as previously, giving very severe wear over a relatively short period, but for this reason restricted to a low maximum load (5 Nm) and a low maximum speed (10 Hz). Figure 7.23 shows the development of absolute TE with time. Ideally, the reference record should correspond to zero or minimal wear, though in this case the rig had been run for 20 minutes before the first record was taken. It could be argued that a reference after initial run-in might be advantageous, in order to remove local high spots, which can distort GTE.

The absolute wear amount (with respect to the reference measurement) is indicated by the mean value of the TE, and in Figure 7.24 this is plotted against test number, for measurements of GTE, STE, and DTE, with speed/load combination [Hz, Nm] respectively [2, 0], [2, 5], and [20, 5]. It is seen that these are very close to each other, and this can be explained by the fact that for constant speed and load, the TE fluctuations around the mean should be much the same (though different for each operating condition), but not have a marked effect on the offset represented by the wear. This requires some qualification with respect to tooth deflection under load, as in this case the very severe wear depth (0.54 mm in 3.14 mm, on the pinion alone, and 0.22 mm on the gear, in the worst case) would have had an effect on the mesh stiffness (which is not normally the case for wear in industrial situations). However, there are two mitigating factors with respect to this test (and test

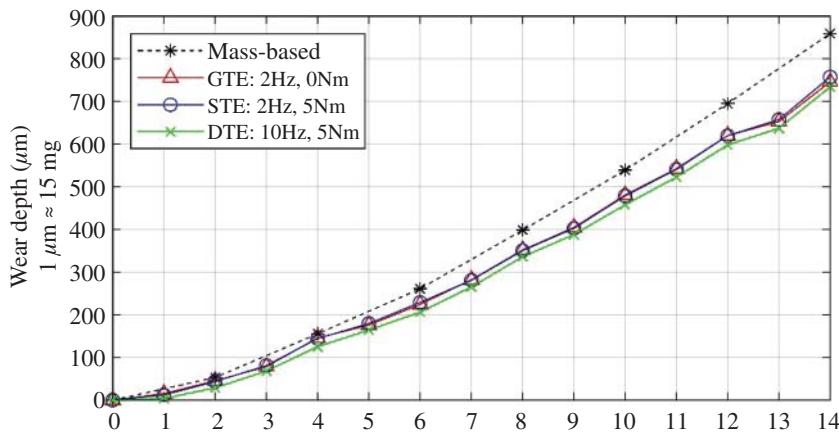


Figure 7.24 Comparison of combined average wear depth calculated using GTE, STE and DTE signals, and from mass loss of the gears.

rig); firstly, the TE is dominated by the unusual flexibility of the shafts, rather than the gearteeth, and secondly the 5 kN load is much lower than the rated load; finally, the maximum speed of 10 Hz does not excite resonant responses (as in Figure 7.20). Even so, it is expected that in the general case, the mean absolute TE would not be very sensitive to operating condition, as long as the measurements were always made at the same speed and load.

The validity of the estimates based on absolute TE is confirmed by a comparison with estimates based on measured mass loss of the gears. The accumulated values of the latter were obtained by weighing all wear particles collected on paper in the gearbox in the short stops after each test. The final value could be checked by measuring the actual mass loss of the gears at the end of the test series, and a discrepancy of 6% was found. This correction was then applied to all the earlier values (in Figure 7.24) as well as the final value.

The mass-based estimates are about 14% higher, but this can be explained by the fact that they assume uniform wear along the whole line of contact across the tooth face, and thus full metal-to-metal contact without gaps, but the worn tooth surfaces did exhibit waviness, which would have led to some gaps and thus reduced measured TE.

It is interesting that the final average wear depth of the pinion (0.54 mm) is greater than that of the gear (0.22 mm) in proportion to the number of contacts on individual teeth (i.e. in inverse proportion to the numbers of teeth, which was 19 : 52) as expected, and this justifies dividing up the total wear, as measured by the absolute TE, in this proportion.

7.2.2.4 Planetary Gear Set

Ref. [13] also included some results from a planetary gear set, where the encoders were not mounted directly on the input shaft (the planet carrier) and the output shaft (of the sun gear), but the input was actually measured on a shaft driving the planet carrier through a parallel stage. It was still possible to extract the TE of the planetary stage from that of the two stages, by using synchronous averaging over a number of meshes corresponding to the number of teeth on a planet gear. However, this would

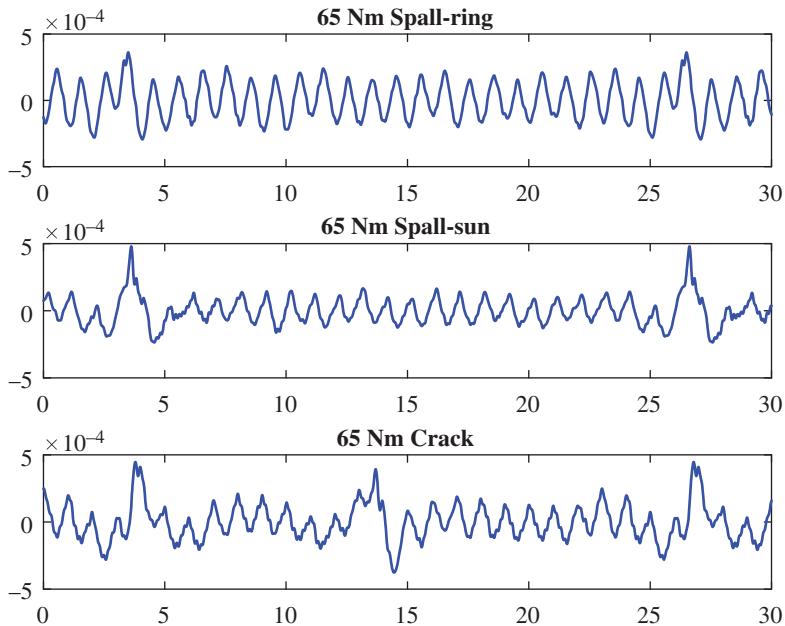


Figure 7.25 Measured TE for three different faults on one planet gear.

have been a composite of all three planet gears, since there is no way of separating the contributions of each of the planets. In this case, faults were only introduced on one planet gear, and Figure 7.25 shows the measured TE for three cases:

- 1) A simulated spall meshing with the ring gear.
- 2) A simulated spall meshing with the sun gear.
- 3) A simulated tooth root crack, whose effects are seen in meshing with both the ring and sun gears, in one case closing the crack, and in the other case opening it.

The original paper shows the results for two loads, but these were not very different and only those for 65 Nm are shown here. Moreover, in [13] the measured results were compared with simulations, but these comparisons are discussed in Section 8.2.4.

7.2.3 Cepstrum Analysis for Gear Diagnostics

There are three main areas of application of cepstrum analysis in gear diagnostics:

- 1) Collecting whole families of uniformly spaced harmonics and sidebands, as well as editing the cepstrum to remove one or more families.
- 2) Helping to separate forcing functions (at the toothmesh) from transfer functions to various measurement points.
- 3) Identifying echoes, and in particular inverted echo pairs, as well as the echo delay time.

Each of these will be treated in turn.

7.2.3.1 Collection of Harmonic and Sideband Families

In Section 2.2.2.1 it is pointed out that uniform wear tends to increase the harmonics of toothmesh frequency, initially the second but later all harmonics, and this can be seen directly in the spectrum. It is also pointed out in Section 2.2.2.2 that variations from the mean, either local or distributed, give changes in all other harmonics of the rotational frequency of the affected gear (e.g. Figure 2.13). Since these are distributed throughout the whole spectrum, and the separate families from each gear are mixed, as well as being modified by widely varying transfer functions over a broad frequency range, it becomes difficult to ‘see the trees for the woods’, when viewing the spectrum over a wide frequency range, or to ‘see the woods for the trees’ when zooming in on a narrow band. The cepstrum has the ability to collect all members of each family into a much more easily interpretable set of rahmonics of each family, of which the first is the most important, since it tells ‘how much on the average each family is protruding above the spectral noise level’. The higher rahmonics are affected by artefacts such as which window has been used in the frequency analysis.

Figure 7.26 illustrates very effectively the different information given by the spectrum and cepstrum for the case of a cement mill gearbox in worn and reconditioned form. After eight years operation, the gearbox was very worn and required repair. The latter consisted primarily in reversing the worn gears so as to make use of the unworn flanks, as well as replacing worn bearings and modifying the supporting structure. Since both sets of flanks were cut on the same machine at the

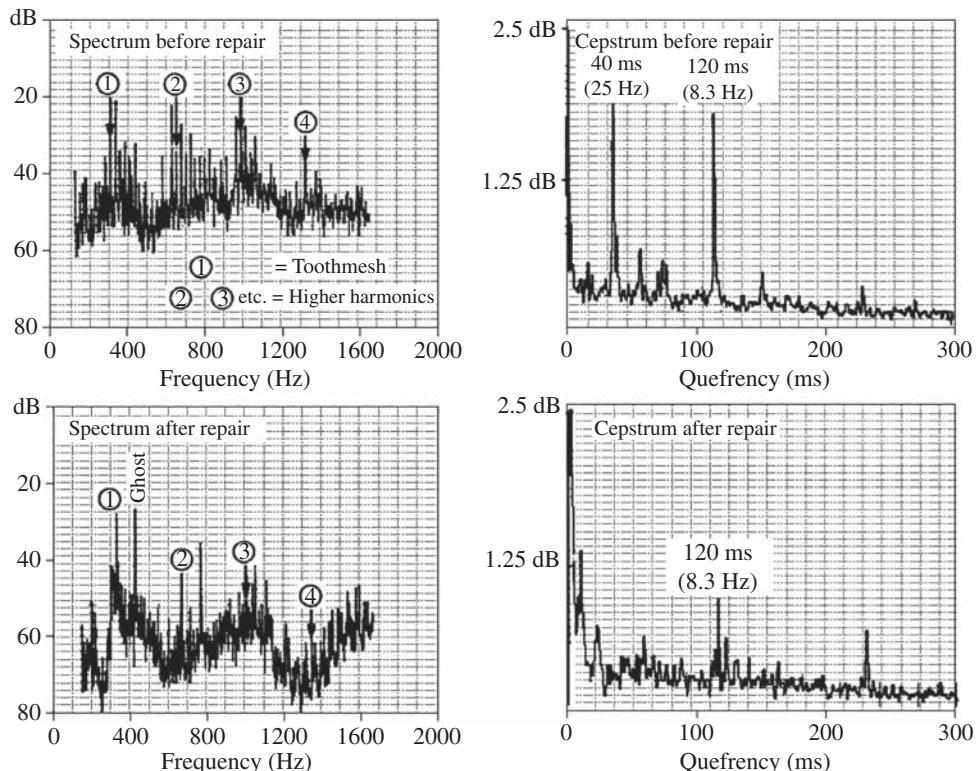


Figure 7.26 Comparison of spectra and cepstra for a worn and reconditioned gearbox. Source: Courtesy Brüel & Kjær.

same time it could be speculated that the spectrum ‘after repair’ would be very similar to the original when the gearbox was put into service, though unfortunately no recordings were made at that time.

If this assumption is made, it is seen that the higher harmonics of toothmesh frequency have all increased with wear (though some are lower than adjacent sidebands). The fact that some sidebands are higher than the carrier components (harmonics of toothmesh) does indicate a substantial amount of frequency modulation, as this is virtually impossible to achieve from pure amplitude modulation. Even though some sidebands are very prominent in the spectrum it is difficult to compare them with other possible families.

The cepstra, however, immediately reveal that dominant sidebands after repair are spaced at 8.3 Hz, which is the speed of the input pinion, and that before repair these are also present (at a higher level) but that there is also a very strong family at 25 Hz (these are the dominant sidebands seen around the toothmesh harmonics in the spectrum). The wear of the pinion was measured, and was found to have a ‘triangular’ pattern, giving a modulation three times per rev. The manufacturers suspected that this was due to a previous policy of lapping the gears for a long time in the workshop. This had likely excited a structural resonance close to 25 Hz and started the triangular wear pattern, which got worse in operation over time. In the repair, the unworn flanks were not lapped, and the structural modification very likely changed the resonance.

In any case, Figure 7.27 makes a similar comparison of spectra and cepstra four years after the repair. It can immediately be seen from the cepstrum that the sideband pattern has changed very little, and that the triangular wear pattern did not redevelop.

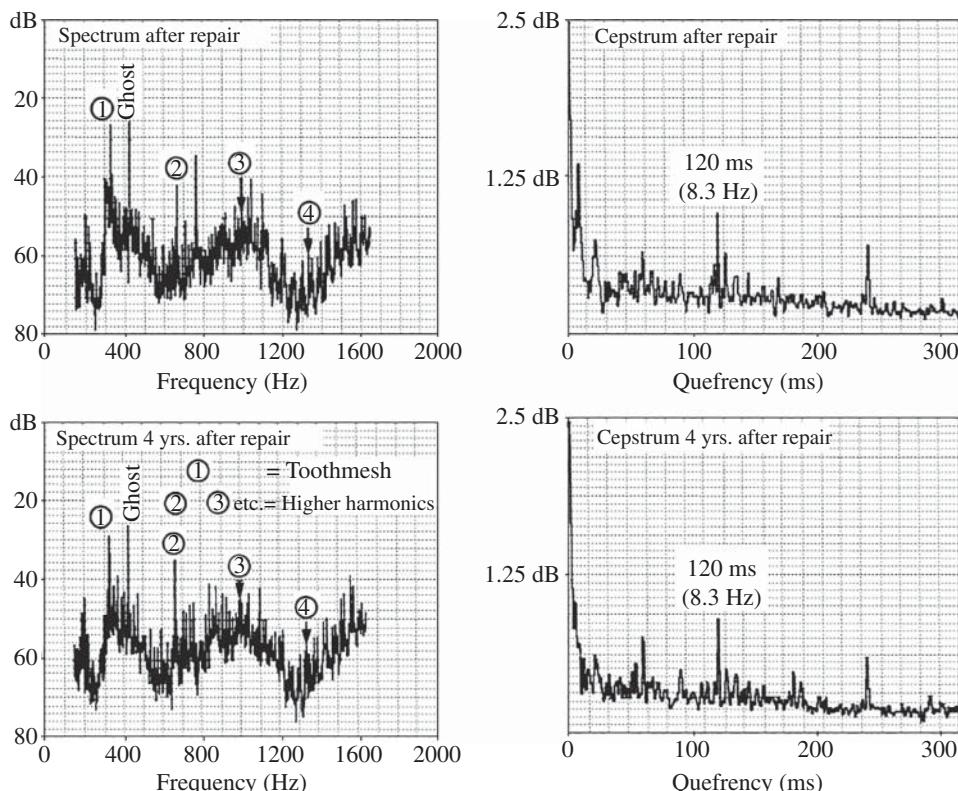


Figure 7.27 Comparison of spectra and cepstra just after and four years after repair. Source: Courtesy Brüel & Kjær.

On the other hand, from the spectra it can be seen that some uniform wear has occurred, because of the increase in the second harmonic of garmesh frequency. Note also that a secondary ghost component (with the same spacing from the second harmonic of toothmesh as the primary ghost component from the first) has also reduced, even though the primary ghost component has not changed.

Thus, it appears that the spectrum is best for detecting uniform faults in gears (such as uniform wear), whereas the cepstrum gives more information about non-uniform faults, as well as indicating the gear on which they are located.

Section 6.2.2 demonstrates that even with the same signal, there are advantages in generating spectra (and corresponding cepstra) in different frequency ranges, since these can give completely different information, in particular with multi-stage gearboxes, such as in helicopters and wind turbines.

Figure 7.28 illustrates another aspect of the use of the cepstrum to collect all members of a given family. The value of the quefrency corresponding to a particular family represents the average spacing of all members of the family, and is thus more accurate than measurements of individual spacings. Roughly the same accuracy could of course be achieved with a harmonic or sideband cursor (even higher for the former) but the cepstrum immediately gives an accurate value without having to find the individual members of the family. The gearboxes in this case were being tested at the end of the production line. One clearly showed a fault in both the spectrum and cepstrum. During the test, first gear was engaged, with input speed 35.6 Hz and output speed 5.4 Hz. The cepstrum indicates that the fault is producing harmonics with a spacing of 10.4 Hz, and this value is sufficiently accurate so as to eliminate the possibility that it is the second harmonic (10.8 Hz) of first gear speed. It did in fact correspond to second gear, which was turning, though not engaged under load. A local 'nick' on second gear was still producing impulses even though it was unloaded.

It should be noted that the version of the cepstrum to be used for this type of application is the amplitude of the 'analytic cepstrum', as described in Section 6.2.1, as this always shows the correct position of cepstrum peaks, independent of whether the uniformly spaced families pass through zero frequency, or are from zoom spectra. Thus, it can be used with edited spectra, to exclude components that are not relevant to the diagnostic problem at hand, or simply to reduce the size of the transform required to calculate the cepstrum.

Figure 7.29 illustrates how editing the spectrum before calculating the cepstrum can aid the diagnostic process. Figure 7.29a shows the spectrum for a single stage gearbox from zero frequency to about 1.5 times the toothmesh frequency, and the corresponding cepstrum. The latter is seen to contain rahmonics corresponding to both shaft speeds.

However, in (b) the spectrum is edited to remove components below about half the toothmesh frequency, and it is then seen that the cepstrum components corresponding to the 121 Hz shaft reduce dramatically. This is presumably because they were due to low harmonics of that shaft speed rather than modulation sidebands around the toothmesh frequency, and thus had nothing to do with gear condition. Figure 7.29c shows a similarly edited spectrum taken one month later, when the 121 Hz shaft alignment had changed. The corresponding rahmonics have reappeared in the cepstrum, and show that this gear is now modulating the garmesh frequency.

However, if the cepstrum is to be edited, for example to remove one family of sidebands or harmonics, then the editing should be carried out on the complex values of the analytic cepstrum, so that the forward transform to the (one-sided) log spectrum can be carried out.

This is illustrated in Figure 7.30, where the harmonics of 50 Hz have been removed from the whole spectrum by nullifying the corresponding component in the cepstrum (as well as a small number of lines around the actual peak value). Note that even though the cepstrum does not give any information about the distribution of sidebands, the edited spectrum may be useful for judging the distribution

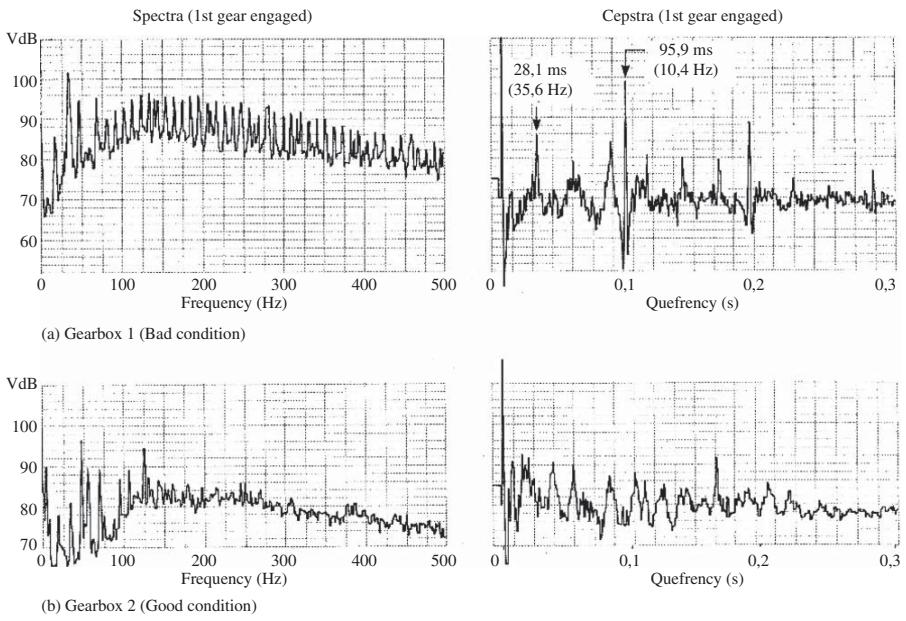


Figure 7.28 Spectra and cepstra for two truck gearboxes, one with a fault. Source: Courtesy Brüel & Kjær.

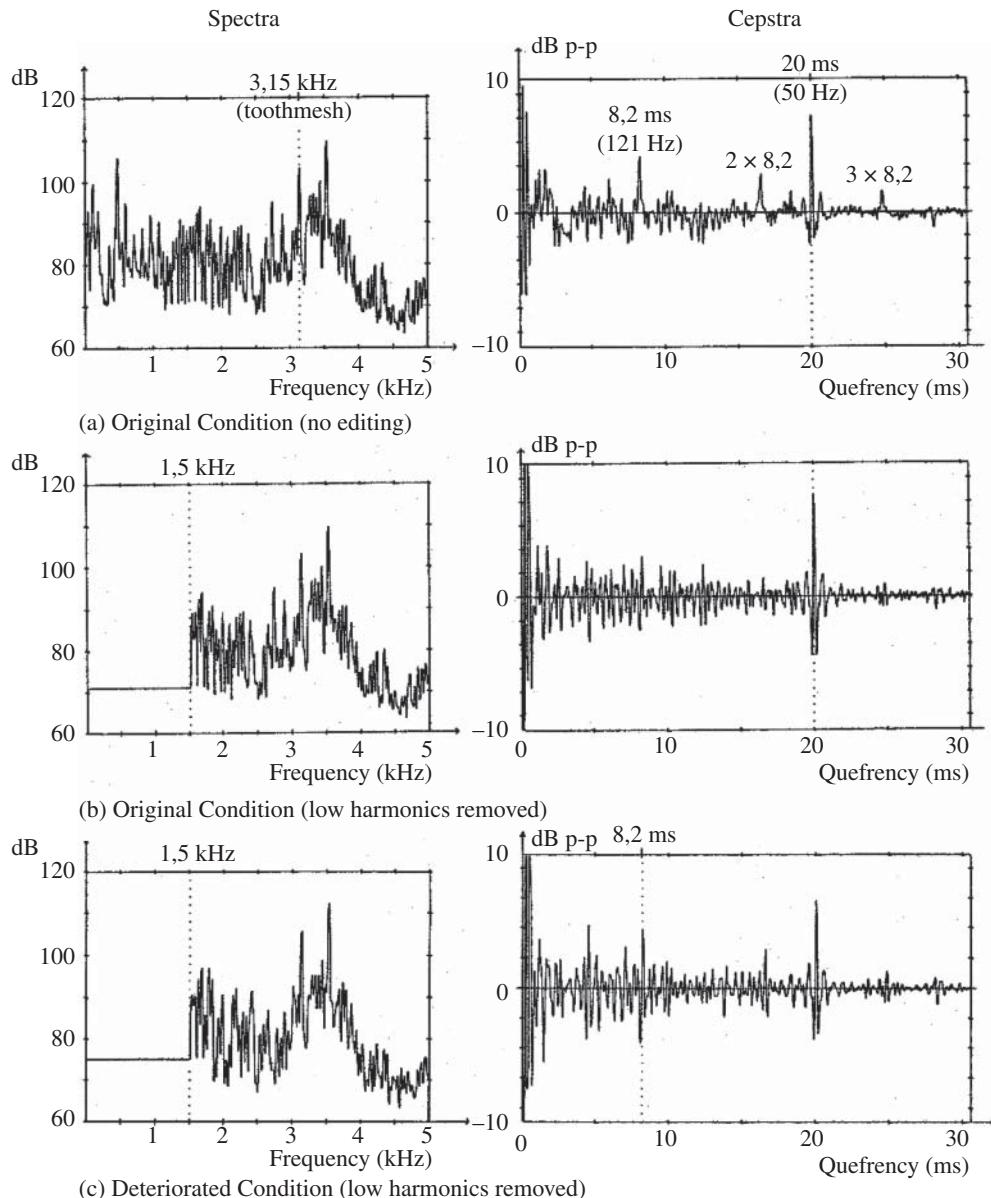


Figure 7.29 Editing in the spectrum to increase the diagnostic power of the cepstrum. (a) original spectrum and cepstrum (b) Spectrum edited to remove low order harmonics (c) Edited spectrum one month later with increased misalignment of the 121 Hz shaft. Source: Courtesy Brüel & Kjær.

without the masking influence of the other family. The same removal could have been achieved using synchronous averaging, but this would have required a tacho signal.

Other examples of harmonic and sideband removal, including choice of type and width of the comb notch lifter to use are given in Section 6.3.1.

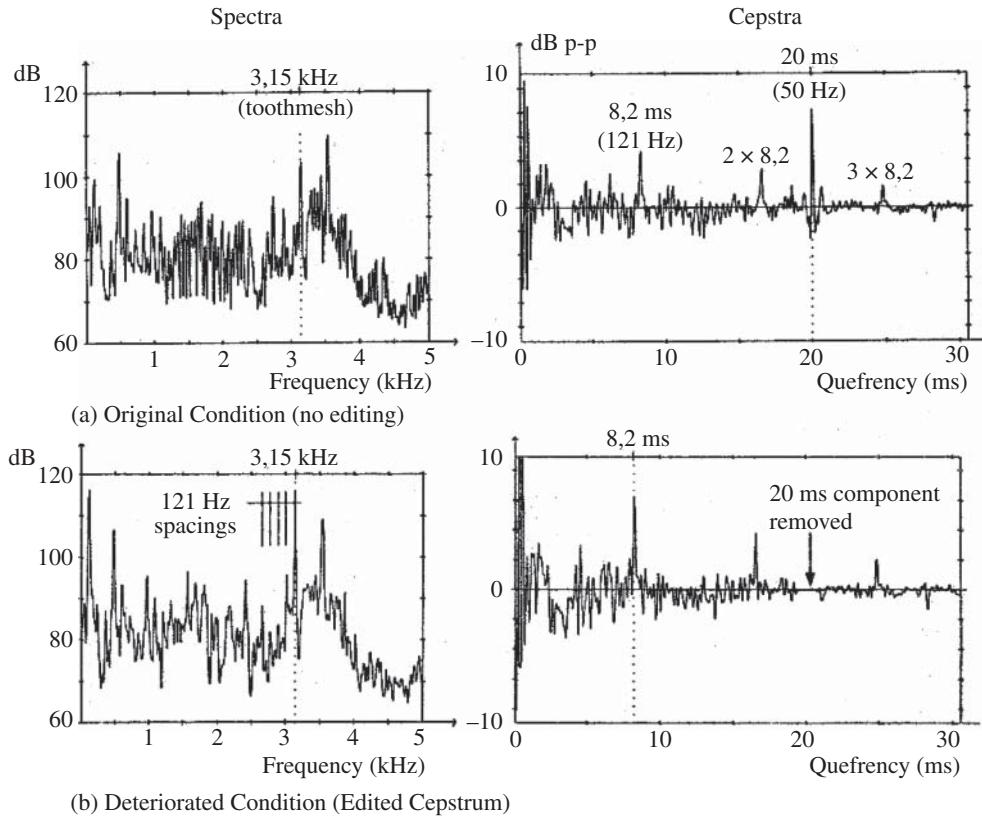


Figure 7.30 Editing in the cepstrum to remove a particular family of harmonics. Source: Courtesy Brüel & Kjær.

7.2.3.2 Separating Forcing Function from Transfer Function

The simplest example of the benefits given by the cepstrum in this respect is the fact that the useful part of the cepstrum representing the forcing function at the garmesh is little affected by the measurement point, and thus the different signal transmission paths to various measurement points.

Figure 7.31 shows such an example of spectra and cepstra for two measurement points on the same gearbox. Even though the spectra on a log amplitude scale at first seem very similar, they differ in detail. For example, at measurement point 1 there is a peak near 2.7 kHz, but a valley at the same frequency for measurement point 2. It can be shown [19] that the low quefrency region will have contributions from both excitation and transfer function, but the high quefrency region is almost entirely dominated by the forcing function, and is seen to be the same for both measurements.

Based on the theory of Ref. [19], Figure 7.32 (from [20]) shows an example where the garmesh excitation was removed from the low quefrency part of the cepstrum, for a case of a gear with and without tooth root cracks, and this shows that the transfer function has changed little for the two cases. In particular, the resonance frequencies of the transfer path appear to be unchanged, thus confirming that the difference is due to a change in the forcing function (the effect of the cracked teeth being primarily to change the force at the mesh). The part of the forcing function located at low quefrency comes from harmonics of the garmesh frequency in the spectrum,

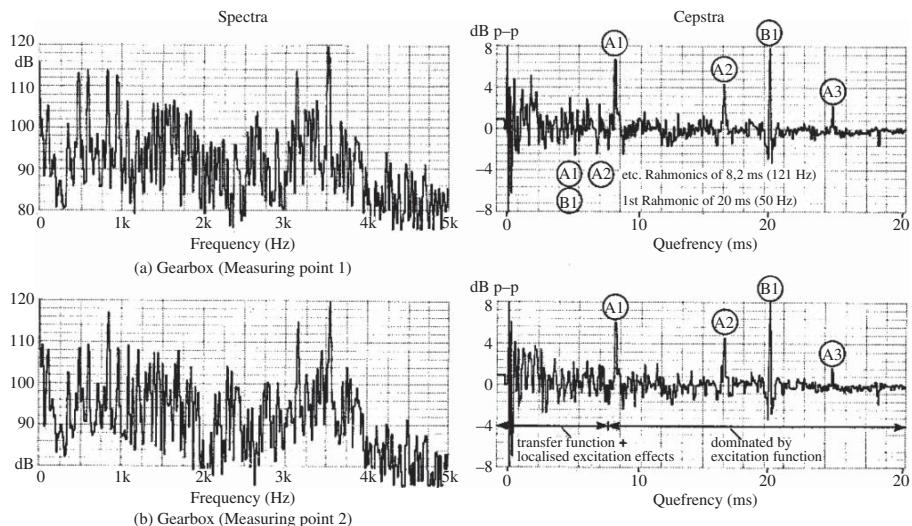


Figure 7.31 Spectra and cepstra for two measurement points on a gearbox. Source: Courtesy Brüel & Kjær.

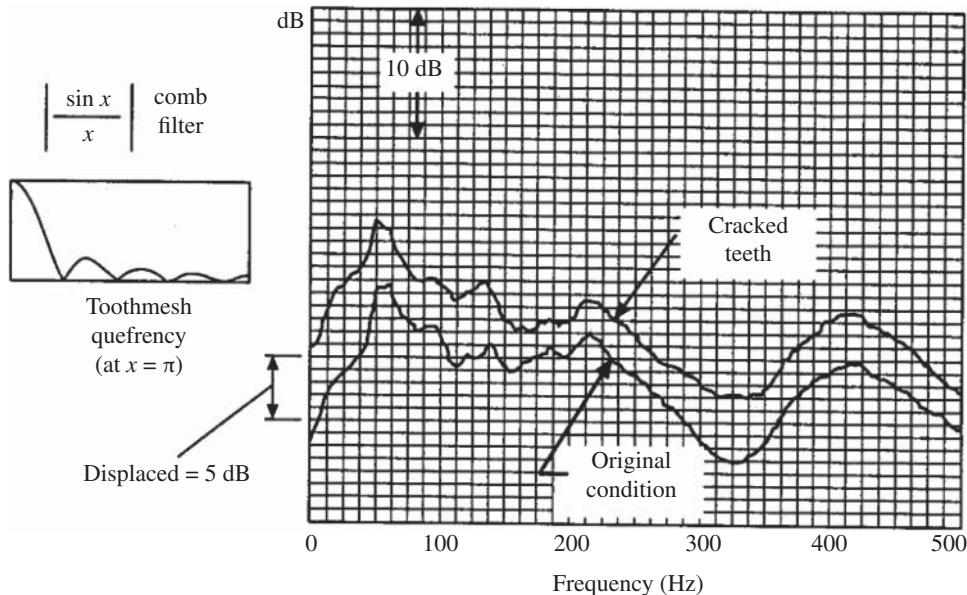


Figure 7.32 Use of filtering in the cepstrum to remove the forcing function component from the low quefrency part of the cepstra for a gear with and without cracked teeth, leaving the part dominated by the structural transfer functions [20]. This illustrates that resonance frequencies have changed little.

and in this case the corresponding rahmonics in the cepstrum were removed by using a ‘comb lifter’ given by a $\frac{\sin x}{x}$ function, with zeros adjusted to the spacing of the rahmonics to be removed.

Section 6.3.2 gives more advanced examples, where not only the spectra, but also time signals corresponding to forcing functions and modal properties can be separated, even using the real cepstrum.

Even though not developed further in this book, the cepstrum can be used for blind determination of the dynamic properties of a structure (operational modal analysis [OMA]), at least when the response is dominated by a single forcing function [21].

7.2.3.3 Identifying Echoes, and Inverted Echo Pairs

In Section 6.1.1 it is pointed out that an echo gives a series of rahmonics in the cepstrum, with a spacing equal to the echo delay time. For a positive echo, the rahmonics are an alternating series, but as shown in Figure 7.33, when the echo is negative, or inverted, all rahmonics are negative. Thus, the sum of all cepstrum components should become more negative when the section of signal analysed contains an inverted echo.

In [22] El Badaoui et al. defined the ‘moving cepstrum integral’ (MCI) to make use of this fact to detect the effects of a spall in vibration signals measured on a gearbox. The idea was that the signal produced when the mating tooth exited from a spall would be an inverted copy of the signal on entry, and would produce an inverted echo. This would be emphasised in the case of acceleration (as illustrated in Figure 7.35(2a–2c)). If a time window were moved along the record, with a length between one and two times the toothmesh period, and then the cepstrum

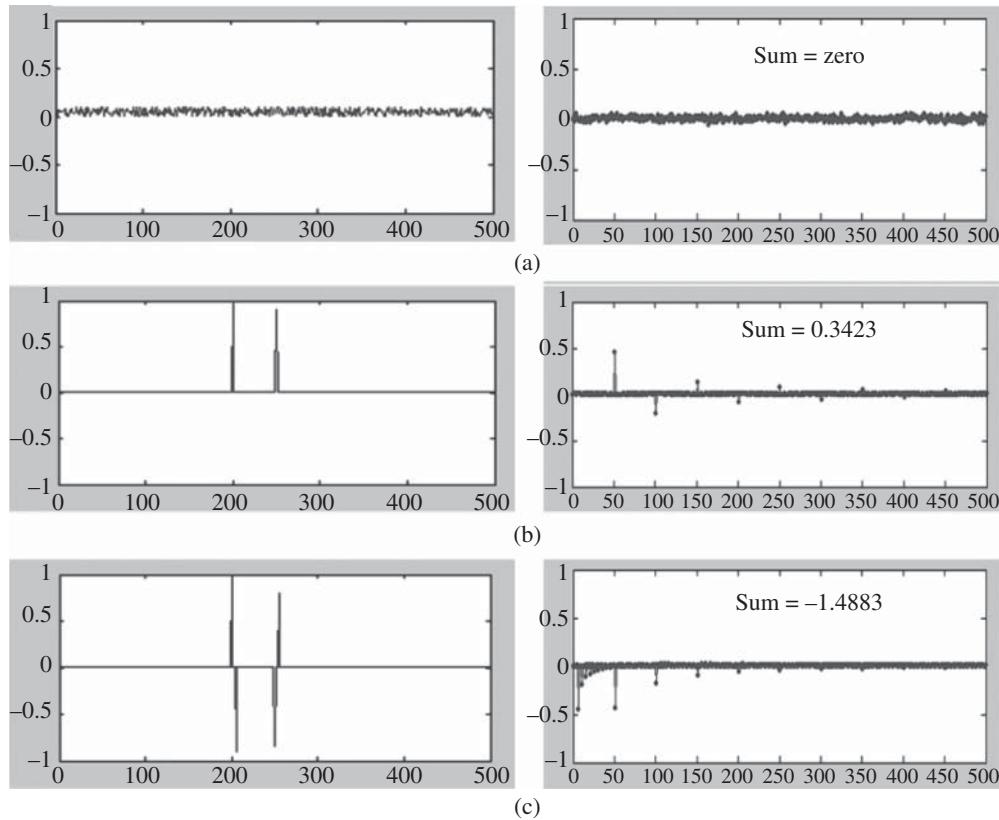


Figure 7.33 Cepstrum for (a) Noise (b) Positive echo (c) Negative echo.

calculated for each position and ‘integrated’ over all quefrency, the MCI should become negative when any inverted echo corresponding to a spall were located within the window. Figure 7.34 from [22] illustrates that this was in fact the case for an actual measured spall. In [23] the same authors, in collaboration with researchers from INSA Lyon in France, demonstrated that the same result was obtained when a spall was simulated on a gear. In other words, the rather simple assumption made in the paper, that the second derivative of the linear displacement of the mating gear tooth in entering and exiting a spall would correspond to the acceleration response, had some basis in fact.

In [24] Endo et al. showed that the MCI also went negative when a tooth root crack was present on a gear, because there is a tendency for this to give two inverted echo pairs. This is discussed further in Section 7.2.4 on separation of spalls and cracks.

7.2.4 Separation of Spalls and Cracks

The methods of Section 7.2.1 allow separation of local and distributed faults, but do not distinguish between spalls and tooth root cracks. This distinction is very important, since cracks would often have a very different prognosis from spalls, with a tendency to fail much more rapidly. The situation is by no means resolved, but a couple of recent studies have made simulations using finite element

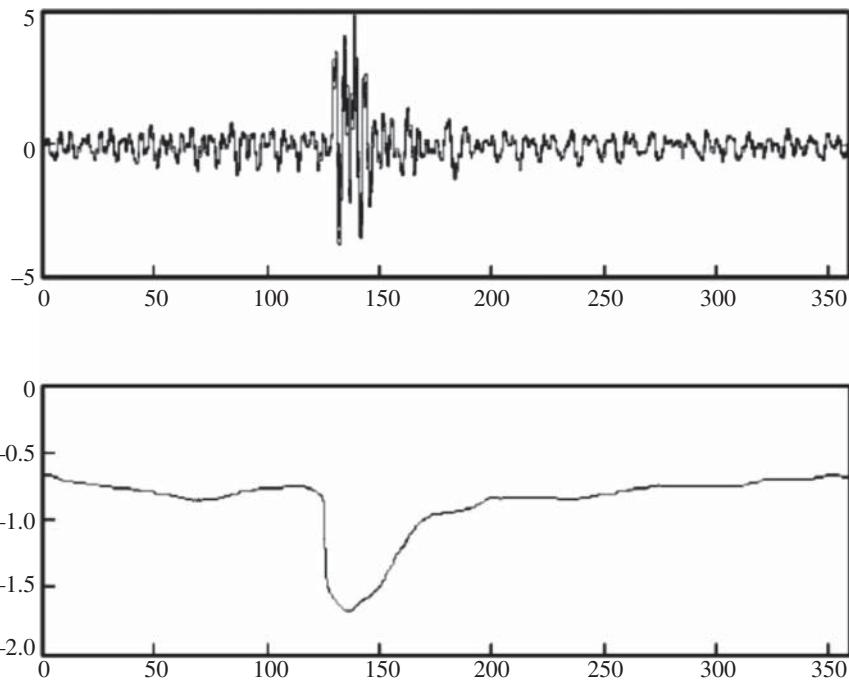


Figure 7.34 Moving cepstrum integral for an actual spall in a gear tooth [22].

analysis to learn of the differences given by spalls and cracks. Both studies treated spur gears only. The actual simulation models are discussed in Section 8.2.

In [24], Endo modelled tooth root cracks as a thin slot at the base of the tooth, this primarily giving a change in stiffness of the affected gear tooth when in mesh. As shown in Figure 7.35(1a–1c), this gives a two-stage deviation from the normal meshing pattern with healthy teeth, the first stage where the load is shared between a healthy pair and one involving the cracked tooth, and then a stage where only the cracked tooth is in mesh. When the resulting transmission error (TE) is double differentiated to give an approximation of the acceleration response, it is seen that there are two pairs of inverted echoes, as mentioned above in Section 7.2.3.3. Endo also simulated spalls of various widths across the tooth and length down the face of the tooth, and came to the conclusion that the TE was governed by the geometric error (virtually independent of load) and dominated by the effect of crowning, which meant that the mating tooth would enter further into the spall the wider the latter spread across the tooth. The resulting simulated TE (and its double differentiation) are shown in Figure 7.35(2a–2c).

From the simulations, Endo found not only that cracks gave a negative deviation of the MCI, but also that the two pairs of inverted echoes were apparent in the windowed cepstrum. Because the time axis depended only on the time the cracked tooth was in mesh, the echo delay time was independent of the crack depth, as shown in Figure 7.36.

On the other hand, the change in TE was found to be proportional to the depth of the crack, because of the reduction in tooth stiffness.

The equivalent for spalls is shown in Figure 7.37. In this case the echo delay time from the cepstrum is proportional to the size of the spall.

In [25] Endo presented the experimental verification of his results, and confirmed that the effect of the spall was independent of the load, whereas the effect of a crack was load dependent. It should

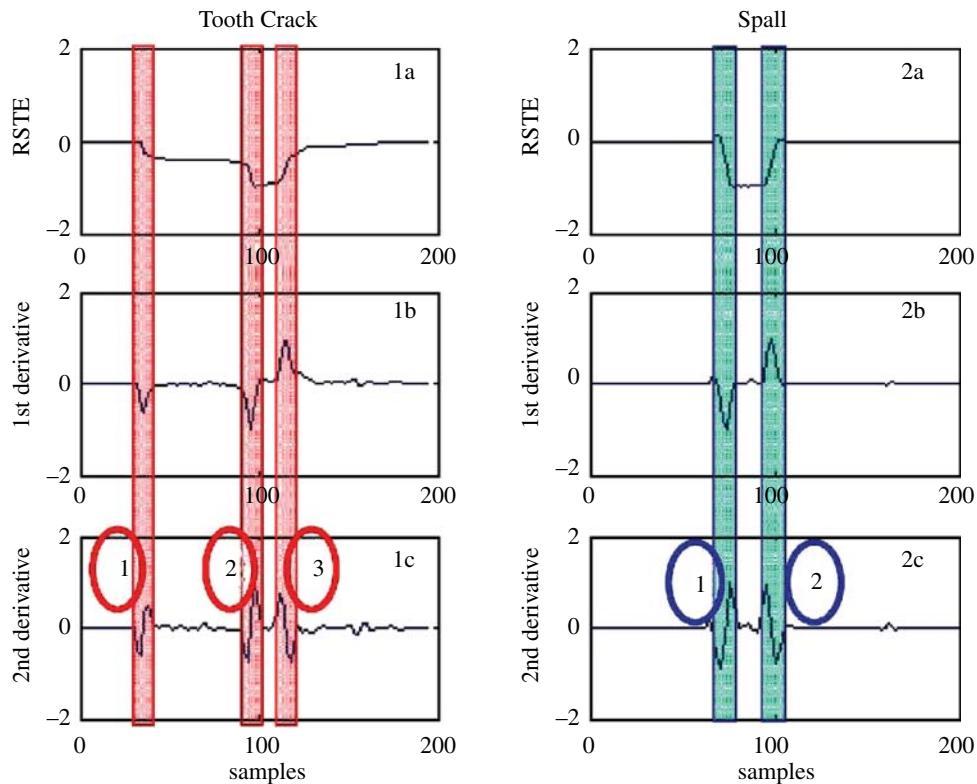


Figure 7.35 Simulated TE for tooth cracks and spalls along with the first and second derivatives to approximate acceleration [23].

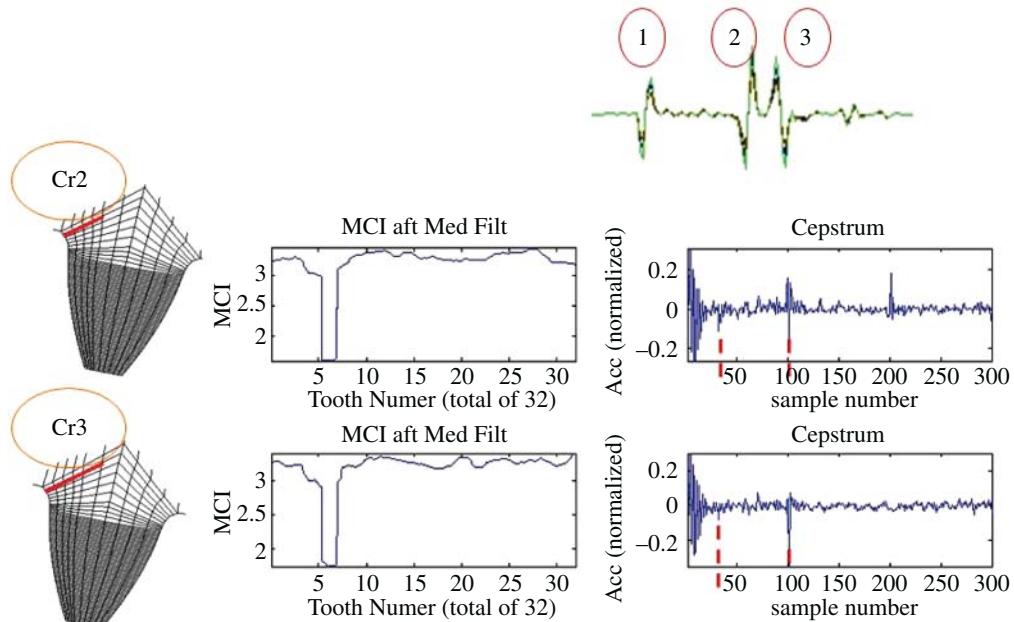


Figure 7.36 MCI and echo delay times for two crack depths, using simulation [23].

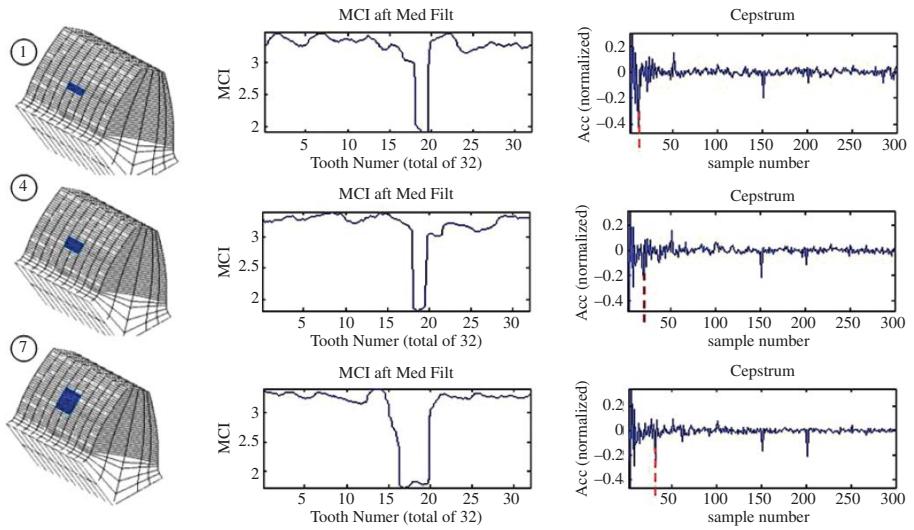


Figure 7.37 MCI and echo delay times for three spall sizes, using simulation [23].

be mentioned that the experimentally simulated crack in this case was formed by electro-discharge machining, and was thus not naturally occurring. Mark and Reagor [16] have shown that for naturally occurring tooth root cracks, the plastic deformation at the crack tip tends to give a fixed geometric deviation of the tooth, which would give a load independent component to the TE in addition to the load dependent component.

Another approach to the simulation of tooth root cracks and spalls is by Jia and Howard [26], who performed the simulations in a different way, in particular of the spall. The latter was modelled as a circular hole in the tooth surface, and because there was no crowning, there was very little geometric deviation. The effect of their simulated spall was thus mainly a change in contact stiffness over the reduced effective face width of the tooth. Even so, they confirmed Endo's finding that the effect of the crack was directly related to the time that the cracked tooth was in mesh, whereas the effect of a spall in general was shorter. They found other differences when the TE and acceleration, demodulated around a harmonic of toothmesh, were plotted in a polar plot. In other words, the real and imaginary parts of the analytic signal (shown in terms of amplitude and phase in Figure 7.8b,c) are plotted against each other.

This simulation work has not yet been extended to helical and other more complex gears, such as spiral bevel and hypoid, where it is evident that it would be more difficult to determine the length of time for which a cracked tooth would be in mesh, in particular when the crack does not extend across the whole width of the tooth.

Since the first edition of this book, other significant developments have however been made in separating the symptoms of spalls and cracks, in particular for planetary gears. Ref. [27] also uses mesh phasing [8] and shows how this can also be used to distinguish between spalls and cracks, in particular on a planet gear. This is because each planet tooth meshes with one flank on the ring gear, but the opposite flank on the sun gear. Thus, a fault which manifests itself when loaded on either flank (e.g. a cracked or chipped tooth) will give two pulses per revolution, while a spall will only show up when meshing with one or the other. Accurate phase measurement can also indicate whether the fault has the same time of engagement, independent of severity, as with a cracked (spur) tooth when in mesh, or whether this time varies with size (and location) and whether meshing with the sun or ring gear (which have different contact ratios), as with a spall.

7.2.5 Diagnostics of Gears with Varying Speed and Load

Many of the abovementioned methods rely on being able to analyse signals recorded under steady speed and load conditions. Where speed varies relatively slowly, over not too large a range, it would often be possible to compensate for it by order tracking as in the case of Figure 7.3. There are a number of situations, for example in the mining industry and with wind turbines, where the load varies over a wide range, sometimes randomly and over relatively short periods of time. This section gives a reference to some of the literature on this specialised topic.

Where the load is still cyclic, but perhaps second order cyclostationary rather than periodic, such as in some mining machinery, it is sometimes possible to obtain a measure of the approximate instantaneous load, for example using some form of time/frequency analysis, and then compensate for it. This was the approach in [28], but based on laboratory tests where the load was varied deterministically. Ref. [29] proposed another approach based on AR modelling, once again for deterministic load variations, a sudden step in load and sinusoidal variation. In [30, 31], Bartelmus and Zimroz

describe the problems encountered with planetary gearboxes on bucket wheel excavators. Here the load is cyclic, as each bucket digs, but with random variations making the load signal cyclostationary. They found that with a worn gearbox (no severe local faults were present) the planet carrier motion was more load sensitive, giving greater modulation, both in amplitude and frequency, which could be used as a condition indicator, or feature. Instantaneous motor speed or current could be used as load indicators, and the slope of the regression line relating the condition feature to load gave a reliable measure of wear condition.

For wind turbines, the load can vary over a very wide range in less than a minute. For the high speed parts of the gearbox, this would probably still allow them to be analysed by conventional techniques using signal sections selected for a particular load range. However, a case is described in [32] where a failure occurred in the low speed part of a gearbox (cracked teeth on the annulus gear), and this was not detected by conventional monitoring techniques. A number of parameters were monitored continuously, and recordings of time signals had been made somewhat randomly. The kurtosis of the overall signal reacted to the fault just before it failed, but no other parameters did. When the recorded signals were analysed, it was found using the fast kurtogram, Section 5.5.3.1, that high values of spectral kurtosis (SK) had occasionally manifested themselves up to several months before the final failure, and this could have been used as failure predictor, in particular if recordings had regularly been made for conditions of higher load. The filter band found by the kurtogram to give the maximum SK was centred on about 11 kHz, and this explains why conventional diagnostic techniques could not be used in a case such as this. The sampling frequency used was 25 kHz, and this was required to encompass the resonance frequency excited by the fault, but since the repetition frequency of the planet carrier around the annulus was about 0.3 Hz, the record length required for TSA would have to contain 75 000 samples. The frequency stability required for successful TSA would be an order of magnitude more than this (1 : 750000), and this is obviously not possible, even if the tacho signal were located in the low frequency part of the system. In fact, it was located on the generator shaft, and calculations indicated that the elastic windup of the input shaft compared with this output shaft would be 0.3° over the load range. The potential for using demodulation of the gearmesh frequency is even less, since this was about 30 Hz for the planetary gearbox, and the maximum allowable modulating frequency would be half this. The conclusion of the paper was that SK provided the best early indication of such a fault, and it was likely that envelope analysis of the optimally filtered signal would point to the source.

Ref. [2] ([4] in Chapter 5) does however give an example where a tooth root crack could be detected, after order tracking, by the method of Figure 7.8 even with a speed variation of $\pm 25\%$, see Figure 7.38.

Since the first edition of this book, more powerful techniques have been developed to compensate better for the effects of variable speed and load, even where order tracking only removes the frequency modulation, but not amplitude modulation, so that TSA may not be valid since it gives the average amplitude function, which when subtracted does not remove the whole effect. Complete compensation can possibly be achieved for two situations, as follows.

7.2.5.1 Effect Is Primarily Due to Passage through Fixed Resonances

Cepstral filtering with an exponential lifter can be used as in Section 6.3.2 to remove the modal part of the response, and retain what may be thought of as the intrinsic forcing function of the gear meshing. Even though this approach only removes the amplitude effects of the transfer functions, an example is given in Figures 6.22 and 6.23, where this appears to give similar results for two cases of $\pm 20\%$ speed variation around two quite different mean speeds, but nominally constant load, and this is very simple to apply.

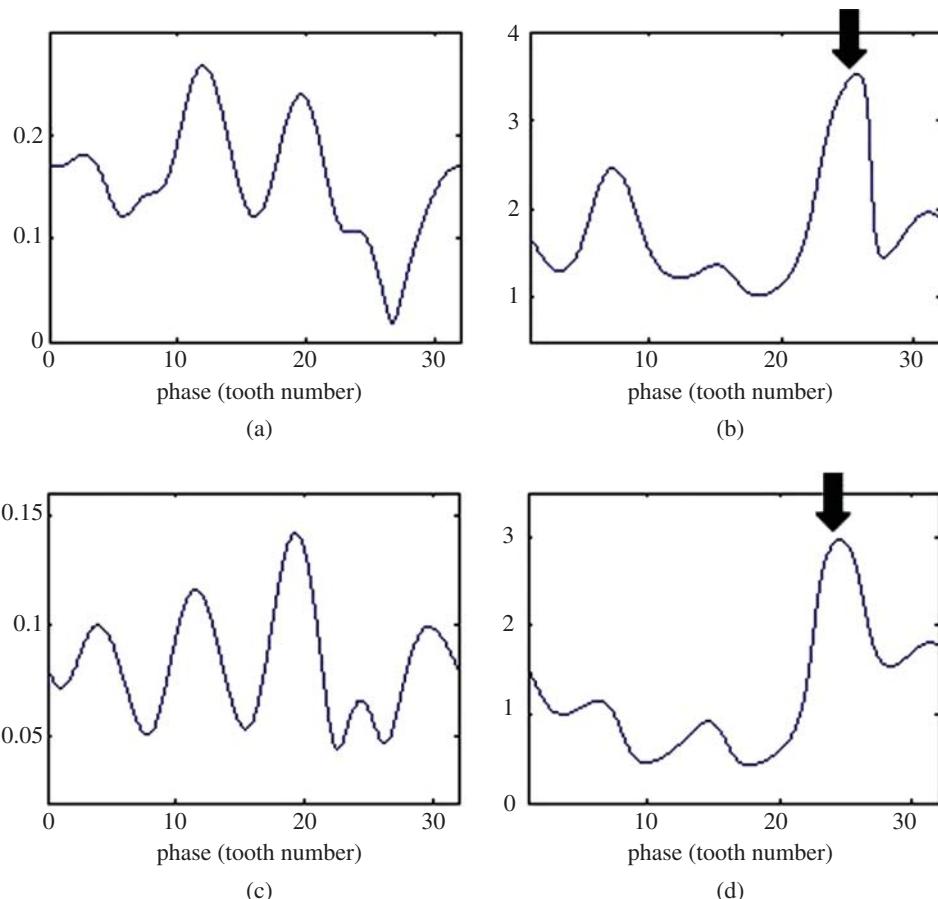


Figure 7.38 Detection of a tooth root crack from the phase demodulation of the garmesh frequency for $\pm 10\%$ and $\pm 25\%$ speed variation: (a, b) 10% case, (c, d) 25% case, (a, c) demodulated amplitude, (b, d) demodulated phase (from [2]).

7.2.5.2 Effect Includes Torque Variations

This is the general case where the actual forcing function varies with time (and speed) so that the response changes with both speed and load, possibly due to angular accelerations and decelerations. A theoretical approach to this problem was outlined in [33], but is yet to be proven in practice. It involves determining the frequency response function (FRF) matrix of the structure (including rotating internal components), which in principle can be done using OMA techniques, with scaling as described in Ref. [21]. Since the cepstral methods described there require SIMO models to make the output equal to the sum of input and transfer function, this would mean that the various sources would have to be separated using a form of blind source separation (BSS). As discussed in Ref. [33] and Ref. [21], this can often be done on the basis of the properties of second order cyclostationary signals, and many machine vibration signals are of this type. Since many sources, such as gearmeshes, would be internal, and not subject to direct measurement, this approach would often involve simulation models, where internal DOFs could be updated on the basis of external measurements. This point is taken up briefly in Section 8.2.

7.3 Rolling Element Bearing Diagnostics

Much of the background for this topic has been given in Section 2.2.3. Of most importance is the fact that the spectrum of the raw signal often contains little diagnostic information about bearing faults, and that over many years it has been established that the benchmark method for bearing diagnostics is envelope analysis, where a signal is bandpass filtered in a high frequency band in which the fault impulses are amplified by structural resonances. It is then amplitude demodulated to form the envelope signal, whose spectrum contains the desired diagnostic information in terms of both repetition frequency (ballpass frequency or ballspin frequency) as well as modulation by the appropriate frequency at which the fault is passing through the load zone (or moving with respect to the measurement point).

However, the envelope analysis technique was devised more than 30 years ago, e.g. [34], and used analogue techniques with inherent limitations. Considerable improvement can be made by taking advantage of digital processing techniques, rather than slavishly following the analogue method in digital form.

A number of benefits arise from performing the amplitude demodulation using ‘Hilbert transform’ techniques as described in Section 3.3. The procedure is illustrated in Figure 7.39 (from [35]). It is analogous to the amplitude demodulation process shown in Figure 3.30d.

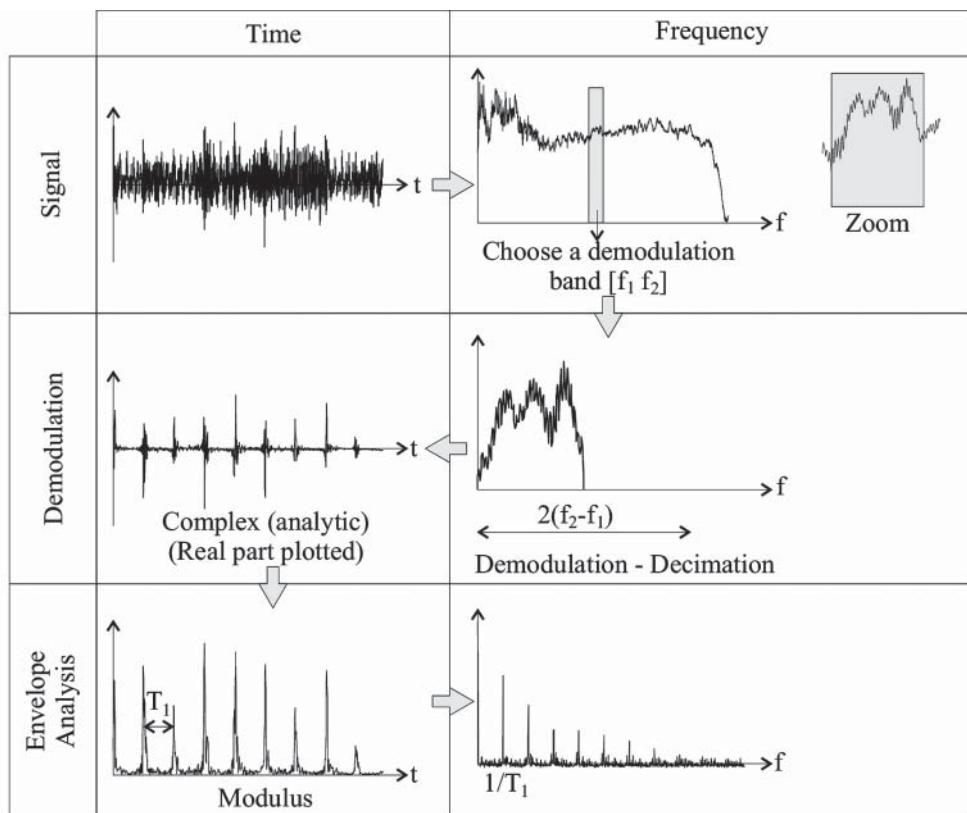


Figure 7.39 Procedure for envelope analysis using the ‘Hilbert transform’ method [35].

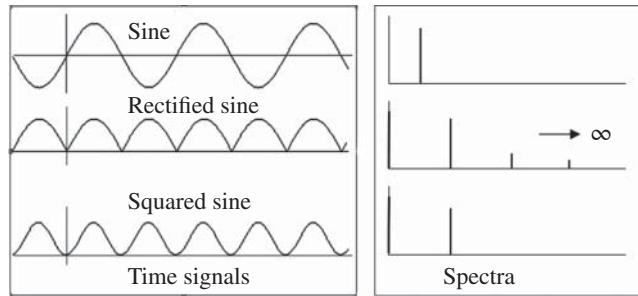


Figure 7.40 Potential aliasing given by squaring and rectifying a sinusoidal signal. With just squaring, but not rectification, aliasing can be avoided by doubling the sampling frequency before squaring.

An immediate benefit is that the extraction of the section of spectrum to be demodulated is effectively by an ideal filter, which thus can separate it from adjacent components that might be much stronger (e.g. gearmesh frequencies). This was not always possible with analogue filters, and real-time digital filters suffer from the same restrictions on filter characteristic. An example of this advantage is given in Ref. [36], which shows that non-causal processing is much better than causal processing, even with digital filters.

Figure 7.39 depicts the envelope as the modulus of the analytic signal obtained by inverse transformation of the selected one-sided frequency band. In fact, it was shown in [37] that it is preferable to analyse the squared envelope signal rather than the envelope as such.

The reason for this is simply explained in Figure 7.40, which compares the spectra of a rectified and a squared sinusoid. It should be noted that mathematically the envelope of a signal is the square root of the squared envelope, and likewise a rectified signal is the square root of the squared signal. The square root operation introduces extraneous components that are not in the original squared signal, and which cause masking of the desired information. In Figure 7.40 it is seen that the rectified signal has sharp cusps, requiring harmonics extending to infinity to reproduce them. Since the whole operation is done digitally, it is not possible to remove these high harmonics by lowpass filtration (as it was for example with an analogue rectifier), and they alias into the measurement range, causing masking. Note that since the squaring doubles the frequency content of a signal, the sampling frequency should be doubled before it is squared or rectified digitally, although as will be seen, this corresponds to the zero padding in Figure 7.39 when the analytic signal is processed.

Finally, the benefits of using the one-sided spectrum are illustrated in Figure 7.41 (from [37]). If the analytic signal (from the one-sided spectrum) is termed $f_a(t)$, its squared envelope is formed by multiplication with its complex conjugate, and the spectrum of the squared envelope will be the convolution of the respective spectra. Thus:

$$\Im\{f_a(t)f_a^*(t)\} = \Im\{f_a(t)\} * \Im\{f_a^*(t)\} = F_a(f) * F_a^*(-f) \quad (7.1)$$

When this convolution is carried out, as illustrated in Figure 7.41a, the result only gives difference frequencies, for example sideband spacings, which contain the desired modulation information. However, for the equivalent real signal $f(t)$, the spectrum of its squared value is simply the convolution of $F(f)$ (the spectrum of $f(t)$) with itself. This is illustrated in Figure 7.41b, and is seen to give the same difference frequency components, but mixed with sum frequencies (the difference of a positive and a negative frequency), which contain no diagnostic information, and only serve to mask the true result.

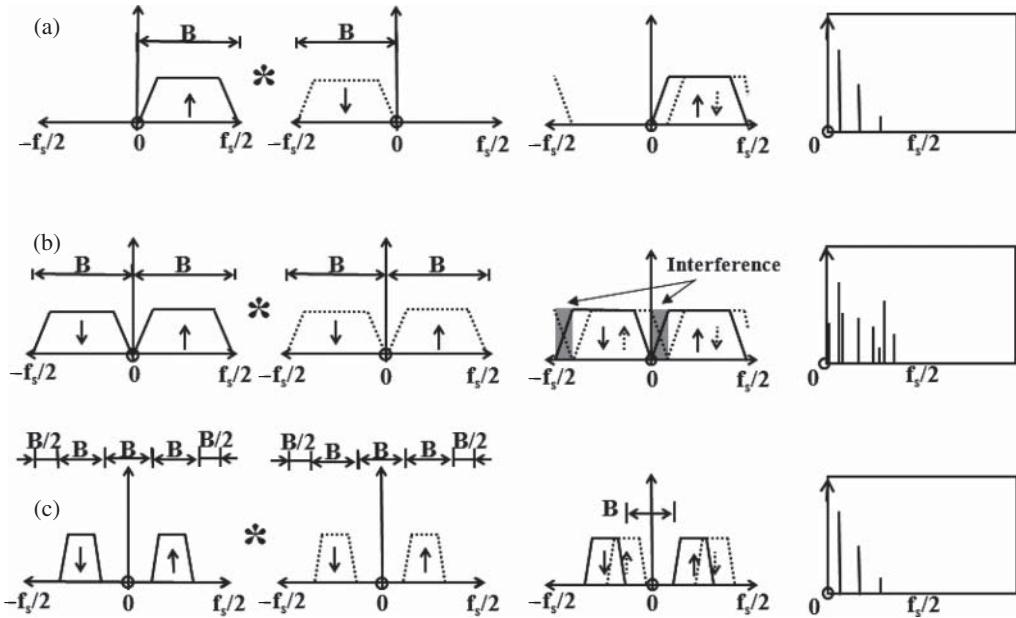


Figure 7.41 Generation of spectrum of squared envelope (or signal) for three cases (a) Analytic signal (b) Equivalent real signal (c) Frequency shifted real signal Downward arrow indicates complex conjugate [34].

As illustrated in Figure 7.41c, this interference can be avoided with a real-valued signal, as long as it is frequency shifted, so as to introduce zero padding around zero frequency as well as around the Nyquist frequency. This effectively means that the sampling frequency has to be doubled for the same demodulation band, so that transform sizes must be twice as big for the same problem. When Ref. [37] was written, the primary means of separating gear and bearing signals was using SANC (Section 5.3.4), which required real-valued signals, but the frequency domain based DRS method (Section 5.3.5) can make use of one-sided spectra, and thus avoid this complication.

Ref. [37] showed that even where the power of the masking noise (random or discrete frequency) was up to three times the power of the bearing signal, in the demodulation band, it was still advantageous to analyse the squared envelope. Using spectral kurtosis, it is usually possible to find a spectrum band where the signal/noise ratio of the bearing signal is much higher.

From the earliest days of envelope analysis there has been a debate on how to choose the most suitable band for demodulation, with many claiming that it is difficult, and some recommending the use of hammer tap testing to find bearing housing resonances. This problem has now been solved for the majority of cases by the use of spectral kurtosis (SK) and the kurtogram to find the most impulsive band (after removal of discrete frequency masking). As shown in Figure 5.37, this gives basically the same information as the dB spectrum difference before and after the appearance of the bearing fault. If the spectrum change has been caused by the bearing fault, it is obvious that the best signal/noise ratio of bearing signal to background noise corresponds to the biggest dB difference, independent of spectrum level. However, this requires reference signals with the bearing in good condition, while the SK method does not.

Since the publication of the first edition of this book, a number of situations have come to light where bearing faults do not necessarily give increased kurtosis, and some alternative approaches are discussed in Section 7.3.3.

7.3.1 Signal Models for Bearing Faults

The optimum way to analyse a faulty bearing signal depends on the type of fault present. A major difference is between initial small localised faults, giving rise to sharp impacts as the rolling elements contact the fault, and extended spalls, in particular if the latter have become smoothed.

7.3.1.1 Localised Faults

For localised faults, the question arises as to the correct way to model the random spacing of the impacts. Perhaps the first publication to model bearing fault signals as cyclostationary was [38], but the results were not very convincing, possibly because the main resonances excited by the faults may have been outside the measured range up to about 6 kHz. Figure 5.37 shows, for example, that localised faults on a very similar sized bearing only manifested themselves at frequencies above 8 kHz. Good results were obtained in [39], by modelling the vibration signals from localised bearing faults as cyclostationary. However, the way of modelling the random variation in pulse spacing (model 1) was later found to be incorrect, and in [40] a more correct model (model 2) was proposed. As illustrated in Figure 7.42, the variation in model 1 was modelled as a random ‘jitter’ around a known mean period, whereas in the correct model it is actually the spacing itself that is the random variable.

In particular, this has implications for the uncertainty of prediction of the location of a future pulse. For model 1, this is constant, and determined by the jitter, whereas in the actual situation, the variation is caused by slip for which the system has no memory, and thus the uncertainty increases with time of prediction into the future (model 2). As pointed out in [40], and put on a firmer mathematical basis in [41], this means that the signals from a localised fault in a bearing are not truly cyclostationary, but

Model 1

Random variable is the jitter δT_i around each period

$$T_i = iT + \delta T_i$$

This gives a truly cyclostationary signal with the uncertainty of occurrence independent of the number of periods into the future

Model 2

Random variable is ΔT_i defined by

$$\Delta T_i = T_{i+1} - T_i$$

This gives a pseudo-cyclostationary signal with the uncertainty of occurrence increasing with the number of periods into the future

Figure 7.42 Two models for the variation in period of pulses from a localised bearing fault.

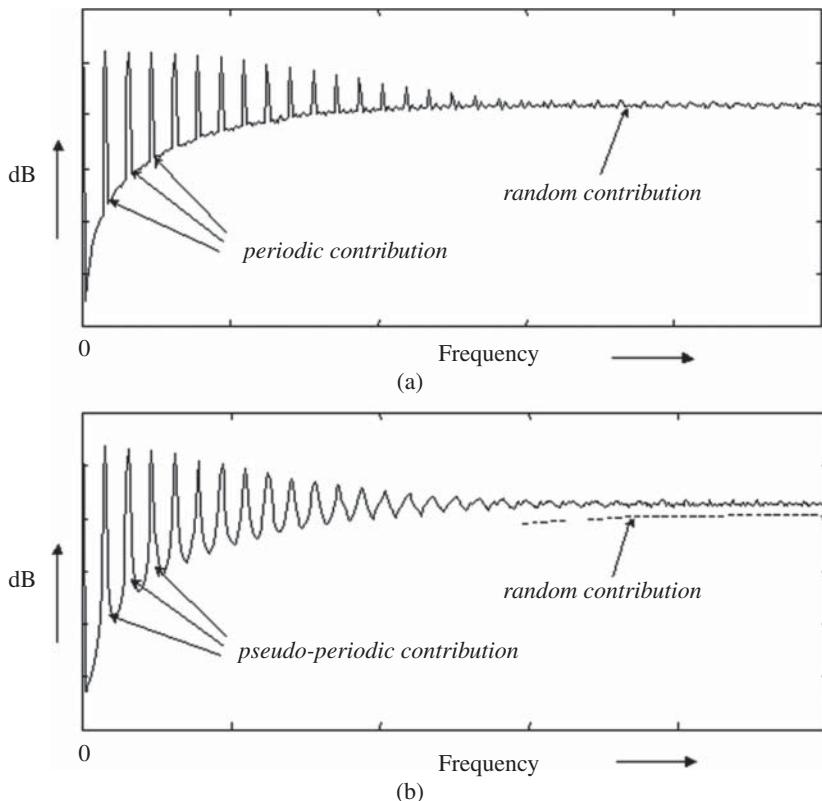


Figure 7.43 Frequency spectra for the two models (a) Model 1 (b) Model 2 [40].

are better termed ‘pseudo-cyclostationary’. Figure 7.43 (from [40]) shows the practical consequences of this for a signal with a small amount of random variation. It is seen that in terms of interpreting spectra, in particular envelope spectra, usually only at the low harmonics, there is little practical difference in treating the pseudo-cyclostationary signals as cyclostationary.

7.3.1.2 Extended Faults

For extended spalls, there will often be an impact as each rolling element exits the spall, and in that case, envelope analysis will often reveal and diagnose the fault and its type. However, there is a tendency for the spalled area to become worn, in which case the impacts might be much smaller than in the early stages. Cases have been encountered where extended spalls no longer give sharp impacts, but they can still be detected and diagnosed if the bearing is supporting a machine element such as a gear, since the fault will generally modulate the otherwise regular toothmesh signal. Figure 7.44 shows a typical modulating signal that would result from an extended spall in the inner race of a bearing supporting a gear. Because the rolling elements are in a different position on the rough spall surface for every revolution of the inner race, it contains both first order (the local mean value) and second order (amplitude modulated noise) cyclostationary components. In Figure 3.40 it is shown that such a mixture would have a spectral correlation with discrete characteristics

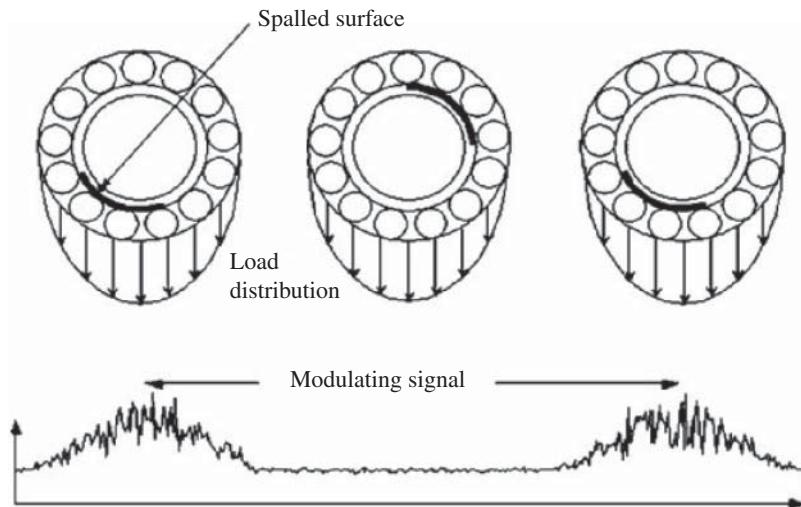


Figure 7.44 Generation of a modulating signal by an extended inner race fault in a bearing supporting a gear.

in the cyclic frequency direction, but a mixture of discrete and continuous characteristics in the normal frequency direction. This is because a periodic signal has a periodic autocorrelation function (in the time lag or τ direction) so the Fourier transform in this direction also gives discrete components. As stated in connection with Figure 3.40, if the modulation of the garmesh signal were by a gear fault it would be periodic, and would produce a spectral correlation with only discrete components in both directions (a ‘bed of nails’). However, if the periodic components are removed, by DRS or one of the other methods of Section 5.3, then only the second order cyclostationary components will be left, and they could only come from an extended bearing fault in a case such as this.

Note that the continuous lines in the spectral correlation of Figure 3.40 are at the low harmonics of shaft speed (Ω) but also in principle at the harmonics of ballpass frequency, inner race (BPFI), and sidebands spaced at shaft speed around them. For an inner race fault the shaft speed is probably the best to use to extract this information, but for an unmodulated outer race fault, components may be found in the spectral correlation at harmonics of BPFO. Note that where the shaft speed is the modulating frequency, the signal is truly second order cyclostationary (since the cyclic frequency is completely determined) whereas if the modulating frequency is BPFO, BPFI or FTF (for a ball fault), the signal would be pseudo-cyclostationary.

Figure 7.45 (from [40]) shows the spectral correlation, for cyclic frequency equal to shaft speed Ω , for two cases of inner race faults in the same type of bearing. For the localised fault, the difference manifests itself at high frequencies above 1000 shaft orders, whereas for the extended fault the differences are concentrated at lower frequencies up to 15 times the garmesh frequency. In the former case the fault was easily detected by envelope analysis, but in the latter case it was much less clear. Figure 7.46 (also from [40]) shows an actual case from the input pinion bearing of a helicopter gearbox, where the extended inner race spall was not detected until very late. There was no on-board vibration monitoring, and metal particles were getting trapped in an oil dam, and not reaching the chip detector. By the time these measurements were made on a gearbox test rig, the spall had become smoothed and did not reveal itself by envelope analysis at BPFI, only at the harmonics of shaft speed, and so could have been misinterpreted as a gear fault if this analysis (with removal of discrete frequency components) had not been done.

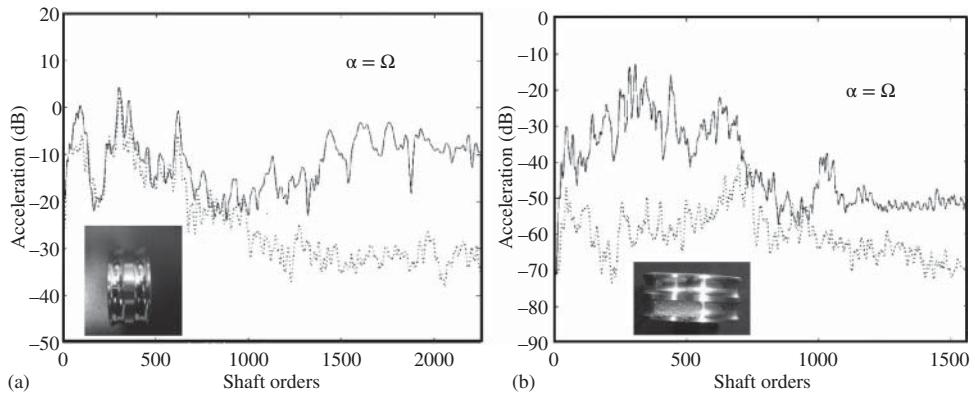


Figure 7.45 Spectral correlation evaluated for cyclic frequency equals shaft speed (a) Localised fault (b) Extended fault.

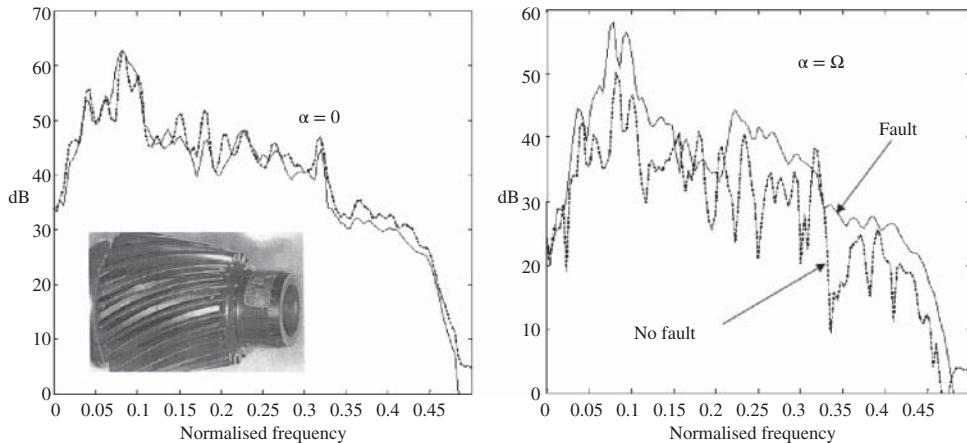


Figure 7.46 Comparison of spectral correlation evaluated for cyclic frequency equals zero (normal power spectrum) and shaft speed Ω for the depicted extended inner race spall. The fault is only apparent at $\alpha = \Omega$. Discrete frequency components were removed using DRS before the spectral correlation analysis.

7.3.1.3 Other Fault Models

The model used in most of this chapter, of a series of impulse responses (IRs), is appropriate when the fault size is small, and more physically correct than that assumed in Figure 3 of Ref. [42], where the signals are treated as a carrier frequency equal to the (single) resonance frequency excited, amplitude modulated by a series of decaying exponential envelopes corresponding to each new impulse response (IR), and with a step in phase at each new IR.

The IR model is inaccurate when the fault size is large, however, since the fundamental reason for the IRs is not a series of repeating impulse forces, but geometric errors caused by faults in the races. Thus, there is a relative motion between the two races, which is divided in some way between them in terms of absolute motion. This division is decided by the stiffness and inertial properties of the individual machine, but it should be remembered that only the absolute motion of the outer race, as

detected by measurements on the casing, will be reflected in the acceleration measured there. In the initial stages, when the faults are likely to be of micron size, the acceleration corresponding to these small displacements, at the relatively low frequencies corresponding to bearing fault repetitions, will be negligible, and the IR model, giving responses corresponding to samples of an ω^2 parabola at low frequency, will not be greatly in error. However, when the faults, and thus relative displacements are larger, and perhaps dominated by the absolute motion of the casing, the vibrations at the low harmonics of fault frequencies might be greater than those from the IR model.

7.3.2 A Semi-Automated Bearing Diagnostic Procedure

In [43], a method was proposed for diagnosing bearing faults that was successful for a wide range of cases, from high speed gas turbine engine bearings to the main bearing on a radar tower, with a rotational period of 12 seconds. It can be said to be semi-automated because only a small number of parameters have to be adjusted for each case, these corresponding to, and including, the dimensions and speed of the bearing. As shown in Figure 7.47, it combines a number of the techniques described in this chapter and Chapter 5. The method was further developed in Sawalhi's PhD thesis [44].

It is generally a good idea to start with order tracking (Section 5.1), as the separation of discrete frequency and random components will not always be possible unless this is done. Ref. [35] describes a case where it was not possible to use DRS to separate gear and bearing signals until order tracking had been carried out. No tacho or shaft encoder signal was available, but it was found possible to extract the instantaneous speed information by phase demodulation of a number of gearmesh frequencies, these being phase-locked to shaft speed. The best mapping of shaft angle vs time was given by averaging a small number of estimates with a similar appearance. In this case the random speed variation was only 0.5% peak-to-peak (1203–1209 rpm).

For separation of discrete frequency and random components (e.g. gear and bearing signals) the best choice is generally DRS (Section 5.3.5) as it poses the minimum problems with regard to choice of parameters. The size of transform N should span 10–20 periods of the minimum frequency to be removed (e.g. the lowest shaft speed) and the delay should be at least three times the correlation length of the bearing signal. Assuming 1% slip, this would correspond to about 300 periods of the centre frequency of the demodulated band. Determining the latter might require one iteration, as it is best decided after the SK procedure.

Minimum Entropy Deconvolution (MED) need only be applied for high speed bearings, where the impulse response of the bandpass filtered resonance is of comparable length to the spacing of the

- (1) Order tracking – Remove speed fluctuation
- (2) DRS, SANC or Linear Prediction - Remove discrete frequencies
- (3) MED – Remove smearing effect of signal transfer path
- (4) SK – Determine optimum band for filtering and demodulation
- (5) Envelope analysis – Determine fault characteristic frequencies

Figure 7.47 Semi-automated procedure for bearing diagnostics.

bearing fault pulses (BPFI would normally be the highest fault frequency and thus the shortest spacing). This can perhaps best be decided by trial and error, based on whether MED gives an increase in SK.

The optimum band for demodulation should be chosen using a fast kurtogram procedure. Note that the kurtogram is sensitive to large random pulses which may be present in some realisations of a signal. If the final envelope spectrum does not reveal periodic components, even though the SK is high, it should be checked whether such random impulses from an extraneous source are dominant in certain frequency bands.

In the final envelope analysis, it should be recognised that modulating effects are important to the diagnosis. In general, inner race faults would be modulated at shaft speed, and rolling element faults at cage speed. For unidirectional load, an outer race fault would not be modulated, but modulation at shaft speed can occur because of significant unbalance or misalignment forces, and modulation at cage speed can result from variations between the rolling elements. Note that with planetary gear bearings, it is the inner race that is fixed with respect to the load, and so inner race faults tend not to be modulated, whereas the signals from outer race faults are modulated by the frequency at which they pass through the load zone. Since planet gears are analogous to rolling elements in a bearing, the modulation frequency can be calculated by an equation similar to that for BSF.

Ref. [43] illustrates the general procedure by its application to three very different case histories, so a brief summary of those results is given here.

7.3.2.1 Case History 1 – Helicopter Gearbox

A test was carried out on a helicopter gearbox test rig at DSTO (Defence Science and Technology Organisation, now DST Group) Melbourne, Australia, where it was run to failure under heavy load. The signals were analysed blind, with no indication of the type of failure. Frequencies corresponding to the planet bearing (which actually failed) are given in Table 7.1, although all other potential bearing frequencies had to be calculated.

Initial analyses of the signals, even at the end of the test where bearing failure was indicated by the growth of wear debris, showed no indication of the fault in either the time signal or spectrum. The latter was dominated by gear components (harmonics of the main meshing frequencies and their sidebands) over the whole frequency range up to 20 kHz. The kurtosis of the raw signal was -0.6 , roughly like noise.

The procedure of Figure 7.47 was applied, and in this case the discrete frequency components were removed using linear prediction. Figure 7.48 compares the residual of the linear prediction process with the original signal, and it is seen that it has become slightly more modulated, (the kurtosis increased to 2.2), but the three ‘bursts’ are related to the passage of the planets (period 58.1 ms).

The Wavelet Kurtogram (Section 5.5.3.1) of the residual signal (of Figure 7.48b) was next produced using a range of filter banks (3, 6, 12, 24 filters/octave) and the results are shown in Figure 7.49. The maximum SK of 12 was obtained using 12 filters/octave (centre frequency of 18 800 Hz and bandwidth 1175 Hz).

Table 7.1 Planet bearing frequencies.

Fault	BPFO	BPFI	Cage	Roller
Frequency (Hz)	77.1	117.8	9.8	37.0

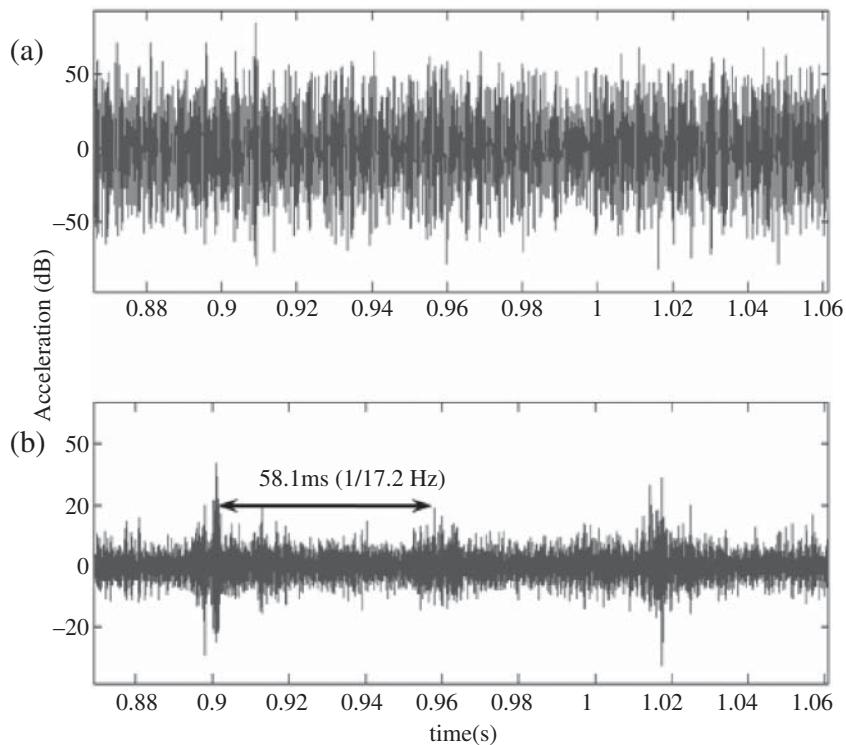


Figure 7.48 Time signals (one rotation of the carrier), (a) Order tracked signal, (b) Residual signal – passage of the three planets is seen.

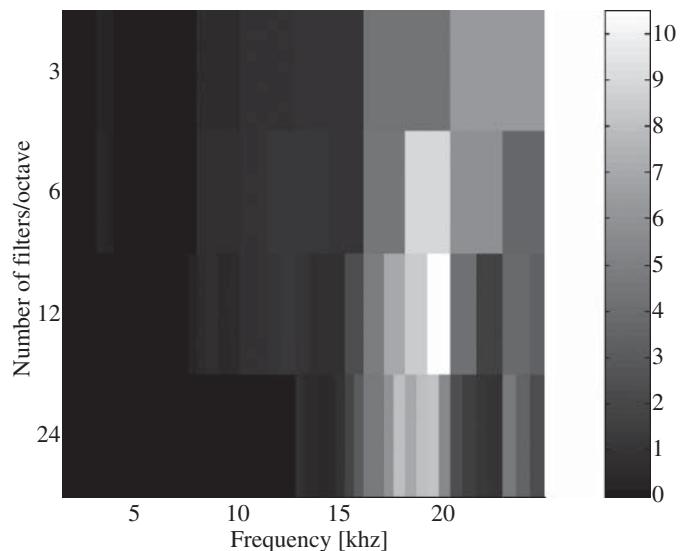


Figure 7.49 Wavelet kurtogram for 4 filter banks; namely (3, 6, 12, 24) filters/octave.

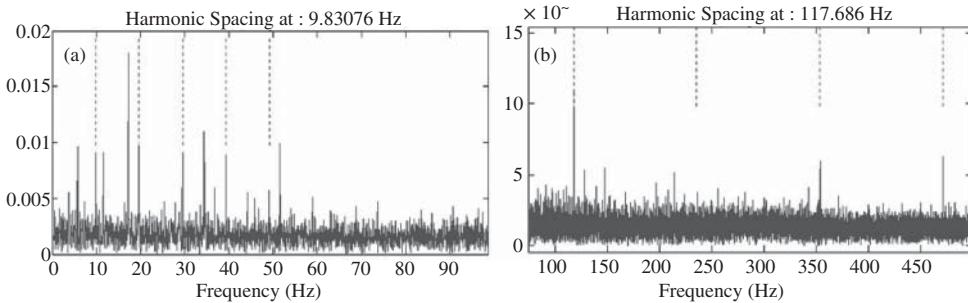


Figure 7.50 Squared envelope spectra showing two fault frequencies (a) Cage speed (9.8 Hz) (b) BPFI (117.7 Hz).

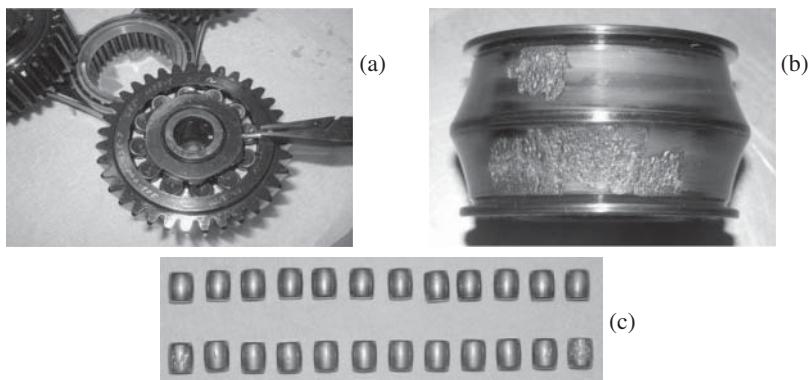


Figure 7.51 Damaged planet gear bearing after final disassembly (a) faulty bearing (b) spalled inner race (c) spalled rollers.

Finally, Figure 7.50 shows the squared envelope spectra in two frequency ranges, for the last measurement. Figure 7.50a shows a strong pattern of harmonics spaced at the cage speed of a planet bearing. This basically gives an indication that there is a variation for every rotation of the cage. This can be a cage fault, but is often an indicator of variation between the rolling elements. Figure 7.50b, in a somewhat higher frequency range, shows a strong component corresponding to the BPFI. Because it is a planet bearing, no modulation is expected for an inner race fault, and no modulation sidebands are found in the envelope spectrum. When the gearbox was disassembled, severe spalling was found on the inner race of one planet bearing and three rollers had minor spalls, explaining the modulation at cage speed. The final damage is shown in Figure 7.51.

The trending of analysis parameters for this case is discussed in Chapter 9 on Prognostics.

7.3.2.2 Case History 2 – High Speed Bearing

Measurements were made on a bearing test rig at FAG Bearings (now Schaeffler) in Germany, on which bearings are tested to failure. At several points through their life, the bearings are dismantled and inspected. The bearing being tested is for a high speed application, and is thus tested at 12000 rpm, typical of gas turbine bearings. The accelerometer used to capture data was mounted

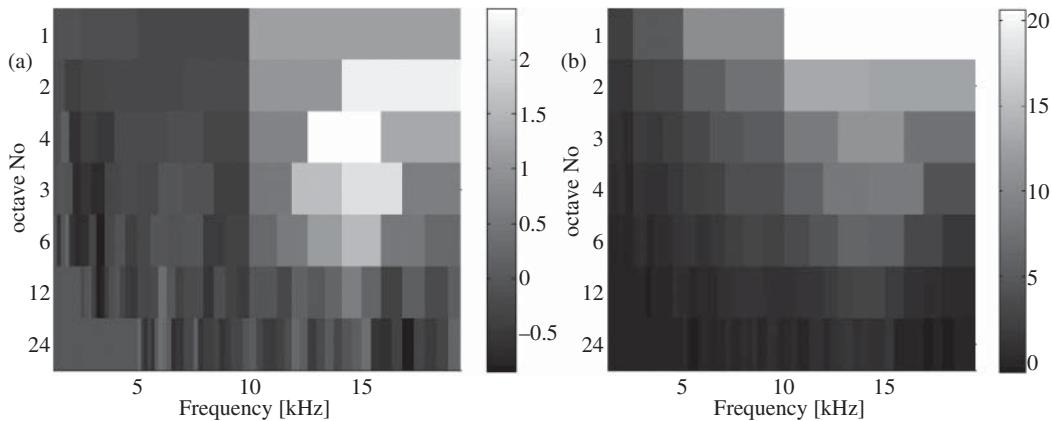


Figure 7.52 Wavelet kurtograms (a) Before application of MED (b) After MED.

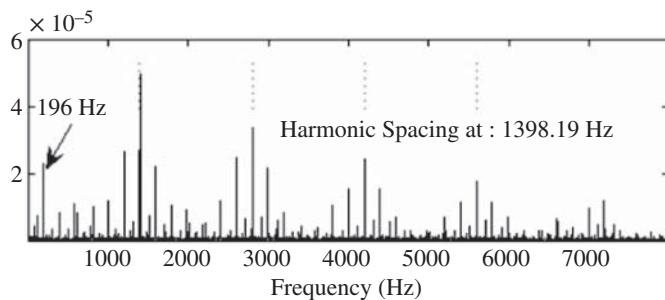


Figure 7.53 Envelope spectrum of signal of Figure 5.35c showing harmonics of BPFI 1398 Hz, and harmonics and sidebands spaced at shaft speed 196 Hz.

using a magnet, and it is suspected (from inspection of the spectra) that the mounting resonance frequency was of the order of 12 kHz.

This is the case depicted in Figure 5.35, where MED gave a considerable improvement in the impulsiveness of the signal because the individual impulse responses were overlapping. The wavelet kurtograms for the signals before and after the application of the MED technique (Figure 7.52) show that the SK has been increased from 2.5 to 20, making the impulsiveness apparent (Figure 5.35).

Figure 7.53 shows an envelope spectrum for a late stage of development of the fault, demodulated in the band indicated in Figure 7.52b. It is a typical envelope spectrum for an inner race fault, with a series of harmonics of BPFI (1398 Hz), together with low harmonics of, and sidebands spaced at, shaft speed (196 Hz).

The trending of analysis parameters for this case is also discussed in Chapter 9 on Prognostics.

7.3.2.3 Case History 3 – Radar Tower Bearing

Measurements were received from before and after a main bearing change on a radar tower. The radar driving system consists of a motor, a gearbox, and a spur pinion/ring gear combination. The motor runs at 1800 r.p.m. (30 Hz) and is connected to a three-stage reduction gearbox. The final tower

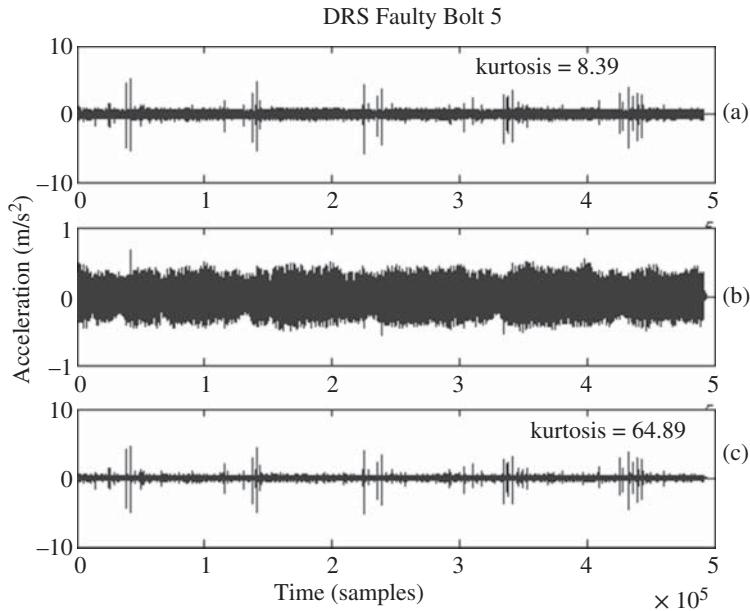


Figure 7.54 Results of applying DRS to signal with faulty bearing (a) Original signal (b) Deterministic part (c) Random part.

rotation period is 12 seconds (0.082 Hz). The ring gear used to drive the tower is an integral part of the bearing and so was changed at the same time, but the drive pinion was left unchanged. The reason for the change was an increase in noise (which may have been partly due to wear of the gears), but the current analysis clearly shows evidence of bearing faults in the old bearing. Compared with the two previous cases, the speed is orders of magnitude less, but even so the same analysis procedure could be used with little operator intervention. Because of the slow speed, the fault impulses were well separated, so there was no need to use MED filtering.

The spectra from before and after replacement gave no indication of the bearing fault, but the raw time signal did indicate the fault, as a series of impulses protruding from the background signal. Application of DRS (Figure 7.54) showed that the latter was dominated by deterministic components, mainly from the gears (which dominated the spectra). The local impulses visible in the time signal from the old bearing (before and after application of DRS) were not present in the signals from the new bearing. Removal of the gear signal increased the kurtosis from 8.4 to 64.9. Even without using SK, envelope analysis of the signal from Figure 7.54c revealed the bearing fault frequency (4.79 Hz) with modulation sidebands at rotational speed 0.082 Hz. Interestingly, it is not possible to determine whether the fault(s) were in the inner or outer race, partly because this is a thrust bearing, and so BPFI and BPFO are the same ($\phi = 90^\circ$ in Eqs. (2.12) and (2.13)).

Moreover, there is a reason why faults in both the fixed and moving races would be modulated at shaft speed, the former because the load was somewhat eccentric, and thus rotating around the bearing, and the latter because of the varying path length to the transducer.

Figure 7.55 shows the result of using a wavelet kurtogram to apply the optimum bandpass filter and extract the signal coming from the bearing fault(s). This was found to be a filter of bandwidth 517 Hz centred on 2755 Hz. The kurtosis has increased to the remarkable value of 541. This makes it clear that the kurtosis is not a parameter that can be used directly as an indicator of fault severity;

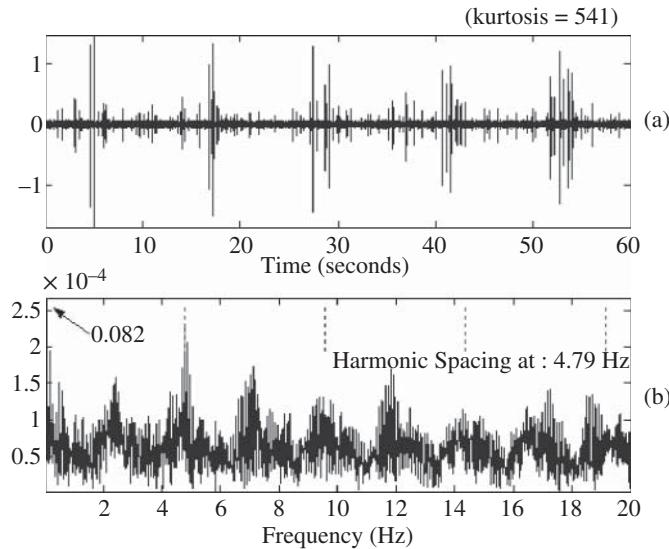


Figure 7.55 Effect of optimal SK filtering (a) Time signal (b) Envelope spectrum.

it is affected not only by the size of the individual fault pulses, but also by the spacing between them. If this machine were run at twice the speed, for example, it is likely that the kurtosis would be approximately halved. Thus, in evaluating the kurtosis corresponding to machine faults, account should be taken of the ratio of fault repetition frequency to typical resonance frequencies excited.

The damping ratio of the impulse responses also has an influence on the kurtosis (hence the benefits of MED). This is also a factor in the case mentioned in Section 7.2.5, of a gear fault in the low speed section of a wind turbine gearbox [32], where relatively high values of SK are found.

The harmonic cursor in Figure 7.55 is set on the ballpass frequency of 4.79 Hz, but it will be seen that there are components at multiples of half this frequency also. The reason for the appearance of the half frequency is that these bearings have a special construction with the races in a V-shape, and the 118 rollers (with the same length and diameter) being alternately mounted with $\pm 45^\circ$ orientation, so that only each alternate roller contacts one side of the V. Thus a fault on one side would give impulses at half the BPF. As mentioned above, there are sidebands spaced at rotational speed (0.082 Hz) around the harmonics of (half) BPF, as well as low multiples of 0.082 Hz.

It is evident that the semi-automated procedure used to extract the bearing signal of Figure 7.55a would have detected the bearing fault at a very early stage, long before it protruded above the background gear signal, as in Figure 7.54a.

7.3.3 Alternative Diagnostic Methods for Special Conditions

Even though bearing faults usually give rise to impulsive responses with high kurtosis, in particular in the early stages, Ref. [45] discusses a number of reasons for this not (or no longer) to be the case:

- 1) Even if the initial very small faults give short impulsive forces with excitation up to a very high frequency, the frequency range has a tendency to decrease as faults become physically larger, and possibly smoother, exciting lower frequencies, though less impulsively.

- 2) The frequency range of the measurement may not be sufficiently high to include the main resonance frequencies excited, but the faults may still manifest themselves at lower frequencies, perhaps even additively when they get sufficiently large, as discussed in Section 7.3.1.3.
- 3) In some cases, the bearing faults may modulate some extraneous carrier frequency (or frequencies) such as the meshing frequencies of gears supported by the bearings, as discussed in connection with Figure 7.44. The bearing faults then require this carrier in order to be observed.
- 4) The period of repetition of the fault is much shorter than the relaxation time of the structural resonances it excites; as a consequence, adjacent impulse responses overlap, which tends to decrease the kurtosis towards the value of a Gaussian distribution. As discussed in Section 5.4, and in Section 7.3.2.2 of this chapter, this can often be counteracted by using MED.
- 5) The bearing faults are not the only, and perhaps not the strongest source of signals with high kurtosis. A typical such case is given by signals from induction motors with a variable frequency drive (VFD), where the associated EMI (electromagnetic interference) signals can be very impulsive and dominant.

One effective solution to the problem of low kurtosis is indicated in [45] and consists in using the spectral coherence diagram instead of the spectral correlation to enhance the bearing fault information over a wider frequency range. Figure 7.56 shows one of the examples from that paper, where an advanced inner race fault in a pump gave low kurtosis, even though the signal was dominated by harmonics of BPFI and sidebands rather than shaft harmonics. The squared envelope spectrum (SES) is of course the integral of the spectral correlation over all frequency, but is not as clear as the integral of the spectral coherence, which gives the so-called ‘enhanced’ or ‘improved’ envelope spectrum.

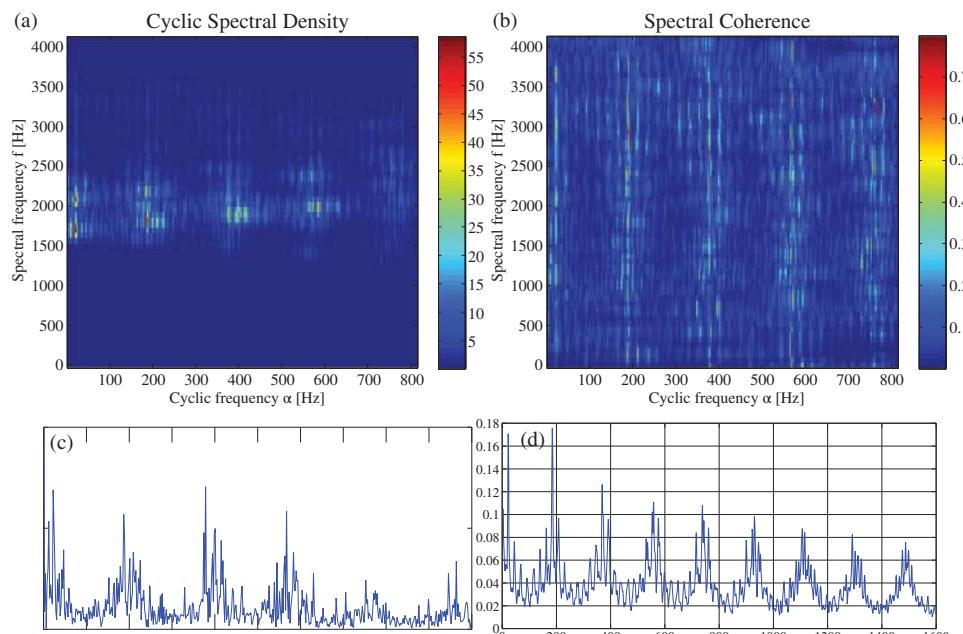


Figure 7.56 Comparison of spectral correlation vs spectral coherence for enhancing the squared envelope spectrum (a) Spectral correlation (b) spectral coherence (c) Squared envelope spectrum (d) Enhanced envelope spectrum.

It should be noted, however, that even though the enhanced envelope spectrum would detect and diagnose a bearing fault earlier than the SES, it would be insensitive to any change in the extent of the frequency range excited (roughly 1500–2500 Hz in Figure 7.56a), and thus not be so useful for tracking the size of a fault, which would typically lower the frequency range excited as it grew physically larger.

Ref. [46] contains a thorough comparative summary of a number of proposed solutions of the problem of choosing an optimum frequency band for demodulation, including the protrogram (which appears to have a limited range of application), the infogram, and the ratio of cyclic content (RCC). The latter solves one problem with the use of kurtosis alone, in that the latter can be dominated by individual large pulses, which do not repeat with a particular cyclic frequency. The RCC [47] is based on the finding that the kurtosis of a signal is given by the normalised sum of the squared amplitude peaks of the SES, and if these are restricted to narrow bands around the harmonics of a known cyclic frequency, corresponding to a bearing fault, it will give the proportion of the total kurtosis corresponding to that cyclic frequency. It means that the operation has to be carried out separately for different cyclic frequencies.

The primary recommendation of Ref. [46] is that insights gained from the studies reported in Refs. [48, 49] on the distinct nature of non-gaussianity and non-stationarity in impulsive signals, have led to a powerful basis for the detection of cyclostationary sources (with known cyclic frequencies such as BPFO, BPFI and BSF) in the presence of impulsive noise, and unrelated cyclic frequencies such as from EMI. The method, based on the log envelope spectrum is known as the ‘log cycligram’ and is shown to give better results in a wider range of cases than alternative methods.

7.3.4 Diagnostics of Bearings with Varying Speed and Load

The vastly increasing application of wind turbines for renewable energy, and other variable speed machines such as mobile equipment and automated machine tools, has led to a need to adapt traditional condition monitoring techniques to the case of varying speed and load, which has brought up a range of problems not encountered with constant speed machines.

As for constant speed machines, there is a need to separate the signals from bearings and gears, with and without faults, and varying speed has emphasised the differences between them. Gear signals are usually dominated by the deterministic forcing functions, at harmonics of shaft speeds and gearmesh frequencies, while bearing faults are usually carried by resonance frequencies, unchanging with speed, even though the repetition frequencies of the impulse responses (IRs), from exciting these resonances, do follow shaft speeds, but with a small amount of random variation due to slip. With nominally constant speed machines, including those driven by, or driving, induction motors or generators, where a speed variation of the order of 2% occurs, it has been common to treat them as constant speed for bearing diagnostics, since the bearing ‘frequencies’ are smeared anyway. Many of the techniques for removing gear signals, such as DRS and cepstral notching, are not very sensitive to small speed variations along the record length. An exception is TSA, which requires that order tracking (OT) be used as a prerequisite, and which converts the ‘frequency’ axis to be in terms of harmonic orders of a specified reference shaft speed. However, this is often just treated as a rescaling of the frequency axis, without a big change in the character of the signals, and is often still scaled in Hz (meaning average frequency over the record length).

However, for variable speed machines, this is no longer possible, since OT only removes the frequency modulation corresponding to the speed change, but not amplitude modulation, for example caused by the passage of gearmesh frequencies through fixed resonances. Thus, the SA will extract a signal with average amplitude, and when repeated periodically and subtracted, it will still leave the part of the original gear signal whose amplitude varies around the mean.

Bearing signals consist of a series of impulse responses of fixed length, independent of speed, though their spacing does vary with speed. Because they excite fixed resonance frequencies, it is standard practice to extract them in the frequency domain, by bandpass filtering a resonance band dominated by a particular bearing fault (usually at very high frequencies), and then obtain the envelope of the impulse response series by amplitude demodulation. It is the envelope signal which contains the required diagnostic information of the (low frequency) repetition rate of the IRs, so with variable speed machines, OT of the envelope signal must be used to obtain these bearing fault frequencies (BFFs) in terms of shaft orders. In principle, the enveloping can be done before or after OT, but it is more efficient to do it before, and this will also allow the envelope (and the tacho signal) to be downsampled before OT, which may be more efficient. Note that the (fixed) carrier frequencies are distorted by the OT, but these are in any case removed in the demodulation. The length of the individual IRs (constant in time) is also distorted by the OT, but this will not have a great effect on the envelope spectrum, which will reflect the slightly modified position of the centre of gravity of the individual pulses (relative to the spacing). This will have minimal effect on the first harmonic, but will give increasing slight smearing of higher harmonics (already smeared by slip).

7.3.4.1 Cases with Limited Deterministic Masking

Where there is little masking of bearing signals by deterministic signals, for example by gears, it may be possible to enhance them first in the time domain, and then perform OT (of the envelope signal) once into the angle domain, where the SES can be obtained.

This was the case for a conference presentation in 2014 [50], for which the approach and results are discussed first. The test rig used was a SpectraQuest® Bearing Prognostics Simulator (BPS), which allows for the testing of bearings over a wide range of speed and load, the latter applied horizontally to the test bearing by a hydraulic cylinder. This test comprised a complete run-up in speed from 0–40 Hz over 120 seconds, controlled by the BPS speed control unit, and the bearing mounted in the test housing had an outer race fault. Because from experience it was known that the signal measured directly on the bearing housing (position A) was very clear, measurements were also made at a couple of other points, including position C on the bedplate, remote from the test bearing.

Figure 7.57 is a spectrogram of the acceleration signal at Point A and of the tachometer signal, obtained by an optical probe sensing a reflective patch on the shaft (every 4th harmonic suppressed).

Horizontal bands in the acceleration spectrogram indicate resonances, of which there are two dominant ones near 1000 Hz and 2400 Hz in this signal. Moreover, the dominant harmonics in this

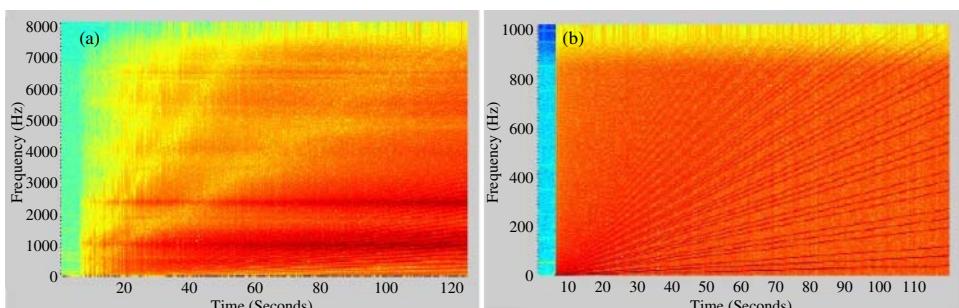


Figure 7.57 Spectrograms (a) Acceleration point A (b) tacho signal, decimated by 8.

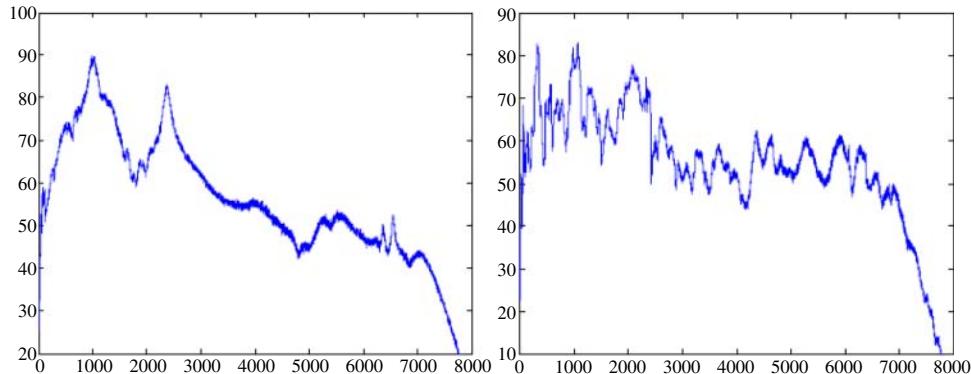


Figure 7.58 PSD spectra over run-up range (a) Acceleration point A (b) Acceleration point C.

signal are in fact the bearing outer race frequency (BPFO), which is unusual at low orders, but explained in this case because the bearing housing support has low resonance frequencies. Analysis was only performed over the range from 10% to 100% of full speed.

Figure 7.58 shows the PSD spectra of the response signals at Points A and C, integrated over the run-up range. For Point A the resonances near 1000 Hz and 2400 Hz are clearly seen, whereas at Point C on the bedplate, there are many more resonances, starting from about 250 Hz.

Three methods of pre-processing were applied separately in the time domain, to enhance the extraction of the bearing signals from the background, as follows:

- 1) Fast kurtogram (Section 5.5.3), to find the optimum passband, based on spectral kurtosis (SK), to maximise the impulsiveness of the bandpass filtered signal.
- 2) Minimum Entropy Deconvolution (MED) (Section 5.4), where an inverse filter is used to counteract the smearing effect of the transmission path from original source impulses to response signals.
- 3) Shortpass cepstral lifting, with an exponential window, to vastly reduce the effects of order related components while retaining modal information so that resonances carrying bearing fault signals would still be preserved (Section 6.3.2).

Figure 7.59 shows the resulting squared envelope signals from the two measurement points at two speeds (22.5% and 90% of full speed), and for the three pre-processing methods as well as the raw signal. This illustrates that the IRs have the same length at both speeds, but the spacing is four times wider at the lower speed. For the unprocessed signal the IRs are just about to overlap at near full speed for Point A, but are well and truly overlapping at Point C, and close to overlapping at 22.5% full speed. This is because there were many more modes on the bedplate, with the lowest frequency ones having longer IRs.

The much more complex transfer function at Point C also explains the different performances of the three processing methods, which are somewhat data dependent. All three performed well at Point A, in shortening the length of the IRs. Method 1 (SK) also performed well at Point C, because the band found by the kurtogram was at sufficiently high frequency (1.8 kHz, compared with 3.6 kHz at Point A) that the IR was only about twice as long. For Method 2, (MED), the number and wide range of resonances at Point C meant it was difficult to find an inverse filter, but even though the base of the IRs was almost as long as for the raw signal, the peaks were narrower and still protruded from

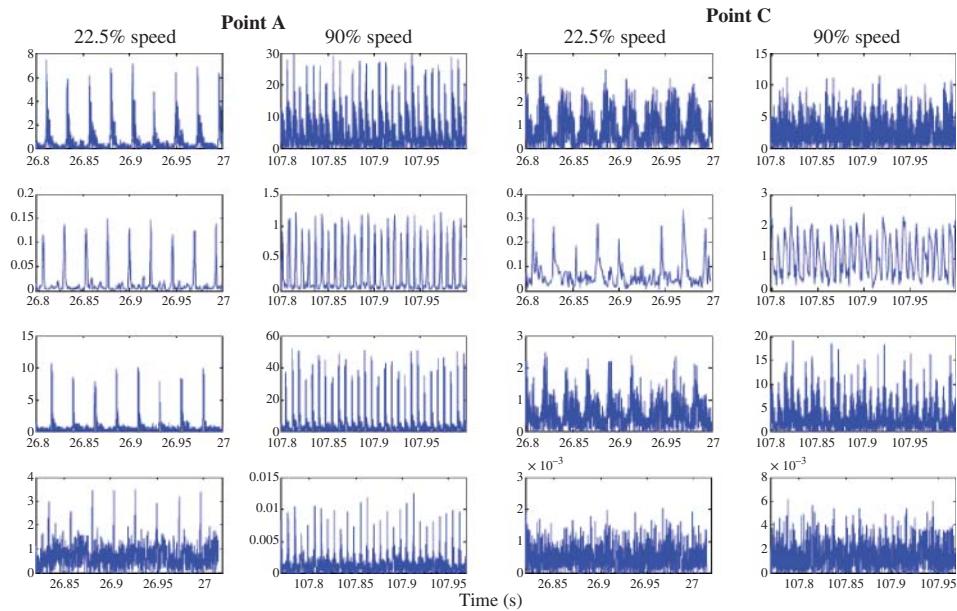


Figure 7.59 Signal envelopes for different methods (Row 1) Original; (Row 2) SK filtered; (Row 3) MED; (Row 4) Exponential liftered.

the base at near full speed. For Method 3 (exponential lifter), even though it gave the shortest IR at Point A, there was a higher noise level than for the other methods. At Point C, the low frequency (long IR) of its first resonance meant that the time constant of the exponential window had to be set even longer, and this appears to have increased the noise at both speeds.

Before and after pre-processing, the signals were all order tracked using the method of Section 5.1.4.1, requiring four frequency bands to cover the 10:1 speed range. The spectrograms of the order tracked raw signals are shown in Figure 7.60, and can be compared with Figure 7.57a. The orders are now horizontal lines, and the resonance bands follow hyperbolic curves. The x-axis is rotation angle in ‘rotations’, and it should be kept in mind that with speed increasing linearly with time, the rotation angle goes up with the square of speed, so that for example, the value at 25% of full scale (550 rotations) corresponds to half speed, i.e. 50% of full scale in Figure 7.57a.

Spectrograms were also formed from the squared envelope signals, and those for the raw signals from Points A and C are shown in Figure 7.61. Despite the appearance of the time signals in Figure 7.59, in particular for Point C, both spectrograms indicate that a satisfactory diagnosis could be made over most of the range, with several harmonics of BPFO (order 3.58) but with some low harmonics and side bands spaced at shaft speed.

The latter are discussed below, since shaft speed sidebands are unusual with outer race faults. Ref. [50] also shows the envelope spectrograms for signals from Point C, pre-processed by MED and exponential liftering, and these also show the bearing fault.

Figure 7.62 shows squared envelope spectra from points A and C without pre-processing, and from Point C after MED and exponential liftering, respectively. These are taken at roughly half speed (550 revolutions), and near full speed. All allow a clear diagnosis of the outer race bearing fault, but with some of the high speed ones also having sidebands spaced at shaft speed, but only at high speed. This is an indication that they are due to unbalance, increasing with the square of speed, and giving an inertial load rotating with the shaft.

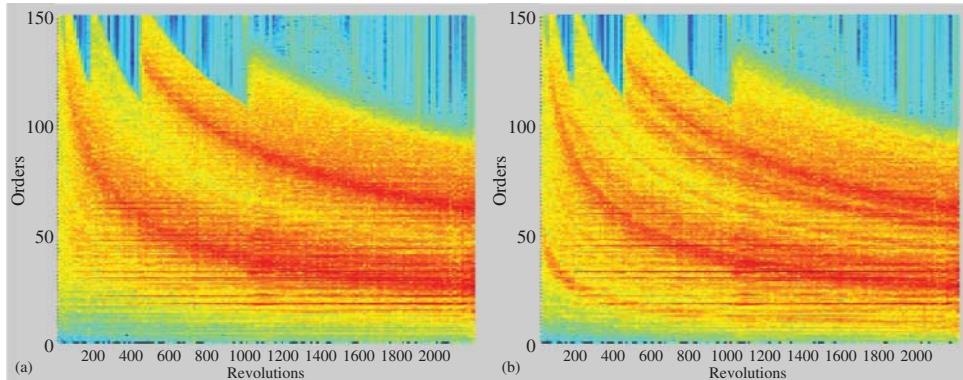


Figure 7.60 Spectrograms of order tracked raw signals (a) Point A (b) Point C.

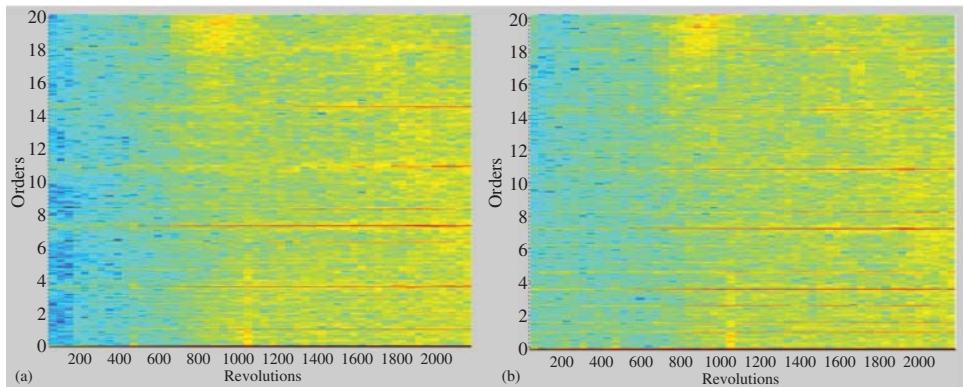


Figure 7.61 Envelope spectrograms (a) Point A (b) Point C.

It is perhaps worth mentioning here that since bearing signals under widely variable speed are (close to) cyclo-non-stationary, an alternative powerful analysis technique would be frequency/order spectral correlation or spectral coherence analysis, as discussed in Section 3.6.4. However, at this point in time, any pre-processing to remove strong masking by deterministic shaft speed related components would have to be done in the time domain beforehand.

7.3.4.2 Cases with Strong Deterministic Masking

A recent publication [51] takes up the challenge of diagnosing bearing faults under widely varying speed, with strong deterministic masking, specifically a tooth root crack in an input pinion of a gearbox. One of the bearings on the output shaft had a local inner race fault. The test rig is shown in Figure 7.63, and is in fact the same as used for the TE results of Figures 7.18, 7.23, and 7.24, with mild steel gears, and a ratio of 19 : 52 (the results in Figures 7.20 and 7.21 used the same casing, but different internals).

The acceleration results shown here were from a Brüel & Kjær accelerometer type 4396, mounted on top of casing, measuring vertically, in the middle on the brake side. A slot was seeded on one

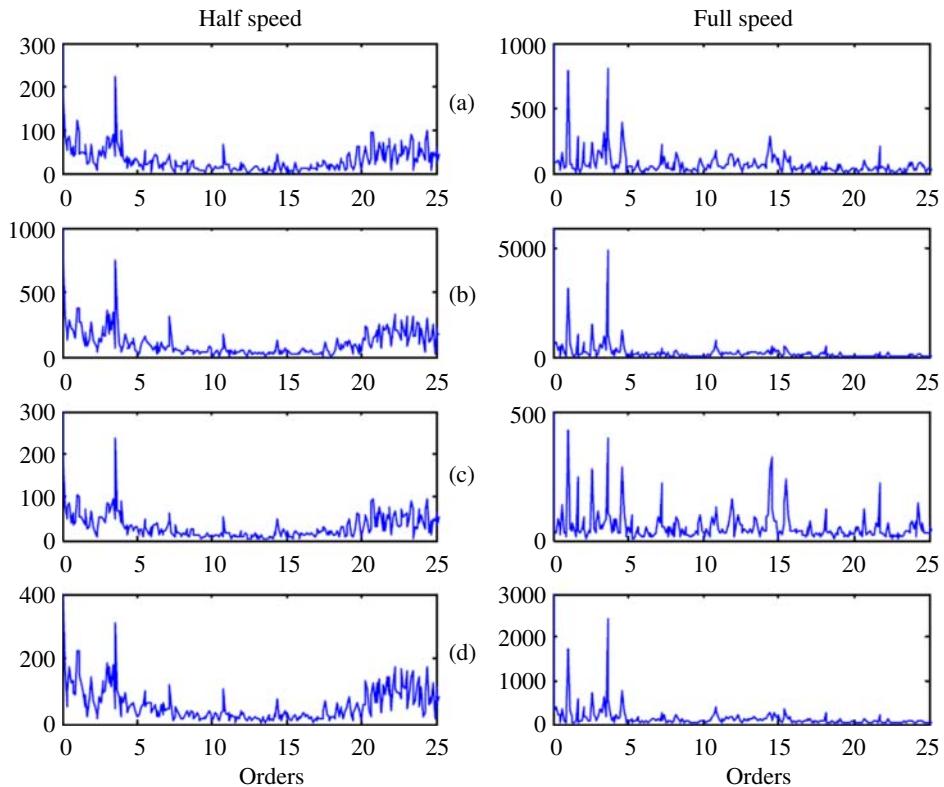


Figure 7.62 Envelope spectra at different positions along record; (a) Point A (b) Point C (c) C after MED (d) C after exponential filtering.

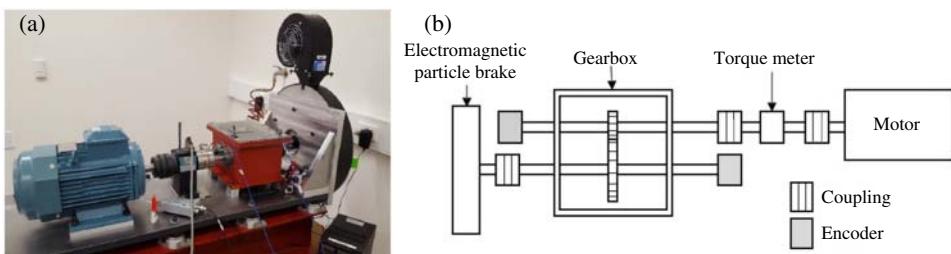


Figure 7.63 Gear test rig (a) Overall view (b) Schematic diagram.

tooth of the 19 tooth input pinion, by EDM machining, to simulate a tooth-root crack extending to the tooth centreline at 45° angle. Shaft encoders were mounted on the free ends of the input and output shafts, but only the 1 ppr (pulses per rev) tacho signals were used in this study, primarily for order tracking.

The faulty test bearing had an EDM-seeded inner race slot (~ 0.5 mm width) in the output shaft bearing at the brake end. It has ball diameter $d = 4.76$ mm, pitch circle dia $D = 23.5$ mm, and No. of rolling elements $n = 9$. The kinematic (no slip) estimate of BFFs are BPFO (outer race) = 3.5885,

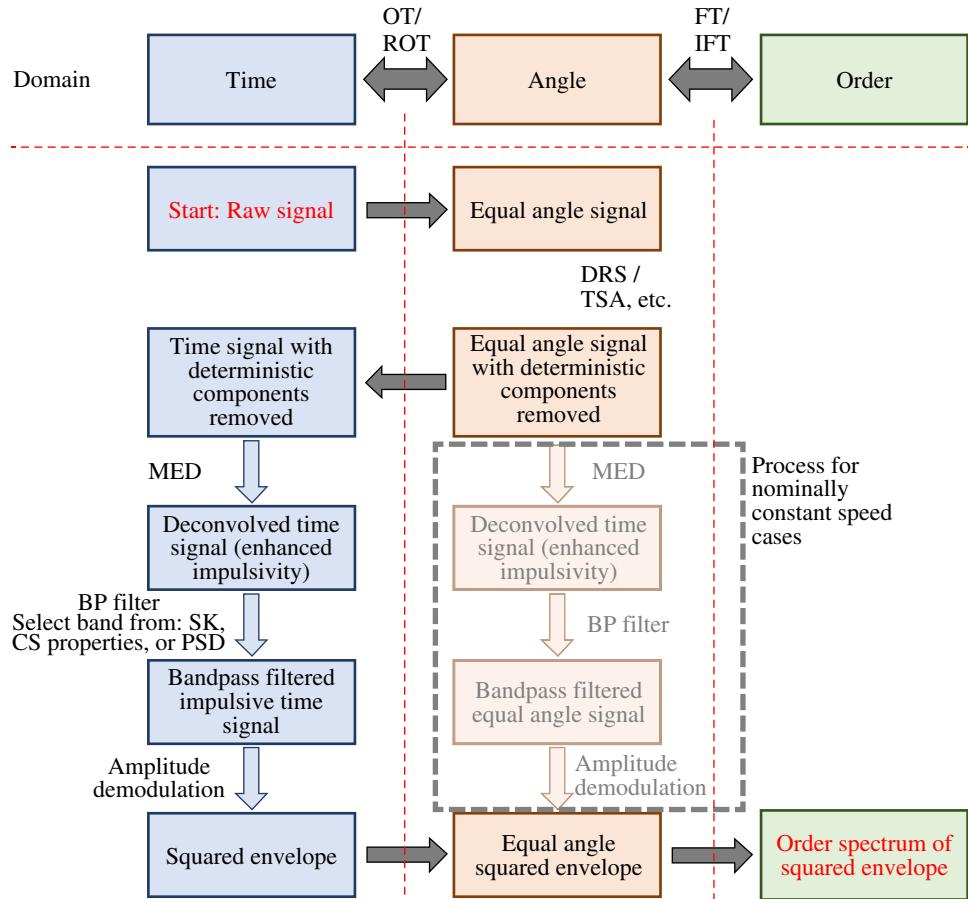


Figure 7.64 Application of various techniques in time or angle domains.

BPFI (inner race) = 5.4115, BSF (ball) = 2.3672 and FTF (cage speed) = 0.3987. Note that these ‘frequencies’ are stated in terms of shaft order, for ease of comparison between fixed speed and variable speed measurements.

Because of the strong gear signals expected (and experienced) from the faulty gear, it was anticipated that it would be necessary to use the procedure depicted in Figure 7.64 (originally from [52] and reproduced in [51]), requiring the signals to be transformed back and forth between the time and angle domains. For example, before being able to separate the bearing signals in the time domain, using MED, exponential filtering to enhance the modal (carrier) frequencies, or bandpass filtering based on a kurtogram or other log cycligram, it would very likely be necessary to first remove the strong gear signals in the angle domain, requiring an OT process, followed by reverse OT. Once the squared envelope of the optimally filtered bearing signal had been formed in the time domain, it would then have to be order tracked again into the angle domain for envelope analysis. Note that in general, synchronous averaging (shown in Figure 7.64 as TSA) is unlikely to be successful, for the reason given above; viz. that it would still leave some of the deterministic components because of amplitude modulation. Cepstral notching and DRS are less sensitive to this.

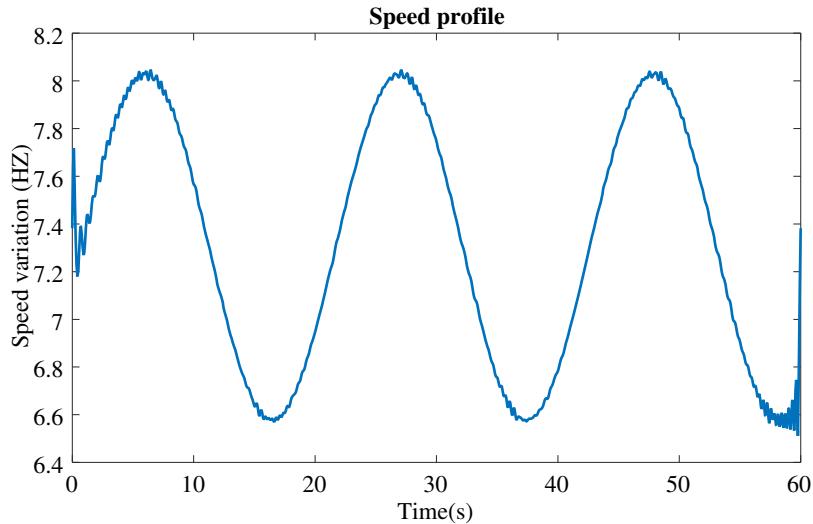


Figure 7.65 Speed profile for one variable speed test.

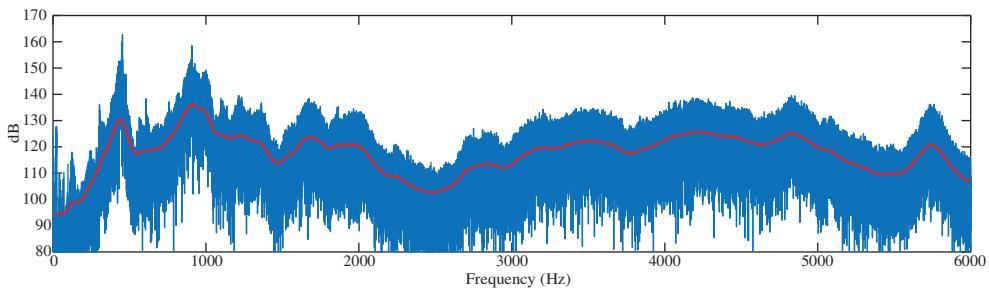


Figure 7.66 Result of applying exponential lifter (light) original (dark) liftered.

The choice of particular techniques for a given situation is data dependent, and only a couple are used for the present demonstration of the general procedure.

A series of tests was done, running the rig at a fixed input speed of 20 Hz, and a number of variable input speeds centred on 20 Hz, but only one of the variable speed cases is reproduced here, since it shows the extra advantage given by applying one technique, as discussed below. The mean speed of the output shaft, on which the faulty bearing was mounted, was 7.2 Hz. The case studied involved a frequency modulation of the output shaft speed by a $\pm 10\%$ sinusoidal variation of about 21 seconds period in a 60 seconds record, as shown in Figure 7.65. An exponential lifter was first applied in the time domain cepstrum, with a time constant of 5 ms, corresponding to a 3 dB bandwidth of 64 Hz. Figure 7.66 compares the original and filtered spectra, and it can be seen that the modal information is enhanced, and the first few (smeared) harmonics below 100 Hz apparently suppressed.

Even so, after order tracking, the order spectrum had clear harmonics of both the input shaft speed and the gearmesh frequency (19th harmonic), so these had not been fully removed by the exponential lifter in the time domain. A comb notch lifter was then applied in the order domain cepstrum, designed to remove the ‘raharmonics’ (evenly spaced components in the cepstrum) of the

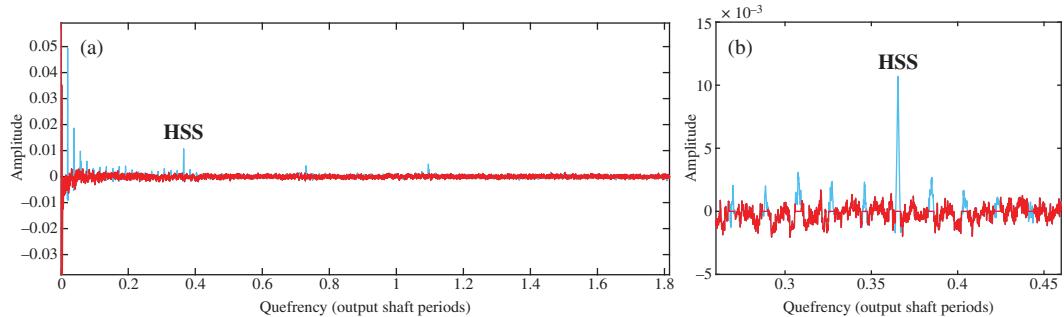


Figure 7.67 Order domain cepstra before and after notch filtering to remove discrete components. (a) Quefrency range including three rahmonics of HSS (b) Zoom around HSS.

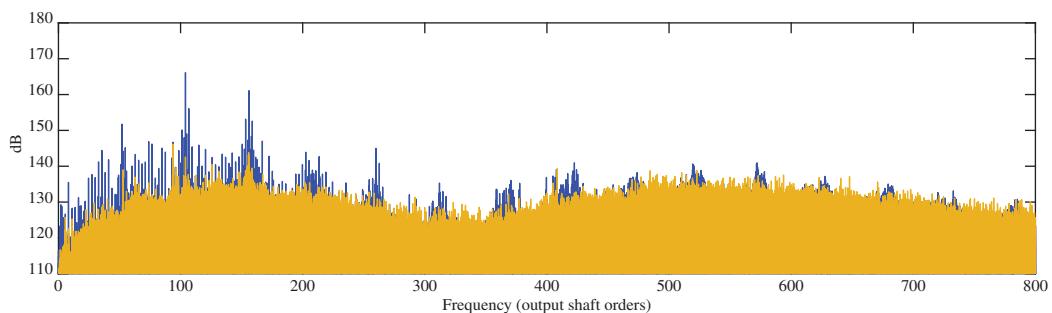


Figure 7.68 Order domain spectra before and after notch filtering to remove discrete components.

gearmesh period. Figure 7.67 compares the cepstra before and after this notch liftering, over a range encompassing the 3rd harmonic of the high speed shaft (HSS), and also a zoom around the 1st harmonic. It can be seen that the harmonic corresponding to the output shaft is negligible, because the tooth fault produces impulses at input shaft speed. It can also be seen that the HSS component is the 19th harmonic of the gearmesh quefrency, as expected, so both have been removed with this notch lifter. The output shaft (1 period) would also have been removed if it had been present, as it is the 52nd harmonic of the gearmesh quefrency.

Figure 7.68 shows the resulting order spectra before and after notch liftering. It is evident that the additive harmonics have been removed.

However, when the SES of this signal is obtained (Figure 7.69), it is found to have considerable contamination by the harmonics of the two shaft speeds. It is possible to make the diagnosis of inner race fault, even though the first harmonic of BPFI (order 5.397 in this case), is very close to the second harmonic of the high speed shaft (HSS), whose exact order, determined by the gear ratio is 5.474, though lower in level than it is. Not only are three harmonics of BPFI identified by the harmonic cursor, but each is surrounded by sidebands spaced at output shaft order 1.

It was thought that the strong input shaft speed components were likely due to multiplicative (modulation), as opposed to additive, effects of the gear fault on the bearing, so it was decided to re-process the signal, as indicated in Figure 7.64, by reverse order tracking to the time domain, once the additive effects had been removed. The (squared) envelope signal could then be processed to try to remove the modulation effects. Exponential filtering was again used in the time domain, to

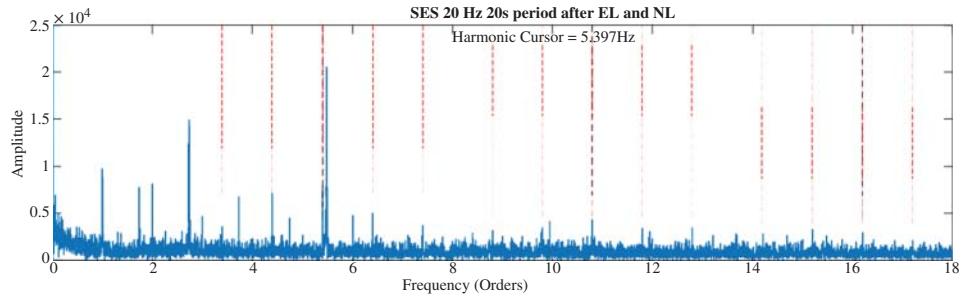


Figure 7.69 SES for 20 Hz variable speed (21 seconds period) after exponential and notch lifters.

further enhance the modal response, and the fast kurtogram was applied to the filtered signal, to see if a particular resonance band might preferentially amplify the bearing fault signal. A band was found, but the kurtosis was not very high. Envelope signals were generated for the signals before and after bandpass filtering according to the kurtogram, and both were order tracked again. In the order domain, the signals were again notch filtered to remove harmonics of the input shaft speed, and the resulting SESs are presented in Figure 7.70.

Although the second harmonic of HSS is still stronger than BPFI in Figure 7.70a, the latter is much stronger than in Figure 7.69, and the higher harmonics and shaft speed sidebands around them much clearer.

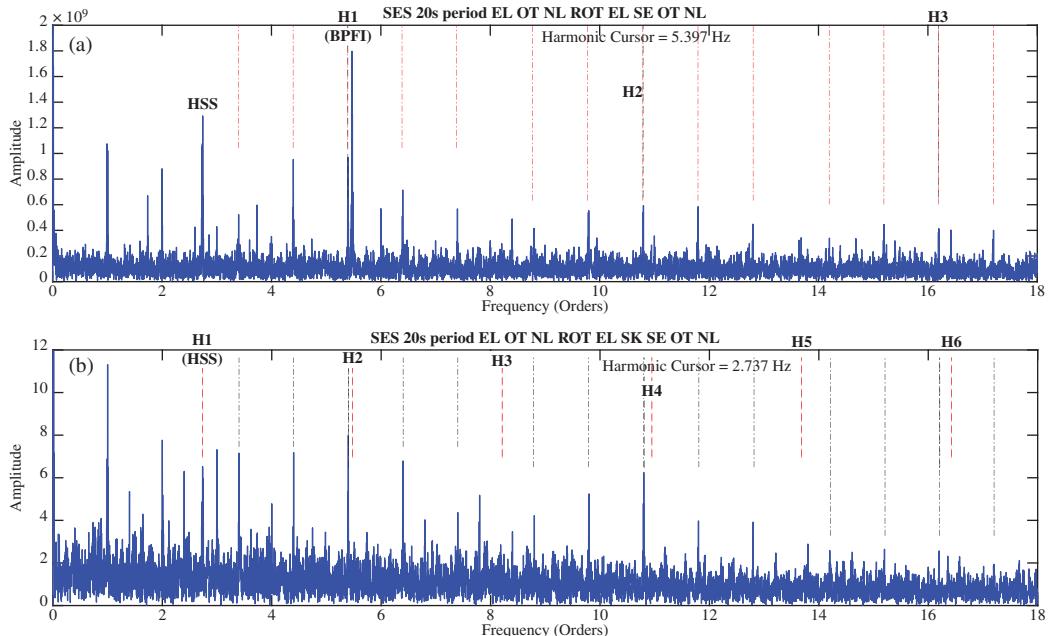


Figure 7.70 Squared envelope spectra (SEEs) for 20 Hz variable speed (20 second period) after second stage processing of the envelope signal (a) not using kurtogram (b) using kurtogram.

However, in Figure 7.70b, based on the bandpass filtered signal from the kurtogram, the harmonic cursor set on HSS finds only the first harmonic, whereas the rest of the SES shows only the bearing fault information, harmonics of BPFI, and sidebands spaced at shaft speed, order 1. Evidently, this procedure has completely isolated the bearing fault, even in the presence of strong masking by the gear fault.

Ref. [51] also presents results for two other speed profiles, one a scaled version of a random profile from a wind turbine, and the other a short length (10 s) at 20 Hz constant input speed, followed by 50 s sinusoidal modulation with a period of 60 s. Both of these gave similar results to Figure 7.70a for the case not using the kurtogram, but the latter introduced more noise, so that the SES was unclear.

Future work will test whether other bearing separation algorithms, such as the methods described in refs. [46–49] of Section 7.3.3, give improved separation of the bearing signal, so that a result as in Figure 7.70b can be achieved more generally.

7.4 Reciprocating Machine and IC Engine Diagnostics

This is far less developed than for rotating machines, but a number of techniques have been developed, based on different approaches. Some are primarily aimed at diagnosing combustion faults or equivalent pressure related anomalies in compressors and pumps, while others detect mechanical faults, or both. One of two main approaches uses some form of time/frequency analysis, and can diagnose pressure related or mechanical faults, while the other approach attempts to reconstruct cylinder pressure (or pressure torque) from external measurements of acceleration or torsional vibration.

More powerful methods have recently been developed, but are based on simulation models of the engine, and so are discussed in Section 8.4.

7.4.1 Time/Frequency Methods

The averaged STFT method described in Section 2.3.1 can be used diagnostically as well as to detect change, as in Section 4.3. As an example, Figure 4.13 shows the differences in both frequency content and the crank angles at which they occur for a gas engine with a misfire in one cylinder. With a knowledge of the timing of various events, and some experience with the machine, it would be recognised that the detected change comes from a greatly reduced exhaust gas flow through the exhaust valve. Similarly, with Figure 4.14 the change in the diagrams is near TDC, but at higher frequency than combustion signals in this diesel engine, so with some experience the difference could be diagnosed as most likely due to increased piston slap.

As mentioned in Section 3.6.3, the Wigner-Ville spectrum (WVS) has the same resolution as the Wigner-Ville distribution (WVD) but with reduced interference terms, as long as the signal is second order cyclostationary (by removal of periodic components). It has the further advantage that the averaging used to reduce the interference simultaneously gives a smoothing of results over several cycles, as for the averaged STFT method just discussed, and so would in general be a better diagnostic tool. Figure 7.71 gives some results from Antoni's PhD thesis (reported in [53]), showing the advantage of the WVS (Figure 7.71b) over the WVD (Figure 7.71a) for the same signal. Figure 7.71c,d, respectively, show the effects of a clogged and an open injector, both in the third firing cylinder.

More recently, in [54] Antoni showed the application to the diagnostics of a reciprocating compressor. The WVS averaged over 102 cycles (Figure 7.72a), has approximately twice as much resolution

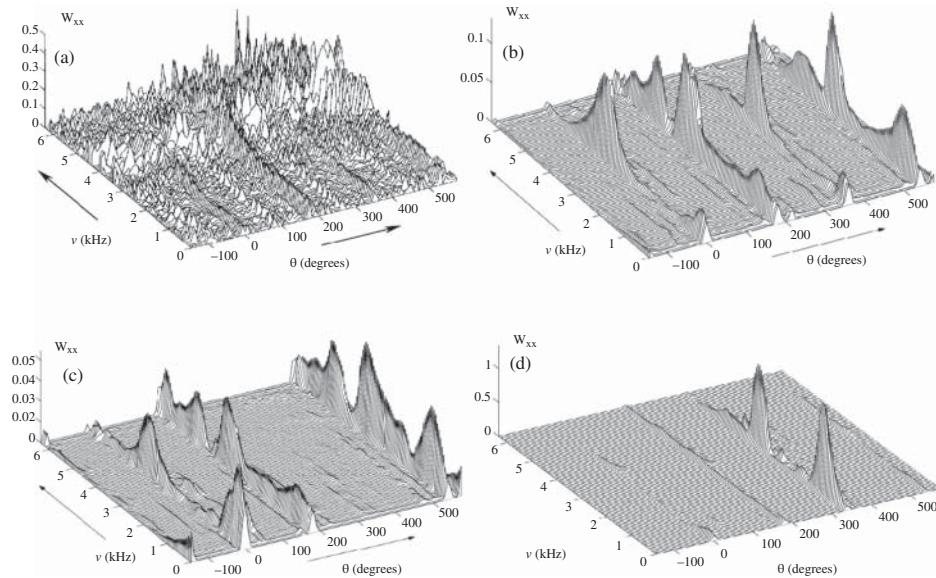


Figure 7.71 Advantage of WVS for diesel engine diagnostics (a) WVD for normal condition (including deterministic components) (b) WVS for normal condition (deterministic components removed) (c) WVS with blocked injector in one cylinder (d) WVS with open injector.

in both time and frequency directions as the STFT of Figure 7.72b. The WVD (for one cycle) of Figure 7.72e has about the same resolution as the WVS, but it is difficult to determine which components represent interference terms. The time/frequency representation of the WVS gives much more information about the different events than the time signal of Figure 7.72d.

Reference [54] shows the compressor in normal condition, while describing the analysis in great detail, but Reference [55] describes the application of the method to a number of faults in a compressor.

The above results show the difficulty in using time/frequency diagrams for diagnosis; even if the eye can see differences in the diagrams, it is difficult to extract the information without resorting to image analysis or similar. A number of people have tried to extract the information in a form that can be used to train neural networks to recognise the condition.

One of the earliest was [56], which used the WVD to visually compare the results for different faults. The local energy in the time/frequency diagram was integrated in a number of patches, and the resulting array used to train the network. Alternatively, the signal was modelled with an evolutive AR model, and the model parameters used to classify the condition.

Another way of extracting time/frequency information, for diagnostics of reciprocating machines, is to use the moments and cumulants of the WVD, as proposed in [53]. Thus, at each point in time the different order moments and cumulants of the frequency spectrum are calculated. For example, the first order moment is the ‘centre of gravity’ and can be considered as the ‘average frequency’. If a high frequency burst suddenly appears the first moment will increase dramatically. The second order cumulant (centred moment) represents the ‘radius of gyration’ about the ‘average frequency’ and thus the ‘average bandwidth’, which for a single frequency peak would be related to damping. In fact, as pointed out in [53], these moments and cumulants of the WVD can be calculated directly from the time signals, without having to produce the full WVD.

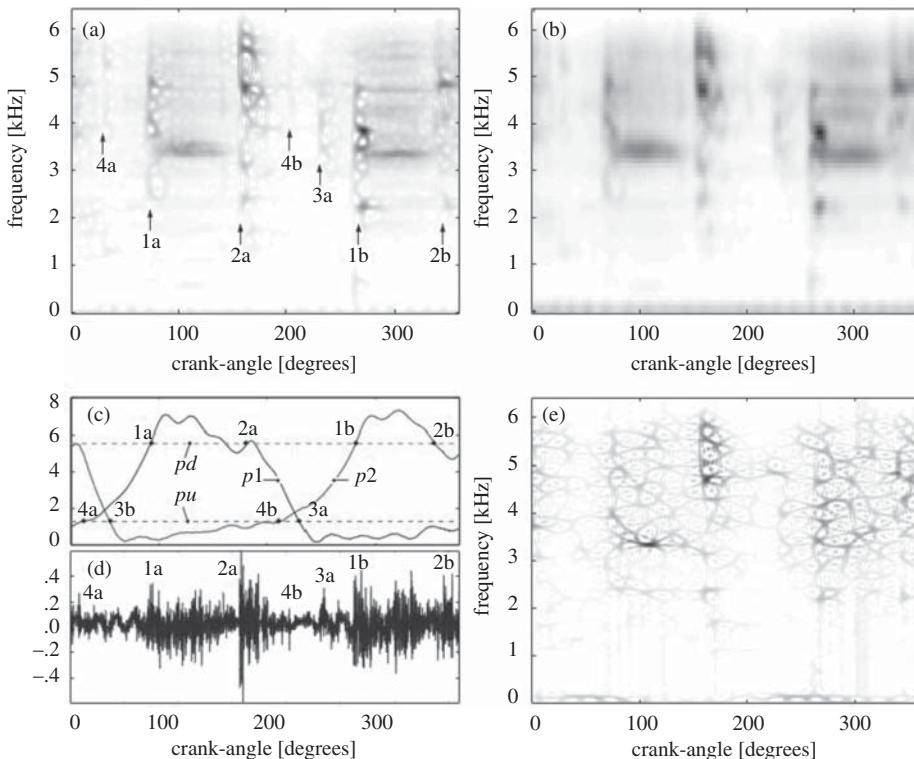


Figure 7.72 Diagnostics of a reciprocating compressor (a) WVS (b) STFT (c) Pressure on forward and backward strokes (d) Accelerometer signal (e) WVD for one cycle. Legend: a – forward stroke, b – backward stroke. 1, 2 – opening, closing of discharge valve. 3, 4 – opening, closing of suction valve.

7.4.2 Cylinder Pressure Identification

Cylinder pressure in an IC engine gives a considerable amount of information about the condition of the engine, in particular with respect to combustion related faults. It would be ideal to measure the cylinder pressure directly, but even though pressure transducers are often mounted in engines in the laboratory, they are not practical in operating engines in the field. Not only are the transducers expensive, but because of the very harsh conditions to which they are subjected, they have a quite limited useful life. A lot of research effort has therefore gone into attempts to determine the cylinder pressure from external measurements. The latter include accelerometer measurements on the engine cylinder head or block, and torsional vibration measurements of the crankshaft.

7.4.2.1 From Acceleration Measurement

One of the earliest attempts to use inverse filtering to determine cylinder pressure was the use of the cepstrum to extract diesel engine cylinder pressure signals from acoustic measurements [57]. A later attempt to achieve a similar result using a Wiener filter was reported in [58].

As discussed in connection with gears in Section 7.2.3.2, and more generally in Section 6.2.3, the cepstrum has the possibility of separating a measured response into its forcing function and transfer function components, as long as the response is dominated by a single source. Determination of an

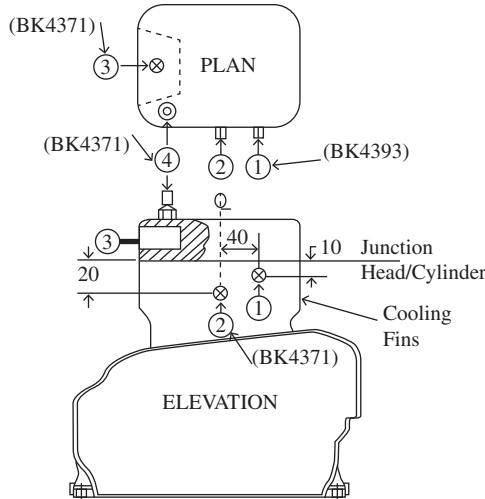


Figure 7.73 Single cylinder air-cooled diesel engine with accelerometer placement.

inverse filter is ideally based on measurement of the pressure signal, but could perhaps be done on a test engine in the laboratory and then applied more widely, to engines without pressure transducers mounted.

The results of some collaborative work between the University of New South Wales and the University of Rome were reported in [59, 60]. They were based on measurements on a single cylinder air-cooled diesel engine, shown schematically in Figure 7.73. The cooling fins on the head and cylinder are not shown, but it meant that it was possible to mount an accelerometer (No. 3) directly on the head, in an access cavity for the injector.

This would not normally be possible on a water-cooled engine. Another accelerometer (No. 4) was placed on top of a cylinder head bolt.

Since accelerometers 3 and 4 would be sensitive to inertial forces in the plane of the crank, two more accelerometers (Nos. 1 and 2) were placed on the cylinder wall normal to this plane. A pressure transducer was mounted in the cylinder.

Figure 7.74 (from [59]) shows a typical pressure signal ($p(t)$), acceleration signal ($a(t)$) and the Hanning tapered rectangular time window ($w(t)$) used to separate them from extraneous noise in other parts of the cycle. The FRFs between the pressure signal and all four accelerometer signals were measured over a number of cycles, along with the corresponding coherence, the latter giving a measure of the linearity of the relationship as a function of frequency. The coherence for accelerometers 2 and 4 is compared in Figure 7.75, and it is seen that it is considerably better at accelerometer 4 (on the head bolt). Accelerometer 3 (on the head) was similar to, but no better than No. 4, while accelerometer 1 gave similar results to accelerometer 2.

The good coherence indicates that the inverted FRF can be used as an inverse filter to convert acceleration into pressure, and Figure 7.76 shows the results of doing this for operation at one speed (2400 rpm) and three loads (50%, 75%, 100%). This shows that the inverse filter generated for any of the loads works well for the others. It was found that such a filter did not work for other speeds, however. Figure 7.77 shows the coherence for an FRF averaged over the same three loads and for two speeds (2400 rpm, 3000 rpm). It is somewhat lower, but stays above 80% over the frequency range up to 3 kHz. Figure 7.78 shows that the corresponding inverse filter performs well for the full range

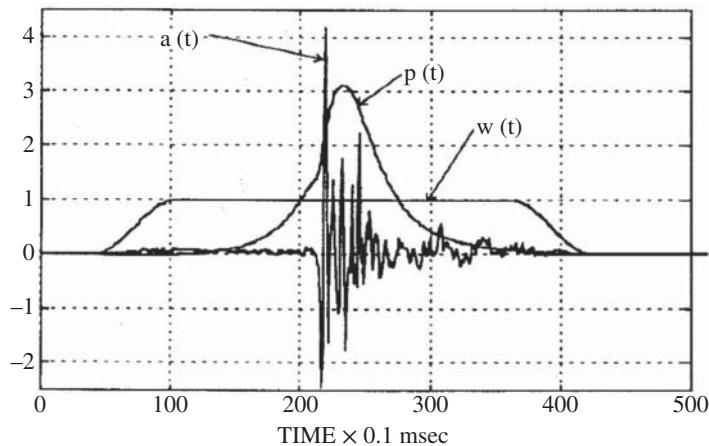


Figure 7.74 Typical pressure and acceleration signals and the time window used to extract them.

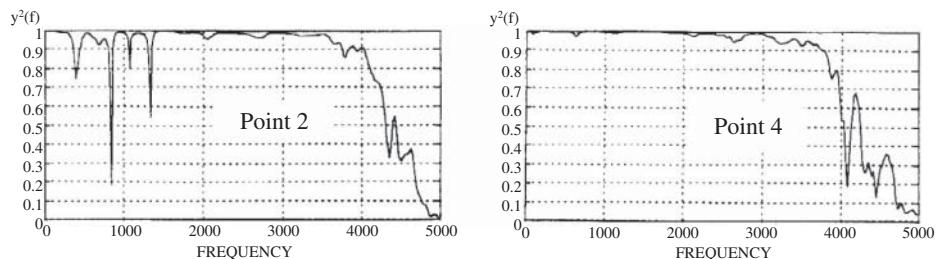


Figure 7.75 Coherence between pressure and acceleration at different measurement points.

of speeds and loads for which it was generated. Once again it did not function for speeds outside this range. It should be kept in mind that part of the pressure signal is related to crank angle rather than time (e.g. compression/expansion), but part is dependent on time (e.g. the chemical reaction during combustion) and so it would be difficult to have an inverse filter which could be used over a speed range.

Because of the potential for avoiding direct measurement of the pressure signal, both [59] and [60] compared the frequency domain methods and cepstral methods for forming the inverse filter. This culminated in the PhD thesis work of Yaping Ren [61], and a result is shown in Figure 7.79 of the use of the differential cepstrum (Eq. (6.11)) to extract the pressure signal from external acceleration measurements, in order to determine the amount of knock. Ren curve-fitted the differential cepstrum of the response for the poles and zeros of the transfer function, and used this to generate an inverse filter.

One of the problems associated with inverse filtering is that it usually involves cancellation of poles with zeros, and if these are not completely aligned, the results may be poor. In [62] Gao and Randall showed that this could be alleviated by using a special smoothing technique. This made use of an exponential window to effectively move the poles and zeros away from the imaginary axis in the Laplace plane, and reduce the effects of the mismatch.

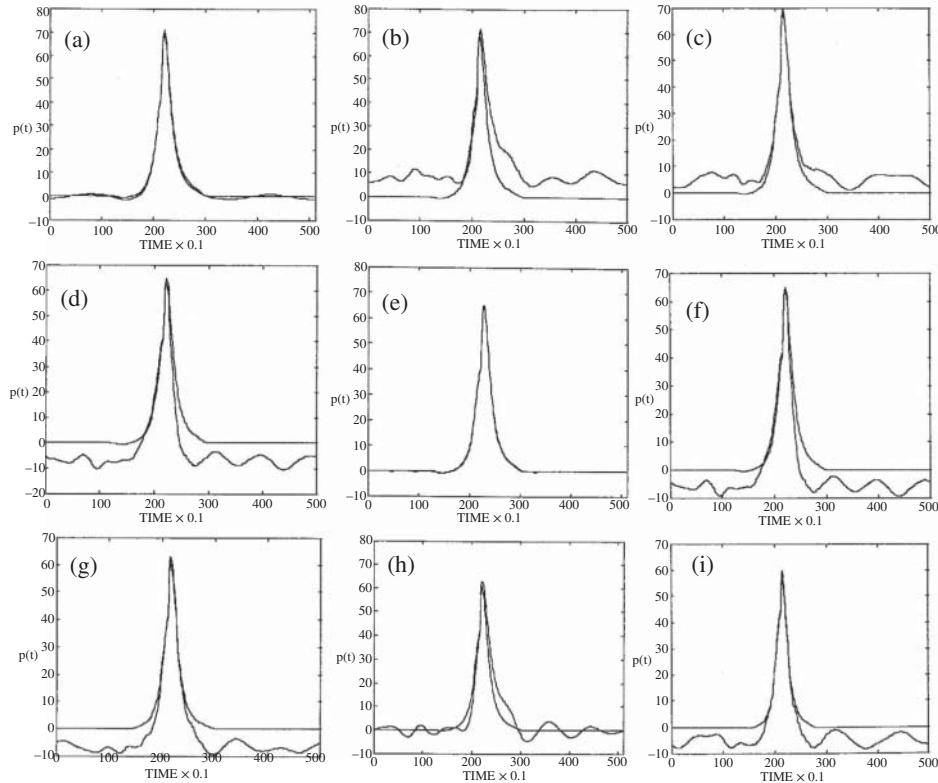


Figure 7.76 Results of inverse filtering for 2400 rpm and three loads (a,d,g) Filter generated at 100% load (b,e,h) Filter generated at 75% load (c,f,i) Filter generated at 50% load (a,b,c) Signal recovered for 100% load (d,e,f) Signal recovered for 75% load (g,h,i) Signal recovered for 50% load.

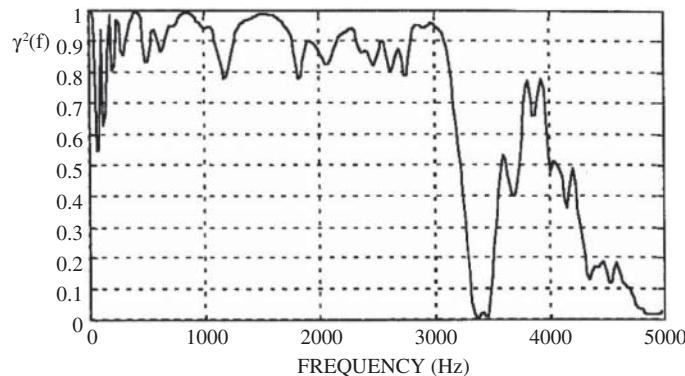


Figure 7.77 Coherence for the FRF averaged over two speeds and three loads.

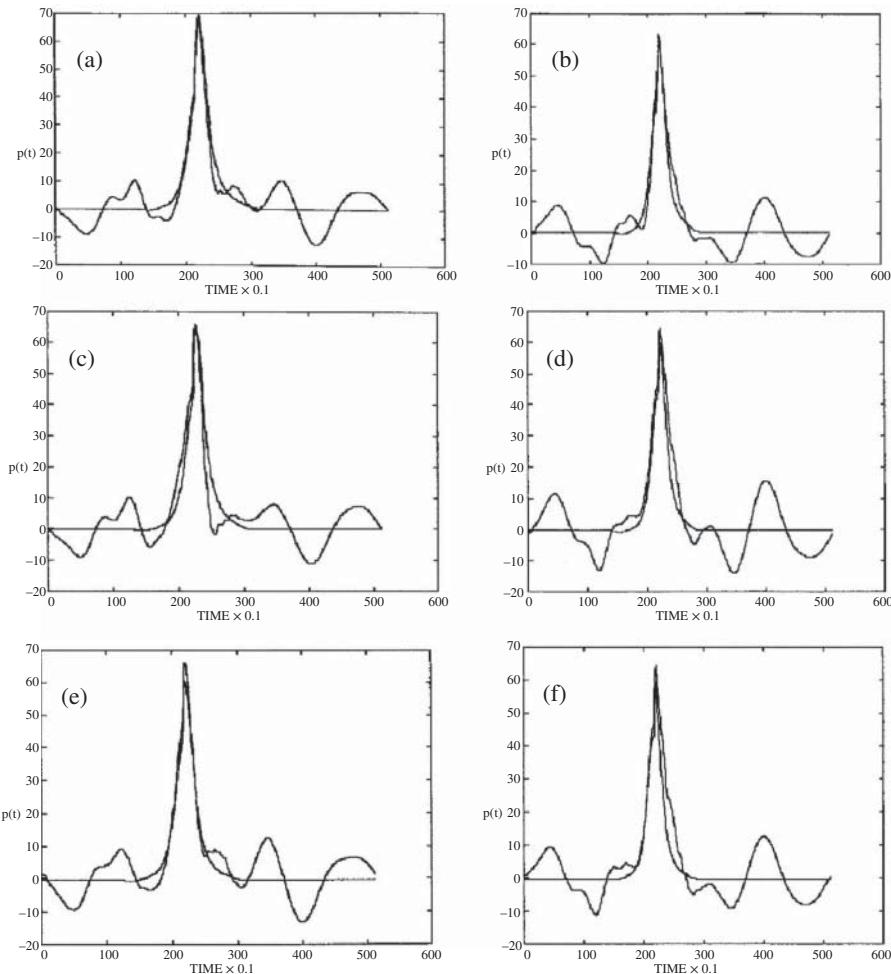


Figure 7.78 Results of inverse filtering for filter averaged over two speeds and three loads (a) 2400 rpm, 100% load (b) 3000 rpm, 100% load (c) 2400 rpm, 75% load (d) 3000 rpm, 75% load (e) 2400 rpm, 50% load (f) 3000 rpm, 50% load.

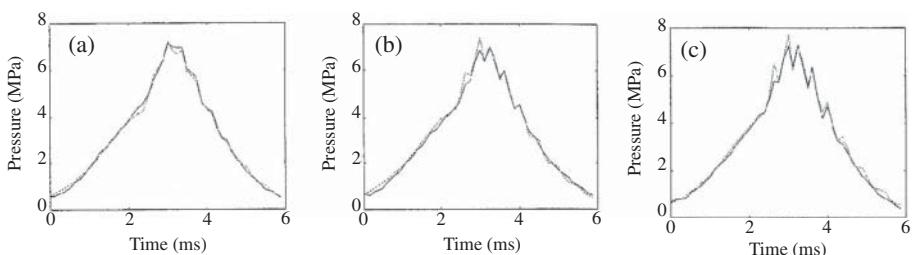


Figure 7.79 Reconstructed (dotted) and measured (solid) pressure signals for (a) Light knock (b) Medium knock (c) Heavy knock. Source: From [61].

Two PhD theses from Luleå University of Technology in Sweden [63, 64] extended this work, in research programs in conjunction with SAAB-Scania Trucks. Aspects of this work can be found in Refs. [63–67]. Zurita [63, 65] extended the cepstral method, using a window similar to that suggested in [62] (combined with a smooth taper at the leading edge to avoid the step at the beginning of an exponential window) and compared the results with those obtained using an approach based on Multivariate Analysis. He showed that the start of combustion could be determined accurately from the reconstructed pressure signals [65]. Johnsson [64, 66, 67] modified Zurita's window and obtained better reconstructions by combining the contributions from accelerometer (high frequency) and crankshaft torsional vibration measurements (low frequency), deciding on a frequency cutoff based on the coherence of the measured pressure with the respective signals. Because of the nonlinear relationship, in particular between pressure and torsional vibration, he employed radial basis function neural networks, originally suggested by researchers at Manchester University [68], to generate the inverse filters, with quite good results. However, even though the neural networks were able to produce good results by ‘interpolation’ among a range of results, it is much less likely that they would be able to reproduce unusual ‘fault’ conditions, deviating from the original training set.

The combination of acceleration and torsional vibration measurements is taken up again in Section 7.4.2.3 after a discussion of the latter.

7.4.2.2 From Torsional Vibration Measurements

It has already been pointed out in Sections 2.3.2 and 4.3.3 that the torsional vibration of the crankshaft reacts directly to any variations in pressure in the cylinders, and so quite a bit of research has gone into attempts to find inverse filters to determine cylinder pressure from torsional vibration.

The situation is complicated by the very nonlinear nature of the relationship between them. The moment arm relating torque to pressure is given by

$$r \sin \theta \left[1 + \frac{r \cos \theta}{\sqrt{L^2 - r^2 \sin^2 \theta}} \right] \quad (7.2)$$

in terms of the crank angle θ , crank radius r and connecting rod length L , and is seen to be very nonlinear, though well-known. It is quite easy to convert pressure to torque, but very difficult to perform the inverse, because the relationship is singular at top and bottom dead centre (TDC and BDC). It can be argued that it is better to use pressure torque rather than pressure as the condition variable.

Another complicating factor is the considerable contribution of inertial torque to the overall crankshaft drive torque, in particular at high speeds. Once again, while highly nonlinear, this is easily calculable however, and Ref. [69] has proposed the use of ‘synthetic’ angular velocity and acceleration, where the inertial effects are removed so as to make the relationship between drive torque and these synthetic variables linear.

Most of the work on relating torsional vibration to drive torque has assumed an effectively rigid crankshaft, so that the torque pulses from any cylinder have the same effect on the angular velocity. For large multi-cylinder engines, however, the crankshaft can have torsional resonances within the frequency range of the significant excitations, so that the effect of pressure variations in one cylinder has a different influence on the response vibrations depending on the measurement point.

This problem was recently investigated [70], but since it uses a simulation model of the engine crankshaft, it is discussed in detail in Section 8.4.1.

As regards measurement of the crankshaft torsional vibrations, there are three main possibilities. The measurements of Ref. [70] were made with a laser torsional vibrometer (Polytec OFV-400) but these are very expensive. A comparison was in fact made in that case with the results of frequency demodulation using a shaft encoder mounted close to the measurement point, and they were very similar. Small discrepancies could be attributed to the fact that the laser vibrometer measured the absolute torsional vibration, including rocking of the whole engine, whereas the shaft encoder gave the relative torsional vibration of the crankshaft to the engine frame. The latter is more relevant in this case, since it is the parameter most directly related to the cylinder pressure. As mentioned in Section 7.2.2, the frequency demodulation of an encoder signal can be done by two methods (illustrated in Figure 7.11). The method based on pulse timing has the advantage that it gives a fixed number of samples per revolution, and thus obviates the need for order tracking. On the other hand, since it gives a map of rotation angle vs time, it can be used if desired to convert from angular to temporal sampling.

7.4.2.3 Using Acceleration and Torsional Vibration Measurements

Despite the theoretical frequency range of torsional vibrations being up to 10 kHz, most of the successful applications of torsional vibration to detection of combustion anomalies have used changes at relatively low order, such as given by complete or partial misfires. As an example, the optimum parameters found for the feature vectors of the neural networks in Ref. [70] were the complex values of the first and fourth harmonics of engine cycle frequency. Johnsson [64, 66, 67] found that the coherence of torsional vibration signals with cylinder pressure was highest at low frequencies up to 20 orders, but that the coherence of acceleration signals (measured on head bolts) was better at higher frequencies, while poorer at lower frequency. It is possible that the coherence of pressure torque with torsional vibration (using synthetic angular acceleration with removal of inertial effects) would be valid to a higher frequency, but it does seem likely that, as recommended by Johnsson, the best way of estimating cylinder pressure from external measurements would use torsional vibration for the low frequency part of the signal, and accelerometer signals for the high frequency part. This remains to be fully demonstrated.

It also seems likely that a more reliable result would be obtained in terms of pressure torque rather than pressure as such, and this might be an equally valid parameter to use in reciprocating machine diagnostics.

7.4.3 Mechanical Fault Identification

Mechanical faults are here interpreted as impacts caused by wear of components giving greater clearances than normal. These would often be detected by the time/frequency methods of Section 7.4.1, but more specific fault information can be obtained by other methods. In comparison with combustion, the mechanical fault responses tend to be more impulsive, and can often be better detected by envelope analysis of the high frequency responses in a similar manner to bearing faults. This approach is used in connection with simulated faults in Sections 8.4.2 and 8.4.3.

A useful approach in some circumstances is to use BSS techniques to separate different components which may indicate a fault. BSS techniques usually require the signal components to be separated to have different statistical properties, and that does apply for example to combustion and piston slap in a diesel engine. In Ref. [71], the separation is made on the basis of the blind least mean square (BLMS) algorithm with Gray's variable norm as cost function. Only one source could be separated at a time, but a parameter could be set to optimise the result if this source were

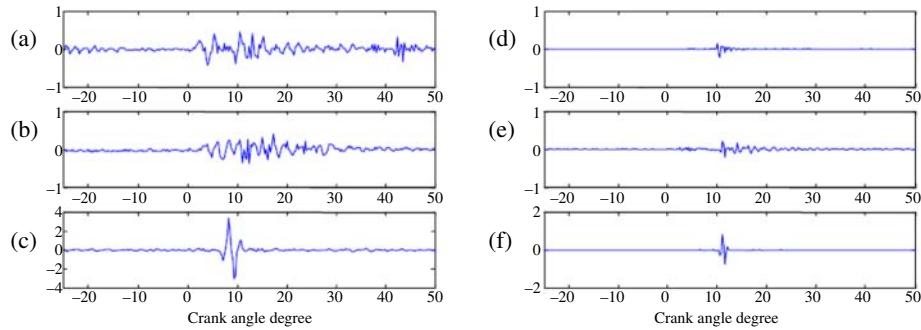


Figure 7.80 Recovered piston slap using faulty piston: a, b; Two of the three measured signals when there are fuel injection and combustion; c; Recovered piston slap; d, e; Two of the three measured signals when there is no fuel injection; f; Recovered piston slap.

super-Gaussian, with high kurtosis, such as piston slap, or sub-Gaussian, such as cylinder pressure. A result is shown in Figure 7.80, for external acceleration signals, which demonstrates that a supposed component due to piston slap could be separated in the presence of combustion, and a very similar result (though with reduced amplitude) was obtained in the absence of combustion in the same cylinder. Piston slap impact velocity is increased with increased cylinder pressure, since this is what drives the piston across the clearance gap. This explains the difference in scaling in the two cases, and also the slightly longer delay for the non-combustion case.

A later paper [72] shows how progressively weaker components can be extracted by successively removing the stronger components using a deflation method.

References

1. Sawalhi, N. and Randall, R.B. (2014). Gear parameter identification in a wind turbine gearbox using vibration signals. *Mechanical Systems and Signal Processing* 42: 368–376.
2. Coats, M.D. and Randall, R.B. (2014). Single and multi-stage phase demodulation based order-tracking. *Mechanical Systems and Signal Processing* 44 (1–2): 86–117.
3. Stewart, R.M. (1977). “Some useful data analysis techniques for gearbox diagnostics. *Proceedings of the Meeting on the Applications of Time Series Analysis, ISVR, University of Southampton*, Southampton, UK (19–22 September 1977), Paper #18.
4. Wang, W. and Wong, A.K. (2002). Autoregressive model-based gear fault diagnosis. *Transactions of the ASME, Journal of Vibration and Acoustics* 124: 172–179.
5. McFadden, P.D. (1991). A technique for calculating the time domain averages of the vibration of the individual planet gears and the sun gear in an epicyclic gearbox. *Journal of Sound and Vibration* 144 (1): 163–172.
6. McFadden, P.D. (1994). Window functions for the calculation of the time domain averages of the vibration of the individual planet gears and sun gear in an epicyclic gearbox. *Transactions of the ASME, Journal of Vibration and Acoustics* 116: 179–187.
7. Forrester, D. and Blunt, D. (2003). “Analysis of epicyclic gearbox vibration.” *DSTO HUMS Conference*, Melbourne (17–18 February).
8. Peng, D., Smith, W.A., Randall, R.B., and Peng, Z. (2019). Use of mesh phasing to locate faulty planet gears. *Mechanical Systems and Signal Processing* 116: 12–24.
9. McFadden, P.D. (1986). Detecting fatigue cracks in gears by amplitude and phase demodulation of the meshing vibration. *Transactions of the ASME, Journal of Vibration, Acoustics, Stress and Reliability in Design* 108: 165–170.
10. Sweeney, P.J. and Randall, R.B. (1996). Gear transmission error measurement using phase demodulation. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 210 (C3): 201–213.
11. Du, S. and Randall, R.B. (1998). Encoder error analysis in gear transmission error measurement. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 212 (C4): 277–285.

12. Du, S., Randall, R.B., and Kelly, D.W. (1998). Modelling of spur gear mesh stiffness and static transmission error. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 12 (C4): 287–297.
13. Randall, R.B., Peng, D., Smith, W.A. (2019). “Using measured transmission error for diagnostics of gears”, *SIRM conference*, Danish Technical University, Copenhagen (13–15 February).
14. Chang, H., Borghesani, P., Smith, W.A., and Peng, Z. (2019). Application of surface replication combined with image analysis to investigate wear evolution on gear teeth – A case study. *Wear* 430-431: 355–368.
15. Randall, R.B., Chin, Z.Y., Smith, W.A., Borghesani, P. (2019). “Measurement and use of transmission error for diagnostics of gears”, *Survishno conference*, INSA de Lyon, France (8–10 July).
16. Mark, W.D. and Reagor, C.P. (2007). Static-transmission-error vibratory-excitation contributions from plastically deformed gear teeth caused by tooth bending-fatigue damage. *Mechanical Systems and Signal Processing* 21 (2): 885–905.
17. Smith, J.D. (2003). *Gear Noise and Vibration*, 2e. NY: Marcel Dekker Inc.
18. Chin, Z.Y., Smith, W.A., Borghesani, P. et al. (2021). Absolute transmission error: a simple new tool for assessing gear wear. *Mechanical Systems and Signal Processing* 146, Ref. 107070.
19. Randall, R.B. (1984). Separating excitation and structural response effects in gearboxes. In: *Third Int. Conf. On Vib. In Rotating Machines*, 101–107. York: I.Mech.E.
20. Randall, R.B. (1997). Advanced machine diagnostics. In: *The Shock and Vibration Digest*, vol. 29, 6–30. Willowbrook, IL, USA: Vibration Institute.
21. Randall, R.B., Antoni, J., and Smith, W.A. (2019). A survey of the application of the cepstrum to structural modal analysis. *Mechanical Systems and Signal Processing* 118: 716–741.
22. El Badaoui, M., Antoni, J., Guillet, F., and Danière, J. (2001). Use of the moving Cepstrum integral to detect and localise tooth spalls in gears. *Mechanical Systems and Signal Processing* 15 (5): 873–885.
23. El Badaoui, M., Cahouet, V., Guillet, F. et al. (2001). Modelling and detection of localized tooth defects in geared systems. *Transactions of ASME, American Society of Mechanical Engineers* 123: 422–430.
24. Endo, H., Randall, R.B., and Gosselin, C. (2004). Differential diagnosis of spalls vs. cracks in the gear tooth fillet region. *Journal of Failure Analysis and Prevention* 4 (5): 57–65.
25. Endo, H., Randall, R.B., and Gosselin, C. (2009). Differential diagnosis of spall vs. cracks in the gear tooth fillet region: experimental validation. *Mechanical Systems and Signal Processing* 23 (3): 636–651.
26. Jia, S. and Howard, I. (2006). Comparison of localised spalling and crack damage from dynamic modelling of spur gear vibrations. *Mechanical Systems and Signal Processing* 20: 332–349.
27. Peng, D., Smith, W.A., Borghesani, P. et al. (2019). Comprehensive planet gear diagnostics: use of transmission error and mesh phasing to distinguish localised fault types and identify faulty gears. *Mechanical Systems and Signal Processing* 127: 531–550.
28. Stander, C.J., Heyns, P.S., and Schoombie, W. (2002). Using vibration monitoring for local fault detection on gears operating under fluctuating load conditions. *Mechanical Systems and Signal Processing* 16 (6): 1005–1024.
29. Zhan, Y., Makis, V., and Jardine, A.K.S. (2006). Adaptive state detection of gearboxes under varying load conditions based on parametric modelling. *Mechanical Systems and Signal Processing* 20 (1): 188–221.
30. Bartelmuß, W. and Zimroz, R. (2009). Vibration condition monitoring of planetary gearbox under varying external load. *Mechanical Systems and Signal Processing* 23 (1): 246–257.
31. Bartelmuß, W. and Zimroz, R. (2009). A new feature for monitoring the condition of gearboxes in non-stationary operating conditions. *Mechanical Systems and Signal Processing* 23 (6): 1528–1534.
32. Barszcz, T. and Randall, R.B. Application of spectral kurtosis for detection of a tooth crack in the planetary gear of a wind turbine. *Mechanical Systems and Signal Processing* 23: 1352–1365.
33. Randall, R.B., Antoni, J., Borghesani, P. (2019). “The potential for obtaining scaled separated forcing functions and scaled transfer functions from operational response vibrations, in particular of rotating machines”, *ICEDyn conference*, Viana do Castelo (June).
34. Darlow, M.S., Badgley, R.H. and Hogg, G.W. (1974). “Application of high frequency resonance techniques for bearing diagnostics in helicopter gearboxes”, US Army Air Mobility Research and Development Laboratory, Technical Report, pp. 74–77.
35. Bonnardot, F., Randall, R.B., and Antoni, J. (2004). Enhanced unsupervised noise cancellation using angular resampling for planetary bearing fault diagnosis. *International Journal of Acoustics and Vibration* 9 (2): 51–60.
36. Randall, R.B. and Smith, W.A. (2019). Uses and mis-uses of energy operators for machine diagnostics. *Mechanical Systems and Signal Processing* 133.
37. Ho, D. and Randall, R.B. (2000). Optimisation of bearing diagnostic techniques using simulated and actual bearing fault signals. *Mechanical Systems and Signal Processing* 14 (5): 763–788.
38. McCormick, A.C. and Nandi, A.C. (1998). Cyclostationarity in rotating machine vibrations. *Mechanical Systems and Signal Processing* 12 (2): 225–242.

39. Randall, R.B., Antoni, J., and Chobsaard, S. (2001). The relationship between spectral correlation and envelope analysis in the diagnostics of bearing faults and other Cyclostationary machine signals. *Mechanical Systems and Signal Processing* 15 (5): 945–962.
40. Antoni, J. and Randall, R.B. (2002). Differential diagnosis of gear and bearing faults. *Transactions of the ASME, Journal of Vibration and Acoustics* 124: 165–171.
41. Antoni, J. and Randall, R.B. (2003). A stochastic model for simulation and diagnostics of rolling element bearings with localised faults. *Transactions of the ASME, Journal of Vibration and Acoustics* 125: 282–289.
42. Randall, R.B. (2016). “Modern envelope analysis for bearing diagnostics”, *Int. J. of Comadem*, 19(3), July. From *Comadem conference*, Buenos Aires, 2015.
43. Sawalhi, N. and Randall, R.B. (2007). “Semi-automated bearing diagnostics – three case studies”, *Comadem Conference*, Faro, Portugal (June 2007).
44. Sawalhi, N. (2007). “Rolling element bearings: diagnostics, prognostics and fault simulations”. PhD Dissertation, University of New South Wales. Available at <http://handle.unsw.edu.au/1959.4/40544> (accessed 02 February 2022).
45. Randall, R.B., Antoni, J., Gryllias, K. (2016). “Alternatives to kurtosis as an indicator of rolling element bearing faults”. *ISMA 2016 conference*, KU Leuven, Belgium (September 2016).
46. Smith, W.A., Borghesani, P., Ni, Q. et al. (2019). Optimal demodulation-band selection for envelope-based diagnostics: a comparative study of traditional and novel tools. *Mechanical Systems and Signal Processing* 134, Ref. 106303.
47. Borghesani, P., Pennacchi, P., and Chatterton, S. (2014). The relationship between kurtosis- and envelope-based indexes for the diagnostic of rolling element bearings. *Mechanical Systems and Signal Processing* 43 (1–2): 25–43.
48. Borghesani, P. and Antoni, J. (2017). CS2 analysis in presence of non-Gaussian background noise – effect on traditional estimators and resilience of log-envelope indicators. *Mechanical Systems and Signal Processing* 90: 378–398.
49. Antoni, J. and Borghesani, P. (2019). A statistical methodology for the design of condition indicators. *Mechanical Systems and Signal Processing* 114: 290–327.
50. Randall, R.B., Smith, W.A., Coats, M.D. (2014). “Bearing diagnostics under widely varying speed conditions”. *CMMNO conference*, Lyon, France (December 2014).
51. Randall, R.B., Smith, W.A. (2020). “Bearing diagnostics in variable speed gearboxes”. ISMA conference, KU Leuven, Belgium (September 2020).
52. Abboud, D., Elbadaoui, M., Smith, W.A., and Randall, R.B. (2019). Advanced bearing diagnostics: a comparative study of two powerful approaches. *Mechanical Systems and Signal Processing* 114: 604–627.
53. Antoni, J., Danière, J., and Guillet, F. (2002). Effective vibration analysis of IC engines using cyclostationarity. Part I: a methodology for condition monitoring. *Journal of Sound and Vibration* 257 (5): 815–837.
54. Antoni, J. (2009). Cyclostationarity by examples. *Mechanical Systems and Signal Processing* 23: 987–1036.
55. Zouari, R., Antoni, J., Ille, J.L. et al. (2007). Cyclostationary modelling of reciprocating compressors and application to valve fault detection. *International Journal of Acoustics and Vibrations* 12 (3): 116–124.
56. Bardou, O. and Sidahmed, M. (1994). Early detection of leakages in the exhaust and discharge systems of reciprocating machines by vibration analysis. *Mechanical Systems and Signal Processing* 8 (5): 551–570.
57. Lyon, R.H. and Ordubadi, A. (1982). Use of cepstra in acoustical signal analysis. *The ASME Journal of Mechanical Design* 104: 303–306.
58. Azzoni, P.M., Minelli, G. and Padovani, E. (1989). “Combustion pressure recovering in spark ignition car engine”. ISATA20, Florence, pp. 667–676.
59. Randall, R.B., Ren, Y., and Ngu, H. (1996). Diesel engine cylinder pressure recovery. In: *Proc. ISMA21*, 847–856. Leuven, Belgium: Katholieke Universiteit.
60. Cassini, G.D., D’Ambrogio, W., and Sestieri, A. (1996). Frequency domain vs Cepstrum technique for machinery diagnostics and input waveform reconstruction. In: *Proc. ISMA21*, 835–846. Leuven Belgium: Katholieke Universiteit.
61. Ren, Y. (1999). Detection of knocking combustion in diesel engines by inverse filtering of structural vibration signals. PhD Dissertation, UNSW, Sydney, Australia.
62. Gao, Y. and Randall, R.B. (1999). Reconstruction of diesel engine cylinder pressure using a time domain smoothing technique. *Mechanical Systems and Signal Processing* 13 (5): 709–722.
63. Zurita, G. (2001). Vibration based diagnostics for analysis of combustion properties and noise emissions of IC engines. PhD Thesis. Luleå University of Technology, Division of Sound and Vibration, Luleå, Sweden.
64. Johnsson, R. (2004). Indirect measurement for control and diagnostics of IC engines. PhD Thesis. Luleå University of Technology, Division of Sound and Vibration, Luleå, Sweden.
65. Zurita, G. and Haupt, D. (2003). A new approach to diagnostics of the combustion process in diesel engines using vibration measurements – part II – detection of the start of combustion using the reconstructed pressure signals. *International Journal of Acoustics and Vibrations* 8 (2): 77–82.
66. Johnsson, R. and Ågren, A. (2004). Cylinder pressure reconstruction from vibration and speed measurements on IC engines. In: *Proc. ISMA 2004*, 965–974. Leuven: Katholieke Universiteit.

67. Johnsson, R. (2006). Cylinder pressure reconstruction based on complex radial basis function networks from vibration and speed signals. *Mechanical Systems and Signal Processing* 20 (8): 1923–1940.
68. Gu, F., Jacob, P.J., and Ball, A. (1999). Non-parametric models in the modelling of engine performance and condition. Part 2, non-intrusive estimation of diesel engine cylinder pressure and its use in fault detection. *Proceedings of the Institution of Mechanical Engineers, Part D* 213 (D1): 135–143.
69. Moskwa, J., Wang, W., and Bucheger, A. (2001). A new methodology for use in engine diagnostics and control, utilizing ‘synthetic’ engine variables: theoretical and experimental results. *Journal of Dynamic Systems, Measurement, and Control* 123: 528–534.
70. Desbazeille, M., Randall, R.B., Guillet, F. et al. (2010). Model-based diagnosis of large diesel engines based on angular speed variations of the crankshaft. *Mechanical Systems and Signal Processing* 24 (5): 1529–1541.
71. Liu, X. and Randall, R.B. (2005). Blind source separation of internal combustion engine piston slap from other measured vibration signals. *Mechanical Systems and Signal Processing* 19 (6): 1196–1208.
72. Liu, X., Randall, R.B., and Antoni, J. (2008). Blind separation of internal combustion engine vibration signals by a deflation method. *Mechanical Systems and Signal Processing* 22: 1082–1091.

8

Fault Simulation

8.1 Background and Justification

The early development of machine diagnostics and condition monitoring was based on measurements from actual failures, but these cannot be predicted or arranged to occur when and where desired. In recent years it has become possible to make simulation models of a machine, such as a gearbox or engine, including the simulation of various faults of different types, severity and location. There are a number of benefits from doing this, the first being to be able to produce sufficient representative signals to train automated fault recognition algorithms such as artificial neural networks, as it is not economically viable to experience the number of actual failures required. The simulation model will normally require updating, to adapt to the properties of an individual machine, typically differing because of dimensional tolerances, even though all might have the same drawings and specifications, and to match the actual response to natural or seeded faults. However, this can be based on a much reduced number of experiments compared with that required to simulate all possible locations and severities of faults in a given machine. Being able to produce signals from faults of different sizes and locations can be useful in the development of diagnostic and prognostic procedures, the latter for example by being able to develop appropriate trend parameters. Finally, the effects of faults in complex machines are often based on nonlinear interactions, which are difficult to foresee, and simulation modelling of the whole machine, including nonlinear and time-varying components, can be very useful to obtain a physical understanding of these complex interactions.

Simulation has long been used for modelling rotors, in particular those operating in a speed range covering a number of critical speeds, and with fluid film bearings, with nonlinear and time varying properties depending on factors such as speed, load, and lubricant viscosity. This is discussed briefly in Sections 2.2.1 and 2.2.5, with references giving more details, including of specific cases such as cracked rotors and electrical machines, with electromagnetic interactions between rotor and stator. This chapter illustrates the advantages of fault simulation, using examples of gears, rolling element bearings, geared systems (including bearings), and internal combustion (IC) engines.

8.2 Simulation of Faults in Gears

Before the faults can be simulated, there must first be a simulation model of the gearbox in healthy condition, into which the faults can be introduced. One of the earliest surveys of gear modelling [1] lists a range of different models, from a single gear pair in torsion with only the gearmesh stiffness as a connection between the two inertias, to complete geared systems including driving and driven machines, and multiple shafts with several gear stages. Many systems included lateral as well as rotational deflections, and some, the rotor dynamics of the shafts involved. Most of the simulation models were for the components in normal condition, but some did model faults in the gears. Most were lumped parameter models (LPMs), but some included finite element (FE) models of some or all components.

8.2.1 Lumped Parameter Models of Parallel Gears

As described in [1], a number of multi-degree-of-freedom (MDOF) models of geared systems, of varying complexity, have been made over the years, though with many designed primarily to assist in the design (e.g. to avoid exciting resonances), and not including modelling of faults.

An idea of the required complexity of an LPM can be gained by considering the way in which the various inertias and stiffness elements interact as the speed of a geared system is increased. Figure 8.1 (from [2]) shows a typical single stage gearbox with driver and driven units, and accessories such as a torque transducer and encoders (lumped into the adjacent inertias, in the following, for simplicity).

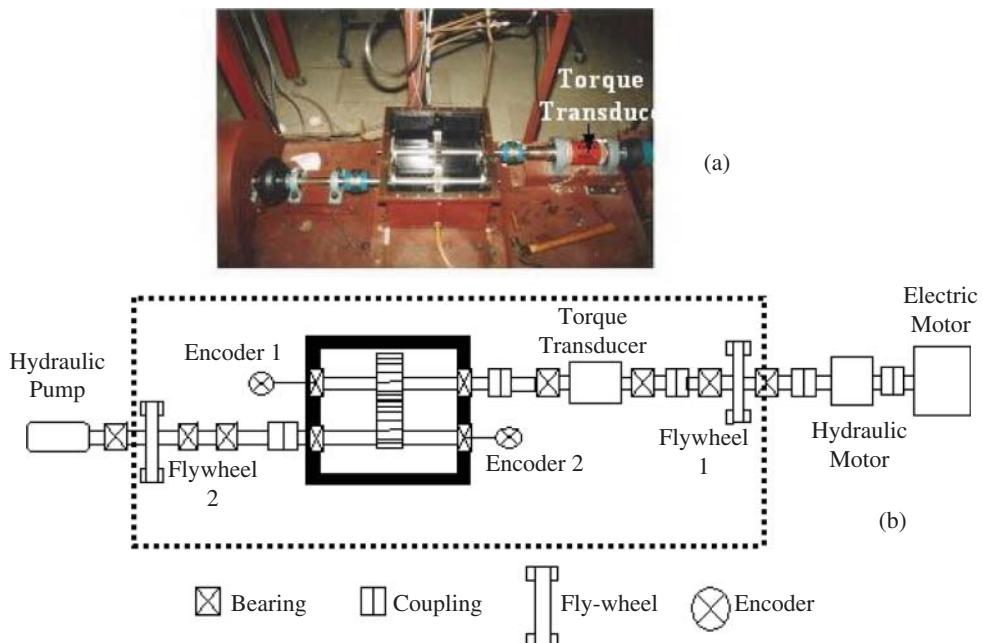


Figure 8.1 (a) Spur gear test rig (b) Schematic diagram of the test rig. Source: From [2].

The ‘driver’ is actually a combination of an electric motor, which controls the speed, and a hydraulic motor which can greatly increase the torque applied to the gearbox without consuming excessive power. The hydraulic pump, which is the driven unit, feeds the hydraulic motor, thus giving a circulating power system. This test rig was used for a number of the demonstrations in this chapter.

Considering torsional modes only, at very low speeds the rig could be considered as a free inertia, which can be rotated to any position with no resistance, and all components on each shaft rotating by the same amount, with the angular displacement of each shaft determined by the gear ratio. It could be modelled as a single rotational inertia at the speed of the input shaft ω_1 , by adding up the kinetic energy (KE) of each element in each shaft string, with its appropriate angular velocity either ω_1 or ω_2 (the speed of the output shaft) and setting the total equal to $I_{eq}\omega_1^2$, where I_{eq} is the equivalent inertia related to the speed of the input shaft. Since the flexible couplings would have the lowest torsional stiffness, at some speed these could all be considered as torsional springs, and all the elements between and outside them could be modelled as single inertias. In the case of Figure 8.1b, there are five couplings with six inertias outside or between them, giving a 6 DOF (degree-of-freedom) system with free-free boundary conditions. At some higher speed, the sections of shaft would begin to act as torsional springs connecting the inertial elements between them, and thus increasing the number of DOFs accordingly. It should be kept in mind that the moment of inertia of a cylindrical object increases in proportion to the fourth power of the diameter (a factor of 16 for just 2 : 1 change in diameter), and directly with the length, while the torsional stiffness also changes with the fourth power of the diameter, but inversely with the length, making it relatively easy to decide whether a cylindrical shape (e.g. shaft section or gear) is primarily an inertia or torsional spring.

For most analysis of gearboxes this is the level at which the components are modelled. Gears, (half) couplings, motors, brakes, and flywheels, etc. can be considered as rigid lumped inertias, while shaft sections can usually be treated as lumped torsional springs, as long as there are no torsional modes in the shaft sections themselves. In modelling the effect of a tooth root crack, for example, this can be treated as a simple change of the mesh stiffness, affecting only those modes involving the inertias of the gears in conjunction with the stiffness of the mesh. The natural frequency of the tooth itself, on its reduced stiffness base, would be far above the frequency range being considered.

The same applies when translational DOFs are also included, and it is still reasonable to model gears etc. as lumped masses, and shaft sections as linear springs. It should be kept in mind, however, that these shaft springs only model the shaft sections as bending in their first bending mode, and in high speed machines there is a possibility that higher order modes might be in the range of excitation of garmesh harmonics, in which case LPMs should be replaced by FE models. When the lumped inertias can tilt, as a result of shaft deflections, the gyroscopic effects can be included in the LPM.

Figure 8.2 is an LPM of the same gearbox as shown in Figure 8.1, restricted to the dashed rectangle of Figure 8.1b, between the two flywheels, assuming that the latter isolate the driver and driven machines.

Note that since it is designed for vibrations caused by the gears only, all transverse forces and vibrations are assumed to be along the line of action of the mesh force, nominated as defining the x-axis. The four bearings are modelled as linear springs (k_2, k_4, k_6, k_7) connected to earth (the bearing supports in the casing). The only lateral DOFs are at and between the bearings, it being assumed that masses outside the casing are isolated by the flexible couplings. There are thus six masses associated with the six transverse DOFs (the x_i) and eight rotational inertias associated with the eight rotational DOFs (the θ_i), making a total of 14 DOFs (the original model, with encoders, had 16).

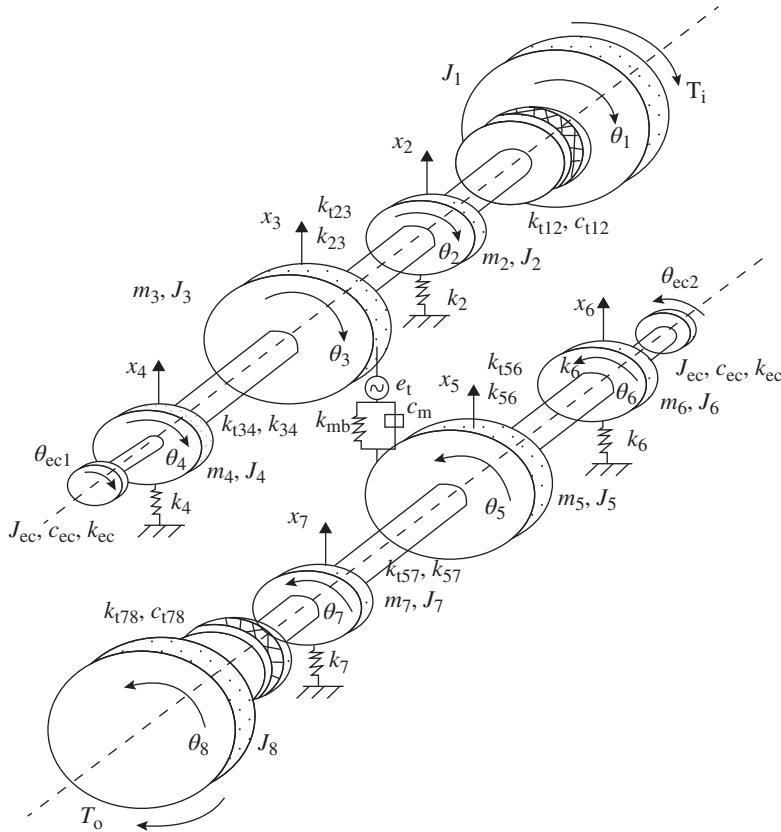


Figure 8.2 LPM of parallel shaft gearbox in Figure 8.1 for gear vibrations only.

The equation of motion (EOM) for an MDOF system such as this (with viscous dampers) is:

$$[M]\{\ddot{x}(t)\} + [C]\{\dot{x}(t)\} + [K]\{x(t)\} = \{f(t)\} \quad (8.1)$$

where in the case of the LPM in Figure 8.2 the ‘displacement’ vector $\{x(t)\}$ is given by $[x_2 \ x_3 \ x_4 \ x_5 \ x_6 \ x_7 \ \theta_1 \ \theta_2 \ \theta_3 \ \theta_4 \ \theta_5 \ \theta_6 \ \theta_7 \ \theta_8]^T$ with the (externally applied) ‘force’ vector $\{f(t)\}$ being the forces (or torques for rotational DOFs) in the same DOFs. The ‘mass’ matrix $[M]$ is a diagonal matrix of the inertias (linear or rotational) at the same DOFs; thus:

$$[M] = \text{diag} [m_2 \ m_3 \ m_4 \ m_5 \ m_6 \ m_7 \ J_1 \ J_2 \ J_3 \ J_4 \ J_5 \ J_6 \ J_7 \ J_8] \quad (8.2)$$

The elements k_{ij} of the stiffness matrix $[K]$, for ‘force’ applied in DOF i and ‘displacement’ measured at DOF j can be determined (for a particular column j) by fixing all other DOFs, and registering the force (or torque) at each DOF i after applying unit displacement (or rotation) at DOF j . The stiffness element will be positive if the force is in the opposite direction to the displacement, but negative if it is in the same direction.

Thus, the stiffness matrix for the system of Figure 8.2 is:

$$[K] = \begin{bmatrix} k_2 + k_{23} & -k_{23} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -k_{23} & k_{mb} + k_{23} + k_{34} & -k_{34} & -k_{mb} & 0 & 0 & 0 & 0 & -k_{mr}r_3 & 0 & -k_{mb}r_5 & 0 & 0 & 0 & 0 \\ 0 & -k_{34} & k_4 + k_{34} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -k_{mb} & 0 & k_{mb} + k_{56} + k_{57} & -k_{56} & -k_{57} & 0 & 0 & 0 & 0 & -k_{mb}r_5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -k_{56} & k_6 + k_{56} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -k_{57} & 0 & k_7 + k_{57} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & k_{12} & -k_{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -k_{12} & k_{112} + k_{23} & -k_{23} & 0 & 0 & 0 & 0 & 0 \\ 0 & -k_{mb}r_3 & 0 & 0 & 0 & 0 & 0 & -k_{23} & k_{23} + k_{34} & -k_{34} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -k_{34} & k_{34} & 0 & 0 & 0 & 0 & 0 \\ 0 & -k_{mb}r_5 & 0 & -k_{mb}r_5 & 0 & 0 & 0 & 0 & 0 & -k_{56} & k_{56} + k_{57} & -k_{56} & -k_{57} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -k_{56} & k_{56} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -k_{57} & 0 & k_{57} + k_{78} & -k_{78} & k_{78} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -k_{78} & 0 & 0 & 0 \end{bmatrix} \quad (8.3)$$

where r_3 and r_5 are the base circle radii of the respective gears. As will be seen, the stiffness matrix is symmetric.

In principle, the damping matrix $[C]$ would have exactly the same form as (8.3), but with stiffness terms replaced by damping terms (since damping forces are related to the relative motion between the ends of the elements, as for springs). However, since the only damping modelled in the system of Figure 8.2 is the mesh damping c_m , all elements would be zero except those in Eq. (8.3) involving k_{mb} which would need to be replaced with c_m , and all other terms set to zero.

Force vector $\{f(t)\}$ can include varying driving (T_i) or driven (T_o) torques, but often geared systems are ‘parametrically’ excited, i.e. the excitation comes from time (or rotational angle) varying parameters, e.g. the stiffness of the gearmesh. The latter is also nonlinear to some extent. It is then common to solve the EOM by dividing the stiffness matrix into the sum of a constant part (the mean values of the stiffness) plus a variable and nonlinear part representing variations around this. The forces from the varying and nonlinear part can then be taken over to the right hand side of the equations, and treated as variable forcing functions applied to linear system equations. These can be solved in the time domain, for example by Runge Kutta methods.

The linear equations for constant parameters can conveniently be solved in the frequency domain. Applying the Laplace transform to Eq. (8.1) gives:

$$([M]s^2 + [C]s + [K])\{X(s)\} = [B(s)]\{X(s)\} = \{F(s)\} \quad (8.4)$$

where the upper case symbols represent the Laplace transforms of the time variables. Eq. (8.4) can be inverted to give the response to a specified force distribution:

$$\{X(s)\} = [B(s)]^{-1}\{F(s)\} = [H(s)]\{F(s)\} \quad (8.5)$$

where $[H(s)]$ is the transfer function matrix. $[H(s)]$ is thus the inverse of $([M]s^2 + [C]s + [K])$, which for each element is a ratio of two polynomials in s , with common denominator of order $2n$, where n is the number of DOFs, and with numerators of lower order, as in Eq. (8.6):

$$H_{ij}(s) = \frac{a_0 + a_1s + a_2s^2 + \dots + a_ms^m}{b_0 + b_1s + b_2s^2 + \dots + b_{2n}s^{2n}} \quad (8.6)$$

The roots of the denominator polynomial are the values of s for which $[H]$ goes to infinity, i.e. the poles, corresponding to resonances, common to all transfer functions, while the roots of the numerator give the zeros, corresponding to antiresonances, different for each transfer function. Eq. (8.6) can be expanded by partial fractions to give the form:

$$H_{ij}(s) = \sum_{k=1}^n \left[\frac{r_{ijk}}{s - p_k} + \frac{r_{ijk}^*}{s - p_k^*} \right] \quad (8.7)$$

the equivalent of a sum of single degree-of-freedom (SDOF) systems. The poles (p_k) are of course still the same, but the r_{ijk} are called the residues (pole strength) and this way of expressing $[H]$ is known as a pole/residue model.

In general, a pole in the Laplace plane is given by the equation:

$$p_k = \sigma_k + j\omega_k \quad (8.8)$$

where the impulse response function (IRF) for the complex conjugate pole pair p_k and p_k^* is given by $2r_{ijk} \exp(-\sigma_k t) \sin(\omega_k t)$.

When damping is light, or ‘proportional’, i.e. $[C] = \alpha[M] + \beta[K]$, it can be shown [3], that the eigenvector matrix and diagonal eigenvalue matrix, obtained by eigenvalue decomposition, diagonalise the mass and stiffness matrices, such that:

$$[\psi]^T [M] [\psi] = [\lambda_m] \quad (8.9)$$

and

$$[\psi]^T [K] [\psi] = [\lambda_k] \quad (8.10)$$

Note that because the eigenvectors $\{\psi_r\}$ are only determined to within a scaling constant, $[\lambda_m]$ and $[\lambda_k]$, (the ‘modal mass’ and ‘modal stiffness’, respectively) are not unique. However, the ratio $\frac{\lambda_k}{\lambda_m}$ is unique and equal to ω_r^2 . Thus, by setting $\{\phi_r\} = \frac{1}{\sqrt{\lambda_m}} \{\psi_r\}$, the Eqs. (8.9) and (8.10) become:

$$[\phi]^T [M] [\phi] = [\lambda_I] \quad (8.11)$$

and

$$[\phi]^T [K] [\phi] = [\omega_r^2] \quad (8.12)$$

Moreover, for proportional damping,

$$[\phi]^T [C] [\phi] = [\lambda_{2\sigma}] \quad (8.13)$$

which is called ‘scaling to unit modal mass’.

Since from (8.5), for zero damping, $[B(\omega)] = ([K] - \omega^2[M]) = [H(\omega)]^{-1}$, using (8.11) and (8.12)

$$[(\omega_r^2 - \omega^2)] = [\phi]^T [H(\omega)]^{-1} [\phi] \quad (8.14)$$

and by inversion

$$[H(\omega)] = [\phi]^T [(\omega_r^2 - \omega^2)]^{-1} [\phi] \quad (8.15)$$

from which a typical frequency response function (FRF) $H_{ij}(\omega)$ would be:

$$H_{ijr}(\omega) = \sum_{r=1}^n \frac{\phi_{ir} \phi_{jr}}{\omega_r^2 - \omega^2} \quad (8.16)$$

showing that the residues of Eq. (8.7) are the product of the eigenvector (mode shape) components for mode r at the points of excitation and response.

One of the main advantages of the orthogonality properties is that they allow decoupling of the equations of motion when the latter are expressed in terms of modal coordinates $\{q\}$ (i.e. how much of each mode is present in the motion). From the ‘modal transformation’ equation:

$$\{x\} = [\phi]\{q\} \quad (8.17)$$

by substituting (8.8), (8.11), (8.12), and (8.13) into (8.1) it will be found that:

$$[\lambda_I]\{\ddot{q}\} + [\lambda_{2\sigma}]\{\dot{q}\} + [\lambda_{\omega_r^2}]\{q\} = \{\Gamma\} \quad (8.18)$$

where the ‘generalised force’ is given by $[\Gamma] = [\phi]^T \{f(t)\}$, i.e. the correlation (dot product) of the force distribution with each mode shape.

Thus, by using modal coordinates, the response of an MDOF system can be obtained as a sum of individual SDOF responses, one for each mode.

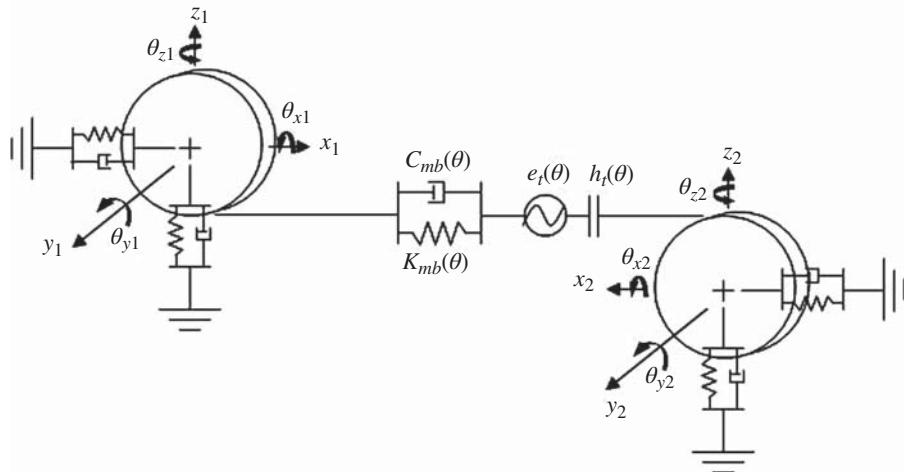


Figure 8.3 Detail of LPM showing the meshing gears. $K_{mb}(\theta)$: Position dependent variable stiffness; C : Damping; $e_t(\theta)$: Combined effect of gear geometrical errors and misalignment; $h_t(\theta)$: A switch representing the contact status of the meshing gear.

8.2.2 Separation of Spalls and Cracks

As discussed in Section 7.2.4 of Chapter 7, the separation of spalls and cracks in gears is very important diagnostically, because of a huge difference in prognosis; cracks can fail rapidly, while spalls may be present for much longer periods (although the increased cyclic stresses may cause crack initiation). The results from Endo's study, extracts of which are reproduced in Figures 7.36–7.38, were obtained using an LPM based on that in Figure 8.2. An extract concentrating on the two gears is shown in Figure 8.3, highlighting the elements that were modified to simulate spall and crack faults. These include the mesh stiffness $K_{mb}(\theta)$, and geometric transmission error (GTE) $e_t(\theta)$. It also includes a switch element $h_t(\theta)$, which is open only when there is no compressive force between the teeth, but otherwise closed. In Endo's work, the switch was permanently closed as the load was always sufficient to maintain positive pressure between the meshing teeth.

Figure 8.4 (from [4], which is Ref. [23] of Chapter 7) defines the geometry of the 2D model used for initial investigation of the effects of tooth root cracks and spalls, the latter assumed across the entire face width of the tooth.

It also shows the FE meshing of this 2D model, the faults being incorporated into the FE models of two whole gears, so as to include the deformation of the gear bodies in the determination of static mesh stiffness against roll angle, of a meshing pair, each with 32 teeth, for different torque loads. Results were obtained for three different crack depths and three different spall widths on a single tooth of one gear.

The results of the 2D analysis, for a single fault size of each type, are shown in Figure 8.5 (from [4]), which indicates the RME (residual motion error), where Endo uses 'motion error' (ME), instead of 'transmission error' (TE), this being the difference in TE from the case with healthy gears. This shows that the residual static TE (STE) for a crack is dependent on load, and due primarily to the reduced stiffness, whereas the STE for the spall is independent of load, and therefore primarily a GTE. Note that the 'crack' in this case was actually modelled as a finite thickness slot, and the later experimental validation of the simulations was also based on a machined slot, rather than a natural

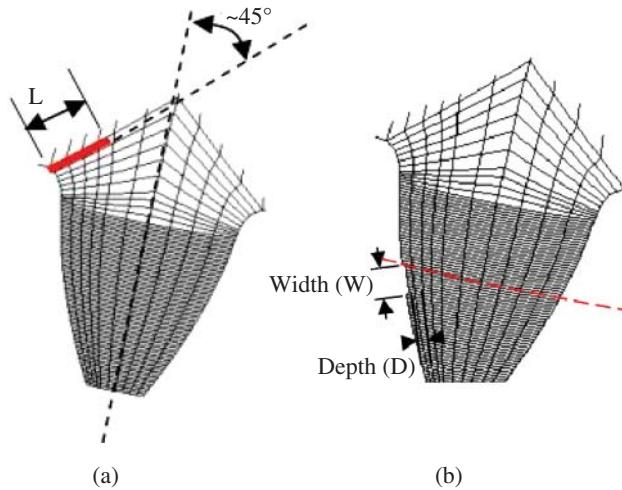


Figure 8.4 Definitions of fault geometry (a) crack (b) spall.

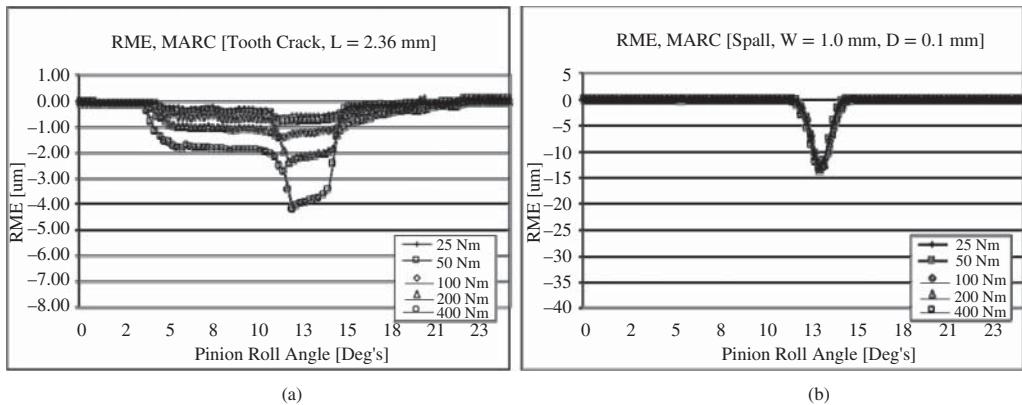


Figure 8.5 The RMEs for gears with (a) a tooth crack and (b) a spall. Source: From [4].

crack. This was done before it was appreciated (from Ref. [25] of Chapter 7) that natural cracks tend to give a GTE (from plastic deformation at the crack tip) in addition to the reduced stiffness.

It was then realised that the 2D model could not represent spalls that did not extend over the full facewidth of the tooth, in particular where the teeth were crowned (as were the experimental gears), so a 3D FE model was then used to do further modelling. Some results from this are shown in Figures 7.36 and 7.37 of Chapter 7, primarily to show how the cracks and spalls differ with respect to the MCI (moving cepstrum integral), this providing one potential way of separating cracks and spalls.

It is interesting to compare the results from the simulation model with experimental measurements, and this was done in Ref. [5] (Ref. [24] of Chapter 7). It was found that because of the special design of the test rig (the shafts were more than an order of magnitude more flexible than the gear-mesh, unlike most gears in practice), it was not possible to compare the results for the two types of fault using the same gears. For the crack, it was necessary to use plastic gears (with Young's modulus two

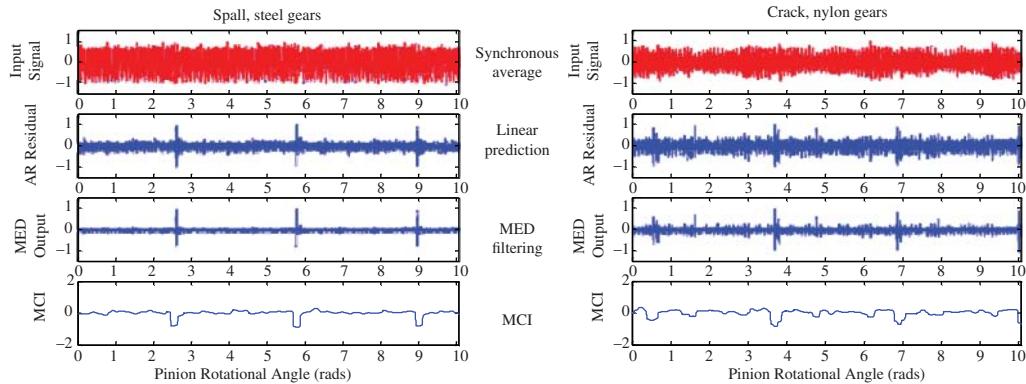


Figure 8.6 Measured properties for spall and crack.

orders of magnitude less than steel) to demonstrate the reduced stiffness of the mesh, while for the spall, the higher modulus of steel was required, to show that the TE was independent of load, and not overwhelmed by Hertzian deformation.

Figure 8.6 presents the measured results for the two types of fault and shows that the MCI for both is much as predicted in the simulated results of Figures 7.37 and 7.38 of Chapter 7. In other words, the MCI in itself does not allow a differentiation of spalls and cracks.

On the other hand, Figure 8.7 validates the other finding of the simulations; viz., that the effect of the spall is independent of the load, even though the garmesh component increases with it, while with the crack, both the garmesh component and the effect of the fault increase with load.

Further details of Endo's work can be obtained from his PhD thesis [2].

8.2.2.1 Modelling of the Stiffness of a Tooth Crack

A number of authors have suggested analytical approaches to estimating the stiffness of a pair of teeth in mesh, including the situation where one has a tooth root crack. One of the most widely cited is Chaari et al. [6], where the stiffness of a healthy tooth is found by combining the compliance components of the two most important contributions. One is the reciprocal of the bending stiffness k_b , where the tooth is modelled as a short beam of varying cross section, with a shear factor because it is stubby, and the other is $1/k_f$, due to the fillet-foundation gear body deflection, which would only be significant where the gear rim is relatively thin. These two compliances are summed to get the compliance of the meshing pair, as a function of roll angle as the mesh point varies from initial contact between the tip of one tooth and root of the other, to loss of contact at the root of the first tooth. The compliance due to contact stress is then added, as the combined value for both teeth, partly because it is due to the difference in curvature between them, which remains relatively constant through the mesh cycle, at least for unworn teeth, even as the individual values vary. Even though the contact compliance is nonlinear, it is usually modelled as a constant value, as given in [6]. The total stiffness at the mesh is the sum of the stiffnesses of the individual tooth pairs in mesh at each roll angle, this varying (for spur gears) by approximately 2 : 1 as the number of pairs in mesh changes from one to two.

Modelling the effect of a crack is more complex, and is often done by FE analysis (as by Endo in [2, 4, 5]). However, in [7] a procedure is given to estimate the reduction of stiffness of a single tooth with a crack, by suggesting a rule for estimation of the effective thickness of the beam in the presence of a crack. The basics are illustrated in Figure 8.8. An FE analysis shows the stress

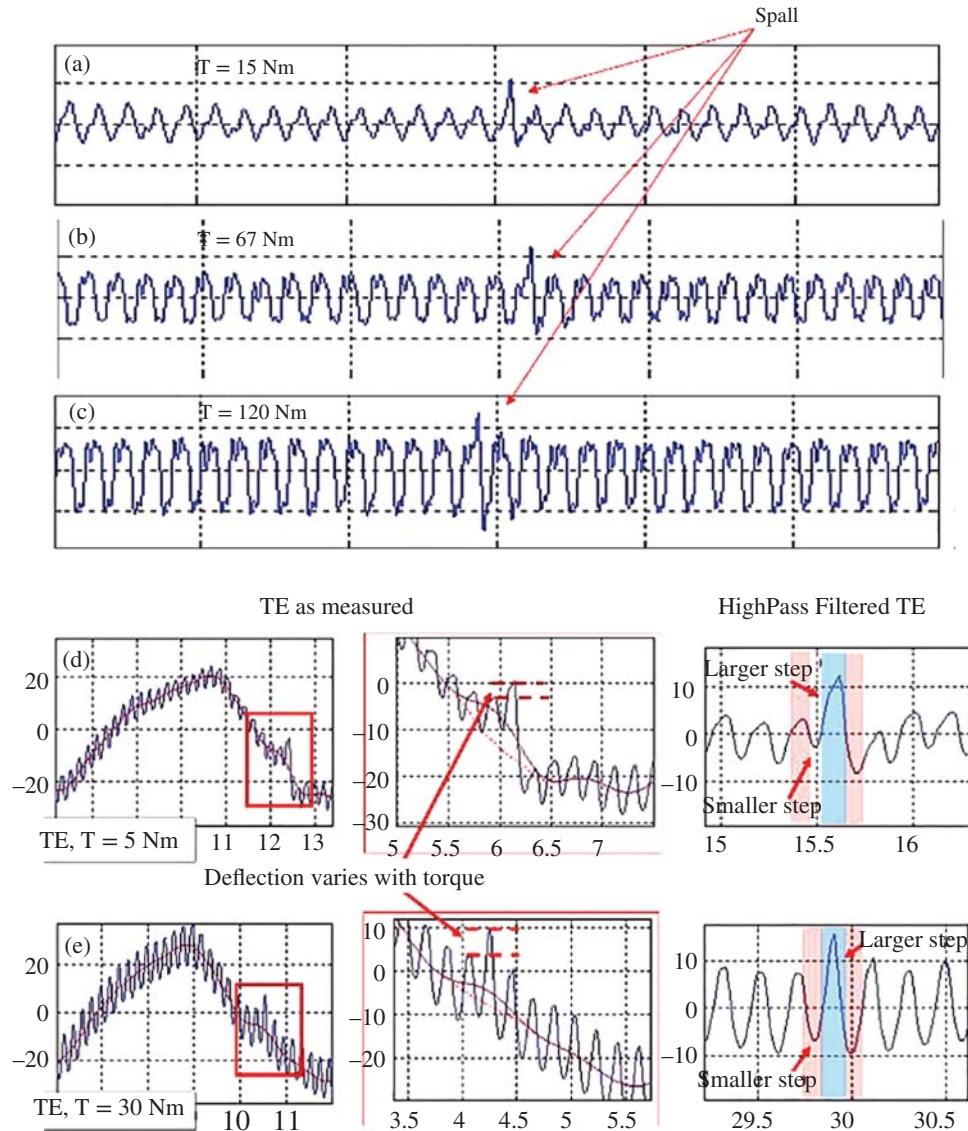


Figure 8.7 Effect of load with spall and crack (a, b, c) Spall (d, e) Crack.

distribution for three different depths of crack, the deepest extending to 50% of the tooth thickness, i.e. to the line of symmetry of the original tooth shape. The stress distributions are for the case with largest bending moment, with force applied at the tip on the crack side and normal to the flank. The line defining the equivalent tooth thickness is shown for the three crack depths. It is a parabola, passing in all cases through the tip corner, where the force is applied in the diagram, and tangential to a radial line through the crack tip. In the intermediate case, the ‘parabola’ is a straight line, but curves in opposite directions for shorter and longer cracks. The justification for this choice is that the neutral axis for the bending, where the bending stress is zero, is seen to be roughly half-way between this line and the profile on the intact side, thus indicating that the moments of the stress

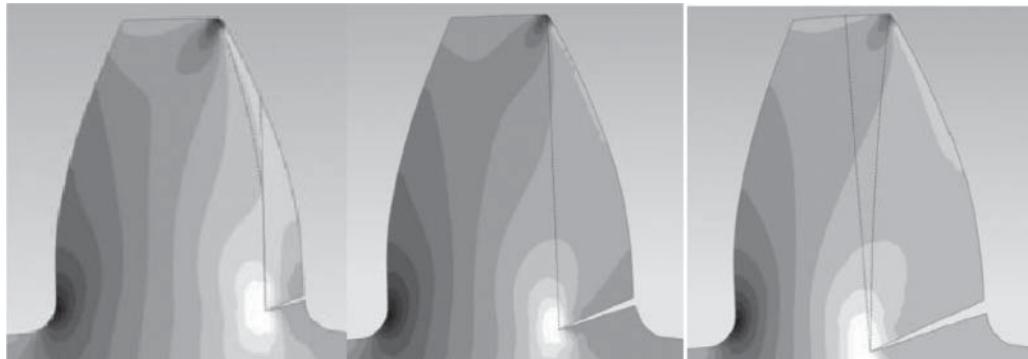


Figure 8.8 Definition of parabolic line defining the effective thickness of the tooth for different crack depths. Source: From [7].

distributions on either side would be balanced. This should give a curvature, and thus deflection, roughly the same as if the tooth were symmetrical about this neutral axis, this being achieved by the suggested line.

Ref. [7] shows that this approach gives the lowest errors of three alternative methods compared with FE analysis, the maximum error being 1.2%. It also gave very comparable simulations of vibration signals, for different crack depths, compared with FE analysis.

8.2.3 Lumped Parameter Models of Planetary Gears

Examples of complex LPMs are given by the work of Kahraman [8] and Parker [9], who analysed planetary gear systems to determine the natural frequencies and mode shapes, for example to demonstrate that mesh phasing could be used to avoid excitation of certain modes. Figure 8.9 shows such a model (from [8]).

These authors also showed that some modes could cause the load distribution between the planets to become so unequal as to completely unload some of them, giving rise to mesh stiffness nonlinearities through loss of tooth contact, and even chaotic behaviour. It had often previously been believed that it would always be possible to maintain tooth contact by applying sufficient torque load on a gear system.

However, when faults are introduced into the various gears, i.e. ring, planet and sun, the resulting vibrations measured at fixed points on the casing become quite complicated, because of the following factors:

- 1) There are multiple paths by which a force at a tooth mesh can reach a response accelerometer. For example, a contact force at a particular tooth on the sun gear, can be transmitted via any or all of the planets to the planet carrier and thence to the casing via the respective meshing teeth on the ring gear. It can also be transmitted to the casing via the bearings supporting the sun gear shaft. In the first case the path length is clearly varying at the speed of the planet carrier, giving corresponding modulations, while the second path varies only in the direction of the forces applied at the bearings, and thus modulating at sun gear speed.
- 2) Planet gears are idlers, meaning that the meshing force with a ring gear tooth is (almost) balanced by an equal and opposite mesh force on the simultaneously meshing sun gear teeth, and vice versa. Note that because the contacts are on opposite flanks of the teeth, the line between these

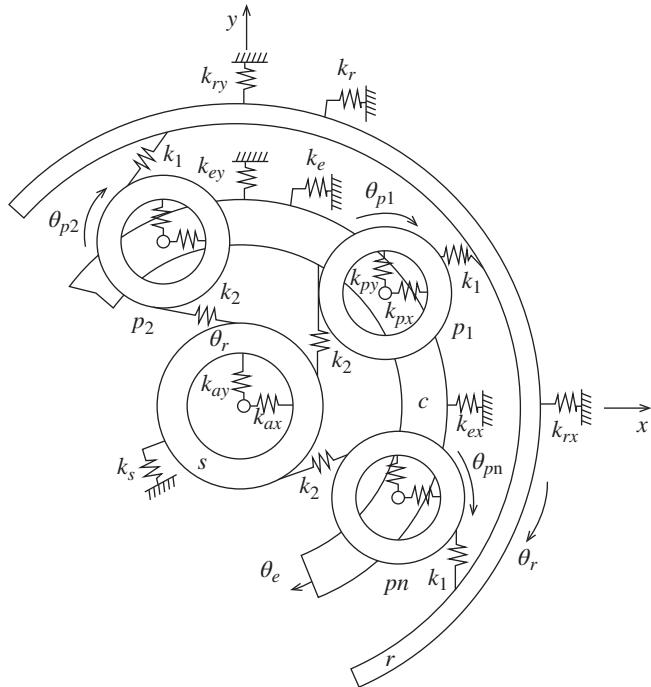


Figure 8.9 Lumped parameter model of a planetary gear (five planets) with both torsional and lateral DOFs.
Source: From [8].

contact points is not diametral, and would be differently divided between one and two tooth pairs depending on whether the number of teeth on the planets is even or odd. The word ‘almost’ is because any difference between these forces is due to the angular acceleration of the planet gear, which would be small at low speeds.

- 3) A fault on one flank of a planet gear tooth, such as a spall, only contacts either the ring gear or the sun, but not both (even though a simultaneous force would be applied to the other as just explained). Faults which have an effect in both directions, such as a change in tooth stiffness (root crack or broken tooth) give a force anomaly in contact with both ring and sun gears. These are not the same (e.g. crack opening or closing) and not evenly spaced, as just explained. The response is also greatly affected by the different contact ratio with the ring and sun gears.
- 4) All the above factors are affected by mesh phasing, which for uniformly spaced planets is either simultaneous or sequential. The first requires tooth numbers on all three gear types to be divisible by the number of planets, whereas the latter requires the sum of the tooth numbers on ring and sun gears to be divisible by the number of planets. A consequence of sequential meshing is that the mesh signals from each planet are offset in phase such that they sum to zero if otherwise uniform, and the largest component in the vicinity of the actual garmesh frequency will be the nearest sideband (spaced at multiples of the carrier frequency) that is a harmonic of the planet pass frequency (carrier speed times number of planets).

Many of these factors, in particular modulation frequencies for the different cases, are explained in Ref. [10].

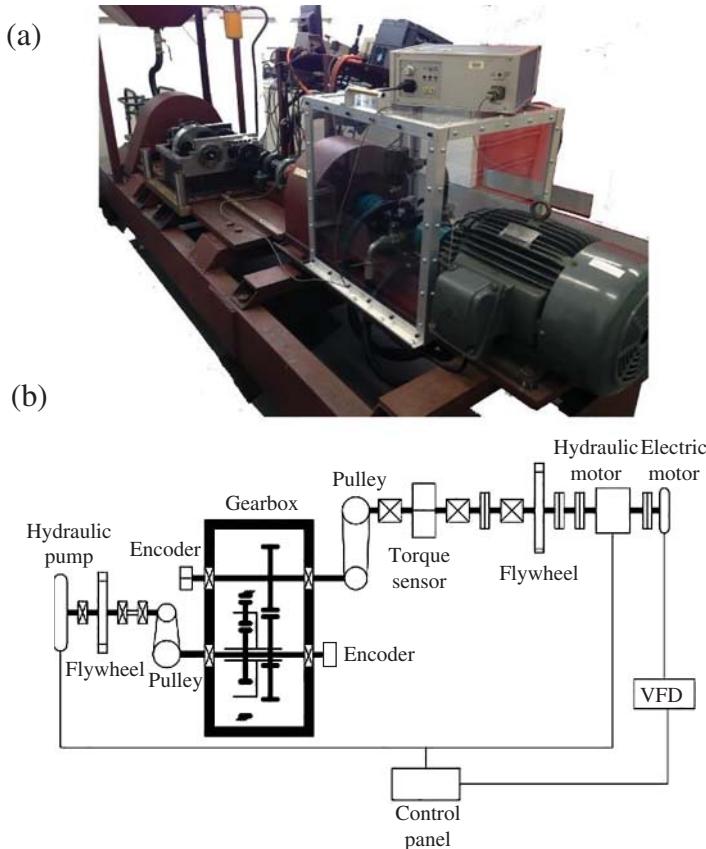


Figure 8.10 The planetary gear test rig at UNSW. (a) photo; (b) schematic diagram ([12]).

8.2.4 Interaction of Faults with Ring and Sun Gears

As mentioned in Section 7.2.2.2, a study was made in Ref. [11] ([13] of Chapter 7), of the benefits of using TE as a diagnostic parameter, for a single stage planetary gear set, driven through a parallel gear input stage. The layout of the experimental rig is shown in Figure 8.10, and the tests are more fully described in Ref. [12].

The planet carrier forms the input to the planetary gear set, which consists of three equi-spaced 23T planets, a fixed 80T ring gear and a 34T sun gear (output). All planetary stage gears are module 2. The parallel gear reduction stage at the input has ratio 42T : 55T. The AC induction motor has a variable frequency drive (VFD), but the circulating power hydraulic motor/pump system is able to apply considerably higher torque to the gearbox, than that delivered by the electric motor itself. The two timing belt drives, introduced for alignment purposes, are in fact 1 : 1 ratio, and do not change the overall transmission ratio from motor to brake.

The two encoders mounted on the free ends of the input and output shafts of the gearbox are both Heidenhain type ROD 426, but that on the input shaft gives 900 pulses per rev (ppr) and that on the output shaft gives 225 ppr, and both also give once-per-rev tacho signals. Accelerometer measurements are taken from the top of the ring gear.

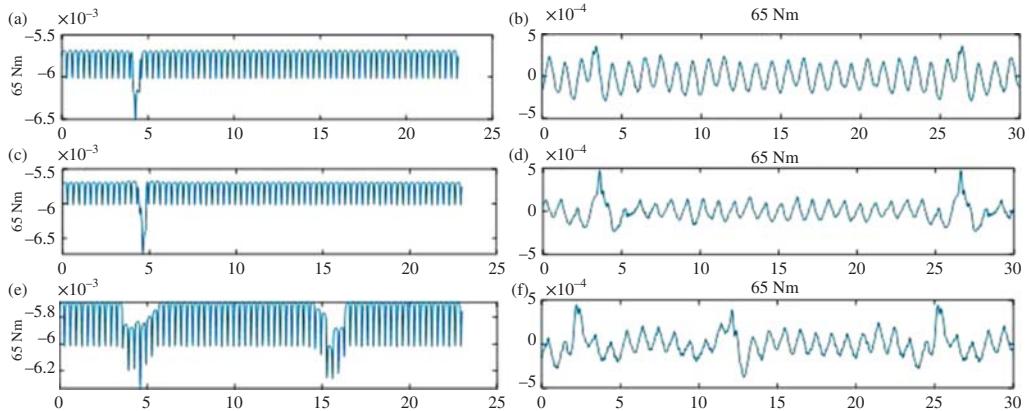


Figure 8.11 Simulated vs measured TE for planetary gear faults (a, c, e) Simulations (b, d, f) Measurements (a, b) Spall – ring gear (c, d) Spall – sun gear (e, f) Crack. Source: From [13].

Figure 7.25, of Chapter 7, gives the results of measured TE for the three cases of a spall interacting with the ring or sun gear, and for a cracked tooth interacting with both. Only the results for one load are presented there but more extensive results are presented in [11, 12]. Ref. [12] also gives details of the simulation models of the planetary gearset and the faults introduced into the gears. The first was a 20 DOF LPM used to obtain simulated vibration signals using Simulink. It included two translational and one rotational DOF for each of the planetary gear components (sun, ring, planet carrier, and three planets) and two rotational DOFs for the input (after the parallel stage) and output (load). The FE models used to simulate the cracked and spalled teeth and the resulting GTE and STE for different loads are described in some detail.

As mentioned in Section 7.2.2.2, comparisons were made between simulations and measurements for the planetary gear rig, but only for relatively low speed, where dynamic effects are not pronounced. This is a limitation of the drive motor and its VFD. The simulations presented here are compared in Figure 8.11 with the measurements presented in Figure 7.25. These results are for input shaft speed close to 2 Hz (planet carrier speed 1.53 Hz, sun gear output speed 5.31 Hz). The input stage to the planetary section was not included in the simulation model, so simulated TE results were estimated and are given for the planetary stage directly, i.e. the TE from the planet carrier shaft to the sun shaft. On the other hand, there was no encoder mounted on the planet carrier shaft, so the measured TE was between the input shaft and the output shaft, and included the input parallel gear stage. However, the TE of the planetary stage could be extracted by using synchronous averaging synchronised with the rotation speed of the planet gears, i.e. a period corresponding to the meshing of 23 teeth. It is necessarily a composite overall TE for the three planet gears, and cannot indicate on which planet gear the fault is located. If there were a fault on more than one gear, they would all show up in this composite average.

Firstly, before making a detailed comparison of the results, there is an immediate obvious difference in the background ripple frequency related to the garmesh frequency. Because of the tooth numbers in the gearset, it has sequential tooth-meshing, with the meshing of the teeth on one planet offset by 1/3 of a tooth spacing compared with each of the others. For there to be synchronous tooth meshing, the numbers of teeth on all three components (sun, planet, ring) would have to all be divisible by 3. Since the simulation is for initially perfect teeth on all gears, apart from that on which

each fault was introduced, the background ripple for the simulations has a fundamental frequency corresponding to the third harmonic of the garmesh frequency (i.e. 69 periods per gear rotation).

The fact that the measured results had clear background ripple at garmesh frequency (23 periods per rotation) could be ascribed to the limited resolution given by the encoder on one shaft, the output shaft, with only 225 ppr, (modified for an earlier experiment, with the pulse count divided by 4). However, it does not significantly detract from these results, showing the effects of local faults on one gear. Even though the TE signal was resampled during the analysis to a higher sample rate, the TE signal was effectively lowpass filtered prior to this at less than three times the garmesh frequency, because of the small number of pulses per tooth spacing. Since the second harmonic of garmesh frequency was virtually not present in any case, the background ripple is dominated by the fundamental garmesh frequency.

It should be noted that the scaling of the TEs has not been unified for the two cases of simulation vs measurement, although this could be done by updating the lumped parameter simulation model to correct natural frequencies, etc. This should in fact be much simpler for the torsional natural frequencies than the much greater number of lateral natural frequencies, which all require updating for a full model. The other obvious difference is that the TE has been measured in the inverse direction for the two cases.

The simulation records are for exactly one planet rotation (23 teeth) so show only one pulse per record for the spalls, but two for the crack, whereas the records for the measurements are displayed over 30 teeth, with a once-per-rotation spacing of 23 teeth.

Comparing now the upper two curves in Figure 8.11, for the spall interacting with the ring gear, there is seen to be a similar pattern for simulation and measurement. In [11] it was shown that there was not much effect of load either. This is because the spall primarily gives a change in GTE, but little change in tooth stiffness, which would be load dependent. In this case, the spall was narrow, and didn't enter into the region with double tooth pair meshing, but the latter would happen with a somewhat wider spall, even though it might stay in the single tooth pair meshing region for the sun-planet interaction.

For the next two curves, with the spall interacting with the sun gear, the results are very similar, though the TE deviations are more triangular, as in [11]. If the spall were wider, and stayed within the single tooth pair meshing region, as mentioned above, this might give more differentiation between the two interactions [11].

The bottom two curves, for a simulated tooth root crack, show obvious differences compared with the spalls. In both simulations and measurements there are now two responses per revolution, one for contact with the sun gear and one with the ring gear, though these are not the same. Neither are they uniformly spaced, both because of the odd number of teeth and the fact that the effective contact points with sun and ring are on the same side of a diametral line through the planet gear. The length of interaction is now greater than for a spall, as it corresponds to the entire time the faulty tooth is engaged, i.e. longer than the tooth spacing, in accordance with [11], although the part with larger deviation extends further for contact with the sun than with the ring [11].

It should be mentioned that the regularly repeating part of the signal can be removed from simulations, to give a so-called ‘residual TE’ or RTE, as shown in [11], and something similar can be achieved in measurements, by using linear prediction, as shown in Figure 8.6.

8.3 Simulation of Faults in Bearings

One of the earliest models of a local fault in a bearing was by Sawalhi [13], which was based on earlier bearing models by Fukata et al. and Feng et al. The former modelled just the bearing in its pedestal, giving estimates of the number of rolling elements supporting the load, for different

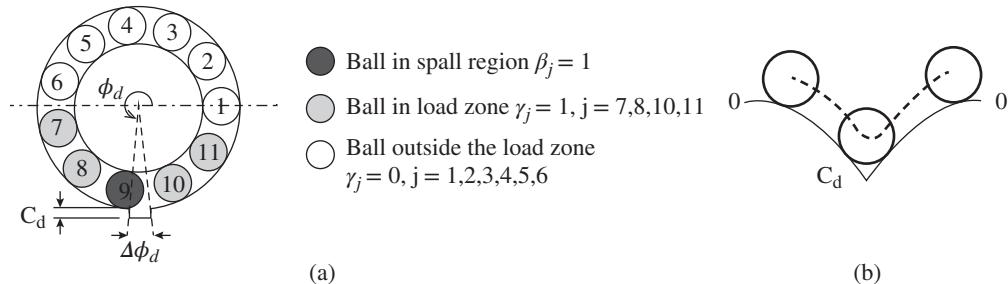


Figure 8.12 Detail of a rectangular simulated spall in the load zone of the outer race. (a) spall geometry and loaded balls (b) trajectory of ball passing through spall.

clearances and loads. Further details are given in Sawalhi's PhD thesis [14] (Ref. [44] of Chapter 7). Random variation in the spacing of the bearing fault responses was incorporated as a uniformly distributed deviation of each rolling element around its kinematic position, in the range $\pm 0.05^\circ$.

The supposed kinematics of the trajectory of a ball as it passed through the zone of a short spall (modelled as a rectangular slot) is shown in Figure 8.12b. Note that this assumes rigid body motion, and neglects the important Hertzian deformation of the sharp edges of the spall. Figure 8.12b (not to scale) shows that the centre of the ball will follow a circular path, with radius equal to that of the ball, as it rolls over the leading edge, until it bridges over the spall, often before touching bottom. It then sharply changes direction to roll out over the trailing edge of the spall.

8.3.1 Local Faults in LPM Gearbox Model

Sawalhi incorporated the bearing/pedestal model into the LPM of the gearbox of Figure 8.2 for all four bearings, but because forces were no longer restricted to the line of action of the gears, lateral DOFs had to include both x and y, doubling them. The encoders of the original model were retained, and an extra sprung mass was added at each bearing to represent a typical casing resonance at 15 kHz, to carry the bearing fault information. The extended LPM thus has 34 DOFs. Endo [2] had made an experimental modal analysis (EMA) of the rig, and the bearing pedestal supports (springs and dampers) were adjusted to give a low order rigid body mode of the casing corresponding to a measured one at 174 Hz.

Figure 8.13 shows typical response spectra from this model, with and without an outer race fault, and compares them with measured spectra.

Even though the power spectra have quite different shapes, largely because the dynamic properties of the casing have not been included, they both show that the fault only shows up (i.e. dominates over the gear vibrations) in the high frequency range where casing resonances are excited (only one in the simulation). When the band with maximum dB increase (determined by a kurtogram) is bandpass filtered and demodulated, the resulting time signals and envelope spectra are very similar, as shown in Figure 8.14.

Inner race spalls, and those on the ball itself, were also simulated, and as expected exhibited modulation at shaft speed, and cage speed, respectively. The spectrum comparisons were quite similar to Figure 8.13, but Figure 8.15 gives time signals and envelope spectra for the ball fault, for comparison with Figure 8.14. The signals are modulated at cage speed 4 Hz, with corresponding low harmonics and sidebands in the envelope spectra. The harmonic cursor is set at $2 \times \text{BSF}$, where BSF equals 'ball spin frequency' (i.e. the rate at which the fault strikes either the inner or

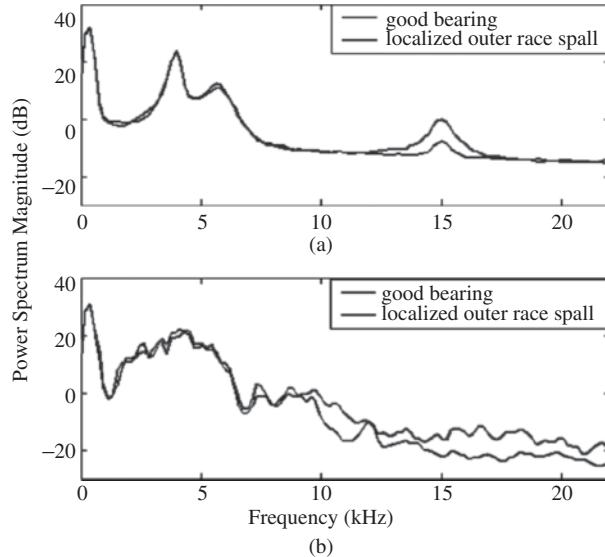


Figure 8.13 Power spectrum comparison for good and outer race defect bearings. (a) simulated, (b) experimental.

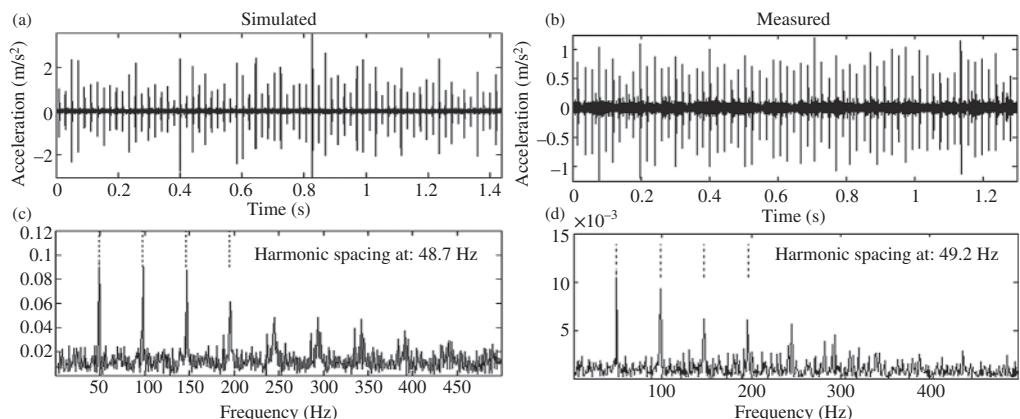


Figure 8.14 Filtered response signals for an outer race fault and their envelope spectra (a, c) Simulated (b, d) Measured (a, b) Time signals (c, d) Envelope spectra.

outer race), but the odd harmonics are also present, since the impacts with outer and inner race are different. For the optimally bandpass filtered signals, the simulated and measured results are very similar.

It should be noted that even though the full range spectra are quite different in Figure 8.13, largely because the properties of the casing are not included in the models, this is not a big restriction in generating envelope spectra corresponding to a bandpass filtered section of the spectrum in a range where a resonance is excited, as the lower frequency part of the spectrum, up to about 7 kHz, is dominated by gear vibrations, which do not directly affect the bearing signals, except to mask them. The scales of the filtered time signals and envelope spectra cannot really be compared, since the

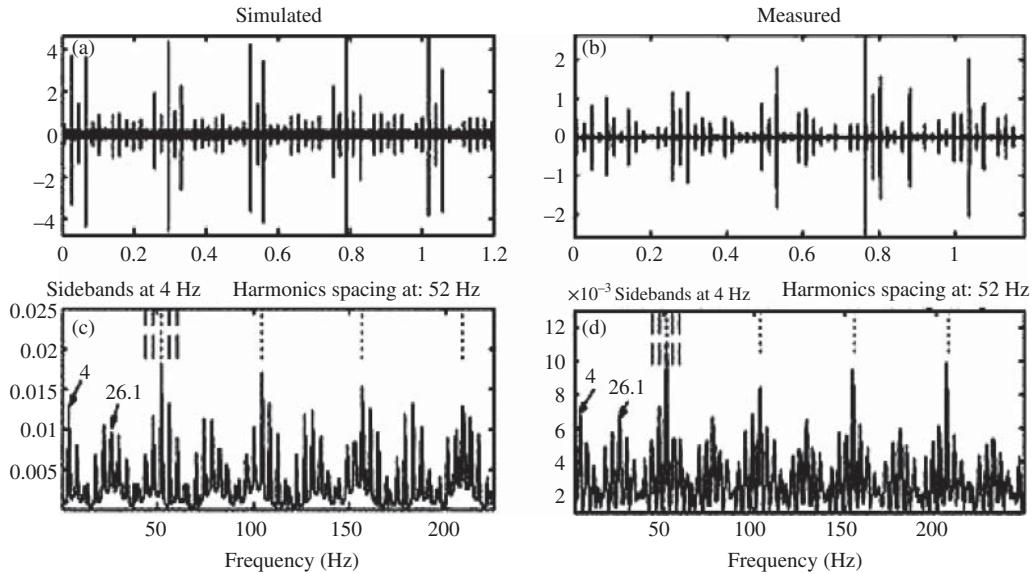


Figure 8.15 Filtered response signals for a ball fault and their envelope spectra. (a, c) Simulated; (b, d) Measured; (a, b) Time signals; (c, d) Envelope spectra.

transfer paths from the bearings to the measurement points are not replicated in the simulation model. However, it is possible to obtain a more realistic model, covering both gear and bearing effects, by combining a simulation model of the casing with that of the internals, as discussed in Section 8.3.3.

8.3.2 Extended Faults in LPM Gearbox Model

Since extended rough faults had been shown to give very different symptoms from local faults, as discussed in Sections 3.6.1 and 7.3.1, a rough fault was simulated in the bearing model and inserted into the 34 DOF LPM. The inner race of the bearing used for experimental comparison, with fault extending over 1/8 of the circumference, is shown in Figure 8.16a. The simulated rough profile was superimposed on a curve representing the approximate local depth of the actual fault, and is shown in Figure 8.16b. It was generated as a random signal, but lowpass filtered in recognition of the fact that the actual trajectory of the ball would have a minimum radius of curvature corresponding to the radius of the ball, regardless of the actual surface. More details are given in [14, 15].

Full results for the simulations will be found in Refs. [14, 15], but the most significant results are shown in Figure 8.17, these being the slices in the spectral correlation diagram (cyclic spectra) corresponding to cyclic frequencies $\alpha = \text{zero}$, and shaft speed 10 Hz. These could be compared, for example, with Figures 7.45 and 7.46 of Chapter 7. As in those cases, the differences are much greater at the cyclic frequency of the fault, but the carrier frequency band is more likely to be dominated by the gear frequency range rather than resonant responses. In this connection, it seems that the somewhat greater response at housing resonance (15 kHz for the simulation in Figure 8.17d), compared with the measurement of Figure 8.17b, is likely because the profile at exit (Figure 8.16b) is somewhat sharper than the actual, giving an exit impulse response.

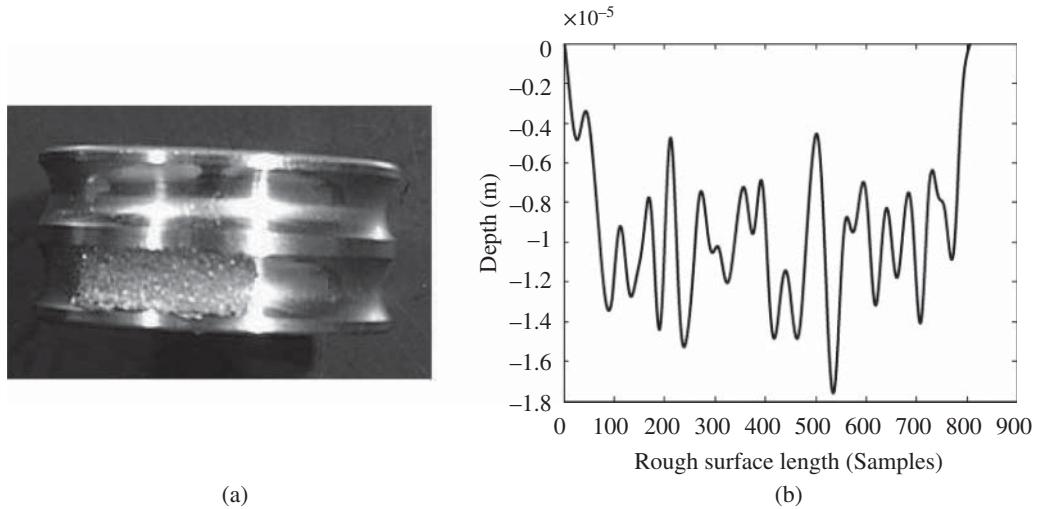


Figure 8.16 Extended rough inner race fault (a) ground surface (b) Modelled profile.

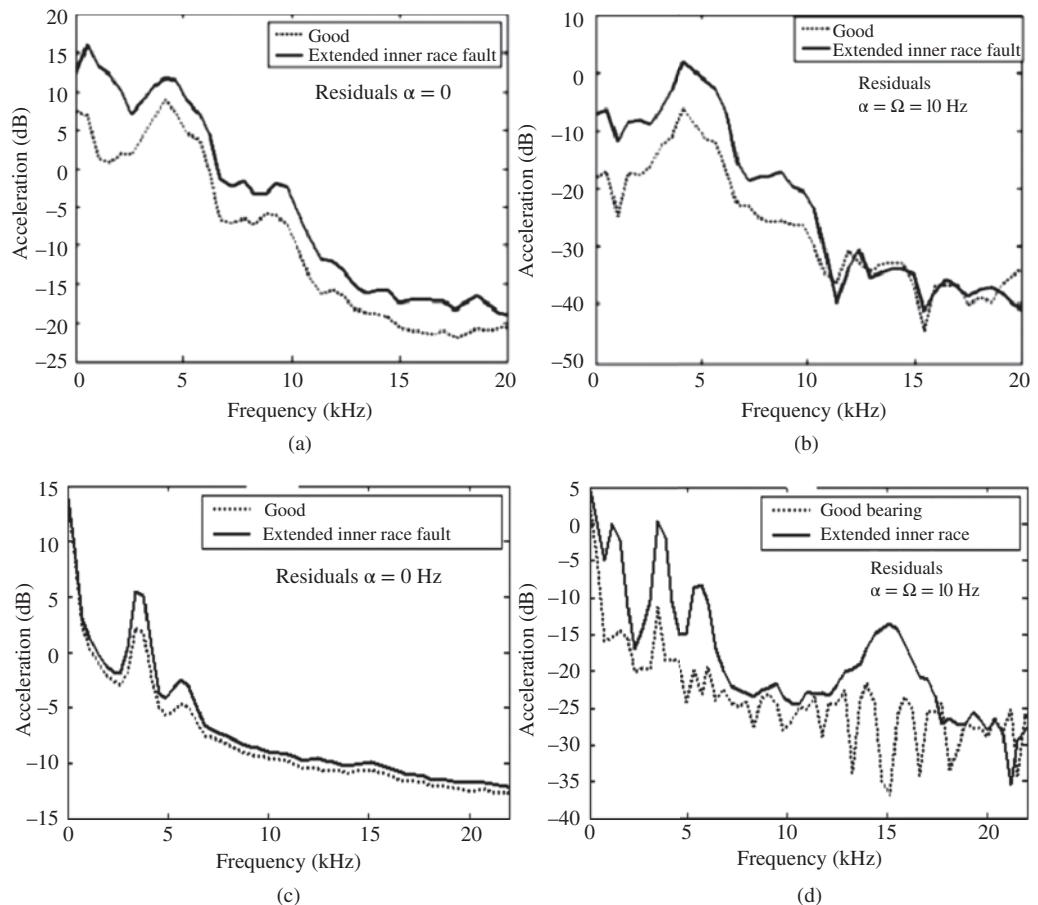


Figure 8.17 Cyclic spectrum comparisons. (a, b) measured. (c, d) simulated. (a, c) $\alpha = 0$, corresponding to power spectra; (b, d) $\alpha = 10$ Hz, rotation speed.

8.3.3 Reduced FE Casing Model Combined with LPM Gear Model

The only practical way of simulating the modal properties of the casing is via an FE model, but this immediately raises the problem that even a relatively small and simple casing, such as that shown in Figure 8.1a, has tens of thousands of DOFs, because of the large number of elements. In principle, the number of vibration modes would be the same, but the vast majority of these would lie well outside the frequency range of interest, i.e. not excited by the dominant forces generated by gear and bearing faults.

Luckily, this problem can be resolved in most cases by using model reduction. This involves dividing the DOFs into master and slave coordinates, and partitioning all matrices such that the problem size can be reduced to something of the order of the master coordinates plus an efficient coupling of the master and slave coordinates.

The results presented here correspond to the method developed by the condition monitoring group at UNSW, as presented by Sawalhi, Deshpande et al. in [16–19], but more detail of the development of the topic will be found in Deshpande’s PhD thesis [20]. Various reduction methods were examined and tried, but in the end, the method most appropriate to our problem was found to be the Craig-Bampton (C-B) method, belonging to the component mode synthesis group, and compatible with the general field of sub-structuring, which is a way of combining the models of structural components developed separately into a whole. The sub-structural components are called ‘super-elements’, and are connected in a similar manner to the individual FEs in a single object.

In the C-B method, the DOFs are divided into master coordinates, usually spatial DOFs, where the substructures join (i.e. common to both structures being joined), and/or where external forces are applied, and modal coordinates, incorporating all the slave coordinates, with the master coordinates fixed. In general, the number of such modal coordinates is vastly smaller than the original spatial coordinates from which they are derived. The advantages of the C-B method can be summarised as follows:

- 1) It generates a combined (hybrid) model where a small number of master coordinates remain as spatial coordinates. Not only does this simplify joining the substructures at these spatial coordinates (rather than having to convert them into multiple modal coordinates) but the excitation due to both gear and bearing faults is largely by geometric mismatch (of the spatial coordinates) between the mating parts of components such as bearing races and gear teeth, rather than forces as such.
- 2) Time varying and nonlinear properties are also largely confined to these interfaces, and can be separated out from an otherwise constant parameter linear model. On the other hand, it is most likely that the modal properties of the casing do not change substantially with fault development in the machine, usually dominated by forcing functions, unless of course a crack develops, but even in that case it is not very onerous to update the modal model.
- 3) The number of required modal coordinates is defined by the desired frequency range of the results, thus defining the number of modes to be included. Plate-like structures (which gearbox casings behave like) have a roughly constant modal density [21], making the number of modal coordinates approximately proportional to the frequency range. Of course the C-B modes, with master DOFs fixed, do not have the same frequencies as the equivalent modes in the combined structure, but the high order modes are not very different, since the combination of sinusoidal and exponential shapes is dominated by the sinusoidal part, as soon as there are more than about three wavelengths of the sinusoidal components between disruptions (usually at the master coordinates) around which there are exponential decays.

The results of this section demonstrate what a huge difference this DOF reduction can make.

8.3.3.1 The Craig-Bampton Reduction Method

This exposition is basically taken from [16], which was one of our first applications of the C-B method.

In the Craig-Bampton reduction method, the dynamic equilibrium of each superelement (substructure), without considering the effect of damping, can be expressed in the EOM, Eq. (8.19):

$$[M]\{\ddot{u}\} + [K]\{u\} = \{F\} \quad (8.19)$$

where $[M]$ is the mass matrix, $[K]$ is the stiffness matrix, $\{F\}$ is the nodal forces, $\{u\}$ and $\{\ddot{u}\}$ are the nodal displacements and accelerations respectively.

The key to reducing the substructure is to split the DOFs into master $\{u_m\}$ (at the connecting nodes) and slaves $\{u_s\}$ (at the internal nodes). The mass, the stiffness and the force matrices are re-arranged accordingly as follows:

$$\underbrace{\begin{bmatrix} M_{mm} & M_{ms} \\ M_{sm} & M_{ss} \end{bmatrix}}_M \begin{Bmatrix} \ddot{u}_m \\ \ddot{u}_s \end{Bmatrix} + \underbrace{\begin{bmatrix} K_{mm} & K_{ms} \\ K_{sm} & K_{ss} \end{bmatrix}}_K \begin{Bmatrix} u_m \\ u_s \end{Bmatrix} = \begin{Bmatrix} F_m \\ 0 \end{Bmatrix} \quad (8.20)$$

The subscript m denotes master, s denotes slave. Furthermore, the slave DOFs (internals) can be written using generalised coordinates (modal coordinates (q)) using the fixed interface method, i.e. using the mode shapes of the super-element by fixing the master DOF nodes (connecting/boundary nodes). The transformation matrix (T) is the one that achieves the following:

$$\begin{Bmatrix} u_m \\ u_s \end{Bmatrix} = T \begin{Bmatrix} u_m \\ q \end{Bmatrix} \quad (8.21)$$

For the fixed interface method, the transformation matrix (T) can be expressed as shown in Eq. (8.22):

$$T = \begin{bmatrix} I & 0 \\ G_{sm} & \phi_s \end{bmatrix} \quad (8.22)$$

where,

$$G_{sm} = -K_{ss}^{-1}K_{sm} \quad (8.23)$$

and ϕ_s is the modal matrix of the internal DOFs with the interfaces fixed.

Applying this transformation, the number of DOFs of the component will be reduced. The new reduced mass and stiffness matrices can be extracted using Eqs. (8.24, 8.25) respectively:

$$M_{reduced} = T^t M T \quad (8.24)$$

and

$$K_{reduced} = T^t K T \quad (8.25)$$

Thus Eq. (8.20) can be re-written in the new reduced form using the reduced mass and stiffness matrices as well as the modal coordinates as follows:

$$\underbrace{\begin{bmatrix} M_{bb} & M_{bq} \\ M_{qb} & M_{qq} \end{bmatrix}}_{M_{reduced}} \begin{Bmatrix} \ddot{u}_m \\ \ddot{q} \end{Bmatrix} + \underbrace{\begin{bmatrix} K_{bb} & 0 \\ 0 & K_{qq} \end{bmatrix}}_{k_{reduced}} \begin{Bmatrix} u_m \\ q \end{Bmatrix} = \begin{Bmatrix} F_m \\ 0 \end{Bmatrix} \quad (8.26)$$

where M_{bb} is the boundary mass matrix i.e. total mass properties translated to the boundary points, K_{bb} is the Interface stiffness matrix i.e. stiffness associated with displacing one boundary DOF while the others are held fixed, and M_{bq} is the component matrix (M_{qb} is the transpose of M_{bq})

If the mode shapes have been mass normalised (typically they are) then:

$$K_{qq} = \begin{bmatrix} \lambda_i & 0 \\ 0 & \lambda_i \end{bmatrix} \quad (8.27)$$

where λ_i are the eigenvalues; $\lambda_i = k_i/m_i = \omega_i^2$ and,

$$M_{qq} = \begin{bmatrix} I & 0 \\ 0 & \omega^2 \end{bmatrix} \quad (8.28)$$

Finally, the dynamic EOM (including damping) using the Craig-Bampton transform can be written as:

$$\begin{bmatrix} M_{bb} & M_{bq} \\ M_{qb} & I \end{bmatrix} \begin{Bmatrix} \ddot{u}_m \\ \ddot{q} \end{Bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 2\zeta\omega \end{bmatrix} \begin{Bmatrix} \dot{u}_m \\ \dot{q} \end{Bmatrix} + \begin{bmatrix} K_{bb} & 0 \\ 0 & \omega^2 \end{bmatrix} \begin{Bmatrix} u_m \\ q \end{Bmatrix} = \begin{Bmatrix} F_m \\ 0 \end{Bmatrix} \quad (8.29)$$

where $2\zeta\omega$ = modal damping (ζ = critical damping ratio).

8.3.3.2 Responses from Combined Reduced Models

This description of the approach is basically taken from [17], which compares the results from two reduced models of the gearbox of Figure 8.1a, one combining the 34 DOF LPM of Section 8.3.1 with a C-B reduced model of the casing, and another where the internals are also modelled by FEs, but with a greatly reduced number of elements than for the casing. Only the details of the latter model are given here, but the results from both are discussed. The method was improved between Refs. [16–19], and different faults were modelled in some papers, so the comparative results shown here are largely from Ref. [18] for time records of local faults, and Ref. [19] for spectral results from local and extended faults. The latter paper includes an improvement of the FE model of the gearbox, using updating techniques based on EMA. This is always to be preferred, as agreement of natural frequencies is critical to a good representation of waveforms, at least up to frequencies where the modes are separated. Figure 8.18a shows the FE model of the casing, with connecting elements to the bearings of the LPM and FE models of the internals, and Figure 8.18b shows the FE model of the internals.

As described in [17] the casing was modelled with both shell and solid elements, whereas the shafts were modelled with beam elements and gears with shell elements. Flywheels and encoders were modelled using ‘mass’ elements with appropriate inertial properties. Before reduction, the FE model of the casing had 104 340 DOFs whereas each shaft-gear assembly (with flywheel and encoder) had 840 DOFs.

Table 8.1 [20] gives details of the DOFs of the reduced model, with a total of 182, of which 40 are spatial coordinates and 142 are modal coordinates.

The actual measurement points on the gearbox were originally ‘internal’ coordinates in the C-B model of the casing, so after solving the equations in the time domain using Simulink® for various fault types in the bearings, the responses at a particular measurement point were originally extracted back from the modal coordinates for comparison with experimental measurements [16]. This is shown as Approach A in Figure 8.19 [20]. However, the range of modes chosen for the reduced models in Table 8.1 correspond to a frequency range of 8.7 kHz for the internals and 4.4 kHz for the

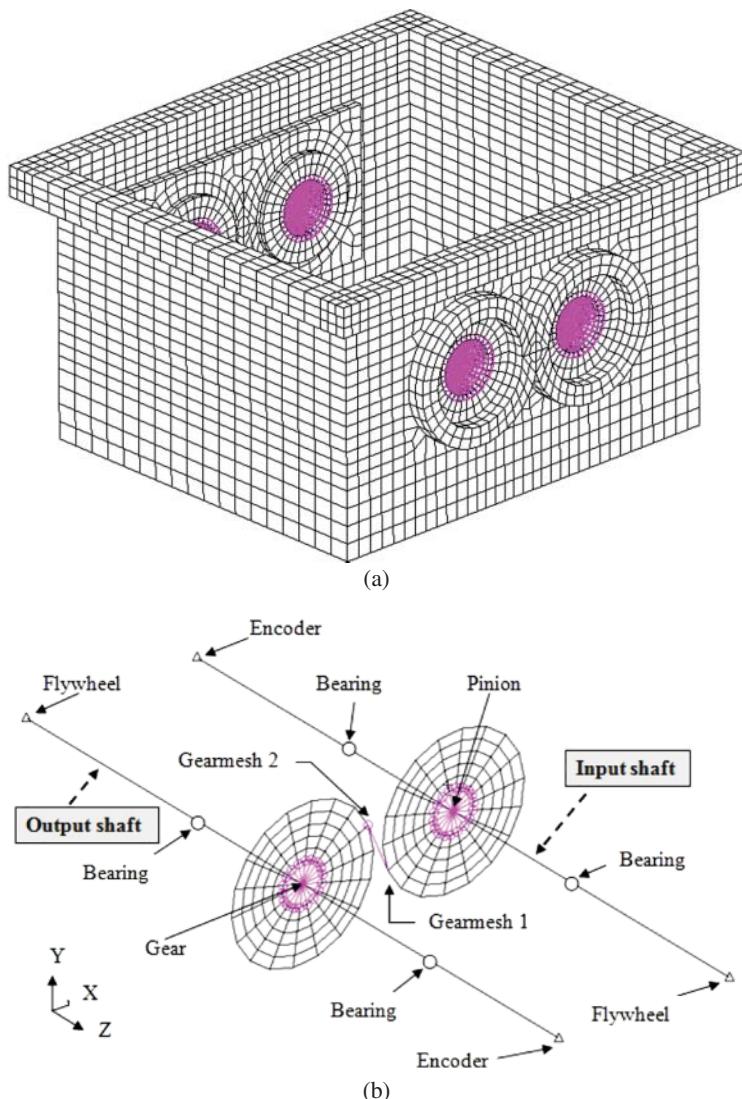


Figure 8.18 FE model – (a) gearbox casing (b) internals.

Table 8.1 Full reduced FE model – summary of DOFs.

Reduced FE model	Physical (spatial) coordinates	Generalised (modal) coordinates	Total DOFs
Gearbox internals	16	42	58
Casing	24	100	124
Entire gear box	40	142	182

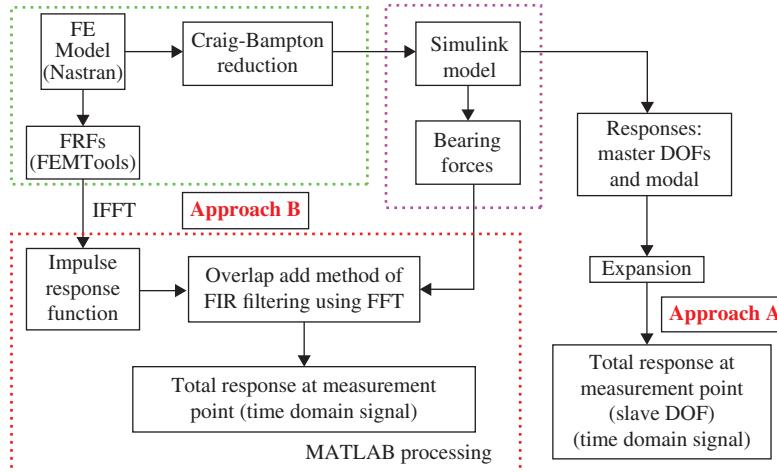


Figure 8.19 Evaluation of total response from reduced models.

casing, which is adequate to study the gear/bearing interaction in the frequency range below 5 kHz, but limits the result outputs to this range in terms of the casing response. It was thought, however, that the forces at the bearings would be little affected by the casing resonances at higher frequency, so these forces could be applied to the full FE model of the gearbox (before reduction), as in Approach B in Figure 8.19. Time invariant FRFs derived from this linearised model, inverse transformed to IRFs, are not computationally onerous, and permit generation of time domain responses up to higher frequencies, in this case up to 20 kHz, including the resonances of the casing up to that frequency. The overlap-add method mentioned in Figure 8.19 is a commonly used efficient convolution method, via multiplication in the frequency domain, at the same time accounting for the circularity of the Fast Fourier Transform (FFT) transform.

Ref. [17] compares power spectra for an extended inner race fault between experimental measurements and three simulations; the LPM as in Figure 8.17, and both the 146 and 182 DOF reduced models. The reduced models gave more realistic results than the LPM, but there was not much difference between them. This is data dependent, and in general an FE model of the internals would be better because it would allow higher order shaft modes. Ref. [18] gives results for localised inner and outer race faults, comparing the test measurements with the LPM and 146 DOF reduced model. It did appear that better spectrum correspondence would be obtained by adjusting the simulation model, both in terms of structural dynamics and in modelling the fault geometry, and this was left for later work. However, excellent results were obtained for simulated time signals, which are reproduced here. These were pre-processed by cepstrum pre-whitening (Section 6.3.3) which de-enhances the effects of spectrum shape, and in particular reduces the masking by low frequency gear signals.

Figure 8.20 gives simulated signals using the 146 DOF reduced model for both a localised inner race fault and a localised outer race fault. In the zoomed figure for a single fault impact, the entry and impact events are indicated, and their timing corresponds to the trajectory of the centre of the ball indicated in Figure 8.20e,f, for which this exact timing is known.

Figure 8.21 gives the corresponding test data for comparison. In this case, the entry and impact can only be inferred, but seem to correspond to the simulations.

For the inner race fault, the gentler entry event can be seen in both simulated and measured signals, but interestingly the entry event for the outer race fault is not clearly defined for either. Since

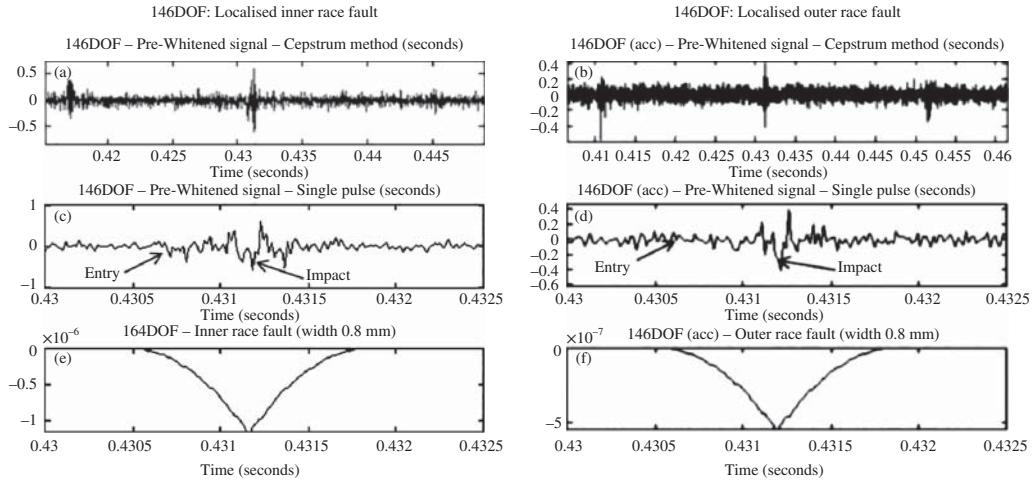


Figure 8.20 Pre-processed time signals from simulated data for localised faults. (a, c, e) Inner race fault; (b, d, f) Outer race fault; (a, b) Three consecutive fault impacts; (c, d) zoom on central event; (e, f) Trajectory of ball centre, showing that impact event occurs at lowest position.

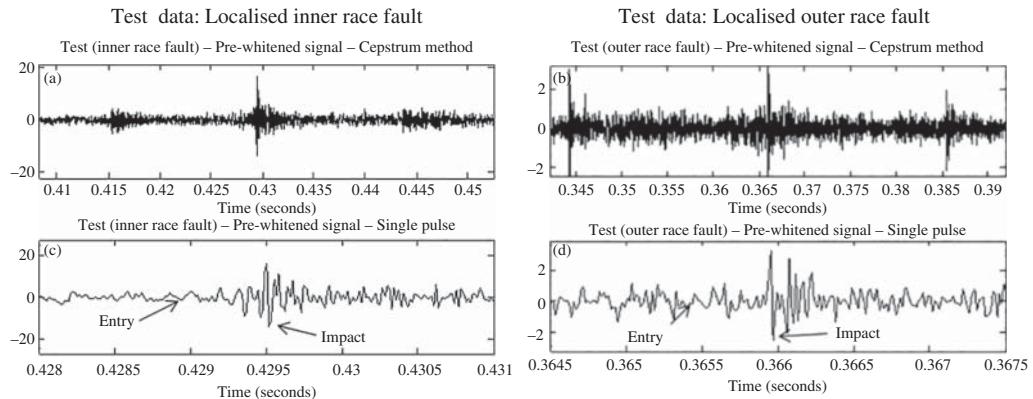


Figure 8.21 Pre-processed time signals from test data for localised faults. (a, c) Inner race fault; (b, d) Outer race fault; (a, b) Three consecutive fault impacts; (c, d) zoom on central event.

these results were published, considerable work has been done on enhancing these entry and impact signals, and this is discussed in more detail in Section 9.2.3, because of its importance in fault prognosis.

It is worth pointing out that the three successive impacts for both simulated and measured inner race faults are amplitude modulated as the fault passes through the load zone, but not for the outer race fault, which is always in the load zone.

As described in [19], the best spectral results were obtained using a reduced FE model which had been updated on the basis of EMA. For the EMA, the gearbox casing was suspended in bungee cords, to give free-free boundary conditions, the easiest to replicate experimentally. The suspension, and positions of the three response accelerometers (x , y , z directions) are illustrated in Figure 8.22.

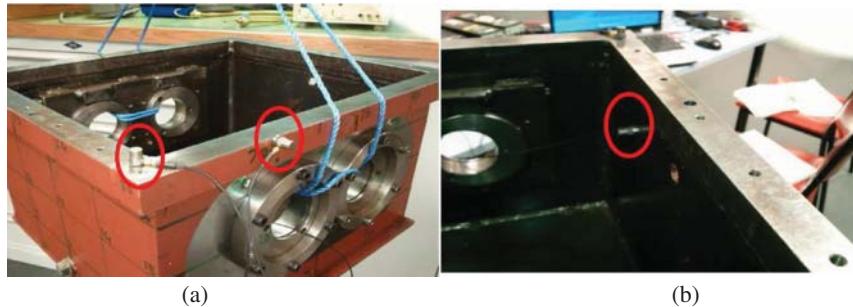


Figure 8.22 Casing suspension and accelerometer locations.

The EMA was done using B&K Pulse hardware and Modal Test Consultant software, using impact hammer excitation, and the measured FRFs were exported to ME'scope software to generate the modal frequency, modal damping and mode shape for each mode. Global curve fitting (polynomial method) was used for the global parameters and the residues were derived using these.

The FE model of the casing was imported into the model updating software FEMTools[®] along with the EMA results. This software achieved a pairing of the coarser EMA mesh with the FE mesh, and mode shape pairing was achieved by correlation using MAC (Modal Assurance Criterion) values. FEMTools modifies specified parameters of the FE model so as to minimise a customer-specified objective function, in this case based on the differences in natural frequencies of equivalent modes. This was done for the frequency range up to 2 kHz, where the modes were still separated. Major errors in the modelling would be due to the lack of full metal-to-metal contact at welded joints between components and the fact that weld fillets themselves were not modelled. Based on previous experience, the parameter chosen to update the model was the thickness of the elements, since these are distributed over the whole model, including adjacent to areas requiring modification. Also, the stiffness of a plate element is proportional to the cube of the thickness, and the mass directly to the thickness, so it is quite sensitive. Figure 8.23 shows the result of the update in terms of the relative change in (hypothetical) thickness of the elements in order to achieve the best match of natural frequencies.

Figure 8.24 compares two typical mode shapes after updating, showing an excellent correspondence.

Ref. [19] contains a complete table of the natural frequencies of the test data and FE model before and after updating. After updating, the difference is less than 0.3% for 8 of 17 modes, with mean (maximum) difference 1.3 (6.75)%, compared with 3 modes <0.3%, and mean (maximum) difference 4.4 (12.35)% before updating.

Figure 8.25 compares test results for a localised inner race fault with simulated results based on a reduced model derived from the updated FE model just described. The reduced model was a 146 DOF model, where the C-B modal coordinates were obtained from the updated FE model by changing the boundary conditions from free to fixed.

Figure 8.25a,c compare power spectral density (PSD) spectra obtained using Method B (of Figure 8.19), with measured spectra in the frequency range up to 20 kHz, with and without the fault. While not perfect, the results are clearly much better than obtained from an LPM. Much of the difference in the frequency range above 5 kHz can be ascribed to the fact that the simulations assumed constant speed, whereas the small random speed variations for the test caused smearing of the higher order shaft harmonics of the gears, so that they did not protrude from the base noise level. Neglecting these discrete frequencies appearing in the higher frequency range, which are removed

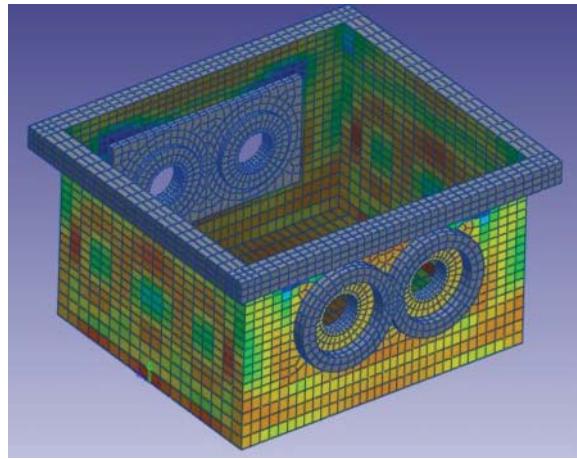


Figure 8.23 Relative thickness change in updated model.

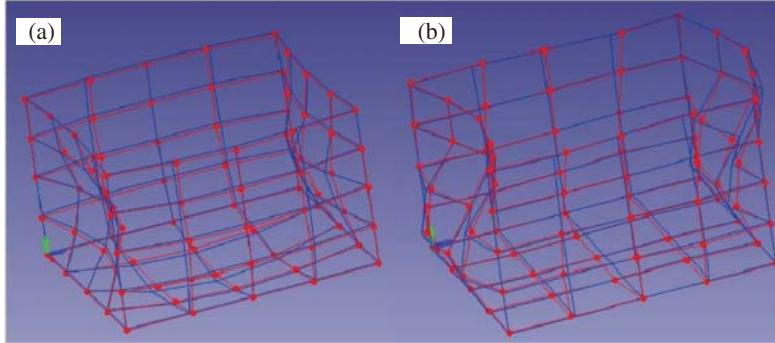


Figure 8.24 Mode shape comparisons, test vs updated FE. (a) Mode shape pair 02; (FEA 350.77 Hz, EMA 345.68 Hz, MAC 88.7%). (b) Mode shape pair 10; (FEA 930.52 Hz, EMA 882.92 Hz, MAC 84.4%).

before envelope analysis in any case, the spectrum differences due to the fault are quite comparable, and this is confirmed by the similarity of the squared envelope spectra in Figure 8.25b,d, containing harmonics of ballpass frequency, inner race, (BPFI) with sidebands spaced at shaft speed.

Ref. [19] also gives results from a simulated extended rough inner race fault, as characterised by spectral correlation density (SCD) at zero cyclic frequency (i.e. normal PSD) and at shaft speed 10 Hz (characteristic of inner race faults). Figure 8.26 compares these SCD spectra up to 10 kHz, where the most important information is contained. This confirms that the extended fault gives much greater differences in the lower frequency range, occupied by gear harmonics, made much more obvious by the removal of the latter, as recommended. Unfortunately, the same analysis had not been done on the test measurements, but comparison of Figure 8.26d can be made with a result taken many years earlier from the same gearbox, but with different gears and bearing shaft speed 15 Hz, and shown in Figure 8.27. In fact, it is the same as Figure 7.45b, but with x-axis recalibrated in Hz, rather than orders.

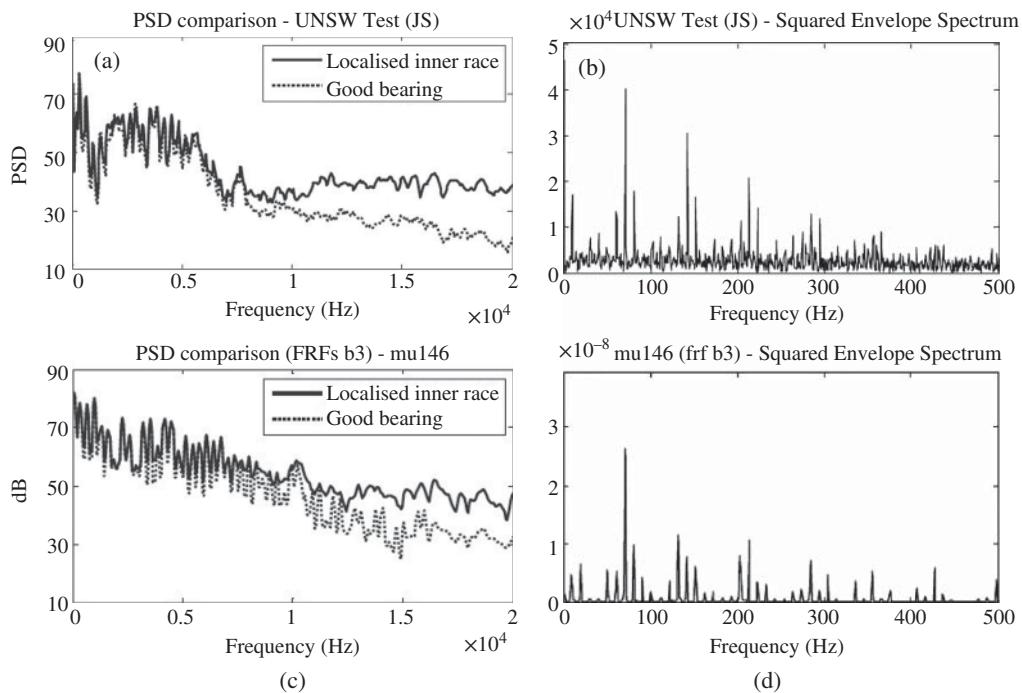


Figure 8.25 Comparison of PSD spectra and squared envelope spectra for local inner race fault.
 (a, b) Test results; (c, d) Simulation; (a, c) PSD spectra (b, d) Envelope spectra.

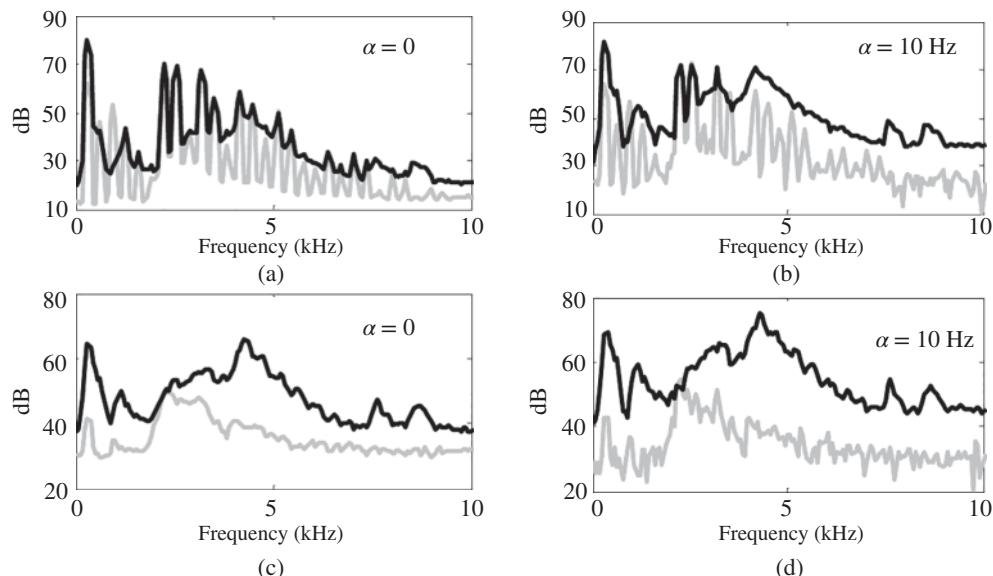


Figure 8.26 SCD comparisons for healthy case and extended inner race fault, before and after removal of discrete gear harmonics. (a, b) before removal (c, d) after removal.

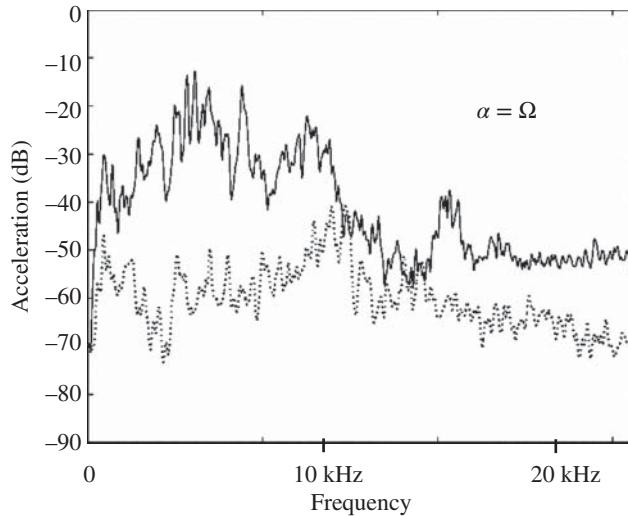


Figure 8.27 Measured spectral correlation density before and after comparable extended inner race fault, though with shaft speed 15 Hz and with non 1 : 1 gear ratio.

8.4 Simulation of Faults in Engines

The simulation of faults in this section covers combustion faults (e.g. misfires) and mechanical faults such as piston slap and bearing knock in internal combustion (IC) engines.

8.4.1 Misfire

8.4.1.1 Large Diesel Engine with Flexible Crankshaft

Simulation of combustion faults in a 20-cylinder diesel engine was mentioned in Chapter 7 (Ref. [70]), and so is taken up again here as Ref. [22].

A number of nominally identical engines were available to give sufficient backup power to shut down the reactor in a nuclear power plant, in the case of power failure. This meant, however, that they were only operated occasionally, primarily for testing purposes, so very little operational measurement data was available. It was thus desirable to have a simulation model of the engine to guide the diagnostics of faults, with one tool being the measurement of torsional vibrations at the accessible free end of the crankshaft. This was done with a torsional laser vibrometer.

Because the torsional natural frequencies of the crankshaft were within the excitation range of combustion frequency harmonics of this large engine, the modal properties had to be taken into account, since the same torque variation in different cylinders would give a different torsional vibration response at different positions along the crankshaft.

Figure 8.28 shows the simplified torsional model of the engine crankshaft and its connections at either end; the generator connected via a coupling to the flywheel and a tuned damper at the free end. As described in [22], it was a V-engine with two pistons per crank, and so the 10 cranks (with connecting rods and pistons) were modelled as fixed rotational inertias, by standard techniques, allocating an equivalent rotational inertia for the reciprocating components. The stiffness of the crankshaft sections between cranks was estimated using FE analysis. The stiffness and inertia of

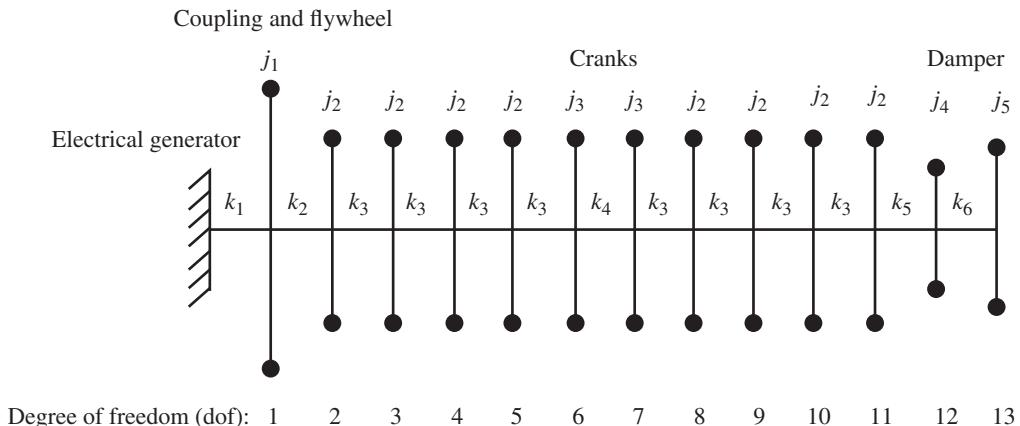


Figure 8.28 Torsional model of crankshaft.

the coupling/flywheel and tuned damper were taken from catalogue values and adjusted to match measured natural frequencies. Since the generator was locked to mains frequency, it was treated as fixed in torsion (at synchronous speed). The laser vibrometer measurement was located at node 12, the inside of the damper at the end of the crankshaft.

Analytical modal analysis (for zero damping) was carried out on this LPM, giving 13 modes with their natural frequencies, of which the first five were found to dominate the waveform of the measured torsional vibrations, and modes 1–4 are depicted in Figure 8.29. The mode shapes can be understood as follows:

Mode 1 is basically the inertia of the whole crankshaft acting on the spring of the coupling. Note that the crankshaft is not completely rigid, with a slight elastic deformation along its length.

Mode 2 is basically the first torsional mode of the crankshaft itself, with a linear deflection along most of it, as for a uniform rod in torsion, offset at the left hand (LH) end by interaction with the flywheel, but with a jump at the relatively soft spring of the damper.

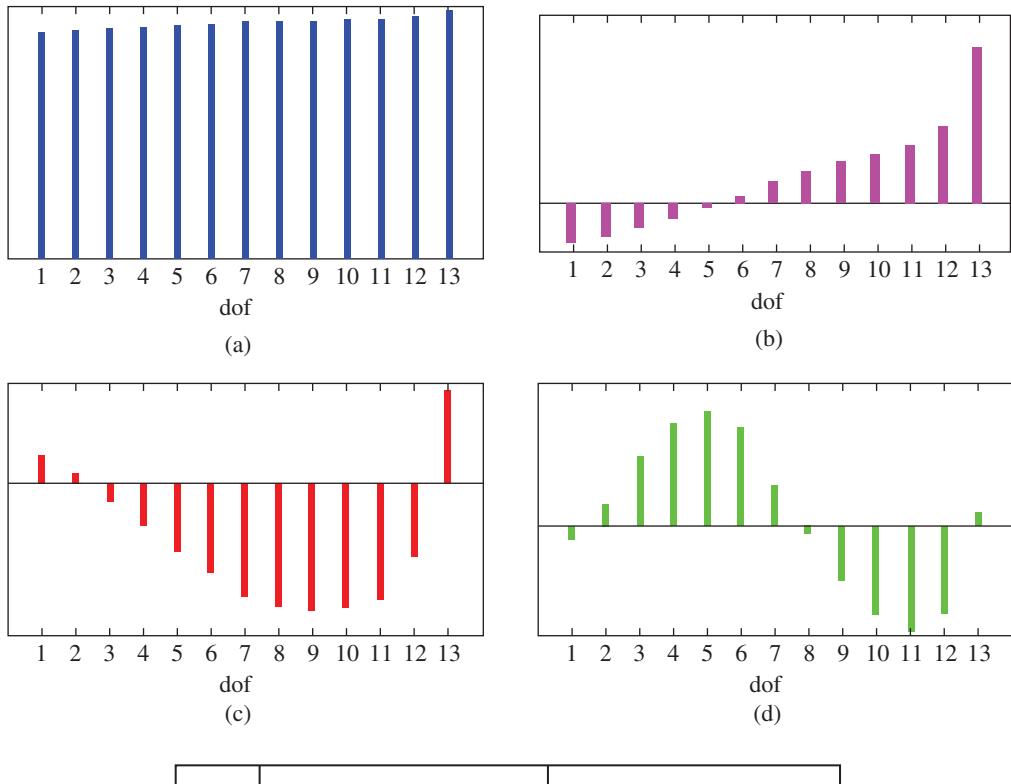
Mode 3 is basically the second elastic mode of the crankshaft, and resembles the second mode of a uniform rod in torsion (a quarter wavelength sinewave), also offset at the LH end, and with a sudden transition at the damper.

Mode 4 is likewise the third mode of the crankshaft, which for a uniform rod would be a $\frac{3}{4}$ wavelength sinewave, but with similar deviations at each end.

The ‘observed frequencies’ in the table were extracted from measurements of a run-up of the engine, for which the spectrogram is shown in Figure 8.30a. When integrated over all time, the resulting spectrum in Figure 8.30b has peaks representing the first few modes, numbered 1–5 for correspondence with the table in Figure 8.29.

Not only the frequencies, but also the damping of each mode was estimated from the 3 dB bandwidth of these peaks in Figure 8.30b, although the values could not be expected to be accurate as they might be broadened by a too fast run-up.

The modal model was used to estimate FRFs from each cylinder to measurement DOF 12 (response measurement point), effectively using Eq. (8.16), but with modal damping, calculated from the estimated 3 dB bandwidths, added into the model. The total response could then be calculated from the torque applied at each cylinder.



Mode	Estimated frequency (Hz)	Observed frequency (Hz)
1	7.2	9.4
2	33.2	37.3
3	63.0	51.8
4	132.4	≈ 125
5	210.2	≈ 200

Figure 8.29 Mode shapes and natural frequencies. (a–d) Mode shapes 1–4, respectively. The table gives the estimated frequencies for modes 1–5, and compares them with observed frequencies (described above).

In order to calculate the input torque for each cylinder, for varying amounts of fuel injected, it was necessary to simulate the curve of cylinder pressure vs crank angle for varying conditions. Use was made of Wiebe's functions for a diesel engine [23], and Figure 8.31 shows the fitted curve for a measured case for 75% of full load. The fitted model then allows predictions for other fuel injection parameters.

The analytical modal model was then used to predict torsional vibration waveforms for a given operating condition, initially with estimated natural frequencies and 'observed' damping (i.e. taken from Figure 8.30), and then with observed frequencies and damping (Figure 8.32a,b). This gave a considerable improvement, but the best result was obtained by optimising the frequencies and damping (using a genetic algorithm) to minimise the differences in waveform, and this gave a near perfect result, as shown in Figure 8.32c, despite using the analytical mode shapes. This emphasises how critically waveforms depend on exact natural frequencies. The same procedure was found to

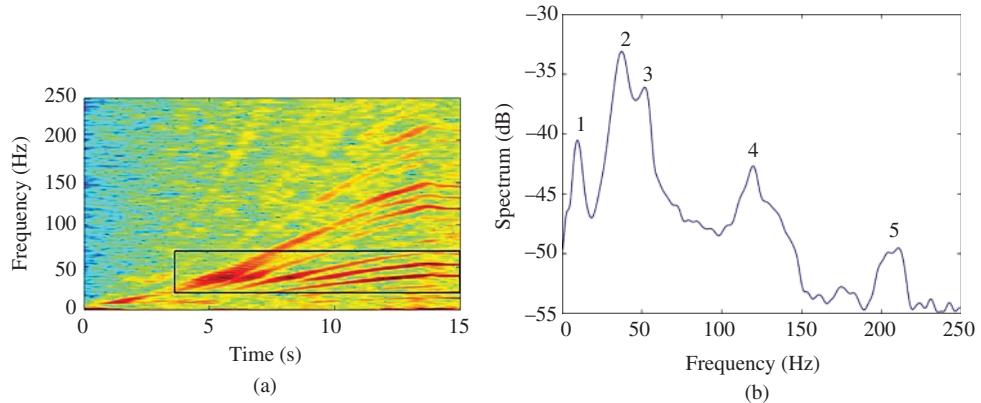


Figure 8.30 Torsional vibration resonances from run-up (a) Spectrogram (b) Integrated power spectrum.

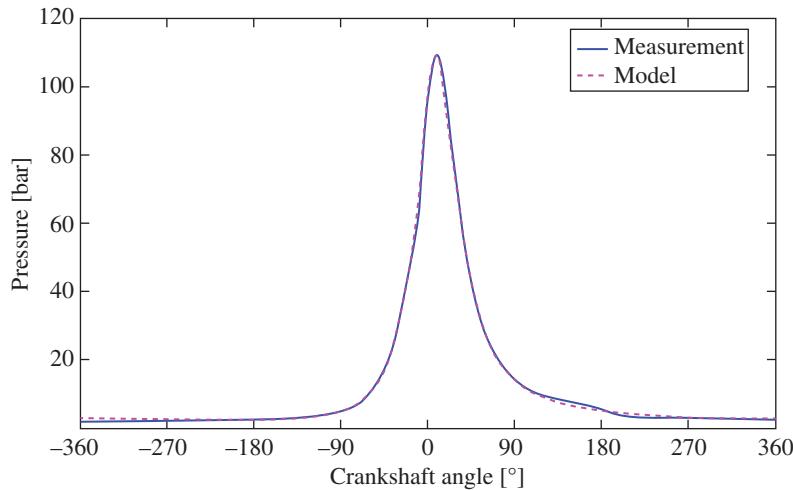


Figure 8.31 Wiebe model vs measured cylinder pressure.

give an equally good match for signals from another diesel engine, which had exactly the same LPM, but quite different waveforms.

Ref. [22] explains how the updated model was then used to generate data for the simulation of a small number of fault conditions introduced into the engine by variation of the heat release during injection. The simulated data was used to train neural networks which were then tested on the corresponding measured data. Three separate networks were trained, rather than a single network for all permutations and combinations:

- 1) Healthy or faulty?
- 2) If faulty, which cylinder?
- 3) Fault severity

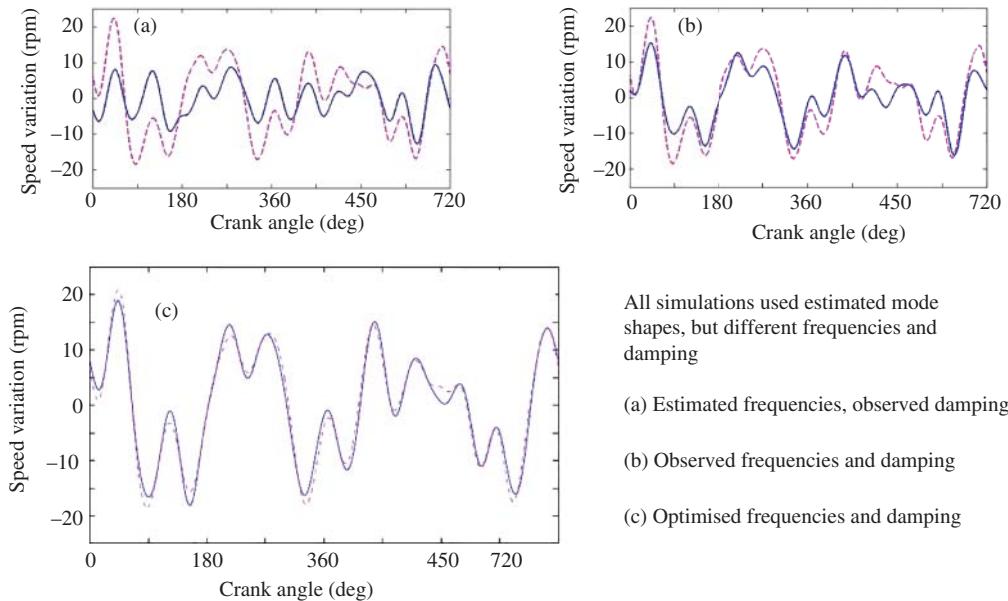


Figure 8.32 Waveforms obtained by adjusting frequency and damping parameters.

Since the output of a simulation model is deterministic, while actual responses are subject to some random variations, a simple adjustment was made for this by basing the latter on two measurements in normal condition two years apart, using the difference as an estimate of the standard deviation (s.d.) of the variations. The networks trained purely on simulated data were successful in diagnosing the source and amount of variation in heat release for all cases where the latter was <75% of that in other cylinders, that limit being for the worst case of a cylinder near the generator end of the crankshaft, where all mode shape components are smaller. The amount of data was however insufficient to be statistically viable, and a later study, described below, gave more reliable results.

8.4.1.2 Four Cylinder Spark Ignition Engine

The success of the method just described for a large diesel engine inspired a much more extensive study of this approach, embodied in the PhD thesis work of Jian Chen described in [24]. The study was financed by an Australian Research Council Linkage grant, with support from the Belgian company LMS in Leuven (now Siemens). As discussed in the next sections, that study also included mechanical faults, represented by piston slap and bearing knock. Even though this study was for a small engine, with effectively rigid crankshaft, the extra complication of modal responses depending on the location of the source was found to affect a second method used to diagnose misfires, viz., variations in engine block angular acceleration. This implies that the approach based on torsional vibration could easily be extended to engines with flexible crankshafts (which was in fact allowed for in the crankshaft model).

The test engine was a four-cylinder Toyota 3S-FE spark ignition engine (two examples), with firing order 1-3-4-2, and is depicted in Figure 8.33. It was equipped with five accelerometers (placement shown in Figure 8.33b), two proximity probes (once-per-rev tacho and ring gear tooth detection to provide timing and act as a shaft encoder for frequency demodulation), and a Kistler cylinder

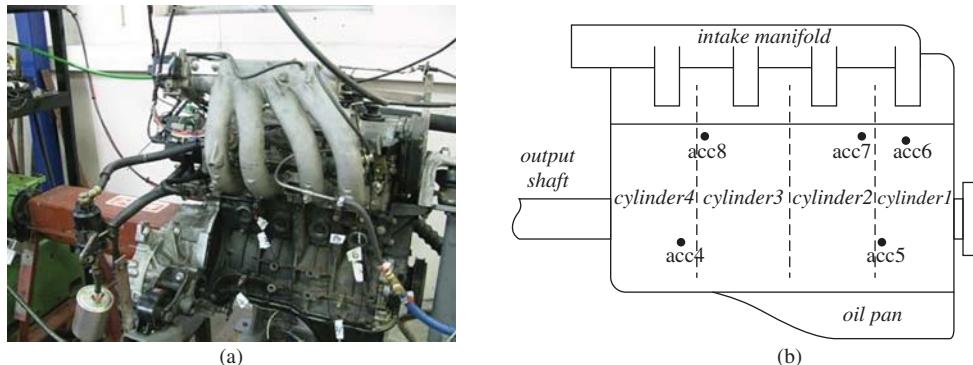


Figure 8.33 Test engine (a) Photograph (b) Accelerometer layout.

pressure sensor in a spark plug, which could be used in any cylinder. The accelerometer positions were limited by available bosses on the engine block casting.

Figure 8.34 shows typical signals from the sensors. The once-per-rev tacho signal is an exact phase reference for the rotation of the crankshaft, but only every second one defines the start of an engine cycle. The pressure transducer signal allows to distinguish between the compression and exhaust strokes, even though the phase of the pressure signal changes with advance/retard of the ignition. It will be seen that the ring gear tooth passage signal has a certain amount of ‘DC drift’ because of minor eccentricity of the flywheel, but this is automatically removed by frequency demodulation around the carrier frequency corresponding to the mean tooth passage frequency.

A somewhat updated version of what was included in the thesis [24] on this topic was published in [25]. Two approaches were used to diagnose and simulate combustion faults in the engine, viz. crankshaft angular velocity, and torsional acceleration of the engine block, which had been suggested by a number of authors, such as [26].

The (separate) simulation models for crankshaft and block motion were built in the LMS AMESim software package. This required a knowledge of the stiffness and inertial properties of the various elements of the LPMs. Some were estimated, such as the stiffness of crankshaft sections by FE analysis, (as for the larger engine in Section 8.4.1.1, but less important here because the crankshaft was effectively rigid), while others were measured directly, or measurements used to update analytical estimates. The latter applied to the inertial properties of the whole engine, which were measured in two different ways, using EMA. In the first, for the bare assembled engine, it was suspended free-free in a soft suspension (using bicycle tubes), and the four inertial properties (one mass and three rotational inertias) were estimated from the rigid body modes, in the mass line region between the suspension resonances and the lowest elastic mode of the engine. The second method involved measuring the rigid body modes of the engine in its mounts, with connections to exhaust system, cooling water, etc. This not only gave the effective inertias as constrained by the mounting and assembly, but also the linearised stiffness parameters of the engine mounts. Excitation for the modal analyses was by impact hammer, with an appropriate interface to limit the excitation frequency range, and using a drift between the hammer and the engine to better control the location and direction of the impact force.

The combustion faults (misfires of 0%, 50%, and 100%) in various cylinders were again simulated using Wiebe’s functions, this time for spark ignition engines [23], as shown in Figure 8.35.

The angular velocity of the crankshaft was measured at the flywheel, by frequency demodulation of the ring gear tooth passage signal shown in Figure 8.34. This was done by the procedure described in Section 4.3.3 (though the method of Section 5.2.1 could have been used). This approach was by

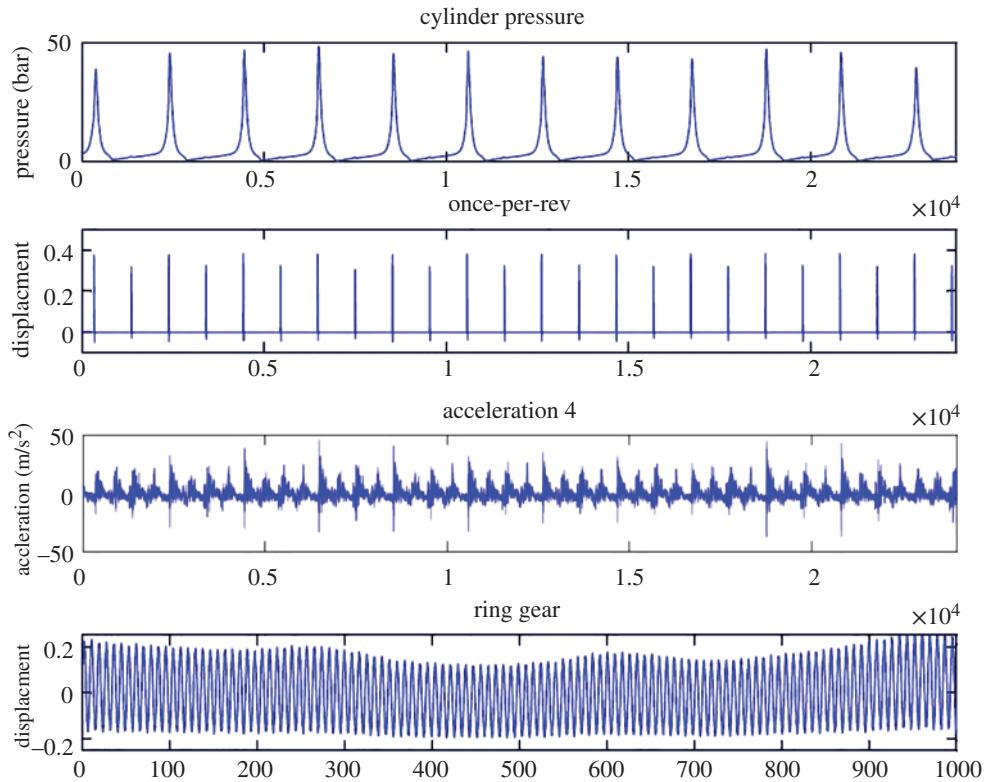


Figure 8.34 Typical signals for the engine at constant speed.

far the simplest, and easiest to simulate, but this was partly because the crankshaft was rigid in the dominant frequency range, and thus the results were little affected by engine speed.

Figure 8.36 shows experimental and simulated crankshaft angular velocities, for one speed (1500 rpm) and load (80 Nm) with and without misfiring cylinder (no. 1), but a number of others are given in [25]. It is obvious from this figure that the healthy signal is dominated by the firing frequency (two times the crankshaft speed for a 4-cylinder, 4-stroke engine), while the signal for a misfire in one cylinder is dominated by the cycle frequency (half the crankshaft speed). This in fact formed the basis of the automated detection of misfires as described in Section 9.3.4 of Chapter 9. The cylinder with the misfire is clearly indicated by the phase of the signal, as this is where the velocity dips instead of increasing. For the (almost) rigid crankshaft, the phasing of these components was independent of load, and almost independent of speed, as shown in Figure 8.37, which compares polar diagrams for two different speeds (and loads) for misfire in Cylinder 1. The 360° of the diagram corresponds to a complete engine cycle, and thus 720° of crankshaft rotation.

It is seen that the phase of the first harmonic, indicating the misfire in Cyl. 1, is independent of the speed and load, but the phase of the fourth harmonic, has shifted a little at this highest speed, presumably because of interaction of the higher harmonics with the fixed natural frequencies of the crankshaft, which are just starting to come into play at this speed. This is discussed further in Section 9.3.4 on automated diagnostics and prognostics.

The other approach involved the angular acceleration of the engine block, but the parameter actually used was the ‘pseudo angular acceleration’ obtained by taking the difference between the linear

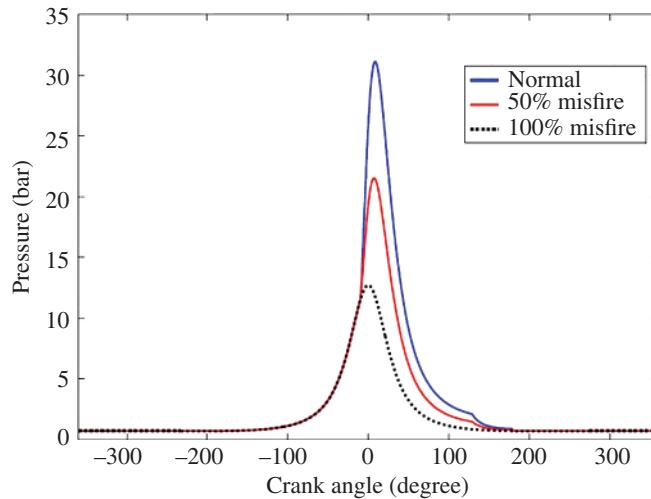


Figure 8.35 Pressure curves for normal and misfire conditions.

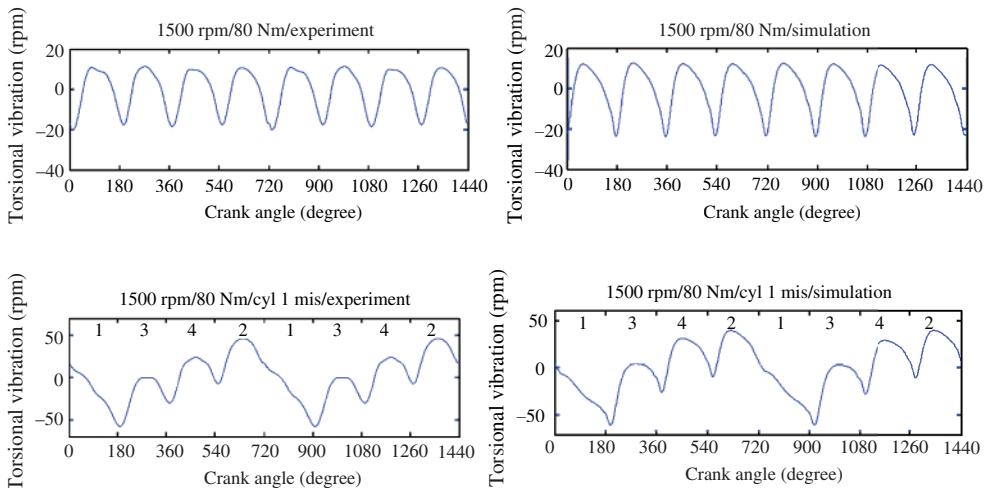


Figure 8.36 Measured and simulated crankshaft angular velocities.

acceleration of accelerometers 7 and 5 in Figure 8.33b. This was dominated by angular acceleration about the longitudinal axis, but was not measured about the centre of gravity, and might also have had a small component about the vertical axis. However, the results shown here and in Section 9.3.4 show that this did not greatly affect the validity of the method.

Figure 8.38 compares pseudo angular accelerations of the block for healthy condition and misfire in Cylinder 1, for one speed and load, for comparison with Figure 8.36. It is seen that the modelling is not quite as good as for torsional vibration, partly because of the greater complexity of the simulation model and the fact that the rigid body modes of the suspension lay within the excitation range of the firing frequency harmonics.

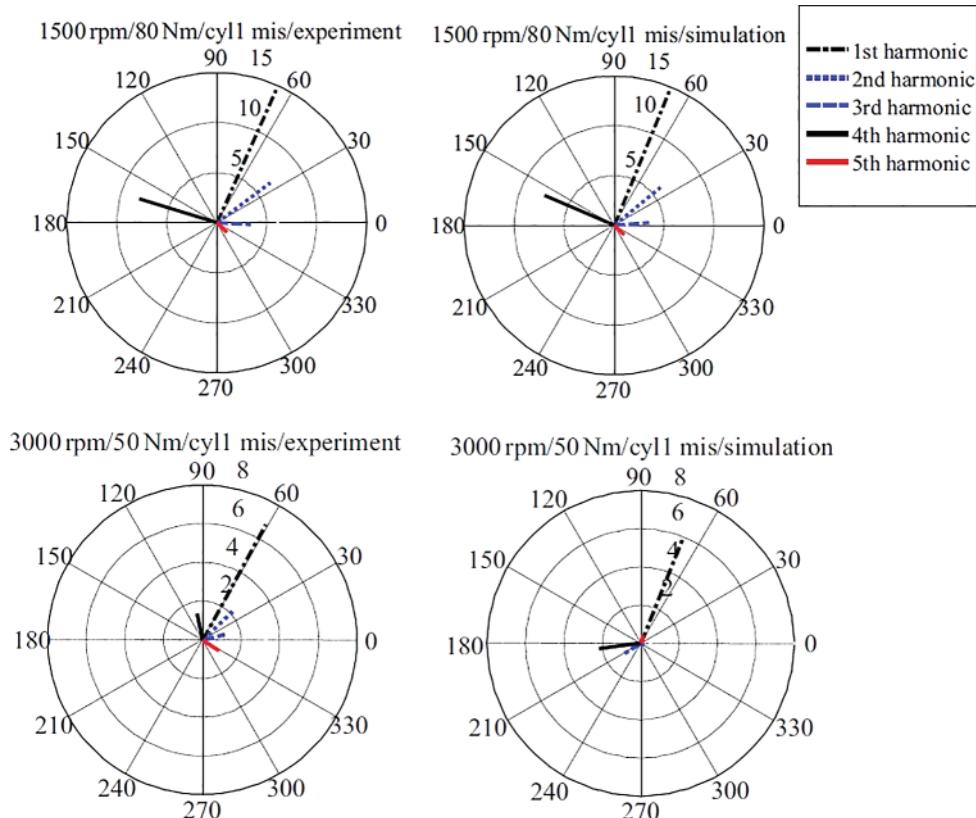


Figure 8.37 Polar diagrams of the experimental and simulated torsional vibration for two torques and misfires in cylinder 1 (for 1500 and 3000 rpm).

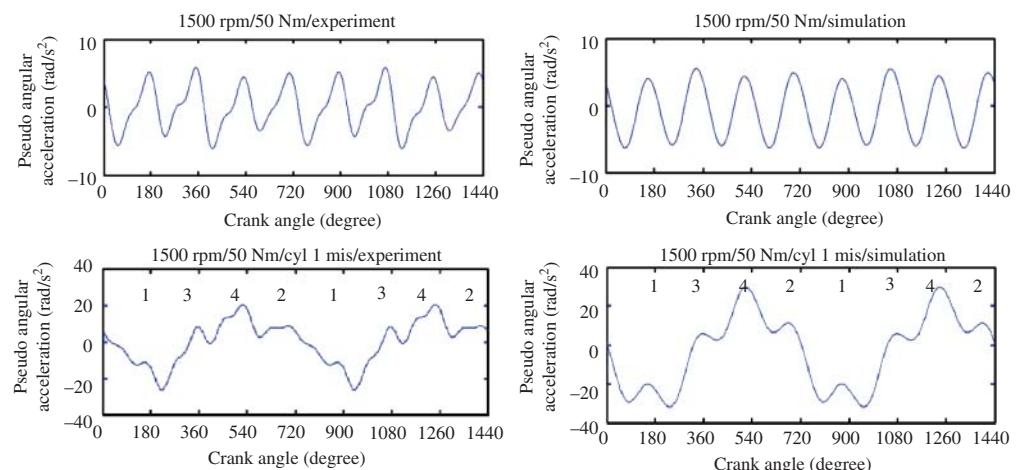


Figure 8.38 Measured and simulated pseudo angular accelerations of the block for healthy condition and misfire in Cylinder 1.

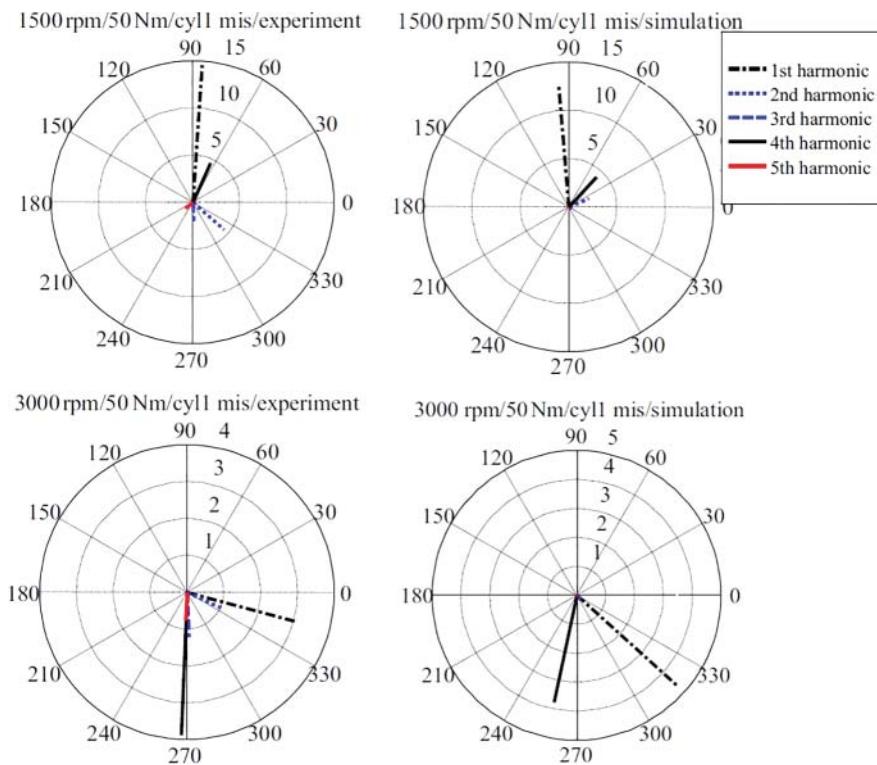


Figure 8.39 Polar diagrams of the experimental and simulated pseudo angular acceleration for misfires in cylinder1 (for 1500 and 3000 rpm).

This is made more obvious by the polar diagrams of Figure 8.39, which can be compared with Figure 8.37. It is seen that even though there is good correspondence between experiment and simulation at each speed, there is now a considerable difference at different speeds. This can be explained by the interaction between the varying forcing frequencies of the engine firing harmonics and the fixed resonance frequencies of the engine suspension modes.

As mentioned above, the same would apply to the torsional vibration responses for engines with a flexible crankshaft (like that in Section 8.4.1.1) or even with the current engine at higher speeds than those tested here (up to 3000 rpm).

As seen in Section 9.3.4, however, there is no great problem in incorporating speed dependence into the automated diagnostics and prognostics.

8.4.2 Piston Slap

As mentioned above, piston slap was one of the mechanical faults studied in Chen's PhD thesis. Various other details are given in references in the thesis, but once again there is a later journal paper, with some updates [27].

Early measurements showed that the dominant response to both piston slap and bearing knock was resonant responses to short impacts, similar to bearing faults, and so a different approach was required compared with combustion faults, where the indicators are additive vibrations rather than being modulations of high frequency carriers. One consequence of this was that simulation models

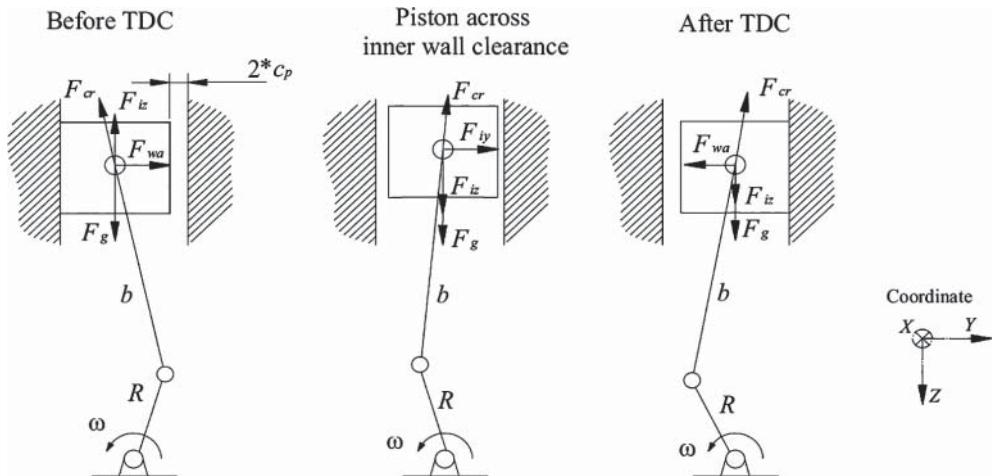


Figure 8.40 Forces causing lateral piston motion after TDC.

did not have to be exactly correct with regard to natural frequencies, as long as their modal density was about right in the excited resonance zones, so as to give correct envelopes. It was found that an indication of the best frequency band to demodulate could be obtained using the fast kurtogram (Section 5.5.3), although since the indications of the latter varied somewhat with speed and load, a broadened band could be chosen encompassing the whole range of operating conditions, while still being insensitive to other excitations. Even though the frequency bands for piston slap and bearing knock were not very different, these two faults could be differentiated largely because of their local responses, near the top of the engine block for piston slap, and near the main bearings of the crankshaft for bearing knock (i.e. accelerometer positions 6, 7, 8 and 4, 5, respectively, in Figure 8.33b).

The mechanism of piston slap is illustrated in Figure 8.40 for the most important case, around top dead centre (TDC), firing stroke. Just before TDC, during the exhaust stroke, the cylinder pressure acting on the piston gives a moment about the crankpin forcing the piston to be in contact with the left cylinder wall. Just after TDC, the moment changes direction, causing the piston to move across the clearance, eventually impacting on the right cylinder wall. The impact is not necessarily simultaneous along the whole cylinder wall, as indicated in the diagram, since the piston can pivot around the gudgeon pin, but might strike first at the top or bottom corner. However, this does not greatly affect the total impulse of the impact, which is basically given by the maximum velocity of the centre of gravity of the piston after accelerating across the clearance space; it just modifies the time over which this impulse is transferred to the cylinder, which may not greatly affect the measured response, as discussed below. Note that piston slaps occur after both TDC and bottom dead centre (BDC), and with all three other strokes, though with smaller impacts since the cylinder pressure is lower.

To model the piston slap, a simulation tool existed already; LMS Virtual.Lab (now Siemens PLM Virtual.Lab), this being provided by the sponsor of the project. There was a template available for normal engines, but this could be modified to insert greater clearance between the piston and cylinder wall. The basic kinematics/kinetics of the reciprocating mechanism of the engine was used to model the time waveform of the impact force on the cylinder wall, this being treated as almost rigid, in the sense that the imparted impulse would not be greatly influenced by the elasticity of the engine block.

The impact with the cylinder wall is not instantaneous, as the piston is slowed down as it encounters the oil film just before impact. This, and the small effect of the piston rings, were modelled by the spring/damper mechanisms shown in Figure 8.41, which act only after the clearance to the pad at

the internal ends of the spring has been used up. The other ends of the spring/dampers were assumed fastened at the outer surface of the cylinder liner, so that the springs would still have finite length at the time of impact with the inner surface. At impact with the wall, the mechanism was changed to another with much greater forces than those given by the spring/dampers at the time of their maximum compression. Rather than being treated by Hertzian contact mechanics, as sometimes done, the impact was modelled using a coefficient of restitution method.

The parameters of these mechanisms were updated to give the best match with measurements, but as already mentioned, their effects would primarily be to modify the duration of the impulse, and not the total value as determined by the change in momentum from an almost fixed initial velocity.

To model the responses at the external accelerometers, the engine block was modelled by an FE model, which allowed calculation of the FRFs of the transmission paths between the point(s) of impact and the measurement point(s) on the external surface of the engine block. The FE model was updated using indirect measurement of these paths, as illustrated in Figure 8.42a. This used the reciprocity principle to measure the FRFs, by applying forces via a shaker at the external measurement points, while measuring responses at the internal points where impact forces were actually applied. The use of the bearing knock result is discussed in Section 8.4.3.

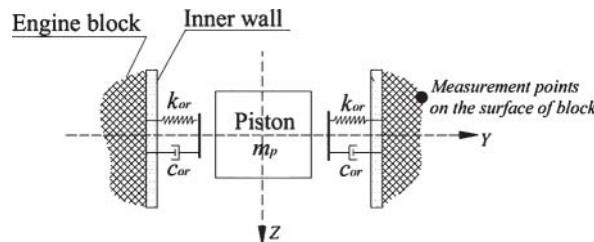


Figure 8.41 Impact model between piston and cylinder wall.

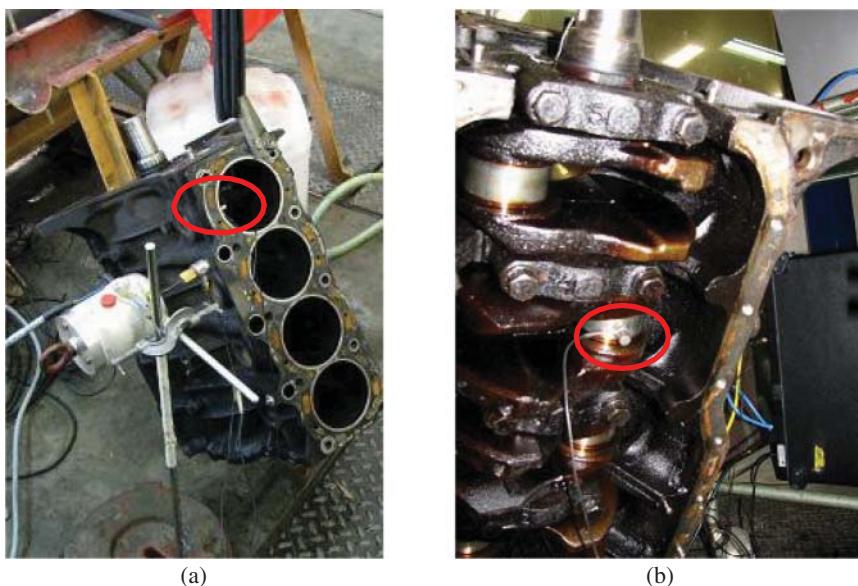


Figure 8.42 Measurement of transfer paths for piston slap (a) and bearing knock (b).

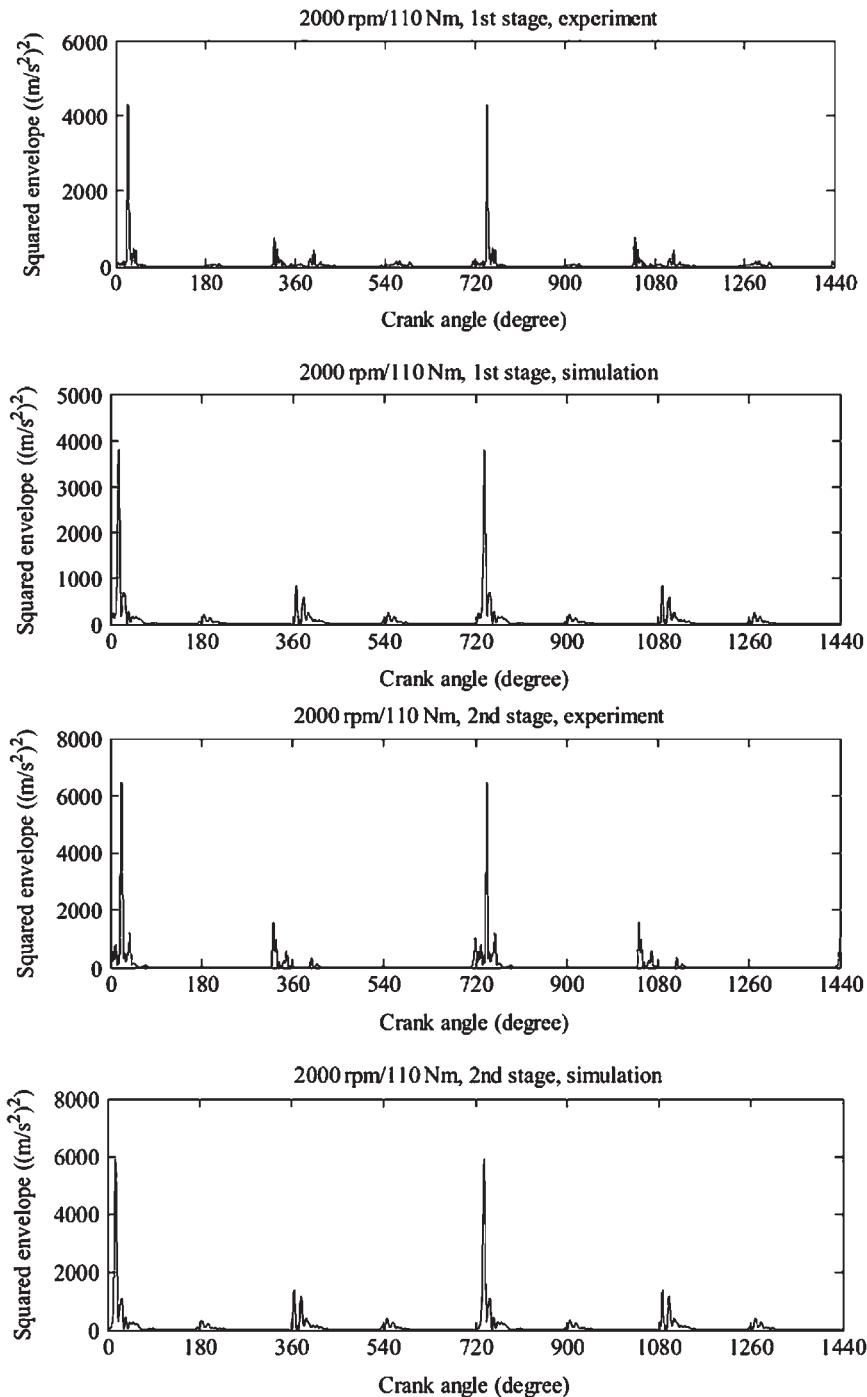


Figure 8.43 Simulated and experimental envelope signals for two levels of excessive clearance (3× and 6×).

Measurements and simulations were made for three different levels of piston/cylinder clearance; (i) at the limit of normal clearance; (ii) $3\times$ normal clearance; (iii) $6\times$ normal clearance. Figure 8.43 compares these for the two levels of excessive clearance, and shows that the models gave very good results. This was confirmed by the automated diagnostics and prognostics described in Section 9.3.4.

For the responses not to be very dependent on the simulated duration of the impact, the effective impulse length should be shorter than a half period of the dominant resonance frequency at the response measurement point, and this was evidently satisfied.

8.4.3 Bearing Knock

In contrast to piston slap, there was no software readily available to simulate this in the engine, so it had to be developed from scratch. In addition to information given in the thesis [24], later journal papers were published with further information, of which Ref. [28] describes the simulation model. It was decided that this could be based on the mechanism for a single cylinder, since at the high response frequencies, there would be little transfer from one cylinder to another, and the timing of occurrences in each cylinder, would be reasonably predictable.

Figure 8.44 shows the modelled piston and connecting rod system for a single cylinder, with only the big end bearing having oversize clearance. The piston/cylinder part of the mechanism was modelled as an ideal slider, and the main bearing had no clearance.

The small circle in (a) at the big end bearing represents the diameter of the crankpin, and the large circle the internal diameter of the big end bearing (exaggerated). In principle, the small diameter circle can be anywhere inside the large one, being constrained only by pressure forces in the lubricating fluid.

As explained in [28], the solution of the problem involved two sets of equations in parallel; the kinematic/kinetic equations for the crank/slider system in Figure 8.44, and the fluid dynamic equations for the lubricant in the big end bearing space, to give the forces F_{y22} , F_{z22} , F_{y31} , and F_{z31} acting on the mechanism. These were formulated using Reynolds equation for short bearings (L/D ratio up to 0.75, which applies in this case) and using the Gumbel π oil film condition, which assumes cavitation in the lubricant to prevent the pressure falling below ambient. As shown in Figure 8.45, at each time step, the updating of the kinematic/kinetic equations, gives the new locations and velocities of the centres of the bearing and crankpin, and these are then processed by the Reynolds equations (dependent only on relative displacements and velocities) to give the updated oil film forces. The equations were solved in Simulink, using the ode45 method.

In this case, the FRFs of the transfer paths from big end bearings to the measurement point(s) were not estimated using an FE model (as for piston slap), but the one required for comparison with the experimental results (big end bearing cyl 2 to acc5) was measured as shown in Figure 8.42b, with the result given in Figure 8.46. This was then used to estimate simulated acceleration responses at acc5, at least up to the maximum frequency (4400 Hz) used for generating the squared envelope spectra used as features for determining the severity of the condition.

The age and condition of the motor used for testing did not allow for measurements to be made at speeds over 3000 rpm, or for clearance $>$ four times the normal maximum, but Figure 8.47 shows that an excellent correspondence was found between simulations and measurements for first stage ($2\times$ normal clearance) and second stage ($4\times$ normal clearance) bearing knock.

Such simulated responses were used to train neural networks to detect, locate and estimate the severity of the bearing knock for a range of situations and the results are given in Section 9.3.4 of Chapter 9.

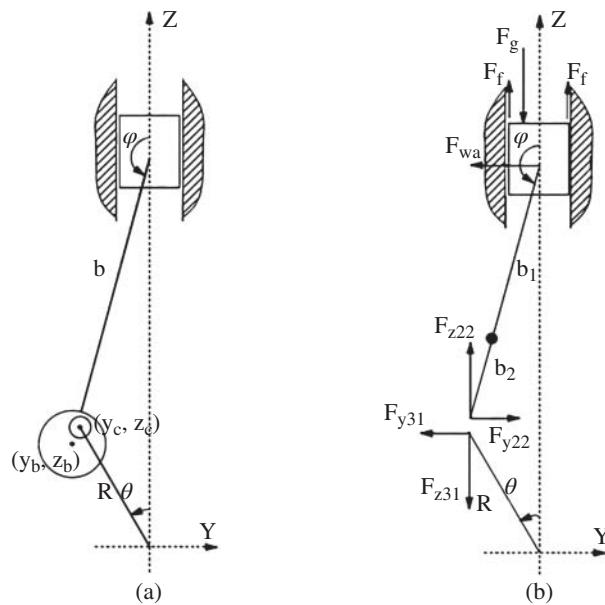


Figure 8.44 Modelled piston connecting rod system (a) oversized big end bearing clearance (b) expanded view showing forces.

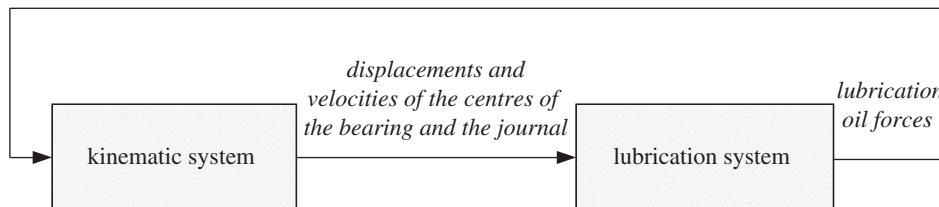


Figure 8.45 Interaction between kinematic/kinetic system and lubrication system.

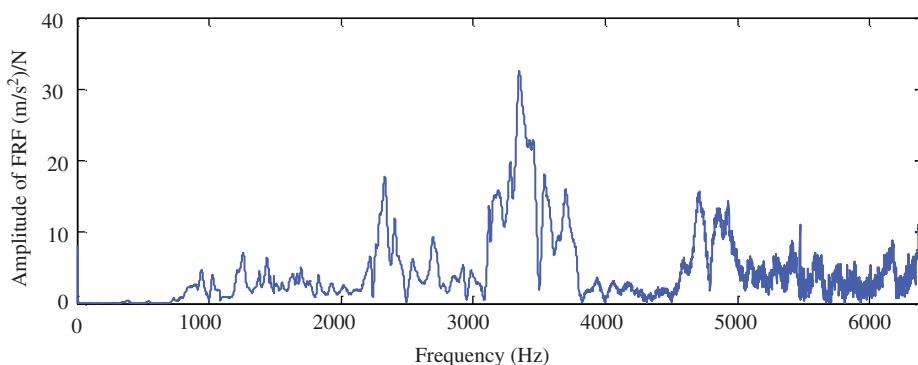


Figure 8.46 Experimental transfer function for bearing knock from cyl2 to acc5.

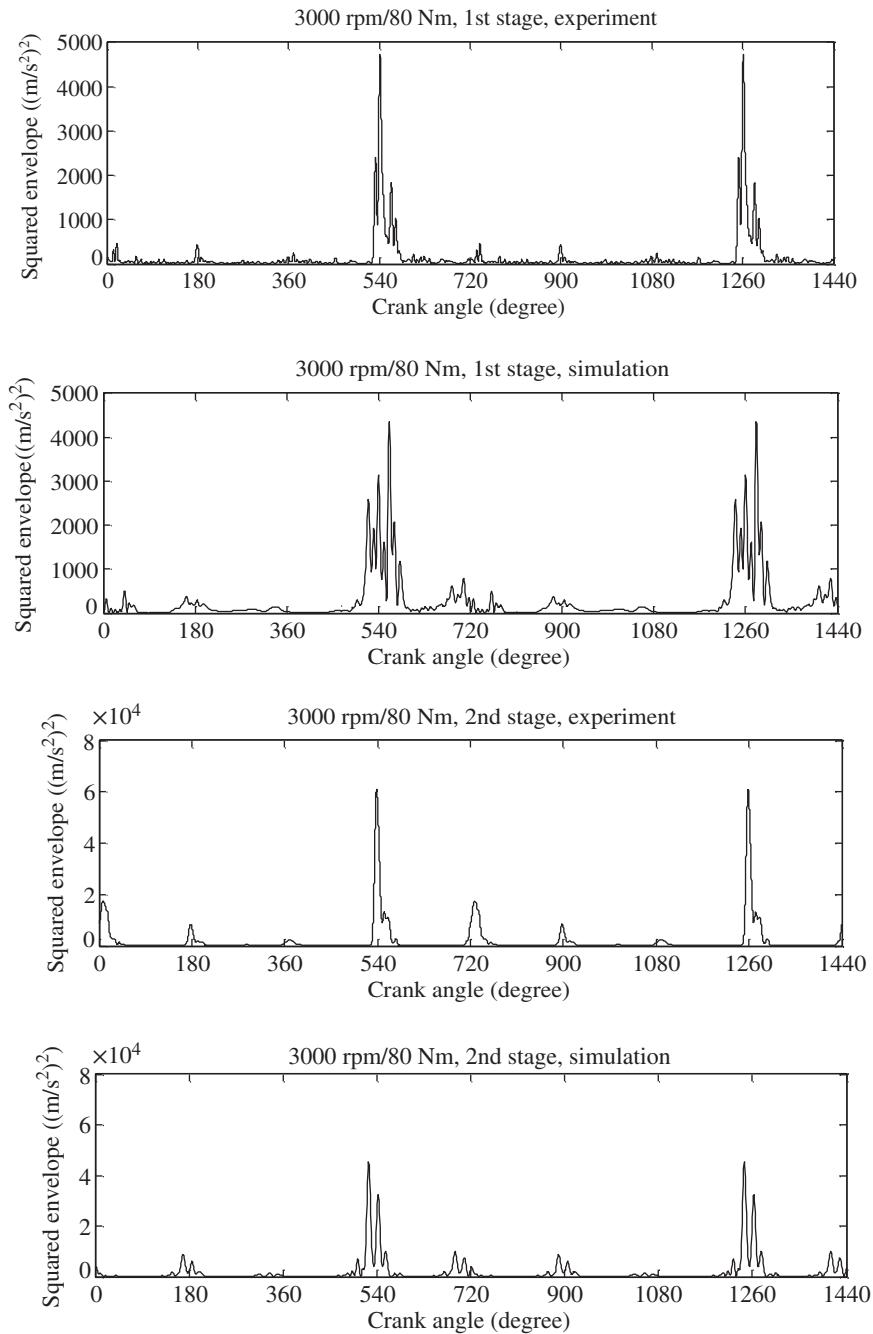


Figure 8.47 Comparison of measured and simulated envelope signals for first and second stage bearing knock, at 3000 rpm and 80 Nm load.

References

1. Özgüven, H.N. and Houser, D.R. (1988). Mathematical models used in gear dynamics—a review. *Journal of Sound and Vibration* 121 (3): 383–411.
2. Endo, H. (2005). A study of gear faults by simulation, and the development of differential diagnostic techniques. PhD Dissertation, UNSW, Sydney.
3. Ewins, D.J. (2009). *Modal Testing: Theory, Practice and Application*, 2e. Wiley.
4. Endo, H., Randall, R.B., and Gosselin, C. (2004). Differential diagnosis of spalls vs. cracks in the gear tooth fillet region. *Journal of Failure Analysis and Prevention* 4 (5): 57–65.
5. Endo, H., Randall, R.B., and Gosselin, C. (2009). Differential diagnosis of spall vs. cracks in the gear tooth fillet region: experimental validation. *Mechanical Systems and Signal Processing* 23 (3): 636–651.
6. Chaari, F., Fakhfakh, T., and Haddar, M. (2009). Analytical modelling of spur gear tooth crack and influence on gearmesh stiffness. *European Journal of Mechanics A/Solids* 28: 461–468.
7. Mohammed, O.D., Rantatalo, M., and Aidanpää, J.-O. (2013). Improving mesh stiffness calculation of cracked gears for the purpose of vibration-based fault analysis. *Engineering Failure Analysis* 34: 235–251.
8. Kahraman, A. (1994). Natural modes of planetary gear trains. *Journal of Sound and Vibration* 173 (1): 125–130.
9. Parker, R.G. (2009). A physical explanation for the effectiveness of planet phasing to suppress planetary gear vibration. *Journal of Sound and Vibration* 236 (4): 561–573.
10. Inalpolat, M. and Kahraman, A. (2009). A theoretical and experimental investigation of modulation sidebands of planetary gear sets. *Journal of Sound and Vibration* 323: 677–696.
11. Randall, R.B., Peng, D., and Smith, W.A. (2019). Using measured transmission error for diagnostics of gears. *SIRM Conference*, Copenhagen (February 2019).
12. Peng, D., Smith, W.A., Borghesani, P. et al. (2019). Comprehensive planet gear diagnostics: use of transmission error and mesh phasing to distinguish localised fault types and identify faulty gears. *Mechanical Systems and Signal Processing* 127: 531–550.
13. Sawalhi, N. and Randall, R.B. (2008). Simulating gear and bearing interactions in the presence of faults: part I. The combined gear bearing dynamic model and the simulation of localised bearing faults. *Mechanical Systems and Signal Processing* 22: 1924–1951.
14. Sawalhi, N. (2007). Rolling element bearings: diagnostics, prognostics and fault simulations. PhD Dissertation, University of New South Wales. <http://handle.unsw.edu.au/1959.4/40544> (accessed 02 February 2022).
15. Sawalhi, N. and Randall, R.B. (2008). Simulating gear and bearing interactions in the presence of faults: part II: simulation of the vibrations produced by extended bearing faults. *Mechanical Systems and Signal Processing*, 22: 1952–1966.
16. Sawalhi, N., Deshpande, L.G., and Randall, R.B. (2011). Improved simulations of faults in gearboxes for diagnostic and prognostic purposes using a reduced finite element model of the casing. HUMS 2011, Melbourne, DSTO.
17. Deshpande, L.G., Sawalhi, N., and Randall, R.B. (2011). Gearbox fault simulation using finite element model reduction technique. *Acoustics 2011*, Australian Acoustical Society, Gold Coast (2–4 November 2011).
18. Deshpande, L.G., Sawalhi, N., and Randall, R.B. (2012). Gearbox bearing fault simulation using a finite element model reduction technique. Comadem Conference, Huddersfield, UK, Comaden International, Birmingham, UK.
19. Deshpande, L.G., Sawalhi, N., and Randall, R.B. (2013). Application of finite element model updating and reduction techniques to simulate gearbox bearing faults. *Australian Journal of Multi-Disciplinary Engineering* 10 (2).
20. Deshpande, L.G. (2014). Simulation of vibrations caused by faults in bearings and gears. PhD Dissertation, University of New South Wales.
21. Lyon, R.H. (1987). *Machinery Noise and Diagnostics*. Butterworth-Heinemann.
22. Desbazeille, M., Randall, R.B., Guillet, F. et al. (2010). Model-based diagnosis of large diesel engines based on angular speed variations of the crankshaft. *Mechanical Systems and Signal Processing* 24 (5): 1529–1541.
23. Ghojel, J.I. (2010). Review of the development and applications of the Wiebe function: a tribute to the contribution of Ivan Wiebe to engine research. *International Journal of Engine Research* 11 (4): 297–312.
24. Chen, J. (2013). Internal combustion engine diagnostics using vibration simulation. PhD Dissertation, University of New South Wales.
25. Chen, J. and Randall, R.B. (2015). Improved automated diagnosis of misfire in internal combustion engines based on simulation models. *Mechanical Systems and Signal Processing* 64–65: 58–83.
26. Ball, J.K., Bowe, M.J., Stone, C.R., and McFadden, P.D. (2000). Torque estimation and misfire detection using block angular acceleration. *SAE Technical Paper no.2000-01-0560*, Society of Automotive Engineers, Warrendale, PA, USA.
27. Chen, J., Randall, R.B., and Peeters, B. (2016). Advanced diagnostic system for piston slap faults in IC engines, based on the non-stationary characteristics of the vibration signals. *Mechanical Systems and Signal Processing* 75: 434–454.
28. Chen, J., Randall, R.B., Feng, N. et al. (2014). Modelling and diagnosis of big-end bearing knock fault in internal combustion engines. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 228 (16): 2973–2984.

9

Fault Trending and Prognostics

9.1 Introduction

As discussed in Chapter 1, perhaps the major benefit from condition monitoring comes from being able not only to detect and diagnose machine faults, but from being able to make reliable predictions as to how much longer a piece of equipment can operate safely, reliably, or economically. End of life is usually defined as when the system no longer meets its design specifications. Determination of ‘remaining useful life’ (RUL) is the basis of prognostics, but it must be acknowledged that the latter is the least developed of the three phases of condition-based maintenance, viz. detection, diagnostics, and prognostics.

Excellent surveys of machine prognostics are given in [1] and [2], and even though these are now quite old, it must be said that there have been few major breakthroughs since that time. In [1] it is stated that condition-based prediction can basically be divided into two main categories, ‘physics-based’ and ‘data-driven’. The first requires a physical or mathematical model of the mode(s) of failure, such as rate of crack growth, and then uses measurements to give an indication of the extent to which a particular failure mode has progressed. The second is based on deriving failure models from statistical processing of measurement data from a number of historical cases, or in the simplest situation by trending features indicative of faults and attempting to extrapolate the trends into the future.

This chapter first examines simple data trending, and then more complex ways of performing prognostics.

9.2 Trend Analysis

Many failures of simple machines do proceed in a reasonably predictable way, and trend analysis of fault features has been used successfully over many years. In Chapter 4, it was discussed how information about fault detection could be extracted by spectrum comparison of machine vibration signals, in particular if they were expressed as CPB (constant percentage bandwidth) spectra on a logarithmic frequency scale, and with the vibration parameter (velocity or acceleration) expressed on a logarithmic or dB scale. In one sense, this is the simplest type of physics-based model, where experience gained over the years, and incorporated in the successive standards, VDI 2056, ISO 2372, ISO 10816, and finally ISO 20816 (Section 4.2.1) has shown that a change in vibration level of 6–8 dB

is significant and a change of 20 dB is serious. Even though the standards were originally intended to apply to overall RMS velocity values in the frequency range 10–1000 Hz, it was suggested in [3] that the criteria for changes in severity could be extended to spectrum changes measured in that way, at least as an initial guide. Because of the relationship of vibration velocity to dynamic stress, it was reasonable to assume that the latter in many cases could change by a factor of ten (20 dB) within the allowance of a design safety factor. Note that it is unlikely that such a factor would apply to total stress, but in most cases the dynamic stress is added to a large static component. It is reasonable to apply such criteria to major spectral components such as low harmonics of shaft speeds, garmesh frequencies, etc., as these are likely to correspond to limiting stresses, but not all spectral changes are equally important. For example, modulation sidebands can change quite a bit more, without really corresponding to increased stress; quite often phase relationships between concurrent amplitude and frequency modulation, which can change with relatively small load variations, mean that sidebands can reinforce on one side of a carrier frequency and partially cancel on the other. Using CPB spectra guards against this to some extent, since the bandwidth would often be sufficient to span over both left and right sidebands, and thus average out the differences, in contrast to FFT spectra where they would normally be separated.

Another situation where changes greater than 20 dB can sometimes be allowed is with changes at high frequency in CPB spectra from faults in rolling element bearings. The reason is that when they are in pristine condition, they generate virtually no vibration, and it is only when a small fault develops that they do produce something measurable, so the first 10 or even 20 dB change may not be very critical. This is very machine and measurement point dependent, since the dB change depends on the initial noise or background spectrum level, and not that produced by the bearing itself. It is recommended that 20 dB change should be used conservatively as an initial guide, but the tolerance could be increased in a given situation based on experience. With bearings in particular, it would be a good idea to use other criteria of fault severity, such as spectral kurtosis, as an additional severity indicator (see Section 9.2.2).

9.2.1 *Trending of Simple Parameters*

The case illustrated in Figure 4.7 of Chapter 4 is a good example of trending of single CPB spectrum components. As mentioned there, the fault was a developing misalignment, found later to be due to failure of the grouting of the gearbox on its foundation, which caused a progressive increase in the level at the output shaft speed (121 Hz). A trend of the spectrum band containing this component is shown in Figure 9.1 for the three months over which the deterioration occurred. The increase at the time of the last measurement is already greater than 20 dB, but it was judged that a limit of 30 dB would be appropriate in this case. A linear regression line (of the dB values) was applied to the last three measurements, and predicts a three week lead time. The repair was in fact carried out within that time.

The question arises as to what type of curve should be fitted to data such as this. The linear curve just applied corresponds to a uniform rate of change in ‘severity’ as inferred in Chapter 4 from the fact that equal changes in the vibration criteria correspond to equal changes on a logarithmic amplitude (or dB) scale. In this case the prediction appears conservative, as the rate seems to be decreasing at the last measurement. There are cases however, where the development of a fault can have a feedback effect, and increase the rate of deterioration, even in terms of dB. Such a case is described in Ref. [4], where it is pointed out that increasing wear of gearteeth increases the dynamic load on the teeth, thus increasing the rate of wear. In such cases, it might be more valid to fit an exponential curve to the data.

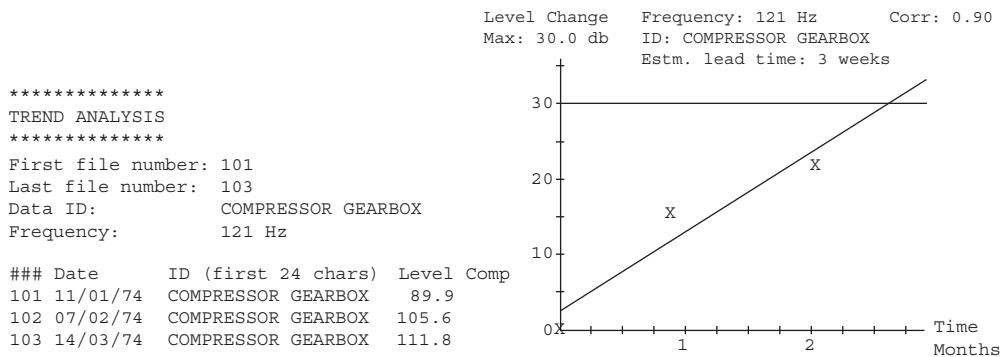


Figure 9.1 A trend of the data from Figure 4.7 of Chapter 4. Source: courtesy Brüel & Kjær.

It is possible to fit other curves to the data, e.g. polynomials, but unless there is a physical reason why the symptoms should evolve in this way, or experience has been gained with polynomials for a particular application, the author recommends that polynomial curve-fitting should be avoided. It is very good for interpolation, but somewhat unpredictable for extrapolation. The choice between linear or exponential trending is perhaps best made on the basis of the correlation coefficient of the fitted curve, but should take previous experience into account.

The trending procedure described in this section, initially proposed by the author, has been incorporated in monitoring systems by the Danish company Brüel & Kjær for many years, and has proved to be quite robust. A number of Application Notes have been published by them, giving details of successful applications e.g. [5–9]. Figure 9.2 shows examples of trending single frequency components (the shaft speed of a Roots blower on an engine) at two different engine speeds [5]. In both cases the trend (in dB) is very linear as shown by the correlation coefficient. The vibration in that case was an indicator of increasing wear in the splines at the end of the drive shaft.

A better trend can sometimes be achieved by combining a number of frequency lines, in particular at higher frequencies, where individual harmonics are no longer isolated in a single CPB band. Figure 9.3 (from [6]), shows the trending of a band (selected manually) that was directly affected by the growth in a bearing fault. The correlation coefficient is 0.854. An attempt was also made to fit an exponential curve to this trend, but the correlation was lower (0.812).

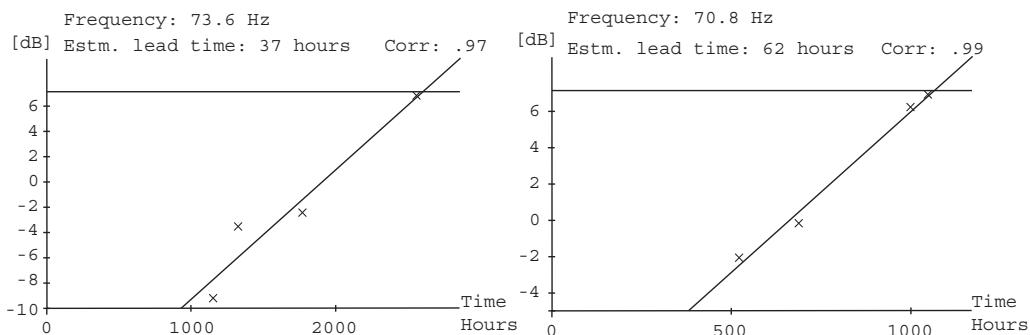


Figure 9.2 Trend curves for two different speeds of a blower drive shaft. Source: courtesy Brüel & Kjær.

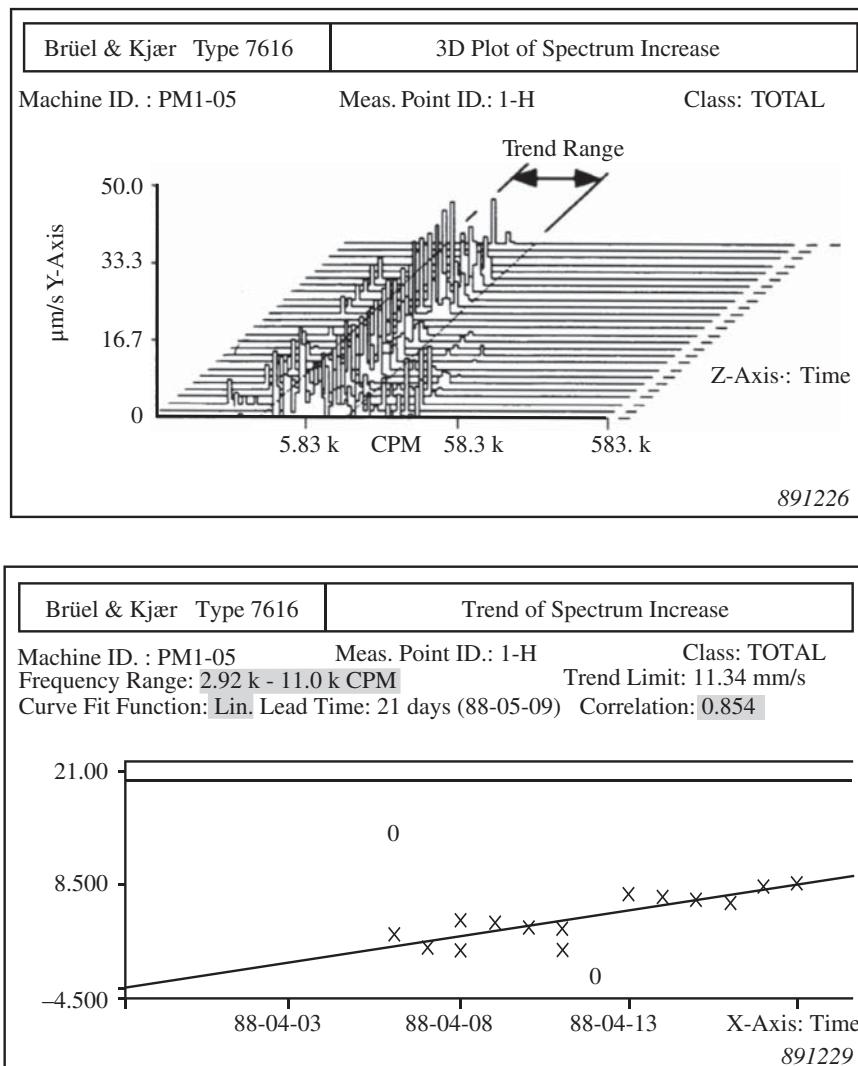


Figure 9.3 Trending of a frequency band indicating a bearing fault. Source: courtesy Brüel & Kjær.

Grouping frequency bands does not necessarily improve the correlation coefficient, in particular when the trend curve is nonlinear, or even non-monotonic. Figure 9.4 shows an example from [10] where a bearing fault in an auxiliary gearbox mounted on a gas turbine driven oil pump gave rise to many harmonics of the ballpass frequency. However, as mentioned above, individual harmonics and sidebands can fluctuate widely, so the trend of two individual components in Figure 9.4a is different and with a large random spread. However, by combining a number of the harmonics in the range 2640–4248 Hz (in between harmonics of garmesh frequencies) a much smoother trend is obtained in Figure 9.4b. Note however, that the correlation coefficient of 0.58 is still low, because the trend is nonlinear, and even decreases towards the end. This is not untypical of bearing faults, for reasons that are discussed below in connection with other possible trend parameters such as spectral kurtosis

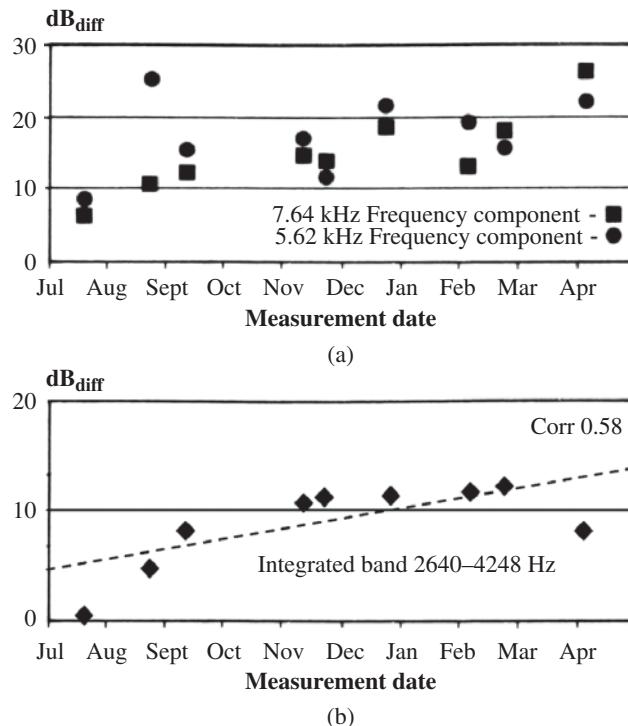


Figure 9.4 Trending of frequency components from a bearing fault. (a) two individual harmonics (b) a band including several harmonics. Source: courtesy Brüel & Kjær.

(Section 9.2.2). However, this example is used here mainly to show how integrating over a band can make the trend more evident, even if not linear.

As a matter of interest, another parameter found to be useful in this case is the cepstrum component corresponding to the bearing fault frequency 206 Hz. Figure 9.5 shows a series of spectra and the corresponding cepstra for the data of Figure 9.4. The growth of the bearing harmonics can be seen by eye in the spectra, and the trend curve of Figure 9.4b shows that there was an increase from July through November 1981, and then the trend levelled out until March of the following year. The machine was closely monitored throughout this period, and in April it was convenient to shut it down for repair. The detailed measurements shown were made by a roving team (who monitored all the pump stations) at roughly one month intervals, but measurements had to be made for whatever load condition applied at the time. In Figure 9.5, the first and third measurements were made at relatively heavy load, while the second and fourth were made at relatively light load. This will be seen to have quite a significant effect on the garmesh components just over 4 kHz, but in fact less on the bearing components, because the torque load had a smaller influence on the radial bearing load. The spectra of Figure 9.5 confirm that the individual bearing harmonics vary widely, as shown in the trend curves of Figure 9.4a, but that the band trended in Figure 9.4b is dominated by a number of them. The curves on the right in Figure 9.5 represent the cepstra (Section 6.2) corresponding to the adjacent spectra. The first harmonic corresponding to BPFO gives a measure of the average strength of the harmonic family with this spacing as it protrudes from the spectrum noise level. Figure 9.6 shows the trend of this parameter, and it is seen to be quite smooth, even though nonlinear.

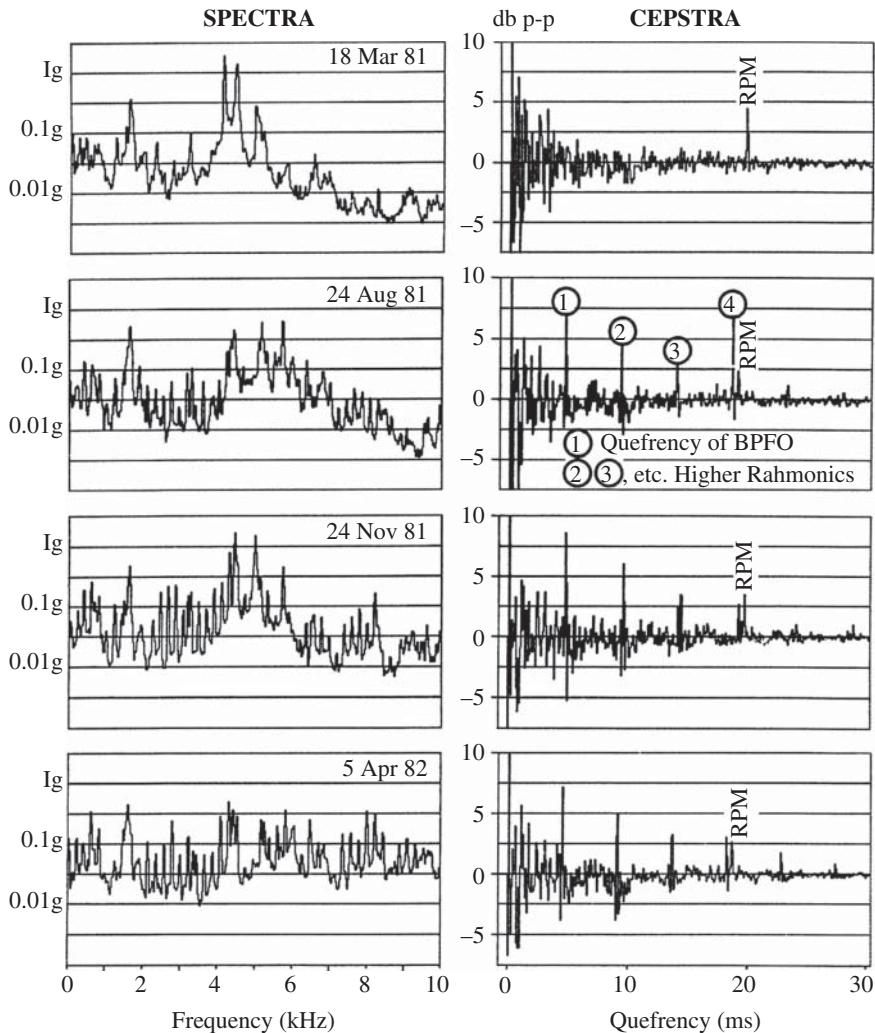


Figure 9.5 Spectra and cepstra for measurements on an auxiliary gearbox [10].

It is in fact very similar to Figure 9.4b, also representing an average over several harmonics. The advantage of the cepstrum is that it collects all harmonics with the same spacing, regardless of other components in the spectrum; for Figure 9.4b it was necessary to find a band not contaminated by other components.

However, in order to use the cepstrum as a trend parameter, the harmonics or sidebands must be separated, as they are in this case. As discussed at length in Sections 2.2.3 and 7.3, harmonics of bearing fault frequencies are often too weak in the low frequency region, where they are separated, and smeared in the high frequency regions where they are amplified by resonances. This is the basis for the dominance of envelope analysis in bearing diagnostics. For the case in point here, envelope analysis performed at least as well, if not better, in making the fault diagnosis. Other parameters for trending bearing faults are discussed in Section 9.2.2.

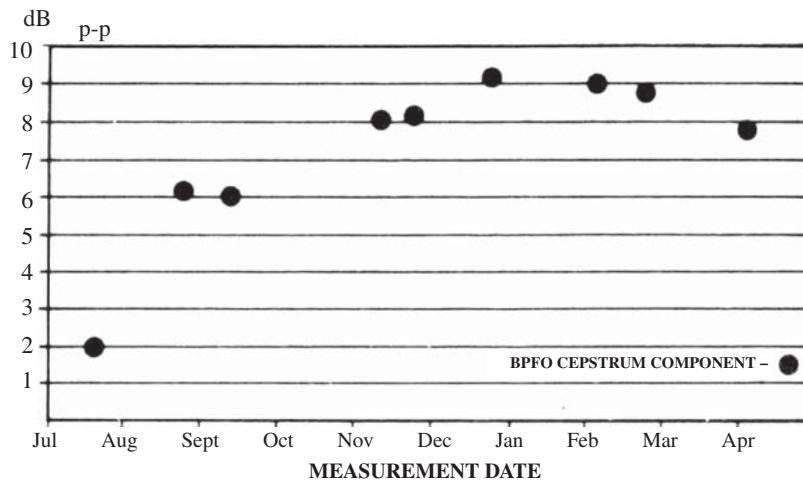


Figure 9.6 Trend of cepstrum values for the case of Figures 9.4 and 9.5. Source: courtesy Brüel & Kjær.

9.2.2 *Trending of ‘Impulsiveness’*

Localised faults in bearings and gears are characterised, at least initially, by generating impulsive components in the response signals, typically impulse responses from a series of sharp impacts with a spall on the surface of a bearing component or gear tooth. Thus, measures of the impulsiveness of a signal have been used over many years as indicators of bearing and gear fault severity.

The simplest measure of impulsiveness is the ‘crest factor’, i.e. the ratio of peak to RMS value of a signal. This does suffer from the problem that the peak value is not very stable, and depends greatly on the section of signal analysed. As discussed in Section 3.1, random signals can in principle have peak values up to infinity (with a probability tending to zero), although for a Gaussian random signal there is a 99.7% probability that the peak value will be less than three times the standard deviation (which is also the RMS value for typical acceleration signals with a mean value of zero).

A more stable measure of impulsiveness is kurtosis, the normalised fourth centred moment of the signal as defined in Section 3.1.2 (sometimes based on cumulants, so that a constant is subtracted from that value). It tends to be more stable than the crest factor, as it is averaged over a signal containing a number of impulsive responses, not just the one defining the peak value. Kurtosis was recommended in the 1970s by Stewart [11, 12] as a fault indicator in gears and bearings, and it forms the basis of some of his ‘figures of merit’, still used in gear diagnostics (Section 7.2.1). A number of other authors have since used kurtosis as one of the components of feature vectors to detect gear and bearing faults using artificial neural networks (ANN), e.g. [13, 14]. However, the signals were typically analysed in raw form, without extraction of that part dominated by the fault. In Section 5.3, a number of techniques are described for separating the signals from faulty gears and bearings from each other and from background masking signals. In the example described in Section 7.2.5 for failure of a gear at the low speed input section of a wind turbine [15], it was found that a high kurtosis value could only be obtained (during the period when the fault was developing) by using spectral kurtosis (SK), as described in Section 5.5, where the most impulsive frequency band is extracted by optimum filtering.

In two of the examples used in Section 7.3.2 to illustrate a semi-automated bearing diagnostic procedure, the spectral kurtosis was shown to correspond well with the degree of degradation of the

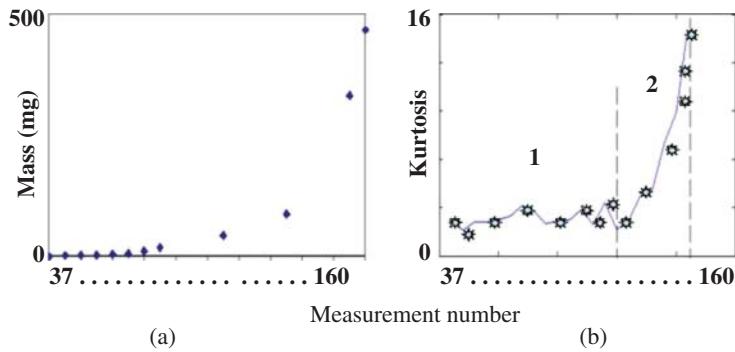


Figure 9.7 (a) Accumulated metal wear debris. (b) Kurtosis of the filtered signal.

bearing. For the helicopter gearbox of Case History 1 (Section 7.3.2.1) Figure 9.7 shows that the kurtosis of the optimally filtered signal follows the same trend as the cumulated oil wear debris, and thus has potential for making a prognosis of RUL. Note that the wear debris gives no indication of the source, whereas the SK corresponds to one of the planet bearings.

In this case, MED (Section 5.4) was tried but gave no benefit, because the failure was in the low speed part of the gearbox.

For the high speed bearing of Case History 2 (Section 7.3.2.2) Figure 9.8 shows how the SK corresponds well with fault size, but only after MED was used to separate the individual impulse responses, because of their close spacing.

In Case History 1, the increase in kurtosis is monotonic, and even accelerates towards the end, but for Case History 2, the development flattens out towards the end and even dips at one point. Both

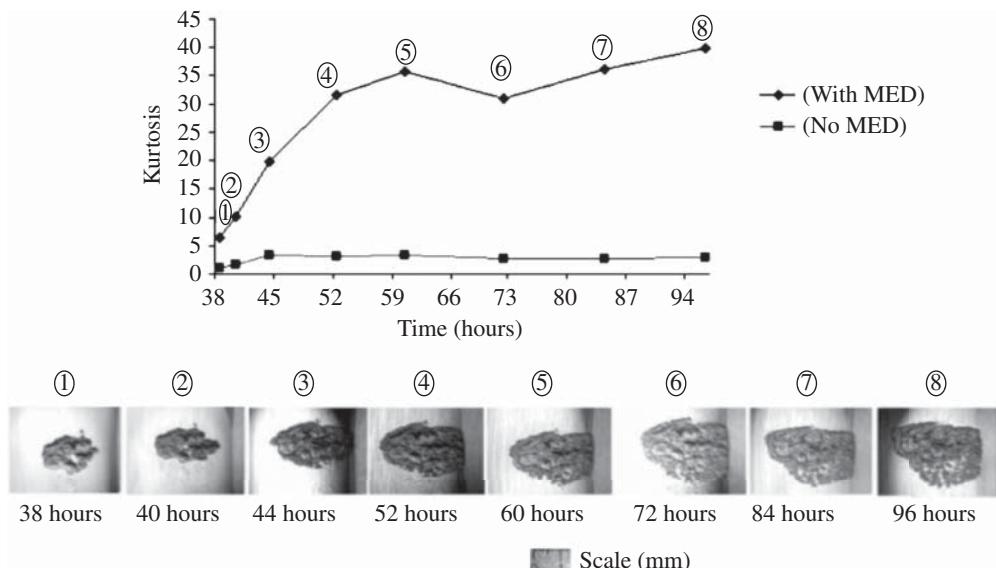


Figure 9.8 The development of kurtosis with and without MED for a high speed bearing, compared against fault size.

cases are for testing on a test rig at close to full rated load, and in that situation it is not uncommon for a single spall to proceed to failure. However, for bearings in practice, it is unlikely that they will be operated continuously at full load, as the latter would quite often occur for only a relatively small part of the time. For military aircraft gas turbine engines, for example, the maximum load, for which the bearing has to be designed, would only occur under certain manoeuvres, for a small percentage of the time. In that situation it is common for more than one spall to develop before the bearing can be considered to have failed. With rail vehicle bearings, it is normally considered that they should not be replaced until the total spall area reaches a certain size. Thus, it is quite common for trend parameters of bearing faults to first increase, and later decrease. This applies to measures of impulsiveness, such as crest factor and kurtosis, and can also apply to the components of envelope spectra, as will be explained.

Figure 9.9 illustrates how peak and RMS values change with typical spall development (when the bearing is not permanently subjected to full load). The crest factor might typically start at about 3, when there are no impulsive components in the signal. As a single small spall develops, its peak values start to exceed the peak values of the background signal, but there is too little added energy to affect the RMS value, so the crest factor increases while the RMS value stays constant. As the spall size increases, the crest factor continues to rise without affecting the RMS value. However, since spall depth tends to be fairly constant, and dependent on the depth of maximum shear stress caused by the Hertzian deformation of the bearing surface by the rolling elements [15], there is a limit to the peak values that are generated, so the peak value and crest factor tend to stabilise. When further spalls start to develop, the peak value is not affected, but the total energy in the impulse responses does increase, to the point where it starts to dominate the overall RMS value of the signal, so that the crest factor starts to diminish again. In the end, the signal is completely dominated by impulse responses, but because the spacings between them become smaller and smaller, the signal becomes more and more stationary. In this situation the crest factor can fall to the initial value again. Note that this means that the envelope signal also becomes more uniform (because the individual impulse responses are no longer separated), so the amplitude of components in the envelope spectrum diminishes, and may even disappear.

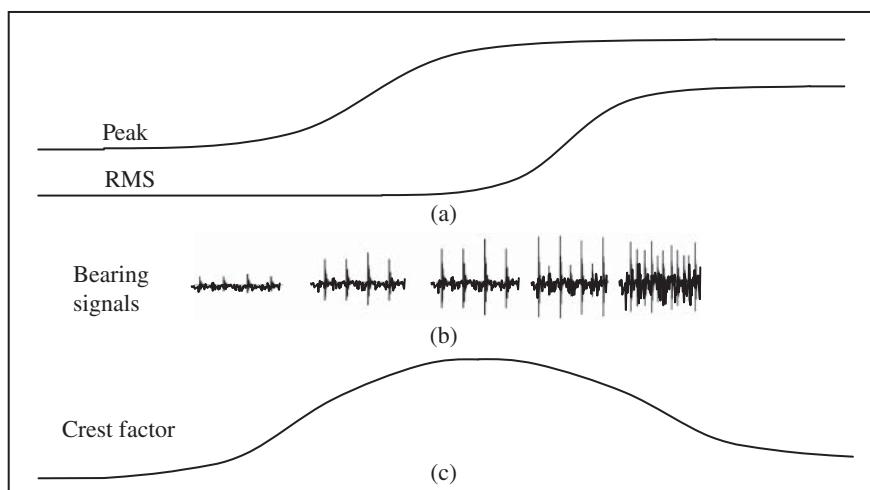


Figure 9.9 Typical trend of crest factor (and kurtosis) with fault development.

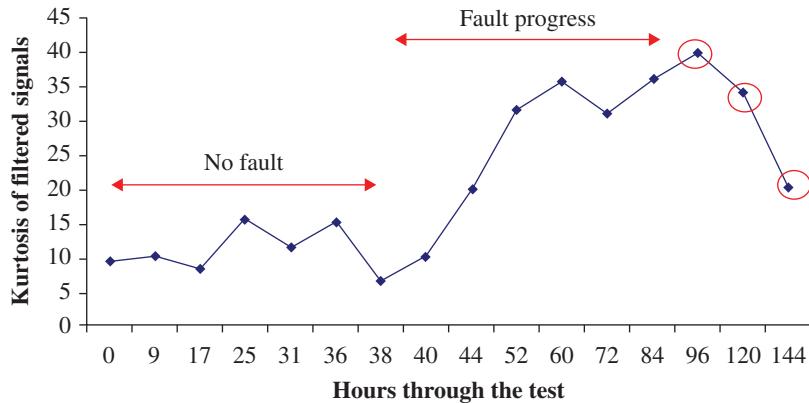


Figure 9.10 Trend of SK for another high speed bearing.

Since kurtosis is a sort of normalised crest factor, it tends to follow a very similar trend, at least in the initial stages of development of a single spall. Experience shows that it can fluctuate very widely towards the end of life, but it is not unusual for it to decrease. The point can perhaps be made that for certain applications, such as commercial aero engine bearings, where safety is of paramount importance, the effective life might be limited to the range where parameters such as kurtosis are still increasing monotonically.

Even where bearings are loaded at full or over capacity, the kurtosis can trend downwards towards the end of life, as illustrated in Figure 9.10 (from the same measurement series as the results of Figure 9.8, and presented in [16]). This is suspected to be due to the fact that the entry and exit edges of the spall would tend to wear. On the other hand, even though the symptoms are decreasing, it is quite likely that subsurface cracking is continuing, so that there might be a further step increase if more metal is lost as the spall suddenly extends. This is also thought to be the case for the failure depicted in Figures 9.4–9.6, where even though the symptoms are trending down, subsurface cracking could still be proceeding.

In other words, trend parameters for bearing faults are unlikely to be increasing linearly, or even monotonically (except RMS values in broad frequency regions dominated by the fault, such as in Figure 9.3) so it is necessary to obtain other indicators of fault progression. The next section describes such an approach to determining spall size in bearings by detecting the entry and exit events and measuring the spacing between them. A method for determining spall size on gear teeth was described in Section 7.2.4, but still needs considerable development.

9.2.3 Trending of Spall Size in Bearings

Ref. [17] is one of the first studies of signal processing methods to detect the separate events corresponding to the entry of a rolling element into a spall, and the exit from it, so as to be able to determine the spall size.

In the literature, two cases had been found which described the separate entry and exit events [18, 19]. Figures 9.11 and 9.12 show typical results from them.

Figure 9.11, from [18], shows measurements from the PhD work of I.K. Epps, under the supervision of H. McCallion. It shows that the entry event has a different character from the exit event. This is presumably because the load-bearing rolling element (RE) must roll over the edge of the spall

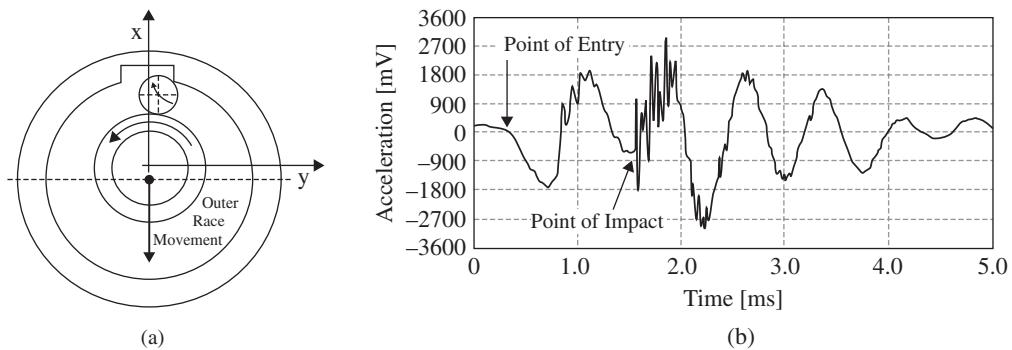


Figure 9.11 (a) Model of rolling element travelling into a fault. (b) A typical measured response [18].

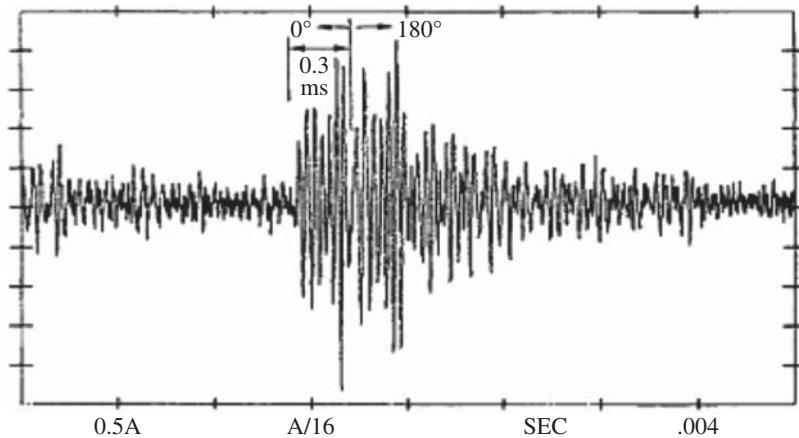


Figure 9.12 Bandpass filtered trace from a helicopter gearbox [19].

in a curved path, and the sudden change of curvature corresponds to a step in acceleration. On the other hand, at exit, the direction of the rolling element must change suddenly, giving a step change in velocity, or impulse in acceleration. The initial step response is obviously dominated by lower frequency resonances than the impulse at exit.

Figure 9.12, from [19], shows how the phase of the signal changes by 180° at exit, compared with that at entry, with the spacing of 0.3 ms corresponding to the fault size. It is interesting that the second event appears to have two peaks, with a separation of the same order as the fault size. The author has experienced similar effects on other machines, which was initially confusing, because it was thought that this spacing indicated the fault size. An example is shown in Figure 9.13, from [17], where it was initially thought that the double pulses revealed by MED filtering resulted from the entry and exit events. For that particular measurement the spacing did correspond approximately to the fault size. It was only when the same rig was run at a number of different speeds later, that it was realised that the spacing was constant in time, not distance.

When the signals depicted in Figure 9.13a were examined in detail, it was found that they also had a step/impulse response character, as shown in Figure 9.14, and this separation also corresponded to

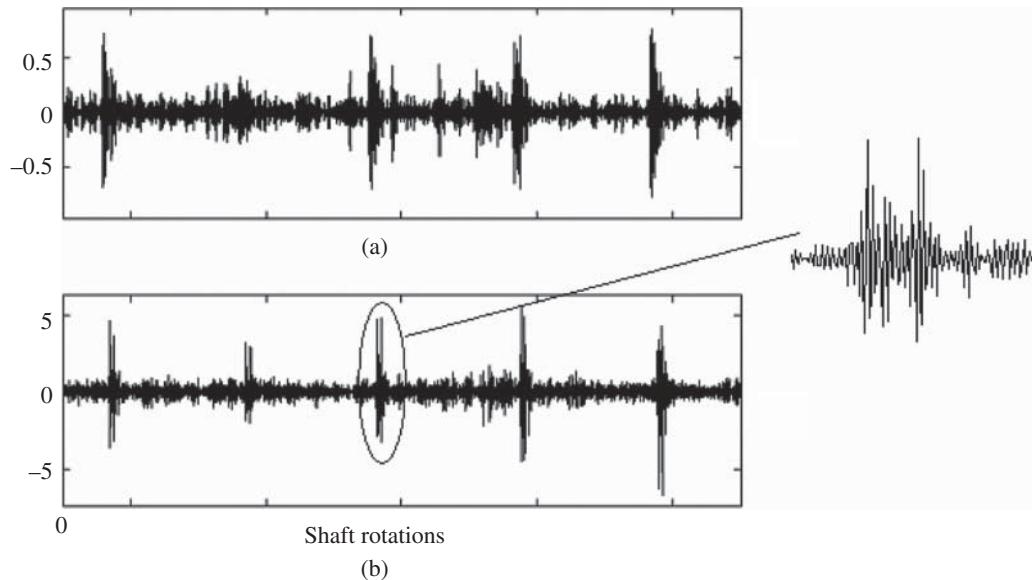


Figure 9.13 Outer race bearing fault signal (a) before MED filtering (b) after MED filtering.

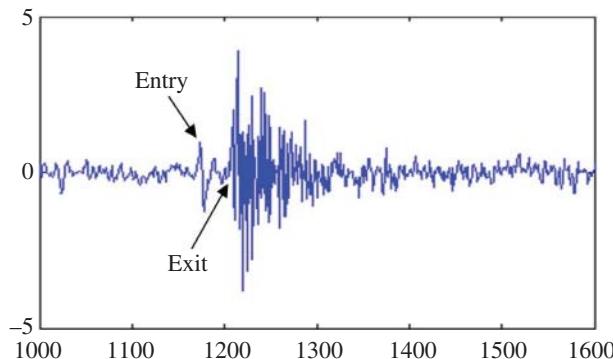


Figure 9.14 A single entry/exit event for a faulty bearing in the same test rig as Figure 9.13.

the fault size. Note that the MED operation (Figure 9.13b) tended to enhance the impulse event at the expense of the step response at entry. It was then surmised that the double pulses found in Figures 9.12 and 9.13, with constant spacing in time, rather than distance, must correspond to constant natural frequencies. It could be explained, for example by a beat between two natural frequencies, spaced by an amount corresponding to the reciprocal of the delay time between them (beat period). These could be due to the stiffness nonlinearity of the bearing, possibly affected by the direction of the loading.

As described in [17], a series of tests was then carried out on another test rig, for which it was easier to vary the speed over a wide range. Simulated spalls of two different sizes (approx. 0.6 and 1.2 mm) were introduced by EDM (electro-discharge machining) in the outer and inner races of separate

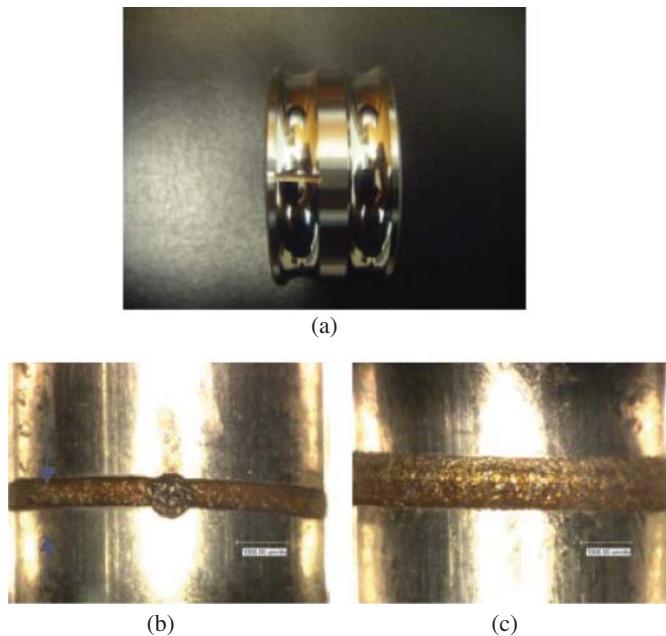


Figure 9.15 Simulated spalls in the inner race (a) Location in one race (b) Small spall (0.6 mm) (c) Large spall (1.1 mm).

bearings, and measurements were made at speeds of 900, 1200, 1800, and 2400 rpm. Figure 9.15 shows the small and large inner race faults in one race of the double row bearing.

Figure 9.16 shows the measured signals for the two inner race fault sizes depicted in Figure 9.15 at one speed. The step response at entry is less marked than in Figure 9.14 for the other rig, but it is seen that the separation from the impulse response is approximately doubled for the large fault size.

It was desired to enhance these events so as to be able to measure the spacing between them, and so initially the following procedure was carried out:

1. Use AR prewhitening to balance the energy at low and high frequency (Section 5.3.2).
2. Apply wavelet filtering to better balance the frequency content of the two events. Morlet wavelets of octave bandwidth were used (Section 3.5.3).
3. Obtain the squared envelope of the filtered prewhitened signal (Sections 3.3.2, 7.3).
4. Apply MED filtering to narrow the envelope peaks (Section 5.4).
5. Use the cepstrum to measure the average spacing between the entry and exit events (Section 6.1.1).

Figure 9.17 shows the result of stages 1 to 3. Prewhitening (Figure 9.17b) is seen to significantly enhance the step response at entry. However, there is still a large difference in frequency content. Octave band Morlet wavelet filtering (Figure 9.17c) gives a better frequency balance, and means that the squared envelope of Figure 9.17d has two reasonably similar pulses representing the entry and exit events. MED was found to be beneficial before applying the cepstrum, as it made the pulses slightly shorter, and thus the (log) spectrum flatter.

The size of the spall was estimated on the basis that the entry would correspond to the point where the RE just started to roll over the leading edge of the spall (step response), and the impact

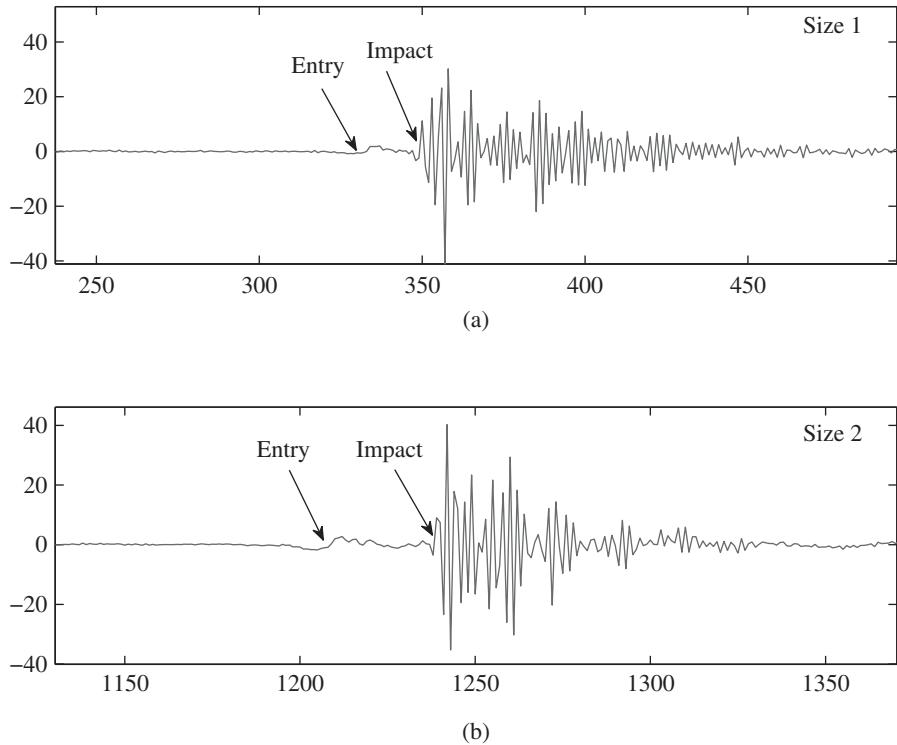


Figure 9.16 Entry/exit events for small and large inner race spalls.

(impulse response) would occur when the RE bridged over the spall, i.e. with its centre halfway through the passage. This assumes that the depth of the spall is sufficient that the RE does not touch bottom, which holds until the length of spall becomes a considerable proportion of the RE diameter. The ‘time to impact (T_i)’ thus corresponds to half the total time of passage (T_{sp}) of the RE centre over the spall length.

Ref. [17] shows that the equation for T_{sp} for an outer race spall of length l_o is given by:

$$T_{sp} = \frac{2l_o D_p}{\pi f_r (D_p^2 - d^2)} \quad (9.1)$$

where f_r is the shaft rotational speed, D_p is the pitch diameter of the bearing, and d is the effective RE diameter. Somewhat surprisingly, the value for an inner race spall is the same, with length l_i substituted for l_o .

In [17] an alternative way of estimating the value of T_i was presented, which involved separating the entry and exit events (based on the fact that the latter was always much larger, and could be detected when the squared envelope of the signal exceeded a threshold). The two events could then be processed separately, giving some advantages, but the end results were comparable.

A problem with the technique was that the size estimate varied with rotational speed of the machine (see Figure 9.18), so various authors proposed other solutions. Significant among these was a group at

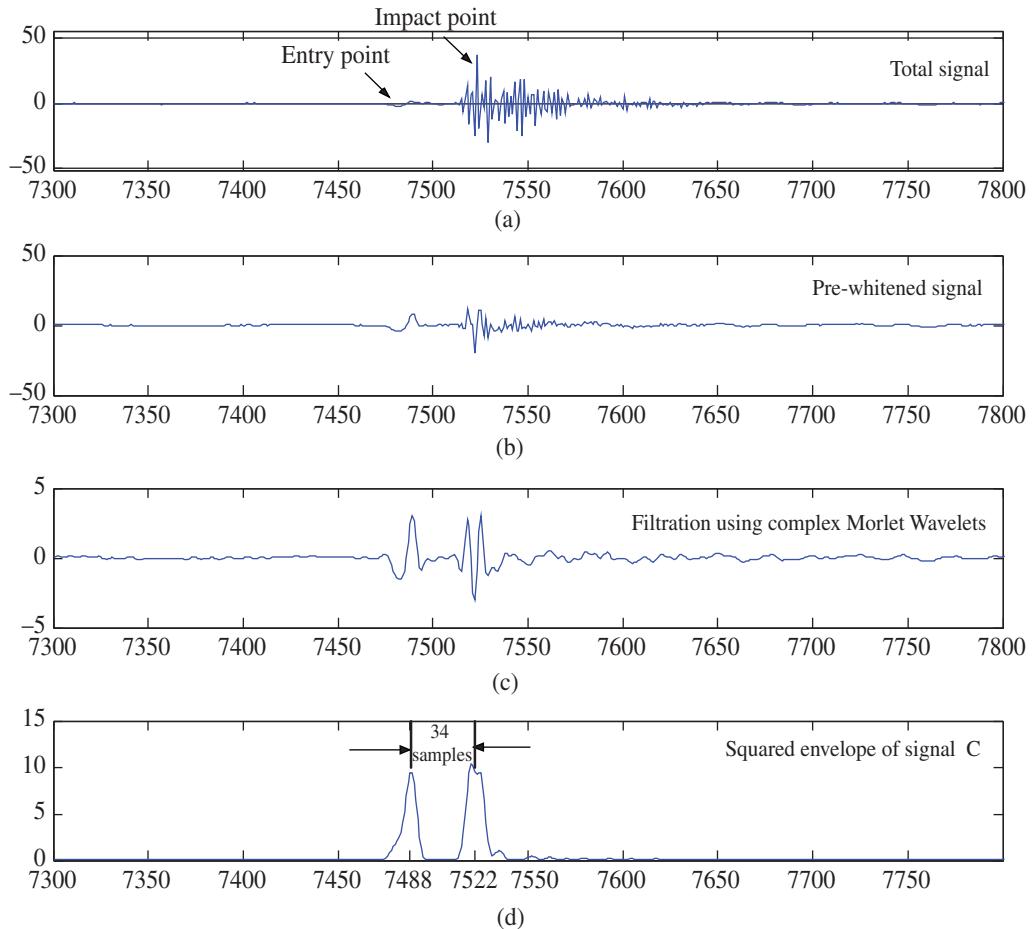


Figure 9.17 Effect of the first three processing stages.

Adelaide University, under the supervision of Prof. Carl Howard, which made a number of advances [20–22], in particular for larger spalls, where significant impacts do not necessarily occur at the midway point of the passage. A major one, described in [22], points out that the low frequency event at entry corresponds to the period during which the RE destresses on entry, and is matched by a symmetrical restressing event at exit. The latter is usually masked by one or more high frequency impact events, but these can be removed by lowpass filtration. Enhancing the low frequency events at entry and exit can give a measure of the total passage time.

Another approach to more reliable detection of the entry event, treating it as a destressing, was proposed in [23]. It was found that this destressing gave rise to a sharp roll-off in the acceleration signal, which could be identified by passing a threshold value, but that the actual entry could best be identified by finding the maximum slope before this trigger point, as indicated by a zero crossing of the derivative of the acceleration signal (evaluated only up until the impact). Using this instead of the method(s) of [17] gave better estimates of the spall size, with smaller influence of speed.

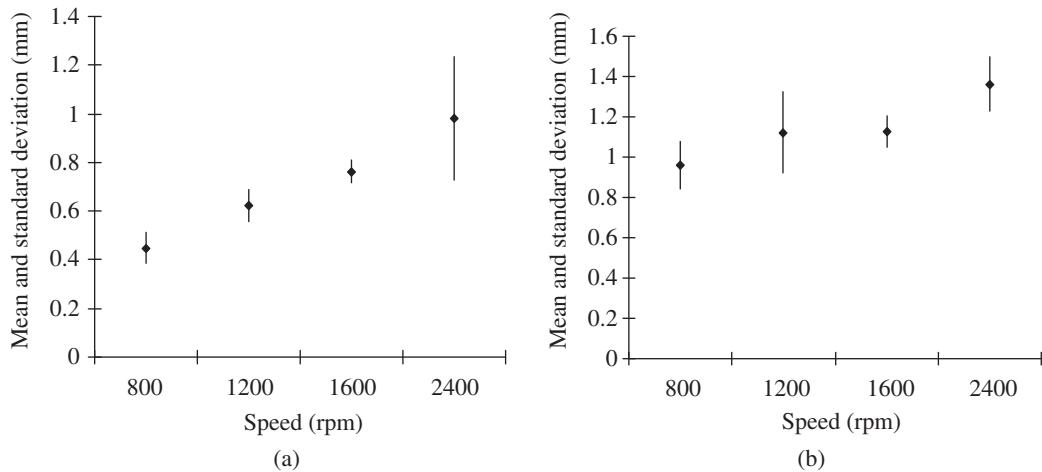


Figure 9.18 Results from alternative procedure (a) small spall (b) large spall.

A recent development [24] proposes an entirely different approach to detection of spall size, based on local changes in natural frequency of the system including the bearing, which appears to have considerable advantages for naturally developed (or extended) spalls. Practically all the prior literature had studied only artificially made ‘spalls’, with sharp entry and exit points, and even in that case the gentle entry events were difficult to detect. Ref. [24] includes a detailed comparison of the earlier literature, including a number of papers not discussed above, but still using the same basic model and demonstrated using artificial ‘spalls’. Note that the approach in [24] is quite different to that published some years ago in [25], which proposed that the change in ‘natural frequency’ of a bearing could be used as a trend parameter to predict RUL. The authors of [25] give no indication of how this natural frequency was measured, nor do they show any measured spectra, but the claim was that the natural frequency changes monotonically with fault severity. Ref. [24] found no evidence of this; only that the natural frequency was perturbed during the passage of the rolling element through the fault.

The approach in [24] uses the Wigner-Ville Spectrum, WVS (Section 3.6.3), to obtain a time-frequency diagram with optimal resolution without interference effects, on the basis that the signal being analysed is second order cyclostationary (CS2). In contrast to the method described in Section 3.6.3 to generate the WVS (using an averaged spectral correlation diagram) the averaging here is over multiple evaluations of the Wigner-Ville Distribution (WVD), thus removing the interference components, which come with random phase.

Since the information being sought is related to rotation angle (of the cage) rather than time as such, the signal must first be order tracked, but the problem arises that the cage motion is not directly tied to shaft speed, because the signal is pseudo-cyclostationary. However, the repetition frequency of impact responses contained in the (squared) envelope of a bandpass filtered signal, centred on a major resonance excited by the bearing fault, will follow small speed changes, even though there will be a small jitter of the actual cage position around the value given by the local mean cage speed, and the averaging of the WVS will slightly reduce amplitude values without bias of the mean cage position values. Note that this only applies where the mean shaft speed is constant, so a different approach will have to be used for variable speed cases. The other reason for this restriction to constant speed is that the order tracking distorts the ‘frequency axis’ of the spectrum (to an order axis), so that the estimated natural frequency, in which the change is being detected, will be slightly smeared.

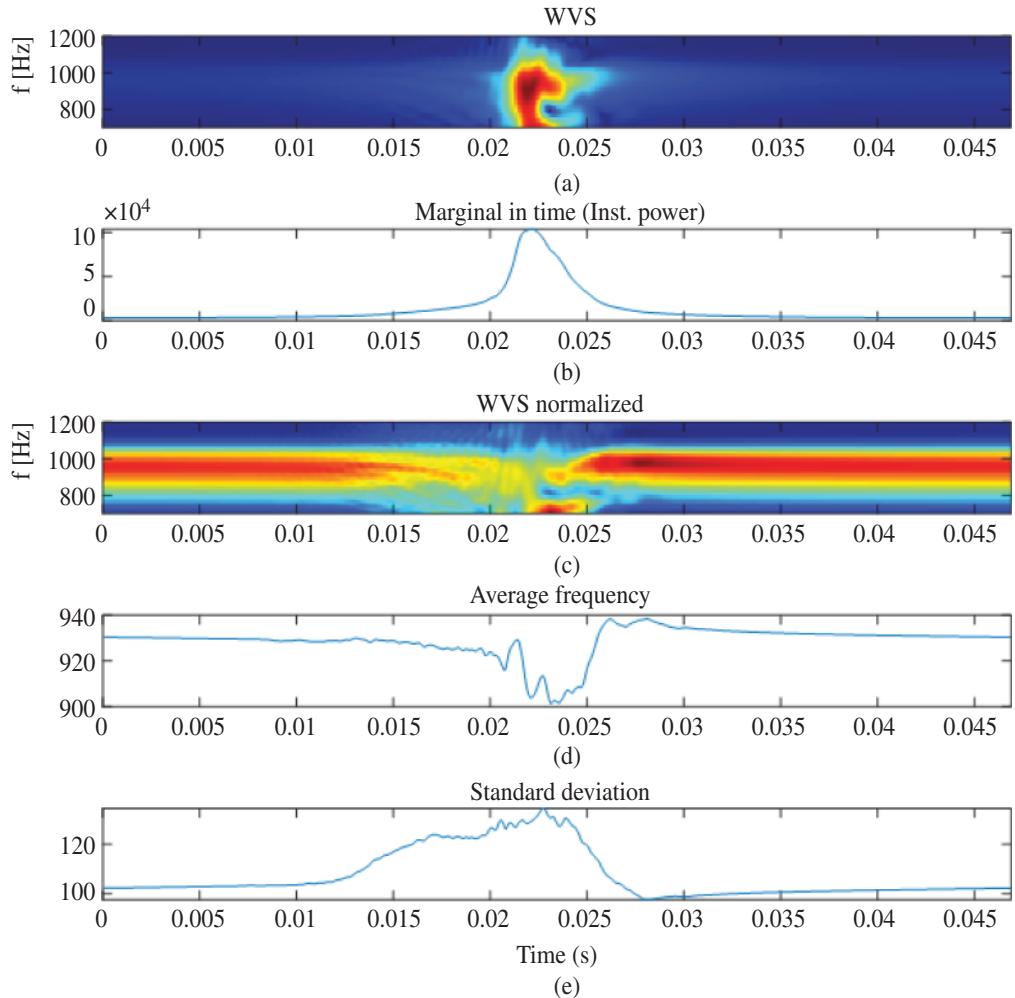


Figure 9.19 The proposed approach applied to a spall of size 1.6 mm (0.0045s in time): (a) WVS averaged over multiple ball pass occurrences; (b) instantaneous power; (c) WVS normalised by instantaneous power; (d) average frequency of the normalised WVS; (e) standard deviation of the normalised WVS. Source: From [24].

Figure 9.19, from [24], shows how the WVS diagram is processed to obtain information about the frequency perturbation which occurs as the RE traverses the ‘spall’ for the particular case of a small artificially produced spall of length 1.6 mm, with sharp edges. Figure 9.19a shows the WVS as produced, but distorts the visualisation of frequency information because of greatly varying amplitude. This is corrected by normalising it with respect to the instantaneous power of the signal (time marginal, or integration over all frequency at each time), in Figure 9.19b. The result in Figure 9.19c gives a much clearer indication of the frequency perturbation, of which two statistics are shown in Figures 9.19d, e. The first is the ‘average frequency,’ or first moment in the frequency direction, while the second is the ‘standard deviation’ or square root of the second moment. The better localisation of the average frequency deviation in this case gives a better indication of the actual fault size of 0.0045s.

Figure 9.20, from [24], compares the frequency perturbation method (average frequency), called ‘WVS method’, with three of the methods reviewed above, called Sawalhi’s method [17], Smith method [23], and Moazen’s method [22]. The WVS method requires the choice of a threshold to give the ‘correct’ result, as described in [24], but this result compares favourably with the three other methods, which give underestimates, in particular the Moazen method, for this particular set of data.

However, when these methods were applied to naturally developed spalls, albeit from an initially inserted fault, the WVS method was found to be superior, as it did not require there to be sharp changes at entry and exit as for the other methods. This is shown in Figure 9.21, where the spall was extended from an initial sharp rectangular notch of length 0.4 mm. In this case, the WVS method is the only one giving an indication of the true final spall size.

Ref. [24] gives further examples of this advantage, including a case where the naturally extended spall arose from a small local dimple in the race. This method of generating realistic spalls is the same as that used by the bearing manufacturer FAG (now Schaeffler) as for Case 2 in Section 7.3.2.2, and Figure 9.8 of this chapter. Ref. [24] also discusses the choice of the resonance to be demodulated, and shows an example where two different bands gave comparable results, even at different stages in the fault development, where the bearing was dismantled sufficiently for the spall size to be directly measured.

The example of tracking spall size in bearings illustrates how information can be extracted from measured signals, to give better indications of fault progression for insertion into physics-based models of failure (Section 9.3.1).

9.3 Advanced Prognostics

This is first discussed in terms of the basic division into physics-based and data-driven failure models, and then hybrid approaches, including where the prognostic process is made a combination of reliability (whole population) estimates, and condition-based estimates, with the emphasis changing from the former to the latter during the life of an individual component [1].

9.3.1 Physics-Based Models

Most of the work reported in [1] in this area is for situations where the degradation is a result of crack growth, such as the development of spalls in rolling element bearings and gears. A number of these are based on the Paris law of crack growth [26] which relates the rate of growth of a crack to the current crack size. Since it is not normally possible to measure crack size in operating machines, different papers propose different methods to relate the size of a fault to measurable vibration parameters. It would appear that incorporation of parameters such as spectral kurtosis, as discussed in Section 9.2.2, and estimated spall size, as discussed in Section 9.2.3 should improve the correlation, but this has yet to be fully demonstrated.

Ref. [27] recommends using data fusion to combine the information from vibration parameters and, for example, wear debris analysis to reduce the uncertainty of prognostic models. Ref. [27] gives physics-based models for both bearing and gear fault propagation, and relates them to vibration parameters.

Specific cases of physics-based models are described in Section 9.3.4 on Simulation-based prognostics.

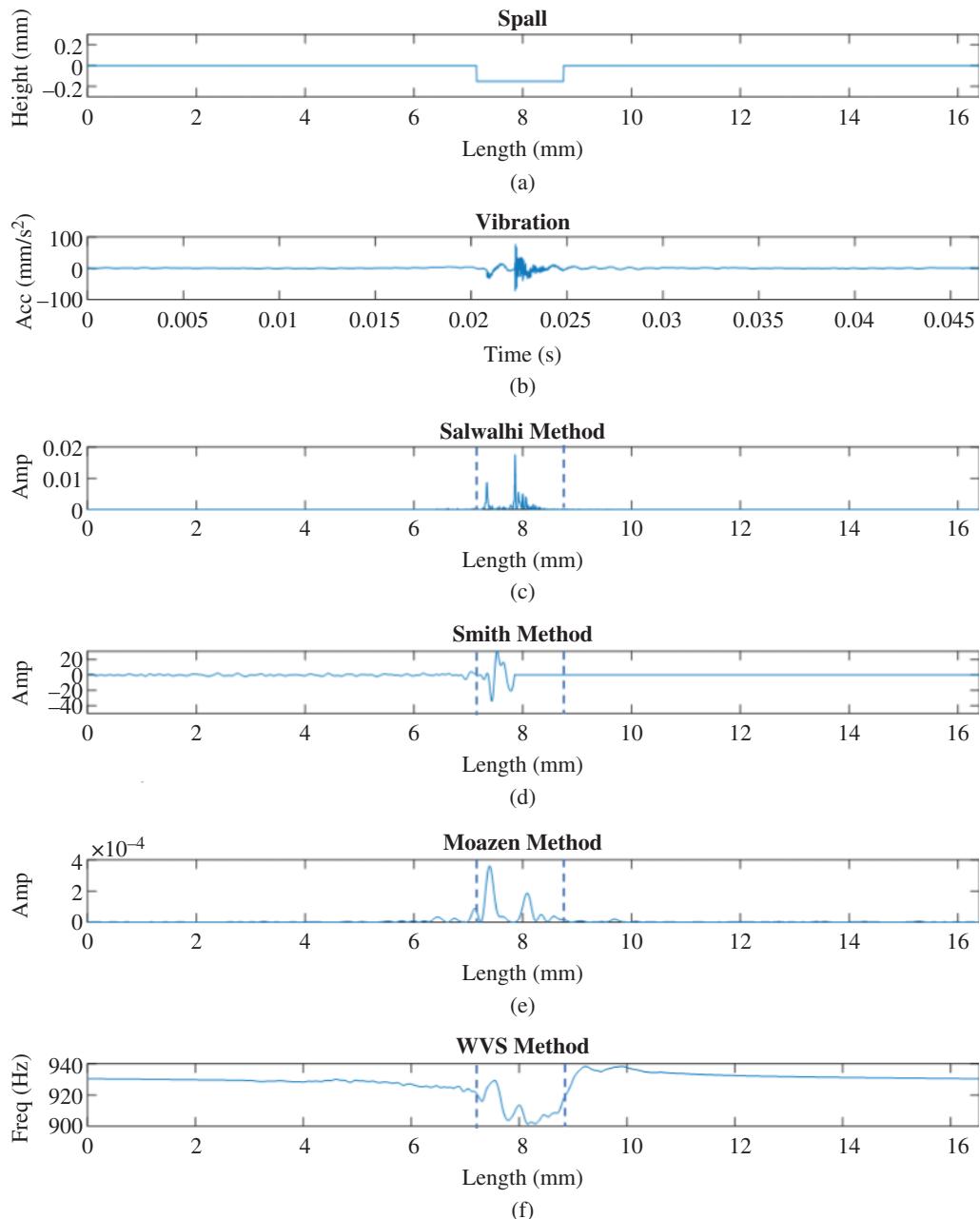


Figure 9.20 Comparison of spall size estimation methods. (a) The spall (aligned approximately according to the impact event in the centre) corresponding to the dotted lines in the remaining figures; (b) The measured vibration signal in time domain; (c) Sawalhi's method to reveal the entry and impact points; (d) Smith's method (gradient); (e) Moazen's method to reveal the entry and exit points; (f) The natural frequency variation method by using WVS. Source: From [24].

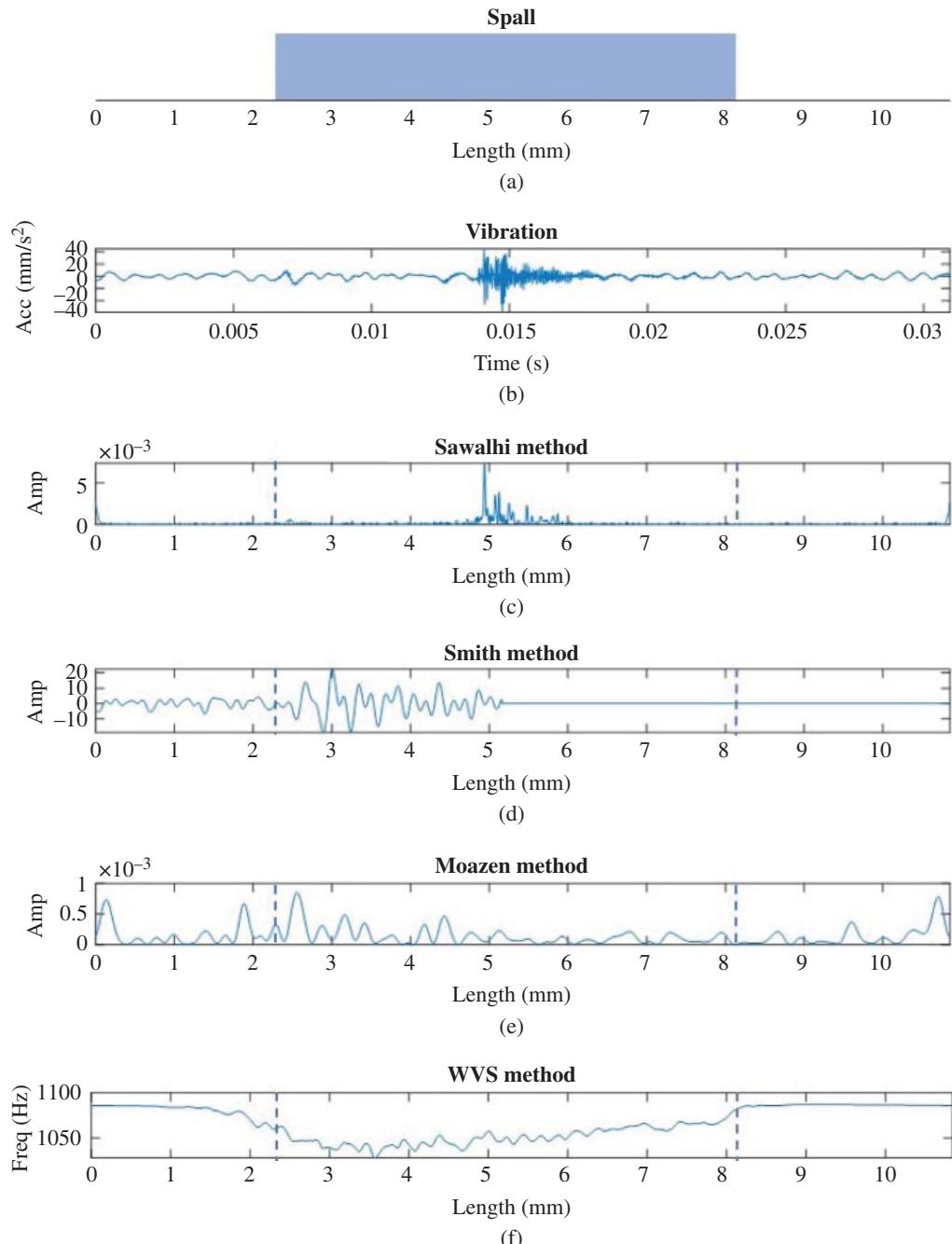


Figure 9.21 Comparison of spall size estimation methods for extended spall (5.92 mm). (a) The size of extended spall; (b) The raw vibration signal, (c) Sawalhi's method to reveal the entry and impact points, (d) Smith's method (gradient), (e) Moazen's method to reveal the entry and exit points, (f) The proposed approach by using WVS. Source: From [24].

9.3.2 Data-Driven Models

The simplest form of data-driven models are the trending models as discussed in Section 9.2, but these do rely on finding measured parameters which trend monotonically (with some random spread) to failure, and where a limiting value can be specified on the basis of experience.

As discussed in [1, 2] more advanced models use multivariate analysis to combine information from several parameters shown to be correlated with component degradation. Artificial neural networks are trained to recognise the complex, often nonlinear and even non-monotonic, relationships between the measured parameters and degradation state. In general, these require large amounts of data, and can therefore not be used for critical and expensive machines for which failure has to be prevented at all costs.

It is quite common to find papers claiming to perform ‘fault diagnosis’ based on classification of a particular data set into categories such as ‘outer race fault’, ‘inner race fault’ or ‘ball fault’. Most often these use part of a particular time record to train an ANN and then classify the other part of the same signal as being in the same group. Many such papers use the Bearing Data Set generated by CWRU (Case Western Reserve University) [28] for the purpose of providing typical bearing fault data, where a number of artificial faults (conical holes of different diameters) were seeded into motor bearings on a test rig consisting only of the drive motor and a dynamometer to provide four different (torque) loads. Many of the papers confuse these ‘loads’ with loads on the bearings, which never change, as they are due to the weight of the rotor. The motor is a four-pole induction motor, so the only effect of the ‘load’ is to give small changes in the speed, from 1797 rpm at zero load down to 1730 rpm at 3 HP load. Remarkably, the attempt is rarely made to classify the data from these four different speeds (but exactly the same bearing condition) as the same, using the proposed techniques. Since each data record is obtained at constant speed and load, the signals could be expected to be stationary, so it is not surprising that for example the second half of each record would fall into the same class as the first half. In fact, a small number of the records appear to be pure noise, but stationary, so the proposed ‘diagnostic’ systems often classify them as characterising the nominal ‘fault’. In any case, it is clear that the only way this approach can be used to make a diagnosis, is if it is known that the signals do actually correspond to the nominated fault. Even in that case, the classification would only apply to that particular fault on that particular machine, at the specified measurement point, and for the particular operating conditions. The trained ANN could not be used to recognise the same type of fault on any other machine, in particular in the normal situation where the vibration signals are masked by other components in the machine and its signals, such as gears, impellers, turbulence, cavitation etc.

As published in Ref. [29], a true diagnosis was made of the entire data set, using physics-based methods as discussed in Section 7.3. It was found that in a high proportion of cases, the signals did correspond to the nominated fault, the diagnosis often only requiring envelope analysis of the raw signal, without any preprocessing. However, a substantial percentage of those cases revealed that the nominated fault was accompanied by evidence of mechanical looseness, which appeared to increase as the tests proceeded (it is not made clear how the different size faults were introduced into the bearings, which were of a type requiring disassembly/reassembly). One such indicator was that the envelope spectra for outer race signals showed modulation at shaft speed (usually expected only for inner race faults), with time signals showing somewhat randomly occurring impulses, but with spacings often corresponding to a shaft revolution. Another indicator was that the outer race frequency was detected even when the fault was placed opposite the nominal load zone, and could only come under load by the shaft bouncing erratically in the clearance space. This diagnosis could be made from analysis of the signals, but a purely data driven method would have ascribed the symptoms of looseness to just the bearing fault.

It should be made clear that data sets such as the CWRU data, are very valuable for testing and comparing different diagnostic algorithms (as was done in Ref. [29]), and possibly the validity of simulation algorithms, as long as the properties of the signals are known, and the criticism here is only of the papers which attempt to use the data for an invalid purpose.

A purely data driven technique could only be used where the state of a fault is known from disassembly and inspection at several stages through the life of a machine, and thus in practice only for test machines in a laboratory. In principle this could be done for a fleet of similar machines, for example helicopters or wind turbines. Partly because of the lack of agreement on what should be measured in HUMS systems for aircraft, this has not yet been achieved for helicopter gearboxes, and even though many successes have been experienced, some serious undetected failures have also occurred. Many manufacturers of wind turbine monitoring systems have had considerable success with hybrid systems, as discussed below, but details are rarely published in full for commercial reasons.

9.3.2.1 Problems with ‘Big Data’

Recent developments with statistical processing of ‘big data’ have been heralded by some as the answer to problems with machine health monitoring in the near future. A lot of data is currently being collected by machine monitoring systems around the world, and much of it is not analysed, but such ‘big data’ is not available for the complete life of critical and expensive machines, partly because they cannot be stopped, dismantled and inspected at multiple stages throughout the fault development, so as to relate ‘health indicators’ (HIs) with the actual current condition, and partly because it is not economically viable to allow such machines to proceed to failure in sufficient quantities to give adequate data for the ‘big data’ techniques. A typical example is given by aircraft engines, where even though there are tens of millions of flights per year, the number of fatal accidents is counted in tens, and the proportion of those caused by mechanical failure much smaller again. When it is considered that each engine contains hundreds of components which might cause catastrophic failure, it is obvious that statistical amounts of data on each possible failure type cannot be gathered.

However, big data is available for many machines in healthy, or near healthy condition, and that could be very valuable for separating the effects of actual condition, and operating conditions, which can have a big influence on vibration responses.

It is possible that this problem may be solved in the future by using simulation models to generate data from simulated faults in different machines and components. The type, location, and severity of the fault can then be varied as desired, rather than waiting for faults to happen. Examples and references are given in Section 2.2.1.3 of how various faults in rotors, such as unbalance, misalignment, cracks, and rub can be modelled. The example of torsional response of a large diesel engine crankshaft described in Section 8.4.1.1 showed how it was possible to generate a simulation model of the crankshaft, update it to agree with measurements using a very limited amount of data, and then train a neural network on the basis of simulated responses for various levels of combustion fault in different cylinders. The ANNs trained using this simulated data were 100% successful in recognising actual combustion faults, as well as being able to identify the faulty cylinder and giving a good estimate of the fault severity (though the amount of data was not statistically viable). In cases such as this, where the root cause of the fault is directly modelled, the approach could be said to be physics-based, rather than data driven. However, in some situations it is a moot point whether the simulation approach should be considered as physics-based or data driven. This is for example the case where a physics-based model of the machine is used, but it generates data corresponding to symptoms of the faults (such as spall size) and information has to be extracted from this in the

same way as for real measured data as to the root cause of the failure, such as crack development. Simulation of responses to localised and extended faults in bearings has been achieved quite successfully, with better results for the former [22, 23]. The model is currently being improved to reflect the difference between the effects of entry into and exit from a spall, as discussed in Section 9.2.3 [24].

Some means has to be provided to introduce a realistic amount of variability into the simulated responses, as the simulation models tend to be deterministic. In the case of the diesel engine crankshaft in Section 8.4.1.1, a random variation of the parameters used to train the ANNs was introduced, with standard deviation based on the differences between two measurements on the same engine (in good condition) at two different times. This question is more fully discussed in Section 9.3.4.

9.3.3 Hybrid Models

An excellent approach to prognostics generally is given in [30], from which Figure 9.22 is taken. It is based largely on experience in the nuclear industry, but also from a wide range of other situations.

Type 1 prognostics is based on reliability techniques, such as Weibull analysis, where statistical information about probability of failure is obtained from actual failure histories and laboratory testing. It gives results for the average component operating under average conditions. It applies to the whole population of machines or components of the same type, and thus in general has a large uncertainty for a given unit.

Type 2 prognostics takes account of factors which have an effect on life, so-called ‘environmental stressors’ such as load, speed, number of cycles of a given operational regime, temperature, cleanliness, etc., which can be monitored for an individual component. These would usually apply to all components on a given machine, operating under a specific condition such as load, but give the possibility to reduce the uncertainty of the prediction of RUL for that whole group.

Type 3 prognostics uses measures of performance and condition to improve the estimates of RUL for each component. This is the best approach to use, in particular in the later stages of life when

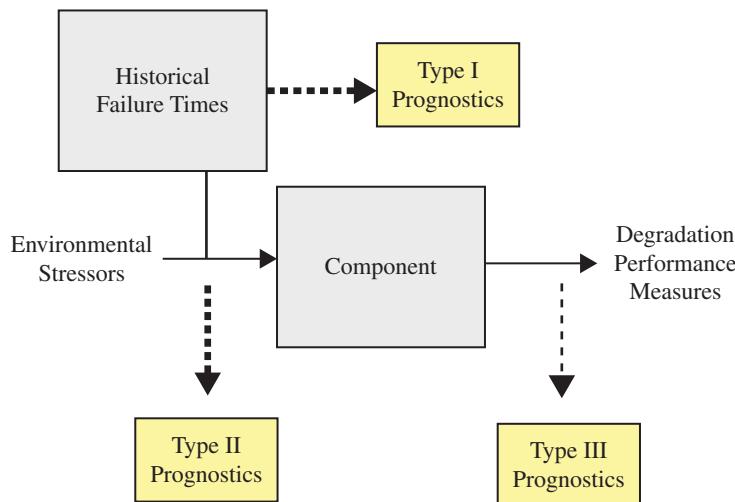


Figure 9.22 Prognostic method types [30].

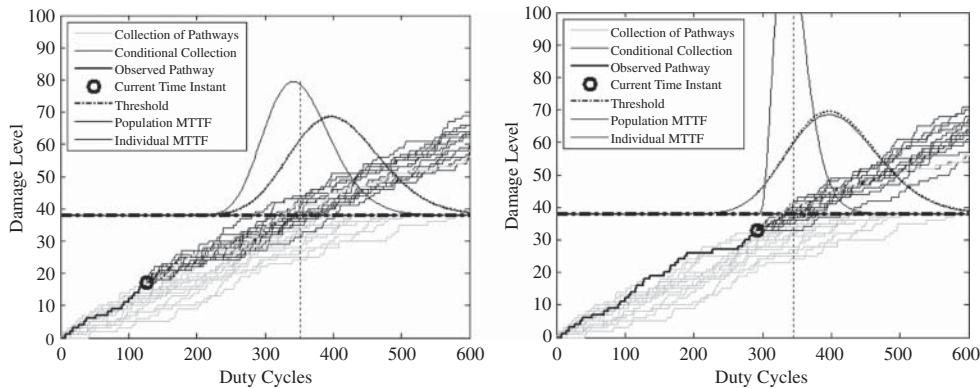


Figure 9.23 Illustration of how the uncertainty of estimation of RUL can be improved on the basis of an estimate of current degradation [31].

clear evidence of deterioration exists in terms of changes in monitored parameters. Figure 9.23 from [31], a tutorial on the same topic as [30], shows how condition information can be used to narrow down the uncertainty of predictions of RUL at two different stages of the life of a particular piece of equipment, as compared with the original reliability-based estimate. This is based on a ‘cumulative damage’ model, which is defined to be irreversible accumulation of damage in components under cyclical loadings [30, 31]. The cumulative damage model was expressed in terms of a discrete time Markov Chain, but [30] also discusses other possibilities such as general path models, and shock models, which may be more appropriate in certain circumstances.

Ref. [30] also describes a procedure which uses Bayesian inference to incorporate prior information from reliability data into the prognostic model.

Ref. [27] also recommends a hybrid method, using reliability-based estimates of RUL until incipient faults become detectable in measured data such as wear debris and vibration.

The hybrid or integrated approach is the third category of prognostic model discussed in [1], where reliability information can be combined with either physics-based or data-driven models. An example of the latter is given in [32], based on data from a large number of centrifugal pumps in the pulp and paper industry in Canada. The paper makes the point that much data about the ‘life’ of equipment is biased because the equipment was removed from service before actual failure; so-called ‘suspended histories’. A method is described (so-called Kaplan–Meier estimator) which can correct for this biased data. The EXAKT software package [33] was used to determine features having significant correlation with bearing failure, the best being the vibration level measured on the bearing in a certain frequency band in the vertical direction, and this was used as the fault indicator. It should be mentioned that it is rare to find such a simple indicator covering the majority of faults in a machine. Figure 9.24 shows the trend of measured parameter vs operating age in days, with insets showing the estimated survival probability at different points throughout the life, based on the prognostic model recommended in the paper. A probability of 0.5 is taken as giving the estimated time of failure. The actual life was 600 days, and at 560 days it was predicted that it would be approximately 575 days. At 580 days the estimate of RUL was less than 10 days. The predicted failure time was thus a little short of the actual, but only by a small amount. It will be noted from insets C and E that even where the trended parameter has stabilised or is falling, the ANN-based prediction model shows a steady decrease in the survival probability, so that it was able to cope with nonlinear and even non-monotonic data.

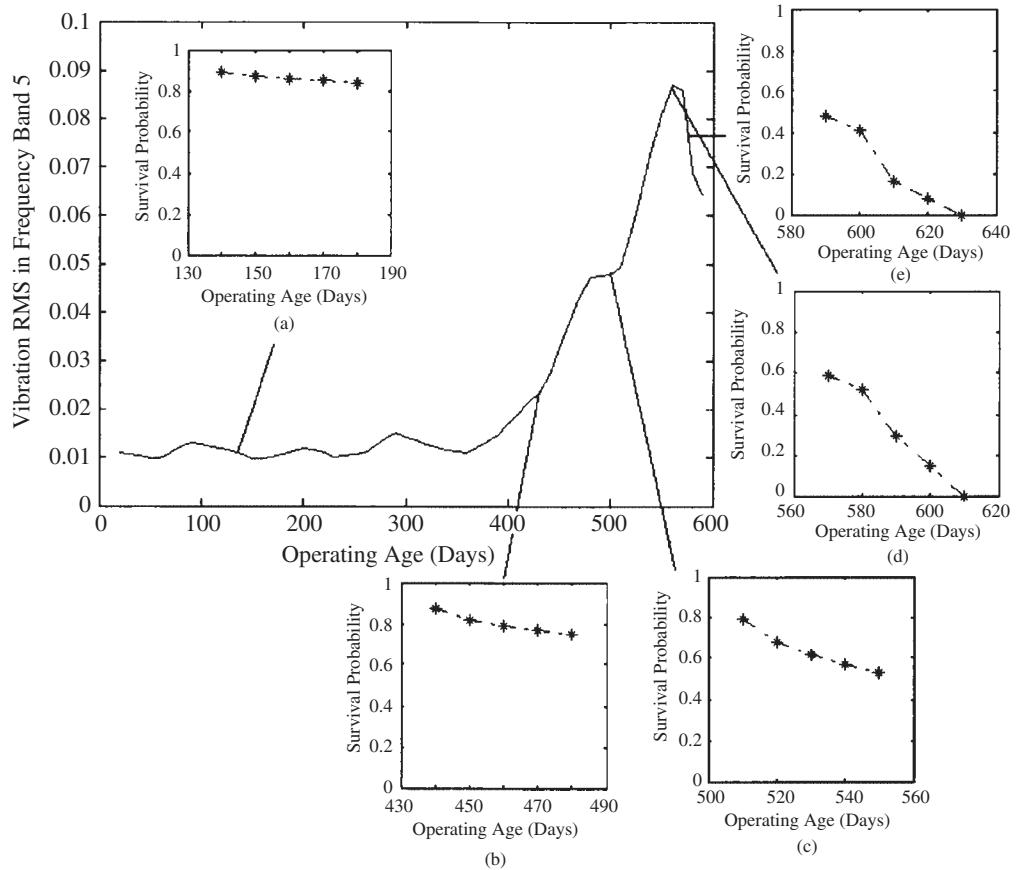


Figure 9.24 Graph of trended parameter vs Operating age for a particular case, and the prediction outputs of survival probability for the proposed prognostic model [33].

As mentioned above, some fleets of similar machines do tend to have limited types of failure, which can be detected sufficiently in advance to apply condition-based maintenance. Ref. [34] describes such a system for remote semi-automated monitoring of Vestas wind turbines.

Data analytics present an approach for extracting fault trends from data with a limited knowledge of the physical setup of the machine. Quite recently there have been a number of successes in applying deep-learning techniques to learn the structure of response data in normal condition and then detect anomalies that come from non-reversible changes in condition. One such example is given in Ref. [35], for improved tracking of a naturally developed crack in a spur gear test gearbox. It uses the so-called ‘Long Short-Term Memory (LSTM)’ augmentation of recurrent neural networks (RNNs). In [35], the signals analysed are the synchronously averaged signals corresponding to a particular gear, and so were already pre-processed. The network, trained on data before the crack development was obvious, produced ‘error’ (i.e. residual) signals, which were similar to those produced by the ‘classic’ residual technique illustrated in Figure 7.6a of Chapter 7. This approach was found to be superior to the pure signal processing techniques for smaller cracks, but only if account were taken of the robustness of the training; a more sensitive network, with earlier anomaly detection, would sometimes give false alarms. Various options for optimising the process are given in the paper.

Trending the growth of the anomalies remains a problem, and it is likely that the primarily data driven techniques will at best be able to give a running window of safe operation of perhaps some weeks into the future. However, this can be very valuable, for example allowing an aircraft to return to home base for more economical maintenance, or for a machine to last until a planned maintenance period.

9.3.4 *Simulation-Based Prognostics*

Chapter 8 gives many examples of fault simulation in gears, bearings and engines, and this section discusses how this can be used for prognostics. There is a wide range of sophistication of the simulation models that can be used for this purpose, ranging from simple analytical models producing typical signals generated by for example bearings, to ‘digital twin’ type models of a complex machine such as a wind turbine gearbox or aero engine, adapted for introduction of faults in sub-components.

The following gives typical examples of the various levels of sophistication of the fault models to be used for prognostics.

9.3.4.1 **Generic Fault Models**

When discussing the CWRU data in Section 9.3.2, it was pointed out that ANNs trained on this data (or any similar data set) could only be used to detect the same fault in the same machine, but an alternative approach was proposed in Ref. [36] where the CWRU data was used to generate fault data typical of the various faults. The simulated data was based on the bearing fault model published in [37], itself tuned to match the CWRU data. It gives a simplified model of the bearing itself, with 3 DOFs corresponding to the ‘masses’ of the three main components, inner race, balls, and outer race, but tuned to give impulse responses similar to the CWRU data. It does incorporate finite size notch ‘spalls’, so the responses can vary according to the length of spall. It also incorporates amplitude modulation for inner race and ball faults corresponding to their position with respect to the load direction. In [36] the simulated signals were generated with random variations of the model parameters such as impulse period and jitter, exponential decay constant and fault size. A broadband noise (chosen as pink) was also added, with signal/noise ratio (SNR) varying from 1 to 10. The limitation of the number of possible resonances to three is not very critical since the signatures extracted from the signals were envelope signals averaged over 25 repetition periods (in the angle domain rather than time domain, to cope with some speed variation) of the expected fault types (in [36] restricted to inner race and outer race). The ASA (angle synchronous averaging) results were all scaled to unit variance with zero mean value, so the only effect of fault size must have been the separation of the entry/exit pulses, and possibly the SNR variation (where the noise rather than the signal would actually be more likely to be constant, though very different for different machines and/or measurement points).

Ref. [36] uses the simulated data to train different classifiers, and then compares the effectiveness of the classifiers when applied to the original CWRU data, but also three other data sets, two from other bearing test rigs with artificially seeded faults (called ‘MFPT’ and ‘SpectraQuest’, and one from a naturally developing (inner race) fault in a high speed shaft bearing of a wind turbine. This approach was called ‘simulation-driven’, and was compared with ‘data-driven’ classifiers, which were directly trained using features extracted from the CWRU data, but tested on combined data from the MFPT and SpectraQuest data. When used for classification based on statistical features, the simulation-driven classifier achieved scores ranging from 87% to 89%, over a range of

different classification algorithms, compared with 67–86% for data-driven, with the best method being logistical regression (LR) for both approaches. Ref. [36] however recommends a different approach to classification, based on convolutional neural networks (CNNs) or nearest-neighbour dynamic time warping (NNDTW). The simulation- driven NNDTW classification gave the best results (94%) compared with the LR feature- based result, for the combined MFPT and SpectraQuest data. When applied to the wind turbine data, the NNDTW simulation-driven classifier detected the fault (i.e. > 50% probability) several days earlier than the CNN or LR classifiers.

However, even though in the last case there was a trend in the probability of the actual fault being present, this method is still primarily a fault detection and diagnosis method, and it would still be difficult to use it for prognostics. The latter depends on a monotonic progression of some feature, and the last data in the wind turbine case corresponds to the ‘discovery’ of a (single) ‘crack’ in the inner race of the bearing, with no indication of how severe it was, or whether the machine could have run for longer.

Ref. [38] presents a somewhat similar idea, with some advantages and some disadvantages. It proposes using actual data for the machine concerned, in nominally healthy condition, but adding simulated fault signals on a purely analytical basis. This thus makes use of the abundant data in normal condition, to model variability coming purely from operating conditions, but means that there is no way of knowing the strength of the fault signals that would occur in practice (presumably differing at different measurement points), so the severity could not be estimated. This could perhaps be adjusted in the case of a fleet with similar machines, after one occurrence of one failure (with data records for the entire life of the machine, or at least several in the period before the failure). Another problem is that the simulated faults corresponded only to those exhibiting separated impulse responses, i.e. a single localised spall in one component or another. The results in the paper show some success in detecting and diagnosing faults in several different data sets, including the CWRU data, but no indication of severity.

9.3.4.2 Cases with Simple Monotonically Changing Features

A good example of this situation is represented by the case of faults in IC engines, as discussed in Section 8.4 of Chapter 8. For the three different types of faults simulated, viz. misfire, piston slap, and bearing knock, physics-based features could be found that characterised the faults, with respect to type, location and severity, and simulations produced signals from which the same features could be extracted. This section describes the way in which separate ANNs were trained, purely

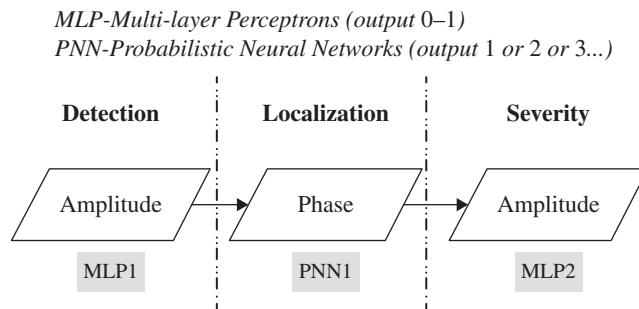


Figure 9.25 Structure of the three-stage ANN system.

on simulated data, to recognise the three following situations (for each type of fault), as shown in Figure 9.25:

- 1) Existence of a fault of a given type.
- 2) Which cylinder?
- 3) Severity.

In the following, the application and results of this approach to the three type of faults, misfires, piston slap, and bearing knock, including the features used, are discussed in some detail. Note that since the output of the simulation models described in Section 8.4.1.2 of Chapter 8 is deterministic, some means had to be found to compensate for the variations found in actual data for a fixed set of operating conditions. One reason for the latter was that measurements were made over a six-month period, where the ambient temperature varied by 20 °C, but the same approach could be used for any variation in measurements due to variations in operating conditions. The simulated data used for training the networks was thus varied by an amount determined by the standard deviation of the measurements for each speed/load condition, with and without faults, over a long period. In practice, much more data would be available for normal rather than faulty condition, but it is reasonable to assume that random variations due to operating conditions would be similar for both.

As seen in Figure 9.25, feed-forward multi-layer perceptron (MLP) networks were used for detection and severity, with outputs in the range 0–1. For detection this was a simple Yes or No question, so that the output would either be 0 for No or 1 for Yes. For fault localization, probabilistic neural networks (PNN) were chosen, because of their advantages in discrete classification, with outputs 1, 2, 3, 4 indicating the faulty cylinder.

For misfire, as discussed in Section 8.4.1.2 of Chapter 8, for uniform firing on all cylinders, the dominant frequency component would be the firing frequency, in this case the fourth harmonic of the cycle frequency (half shaft speed). Any anomaly in the combustion between cylinders would repeat at cycle frequency, so the appropriate feature was the amplitude ratio of 1st to 4th harmonic, and this was used both for detection and severity. MLP1 (detection) consisted of three layers: input, hidden and output, with the number of neurons in the hidden layer determined by trial-and-error. A nonlinear sigmoid transfer function was used to force the result to be close to zero or one.

Originally (including in Ref. [39]) the same approach was used for MLP2 (severity), with the output forced to be one of the three conditions used for training, viz., zero, 50% or 100% misfire, but in Ref. [40] this was corrected to correspond to the desired result in practice; that the result should be on a sliding scale from 0 to 1. Thus, as described in detail in [40], saturating linear transfer functions, $\text{satlin}(x)$, were introduced into both layers, to give a linearly weighted result in the range 0–1. To get the best results for the severity, it was found necessary to train a separate MLP2 network for each speed, as described in [40]. The same procedure was applied to both methods discussed in detail in Section 8.4.1.2, viz. torsional vibration of the crankshaft, and pseudo angular acceleration of the block.

As mentioned in Section 8.4.1.2 (and in [40]), the faulty cylinder was indicated by the phase of the 1st harmonic, so this was the feature used to determine the localization of the fault. For the torsional vibration, the phases of the 1st harmonic with misfire in a certain cylinder are nearly fixed for any speed/load condition, so the inputs to the PNN for torsional vibration had only one element (the phases of the 1st harmonic), but for the pseudo angular acceleration methods, the phases of the 1st harmonic with misfire are only fixed at a certain speed, so the inputs to the PNN for the pseudo angular acceleration method included the phases of the 1st harmonic and the speed. Note that the same would apply to torsional vibration with a flexible crankshaft, where fixed modal frequencies interact with the varying shaft harmonics, as discussed in Section 8.4.1.1.

Table 9.1 Distribution of data with networks from simulation and experiment.

For training (simulation)			For test (experiment)		
Normal	100% misfire	50% misfire	Normal	100% misfire	50% misfire
180 in total	180 in total, 45 in cylinder 1 45 in cylinder 2 45 in cylinder 3 45 in cylinder 4	180 in total, 45 in cylinder 1 45 in cylinder 2 45 in cylinder 3 45 in cylinder 4	15 in total	19 in total, 9 in cylinder 1 7 in cylinder 2 3 in cylinder 3	2 in total, 2 in cylinder 1

The networks were trained entirely with simulated data (total of 540 data sets, with 180 for each of the three amounts of misfire zero, 50% and 100%), and had a 100% success rate when tested on the experimental data shown in Table 9.1.

The ranges of the results for the new MLP2 design, for 50% misfire, were (0.495–0.512) and (0.488–0.523) for torsional vibration and angular acceleration, respectively, and for 100% misfire, were (0.991–1.00) and (0.984–1.00), respectively [40].

This study was based on misfires in one cylinder at a time only, but could most likely be extended to multiple cylinders. It would mean that the amplitude and phase information would have to be combined, and that fault location and severity (in each cylinder) would have to be determined at the same time.

For piston slap, for which the simulation models and signal processing are described in Section 8.4.1.2, updated results for the ANNs used for automated diagnosis were published in Ref. [41]. The approach was very similar to misfire, except that (squared) envelope signals were used instead of the direct vibration signals, since the diagnostic information was contained in the responses to impacts, as for rolling element bearings. Thus, a frequency band had to be chosen to demodulate, and as described in Section 8.4.1.2 this was done using a kurtogram as a guide, as for bearings. This gave two benefits, since the lower cutoff frequency of the demodulation band excluded additive components affected by misfire, but also the simulation model did not have to reproduce resonance frequencies exactly, since these were removed by the demodulation, and did not greatly affect the envelope signals.

For detection, the structure of the MLP1 network was effectively the same as for misfire, using log sigmoid transfer functions. For severity, the harmonic amplitude information from the envelope signals was found to be sensitive to both speed and load (in that case almost proportionally), so once again separate MLP2 networks were trained for the three speeds, and moreover the input amplitude features of the individual MLPs were scaled (divided) by the load values. The training data was once again entirely simulated, that for ‘no piston slap’ and ‘piston slap’ (at two different levels) including some cases with no other fault (the first groups in Figures 9.26 and 9.27), and some including added misfire faults (the remaining cases). As will be seen, the latter had no influence on the results, demonstrating the robustness of the method.

Figure 9.26 shows that the detection results for MLP1, trained on simulated data and tested on experimental data (28 cases with piston slap, and 21 cases without). The success rate is 100%.

As for the misfire classification, a linear transfer function was required in the output layer of MLP2, but in contrast to misfire (with maximum value 100%) the linear function was open-ended (so-called `purelin(x)`). Only three levels of piston clearance were used for simulation and testing; normal clearance, 3× normal clearance and 6× normal clearance, denoted by 0, 0.5, and 1.0, respectively. With the linear transfer function, this would mean that 9× normal clearance would be represented by 1.5.

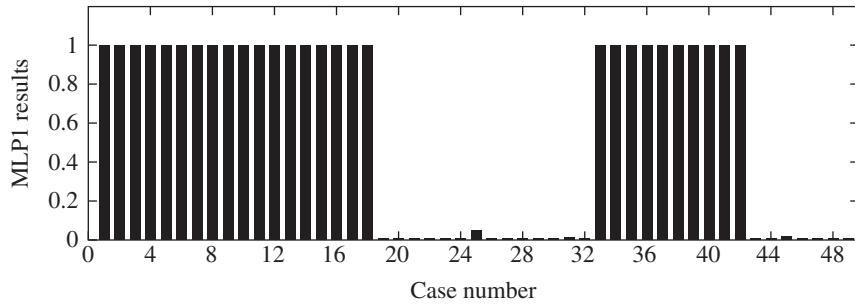


Figure 9.26 Output of MLP1 for the piston slap faults.

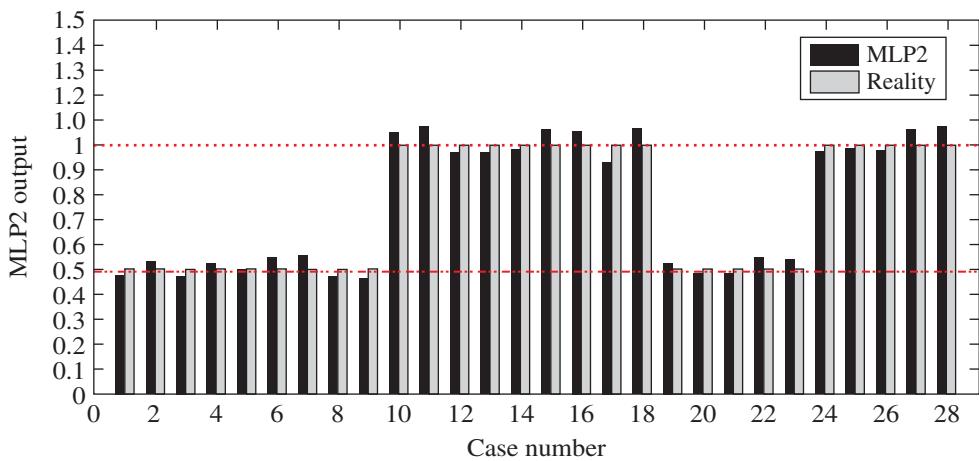


Figure 9.27 Output of MLP2 for the piston slap faults.

Figure 9.27 gives the results achieved by the improved MLP2 network, tested on 14 data sets with 3 \times normal clearance, and 14 with 6 \times normal clearance, each including five data sets with combustion faults. Once again, the results from training purely with simulated data and testing with experimental data are close to ideal.

In this case, the phase features required for fault localization were somewhat more complex than for misfire. There was some variation of the first harmonic, but the second harmonic was more stable. To differentiate the cylinder in which the slap was occurring, it was found necessary to include the phases of the third, fifth, and seventh harmonics to get the best results [41].

For bearing knock, for which updated classification results were published in Ref [42], the approach was similar to that for piston slap, except that the number of cases to be used for testing was more limited. Since the test engine was quite old, the maximum speed was limited to 3000 rpm, and this was in fact the minimum speed at which the knock was clearly evident (it would have been much more evident at the original rated speed of 6000 rpm). For the same reason, the maximum clearance was limited to 4 \times normal clearance, so the severity was scaled to 0, 0.5, and 1.0, corresponding to normal clearance, 2 \times normal clearance, and 4 \times normal clearance, respectively. The simulations were performed as described in Section 8.4.1.2, and also used the squared envelope signals, with a

slightly different but overlapping demodulation band, but with a different measurement point near a crankshaft main bearing. The output transfer function for MLP2 was slightly different from that for piston slap, in that it once again used a saturated linear function, but with saturation level well outside the expected range, because it was more computationally efficient.

The phase information was somewhat similar to that for piston slap, so once again PNN1 used higher harmonics of the envelope signals in addition to the first, in this case the second and fifth harmonics.

Because of the speed limitation, there were a limited number of experimental cases on which to test the ANNs, and all increased clearances were in Cylinder 2, but even so the results of testing were excellent.

Figure 9.28 gives the results for MLP1 based on 10 cases with increased clearance at two levels, and 27 with normal clearance, of which cases 32–37 had oversize piston clearance as well. Even though case 37 gives a result of 0.30, it would still have been classified as normal, taking the cutoff value as 0.5.

Figure 9.29 shows that the updated MLP2 gave close to ideal results for severity.

Everything demonstrated so far in this section is basically for diagnosis of a particular condition, rather than prognostics as such, but it has been shown that the effects of fault severity can be predicted

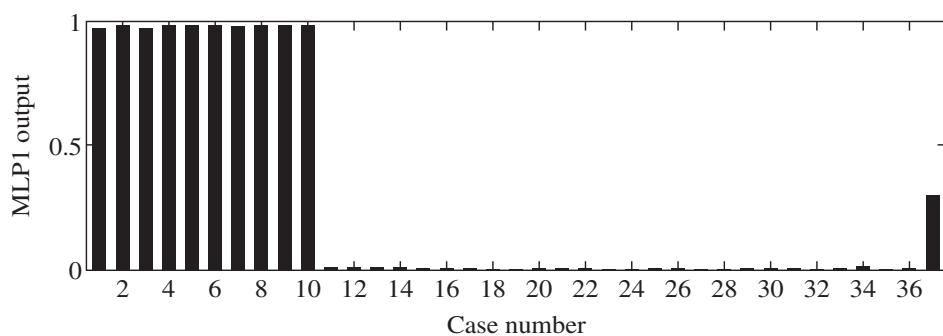


Figure 9.28 Output of MLP1 for the bearing knock faults.

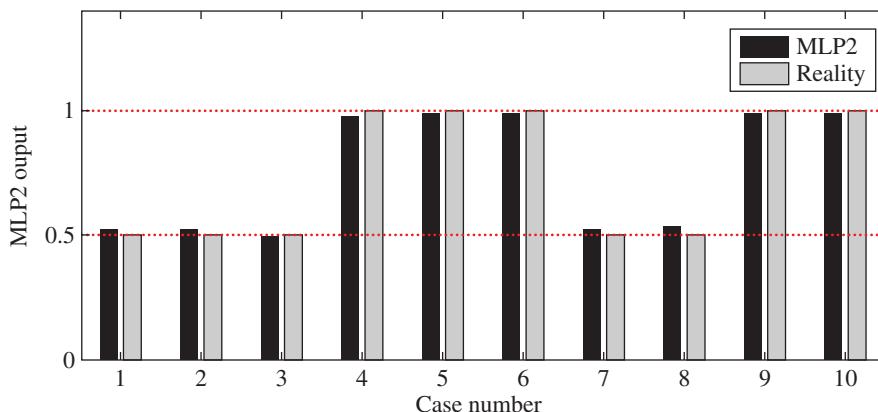


Figure 9.29 Output of MLP2 for the bearing knock faults.

with reasonable accuracy, and could be extended to larger fault sizes with reasonable confidence. Thus, this technique can be used for assistance with prognostics, by matching features of developing faults with those predicted with simulation.

A more direct approach to this problem was presented in Refs. [43, 44], where the rate of growth of bearing knock faults was studied, based on a wear model updated using feedback from vibration measurements. The basic software developed by Chen [39] was used for the kinetics/kinematics of the connecting rod and crankshaft, as was the basic Reynolds equation software for the big end bearing, with the difference that the bearing geometry was no longer kept cylindrical, of different diameters, but was adjusted by the estimated wear profile. Moreover, since it was discovered that wear would be very low where the bearing maintained full hydrodynamic lubrication (assumed by Reynolds equation), a test for the latter was inserted, based on calculated oil film thickness, and the model changed to one of boundary or dry lubrication when there was not full lubrication.

In [43], the test was simply the oil film thickness given by Reynolds equation becoming negative, whereupon the dynamic force was calculated using a contact pressure equation for boundary lubrication [45]. The resulting wear rate was based on Archard's law, a simple equation stating that the wear rate is proportional to contact force and sliding velocity through a coefficient depending on factors such as surface roughness and material hardness, often determined empirically. This first approach gave a reasonable match to an experimental result in the literature in terms of total wear [46], by adjustment of Archard's constant, where it was found that the predicted wear profile also matched fairly well. The predicted wear profile did not fully take account of the need for the worn surface (on the soft bearing) to be locally conformal with the virtually unworn surface of the hard journal.

Ref. [44] improved the model in a number of ways, including surface conformity, but mainly by introducing a transition mixed lubrication phase between the full hydrodynamic lubrication and dry contact [47]. This allowed for varying asperity contact depending on the oil film parameter λ , this being the ratio of oil film thickness (calculated by Reynold's equation) to the standard deviation of the surface roughness. In the range $3 > \lambda > 0$ the load was assumed to be divided between the oil film and the surface asperities in a ratio decreasing from 1 to zero, giving a more gradual change to the dry contact formula. Figure 9.30 compares the simulated wear profile using this approach with the measured profile from [46].

Despite the bearing profile no longer being cylindrical, the simulated vibration responses resulting from a given clearance along the diameter defined by the impact point were found to agree very well with those estimated by Chen for a cylindrical profile, the impact forces primarily being determined by the change in velocity across this clearance. This approach thus gives the possibility to perform genuine physics-based prognostics, using the current measured vibrations to infer the clearance given

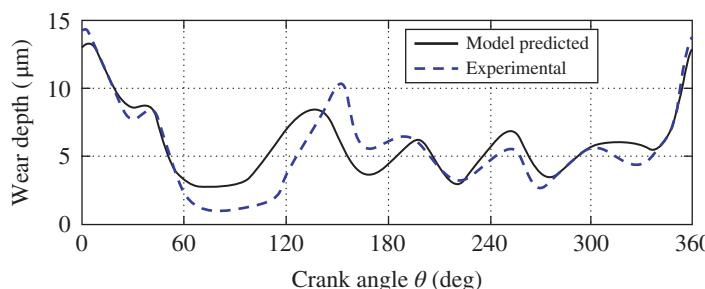


Figure 9.30 Simulated vs experimental wear profiles [42].

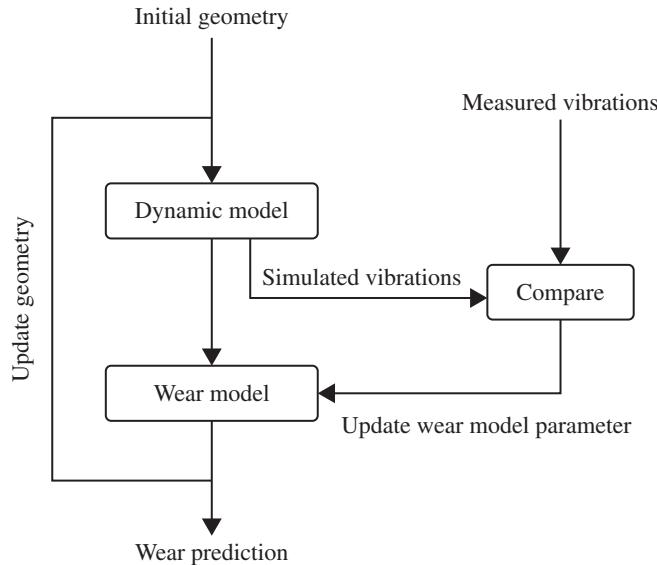


Figure 9.31 Proposed vibration-based scheme for updating wear prediction [46].

by the modified wear profile. Predictions of future development can thus be continuously updated by using the current measured vibrations to update the wear prediction parameters on a continuing basis.

The case of engines is perhaps somewhat simpler than machines such as gearboxes, where there is more interaction between the components such as bearings, gears and rotors, and their connection with the casing. Full application has yet to be demonstrated, but a start has already been made on the inclusion of a wear prediction model in a gearbox. Ref. [48] also uses Archard's law to predict the wear profile of gear teeth subject to dry wear (the same data as used in Section 7.2.2.2, Figure 7.23, in Chapter 7).

The increasing geometric TE caused by this wear resulted in increasing vibration, which could be used as a measure of the wear development, as indicated in Figure 9.31. The 'dynamic model' shown there was a simplified model, without inclusion of the casing, similar to that in Figure 8.2 of Chapter 8. However, with suitable scaling based on experimental measurements, the simulated response vibrations were sufficiently similar to the measurements that they could be used for updating the parameters of the wear equation. The results in [48] were reasonably successful, but the gearbox modelling and choice of wear severity indicators is being continuously improved at the time of writing. Techniques are also being developed to use vibration measurements to distinguish between different wear modes, abrasive wear as in the above case, and surface pitting, as in the data of Figure 7.18 of Chapter 7, to enable the use of different wear equations for the two cases [49].

9.4 Future Developments

9.4.1 Advanced Modelling

It is now becoming more common for manufacturers (OEMs) of machines to have so-called 'digital twins' of their products, i.e. CAD, FEM, and multi-body dynamic models of their different

machine types. Such models are developed during the design/development phase, but in some cases can be adapted for individual machines throughout their service life, in particular when the OEM is responsible for their efficient performance, ongoing maintenance etc. Such models would not normally include fault simulation in the various components, such as bearings and gears, but it should be possible to add these using substructuring and model reduction techniques, as described in Section 8.3.3. The simulation models could be used to simulate the effects of faults of increasing severity compared with those for which the model was validated, and also in other, but similar, components, e.g. other cylinders in the same engine, as discussed in Section 9.3.4.2 for IC engines. As demonstrated there, also for gearboxes, it should be possible to add wear prediction models, with feedback on the current state of wear, and updating of wear prediction parameters, by using vibration measurements.

In many cases it is possible to use simpler simulation models of machines, or indeed of sub-assemblies, to predict the development of faults, as demonstrated in Section 9.3.4. Compensation for modal effects can also often be made neglecting the phase of transfer functions, as demonstrated in Section 6.3.2 using the real cepstrum only. However, to achieve true inverse filtering of transfer function effects, for example to reproduce complex time waveforms, it is necessary to include both amplitude and phase effects, something which is becoming possible by the extension of OMA (operational modal analysis) techniques from structures to rotating machines. The main difference between structural health monitoring (SHM) and machine health monitoring (MHM) is that the former is mainly based on changes in the modal properties of the structures, whereas machine faults are perhaps 70% indicated by changes in forcing functions, and only 30% by changes in modal properties. Thus, OMA is used largely in MHM to generate inverse filters to extract forcing functions from responses, whereas in SHM the forcing functions are not of great interest, being broadband ambient excitations such as wind, waves, road traffic etc., with no information about the structural health.

It does appear that cepstral methods of OMA [50] have some advantages in their application to machines, partly because of the simplicity of using exponential ‘lifters’ to remove much extraneous masking by complex forcing functions, see for example Figure 6.12 of Chapter 6, while only adding a known small amount of damping to the modal model (which can be compensated for). Notch lifters are also very useful for removing the large families of discrete frequency components typical of machines, as compared with structures.

Ref. [51] describes an ongoing research project (at the time of writing) with the aim of developing these techniques, and others such as blind source separation (BSS), to achieve full extraction of source signals from external vibration measurements, using a gearbox as a typical complex machine with interacting effects of gears, bearings, and rotors.

Figure 9.32 is a schematic diagram of the proposed procedure. Short- and long-pass filtering of the cepstrum is first used to separate into low and high quefrency components, where the former will

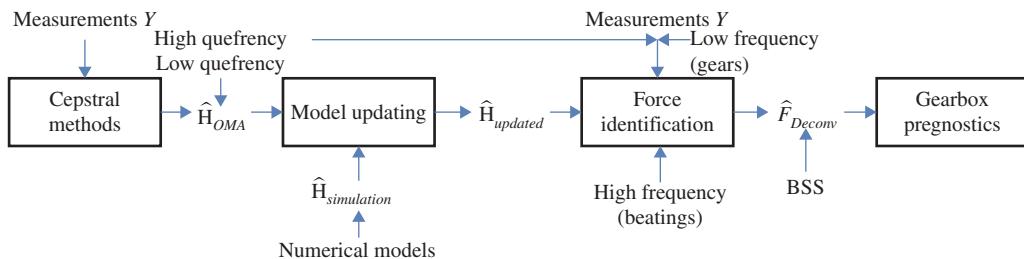


Figure 9.32 Proposed flow diagram for processing.

be dominated by the modal properties, and the latter for example containing gear information (as in Figure 6.22b). The important part of this gear information will be at relatively low frequencies, up to five or so harmonics of the garmesh frequencies, where the modes are still often separated, and where OMA can be used to generate modal models (frequency, damping, and mode shapes lacking an overall scaling factor).

The modal parameters will be used to update FEM models, as in Section 8.3.3.1, which can then be used to scale the modal models [50] to give scaled responses, for both current and future predictions (after simulated fault evolution). The cepstral methods of OMA require the responses of different sources to be separated, and methods will be developed to achieve this. Ref. [52] showed that if there is no more than one (CS2) source with a particular cyclic frequency, the cepstrum could be applied to the spectral correlation (Section 3.6.1) at this cyclic frequency to obtain the transfer function for that particular source, and this is one potential method of BSS for that situation. In [52] that approach was used to separate the response to excitation by the engine of a diesel railcar, in the presence of other excitations such as air flow, rail roughness, and air conditioning unit.

The scaled modal matrix can then be used to obtain scaled forcing functions from the separated high quefrency part of the overall response, at least in the lower frequency range dominated by gears, where the modes are separated. It is possible that the FRFs for the higher frequency part, with overlapping modes, but more likely containing the resonances excited by bearing faults, may be extracted using simpler methods (so-called mid-frequency) with a mixed modal/SEA (statistical energy analysis) approach, since it is only required that the bandpass-filtered envelope signals are correctly scaled. Further BSS may be required here to separate different sources of the same type, such as different bearings on the same shaft.

While many of these procedures need further development, many sub-components have already been demonstrated. It is thought that this approach will be applicable to the most critical machines, with detailed digital twin models, allowing extraction of forcing functions and continuously updated modal properties, and when combined with wear models, will be able to predict future signals so as to give a true physics-based prognostic capability.

9.4.2 *Advances in Data Analytics*

Analysis of Big Data is very powerful, and will certainly improve in the years ahead. Thus, it is expected that where there is a lot of data, viz. from machines in, or close to, normal condition, it will be possible to determine with a high degree of certainty when they start to deviate from normal condition, and by how much, at least for limited changes. It should be possible to learn the patterns of deviation of vibration responses, while in normal condition, caused even by wide variations in operating conditions, and apply these also to simulated responses, even for faults which have not actually been experienced.

A situation can be foreseen where the most powerful prognostic systems will combine data analytics with physics-based models, initially relying primarily on simulated faults, but updated as data is gathered from experience with actual faults. One of the first advances is expected to apply to fleets of similar machines, where laboratory experiments with seeded faults on a test unit could be used to update and calibrate the simulation models, though still just for the limited number of types and locations of faults that could be economically tested.

Developments in data analytics should help to unify the interpretation of naturally occurring vs artificially seeded faults, and also to find health indicators and features that are less sensitive to the actual machine(s) on which the systems are trained.

The trend towards Industry 4.0 and the Internet of Things, with many more inbuilt and well-maintained transducers will certainly be of much benefit in this regard.

At the same time, it is expected that the vast number of machines where online monitoring cannot be justified, will still be serviced by field portable data collectors, though this type of operation has the potential to be greatly improved. The vast majority of collectors have changed little in the last twenty years with respect to developments in signal processing tools. However, there is an increasing demand by users to get access to the raw data collected (*inter al.* to be able to apply Big Data tools), rather than just the processed results the individual manufacturers happen to provide. At least one supplier is known to already give such access. This would open a market for providers of offline signal processing tools (virtual analysers) to cover the large number of new developments already missing from the standard units, or yet to be developed. The data collector OEMs would probably still dominate the market for supply of the hardware, and basic analysis tools, because of their long experience in designing for ruggedness in difficult operating environments, explosion hazards etc., as well as ease of operation.

References

1. Heng, A., Zhang, S., Tan, A.C.C., and Mathew, J. (2009). Rotating machinery prognostics: state of the art, challenges and opportunities. *Mechanical Systems and Signal Processing* 23 (3): 724–739.
2. Jardine, A.K.S., Lin, D., and Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing* 20: 1483–1510.
3. Randall, R.B. (1985). Computer aided vibration spectrum trend analysis for condition monitoring. *Maintenance Management International* 5: 161–167.
4. Thompson, R.A. and Weichbrodt, B. (1969). ‘Gear diagnostics and wear detection’. *ASME Paper 69-VIBR-10*, Vibration Conference, Philadelphia (March 1969).
5. Brüel & Kjær (1988). ‘Machine-condition monitoring using vibration analysis – a case study from an iron-ore mine’. *B&K Application Note* BO 0178-11, Brüel & Kjær, Copenhagen.
6. Brüel & Kjær (1989). ‘Systematic machine-condition monitoring – a case study from Parenco Paper Mill in Holland’. *B&K Application Note* BO 0299-11. Brüel & Kjær, Copenhagen.
7. Brüel & Kjær (1988). ‘Machine-condition monitoring using vibration analysis – permanent monitoring of an Austrian Paper Mill’. *B&K Application Note* BO 0247-11. Brüel & Kjær, Copenhagen.
8. Brüel & Kjær (1988). ‘Machine-condition monitoring using vibration analysis – the use of spectrum comparison for bearing fault detection – a case study’. *B&K Application Note* BO 0253-11. Brüel & Kjær, Copenhagen.
9. Brüel & Kjær (1986). ‘Six weeks’ advance warning of breakdown’. *B&K Application Note* BO 0230-11. Brüel & Kjær, Copenhagen.
10. Bradshaw, P. and Randall, R.B. (1983). ‘Early detection and diagnosis of machine faults on the trans Alaska pipeline’. *MSA Session, ASME Conference*, Dearborn MI (September 1983).
11. Stewart, R.M. (1977). ‘Some useful data analysis techniques for gearbox diagnostics’. *Proceedings of the Meeting on the Applications of Time Series Analysis*, ISVR, University of Southampton, Southampton, UK (19–22 September 1977), Paper #18.
12. Dyer, D. and Stewart, R.M. (1977). ‘Detection of rolling element bearing damage by statistical vibration analysis’ *ASME Paper* (26–30 September).
13. Samanta, B. (2004). Gear fault detection using artificial neural networks and support vector machines with genetic algorithms. *Mechanical Systems and Signal Processing* 18: 625–644.
14. Samanta, B. and Al-Balushi, K.R. (2003). Artificial neural network based fault diagnostics of rolling element bearings using time-domain features. *Mechanical Systems and Signal Processing* 17: 317–328.
15. Ding, Y. and Rieger, N.F. (2003). Spalling formation mechanism for gears. *Wear* 254: 1307–1317.
16. Sawalhi, N. and Randall, R.B. (2008). ‘Novel signal processing techniques to aid bearing prognostics’. *IEEE PHM Conference*, Denver.
17. Sawalhi, N. and Randall, R.B. (2011). Vibration response of spalled rolling element bearings: observations and signal processing techniques to track the spall width. *Mechanical Systems and Signal Processing* 25: 846–870.
18. Epps, I.K., (1991) ‘An Investigation into vibrations by faults in rolling element bearings,’ PhD dissertation, University of Canterbury, NZ.

19. Dowling, M.J. (1993). Application of non-stationary analysis to machinery monitoring. In: *IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 1, 59–62.
20. Petersen, D., Howard, C., Sawalhi, N. et al. (2015). Analysis of bearing stiffness variations, contact forces and vibrations in radially loaded double row rolling element bearings with raceway defects. *Mechanical Systems and Signal Processing* 50–51: 139–160.
21. Petersen, D., Howard, C., and Prime, Z. (2015). Varying stiffness and load distributions in defective ball bearings: analytical formulation and application to defect size estimation. *Journal of Sound and Vibration* 337: 284–300.
22. Moazen Ahmadi, A., Howard, C.Q., and Petersen, D. (2016). The path of rolling elements in defective bearings: observations, analysis and methods to estimate spall size. *Journal of Sound and Vibration* 366: 277–292.
23. Smith, W., Hu, C., Randall, R.B., and Peng, Z. (2015). Vibration-based spall size tracking in rolling element bearings. *Mechanism and Machine Theory* 21: 587–597.
24. Zhang, H., Borghesani, P., Smith, W.A. et al. (2021). Tracking the natural evolution of bearing spall size using cyclic natural frequency perturbations in vibration signals. *Mechanical Systems and Signal Processing* 151, Ref. 107376.
25. Qiu, J., Seth, B.B., Liang, S.Y., and Zhang, C. (2002). Damage mechanics approach for bearing lifetime prognostics. *Mechanical Systems and Signal Processing* 16 (5): 817–829.
26. Paris, P.C. and Ergonon, F. (1963). A pitting model for rolling contact fatigue. *ASME Transactions Journal of Basic Engineering* 85: 528–534.
27. Roemer, M.J., Kacprzynski, G.J., Orsagh, R.F., and Marshall, B.R. (2005). Prognosis of rotating machinery components', Chapter 19. In: *Damage Prognosis for Aerospace, Civil and Mechanical Systems* (eds. D.J. Inman, C.F. Farrar, V. Lopes Jr. and V. Steffen Jr.). Wiley.
28. Case Western Reserve University Bearing Data Center Website (<http://csegroups.case.edu/bearingdatacenter/home>) (accessed 16 November, 2020).
29. Smith, W.A. and Randall, R.B. (2015). Rolling element bearing diagnostics using the Case Western Reserve University data: a benchmark study. *Mechanical Systems and Signal Processing* 64–65: 100–131.
30. Hines, J.W. and Usynin, A. (2008). Current computational trends in equipment prognostics. *Int. J. Comput. Intell. Systems* 1 (1): 95–109.
31. Hines, J.W. (2009). 'Empirical methods for process and equipment prognostics', Tutorial 2, *MFPT 2009 Conference*, Dayton, Ohio (28–30 April 2009).
32. Heng, A., Tan, A.C.C., Mathew, J. et al. (2009). *Mechanical Systems and Signal Processing* 23 (5): 1600–1614.
33. Jardine, A.K.S., (2007). "EXAKT condition-based optimization software", BANAK Inc., Toronto, ONT. http://www.banak-inc.com/exakt_factsheet.pdf accessed 21/03/2021.
34. Christensen, J.J., Andersson, C., and Gutt, S. (2009). Remote condition monitoring of Vestas turbines. In: *Proceedings of European Wind Energy Conference*, Mar. 2009, 1–10.
35. Wang, W., Galati, F.A. and Szibbo, D. (2019). "LSTM residual signal for gear tooth crack diagnosis". *Proceedings of COMADEM2019*, Huddersfield, UK (3–5 September 2019).
36. Sobie, C., Freitas, C., and Nicolai, M. (2018). Simulation-driven machine learning: bearing fault classification. *Mechanical Systems and Signal Processing* 99: 403–419.
37. Sassi, S., Badri, B., and Thomas, M. (2007). A numerical model to predict damaged bearing vibrations. *Journal of Vibration and Control* 13 (11): 1603–1628.
38. Hemmer, M., Klausen, A., van Khang, H. et al. (2019). Simulation-driven deep classification of bearing faults from raw vibration data. *International Journal of Prognostics and Health Management* 10 (Special Issue on Deep Learning and Emerging Analytics) 030: 1–12.
39. Chen, J. (2013) 'Internal combustion engine diagnostics using vibration simulation'. PhD Dissertation, University of New South Wales.
40. Chen, J. and Randall, R.B. (2015). Improved automated diagnosis of misfire in internal combustion engines based on simulation models. *Mechanical Systems and Signal Processing* 64–65: 58–83.
41. Chen, J., Randall, R.B., and Peeters, B. (2016). Advanced diagnostic system for piston slap faults in IC engines, based on the non-stationary characteristics of the vibration signals. *Mechanical Systems and Signal Processing* 75: 434–454.
42. Chen, J. and Randall, R.B. (2016). Intelligent diagnosis of bearing knock faults in internal combustion engines using vibration simulation. *Mechanism and Machine Theory* 104: 161–176.
43. Haneef, M.D., Randall, R.B., and Peng, Z. (2016). Wear profile prediction of IC engine bearings by dynamic simulation. *Wear* 364–365: 84–102.
44. Haneef, M.D., Randall, R.B., Smith, W.A., and Peng, Z. (2017). Vibration and wear based analysis of IC engine bearings by numerical simulation. *Wear* 384–385: 15–27.
45. Nikolic, N., Torovic, T., and Antonic, Z. (2012). A procedure for constructing a theoretical wear diagram of IC engine crankshaft main bearings. *Mechanism and Machine Theory* 58: 120–136.

46. Ushijima, K., Aoyama, S., Kitahara, K., Okamoto, Y., Jones, G., Xu, H., (1999). 'A study on engine bearing wear and fatigue using EHL analysis and experimental analysis', *SAE Int.* 1999-01-1514. <https://www.sae.org/publications/technical-papers/content/1999-01-1514/> (accessed 16 November 2020).
47. Greenwood, J.A. and Tripp, J.H. (1970). The contact of two nominally flat rough surfaces. *Proceedings of the Institution of Mechanical Engineers* 185 (1): 625–633.
48. Feng, K., Borghesani, P., Smith, W.A. et al. (2019). Vibration-based updating of wear prediction for spur gears. *Wear* 426–427: 1410–1415.
49. Feng, K., Smith, W.A., Borghesani, P. et al. (2021). Use of cyclostationary properties of vibration signals to identify gear wear mechanisms and track wear evolution. *Mechanical Systems and Signal Processing* 150, Ref. 107258.
50. Randall, R.B., Antoni, J., and Smith, W.A. (2019). A survey of the application of the cepstrum to structural modal analysis. *Mechanical Systems and Signal Processing* 118: 716–741.
51. Randall, R.B. (2019). 'The potential for obtaining scaled separated forcing functions and scaled transfer functions from operational response vibrations, in particular of rotating machines'. Keynote paper. *ICEDyn conference*, Viana do Castelo, Portugal (June 2019).
52. Hanson, D., Randall, R.B., Antoni, J. et al. (2007). Cyclostationarity and the cepstrum for operational modal analysis of MIMO systems—part I: modal parameter identification. *Mechanical Systems and Signal Processing* 21 (6): 2441–2458.

Appendix

Exercises and Tutorial Questions

Introduction

Much of the material in this book has been used as the background for a course in Machine Condition Monitoring taught by the author at the University of New South Wales, Sydney, Australia. This appendix gives details of a number of typical tutorial and examination questions on the material, as well as exercises that can be used, for example, as the basis of assignments. The answers might be based solely on material given in the chapter, or perhaps (for tutorial questions and assignments) on an internet search of the appropriate topic. The exercises are divided into the material corresponding to each chapter, so that Section A.1 corresponds to Chapter 1, for example. Note that there are no questions relating to chapters 5 and 8.

Data for some exercises will be found on the following website:
(<http://www.wiley.com/go/randall>)

Answers to selected questions and assignments will also be found on this website. Data sets may be expanded from time to time in addition to those referred to in this Appendix.

A.1 Introduction and Background

This chapter is largely descriptive, so examples are given of a number of typical descriptive questions.

A.1.1 Exam Questions

1. Write an essay on the use of condition monitoring techniques as an aid to maintenance, comparing condition based maintenance with other strategies, and discussing the economic factors which affect the choice of maintenance strategy. Discuss the range of available condition monitoring techniques (i.e. in addition to vibration analysis) and where they are most applicable. Compare the application of permanent and intermittent monitoring.
2. Describe a number of condition monitoring techniques based on analysis of the lubricating oil, and discuss the situations where oil analysis would be used as compared with vibration analysis.
3. Compare the properties of proximity probes and accelerometers for measuring the vibrations of rotating machines, and indicate how this affects their application in machine condition monitoring.

4. ‘Dynamic range’ gives a measure of the ratio of the largest to the smallest signal components that can be measured in a signal. This can however be interpreted in different ways. It can mean the largest and smallest values that can be measured by a given transducer, but can alternatively mean the difference in level between the largest and smallest components in a frequency spectrum. Explain how these might be different, using the example of a proximity probe signal.
5. Discuss the pros and cons of the three main vibration transducers used for condition monitoring, viz. proximity probes, velocity probes and accelerometers, in terms of dynamic range, frequency range and signal fidelity.
6. Vibration velocity gives the best measure of vibration ‘severity’. What is the best transducer to use for the measurement of velocity, in terms of frequency range, dynamic range and signal fidelity? Explain your recommendation.
7. Explain how the absolute motion of a shaft supported in fluid film bearings can be measured, and situations where this might be advantageous.
8. Explain what is meant by ‘runout’ in connection with proximity probes, including mechanical runout and electrical runout. Discuss the extent to which runout can be compensated for.
9. Discuss the factors which limit the low frequency range of piezo-electric accelerometers, and what can be done in the design to ameliorate the problems.
10. Discuss the amplitude and phase response of transducers such as accelerometers and velocity probes, and how it affects signal fidelity in terms of waveforms and frequency spectra.
11. Discuss the use of laser Doppler instruments for the measurement of linear and rotational vibrations.
12. Describe how shaft encoders can be used to measure rotational vibrations in terms of angular displacement, velocity and acceleration. Describe two ways of obtaining such signals as a function of rotational angle instead of time, this corresponding to ‘order tracking’.
13. Give some simple examples of how changes in a measured signal can be ascribed to a change in the forcing function or in the signal transmission path.

A.2 Vibration Signals from Machines

A.2.1 Exam Questions

1. Misalignment is often said to produce a strong effect at the second harmonic of the shaft speed. Explain how this can be the case for two types of coupling, viz a Hooke’s (universal) joint and a gear (spline) coupling. How are the torsional vibrations which are generated converted into linear vibrations at the bearings of the machine.
2. For an induction motor (asynchronous), describe how local faults on the stator and rotor affect the vibration signals. For a rotor fault, how would the electrical unbalance distinguish itself from mechanical unbalance? Give formulas for the frequencies of the characteristic features in the signals in the cases of stator and rotor faults.
3. Discuss how the critical speeds of a rotor are related to, and how they differ from, the vibration modes of the non-rotating rotor in lateral bending.
4. Describe the physical reasons for, and the vibration characteristics of, oil whirl, oil whip, hysteresis whirl and dry friction whirl.
5. Explain how periodically varying support stiffness can give rise to exact subharmonic response vibrations, and how this might be related to looseness and/or rubbing of a rotor on the stator. How can this be distinguished from oil whirl?
6. Describe a number of methods by which cracks in rotors can be detected. What is a ‘breathing’ crack, and what influence does it have on the rotor vibrations?
7. Describe what is meant by ‘ghost components’ in gear vibrations, and how they manifest themselves in response vibrations. What are their characteristics and how can they be used diagnostically?

8. Describe why there is usually some vibration at gearmesh frequency even with a gear set having perfect involute profiles. How is the vibration at gearmesh frequency affected by load?
9. How do tooth root cracks and spalls on individual gear teeth affect the vibration signal at the toothmesh frequency?
10. Explain why uniform wear of gear teeth is often first detectable at the second harmonic of gearmesh frequency?
11. With the aid of sketches illustrate the effect of a localised fault on the spectrum of the vibration signal from a gearbox. Give some examples of such localised faults which would manifest themselves in the spectrum in this way, and also of more distributed faults.
12. Rolling element bearing signals often manifest themselves as series of high frequency bursts. Explain the mechanism, and why the series of bursts are often modulated at lower frequencies. Do the bearings of planet gears differ from other bearings in this respect?
13. Explain how and why actual frequencies of bearing faults usually differ from the nominal frequencies given by kinematic formulae. How do cylindrical roller bearings differ from other roller bearings and ball bearings?
14. Figure A.2.1(a, b) shows two FFT spectra of the vibration signal measured on an induction motor that drives a screw compressor. In a screw compressor, air is compressed in passing along the rotor from the inlet to the outlet, but there is still some pulsation at the rate at which the rotor lobes pass the outlet valve. In this compressor, there are two screws, one driven by the other. The directly driven rotor of the compressor has four lobes (effectively helical teeth) and the other rotor has six lobes, so that it is like a four tooth gear meshing with a six tooth gear. The motor has two poles and thus a running speed just less than mains frequency 50 Hz.

From Figure A.2.1(a) describe the features of the spectrum that indicate that there is some transmission of vibration from the compressor back to the motor (related to the meshing of the lobes).

From Figure A.2.1(a) it appears that the second harmonic of shaft speed at about 100 Hz is slightly stronger than the first harmonic. Figure A.2.1(b) is a zoom analysis in a narrow band around 100 Hz to investigate this phenomenon (note the logarithmic amplitude scale of both spectra). From the zoom spectrum, comment on the likelihood of misalignment of the motor shaft, and the type of fault that would give the pattern shown in Figure A.2.1(b).

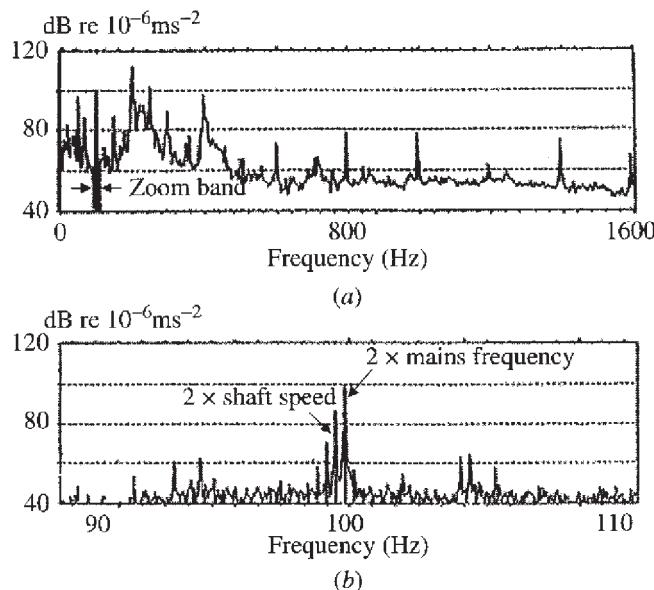


Figure A.2.1 Vibration spectra of an induction motor driving a screw compressor.

A.3 Basic Signal Processing

A.3.1 Tutorial and Exam Questions

A.3.1.1 Fourier Series

- For the periodic signals in Figure A.3.1, find the Fourier series components both in terms of cosine/sine components, Eqs. (3.15)-(3.17) and using the complex version, Eq. (3.21). Different students can be assigned different signals.

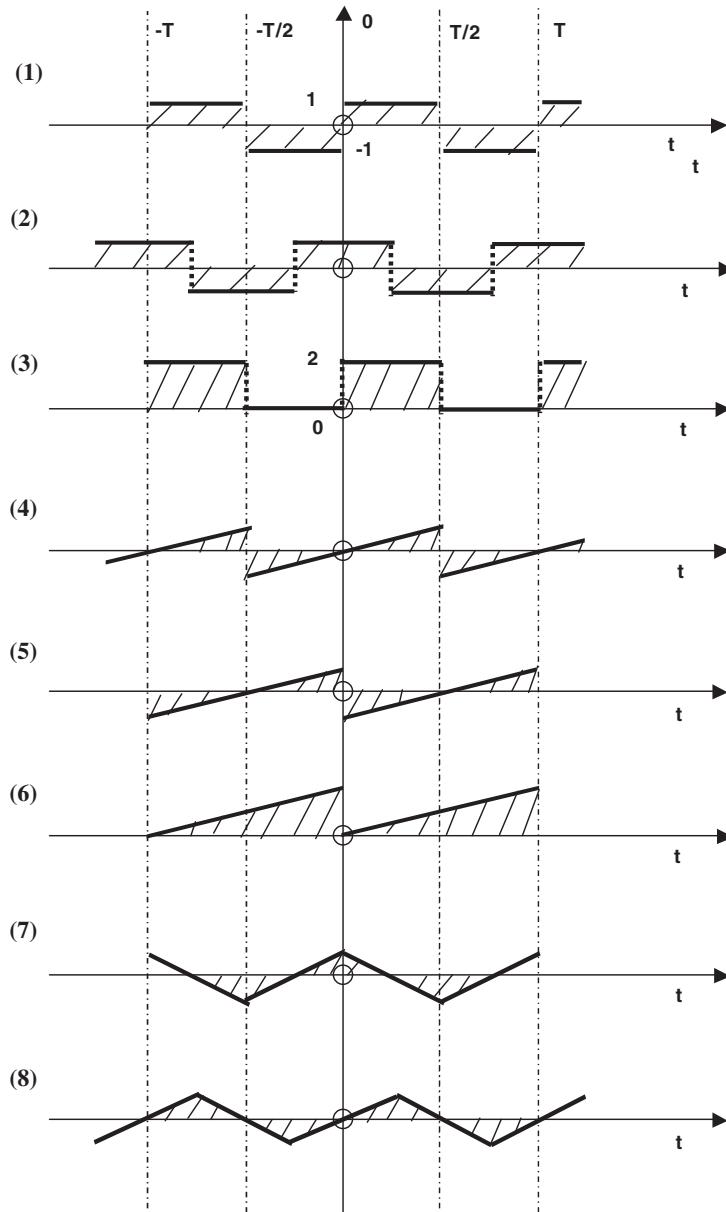
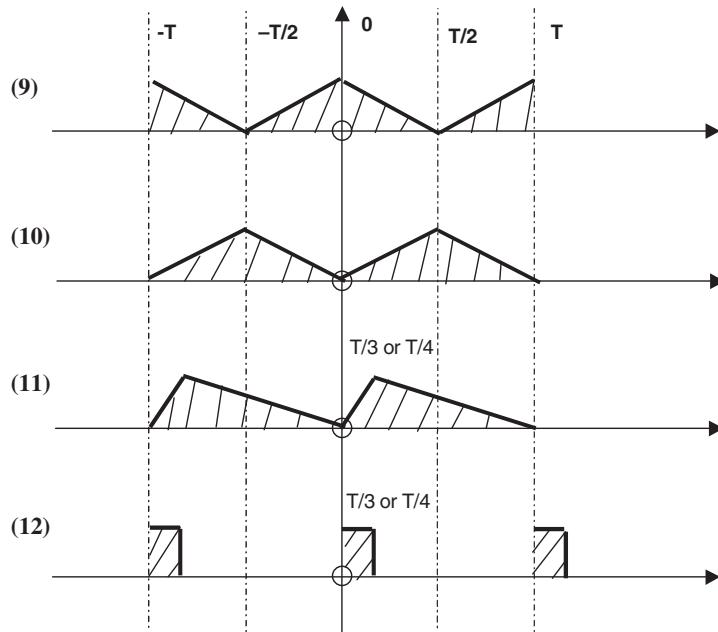
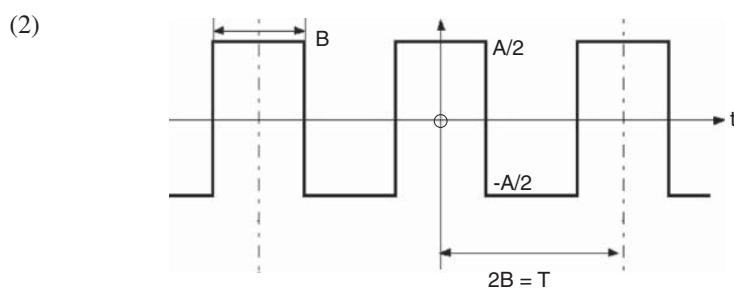
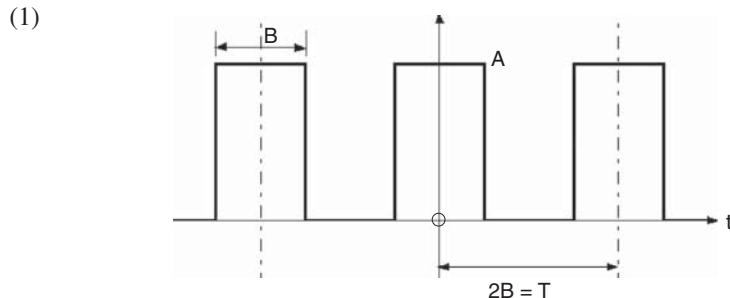


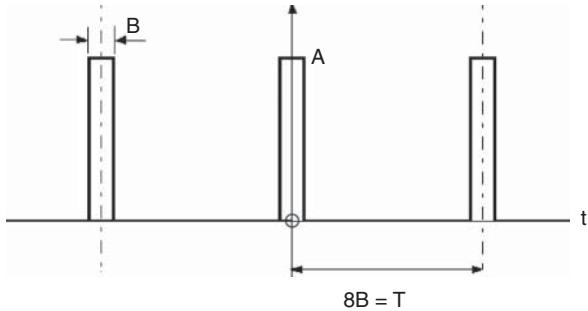
Figure A.3.1 Signals for Fourier series analysis.

**Figure A.3.1 (Continued)**

2. Perform Fourier series expansion of the following periodic signals.



(3)



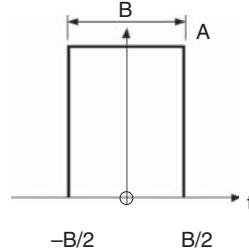
Obtain values for first few harmonics ($k = 0, 1, 2, 3, 4$) and sketch spectra.

Hint: $f_k = kf_1 = k/T$; T = period.

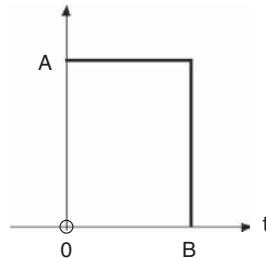
A.3.1.2 Fourier Transforms

3. Find the Fourier transforms of the following functions.

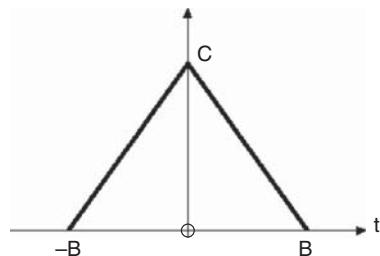
(1)



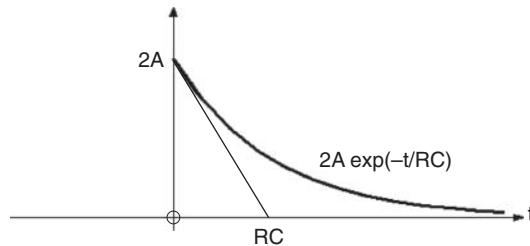
(2)



(3)



(4)



Sketch amplitude and phase functions for Nos. (2) and (4).

A.3.1.3 Convolution Theorem and Other Methods

NB: Linearity property of Fourier transform:

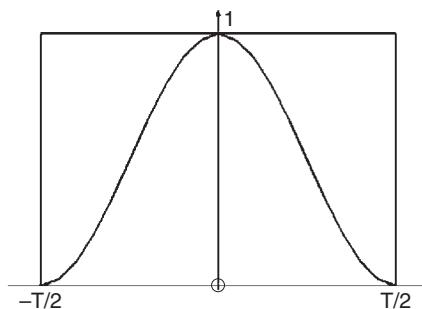
$$\begin{aligned} \text{If } x(t) = x_1(t) + x_2(t) \text{ then } \mathfrak{F}\{x(t)\} &= \mathfrak{F}\{x_1(t)\} + \mathfrak{F}\{x_2(t)\} \\ \text{or } X(f) &= X_1(f) + X_2(f) \end{aligned}$$

Time shift property of Fourier transform:

$$\mathfrak{F}\{x(t - t_0)\} = X(f) \exp(-j2\pi f t_0)$$

This changes the phase of the spectrum, but not the amplitude. It can be understood as the phase angle through which each rotating vector at frequency f would have rotated in time t_0 , and the origin is being moved backwards by this amount.

4. Use the linearity property of the Fourier Transform to generate the result of Question 2. (2) from that of 2. (1).
5. Use the convolution theorem, and the fact that the spectrum of a sequence of unit impulses, spacing T , is a uniform set of harmonics, spacing $1/T$, and with value $1/T$, to derive the result of Question 2. (1) and 2. (3) from that of Question 3. (1).
6. Use the time shift property of the Fourier Transform, to derive the result of Question 3. (2) from that of 3. (1).
7. Use the convolution theorem to derive the result of Question 3. (3) from that of Question 3. (1).
Hint: the convolution of a rectangular function with itself is a triangular function.
8. From the given figure



- (i) By inspection sketch the spectra of the (continuous) raised cosine and rectangular functions shown (use the result of Question 3. (1)).
- (ii) Use the convolution theorem to determine the spectrum of the truncated raised cosine.
- (iii) Use the time shift theorem to determine the spectrum of a Hanning window (obtained by shifting the truncated cosine by an amount $T/2$).
- (iv) In the FFT process, the Hanning window is repeated periodically. Use the method of Question 5. to determine the resulting spectrum values. These could be applied as convolution coefficients to apply the Hanning weighting in the frequency domain rather than the time domain.
9. (a) Given that the Fourier transform of a rectangular pulse of height A and length B , centred on zero time, is given by:

$$\frac{AB \sin(\pi f B)}{(\pi f B)}$$

show that the Fourier series components $G(f_k)$ of the signal $g(t)$ shown in the Figure A.3.2 are given in terms of their real and imaginary components by:

$$\begin{aligned}\text{Re}[G(f_k)] &= \frac{A}{2\pi k} \sin(2\pi kB/T) \\ \text{Im}[G(f_k)] &= \frac{A}{2\pi k} \{1 - \cos(2\pi kB/T)\}\end{aligned}$$

- (b) The speed of a DC motor is controlled by applying a voltage signal of the form shown in Figure A.3.2, where A and T are constant, and B is varied to vary the DC component and thus the mean speed. If $T = 100$ ms, $B = 75$ ms and $A = 100$ V, determine the mean voltage, and that the amplitude of the sinusoidal voltage oscillations at 10 and 20 Hz is given by $\frac{100\sqrt{2}}{\pi}$ V and $\frac{100}{\pi}$ V, respectively.

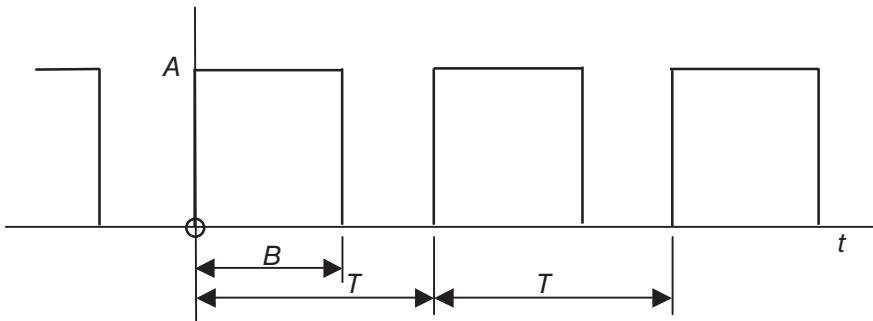


Figure A.3.2

10. Figure A.3.3 shows a simplified model of the torque produced by a turbine wheel with one of its 12 blades missing. Use the convolution theorem and the Fourier transform of a rectangular pulse (as given in Question 9.a) to show that the Fourier series components of the torque have the following values (for zero and positive frequency harmonics):

$$F(0) = 0.917A$$

$$F(1) = -0.0824A$$

$$F(2) = -0.0796A$$

$$F(3) = -0.0750A$$

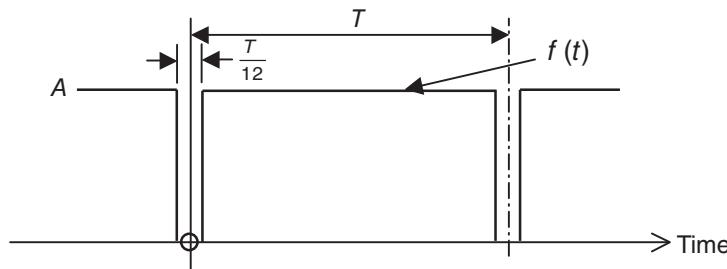


Figure A.3.3

$$F(4) = -0.0689A$$

$$F(5) = -0.0615A$$

$$F(6) = -0.0531A$$

Hint: It may be found convenient to model the signal as a constant value minus a series of short rectangular pulses.

11. (a) For the signal $g_1(t)$ as shown in Figure A.3.4, show by direct integration that its Fourier transform $G_1(\omega) = \frac{A}{\sigma + j\omega}$, where $\omega = 2\pi f$.

A damped sinewave is obtained by multiplying $g_1(t)$, by $\sin(\omega_d t)$ whose Fourier transform may be represented by a pair of delta functions, one of value $-j/2$ at $+\omega_d$ and one of value $+j/2$ at $-\omega_d$. Use the convolution theorem to find an expression for $G_2(\omega)$, convolving $G_1(\omega)$ with each of the delta functions in turn.

What is the value of $G_2(\omega)$ at zero frequency?

- (b) The transfer function for a SDOF system can be expressed as:

$$H(s) = \frac{1}{ms^2 + cs + k} = \frac{1/m}{(s - s_1)(s - s_2)}$$

where s_1 and s_2 are the conjugate poles $-\sigma \pm j\omega_d$.

Obtain the frequency response function $H(\omega)$ by substituting $s = j\omega$ in the equation for $H(s)$, and find the value at zero frequency. Since $H(\omega)$ is the same as $G_2(\omega)$ of part a), use this to express scaling constant A in terms of m and ω_d .

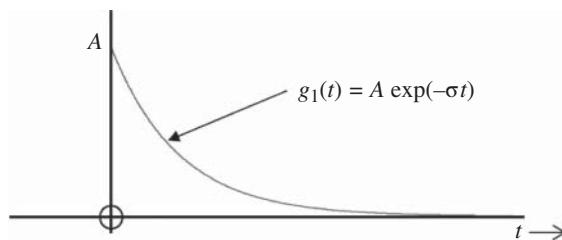


Figure A.3.4

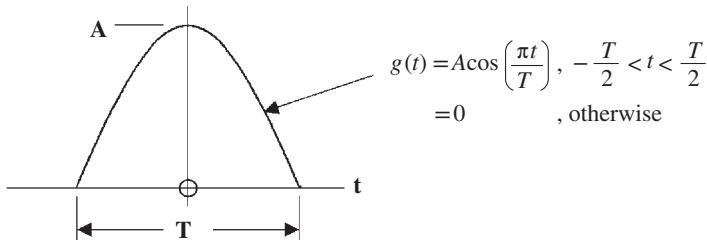


Figure A.3.5

12. (a) The force pulse applied by an instrumented hammer has the half cosine shape illustrated in Figure A.3.5.

Considering this to be the product of a cosine and a rectangular function, use the convolution theorem (with suitable sketches) to show that the spectrum $G(f)$ is given by:

$$G(f) = \frac{AT}{2} \left[\frac{\sin\left(\pi ft - \frac{\pi}{2}\right)}{\left(\pi ft - \frac{\pi}{2}\right)} + \frac{\sin\left(\pi ft + \frac{\pi}{2}\right)}{\left(\pi ft + \frac{\pi}{2}\right)} \right]$$

Sketch this spectrum (noting that it is the sum of two displaced $\frac{\sin x}{x}$ functions), in particular showing:

- (i) The value at zero frequency
 - (ii) The frequency of the first zero crossing in the spectrum
 - (iii) The spacing of the subsequent zeros in the spectrum
- (b) If $T = 2$ ms, $A = 100$ N, determine the energy spectral density (ESD) in $N^2 s/Hz$ at zero frequency and at 400 Hz. What is the dB difference between these two values?

13. (a) Figure A.3.6 shows the idealised torque signal applied by a hydraulic motor. The period of each half cosine pulse is 20 ms.

Using the result from Question 12. (a), determine:

- (i) The average (i.e. zero frequency) torque applied by the motor.
 - (ii) The values of the first four harmonics of the fundamental repetition frequency (stating their frequencies) expressed as Nm RMS (i.e. including the negative frequency contributions).
- (b) Show with sketches how you could determine the Fourier series spectrum of the signal if one of the five pistons of the motor were not functioning, i.e. if every fifth half cosine pulse were removed. What is the fundamental frequency of this new signal? Determine the new

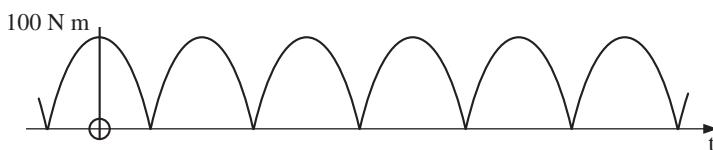


Figure A.3.6

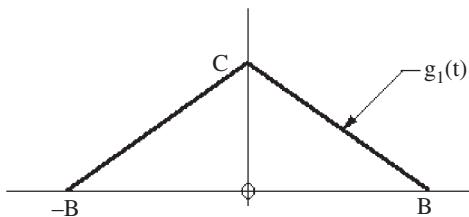


Figure A.3.7

torque values at the first and second harmonics of the old fundamental frequency (i.e. the fifth and tenth harmonics of the new fundamental frequency).

14. (a) Figure A.3.7 shows a triangular function of length $2B$ centred on time zero.

Using suitable sketches show how this can be generated by convolving two rectangular functions of length B and height A . From this derivation express the height of the triangular function C in terms of A and B . Using the convolution theorem, show that the Fourier transform of the triangular function is given by the expression:

$$G_1(f) = CB \frac{\sin^2(\pi f B)}{(\pi f B)^2}$$

- (b) Figure A.3.8(a) shows a sampled time function, and Figure A.3.8(b) a ‘sample and hold’ version of it produced by a DA converter (shifted to centre each step on the samples).

It can be modelled as the result of a convolution of the sampled signal with a rectangular function of length Δt (the sample spacing), and unit height. Figure A.3.8(c) shows a version produced by linear interpolation between the samples. It can be modelled as the result of a convolution of the sampled signal with a triangular function (of the form of Figure A.3.7) of unit height, where again $B = \Delta t$. The convolution in the time domain results in a low-pass filtration in the frequency domain by (multiplication by) the Fourier transform of the convolving function. Keeping in mind that the sampling frequency $f_s = 1/\Delta t$, use the formulas for the Fourier transforms of the rectangular and triangular functions to determine the attenuation in dB of the two lowpass filters at 40% of the sampling frequency, which is the highest valid frequency in the sampled signal.

Hint: The first zero crossing in both Fourier transforms occurs at a frequency equal to $1/B$, where $B = \Delta t$.

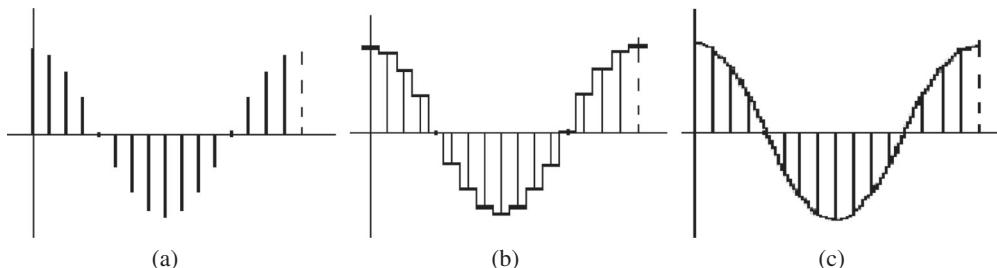


Figure A.3.8

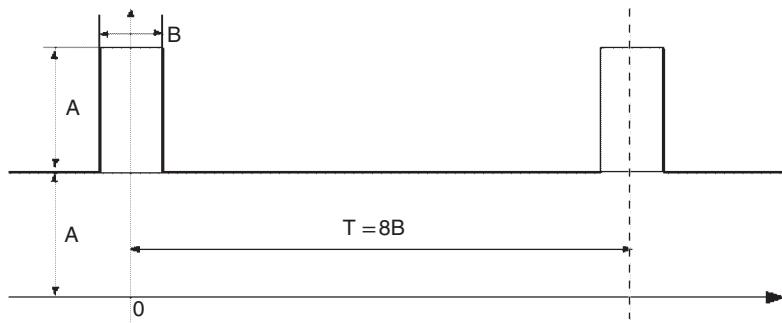


Figure A.3.9

15. (a) Given that the Fourier transform of a rectangular pulse of length B and height A , centred on zero time, is:

$$\frac{AB \sin(\pi f B)}{(\pi f B)}$$

make use of the convolution theorem to derive the Fourier series spectrum of the periodic time signal obtained by repeating such a pulse with period $8B$. Next make use of the linearity property of the Fourier transform to derive the Fourier series spectrum of the signal $g_1(t)$ shown in Figure A.3.9. Sketch the two-sided spectrum over a range encompassing at least the first 16 harmonics.

- (b) An idealised model of the vibration signal produced by a local defect on one tooth of an eight tooth gear is shown in Figure A.3.10.

Considering the carrier component to be a distorted sine wave with three (sine) harmonics of decreasing amplitude, sketch the spectrum of the uniform carrier component (remember that the sine function has a purely imaginary spectrum and can thus be represented in the imaginary plane) and then make use of the convolution theorem and the results of part (a) (up to the first zero crossing) to sketch the (two-sided) spectrum of the signal in Figure A.3.10.

Finally, sketch the corresponding one-sided amplitude spectrum.

16. Figure A.3.11 represents the impulse response function $h(t)$ of an RC averaging circuit, whose Fourier transform $H(f)$ is the frequency response function (FRF).

Find $H(f)$ by direct application of the Fourier transform. It is effectively a low-pass filter, which transmits the zero frequency (or average) value, and attenuates higher frequency ripple components.

- (a) What value of K is required so that the amplitude of the DC (zero frequency) component is unchanged (i.e. unit zero frequency gain)?

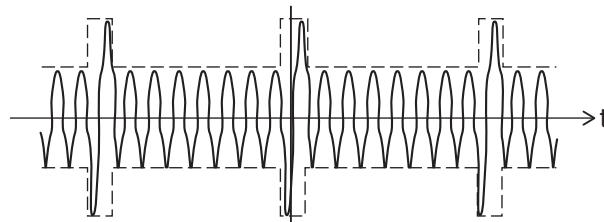
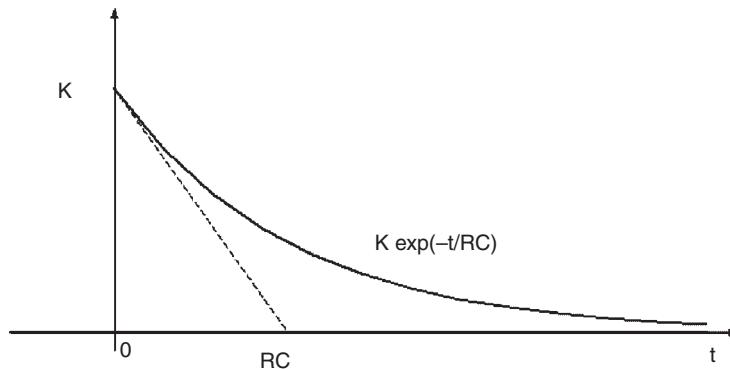


Figure A.3.10

**Figure A.3.11**

- (b) With this value of K show that the amplitude and phase characteristics are given by:

$$|H(f)| = \frac{1}{\sqrt{1 + (2\pi RCf)^2}} \quad (\text{A.1})$$

$$\angle H(f) = -\tan^{-1}(2\pi RCf) \quad (\text{A.2})$$

- (c) The impulse response function of a running linear averager is a rectangular function of height A and length (i.e. averaging time) T . The amplitude of its FRF is given by:

$$|G(f)| = \left| \frac{AT \sin(\pi fT)}{(\pi fT)} \right|$$

What value of A is required to give unit zero frequency gain?

- (d) These two averaging circuits are deemed to be equivalent for noise signals when their ‘noise bandwidth’ is the same. Find the relationship between T and RC for this to be the case, which is when the integral over all frequency of the squared amplitude is the same (for unit zero frequency gain). You may use the standard results:

$$\int_{-\infty}^{\infty} \frac{\sin^2 x}{x^2} dx = \int_{-\infty}^{\infty} \frac{dx}{1+x^2} = \pi$$

17. (a) Figure A.3.12 shows how a half wave rectified cosine may be modelled as a periodic repetition of a half cosine pulse obtained by convolution with a pulse train of period $2B$.

Use the convolution theorem to derive the Fourier series components of such a half wave rectified signal, expressing them in terms of k , the harmonic number. You may use the result of Question 12. (a) for the Fourier transform of a half cosine pulse.

- (b) If the signal is obtained by half wave rectifying a 240 V (RMS), 50 Hz mains signal, calculate the value of the zero frequency component and the RMS value of the components at 50 and 500 Hz.
- (c) If the signal from (b) is passed through an RC-averaging circuit (as in Question 16.) with RC time constant 0.125s, use Eqs. (A.1, A.2) of Question 16. to calculate the amplitude and phase of the transmitted components at zero and 50 Hz.

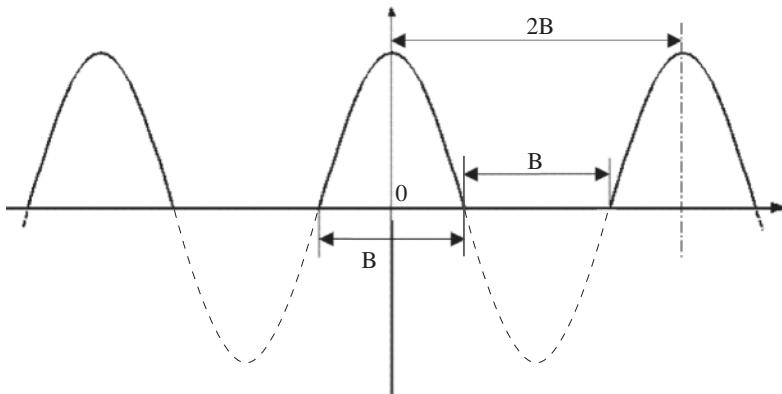


Figure A.3.12

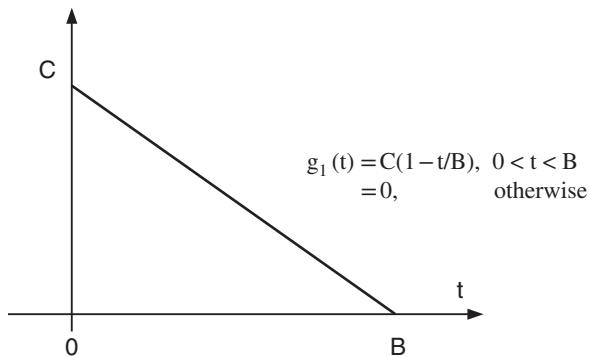


Figure A.3.13

18. (a) Find by direct integration, the Fourier transform $G_1(f)$ of the time function $g_1(t)$ shown in Figure A.3.13. You may use the standard integral:

$$\int te^{at} dt = \frac{e^{at}}{a} \left(t - \frac{1}{a} \right) + \text{const}$$

- (b) If $C = 10 \text{ V}$, $B = 1 \text{ ms}$, calculate the ESD at 50 Hz , in $\text{V}^2/\text{s}/\text{Hz}$

If $g_2(t)$ is the trapezoidal function shown in Figure A.3.14, make use of the properties of the Fourier transform to show that $G_2(f)$, the Fourier transform of $g_2(t)$, may be written in terms of $G_1(f)$ from part a) as follows:

$$G_2(f) = G_1(f)e^{-j\pi f(A-B)} + G_1(-f)e^{j\pi f(A-B)} + \frac{C(A-B) \sin \pi f(A-B)}{\pi f(A-B)}$$

- (c) Use suitable sketches to show that the time function $g_2(t)$ of part b) can alternatively be obtained by convolving the two rectangular functions shown in Figure A.3.15.

What is the value of scaling factor C in terms of A , B , and D ?

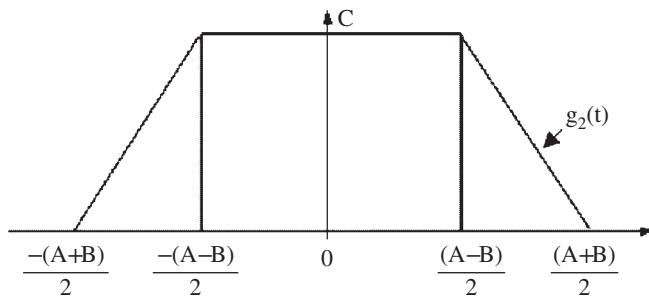


Figure A.3.14

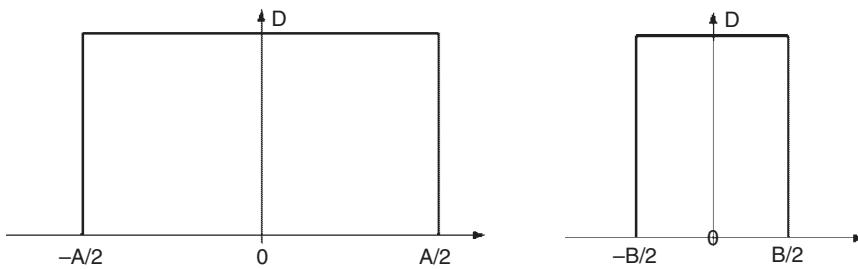


Figure A.3.15

Hence, use the convolution theorem to show that $G_2(f)$ may alternatively be expressed as:

$$G_2(f) = \frac{C \sin(\pi f A) \sin(\pi f B)}{B \pi^2 f^2}$$

19. Figure A.3.16 shows a schematic diagram of the use of adaptive noise cancellation (ANC) for separating a faulty bearing signal from a background gear signal. Give a simple explanation of how this works, and how the measurement points for the Primary and Reference inputs would be selected.

In a situation such as a planetary bearing, where the conditions for selecting a suitable reference input do not exist, explain how the basic ANC algorithm can be converted into self-adaptive

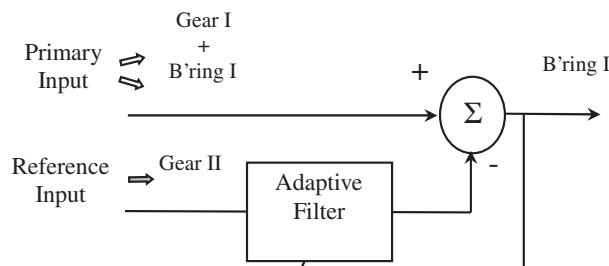


Figure A.3.16

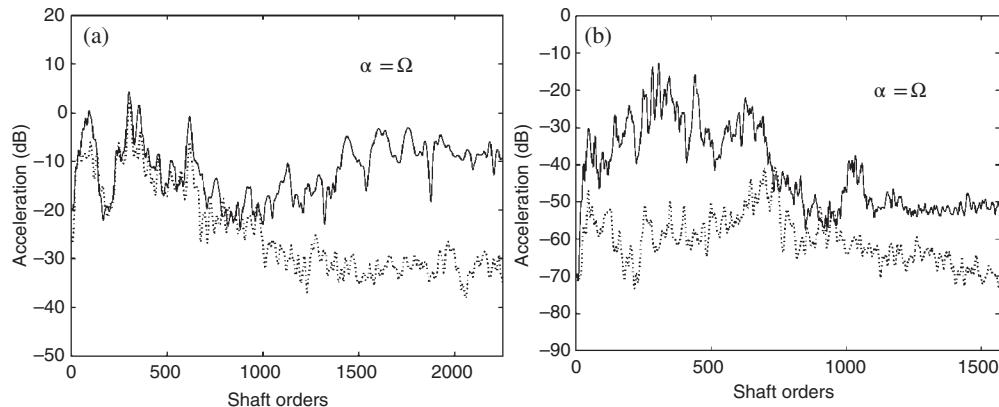


Figure A.3.17

noise cancellation (SANC), using only one signal, and explain at the same time how this works to separate gear and bearing signals.

20. Figure A.3.17 shows the spectral correlation for two bearings with faults, calculated for cyclic frequency α equal to shaft speed Ω . Before analysis, the signals were subjected to self-adaptive noise cancellation to remove discrete frequency components. Note that the frequency axis is expressed in terms of shaft orders. The bearings were mounted in a gearbox for which significant garmesh harmonics extended up to about 500 orders.

Explain, giving reasons, what are the types of faults indicated in Figure A.3.17(a) and (b), and which of them is localised and which is an extended fault. Which is most likely to be detected in an envelope analysis? Explain why.

A.4 Fault Detection

A.4.1 Tutorial and Exam Questions

1. Discuss the reasons why velocity is the best parameter to indicate vibration severity, in particular when a total RMS value over a frequency band is used as an indicator. How is this affected when comparing vibration spectra?
2. Explain the problems involved in digital comparison of vibration spectra. Your discussion should include the sampling of discrete frequency peaks, and the possibility of even minor speed variations. How can these problems be overcome?
3. Discuss the frequency range that can be covered by an FFT spectrum. Can such a spectrum be adjusted for speed variation, and if so, how? What frequency range should typically be monitored for fault detection on typical rotating machines? Give examples.
4. Figure A.4.1 shows the CPB (constant percentage bandwidth) spectrum of the vibration signal from an auxiliary gearbox in September, 1981, as well as the difference spectrum in dB obtained by comparison with a mask formed from a reference spectrum. The cursor line at 609 Hz is placed on the frequency component which exhibits the biggest change from the reference condition (16 dB according to the dB difference scale on the right).

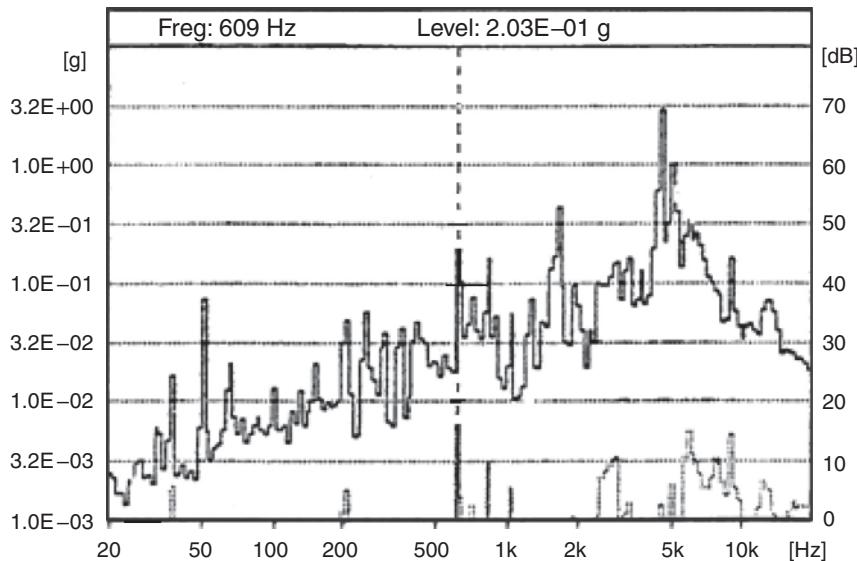


Figure A.4.1 (upper) New CPB spectrum, September 81. (lower) dB difference spectrum from reference mask.

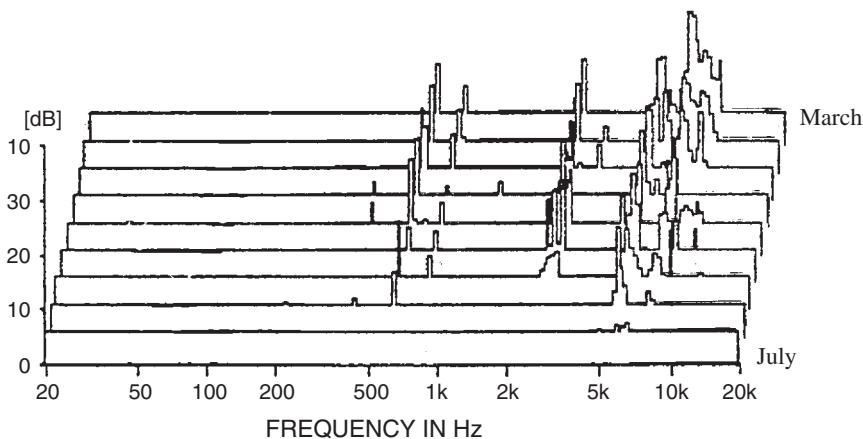


Figure A.4.2 Waterfall plot of a series of difference spectra ranging from July, 1981, when the first significant change was detected, at one month intervals until March, 1982. The baseline of each of these spectra is set at the tolerance of 6 dB, so that only components above this value are seen.

Figure A.4.2 shows a series of such difference spectra, ranging from July, 1981, when a significant difference was first detected, at one month intervals until March the following year. Thus the difference spectrum of Figure A.4.1 is the third from the front.

Figure A.4.3 shows spectra on linear frequency axes from before and after the fault developed, and the corresponding cepstra. The lines at 20 ms in the latter (marked 'RPM') represent the 50 Hz speed of the shaft involved in the problem.

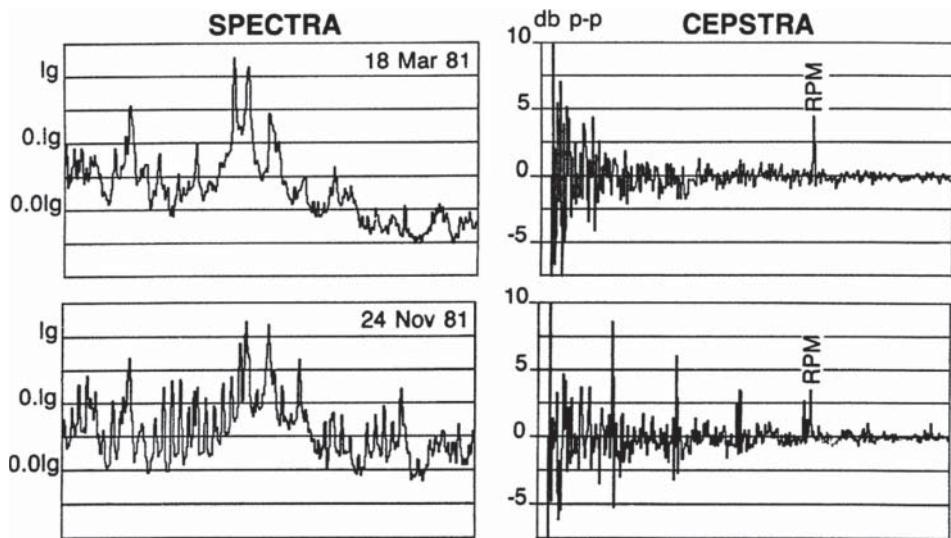


Figure A.4.3 Spectra and cepstra on linear x-axes for the same case before the fault developed (March '81) and after it developed (November, '81).

- (a) From the spectra in Figure A.4.1, comment on the degree of severity of the fault in the gearbox.
 - (b) From the spectral and cepstral information in Figure A.4.3, determine the most likely fault which has caused these changes, giving reasons. The bearings on the shaft of speed 50 Hz have 10 balls, with a ratio of ball diameter to pitch diameter of 0.17 and a fully radial load.
 - (c) From the trend plot of Figure A.4.2, comment on the decision to leave the machine running for the period shown (it was repaired in April 1982). You might like to take into account the diagnostic information of part (b).
5. Figure A.4.4 shows a diagram of torsional vibration (angular velocity) of a spark ignition engine with a fault. The numbers at the top of the figure are placed after the ignition time for each cylinder, and indicate the order. Explain the appearance of the angular velocity diagram, indicating the general type of fault that would cause this.

Figure A.4.5 is for a different condition and also shows some lack of uniformity in the angular velocity diagram. Give a possible explanation for this.

For training neural networks to recognise patterns such as those in the two figures, the number of samples per engine cycle should always be the same. Describe two procedures by which the data could be obtained in this form from a shaft encoder signal.

6. Figure A.4.6 shows typical angular velocity diagrams for a four-cylinder spark ignition engine operating with two different faults.

By inspection, indicate the general type of fault present in each case, and with which cylinder it is associated. Suggest a possible explanation for the difference in detail between the two signals. In this regard, pay particular attention to the section sloping down to the right between numbers 1 and 3.

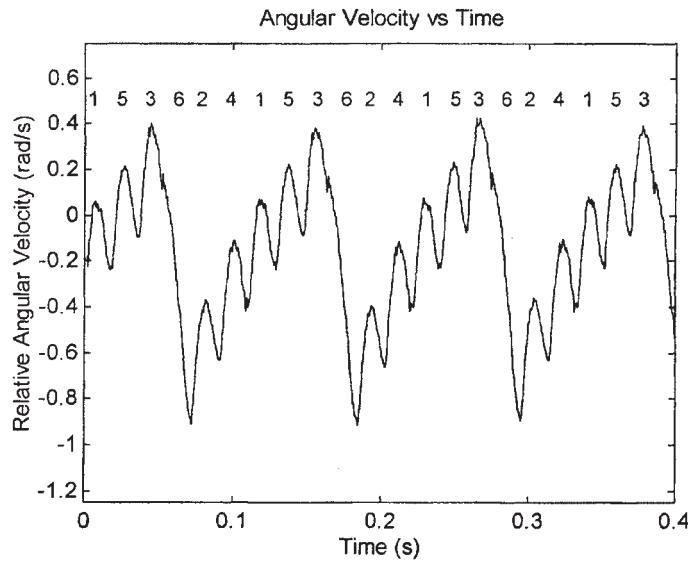


Figure A.4.4

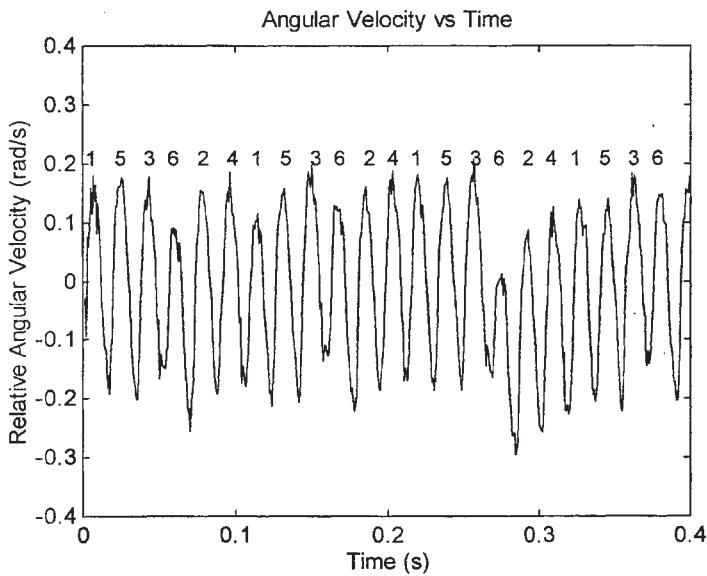


Figure A.4.5

Describe a technique that could be used to obtain the angular velocity diagrams in Figure A.4.6, and in particular how to avoid amplifying high frequency noise in the differentiation from angular displacement to angular velocity. What transducers are required, and how important is their amplitude calibration?

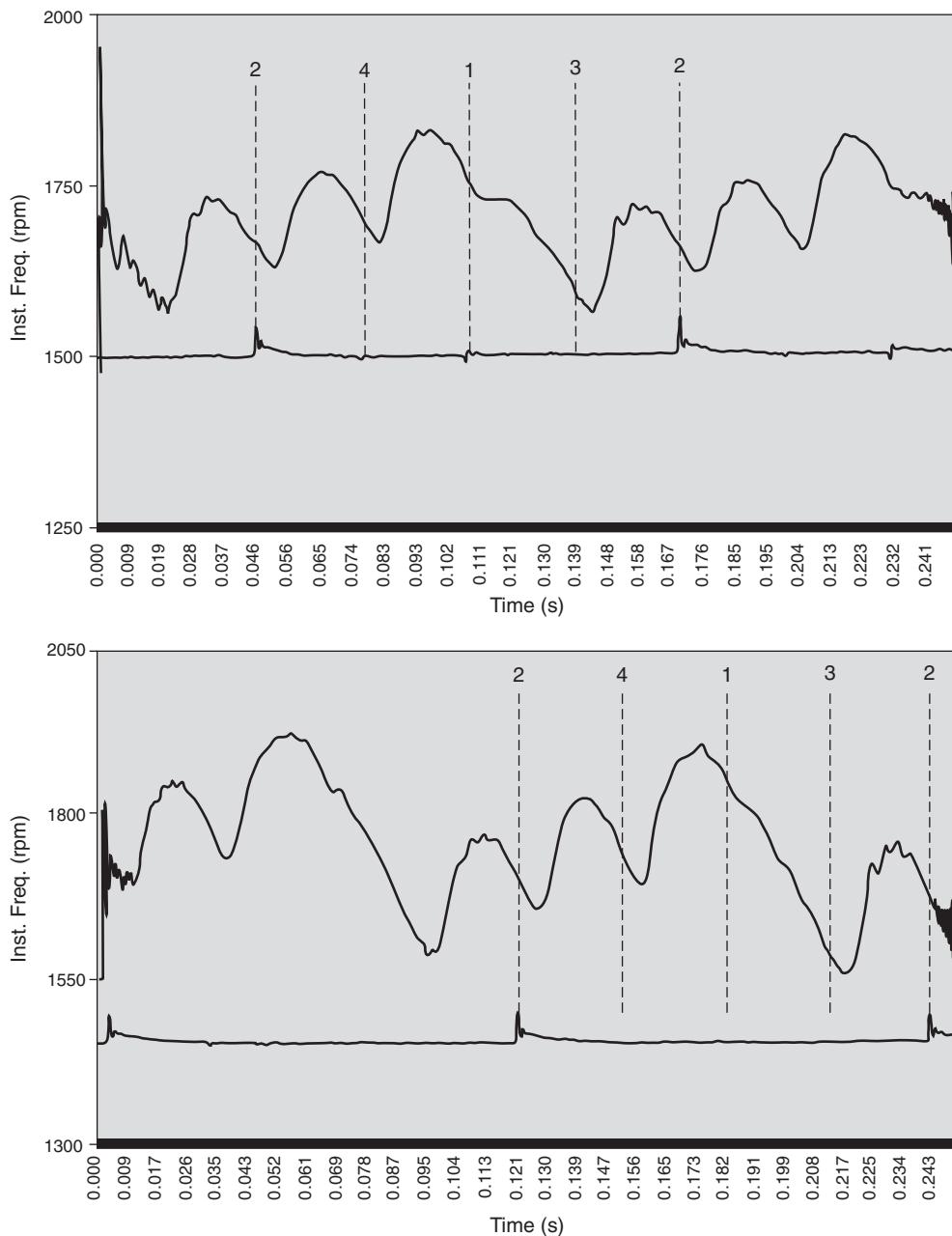


Figure A.4.6 Crankshaft angular velocity vs time for two conditions of a spark ignition engine. The lower curve in each figure is a spark timing signal for cylinder 2. The vertical dotted lines indicate firing on the other cylinders, indicated by their number.

A.4.2 Assignment

The signals to be used for this assignment are stored as .mat files on the website:

(<http://www.wiley.com/go/randall>)

The first series of files are the seven columns of the file vdB.mat and are Matlab® data files representing 1/18-octave (4% bandwidth) spectra expressed as dB levels re 1e-3 mm/s. They were recorded on a gearbox at what can be taken as one month intervals. The corresponding log frequency scale (in Hz) is in the file logf.mat. The spectra can be plotted using the form semilogx(logf, vdB1), where vdB1 is the first spectrum from the first column of vdB.

The second series of files are the seven columns of the file timesig.mat and are 8K (8192 sample) time records corresponding in the same order to the CPB (constant percentage bandwidth) spectra. They are of velocity in mm/s, and the sampling frequency is 2048 Hz. Linear frequency spectra can be produced from the time records using FFT analysis. In fact, averaged spectra will be produced using the Matlab function PWELCH which uses overlap averaging of spectra of shorter records. It is suggested that you use a transform size of 2K (2048) samples, Hanning window, and 50% overlap of successive records. PSD produces an amplitude squared spectrum from zero to the Nyquist frequency, and outputs the corresponding frequency scale as well. These can be converted to dB amplitude scale (over 100 dB dynamic range) for easier comparison with the CPB spectra. The 100 dB level should correspond to the nearest round number greater than the largest value in all the spectra, and can be set by dividing all spectra by this value, multiplying by 10^{10} and taking $10 \log_{10}$ of the result (remembering that the PSD spectra are of amplitude squared). Set any negative values to zero to limit the dynamic range.

The idea of the assignment is to detect significant changes (representing possible faults) by comparison of the CPB spectra, and to learn of the difficulties associated with attempting the same thing with the FFT spectra.

- (a) Make a mask spectrum from the first 4% spectrum by smearing one step to each side as described in Fig. 4.5 (Chapter 4).
- (b) Compare each of the later spectra with the mask, plotting the two on the same graph, and also the spectrum of positive differences (i.e. new – old). Insert suitable texts and scales on the axes, the frequency scale corresponding to the reference spectrum.
- (c) Before making the comparisons in b), the new spectra should be shifted laterally an appropriate number of 4% bandwidths, depending on the new speed. The speed for each case is to be determined from the corresponding PSD spectrum.
- (d) In interpreting the PSD spectra, use the fact that one of the shafts in the gearbox had a speed of approx. 50 Hz, and the other approx. 120 Hz. Since the frequency range of the PSD spectra is 1024 Hz (of which up to 800 Hz is valid), you will obtain a more accurate result by locating higher harmonics of these shaft speeds and dividing their measured frequency by the harmonic order number. You should plot a ‘harmonic cursor’, with spacing equal to the determined harmonic spacing, on each spectrum to demonstrate that the speed has been determined correctly.
- (e) The resulting six difference spectra should be plotted out in 3-dimensional form to see if there are any trends in the fault development. Make comments on these trends. A tutorial on 3-D plots can be found on the website.
- (f) Even though the PSD spectra cover a much more restricted frequency range, try to make direct digital spectrum comparisons (i.e. subtract the dB values of the first spectrum as reference) of the PSD spectra for the two cases of minimum and maximum difference in condition as determined from the CPB spectra. Comment on the results. Can you suggest a way of making a valid comparison of these linear frequency spectra?
- (g) Copies of your Matlab programs should be included in an appendix.

A.6 Cepstrum Analysis Applied to Machine Diagnostics

A.6.1 Tutorial and Exam Questions

The real or power cepstrum is normally defined as the inverse Fourier transform of the logarithmic power spectrum. Since the power spectrum is a real, even function of frequency, the cepstrum would thus be a real, even function of quefrency (i.e. time). In order to obtain better resolution in the spectrum, and limit the range over which the cepstrum is calculated, so as to emphasise effects concentrated in this frequency range, cepstra are sometimes calculated from zoom spectra over a restricted frequency band. An example is given in Figure A.6.1 of such a case, where cepstra have been calculated from two slightly displaced zoom bands of the same basic spectrum. Note that even though the zoom spectra are very similar, the corresponding cepstra show considerable differences.

Explain these differences, and discuss the difficulties of estimating sideband spacings from them (the same results should come from both cepstra, since they are from the same signal). Propose a modified definition of the cepstrum, from which the sideband spacings could be obtained more easily, and from which the same results should be obtained for the two cases.

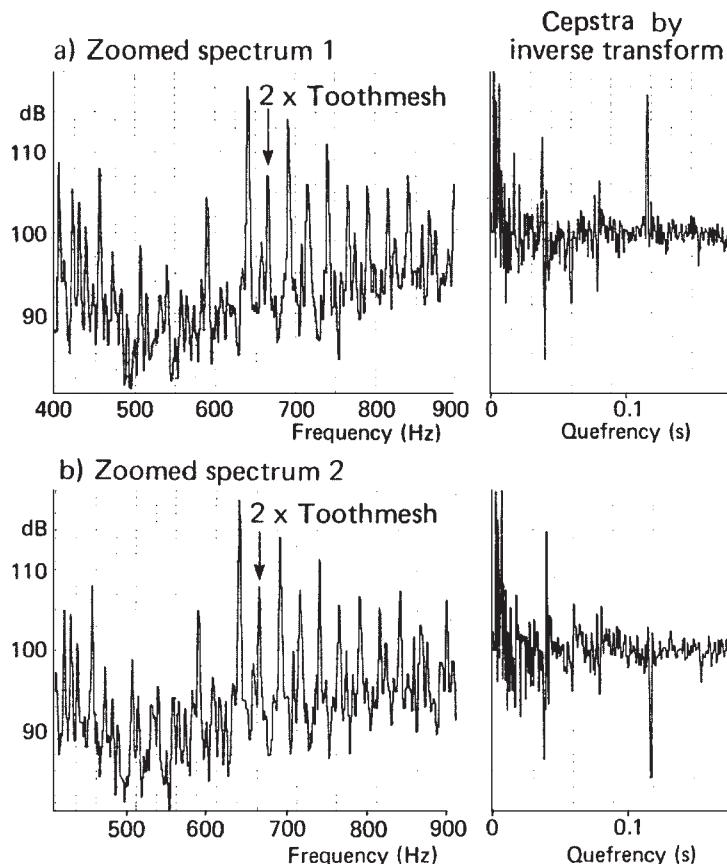


Figure A.6.1 Cepstra obtained from two zoom bands of the same spectrum.

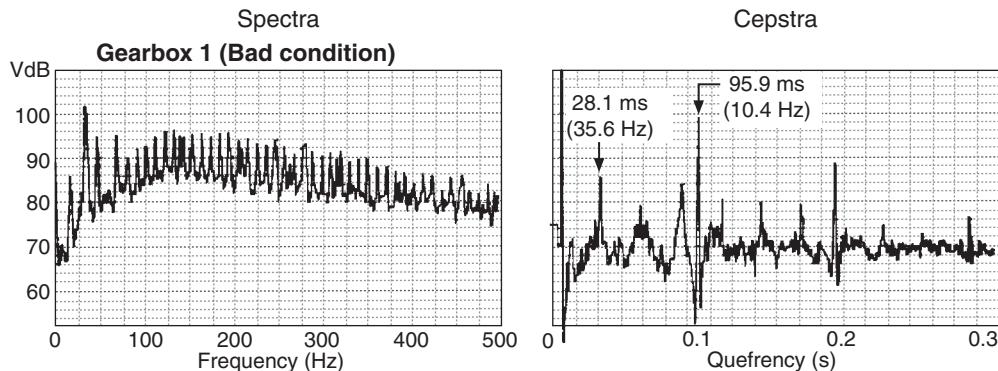


Figure A.7.1 Spectrum and cepstrum for a gearbox with a fault on test in the manufacturing plant.

A.7 Diagnostic Techniques for Particular Applications

A.7.1 Tutorial and Exam Questions

1. Figure A.7.1 shows the spectrum and corresponding cepstrum for a gearbox on test in the manufacturing plant. It has a fault which shows up in both the spectrum and cepstrum. In the latter it manifests itself as a series of rahmonics with a spacing of 95.9 ms (equivalent to 10.4 Hz). The input shaft speed component corresponding to 35.6 Hz is also shown. All values shown are accurate to the three figure accuracy shown. Even though first gear is engaged, it is only lightly loaded, and all gears are in mesh, though only first gear is connected to the output. The gear ratios of input shaft speed to gear speed (to four figure accuracy) for the various gears are 6.593, 3.423, 1.851, and 1.000 for gears one to four, respectively.
 - (a) Taking account of the accuracy of the figures given, suggest the most likely location of the fault in the gearbox under test, giving reasons.
 - (b) From the spectral and cepstral information given, state whether the fault is most likely to be localised or distributed around the gear, giving reasons.
 - (c) Explain how the cepstrum can be useful more generally in all three main phases of condition monitoring, namely detection, diagnosis, and prognosis (life prediction through trend analysis).
2. Figure A.7.2 compares spectra and cepstra from a gearbox driving a cement mill before and after a repair made after eight years of operation. Since the repair consisted largely in reversing the gears so that the unused flanks were now in mesh, and since both flanks were cut on the same machine at the same time, it could be assumed that the original spectrum when the gears were new would have been similar to that after repair. Making that assumption, comment on changes in the spectrum that indicate uniform wear of the gears. Using the spectrum and cepstrum, comment on non-uniform wear and other damage of the gears. If the pinion speed was 8.3 Hz, comment on the distribution of the non-uniform wear. Comment on the usefulness of the cepstrum vs the spectrum in diagnosing respectively uniform and non-uniform wear.
3. Figure A.7.3 shows the envelope spectrum of a signal from a rail vehicle bearing which has been overloaded on a test rig so that faults have developed. The speed of the shaft in the test rig was approximately 565 rpm at the time of recording. Note that calculated bearing frequencies may be in error by a few percent because of slip. The bearing parameters are as follows:

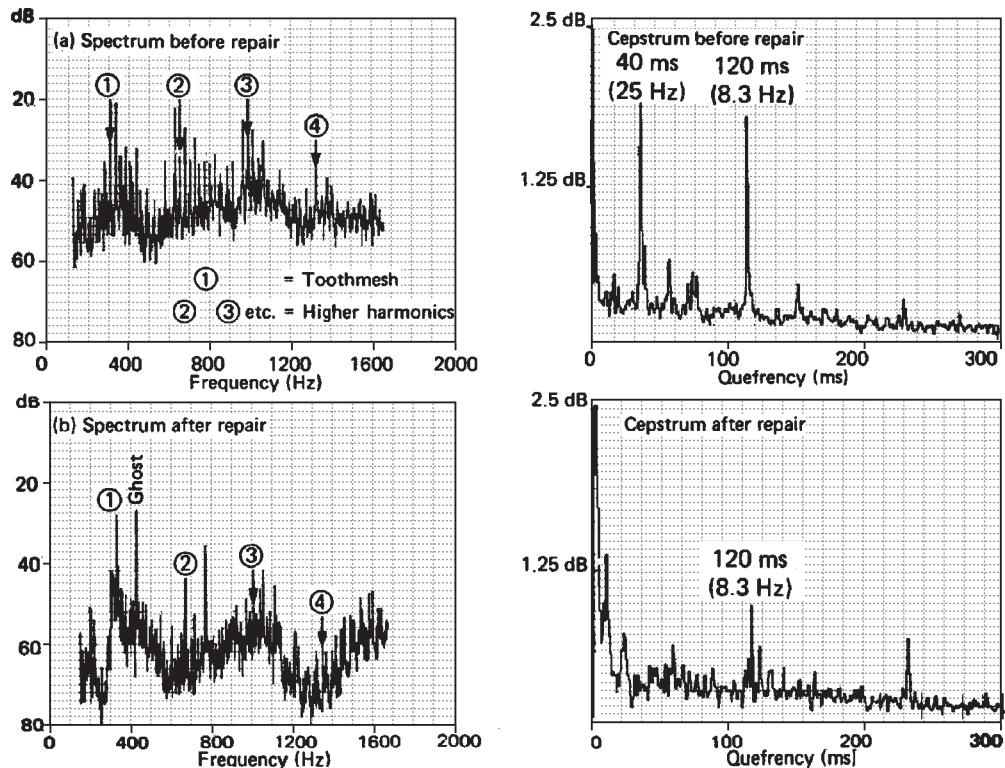


Figure A.7.2

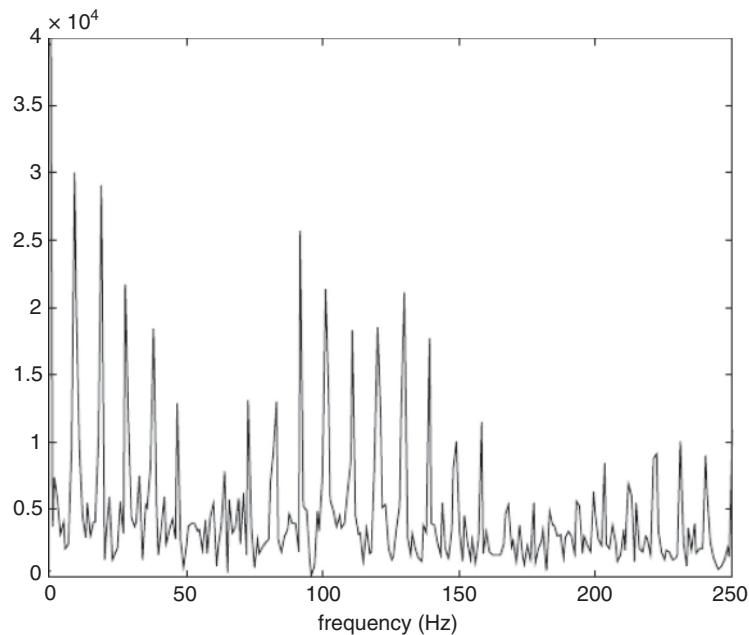


Figure A.7.3

Roller diameter	$d = 21.4 \text{ mm}$
Pitch circle diameter	$D = 203 \text{ mm}$
No. of rolling elements	$n = 23$
Contact angle	$\phi = 9.0^\circ$

4. Calculate the likely bearing fault frequencies for this bearing, and by scaling from the figure, determine the most likely fault, giving reasons.

Describe how such an envelope analysis could be carried out using Matlab®, on a reasonably long digitised signal, containing frequencies up to several kHz.

5. Figure A.7.4. shows an envelope spectrum obtained from a test rig, with the following bearing properties:

Ball diameter	$d = 7.12 \text{ mm}$
Pitch circle diameter	$D = 38.5 \text{ mm}$
No. of rolling elements	$n = 12$
Contact angle	$\phi = 0^\circ$

6. The envelope spectrum is typical of a ball fault, with both odd and even harmonics of ball spin frequency (BSF), surrounded by sidebands with a spacing equal to the rate at which the ball fault passes through the load zone. By scaling from the graph, using only the sideband spacing,

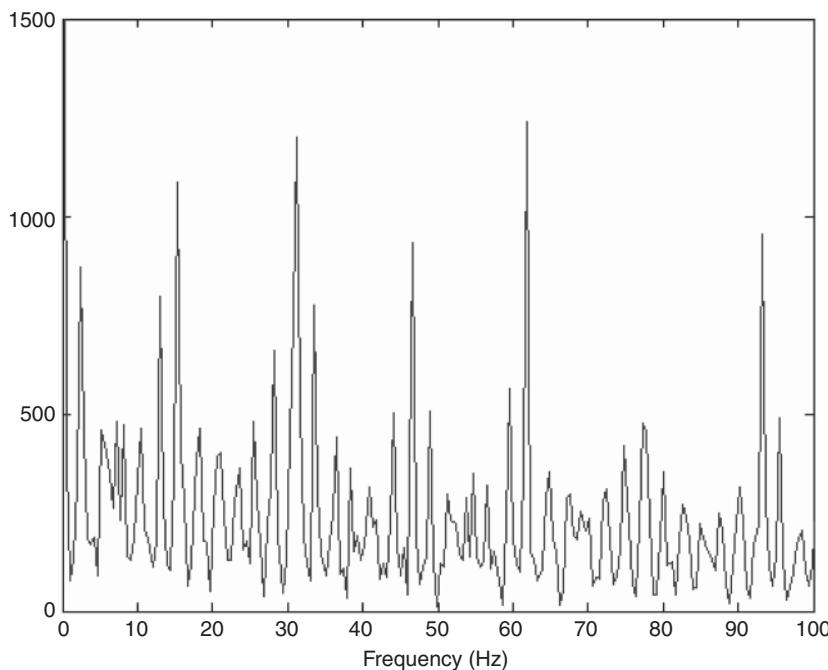


Figure A.7.4. Envelope spectrum from a bearing fault.

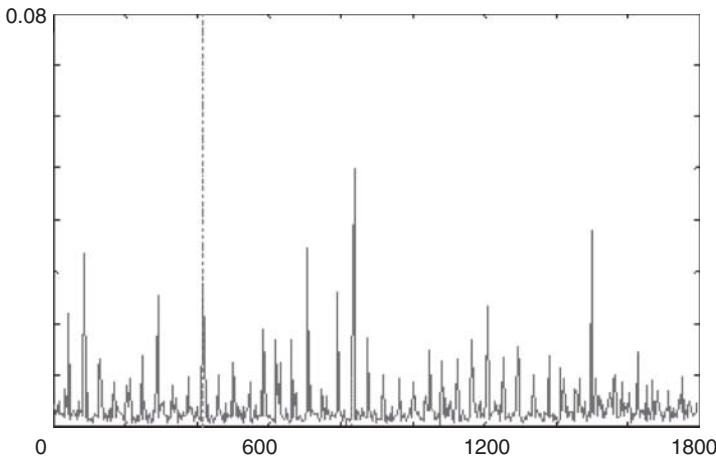


Figure A.7.5 Envelope spectrum from a helicopter gearbox.

estimate the fundamental shaft frequency (not necessarily equal to the sideband spacing). Indicate clearly the range(s) over which the sideband spacings have been determined.

Using this estimate of shaft speed, calculate the corresponding value of BSF. Locate the actual harmonics of BSF in the spectrum, and use the highest harmonic to obtain a more accurate estimate of the shaft speed and the sideband spacing. State which harmonic has been used to obtain the more accurate estimate.

7. Figure A.7.5 shows an envelope spectrum obtained from a helicopter gearbox, with the following bearing properties:

Shaft speed: 95.8 Hz

No. of rollers: 23

Ratio d/D : 0.11

Load angle ϕ : 15°

A cursor indicates a dominant fault frequency. By scaling from the graph, estimate the fundamental frequency, and the spacing of the other important harmonics and sidebands in the spectrum. Indicate clearly the range over which the spacings have been determined. From these results, determine the nature of the fault, explaining all the features leading to this conclusion.

The second harmonic of the fundamental frequency indicated by the cursor is larger than the first. Explain the most likely reason for this.

A.7.2 Assignments

1. Bearing Diagnostics

This assignment involves the diagnosis of rolling element bearing faults using digital envelope analysis.

The signals to be analysed are time history signals from bearings on a gear test rig. They are denoted `goodbear.mat`, `fault1.mat`, `fault2.mat` and `fault3.mat` for the good bearing and for the bearings with faults (either outer race or inner race or rolling element). The Matlab files contain two columns with different signals. Channel 1 is from an accelerometer near the bearing, Channel 2 from a once-per-rev tacho. The acceleration signals were scaled in ms^{-2} at the time

of downloading. The speed of the shaft in the test rig was approximately 6 Hz for all signals, but should be identified by analysing the tacho signal (use the mean time between pulses rather than a frequency spectrum). Note that calculated bearing frequencies may be in error by a few percent because of slip. The sampling frequency of the signals is 48 000 samples/sec, and the record length is 100 000 samples for each record. The bearing parameters are as follows:

Ball diameter	$d = 7.12 \text{ mm}$
Pitch circle diameter	$D = 38.5 \text{ mm}$
No. of rolling elements	$n = 12$
Contact angle	$\phi = 0^\circ$

This part of the assignment is to determine the fault(s) in the faulty bearings using the following procedure:

- (1) First compare the baseband spectrum of the faulty bearing signal to the reference condition. This can be done using the PWELCH command, which forms an average power spectrum from overlapped time records. It is suggested to use a 1024 point transform with Hanning (or Hamming) window and 50% overlap. The comparison can be done on both linear and logarithmic (dB) amplitude scales, to see what difference this makes.
- (2) Demodulate an appropriate band to form the envelope signal for the envelope analysis. This should be a band where a large change occurred with respect to the reference spectrum. The demodulation is carried out on the complex spectrum obtained by transforming the whole time record, not on the PSD spectrum, which has reduced resolution and no phase information. You should compare the results from a couple of different bands. The width of the band to be demodulated should be chosen as at least equal to the required bandwidth of the final envelope spectrum. The latter should contain at least three harmonics of the highest potential bearing fault frequency (i.e. > 3.5 times this spacing).
- (3) The band to be demodulated should be placed in a new buffer with the lowest frequency in the band at zero frequency, and padded with zeros to at least double its size (so that the new sampling frequency is at least double the desired frequency range). Inverse transforming this signal gives a complex (analytic) time signal whose amplitude is the required envelope. Obtain the amplitude spectrum of this envelope signal and of the squared envelope.
- (4) Search the envelope spectrum for the suspected bearing fault frequencies (and sideband patterns) using a harmonic/sideband cursor, and make conclusions about the fault(s). Analyse the envelope signal of the reference signal by way of comparison.
- (5) Your report should identify which of the three faulty bearings corresponds to which fault, giving your reasons, and should compare and discuss the results for the envelope and the squared envelope, and for the effect of choice of band, including whether the comparison should be made on linear or logarithmic amplitude scales.

2. Gear Diagnostics

This assignment uses demodulation analysis to give diagnostic information about a gearbox. The Transmission Error (TE) signal is compared with the vibration acceleration signal. There are two records of length 1 000 000 samples, named gearf.mat and gearg.mat, and representing a faulty gear and a good gear in mesh with a good gear, respectively. The faulty gear has a simulated tooth root crack at one position (a machined slot using spark erosion). The sampling frequency is 24 000 samples/s for all recordings. Each record has four signals (one per column) as follows:

Column 1 – vibration acceleration signal (scaling unknown, but the same for both signals)

Column 2 – shaft encoder 1
 Column 3 – shaft encoder 2
 Column 4 – tacho (once per rev)

Both gears have 32 teeth and their shaft speed is approx. 6 Hz although this should be determined accurately using the tacho signal.

First determine the TE signal by subtraction of the torsional vibration signal of one gear from that of the other. The torsional vibration signal (in terms of angular displacement) is obtained by phase demodulating the shaft encoder signal as follows:

Perform the fft of a long section (preferably a power of 2 such as 65 536). By plotting the absolute value of the spectrum (against line number), and using ZOOM and/or GINPUT, determine the encoder pulse frequency which is to be demodulated, and the width of band which contains significant sidebands. Set up a new shifted spectrum of reduced size (say 8192 samples) first filled with zeros, and then with the low positive frequencies taken from the encoder spectrum starting at the carrier frequency and with the significant (complex) samples to the right of it. The corresponding negative frequencies (located just below the sampling frequency at the right hand end of the buffer) are taken from the original spectrum to the left of the carrier frequency. The inverse fft of this spectrum gives a complex time signal, whose unwrapped phase angle is the torsional vibration signal. You may use the Matlab command UNWRAP. Since the gears have the same number of teeth, the TE signal is simply the difference between the two torsional vibration signals. Note that even if the two phase signals wander about a bit, because of slightly varying speed, the TE signal should be very repeatable. If the torsional vibration signals have excessive slope (indicating that the true carrier frequency has not been used) it is best to adjust the demodulation centre frequency accordingly, as this makes the phase unwrapping more reliable.

By plotting the TE signal over about four periods, comment on the likely position of the crack. The position of the crack is to be specified as a percentage of the spacing between successive tacho pulses.

To help locate the crack, the toothmesh frequency of the TE signal is next demodulated using an almost identical process (except for the numbers of samples involved). In this case both the amplitude and phase modulation signals will be of interest.

The results obtained by analysing the TE signal are to be checked by analysing the acceleration signal in the same way. Both the raw signal and a synchronously averaged acceleration signal should be used for comparison. The toothmesh frequency of the raw signal (or a higher harmonic if it is stronger) can be demodulated in the same way as the TE signal. To facilitate this, the raw signal can first be decimated to the same sampling frequency as the TE signal. Note that the Matlab function DECIMATE uses appropriate lowpass filtering before downsampling. Before synchronous averaging it will be necessary to resample each section to be averaged (at the original sampling frequency) to a fixed (greater) number of samples (4096 suggested). This will be done to the nearest sample by detecting the first sample in each period where the tacho signal exceeds a specified value. The FIND function can be used for this, although only the first sample in each group satisfying the trigger condition is to be retained. The DIFF function can then be used to determine the number of samples in each rotation (which is to be resampled to 4096 using INTERPFT). The synchronous average can be obtained using the form:

$$\text{xav} = \text{mean}(\text{x}, 2);$$

where x is a rectangular matrix containing the vectors (of length 4096) to be averaged, as columns, (the parameter 2 indicates that the averaging is to be done over the columns instead of the normal rows). The result will of course only contain one period of rotation, but in addition to seeing

whether the position of the crack is now clearer, this signal should also be demodulated around the toothmesh frequency (or harmonic) to see if this gives a clearer result.

Your report should compare all the methods used.

3. Engine Diagnostics

This assignment uses three groups of signals denoted engnorm.mat, engfault1.mat and engfault2.mat. The first represents the engine running in normal condition, and the other two the engine running with misfire(s) on some cylinder(s). Each group has three signals as follows:

Column 1 – acceleration signal from the engine block near cyl.6

Column 2 – once-per-cycle tacho pulse from the distributor (cyl. 1 firing TDC)

Column 3 – (shaft encoder) signal detecting passage of the ring gear teeth

There are 157 teeth on the ring gear, giving 314 pulses per cycle for the 4-stroke motor. The sampling frequency is 24 000 Hz.

Since the engine speed fluctuates somewhat, in particular for the two fault conditions, it is best to work with shorter sections of the signals, comprising a small number of cycles (8192 samples suggested). If a phase demodulation is made of the whole record, the constant speed sections will have constant mean slope. If the section chosen has significant slope, the carrier frequency should be adjusted accordingly when demodulating the shorter section. The average engine speed was nominally 1500 rpm, but varied between the different recordings. It should be determined by analysing each tacho signal.

The first thing to be generated is the torsional vibration signal of the crankshaft, generated by phase demodulation of the encoder signal, which is then differentiated to angular velocity. The phase demodulation is done in basically the same way as for the encoder signals of Part 2, i.e. fft transformation to a complex spectrum, whose amplitude is viewed against line number to determine the carrier frequency and bandwidth to be demodulated. The spectrum to be inverse transformed can for example have 2048 frequency lines total, with 256 each of positive and negative frequency components inserted. The differentiation of the (real) phase modulation (PM) signal is best done by a $j\omega$ operation in the frequency domain. In this way a bandpass filtered differentiated signal can be formed in one operation. First, an fft operation is performed on the PM signal, and its log amplitude spectrum viewed to determine the upper frequency where the periodic (i.e. discrete frequency) components approach the base noise level in the spectrum (the upper noise limit). At the same time it can be checked which line number the engine cycle frequency corresponds to. Frequency components below half the cycle frequency and above the upper noise limit should be set to zero. It is best to set all negative frequency components to zero at the same time, making an analytic signal whose real part will be taken as the final answer. The non-zero components should be multiplied by $j*\omega$, the radian frequency corresponding to each line. For an unscaled result, ω can be taken as (line number – 1). The angular velocity of the crankshaft is thus the real part of the ifft of this one-sided complex spectrum.

One of the fault signals has a misfire on one cylinder. The other has two ignition cables swapped giving a misfire on two cylinders. Determine which case corresponds to which fault signal, and which cylinders are misfiring, given that the firing order is 1-5-3-6-2-4.

By way of comparison, generate synchronously averaged accelerometer signals (over the whole record) in the same way as described for Part 2, taking two engine cycles as a basic period. In this way, the effectiveness of the averaging can be seen from the similarity of the two cycles in the basic period. Having done this check, the two individual cycles (i.e. the two half records) can be averaged together to gain a small further improvement.

Compare and discuss the results obtained from torsional vibration and acceleration.

Matlab code generated should be listed in an Appendix.

A.9 Prognostics

A.9.1 Tutorial and Exam questions

1. Estimates of remaining useful life (RUL) can often be made by trending changes in vibration parameters. Discuss the choice of parameter for simple situations such as developing unbalance, misalignment, etc., and what amount of change can be considered significant and/or serious. Give the background for your recommendations.
2. Discuss the choice of the type of curve that should be fitted to the trend data, e.g. linear, exponential, polynomial, and a suitable basis on which this choice can be made.
3. Explain how the cepstrum is sometimes useful as a trend parameter.
4. Trend parameters for faults in rolling element bearings sometimes first trend up, and then down. Give a physical explanation for this phenomenon.
5. Discuss the use of spectral kurtosis (SK) as a trend parameter for bearing and gear faults. The SK is a dimensionless number; does it directly indicate severity? If not, what factors should be taken into account to make the SK values comparable. How is the value of SK affected by the speed of a machine? What can be done to improve the validity for high speed machines?
6. Explain the difference between physics-based and data-driven approaches to prognostics. If fault simulation is used to produce data, which approach does it fall into?
7. Describe the use of hybrid prognostics models to reduce the uncertainty of prediction of RUL as wear proceeds, and measurement data becomes available to give a measure of current condition of individual units.

Index

- Accelerometers, 14–16
application, 4–10, 17, 18, 38, 51, 56–59, 127, 208, 280, 289, 320, 322, 335
applied to engines, 56–9, 297, 298, 302, 303, 342–9
DC, 10
Acoustic emission (AE), 4–5
Adaptive noise cancellation (ANC), 178, 183
Akaike information criterion (AIC), 183
Aliasing, 5, 84–6, 158–9, 163, 271
Amplitude demodulation, 11, 46, 99, 111, 169, 208, 270, 291
Amplitude modulation, 29, 40, 43, 94–9, 111, 152, 163, 170–1, 223–4, 232, 238, 244, 256, 268, 285, 291, 380
Analytic signal, 68, 94–9, 168, 171, 186, 207, 267, 271
Annulus gear, 236–8, 268
Anti-aliasing filter, 150
Archard’s (wear) law, 386–7
Artificial neural network (ANN), 106, 146, 296, 302–3, 309, 341, 351, 361, 375–6, 379–82
Autocorrelation function, 29, 76, 82–3, 112–9, 181, 188, 199, 203, 208–9
angle-time, 118
Autoregressive (AR) model, 181
Autospectrum, 82–3, 186, 199, 214

Backward whirl, 31
Ball/roller spin frequency (BSF), 47–8, 278, 285, 291, 325, 417
Ballpass frequency, inner race (BPFI), 47–8, 116, 155, 219, 227–8, 275, 278, 280–5, 291–5

Ballpass frequency, outer race (BPFO), 47–8, 208, 275, 278–90, 359–61
Base strain, 16
Bent shaft, 30, 34
Big data, 376, 389–90
Bladed machines, 50, 208–10
Blind source separation (BSS), 23, 66, 189, 224, 269, 303, 388–9
Breathing crack, 35–6
Brinelling, 50
Bucket wheel excavators, 268

Cardan joint. *See* Hooke’s joint
Carson’s rule, 151–2, 161
Case Western Reserve University (CWRU) bearing data set, 375–6, 380–1
Causal function, 93, 204
Causal signal processing, 7, 170
causal filters, 7, 103
Cavitation, 25, 37, 91, 351, 375
Cepstrum, 66, 200
analytic cepstrum, 207
applied to bearings, 50, 202, 209, 233
applied to engines, 217
applied to gears, 202, 207–13
applied to prognostics, 234, 242
applied to separation of forcing and transfer functions, 199, 202, 214–6
obtain complex cepstrum from real cepstrum, 216–7
differential, 203–4
editing time signals using real cepstrum, 216–25
practical considerations, 205–8

- Cepstrum analysis, 51, 178, 199–229
 Cepstrum prewhitening, 225–8
 Characteristic functions (moments and cumulants), 65–6
 Charge amplifier, 14–18
 Cohen’s class, 104
 Complex wavelet, 105–8
 Constant percentage bandwidth (CPB), 92, 101–2, 105, 128–36, 195, 355–7
 spectrum comparison, 128–34
 Contact ratio (CR), 42–3, 267, 321
 Convolution, 20–1, 66, 73–83, 86, 88, 90–3, 99, 102, 105, 113, 148, 150–1, 156, 163, 178, 186, 200, 202, 271, 333
 Convolution theorem, 69, 76–84, 93
 Couple unbalance, 33
 Craig Bampton model reduction, 329–33
 Crest factor, 226, 361, 363–4
 Crownning, 41, 264, 267
 Cubic spline, 109, 150
 Cumulant, 64, 66, 191, 296, 361
 Cumulative damage model, 378
 Cyclo-non-stationary signals, 25–7, 30, 222, 289
 analysis, 116–9
 Cyclostationary analysis, 111–6, 370
 Cyclostationary signals, 23, 25–30, 51, 178, 267–9, 273–4, 295
 bearings, 48, 50, 178, 273–5, 285
 Cylinder pressure, 59, 146, 303–4, 340–1, 344, 348
 reconstruction, 295, 297–303
- Damping matrix, 314
 Data analytics, 379, 389
 Data logger (Data collector), 8, 390
 Data windows, 87–90
 Data-driven model, 355, 372, 375–8, 380–1
 Deconvolution, 202,
 Delta function, 73–4, 80–1, 112, 200,
 convolution with, 76–7, 80
 in a spectrum, 78, 80–1, 86, 92, 401
 Delta function train, 76, 80–2, 148, 178
 Demodulation, 96–103, 185
 Deterministic signals, 26, 28, 178, 286
 Diesel engines, 20, 55–57, 59, 105, 135, 137, 140–1, 295–8, 303, 338, 340–2, 376–7
 Digital filter, 84, 101–4, 271
 FIR, 101–3
 IIR, 101–3
 Digital filtering, 101–04
 Digital twins, 380, 387, 389
 Discrete Fourier transform (DFT), 41–2, 70–3, 83–5, 90–2, 186
 Discrete/Random separation (DRS), 114, 178, 185–7, 226–8, 272, 275–7, 282, 285, 291
 Dry friction whip, 38
 Dry friction whirl, 38
 Dual vibration probes, 18
 Duhamel integral, 74
 Dynamic range, 11, 14, 33, 56, 86, 99, 128, 130, 143, 151, 206
 accelerometers, 16–18
 transducers, 10–12, 18
 Dynamic stress, 356
- Echo delay time, 199, 254, 265–6
 Echoes, 199–201, 254, 262–4
 use of the cepstrum, 201–2, 262
 Electrical machines, 34, 51–5, 309
 Electrical runout, *See* Runout
 Empirical mode decomposition (EMD), 108–11, 109–11
 EEMD, 110–11
 Energy operators, 163, 168–70
 frequency domain (FDEO), 169, 172
 frequency weighted (FWEO), 170
 Teager Kaiser (TKEO), *See* Teager Kaiser energy operator (TKEO)
 Energy spectral density (ESD), 92–3
 Envelope analysis, 48, 50, 95, 104, 187, 193, 202, 208, 218, 223, 268, 270–8, 282, 291, 303, 336, 360, 375
 Environmental stressors, 377
 Epicyclic gearbox. *See* Planetary gearbox
 Ergodic, 26, 29, 83
 Extended spalls, 273–6, 370
- Fast Fourier transform (FFT), 7, 53, 72–3, 128, 333
 zoom, 83–4, 88, 91–2, 99, 101, 159, 205–7, 231–5, 240–1, 257
 Fault detection, 3, 123–46, 355, 381
 Fault diagnosis, 3, 109, 208, 360, 375
 Fault prognosis, 3, 214, 263, 316, 334, 362
 Fault simulation, 309–53, 372, 376–7
 bearing knock (IC engines), 351–3
 bearings, rolling element (local and extended faults), 324–38
 engines (diesel and spark ignition), 144, 338–53
 gears (parallel and planetary), 310–24
 misfire (IC engines), 338–47
 piston slap (IC engines), 347–351
 Ferrography, 5
 Figures of merit, 236, 361

- Finite element model, 36, 224, 263, 310
model reduction, 329–33
model updating, 36, 309, 331, 335
- Flat top window, 88–9
- Flexible couplings, 19, 32, 311
- Fluid film bearing, 4, 10–11, 31, 34, 37, 128, 309
- Forcing function, 2, 21–2, 38, 73–4, 146, 172, 181, 199, 214, 216, 223–5, 254, 260–2, 268–9, 285, 297, 314, 329, 388–9
- Fourier integral transform, 68–9, 71, 84
- Fourier series, 67–71, 76, 78, 80–82, 91, 112–3, 118, 200
- Fourier transform, *See* Fourier integral transform
- Four-stroke engine, 55, 344
- Frequency demodulation, 20, 98, 104, 140, 143, 168, 241, 303, 342–3
- Frequency modulation (FM), 43, 94–8, 128, 168, 170, 218, 220, 223, 240, 244, 256, 268, 285, 292, 356
- Frequency range, 4, 7, 59, 83, 99, 101, 106, 123–4, 129, 134, 148, 150–1, 159, 172, 195, 206, 208–13, 219, 231–3, 257, 283–5, 302–3, 311, 325, 327–336, 343–4, 356, 389
- accelerometers, 5, 15–17,
bearings, 22, 49–51
transducers, 4, 10, 12–13
- Frequency response function (FRF), 20–1, 69, 79–81, 103, 172–3, 186, 203–4, 214, 269, 298, 315, 333, 339, 349–352
- Fundamental train frequency (FTF), 47–8, 227–8, 275, 291
- Gas turbines, 4, 6, 16, 28, 33, 49, 189, 214, 277, 280, 363
- Gaussian
distribution, 64–5, 284
random signal, 63, 66, 91, 189, 191, 285, 304, 361
window, 104, 106–7
- Gear couplings, 32–3
- Gearbox, 19, 33, 50, 94, 106, 111, 130–3, 148, 152, 172, 176, 376
- Gears, 32
diagnostics, 50, 105–6, 108–11, 134, 178, 186, 188–9, 202, 210–6, 221–5, 231–269, 310, 372, 379
varying speed and load, 116, 154, 159–63, 172, 179–80
vibrations, 3–4, 19–23, 25, 39–47, 94, 118, 128, 356–64
- General machinery criterion chart, 124–5
- General path model, 378
- Ghost components, 44–7, 257
- Gyroscopic effects, 31, 311
- Hanning weighting/window, 21, 56, 87–92, 104, 157, 186, 206, 237
- Harmonic cursor, 23, 210, 228, 231–5
application to determination of numbers of gearteeth, 233–5
combination with order tracking, 154, 232–5
- Harmonic wavelets, 131–2
- Helicopter gearbox, 186–7, 226, 236, 275, 278–80, 362, 365, 376
- Hertzian deformation, 40, 318, 325, 349, 363
- Hilbert transform, 68, 93–4, 99, 106–7, 144, 159, 168–70, 204, 207, 238, 270
- Hooke’s joint, 32–3
- Hot box detectors, 6
- Hunting tooth, 210, 221, 232, 235, 249
- Hybrid models, 329, 377–80
- Hydrodynamic bearings. *See* Fluid film bearings,
- Hydroelectric power plants, 3
- Hysteresis whirl, 37–8
- IC engines, 4, 26, 55, 309, 338, 342–3, 347–8, 380–1, 388
- Imbalance, *See* Unbalance
- Impulse response functions, 69, 187
- Impulse wavelets, 108–9
- Impulsiveness, 44, 50, 189, 192, 281, 287
trending of, 361–4
- Induction generator, 54
doubly fed, 55
- Induction motor, 52–4, 133, 232, 284–5, 322, 375
- Industry 4.0, 1, 390
- Inertial torque, 302
- Instantaneous machine speed, 32, 117, 139, 156, 163–177
PDF method, 176–7
- Intermittent monitoring, 6–9, 19
- Internet of Things (IoT), 1
Industrial (IIoT), 1
- Interpolation, 100, 148–51, 182–3, 302, 357
- Intrinsic mode functions (IMFs), 108–11
extraction, 109
- Inverse filter, 187–8, 287, 297–302, 388
- Inverted echo pairs, 254, 262–4
- ISO 10816 criteria, 124, 135, 355
- ISO 20816 criteria, 124–7, 135, 355
- ISO 2372 criteria, 14, 124, 126, 355
- Jeffcott rotor, 31
- Journal bearing. *See* Fluid film bearing,

- Kaiser-Bessel window, 88–9
 Kurtogram, 189, 193–4
 fast, 193, 195–6, 268, 278, 287, 294, 348
 wavelet, *See* Wavelet kurtogram
 Kurtosis, 65–6, 107, 187–93, 226, 236–7, 268, 273,
 278, 282–5, 294, 304, 361–4
 spectral, *See* Spectral kurtosis
- Laplace transform, 69, 73, 80, 102, 314
 Laser vibrometers, 18–19
 scanning, 19
 torsional, 20, 59, 303, 338–9
 Laval rotor. *See* Jeffcott rotor
 Leakage (windowing effect), 84, 86, 91–2
 Levinson-Durban recursion (LDR) algorithm, 181
 Lifter, 202, 216, 216, 221, 388
 exponential, 172, 176, 214–5, 223, 225, 268,
 287–8, 290–3, 388
 comb notch, 216, 218, 220–3, 259, 262, 292–4
 Linear prediction, 102, 178, 180–3, 188, 226,
 236–7, 277–8, 318, 324
 Looseness, 38, 375
 Lubricant analysis, 2–3, 5
 Lumped parameter models (LPMs)
 parallel gears, 310–20
 planetary (epicyclic) gears, 320–4
- Machine tools, 1, 285
 Maintenance
 condition-based, 1, 3, 355, 379
 predictive, 2
 preventive, 1–3
 run-to-break, 2–3
 Mask spectrum, 128, 130, 132
 Mass matrix, 312, 330–1
 Master coordinates, 329–33
 Maximum entropy method (MEM), 181
 Mean differential cepstrum, 204
 Mechanical filter, 16
 Mechanical looseness, *See* Looseness
 Mechanical runout, *See* Runout
 Mechanical signature analysis, 3
 Mesh phasing, 267, 320–1
 Mesh stiffness, 249, 252, 310–1, 316, 320
 with tooth root crack, 249
 Mesh transfer function, 41–3
 MIMO system, 20, 21, 199, 214
 Minimum entropy deconvolution (MED), 187–9,
 277, 287
 Minimum phase, 204
 Misalignment, 19, 30, 32–36, 45, 51–2, 133, 206,
 259, 278, 316, 356, 376
 Misfire, 59, 135, 137, 139, 142, 144–6, 295, 303,
 338–47, 381–4,
 Mobile equipment, 8, 285
 Mobility, 21
 Mode mixing (EMD), 110–11
 Modulation sidebands, 43, 54, 97–9, 151, 153, 183,
 206, 210, 220–1, 247, 257, 280, 282, 356
 Morlet wavelets, 106–9, 195, 367, 369
 Moving cepstrum integral (MCI), 262, 264, 317
 Multi degree-of-freedom (MDOF), 310, 312, 315
 Neural networks, artificial (ANN), 106, 146, 296,
 302–3, 309, 341, 351, 361, 375, 379, 381–2
 Noise bandwidth, 76, 87–8, 92, 178, 186
 Non-causal signal processing, 7, 18, 103–4, 106,
 168, 170, 172, 186
 Non-stationary signals, 26–27, 30, 91
 Normal distribution, *See* Gaussian distribution
 Nyquist frequency, 72, 80, 83–4, 86, 94, 144, 149,
 151, 163, 272, 413
 Nyquist plot, 81
 Oil analysis, 4–6
 online measurement and analysis, 6
 Oil wear debris, 5–6, 278, 362, 372, 378
 Oil whip, 37
 Oil whirl, 37, 128
 One-sided spectrum, 68, 93–4, 99, 106–7, 168, 186,
 207, 257, 271–2
 Operational modal analysis (OMA), 23, 204, 224,
 262, 388
 Order tracking, 20, 87, 98, 116–7, 147–64, 174,
 178–9, 186, 222–4, 228, 232–6, 240, 249,
 267–8, 277, 285, 290, 292–3, 303, 370
 computed order tracking (COT), 147–163
 improvement by iteration, 152–4
 over wide speed range, 156–63
 phase demodulation based, 151–6
 response signal as reference, 154–6
 Parseval's theorem, 92, 226
 Partial misfire, 144–5, 303
 PeakVue method, 4–5
 Performance analysis, 3, 6
 Permanent monitoring, 8–9, 103, 128
 Phase demodulation, 20, 99–100, 143, 151–63,
 239–46, 269, 277, 421
 Phase modulation, 52, 59, 94–9, 144, 151, 153, 163,
 238, 243–5, 420–1
 Phase unwrapping, 96, 99–101, 143, 161, 420
 Phase-locked loop, 147
 Physics-based model, 355, 372, 375–6, 378, 386,
 389

- Picket fence effect, 84, 86–90
Picket fence error, 87–8, 90–1
Piezoelectric transducers, 4, 15–16
Pitchline pitting, 44
Pitfalls of the FFT process, 84–7
Planet carrier, 210, 238, 253, 268, 320, 322
Planet gear, 180, 236–8, 253–4, 267, 278, 280, 320–4, 395
Planetary gearbox, 180, 238, 268
Power cepstrum, 199–200, 202–3
Power spectral density (PSD), 92, 118, 192, 287, 335–7
Pressure torque, 295, 302–3
Prewhiting, 181, 189, 225–8, 367
Probabilistic Neural Networks (PNN), 381–2
Probability distribution/density, 63–5, 153, 176
Profile errors, 43
Prognostics, 1, 4, 123, 146, 355–90
 advanced, 372–387
 simulation-based, 380–387
Proportional damping, 315
Proximity probes, 7–13, 18, 37, 342, 393–4
Pseudo-cyclostationary, 113, 273–5, 370
Pseudo-random signals, 28, 41
Pulse timing, 239–41, 303

Quefrency, 202–6, 214, 220–1, 225, 228, 257, 260, 262–3, 388–9

Rahmonic, 202, 206–12, 216–23, 255–7, 262, 292–3, 359–60
Rail vehicle bearings, 363
Random signals, 25–8, 63, 66, 111, 178, 206, 361
 stationary, 25–6, 28–9, 70, 83, 91–2, 111, 216
Rathbone criterion chart, 123–4
Real-time monitoring and processing, 7, 84, 101, 103, 168, 170, 172, 271
Reciprocating compressors, 6, 55, 125
Reciprocating machines, 19, 123, 135–46
 diagnostics, 214–25, 296
 vibrations, 25, 55–60, 240
Rectangular window, 56, 86–8, 91, 93, 103, 151, 186, 206, 208, 237
Relative vibration, 4, 9–10
Remaining useful life (RUL), 3, 355, 362, 370, 377–8
Residual signal, 109, 218–9, 236–7, 278–9, 379
Reynolds' equation, 10, 36, 351, 386
Rolling element bearings, 2, 4, 7, 25, 31, 111, 128, 178, 309, 356, 372
 diagnostics, 270–295
 vibrations, 46–50, 95, 113, 324–383
Rotor dynamics, 10, 32, 36, 55, 310
Rubbing, 15, 31, 34, 38, 394
Runout, 11–13, 44, 394
 mechanical, 11, 394
 electrical, 11, 394
Runout subtraction, 11

Sample and hold, 148, 150
Sampled time signals, 70
SDOF system, 38, 79–81, 314
 response, 79–81, 114, 117, 204, 315
Seismic probe, 18
Self adaptive noise cancellation (SANC), 178, 183–5
Separation of forcing and transfer functions, 214–6
Separation of gear and bearing signals, 183–5
Separation of spalls and cracks, 263–7, 316–20
Shaft bow, 11
Shaft encoder, 19–20, 59, 140, 147–8, 238–9, 241, 246, 251, 290, 303, 342
 error, 242–3
Shock model, 378
Shock Pulse Method (SPM), 5
Short time Fourier transform (STFT), 55, 104–6, 189, 191, 195, 295–7
Sideband cursor, 23, 54, 231–2, 257
Sign function, 93
SIMO system, 199, 214, 269
Simulated wear
 gear teeth, 387
 IC engines, 386
Simulation models, 23, 36–7, 264, 269, 295, 302, 309–10, 317, 323–4, 327, 333, 338, 342–3, 345, 351, 376–7, 380, 382–3, 388–9
Single degree-of-freedom (SDOF), 38, 79–81, 114, 117, 204, 314–5
Skewness, 65
Slave coordinates, 329–33
Slip frequency, 52–4, 232
Smart factory, 1
Smoothed pseudo Wigner-Ville distribution, 105
Spalls, 4, 44, 50, 187, 210, 263–5, 267, 273–4, 280, 316–8, 324–5, 363, 366–8, 376–7, 380–
 determination of size, 364–74
Spark ignition engines, 55, 59, 142, 145, 342–3, 410, 412
Spectral correlation, 111–7, 274–6, 284, 289, 327, 336, 338, 370, 389
 order-frequency, 119
Spectral kurtosis (SK), 66, 189–94, 226, 268, 272, 287, 356, 358, 361, 372

- Spectrum averaging, 90–1
 Spectrum comparison, 128–37, 192, 325–6, 328, 355,
 Spectrum scaling, 91–2
 Spherical roller bearings, 48
 Stationary signals, 25–7, 29, 87–8, 91, 216
 deterministic, 26, 28
 Stiffness matrix, 312, 314, 330–1
 Subharmonic whirl, 37–8
 Subharmonics, 30, 38
 Subsurface cracking, 364
 Sun gear, 180, 235–8, 253–4, 267, 320–4
 Surface roughness, 386
 Synchronous whirl, 31, 35
- Taper roller bearings, 48
 Teager Kaiser energy operator (TKEO), 163, 168–73, 175
 Thermal bow, 34
 Thermography, 3, 6
 Time synchronous averaging (TSA), 178–9, 218, 220–2, 236–8, 268, 285, 291
 Time/frequency analysis, 27, 104–11, 267, 295
 Time-frequency diagrams, 55–9, 138
 Tip relief, 39
 Tooth root crack, 22, 44, 106, 111, 239, 244, 247–9, 254, 260, 263–4, 267–9, 289–90, 311, 316, 318, 324
 Toothmesh (TM), 20, 22, 45–7, 216, 237, 254, 262, 274
 frequency, 23, 40–1, 43–5, 130, 134, 231–3, 238, 242, 244, 255–7
 harmonics, 207, 236, 240, 246, 256–7, 259, 267
 signal, 188, 238
 Torsional laser vibrometer. *See* Laser vibrometer: torsional
 Torsional vibration, 4, 19–20, 32, 59–60, 94, 98, 135, 139–46, 239–40, 295, 297, 302–3, 338–42, 345–7, 382–3
 Torsional vibration transducers, 19–20
 Transducers, 2, 4, 9–20, 37, 46, 236, 297–8, 390
 permanently mounted, 6–8
 velocity, 13–14
 Transfer function, 2, 41–4, 46, 69, 102–3, 154, 170, 172–3, 181–4, 186, 199, 202, 214–6, 223–4, 244–5, 254–5, 260–2, 268–9, 287, 297, 299, 314, 352, 382–3, 385, 389
 Transmission error (TE), 19, 39–40, 238–254, 264, 316
 absolute TE (total wear), 249–53
 as a diagnostic tool, 238–54
 planetary gears, 253–4
 Transverse crack, 34–5
 Trend analysis, 355–72
 Trending of spall size in bearings, 364–74
 Triboelectric noise, 15
 Turbulence, 25, 51, 91, 375
 Two-stroke engines, 55
- Unbalance, 7, 22, 30–6, 51, 53, 278, 288, 376
 Unit impulse, 73–4, 76
 Universal joint. *See* Hooke's joint
- Variance, 65, 112–3, 183, 380
 VDI 2056 criteria, 14, 124, 126, 355
 Velocity pickup, *See* Transducers, velocity
 Vibration criteria, 123–7, 135–6, 356
 reciprocating machines, 135–6
 rotating machines, 123–7
 Vibration transducers, 7, 9–19
 Virtual analysers, 390
- Wavelet analysis, 105–8, 195
 Wavelet denoising, 106–7
 Wavelet kurtogram, 195, 278–9, 281–2
 Weibull analysis, 377
 Wear debris analysis, 6, 372
 Wear prediction, 387–8
 Wear profile, 43, 387
 Wiebe's functions, 340, 343
 Wigner-Ville distribution (WVD), 104–5, 113, 117, 295–7, 297, 370
 Wigner-Ville spectrum (WVS), 113–5, 117, 136, 295, 370–2
 Wind turbines, 3, 8, 54–5, 111, 163, 176, 208, 220, 222, 233, 236, 246, 257, 267–8, 283, 285, 295, 361, 376, 379–81
 Window effects, 84
- Yates criterion chart, 123–4
 Yule-Walker equations, 181
- Zero padding, 99, 149, 151, 271–2
 Zoom processor, 84, 99, 101, 240